PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Desenvolvimento de um filtro de descritores moleculares geométricos para gerar um ranqueamento em Banco de Dados de ligantes

Christian Vahl Quevedo

Dissertação apresentada como requisito à obtenção do título de Mestre em Ciência da Computação da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Osmar Norberto de Souza

Porto Alegre

2011

Dados Internacionais de Catalogação na Publicação (CIP)

Q5d Quevedo, Christian Vahl
Desenvolvimento de um filtro de descritores moleculares geométricos para gerar um ranqueamento em banco de dados de ligantes / Christian Vahl Quevedo. – Porto Alegre, 2011.
84 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Osmar Norberto de Souza.

1. Informática. 2. Biologia Computacional. Banco de Dados. I. Souza, Osmar Norberto de. II. Título.

CDD 005.74

Ficha Catalográfica elaborada pelo

Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada **"Desenvolvimento de um Filtro de Descritores Moleculares Geométricos para gerar uma Ranqueamento em Banco de Dados de Ligantes"**, apresentada por Christian Vahl Quevedo, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Bioinformática e Modelagem Computacional , aprovada em 22/03/2011 pela Comissão Examinadora:

Prof. Dr. Osmar Norberto de Souza– Orientador — PPGCC/PUCRS

Prof. Dr. Duncan Dubugras Alcoba Ruiz - — PPGCC/PUCRS

Prof. Dr. Hermes Luís Neubauer de Amorim – — ULBRA

Homologada em..16../..11.../..11....., conforme Ata No. 022..... pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

AGRADECIMENTOS

É com extrema satisfação que deixo meus agradecimentos a todas as pessoas que acreditaram no meu trabalho e que foram essenciais para mais esta conquista. Entre estas pessoas faço questão de destacar:

Minha namorada Andréia pela cumplicidade e apoio neste tempo de ausência;

Meus pais, padrinhos, avós e meu irmão por serem essenciais na minha formação humana;

Aos Prof. Duncan e Osmar pela oportunidade de crescimento profissional e pelas pacienciosas e esclarecedoras orientações;

GPIN: Aos amigos Ana, Luciano, Rita, Barros, Peterson, Nelson, Davide e Patrícia pelas trocas de conhecimento;

LABIO: Aos amigos Karina, Dani, Elisa, Anderson, André, Anderson e Renata pelo companheirismo de cursos de bioinformática;

Aos amigos do Cafedas16, que foram grandes pessoas com que tive o prazer de conhecer nestes dois anos de mestrado;

Os meus amigos Castañeda, Michele, Vitor e em especial ao Roger Granada pelo incentivo e compreensão nos momentos de indisponibilidade e ajuda extra.

Um trabalho desta complexidade acaba sendo muito mais complexo para alunos apenas com uma base computacional e com poucas cadeiras no mestrado voltadas para o tema da bioinformática. A dedicação e o conhecimento dos colegas de laboratório foram fundamentais para o desenvolvimento do conteúdo apresentado.

Por fim, agradeço ao PPGCC e a CAPES pela oportunidade, financiando meus estudos.

RESUMO

Bancos de dados de ligantes de acesso público oferecem atualmente mais de 20 milhões ligantes para os usuários. Em contrapartida, a realização de testes *in silico* com esse elevado volume de dados é computacionalmente muito custoso, que vem demandar o desenvolvimento de novas soluções para a redução do número de ligantes a ser testado em seus receptores alvo. No entanto, ainda não há método para efetivamente reduzir esse número elevado em um valor gerenciável, constituindo-se assim, um grande desafio do Planejamento Racional de Fármacos. Este trabalho tem o objetivo de desenvolver uma função heurística para realizar uma triagem virtual com ligantes disponíveis, cuja intenção é selecionar os candidatos mais promissores. A função desenvolvida é baseada na geometria da cavidade do substrato do receptor, filtrando apenas os ligantes compatíveis com esta cavidade considerando as variações 3D do modelo totalmente flexível do receptor. Para testar a eficácia da função proposta foram feitas duas avaliações utilizando como estudo de caso a enzima do *Mycobacterium tuberculosis*, a InhA. Os resultados obtidos deste filtro melhoraram o processo de triagem virtual, descartando a realização dos testes de docagem molecular dos ligantes que não se encaixam na cavidade do substrato do receptor.

Palavras-chave: Bioinformática, Planejamento Racional de Fármacos, Banco de dados de ligantes, Seleção de ligantes.

ABSTRACT

Public databases provide currently over 20 million ligands to users. In contrast, testing *in silico* with such a high volume of data is computationally very expensive, which demands the development of new solutions for reducing the number of ligands to be tested on their target receptors. However, there is no method to effectively reduce that high number in a manageable amount, thus becoming, as a major challenge of rational drug design. This work aims to develop a heuristic function to perform a virtual screening with available ligands, whose intention is to select the most promising candidates. The function is developed based on the geometry of the substrate of the receiver, filtering only the binder compatible with the cavity, derived from a fully flexible model of the receiver. To test the effectiveness of the proposed function a case study with the enzyme of Mycobacterium tuberculosis, InhA, is evaluated. The results of this filter improved the virtual screening using molecular docking, avoiding the testing of ligands that do not fit the substrate of the receptor binding pocket.

Keywords: Rational drug design, databases of ligands, ligand filtering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3D – Tridimensional

BD – Banco de Dados

DM – Dinâmica Molecular

ETH – Etionamida

HBA – *Hydrogen Bond Acceptors* ou aceitadores de ligações de hidrogênio

HBD – *Hydrogen Bond Donors* ou doadores de ligações de hidrogênio

INH – Isoniazida

InhA – enzima 2-trans-enoil ACP(CoA) *Redutase de Mycobacterium tuberculosis*

LogP – coeficiente de partição octanol-água

MTB – *Mycobacterium tuberculosis*

NAR – *Nucleic Acids Research*

NRB – *Number of Rotatable Bonds* ou número de ligações rotacionáveis

OMS – Organização Mundial de Saúde

PDB – *Protein DataBank*

SDF – *Structure Data Format* ou formato de estrutura de dados

SMARTS – *Smiles Arbitrary Target Specification*

SMILES – *Simplified Molecular Input Line Entry Specification* - é uma palavra usada para descrever a natureza e a topologia de estruturas molecular

TCL – Triclosano

Mwt – Peso molecular

SUMÁRIO

1- INTRODUÇÃO

As indústrias farmacêuticas estão sempre procurando reduzir o tempo necessário para desenvolver novos medicamentos e, embora os avanços tecnológicos tenham contribuído para acelerar o processo de descobrimento de novos fármacos, o processo ainda continua sendo demorado, podendo durar de 10 até 15 anos [1], com um custo de, aproximadamente, 1,2 bilhão de dólares [2].

Com a finalidade de reduzir custos e melhorar o tempo necessário para o desenvolvimento de novos fármacos, pesquisadores têm feito o uso de técnicas computacionais para realizar a triagem virtual de novos compostos obtidos de Bancos de Dados (BD) de ligantes, metodologia conhecida na literatura como planejamento de fármacos assistido computacionalmente.

Nesta última década, o uso das técnicas de triagem virtual baseadas em estrutura para o descobrimento de novos fármacos tem alcançado resultados promissores e, assim, vem sendo amplamente utilizadas, embora ainda se encontrem em um estado de refinamento [3]. Estes resultados são obtidos através de pesados investimentos das indústrias farmacêuticas, fomentando o desenvolvimento de diversas abordagens e possibilitando melhorar a qualidade dos compostos candidatos a fármacos [4].

As melhorias da tecnologia computacional, combinadas com avanços na área de genômica estrutural, permitiram um grande aumento do número de estruturas de proteínas e pequenas moléculas disponíveis em BD químico-biológicos [5]. Atualmente, existem diversos BD de acesso público disponibilizando milhões de pequenas moléculas tridimensionais (3D). Um exemplo destes BD é o ZINC, cujo número de compostos disponibilizado é superior a 20 milhões [6]. Um dos principais desafios está, justamente, em manipular a grande quantidade de estruturas químico-biológicas disponibilizadas [7].

1.1 Definição do problema

Uma forma computacional muito utilizada para acelerar o processo de se identificar bons candidatos a fármacos é conseguir selecionar os ligantes que apresentam

uma boa interação molecular com o receptor alvo [8]. Para avaliar se um ligante possui uma boa interação com o receptor é necessário realizar um teste, onde o ligante precisa assumir diversas posições dentro do sítio ativo da molécula receptora. Assim, o tempo computacional necessário para executar estes experimentos é bastante custoso, consumindo em média 1 minuto[1] para cada experimento de um receptor rígido com um ligante.

Sabe-se que as moléculas não são estruturas rígidas, devendo, portanto, ser considerada a flexibilidade do receptor e do ligante. Para o receptor, foram feitas simulações das conformações da proteína, método que é definido como simulação por Dinâmica Molecular (DM). Neste trabalho foram utilizadas duas simulações por DM com 3.100 conformações. Avaliando o custo computacional de uma trajetória com 3.100 conformações realizando os testes para a verificação da interação com todos os ligantes disponibilizados pelo BD ZINC, quantidade superior a 20 milhões de compostos, o tempo necessário para a execução seria de, aproximadamente, 118 mil anos.

Com isso, diversos pesquisadores vêm trabalhando especificamente com o objetivo de reduzir o tempo necessário para a realização do processo de seleção de ligantes, objetivando reduzir a quantidade de ligantes a serem testados em experimentos. Em virtude disto, muitos métodos heurísticos para classificar candidatos a fármacos de forma rápida têm sido criados na última década. Estes métodos, também conhecidos como métodos híbridos, inicialmente servem para filtrar um conjunto de ligantes de todo o BD de forma rápida, reduzindo o número de candidatos a serem testados para uma faixa de aproximadamente 10^3 ligantes. Com uma quantidade menor de candidatos a fármacos, torna-se possível aplicar testes mais rigorosos para identificar os ligantes com maior afinidade com o receptor alvo [9]. No entanto, deve-se considerar que esta elevada redução do conjunto de ligantes tem como consequência uma baixa no nível de confiabilidade dos ligantes eliminados (Figura 1).

Um dos métodos desenvolvidos mais relevantes foi apresentado em Lipinski [10], conhecido na literatura como a “regra dos cinco”, onde foram estipulados parâmetros para

[1] Utilizando um computador com processador Core2Quad, memória de 8GB RAM, SO Linux Fedora 10 e AutoDock 3.0.

quatro características moleculares. Este método avalia as propriedades ADME (absorção, distribuição, metabolismo e excreção) classificando se os ligantes terão sucesso como fármacos nas fases *in vivo*. Embora esse método seja amplamente utilizado, ele pode gerar muitos resultados incorretos [6,11]. O grande problema deste método foi que seu desenvolvimento foi feito considerando apenas as propriedades farmacocinéticas do ligante, desconsiderando os aspectos farmacodinâmicos, o que acaba não garantindo se a molécula é um fármaco, segundo o estudo de [11].

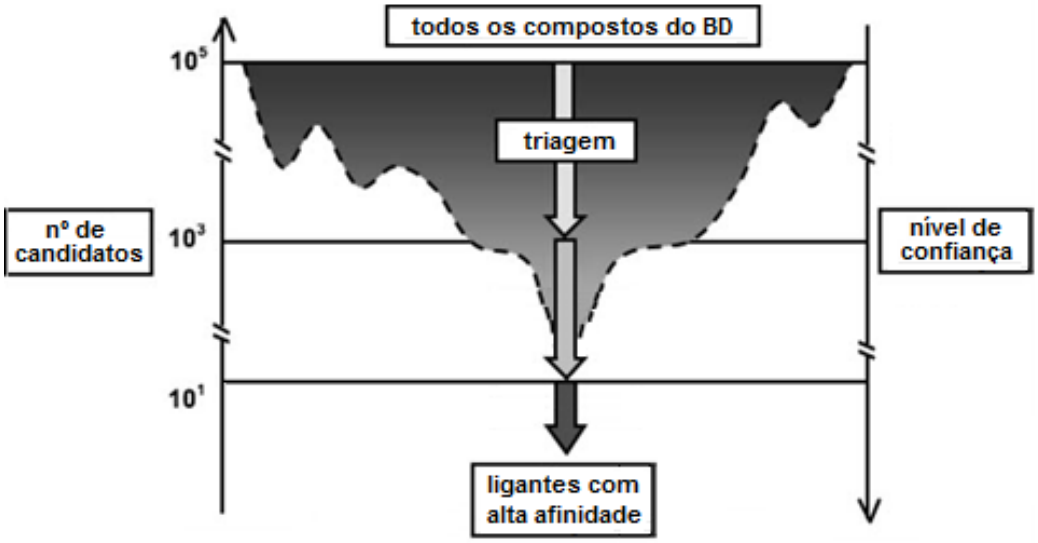


Figura 1 - Existem BD de ligantes capazes de disponibilizar milhões de ligantes e, por ser computacionamente inviável aplicar testes de docagem molecular de um receptor com todos ligantes, diversos métodos híbridos têm surgido, reduzindo de forma rápida a quantidade de ligantes disponíveis, permitindo aplicar métodos mais rigorosos quando há um conjunto menor de ligantes. Adaptada de Amaro and Li, 2010 [12].

O desenvolvimento de métodos híbridos mais eficazes ainda é uma necessidade nos dias de hoje. Este trabalho busca evoluir neste sentido, desenvolvendo uma função heurística capaz de reduzir a quantidade de compostos fornecida pelos BD; utilizando as restrições geométricas da cavidade alvo da molécula pesquisada, contribuindo no cenário do Planejamento Racional de Fármacos.

1.2 Motivação

O tempo computacional elevado, combinado com os altos custos financeiros tem incentivado as indústrias farmacêuticas a desenvolver novos métodos para acelerar o processo de descoberta de novos fármacos. Com isso, pesquisadores têm buscado

desenvolver novos métodos capazes de filtrar essa quantidade de pequenas moléculas a serem testadas *in silico*, fato que se tornou um dos grandes desafios do Planejamento Racional de Fármacos dos últimos anos. Desta forma, desenvolver um método capaz de melhorar a etapa de seleção de ligantes, considerando as propriedades do receptor a ser resolvido na pesquisa, dispensando os ligantes que não estabelecem interação com o receptor, pode gerar um ótimo ganho computacional, fato que demonstra a relevância científica deste tema.

No estudo de caso, pretende-se trabalhar com a enzima do *Mycobacterium tuberculosis* (InhA), que representa um alvo interessante para o desenvolvimento de novos fármacos anti-tuberculose. A tuberculose é considerada uma doença negligenciada por empresas farmacêuticas que não investem no desenvolvimento de novos medicamentos já que esta doença não oferece um retorno lucrativo. Segundo a Organização Mundial de Saúde, foram estimados 9,4 milhões de novos casos de tuberculose em 2009 no mundo [13]. Uma das principais causas apontadas para a ocorrência deste índice é o crescimento do número de casos de pacientes com tuberculose que acabam adquirindo resistência a isoniazida, que é o principal fármaco utilizado no combate desta enzima da InhA [14]. Assim, este trabalho pretende atuar na busca para identificar novas classes de inibidores específicos para esta doença.

1.3 Objetivo e contribuições

O objetivo geral desta pesquisa é o de desenvolver uma função heurística capaz de realizar uma triagem virtual de ligantes disponibilizados em banco de dados de pequenas moléculas e gerar um ranqueamento das moléculas mais promissoras para produzir fármacos para um determinado receptor. Esta função heurística serve como um método híbrido, que apóia pesquisadores descartando ligantes que não possuem características de encaixe no receptor, contribuindo para a aceleração do processo de seleção dos candidatos à docagem molecular de receptores.

Os objetivos específicos são:

- Realizar um estudo aprofundado dos conceitos biológicos necessários para o desenvolvimento desta pesquisa.

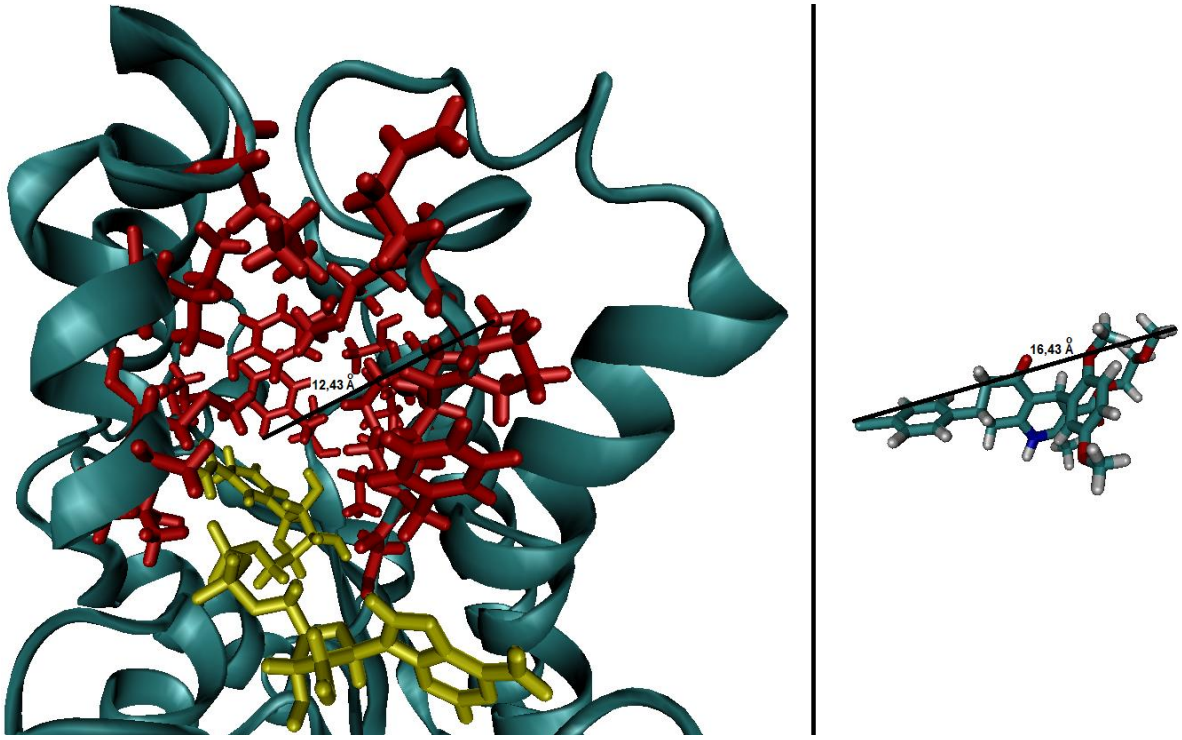
- Verificar a trajetória gerada em Schroeder et al. [14] pesquisar como extrair os arquivos PDB com a co-enzima NADH presente.
- Criar um método capaz de capturar as informações das cavidades existentes na estrutura do receptor em cada conformação da trajetória.
- Gerar uma heurística para determinar qual é a cavidade do substrato durante as variações ocasionadas devido a simulação da flexibilidade da proteína.
- Pesquisar detalhadamente quais os principais BD utilizados pela comunidade científica, verificar quais as principais diferenças entre as informações disponibilizas e como estes BD foram populados, identificando os principais programas envolvidos no processo. Identificar também se os mesmos nomes de compostos disponibilizados possuem as mesmas características.
- Implementar uma função heurística para avaliar as propriedades geométricas da cavidade do substrato de do ligante e, assim, identificar a possibilidade ou não de ocorrer um encaixe entre o ligante e o receptor.

Outro fator a ser destacado é que os sítios dos BD de ligantes mais conhecidos na literatura não apresentam nenhum tipo de filtragem de ligantes que considere as características da cavidade alvo do receptor de interesse. As possibilidades de selecionar ligantes de forma mais específica acaba sendo através de uma pesquisa utilizando a similaridade com um ligante que possui uma boa docagem. Neste caso, somente serão selecionados os candidatos com uma alta similaridade estrutural, não capturando os ligantes não similares que poderiam ser ótimos inibidores.

Para realizar os testes de docagem, o arquivo precisa ser baixado da base de dados para a máquina local. Desta forma, para os pesquisadores testarem toda a base de dados, é necessário baixar todo o conteúdo do BD. Considerando que as informações dos BD de ligantes podem ser alteradas todos os dias, pesquisadores acabam tendo a necessidade de baixar 10^2 GB de dados para manter um banco de dados local sempre atualizado. Então, o desenvolvimento de um filtro para seleção de ligantes capaz de analisar as propriedades geométricas da cavidade de interesse no receptor a ser

resolvido nesta pesquisa pode descartar os ligantes que não tenham possibilidade de encaixe. Um exemplo pode ser visto na Figura 2, onde é apresentada uma comparação entre o maior comprimento de um ligante e a maior distancia da cavidade alvo do receptor em uma determinada conformação.

Baseado no método de chave-fechadura apresentado em [15], o exemplo apresentado na Figura 2 mostra, de forma clara, que o ligante ilustrado na Figura 2b não teria espaço geométrico para encaixar dentro da cavidade exibida na conformação 2 ns do receptor (Figura 2a), sendo descartado da seleção de ligantes para esta conformação. Embora considerado descartado nesta conformação, ainda seria necessário aplicar os testes nas demais conformações para averiguar se o ligante realmente não possui possibilidades de encaixe. Este tipo de comparação foi feita pelo fato deste trabalho considerar as informações de um ligante rígido. Espera-se também que as lições aprendidas neste trabalho, que é uma aplicação não convencional de impacto social importante, possam ser úteis no desenvolvimento de futuros trabalhos interdisciplinares.



(a) (b)

Figura 2 – Comparação entre as maiores distâncias obtidas dentro de uma cavidade do receptor InhA e a de um ligante mostrando não haver a possibilidade de encaixe nesta cavidade do receptor quando o ligante é considerado rígido (imagens geradas com o software VMD). (a) Conformação em 2 ps da proteína InhA obtido da dinâmica molecular [14]. A maior distância apresentada pela cavidade é de 12,43 Å. (b) Ligante obtido do BD ZINC (ZINC2063189) com a maior distância entre os átomos de 16,43 Å.

1.4 Organização do documento

Este trabalho está organizado em sete capítulos. Os problemas enfrentados com uma visão dos seus antecedentes e do cenário atual são abordados nos capítulos 1 e 2. Assim, o primeiro capítulo contém a introdução da dissertação, apresentando o assunto abordado, a caracterização do problema, a motivação e os objetivos esperados, demonstrando a relevância do tema.

O segundo capítulo resgata alguns conceitos fundamentais necessários para um melhor entendimento da área que é o foco deste trabalho, o planejamento racional de fármacos. Este capítulo também destaca características dos métodos de docagem molecular e da simulação por Dinâmica Molecular, responsável por simular a flexibilidade das moléculas. Além disso, no final deste capítulo são apresentados dois formatos de arquivos texto que armazenam as informações das moléculas. A descrição destes formatos é importante para a correta captura das informações moleculares, muito utilizadas nos programas desenvolvidos ao longo do trabalho.

No capítulo seguinte, o Capítulo 3, são abordadas as características dos principais BD de ligantes utilizados pela comunidade científica. Destes, são definidas propriedades necessárias para a escolha do BD mais adequado a ser utilizado neste trabalho. O estudo destas estruturas é de fundamental importância para o entendimento dos sistemas biológicos, assim como para o desenvolvimento mais eficiente de novos fármacos.

O quarto capítulo apresenta a busca pela determinação da estrutura tridimensional da cavidade alvo definida por um especialista. Como já citado anteriormente, a determinação desta estrutura deve contemplar todo modelo flexível. Na literatura encontram-se diversos software capazes de identificar os átomos responsáveis por determinar cavidades no interior da molécula, mas estes software realizam o processamento unitário de arquivos. Este capítulo descreve o programa desenvolvido para submeter todo o modelo flexível de forma automatizada, identificando as cavidades existentes na molécula receptora. Após, é descrita a função heurística criada para identificar e armazenar as informações de todas as cavidades alvo do modelo flexível. Finaliza-se este capítulo apresentando-se as principais dificuldades encontradas no desenvolvimento destas etapas.

Os métodos desenvolvidos na criação da função heurística para realizar a filtragem dos ligantes estão descritos no capítulo cinco. Nele se constrói um conjunto de vetores que contem os volumes atômicos do modelo flexível e dos ligantes distribuídos em esferas concêntricas. A comparação dos comportamentos descritos nos vetores é interpretada pela função heurística, identificando os ligantes que possuem ou não estrutura geométrica para se encaixar em pelo menos uma cavidade alvo do modelo flexível.

O Capítulo 6 traz detalhes dos resultados da validação do estudo de caso testando o encaixe com um conjunto de ligantes extraídos do BD selecionado no segundo capítulo. Estes resultados são classificados e ranqueados conforme o volume do ligante se sobrepõe ao volume das conformações. Finaliza-se esta dissertação com o Capítulo 7, discutindo as considerações finais e as possibilidades de seguimento deste trabalho.

2- PLANEJAMENTO RACIONAL DE FÁRMACOS

Inicialmente, a metodologia empregada no planejamento de fármacos ocorria com o uso de testes *in vitro* de maneira aleatória e, desta forma, resultando em um processo demorado e bastante custoso. Evidentemente, este tipo de pesquisa não apresentava uma relação de custo-benefício adequada aos interesses das empresas farmacêuticas [17]. Com o avanço da ciência, passou-se a investir em procedimentos mais lógicos, metodologia conhecida como Planejamento Racional de Fármacos [18].

O Planejamento Racional de Fármacos é definido como um estudo que trata do reconhecimento de moléculas capazes de ter uma afinidade com determinados receptores. O princípio fundamental deste processo baseia-se na interação molecular [19]. Segundo Kuntz [18], o planejamento racional de fármacos consiste, basicamente, em quatro etapas:

- 1º: deve-se identificar a doença a ser tratada e isolar um alvo específico, determinando este como o receptor a ser tratado. Com a análise da estrutura 3D deste receptor é possível identificar diversas cavidades. Com o auxílio de um especialista, se identifica a cavidade alvo a qual se deseja investigar a ocorrência de interação com ligantes;

- 2º: baseado nas prováveis regiões de ligação identificadas na etapa anterior é selecionado um conjunto de candidatos a ligantes que podem interagir com a região identificada no receptor. As diferentes posições que determinado ligante pode assumir dentro da cavidade alvo do receptor podem ser simuladas por um software de docagem molecular, que classifica com um escore a interação do complexo receptor-ligante;

- 3º: os ligantes que teoricamente obtêm melhores resultados nas simulações são experimentalmente sintetizados e testados;

- 4º: com base nos resultados experimentais, o medicamento é gerado ou o processo retorna à primeira etapa, de maneira iterativa, com pequenas modificações nas características na busca dos ligantes.

Existem muitas formas de solucionar a segunda etapa, selecionando um grupo de ligantes candidatos a fármaco. Dentre elas, existem duas estratégias muito utilizadas e são citadas abaixo:

- Planejamento Racional de Fármacos baseado na estrutura do ligante: esta estratégia costuma ser utilizada quando há o conhecimento da estrutura de um ligante que possui uma boa interação com o receptor. Com este ligante é feita uma pesquisa nos BD de ligantes para identificar outros compostos com propriedades similares a este ligante. O resultado apresentado é uma lista com aqueles que possuem um bom índice de similaridade estrutural. Destes ligantes escolhem-se, por exemplo, os primeiros 1.000 melhores ranqueados, realizando com este pequeno grupo os testes de docagem molecular. Porém, em Martin et al.[20], foram apresentados experimentos que demonstraram que a similaridade estrutural não é garantia de similaridade de função biológica.

- Planejamento Racional de Fármacos baseado na estrutura do receptor: para utilizar esta estratégia é necessário conhecer a estrutura do receptor. Estas informações da estrutura são utilizadas para a avaliação de uma ampla gama de compostos e seleção daqueles que melhor se ligam a cavidade alvo [4]. Estes testes de docagem são realizados visando encontrar o ligante que consiga o melhor encaixe com o receptor. Neste caso, é necessário avaliar a totalidade de ligantes disponível no BD, ou seja: milhões de ligantes a serem testados. Há muitas abordagens distintas baseadas na estrutura do receptor disponíveis na literatura e, um fator determinante a ser considerado, é a capacidade computacional disponível.

Atualmente existem diversas ferramentas computacionais desenvolvidas para auxiliar no processo de descobrimento de novos fármacos. A seguir são descritas as técnicas de docagem molecular e a simulação por dinâmica molecular, que têm como principal objetivo o de quantificar a ocorrência de interação receptor-ligante e simular a flexibilidade do receptor, respectivamente.

2.1 Docagem Molecular

A docagem molecular é o método de simulação computacional que avalia o nível de afinidade entre o receptor e o ligante, onde, para cada teste do ligante com o receptor, é calculado um escore classificando com valores mais baixos quando há uma boa interação [8]. Existe, aproximadamente, 30 programas de docagem molecular disponíveis, podendo estes possuir a licença paga ou sem custos.

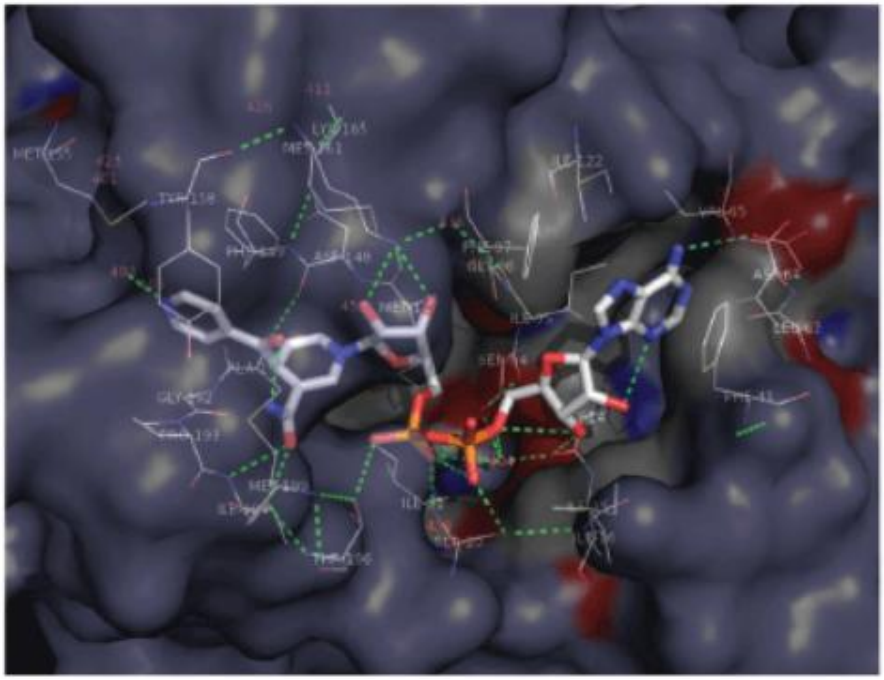


Figura 3 - Docagem molecular feita entre uma conformação da InhA e a co-enzima NADH. As linhas tracejadas verdes demonstram a ocorrencia de interação entre os resíduos da InhA com os átomos do ligante. Figura retirada de Andrade et al.[21].

Para realizar um processo de docagem molecular, o ligante a ser testado assume diversas posições dentro do sítio ativo da molécula receptora e, desta maneira, o tempo computacional necessário para executar estes experimentos é bastante custoso, consumindo em média 1 minuto[2] para cada experimento de um receptor com um ligante. A Figura 3 mostra a melhor posição encontrada pela docagem molecular feita entre uma conformação da InhA e a co-enzima NADH.

[2] Utilizando um computador com processador Core2Quad, memória de 8GB RAM, SO Linux Fedora 10 e AutoDock 3.0.

De modo geral, os programas de docagem molecular consideram a estrutura da molécula do receptor como rígida, enquanto que a molécula do ligante pode ter uma flexibilidade parcial, variando os ângulos de torção. Por considerar a estrutura do receptor rígida, torna-se necessário processar métodos que simulem a flexibilidade da proteína, visto que em meio natural, as moléculas não são estruturas rígidas. O sub-capitulo 2.2 a seguir apresenta a simulação por Dinâmica Molecular.

2.2 Simulação por Dinâmica Molecular

Diversos autores tem comprovado que não considerar a flexibilidade da proteína no processo de seleção de candidatos à fármacos acabam limitando significativamente a precisão dos métodos de docagem molecular [22,23]. Desta forma, com o objetivo de manter boas possibilidades de acerto, este trabalho buscou utilizar a simulação da flexibilidade da proteína definida como estudo de caso.

Na literatura é possível encontrar uma grande quantidade de métodos capazes de simular parte da flexibilidade da proteína. Dentre eles, a simulação por DM é considerada a melhor técnica para obter um conjunto de possíveis conformações de um receptor [24,25].

A enzima InhA é considerada uma molécula flexível e, neste trabalho, foi necessário utilizar duas simulações por DM diferentes. A primeira contendo somente a proteína InhA [14] e a segunda com a proteína InhA interagindo com a co-enzima NADH dentro da cavidade do sítio ativo (estas DM são apresentadas com maiores detalhes no Capítulo 5).

Para simular a flexibilidade da proteína com o método da simulação por DM foi utilizado o software Amber 3.0 [26]. As simulações geradas apresentaram, cada uma, uma trajetória de 3.100 picosegundos (ps) [14]. Uma trajetória com esta duração é considerada significativa, visto que trajetórias relativamente curtas de macromoléculas têm sido utilizadas com razoável sucesso por uma série de comparações experimentais [27].

Por padrão, o software Amber gera arquivos de resultados da simulação a cada 0,5 ps, gerando um total de 6.200 arquivos. Para esta pesquisa foi considerado não haver

necessidade da captura de dados da simulação em um espaço de tempo tão curto. Assim, foram descartados 50% dos arquivos, selecionando os arquivos com intervalos de 1 ps, resultando no total de 3.100 arquivos. Cada um destes arquivos contém as informações das estruturas 3D da macromolécula dispostas no formato PDB. Na próxima seção são detalhados os formatos de arquivos utilizados neste trabalhados.

2.3 Formato dos arquivos PDB e mol2

Tanto as informações 3D do receptor quanto as informações 3D do ligante, precisam estar armazenadas em arquivos para facilitar o acesso aos dados. Além disso, existe a necessidade destas informações serem lidas por diversos sistemas computacionais. Estes fatores exigem o desenvolvimento de padrões para descrever as informações moleculares em arquivos. Com isso, surgiram diversos formatos de arquivo e, nos dias atuais, alguns formatos têm sido amplamente utilizados.

O formato de arquivo PDB é o padrão mais utilizado para descrever estruturas macromoleculares. Este padrão foi criado pelo grupo responsável pelo repositório de dados Protein Data Bank [28]. Este BD disponibiliza estruturas macromoleculares e é amplamente usado pela comunidade científica. Com as estruturas dos ligantes ainda não existe um padrão predominante, havendo muitos formatos como o mol, mol2, sdf, flexibase, etc. A dificuldade deste trabalho em utilizar vários formatos é o tempo de programação necessário para recuperar as informações dos arquivos, portanto este trabalho limitou-se a utilizar um formato para cada tipo de estrutura:

- formato PDB: formato escolhido para a macromolécula;
- formato mol2: formato utilizado para o ligante.

2.3.1 Formato do arquivo PDB

O formato de arquivo PDB é um formato de arquivo texto que descreve as estruturas 3D das macromoléculas, apresentando uma lista dos átomos que compõe a macromolécula e suas respectivas coordenadas atômicas, descrevendo a posição no espaço tridimensional. Este arquivo pode conter um grande cabeçalho para descrever as informações sobre os autores que determinaram esta estrutura da proteína e diversos

detalhes experimentais, tais como: a resolução, a temperatura, o pH, o número de cristais utilizados, etc. Um detalhe importante é que este formato possui posições de caracteres específicas para cada informação, não sendo os dados reconhecidos caso ocorra a modificação das posições. A Figura 4 apresenta parte de um arquivo PDB, onde é possível identificar as coordenadas atômicas de alguns átomos bem como outras características dos átomos.

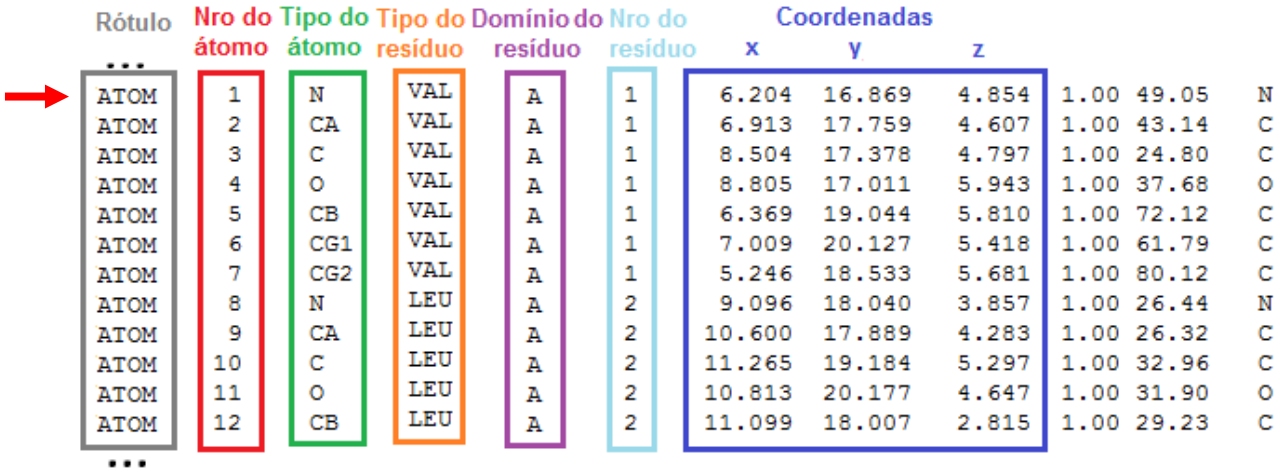


Figura 4 - Parte da estrutura de um arquivo no formato PDB descrevendo as coordenadas 3D de átomos que fazem parte da proteína (código PDB: 4HHB). A linha do primeiro átomo acima descreve as informações de um átomo de nitrogênio presente no resíduo Valina na cadeia A. Os três primeiros números de ponto flutuante são as coordenadas x, y e z e estão em Angströms. As próximas três colunas são a ocupação, o fator temperatura, com o nome do elemento, respectivamente [28].

2.3.2 Formato do arquivo mol2

O arquivo texto mol2 pode ser descrito em dois formatos, sendo utilizado neste trabalho o formato originalmente criado pela Tripos. Este formato, assim como o formato PDB, armazena as informações estruturais de pequenas moléculas como as coordenadas atômicas, ligações entre os átomos, informações das subestruturas e a sua carga parcial. A Figura 5 mostra um exemplo de uma parte de um arquivo mol2.

Um detalhe importante é a possibilidade de um único arquivo mol2 poder conter a descrição de diversos ligantes, fato que ocorre em alguns repositórios de ligantes.

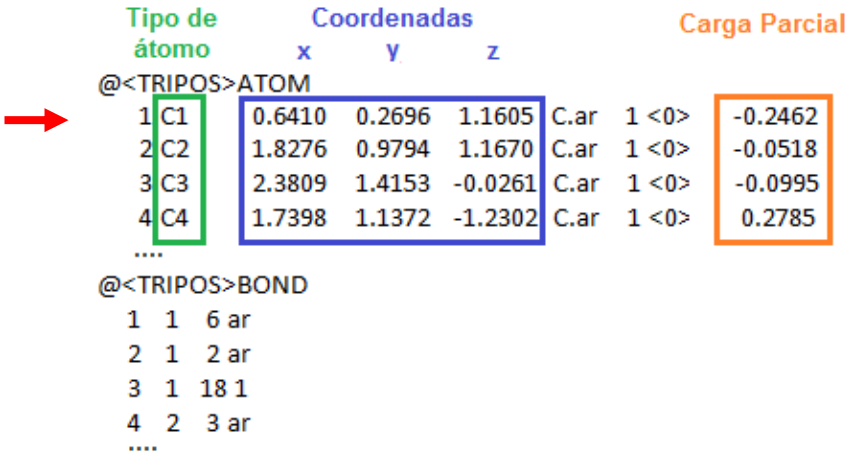


Figura 5 - Parte da estrutura de um arquivo no formato mol2 descrevendo as coordenadas 3D de átomos que fazem parte do ligante triclosano (código ZINC: ZINC00002216). A linha do primeiro átomo descreve as coordenadas atômicas do átomo de carbono com uma carga parcial de -0,2462 [29].

2.4 Considerações finais

Neste capítulo foram apresentados alguns conceitos do processo do Planejamento Racional de Fármacos, auxiliando no entendimento de forma mais precisa sobre as informações utilizadas na solução deste trabalho. Conforme visto na definição feita por Kuntz [18], o escopo deste trabalho está inserido nas duas primeiras etapas, sendo definida uma estrutura alvo para o estudo de caso e a pesquisa visando identificar um conjunto promissor de ligantes candidatos a fármaco.

A molécula considerada no estudo de caso é uma molécula flexível e, pelo fato dos programas de docagem em geral realizarem testes considerando receptores rígidos, tornou-se necessário a aplicação do método de simulação por dinâmica molecular para simular parte da flexibilidade natural da proteína.

Devido à grande quantidade de informações a ser investigada, o uso de aplicações computacionais traz um ganho significativo de processamento. Para utilizar as informações das moléculas de forma automatizada, é necessário compreender como as informações das macromoléculas e dos ligantes estão estruturadas dentro dos arquivos texto. No entanto, há muitos formatos disponíveis na literatura e, para este trabalho, foram selecionados e explanados sobre os formatos PDB e mol2.

3- BANCOS DE DADOS DE LIGANTES

Os bancos de dados de ligantes são repositórios capazes de armazenar as informações de pequenas moléculas. O principal motivo para o surgimento destes BD foi a necessidade dos pesquisadores terem acesso a dados biológicos de forma ágil. Um BD ideal é aquele que possui o maior número possível de informações, provê um acesso fácil às suas informações, fornecem respostas rápidas às requisições e a disponibilização dos dados é feita em formatos que são acessíveis por um grande número de sistemas de computação [30].

Atualmente, existe uma grande quantidade destes BD disponíveis na comunidade científica e esse número está crescendo consideravelmente nos últimos anos, conforme pode ser observado nos volumes anuais divulgado pelo Nucleic Acids Research (NAR), que divulga uma relação destes BD [31]. Embora existam muitos BD de ligantes, todos possuem determinadas particularidades, tornando necessária uma análise criteriosa a fim de compreender quais são as propriedades moleculares armazenadas e a forma como estas informações armazenadas de cada ligante são geradas.

Outra particularidade existente trata da forma de acesso a estes BD, que pode ocorrer de forma pública ou privada. Os BD de ligantes privados possuem licenças com um custo elevado, inviabilizando sua aquisição para muitos grupos de pesquisa [30]. Um exemplo de BD privado é o Cambridge Structural Database, que possui atualmente cerca de 500 mil estruturas armazenadas. Por conseqüência das elevadas licenças, muitos grupos de pesquisa têm feito o uso de BD públicos. Existe um grande número de BD públicos disponíveis e estes, por sua vez, devem ser muito bem avaliados para decidir se a sua utilização atende as necessidades da pesquisa envolvida. Inicialmente foi investigada a qualidade dos BD de ligantes de acesso público disponíveis e, a seguir, são apresentadas algumas avaliações feitas buscando identificar características que auxiliem a elucidar qual o BD mais indicado para o tipo de pesquisa que é o foco deste trabalho.

3.1 Avaliando os BD de ligantes públicos

Devido à grande quantidade de BD de ligantes de acesso público disponíveis, realizar uma leitura criteriosa com todos BD demandaria um tempo consideravelmente longo. Por isso, na busca para definir qual o BD apresenta as características mais indicadas para este trabalho, foram pesquisados na literatura quais os BD mais conceituados pelos grupos de pesquisa. A relação abaixo apresenta os BD de ligantes indicados:

- ChemBank [32];
- ChemDB [33,34];
- MMsINC Database [35];
- NCI Database [36];
- PubChem [37];
- ZINC [30].

A primeira medida criada para avaliar estes BD foi a de desenvolver um estudo comparativo buscando identificar se os dados que estão disponibilizados em um BD possuem as mesmas propriedades que os dados presentes nos demais BD. Para realizar esta análise foram utilizados como exemplo três ligantes considerados como potenciais inibidores da enzima InhA.

Os parâmetros coletados para realizar a comparação são referentes às características armazenadas em comum nos BD, sendo os seguintes descritores moleculares coletados:

- coeficiente de partição (LogP e XLogP);
- número de doadores de ligações de Hidrogênio (HBD);
- número de aceitadores de ligações de Hidrogênio (HBA);
- peso molecular (Mwt);
- número de ligações rotacionáveis (NRB).

Para encontrar as informações deste grupo de ligantes foi necessário acessar as ferramentas de busca de cada sítio dos respectivos BD e utilizar como chave de busca o código SMILES [38] obtido no DrugBank [39]. As Tabelas 1, 2 e 3 apresentam os resultados das buscas feitas para as moléculas isoniazida (INH), triclosano (TCL) e etionamida (ETH).

Tabela 1 - Descrição das características para a molécula isoniazida.

Banco de Dados	LogP	XLogP	HBD	HBA	Mwt	NRB
ChemBank	-0,81	-0,51	2	3	137,139	1
ChemDB.com CHEMISTRY DATABASES	-0,63	-0,82	3	3	137,139	2
MS	-0,31	-	2	3	137,142	-
NCI	-0,81	-0,70	2	3	137,140	1
PubChem Compound	-0,70	-	2	3	137,139	1
ZINC is not commercial	-0,97	-	3	4	137,142	1
	-0,70	-	3	4	137,142	1
	-0,70	-	3	4	137,142	1

Tabela 2 - Descrição das características para a molécula triclosano.

Banco de Dados	LogP	XLogP	HBD	HBA	Mwt	NRB
ChemBank	arquivo em 2D					
ChemDB.com CHEMISTRY DATABASES	4,75	4,96	1	2	289,541	2
MS	5,14	-	1	1	289,545	-
NCI	molécula não disponível					
PubChem Compound	5,00	-	1	2	289,541	2
ZINC is not commercial	5,13	-	1	2	288,545	2
	5,13	-	0	2	288,537	2

Tabela 3 - Descrição das características para a molécula etionamida.

Banco de Dados	LogP	XLogP	HBD	HBA	Mwt	NRB
ChemBank	1,50	1,53	1	2	166,243	2
ChemDB.com CHEMISTRY DATABASES	0,92	0,79	2	1	166,244	2
MS	1,27	-	1	2	166,248	-
NCI	1,52	-	1	1	166.240	2
PubChem Compound	1,10	-	1	1	166,243	2
ZINC is not commercial	1,46	-	2	2	166,249	2
	1,46	-	3	2	167,257	2

Como é possível verificar nas Tabelas 1, 2 e 3, os ligantes avaliados apresentaram diversas variações nas informações fornecidas pelos BD. As principais variações ocorreram no coeficiente de partição, no número de doadores e aceitadores de ligações de hidrogênio e na informação sobre o número de ligações rotacionáveis. Para

compreender a origem destas variações, foram pesquisados os métodos como cada um dos BD foi populado.

Os resultados indicam que diversos BD utilizam os software da empresa *OpenEye* [40] para predizer as suas coordenadas 3D e diferentes métodos para calcular os descritores moleculares. Desta forma, possivelmente esta variação de métodos utilizados pode ser a causa da diferença dos valores encontrados, visto que a estrutura 3D do ligante pode ser idêntica, apenas as coordenadas atômicas em espaços 3D diferentes. Para verificar tal teoria, foram pesquisadas as coordenadas atômicas fornecidas por cada BD e submetidas ao mesmo software para calcular os descritores moleculares dos ligantes. Submeter as moléculas ao software para a avaliação deste processo é temporalmente custoso. Desta maneira, foi realizado um experimento para os mesmos três ligantes utilizados anteriormente e, para cada ligante testado foram capturados os arquivos fornecidos pelos BD em diferentes formatos (mol2, mol e sdf), conforme a disponibilidade do sítio. Estes arquivos foram submetidos ao software Dragon [41] para calcular as mesmas propriedades detalhadas nas três tabelas anteriores.

As Tabelas 4, 5 e 6 apresentam os resultados obtidos do software Dragon, que avaliou os arquivos disponibilizados pelos respectivos BD. O arquivo do ligante triclosano do BD Chembank foi baixado, entretanto, as informações disponíveis estavam em 2D, não sendo o formato adequado para a submissão ao Dragon. Já no BD NCI não foi encontrado o arquivo deste mesmo ligante.

Tabela 4 - Descrição das características da molécula isoniazida no Dragon.

Banco de Dados	LogP	HBD	HBA	Mwt	NRB
ChemBank	-0,81	3	4	137,160	1
ChemDB.com CHEMISTRY DATABASES	-0,81	3	4	137,160	1
MS	-0,81	3	4	137,160	1
NCI	-0,81	3	4	137,160	1
PubChem Compound	-0,81	3	4	137,160	1
ZINC is not commercial	-0,81	3	4	137,160	1

Tabela 5 - Descrição das características da molécula triclosano no Dragon.

Banco de Dados	LogP	HBD	HBA	Mwt	NRB
ChemBank	arquivo em 2D				
ChemDB.com	5,11	1	2	289,540	2
MS	5,11	1	2	289,540	2
NCI	molécula não disponível				
PubChem Compound	5,11	1	2	289,540	2
ZINC is not commercial	5,11	1	2	289,540	2

Tabela 6 - Descrição das características da molécula etionamida no Dragon.

Banco de Dados	LogP	HBD	HBA	Mwt	NRB
ChemBank	1,49	1	2	166,243	2
ChemDB.com	1,53	2	2	166,270	2
MS	1,53	2	2	166,270	2
NCI	2,71	2	3	167,230	2
PubChem Compound	1,53	2	2	166,270	2
ZINC is not commercial	1,53	2	2	166,270	2
	1,42	3	1	167,280	2

Analisando as Tabelas 4, 5 e 6 nota-se que os valores dos descritores moleculares gerados pelo software Dragon foram semelhantes. Analisando todas as informações geradas da Tabela 1 até a Tabela 6 pode-se concluir que as variações encontradas nas três primeiras tabelas ocorrem devido à utilização de software diferentes para a predição dos descritores moleculares, o que parece tornar inadequado a utilização das informações dos descritores moleculares de uma molécula de um BD. Contudo, uma análise mais cuidadosa das diferenças encontradas requer um estudo mais aprofundado, sendo necessário resgatar as fontes bibliográficas que definem os descritores e averiguar se as diferenças são relevantes.

Além da comparação para verificar a origem e o grau de semelhança das informações dispostas nestes BD, ainda é necessário avaliar o quanto as particularidades oferecidas por cada BD podem atender às necessidades da pesquisa. Abaixo é apresentada uma relação com os principais fatores avaliados:

- disponibilidade de moléculas com estruturas 3D. A utilização da técnica de triagem virtual baseada em estrutura baseia-se na utilização de moléculas com a estrutura 3D;
- a quantidade de ligantes disponíveis. Neste trabalho a estrutura dos ligantes é considerada rígida, assim, os BD que armazenam o maior número de ligantes, mesmo que possuam arquivos com variações de um mesmo ligante, tem mais chances de obter resultados promissores;
- testes de similaridade estrutural entre os ligantes. A seleção de ligantes similares pode auxiliar na descoberta da identificação de grupos de ligantes sem possibilidade de encaixar na cavidade alvo;
- permitir baixar os arquivos de todos os ligantes disponibilizados no BD. Desta forma, é possível aplicar testes de forma local, acelerando o processo de pesquisa;
- fornecer moléculas preparadas para executar em programas de docagem molecular. Para validar os testes é necessário executar a docagem molecular, assim o processo é mais ágil quando a molécula já possui as cargas dos seus átomos calculadas.

Com estes fatores definidos, foi feita uma pesquisa bibliográfica e também uma verificação em paralelo das informações atuais disponibilizadas nos sítios dos BD. O resumo desta avaliação está apresentado na Tabela 7.

Analisando a Tabela 7, nota-se que o PubChem é o BD que apresenta a maior quantidade de ligantes disponibilizados e também é o único a informar o volume do ligante (no entanto não há qualquer descrição indicando qual o método utilizado para gerar o volume). O principal problema encontrado neste BD é o fato dos dados não estarem prontos para executar em programas de docagem molecular, gerando um grande problema temporal. Isto impossibilita o uso deste BD neste trabalho, pois para executar os testes de docagem molecular seria necessário descobrir a carga de todos os átomos para cada ligante disponível.

O ZINC é o segundo maior BD em quantidade de ligantes disponíveis. Possui alguns fatores negativos como, por exemplo, não permitir a busca pelo nome da molécula, obrigando seus usuários a desenhar a molécula ou acessar outros BD para encontrar o seu código SMILES [38]. Contudo, os problemas encontrados não inviabilizam a pesquisa, visto que a função heurística não irá utilizar como parâmetro o nome da molécula.

Tabela 7 – Resumo das principais características dos BD de ligantes pesquisados.

	ChemBank	ChemDB	MMsINC	NCI	PubChem	ZINC
Baixar subconjuntos	sim	sim	sim	sim	sim	sim
Baixar todos os dados	sim	sim		sim	sim	sim
Busca por nome	sim	sim	sim	sim	sim	
Busca por código SMILE	sim	sim	sim	sim	sim	sim
Busca por estrutura exata	sim		sim	sim		sim
Busca por subestrutura	sim	sim	sim	sim	sim	sim
Busca por Descritores Moleculares	sim	sim	sim	sim	sim	sim
Busca por similaridade	sim	sim		sim	sim	sim
Prontos para docagem[3]						sim
Volume					sim	
Quantidade de ligantes	4,5 M	5 M	4 M	0,25 M	27 M	20 M
Formatos de arquivos disponíveis	sdf	mol, mol2, mol2h, sdf e PDB	mol, sdf e PDB	mol2 e sdf	ASN.1, XML e sdf	mol2, sdf e flexibase

$M = 10^6$

Os BD ChemBank, ChemDB e MMsINC, possuem algumas características que nem o PubChem e o ZINC atendem, como por exemplo, a pesquisa pelo nome do composto (ZINC) e a pesquisa pela estrutura exata (PubChem). Em compensação, o número de compostos é relativamente baixo quando comparados aos dois BD maiores, variando de 4 até 5 milhões de ligantes. Outros aspectos negativos são que o BD MMsINC não permite baixar todos ligantes disponíveis e o ChemDB (assim como o PubChem), não permite a pesquisa de estrutura exata.

O BD NCI (National Cancer Institute) possui pouco acima de 250 mil estruturas. Embora estes dados sejam de drogas já existentes, o fato desta pesquisa utilizar ligantes rígidos enfatiza muito a diferença da quantidade de estruturas disponibilizadas entre estes BD. Como visto na tabela, os maiores BD possuem um número superior a 27 milhões de compostos, sendo possível afirmar que este BD é bastante restrito. Assim, a quantidade relativamente baixa de dados deste BD acaba sendo um limitador, sendo o seu uso descartado nesta pesquisa.

[3] Os dados fornecidos são considerados como já prontos para executar em programas de docagem quando os arquivos já contiverem os valores das cargas.

Alguns destes BD armazenam diversos arquivos do mesmo ligante, no entanto estes arquivos apresentam variações da composição ou da estrutura. As diferentes estruturas 3D do mesmo ligante são obtidas a partir dos ângulos de rotação. Testar estas diferentes conformações é uma forma de considerar parte da flexibilidade do ligante, necessário diante deste trabalho que considera os ligantes rígidos. Em testes é possível encontrar ligantes que não encaixam em uma conformação e em outra ser aprovada.

Estabelecendo uma relação entre as características de cada BD com as necessidades da pesquisa a ser desenvolvida, todos apresentaram características positivas e negativas. Entretanto, os critérios apresentados sugerem o BD ZINC, que foi escolhido para o desenvolvimento desta pesquisa. A próxima seção apresenta detalhes do BD escolhido.

3.2 Caracterização do BD de ligantes ZINC

O ZINC é um BD publicado em 2005 e que está em constante atualização, estando atualmente na versão 11. Este BD disponibiliza uma quantidade superior a 20 milhões de moléculas [6] e, por ter uma quantidade elevada de dados, foram criados diversos subconjuntos para facilitar o acesso dos pesquisadores a dados mais específicos. Cada subconjunto obedece a uma série de regras de filtragem dos ligantes que disponibiliza. Abaixo são descritos os principais subconjuntos:

- Lead-like: baseado nas regras de Teague et al. [42] para identificar compostos líderes;
- Fragment-like: as regras deste subconjunto são descritas em Carr et al. [43];
- Drug-like: baseado na “regra dos 5” de Lipinski [10], mas possui muitos compostos que são exceções;
- All-purchasable: subconjunto que possui todas as moléculas que podem ser adquiridas de fornecedores;
- Everything: subconjunto com todas as moléculas, até as moléculas que não podem ser adquiridas.

Caso o usuário não encontre um subconjunto que atenda suas necessidades, o ZINC permite que o usuário crie seus próprios subconjuntos para testar os experimentos,

ficando disponível por um período de uma semana. Outra facilidade criada pelo ZINC é que para cada subconjunto estão disponibilizadas as informações sobre o número de moléculas similares, classificando os grupos em 60%, 70%, 80% e 90% de acertos com a similaridade de Tanimoto [44].

Um dos objetivos desta investigação foi de identificar qual a origem dos dados armazenados no ZINC. Os ligantes criados para popular este BD foram gerados através da aplicação de um conjunto de software da empresa OpenEye [40]. Estes software coletam as informações das estruturas 2D, no formato SDF, de fornecedores de pequenas moléculas e, utilizando a predição, entregam as informações das estruturas 3D como resultado. Por último, foi utilizado o software CORINA para calcular as propriedades físico-químicas das moléculas que são o peso molecular, as energias de solvatação polar ou apolar, a área de superfície polar, o número de átomos aceitadores ou doadores de Hidrogênio, o número de ligações rotáveis, o coeficiente de partição e a carga do átomo [44].

Após identificar as propriedades físico-químicas é feita uma seleção descartando os ligantes que apresentem valores acima dos determinados pelos administradores de BD. A Tabela 8 apresenta as principais restrições deste BD.

Tabela 8 - Os ligantes que satifazerem algum dos requisitos desta tabela serão descartados pelo BD ZINC. Existem algumas exceções, por exemplo, para incluir drogas reais que violam estas restrições [30].

Propriedades	**Restrições**
Peso molecular	> 700 g/mol
LogP	> 6 **ou** < - 4
Nº de doadores de H	> 6
Nº de aceitadores de H	> 11
Nº de ligações rotacionáveis	> 15
Tipo de átomo	Qualquer átomo sem ser: H, C, N, O, F, S, P, Cl, Br ou I

O sítio deste BD disponibiliza uma ferramenta de pesquisa *online* para consultar os dados armazenados. Caso encontre a molécula procurada, é possível baixá-la nos

formatos de arquivos mol2, sdf e Flexibase. Se a busca é por um conjunto de moléculas, há uma série de filtros de propriedades físico-químicas para especificar o conjunto. Para realizar uma busca por moléculas similares, é necessário inserir o código ZINC, ou inserir o código SMILES, ou desenhar a molécula no editor molecular. Uma recente atualização da ferramenta de pesquisa permitiu a escolha de alguns valores de similaridade (Figura 6).

Em síntese, este BD armazena as informações sobre as coordenadas atômicas, as cargas atômicas parciais de cada elemento químico, as nove propriedades físico-químicas (ver a Figura 6), dados sobre os fornecedores das moléculas e uma série de atributos de identificação da molécula. Estes atributos são tanto códigos de padrões internacionais para designar as moléculas (SMILES e SMARTS), quanto os códigos do próprio BD para identificar as moléculas.

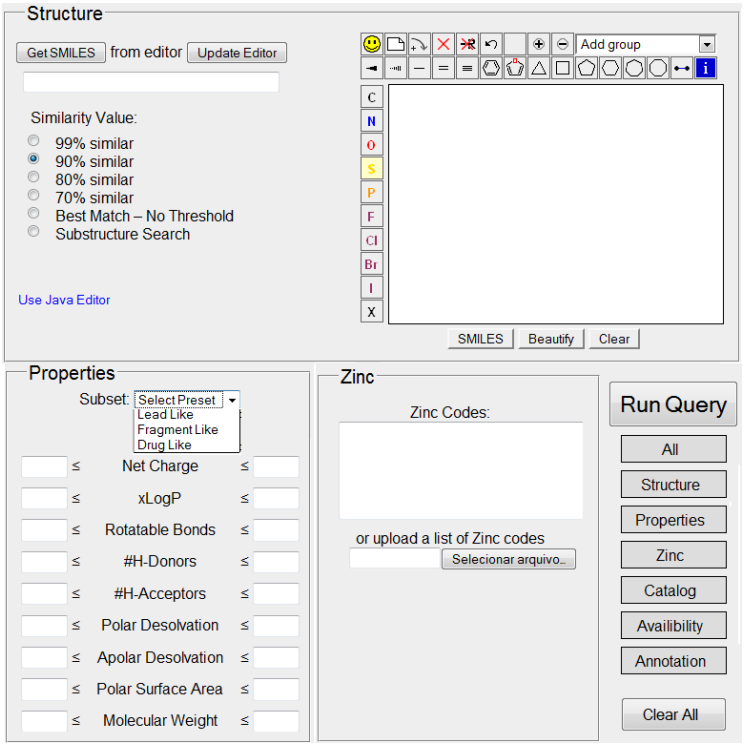


Figura 6 - A ferramenta de busca do BD ZINC permite visualizar algumas propriedades físico-químicas armazenadas por este BD. Pode-se observar também o aplicativo Java que permite desenhar a estrutura molecular. Recentemente esta ferramenta pemitiu a possibilidade de pesquisar diretamente nos subconjuntos e também de escolher o percentil da similaridade de Tanimoto [6].

3.3 Considerações finais

Neste capítulo foi possível identificar as principais particularidades envolvendo os BD de pequenas moléculas. Detalhadamente, foi apresentada uma comparação enunciando os aspectos positivos e negativos das características de cada BD diante das necessidades desta pesquisa. Ao final desta avaliação, ficou constatado que o BD ZINC é o mais indicado para ser utilizado neste trabalho. Como apresentado anteriormente, este BD possui a segunda maior quantidade de pequenas moléculas curadas dentre todos os BD de pequenas moléculas, o que torna este BD muito promissor para se encontrar fármacos. Como um aspecto negativo, possui o fato de não permitir a busca pelo nome da molécula, obrigando seus usuários a desenhar a molécula ou acessar outros BD para encontrar o seu código SMILES. No entanto, este problema não interfere na seleção de ligantes, já que a função heurística não irá utilizar o nome do composto como parâmetro de seleção.

4- DETERMINAÇÃO DAS PROPRIEDADES GEOMÉTRICAS DA CAVIDADE ALVO CONSIDERANDO O MODELO FLEXÍVEL

Neste capítulo é definida a cavidade alvo da molécula em estudo, fazendo uma descrição de a sua importância biológica na busca de novas soluções capazes de inibir esta enzima. Definida a cavidade de interesse, é necessário identificar as suas propriedades geométricas, sendo necessário realizar uma avaliação da estrutura 3D. Esta avaliação da cavidade alvo não pode considerar somente uma única conformação, pois devido à flexibilidade natural da molécula, provavelmente, deve ocorrer alterações na estrutura 3D da cavidade alvo ao longo da trajetória. Portanto, a avaliação desta cavidade deve ser acompanhada durante todo o modelo flexível.

Para fazer a avaliação da estrutura 3D de uma conformação foram pesquisados na literatura alguns dos principais programas capazes de identificar cavidades em uma molécula. Estes programas, além de indicar as cavidades encontradas, também informam os átomos que delimitam estas cavidades. Devido ao fato deste estudo rastrear uma cavidade específica, foi necessário desenvolver métodos computacionais capazes de identificar esta cavidade de forma automática. A dificuldade aumenta quando se considera a flexibilidade da enzima, tornando necessário localizar a cavidade alvo em todas as conformações, o que acaba gerando um grande volume de dados a ser processado e analisado.

4.1 A cavidade alvo: o substrato do complexo InhA–NADH

Inicialmente, esta pesquisa buscou identificar as propriedades da região do sítio ativo da enzima InhA com base nas informações disponibilizadas em Schroeder et al. [30]. Contudo, as primeiras análises das cavidades do sítio ativo apresentaram um grande volume, chegando algumas conformações a apresentar volumes superiores a 1.900 $Å^3$. A Figura 7 apresenta um acompanhamento do volume durante a dinâmica molecular (DM). A fim de avaliar o tamanho desta cavidade em relação aos ligantes disponíveis nos BD, foram pesquisados 20 ligantes com estruturas consideradas volumosas no BD ZINC. Utilizando o Swiss-PDBViewer [16], um software de visualização de moléculas, foi

verificada manualmente a possibilidade de encaixe entre o receptor e os ligantes em algumas conformações da DM. O resultado indicou que somente 2 ligantes não teriam possibilidade de encaixe. Desta forma, concluímos que o uso de uma cavidade com estes volumes não resultariam em soluções satisfatórias, visto que a cavidade alvo apresentou constantemente valores elevados de volume, o que não restringiria muitos ligantes.

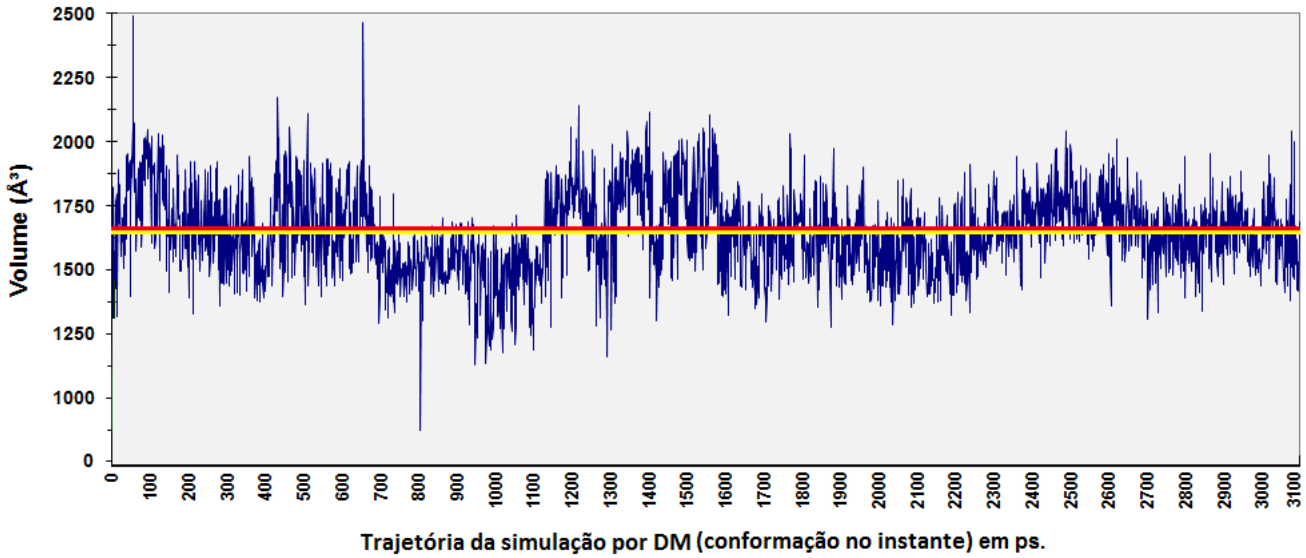


Figura 7 - Análise do volume da cavidade do sítio ativo do modelo flexível da InhA geradas com o CASTp [45]. O volume da cavidade é mostrado em função da conformação no instante ao longo da trajetória da simulação por DM [14]. A média dos volumes obtidos durante toda trajetória foi de 1.647 Å³ (amarelo), valor bastante similar ao volume da estrutura cristalina (código PDB: 1ENY) de 1.657 Å³ (vermelho).

Devido à ocorrência deste problema, buscou-se uma cavidade alvo com volumes inferiores ao sítio ativo da enzima apresentada em [14]. Aliado a esta necessidade, um estudo apresentado em Quémard et al. [46] mostrou que através da formação de um complexo InhA-NADH é possível encontrar ligantes que tornem a enzima InhA inativa. Além disso, também é possível encontrar na literatura casos de potenciais fármacos como isoniazida, etionamida e triclosano que atuam interferindo nas ligações covalentes entre a enzima e o NADH, impedindo a formação do complexo [14]. Desta forma, pesquisar novos métodos focados nesta cavidade do substrato contendo o complexo InhA-NADH pode resultar em conjuntos promissores de candidatos a fármaco.

Em 1999 foi disponibilizada no Protein Data Bank uma estrutura cristalina da InhA com o cofator NADH e um ligante análogo da cavidade do substrato (código PDB: 1BVR) (Figura 8) [47]. As chances de resultados promissores são reduzidas quando somente são consideradas as estruturas cristalográficas, devendo ser feito um acompanhamento da

flexibilidade para identificar as possíveis variações da cavidade do substrato.

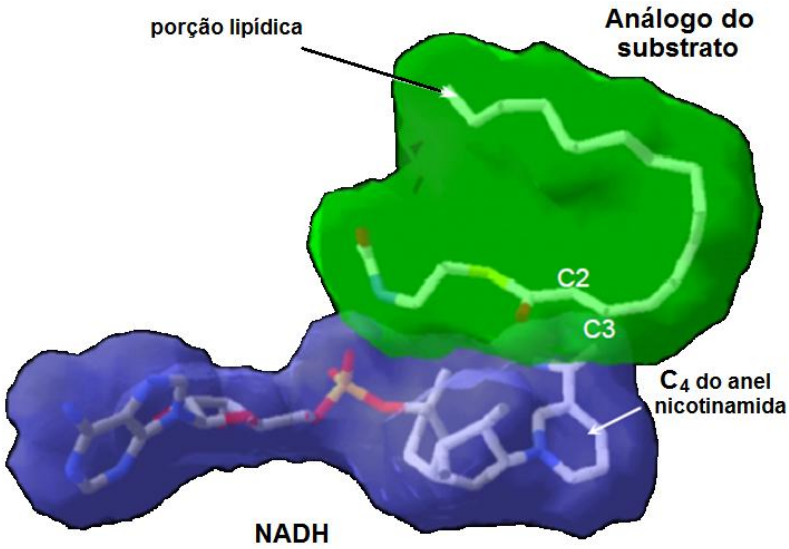


Figura 8 – Cofator NADH e um ligante análogo a cavidade do substrato extraídos da estrutura cristalográfica da 1BVR. Em verde está a região do ligante análogo a cavidade do substrato e, em azul, a região ocupada pelo NADH. Nota-se que a região do substrato é está situada logo acima do anel da nicotinamida do NADH [14].

Para criar os arquivos contendo as informações do complexo InhA-NADH foi necessário executar o PTRAJ (software pertencente ao pacote AMBER) para armazenar as informações da trajetória em arquivos texto no formato PDB [26][4]. Desta forma, foi criada uma simulação de 3.100 ps com intervalos de 1 ps para cada captura, resultando em um total de 3.100 arquivos.

4.2 Identificando os átomos que determinam a estrutura da cavidade alvo em cada conformação

Um grande desafio enfrentado atualmente é a determinação da estrutura 3D da cavidade alvo em uma DM. Descobrir os limites desta cavidade é importante para identificar se as dimensões da cavidade possibilitam ou não o encaixe de um determinado ligante dentro deste espaço 3D. Este passo é fundamental e sua análise deve ser bastante criteriosa para definir as propriedades geométricas das cavidades, já que a

[4] Para realizar este processamento do complexo InhA-NADH feito no PTRAJ contei com o auxílio de colegas do Laboratório LABIO.

determinação desta estrutura de forma errônea poderá acarretar diversos problemas na classificação dos ligantes.

Conforme o funcionamento de alguns algoritmos de docagem é possível que um determinado ligante consiga uma boa docagem em uma única conformação da DM e, mesmo com apenas um resultado, obter uma ótima classificação. Neste sentido, devem-se considerar todos os ligantes que tenham pelo menos uma solução de encaixe com a cavidade alvo, desenvolvendo um processo que avalie todas as variações da cavidade.

Para identificar os átomos que delimitam a cavidade alvo, é necessária uma série de etapas que vai desde encontrar a cavidade alvo corretamente durante a DM até identificar quais átomos determinam esta cavidade em cada conformação. Os subitens a seguir apresentam os principais passos.

4.2.1 Programas que identificam cavidades moleculares

Atualmente existem diversos programas disponíveis na literatura capazes de identificar regiões de cavidades em macromoléculas utilizando critérios baseados na geometria. Entre os programas mais citados estão o CASTp [45], VOIDO [48], APROPOS [49], pvSOAR [50] e o SURFNET [51]. Dentre estes, o CASTp foi definido como o programa a ser utilizado neste trabalho devido a experiência de uso desta ferramenta e pela metodologia abordada. O CASTp é um programa *online* e, devido ao tempo necessário para a comunicação de rede, os processos acabam tornando-se mais demorados que os programas com funcionamento no próprio computador. Entretanto, o fato de disponibilizar o acesso ao código fonte da página dos resultados tornou-se uma considerável vantagem em relação aos demais programas. Esta página de resultados permite o acesso à informação de todas as cavidades de forma bastante estrutura.

Os principais objetivos do CASTp são o de possibilitar os estudos das características da superfície e das regiões funcionais das proteínas. Parte deste processo trata da identificação de todas as cavidades presentes no receptor. Além disso, ele também apresenta todos os átomos que delimitam as cavidades e, por fim, disponibiliza a medidas do volume e da área de cada cavidade [45]. Para encontrar as cavidades, o programa é baseado no método *alpha-shape* desenvolvido por Edelsbrunner and Mücke [52], e utiliza tanto o modelo da superfície acessível ao solvente desenvolvido por Richard

[53], quanto o modelo da superfície molecular desenvolvido por Connolly [54]. A Figura 9 apresenta o comportamento do modelo desenvolvido por Richard.

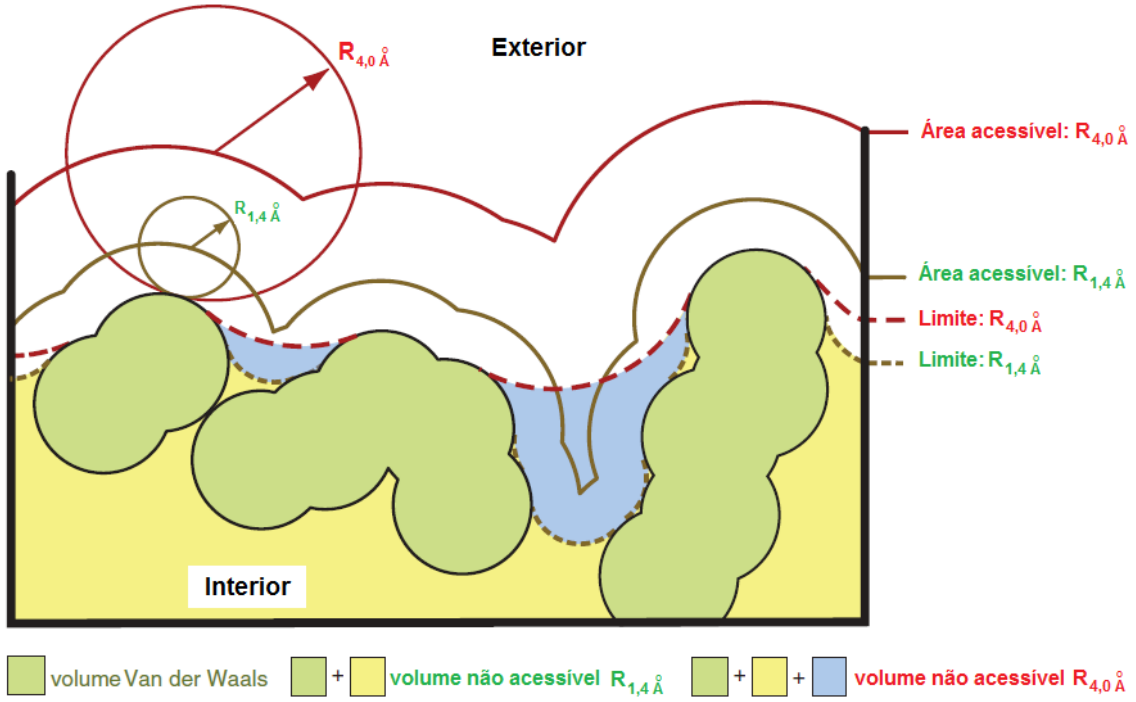


Figura 9 – Exemplo do modelo de superfície acessível ao solvente desenvolvido por Richard. O raio de 1,4 Å representa o raio do solvente (molécula de água). Nota-se que o volume não acessível ao solvente é a área considerada como o volume da molécula. Também é possível perceber que quanto maior o raio de prova, menor será a área acessível a molécula. Adaptado de Voss and Gerstein [55].

O processo para pesquisar cavidades de uma molécula no CASTp é bastante intuitivo. Existem duas possibilidades de submeter o arquivo em um receptor. A primeira é quando o arquivo do receptor está depositado no Protein Data Bank, para isso, basta informar o código PDB e o CASTp irá recuperar os dados diretamente do sítio. A segunda possibilidade é submeter ao sítio um arquivo PDB.

O resultado pode ser gerado de 3 formas:

- visualizador molecular Jmol: é a opção padrão do site para visualizar as cavidades no próprio sítio.
- plugin Chime: é um plugin para navegador semelhante ao item anterior.
- e-mail: são enviados 5 arquivos contendo as informações das cavidades para o e-mail informado e nenhuma visualização aparece no sítio.

A Figura 10 apresenta a interface do CASTp exibindo o resultado da consulta da estrutura cristalográfica (código PDB:1ENY) da enzima InhA utilizando o visualizador Jmol.

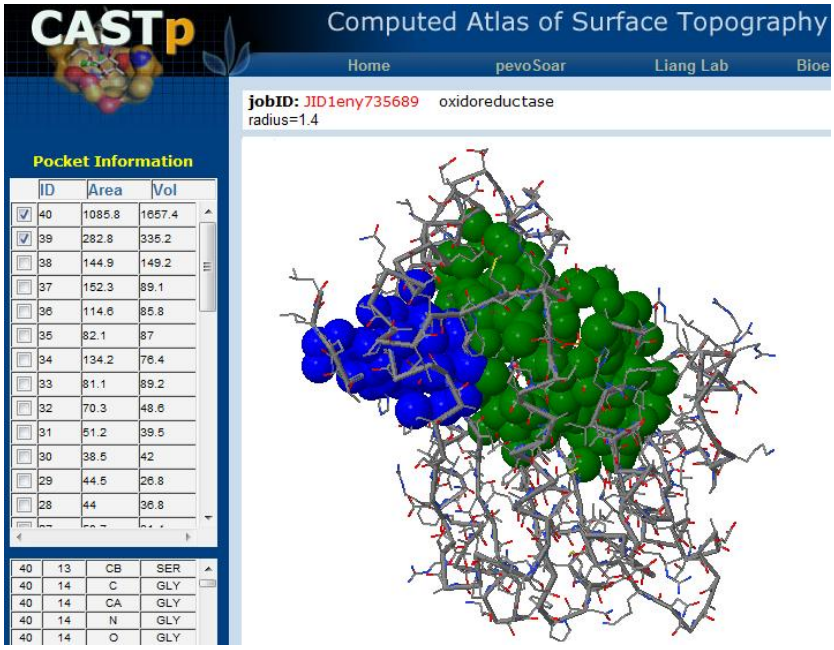


Figura 10 - Tela de resultado da pesquisa do CASTp utilizando o visualizador Jmol para exibir a estrutura 3D da molécula. Na tabela são exibidas as informações do código gerado pelo programa, da área e do volume de cada cavidade. São exibidas duas cavidades da molécula InhA (código PDB: 1ENY), em verde a cavidade 40 e, em azul escuro, a cavidade 39.

O tempo utilizado desde o processo de abertura do sítio e da submissão da molécula, somados com o tempo necessário para o processamento dos cálculos, resultam em aproximadamente 30 segundos. Como visto anteriormente, um receptor flexível pode contabilizar milhares de conformações. Deste modo, realizar esta tarefa manualmente para todas as conformações de um modelo flexível é oneroso, além da grande possibilidade de ocorrer falha humana durante a execução do processo. Além disso, cada conformação pode conter dezenas de átomos delimitando a cavidade.

Com estas adversidades, a solução encontrada para agilizar este processo foi desenvolver um programa para automatizar a submissão de cada arquivo PDB da DM do receptor no CASTp e recuperar as informações geradas.

4.2.2 Desenvolvimento de um algoritmo para a automatização da pesquisa das cavidades moleculares de uma DM no CASTp.

Para automatizar o processo de submissão e recuperação das informações do modelo flexível no sítio do CASTp, foi desenvolvido um conjunto de programas computacionais. Devido a complexidades deste processo, foram utilizadas linguagens de programação Ruby e C. O algoritmo em linguagem Ruby [56] submete os arquivos PDB do modelo flexível ao sítio do CASTp e realiza a recuperação dos dados, salvando o código fonte da página gerada como resultado. O algoritmo em linguagem C pesquisa os dados salvos em formato textual e recupera as informações da cavidade alvo de cada conformação. A seguir é feita uma descrição das etapas desenvolvidas:

- 1º: para cada arquivo PDB deve haver um programa em Ruby, que vai realizar a abertura do navegador, acessar o sítio do CASTp informando o nome do arquivo PDB e o raio de prova a serem submetidos. Após, irá salvar o código fonte do resultado em arquivo texto.

- 2º: para gerar todos os arquivos Ruby do modelo flexível foi criado um programa em linguagem C que pesquisa o diretório contendo a DM e gera os respectivos arquivos Ruby das conformações. A seguir, o programa começa a execução de todos os arquivos Ruby, submetendo os arquivos PDB ao CASTp e gerando os arquivos texto com os resultados da consulta.

Durante o processo de captura dos dados do sítio podem ocorrer problemas que inviabilizam a captura dos dados de forma correta. Deste modo, foi desenvolvido um programa de verificação da corretude dos arquivos texto gerados. Para cada arquivo PDB é feita uma pesquisa do arquivo texto correspondente, testando a inexistência ou a incompletude dos dados. Em ambos os casos, o programa executa o processo de submissão ao sítio do CASTp, armazenando em *log* o diagnóstico dos problemas encontrados.

Com todos os arquivos da DM verificados, começa-se o processo de captura das informações do volume, da área e dos os átomos que delimitam a cavidade alvo. A Figura 11 apresenta a estrutura de um arquivo texto recuperado do CASTp. É possível observar

que as variáveis area_ms, vol_ms e pockets[*][5] são vetores que armazenam as informações das áreas, dos volumes e dos átomos delimitadores das cavidades, respectivamente. Como as informações das cavidades estão dispostas em ordem crescente, podemos concluir que os últimos valores de cada vetor contem as informações da cavidade com maior volume.

Quando os dados referentes à cavidade alvo são encontrados, ocorre o salvamento destes dados em seus respectivos arquivos PDB, atribuindo rótulos que não inviabilizem o uso dos arquivos PDB por outros programas. Então, cada arquivo PDB é aberto e, após a última linha do arquivo é inserida a palavra END. A partir deste ponto, as informações salvas serão as obtidas do CASTp. Em cada arquivo PDB são adicionadas as informações do volume, da área e dos átomos que delimitam a cavidade alvo. Das informações dos átomos delimitadores da cavidade alvo apenas o número da cavidade não é transposto para o arquivo PDB. Na Figura 12 é possível observar as informações geradas pelo CASTp e salvas dentro de um arquivo PDB da DM.

```
<html><head>
<!-- ============================================================================
        Javascript and php script for webpage initialization
============================================================================ -->
<script language="javascript">
var chains = new Array();
var sequences = new Array();
var seqnumref = new Array();
chains[0] = '0';
seqnumref[0] = new Array
('1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18'
'34','35','36','37','38','39','40','41','42','43','44','45','46','47','48','49','
54','255','256','257','258','259','260','261','262','263','264','265','266','267'
sequences[0] = new Array
('A','G','L','L','D','G','K','R','I','L','V','S','G','I','I','T','D','S','S','I',
P','V','A','K','T','V','C','A','L','L','S','D','W','L','P','A','T','T','G','D','I
area_ms = new Array
(17.8,13.2,27.8,25.1,26.9,28.3,27.1,28.7,26.9,9.5,3.7,14.3,27.5,11.0,18.1,32.7,26
,47.5,54.6,52.9,47.3,114.7,45.6,80.9,65.7,85.3,87.3,123.3,266.2,269.4,1118.0);
vol_ms = new Array
(9.0,7.3,12.2,11.8,13.0,12.8,12.7,14.4,12.9,5.1,2.8,8.8,13.3,6.9,7.6,16.6,12.4,6.
3.5,34.0,29.4,64.9,61.8,52.9,58.1,62.8,100.2,213.2,187.0,356.7,1589.4);
annPocketAtoms = new Array();
annPockets = new Array();
annPockets[0]=0;
annResidues = new Array();
pockets = new Array();
pockets[0]=new Array('1 ALA 243 CB 3636','1 ILE 257 CG2 3842','1 LEU 244 CD2 3655
pockets[1]=new Array('2 ASP 149 CA 2234','2 ASP 149 OD1 2240','2 GLU 168 OE2 2534
....
pockets[45]=new Array('46 ALA 190 CB 2868','46 ALA 190 O 2873','46 ALA 197 C 2979
 46 ALA 205 CB 3072','46 ALA 21 N 300','46 ARG 194 CD 2924','46 ARG 42 C
```

Figura 11 - Parte de um arquivo contendo o código fonte da página de resultado do cálculo das cavidades do programa CASTp. As informações sublinhadas representam a área, o volume e a variável que contem os átomos que delimitam a cavidade com maior volume.

[5] Sendo * um valor numérico inteiro maior ou igual a 0.

```
ATOM  4072 N6A NAH  269     0.338 13.032  1.313 0.00 0.00
ATOM  4073 H61 NAH  269     1.046 13.621  0.899 0.00 0.00
ATOM  4074 H62 NAH  269    -0.467 13.449  1.759 0.00 0.00
ATOM  4075 N1A NAH  269     1.776 11.366  0.970 0.00 0.00
ATOM  4076 C2A NAH  269     2.164 10.114  1.203 0.00 0.00
ATOM  4077 H2A NAH  269     3.098  9.823  0.746 0.00 0.00
ATOM  4078 N3A NAH  269     1.538  9.167  1.893 0.00 0.00
ATOM  4079 C4A NAH  269     0.356  9.590  2.429 0.00 0.00
END
AREA_CAVIDADE  469.20
VOLUME_CAVIDADE  582.00
ATOM_CAVIDADE  ALA  197 C  2979
ATOM_CAVIDADE  ALA  197 CB  2975
ATOM_CAVIDADE  ALA  197 O  2980
ATOM_CAVIDADE  GLN  99 CA  1509
ATOM_CAVIDADE  ILE  201 CA  3021
ATOM_CAVIDADE  ILE  201 CB  3023
ATOM_CAVIDADE  ILE  201 CG2  3025
ATOM_CAVIDADE  ILE  201 N  3019
```

Figura 12 - Parte de um arquivo PDB de uma conformação da DM alterado pelo programa de recuperação das informações do programa CASTp gerado neste trabalho. Após as linhas que contém o rótulo “ATOM” são as linhas gravadas pelo programa de recuperação descrevendo a área, o volume e quais são os átomos que o CASTp determinou como sendo os delimitadores da cavidade alvo.

Para finalizar, ainda é necessário encontrar a estrutura da cavidade alvo e extrair as informações de todos os arquivos textos da página de resultado do CASTp e adicionar estas informações nos respectivos arquivos PDB da DM. Análises estatísticas revelam que, na maioria das vezes, a cavidade com o maior volume é a cavidade do sítio ativo. Além disso, estudos empíricos têm comprovado a ocorrência do sítio de ligação pertencendo à maior cavidade encontrada da proteína [57,58,59]. No entanto, já existem alguns casos de exceções relatadas em Liang et al. [60].

Visando observar o comportamento da cavidade do sítio ativo, foram capturadas as informações do sítio ativo da DM apresentada em Schroeder et al. [14]. Aproximadamente 150 conformações foram testadas manualmente e, para todos os casos testados, a proteína InhA apresentou sempre a cavidade com o maior volume como sendo a cavidade do sítio ativo. A Figura 7, que apresenta o volume da cavidade do substrato, foi gerada a partir desta avaliação.

Embora os testes tenham comprovado a eficácia em determinar a cavidade do sítio ativo, a mesma teoria não foi válida quando a pesquisa foi feita no modelo flexível buscando identificar a cavidade do substrato. Neste caso, a teoria para este

comportamento está no fato de que a cavidade do substrato é apenas uma parte da região do sítio ativo da InhA. Desta forma, buscar somente a cavidade com o maior volume de cada conformação foi um erro. Para conseguir identificar a cavidade do substrato foi necessário pesquisar uma estrutura cristalográfica e desenvolver uma heurística específica para este problema.

4.2.3 Heurística desenvolvida para identificar a cavidade do substrato durante a simulação da DM

No sub-capítulo 4.1 foi apresentada a estrutura cristalográfica da InhA com o cofator NADH e um análogo do substrato (código PDB: 1BVR) [47]. Este análogo do substrato identifica o sítio de ligação de interesse desta pesquisa. Portanto, para desenvolver uma função heurística capaz de identificar a cavidade do substrato durante a simulação da flexibilidade, foram pesquisados os resíduos que delimitam a fronteira com o análogo do substrato da estrutura cristalográfica 1BVR. Assim, o arquivo PDB foi editado sendo retirados os dados referentes ao ligante análogo do substrato. Este arquivo editado foi submetido ao CASTp e todos os resíduos que determinaram a cavidade do substrato foram identificados.

Para compor a função heurística também é necessário observar o fato da cavidade do substrato estar situada acima do cofator NADH, mais especificamente utilizando o anel da nicotinamida do NADH como base da cavidade. A Figura 8 e a Figura 13 mostram exatamente o anel da nicotinamida do cofator NADH sendo a base da cavidade alvo.

Baseada nos critérios citados nos dois parágrafos anteriores, a função heurística foi desenvolvida utilizando um sistema de pesos. Onde é atribuído o peso 1 para cada ocorrência dos átomos pesados que formem os resíduos que determinam o análogo do substrato na estrutura cristalina e, o peso 3, para cada ocorrência dos átomos pertencentes ao anel da nicotinamida do NADH.

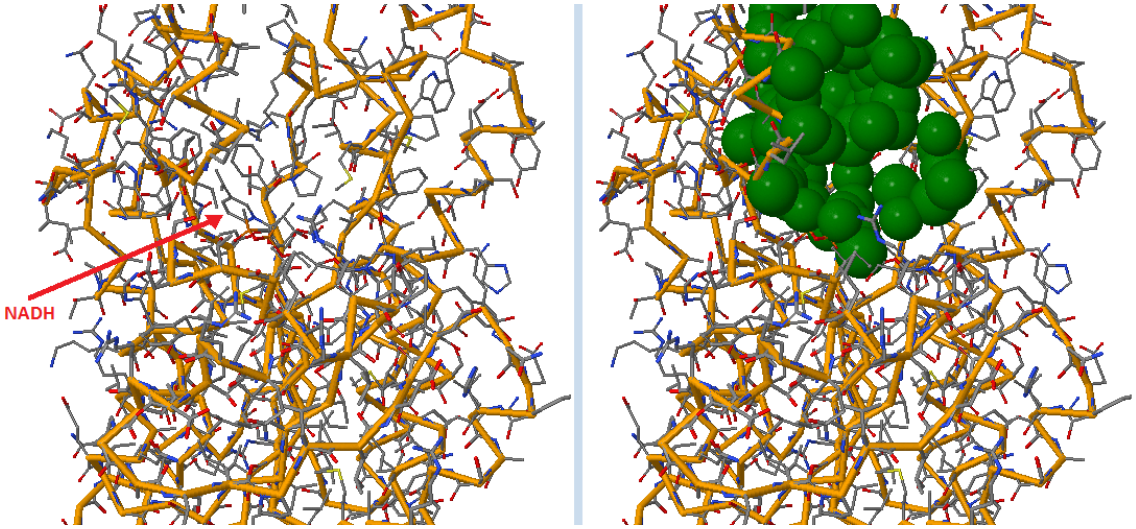


Figura 13 – Estrutura modificada da 1BVR sem o análogo do substrato. (a) A seta aponta o anel da nicotinamida do cofator NADH. (b) Visualização da cavidade do substrato feita no CASTp. Observa-se que o volume do substrato disponível está situado bem acima do anel da nicotinamida do cofator NADH.

A Tabela 9 descreve os resíduos e o número destes resíduos que compõe a função heurística. A tabela também descreve o peso atribuído a cada átomo pesado destes resíduos selecionados. Um particularidade do CASTp é selecionar somente átomos pesados para determinar as cavidades, devido estes átomos possuírem um raio maior. Desta forma, a maior pontuação que pode ser atingida por uma cavidade que seja determinada por todos estes átomos é de 195 pontos.

Tabela 9 - Resíduos que delimitam a cavidade do substrato na estrutura cristalina 1BVR e o cofator NADH, que foram utilizados na função heurística para encontrar a cavidade do substrato no modelo flexível.

Resíduos/Cofator	Peso de cada átomo	Nº de átomos no total	Nº de átomos de pesados
GLY 96	1	6	4
PHE 97	1	20	11
MET 98	1	17	8
GLN 100	1	17	9
MET103	1	17	8
GLY 104	1	7	4
PHE 149	1	20	11
MET 155	1	7	8
PRO 156	1	14	7
ALA 157	1	10	5
TYR 158	1	21	12
MET 161	1	17	8
PRO 193	1	14	7
THR 196	1	14	7
ALA 198	1	10	5
MET 199	1	17	8

ALA 201	1	10	5
ILE 202	1	19	8
LEU 207	1	19	8
GLN 214	1	17	9
ILE 215	1	19	8
LEU 218	1	19	8
NADH	3	80	9
TOTAL DE PONTOS POSSÍVEIS			**195**

Ao final da execução do algoritmo desenvolvido para esta função heurística, cada conformação do modelo flexível resultou em um conjunto de cavidades com um *e*escore atribuído, resultado do somatório dos pesos. A cavidade que apresenta o maior *e*escore, é a cavidade definida como a cavidade alvo da conformação. A Tabela 10 apresenta parte de um arquivo com os resultados da avaliação das cavidades.

Tabela 10 – Visualização das informações captadas do algoritmo desenvolvido para a função heurística de seleção das cavidades alvo do modelo flexível. A tabela apresenta os resultados das três primeiras conformações do modelo flexível, exibindo as cavidades definidas como cavidade alvo e seus respectivos *e*escores.

Conformação	Cavidade	*E*escore	Átomos que atribuíram peso para a cavidade								
1	38	3	4013								
	43	16	1458	1472	2346	2348	4009	4017	4019	4018	
2	42	6	4013	4011							
	44	17	2330	2346	2348	2361	2882	4017	4024	4019	4018
3	45	6	4013	4011							
	46	14	1451	1458	2327	2346	2348	4017	4019	4018	

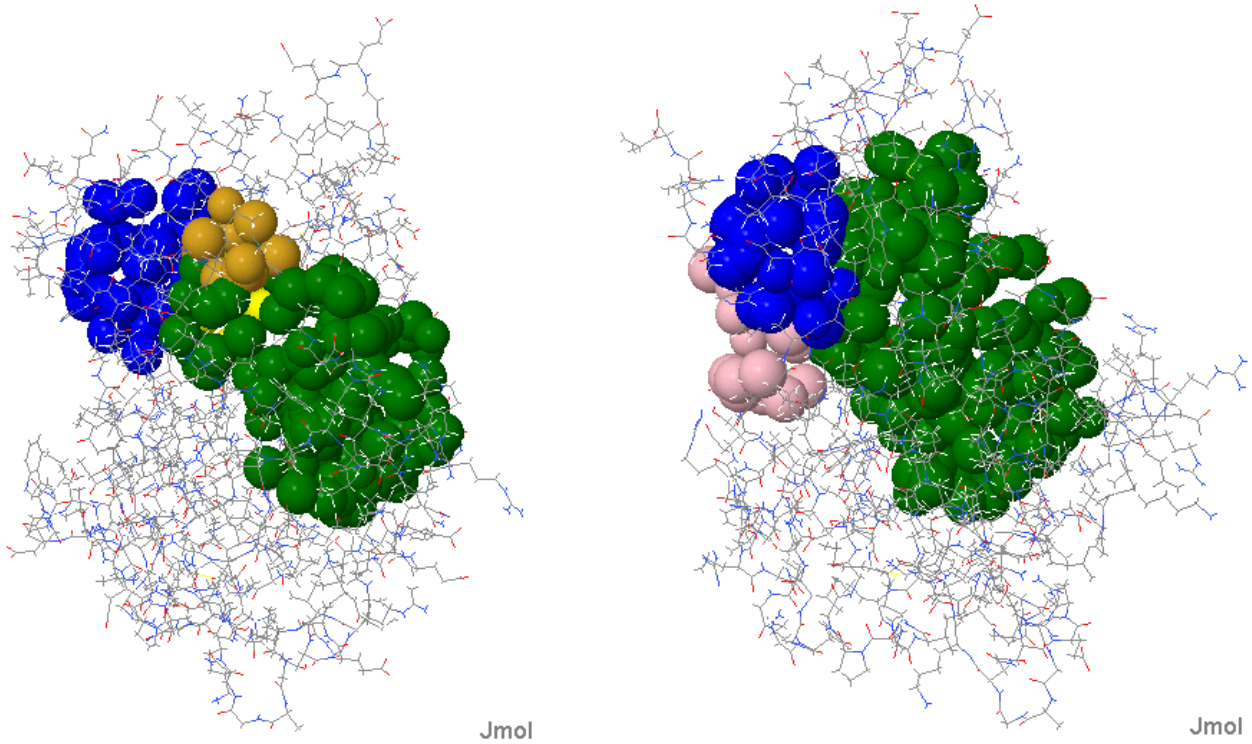
Percebe-se que a conformação 2 apresentou duas cavidades com átomos da região de interesse (42 e 44). Destas, a conformação 44 foi definida como a cavidade alvo, atingido o maior *e*escore desta conformação.

Para verificar a eficiência da heurística, foram testadas aproximadamente 200 conformações no CASTp. Em todas as conformações verificadas, a cavidade alvo definida tinha como base o anel da nicotinamida do cofator NADH. Entretanto, a região da cavidade alvo apresentou diversas fragmentações e, por questões de pesquisa, este trabalho busca cavidades com uma estrutura semelhantes a cavidade do substrato da estrutura cristalográfica. A Figura 14 demonstra duas conformações onde aparece claramente a região da cavidade alvo dividida em duas ou mais cavidades.

Devido a este problema, foi necessário investigar novas formas de capturar as informações da cavidade alvo, a fim que esta não estivesse fragmentada. Com isso, foi

feita uma avaliação dos parâmetros utilizados pelo programa CASTp para calcular as cavidades e, observou-se que este programa utiliza o raio de prova com um padrão de 1,4 Å, que é o valor do raio da área acessível a superfície da molécula pelo solvente água.

Este trabalho busca identificar a maior superfície acessível por átomos, que devido à utilização de um raio menor que o raio do solvente, certamente a área acessível a superfície será maior que a área acessível da molécula de água. A Figura 9 demonstra um exemplo da área de superfície acessível de dois raios diferentes.



(a) (b)

Figura 14 – Nas conformações 1.942 (a) e 2.180 (b), as cavidades do substrato foram fragmentadas em diversas cavidades pelo CASTp. É possível constatar que devido a flexibilidade da molécula, nas conformações acima ocorreu um estrangulamento da cavidade do substrato, fragmentando a cavidade alvo.

Deste modo, foi feita uma pesquisa procurando o menor raio de *van der waals*, cujos valores dos raios foram baseados na tabela de Bondi [61], com exceção do átomo de hidrogênio, que foi baseado em Rowland and Taylor [62]. O menor raio encontrado foi de 1,09 Å, correspondente ao átomo de Hidrogênio. Desta forma, este raio foi adotado como o raio de prova utilizado no sítio do CASTp.

A aplicação de todas as etapas novamente utilizando o raio de prova 1,09 foi um processo demorado, mas além de contribuir aumentando a área acessível da superfície,

este processo também resolveu muitos casos de conformações que tinham a cavidade do substrato fragmentada. O sub-capítulo 4.3 aborda a pesquisa por cavidades possuam semelhança com a cavidade do substrato da estrutura cristalográfica da 1BVR, descartando as conformações que ainda possam ter sua cavidade do substrato fragmentada.

4.3 Definindo o conjunto de snapshots

Após aplicar a função heurística para encontrar as cavidades alvo, foram avaliadas manualmente as informações de, aproximadamente, 300 conformações verificando as variações do volume e do escore, resultado da função heurística.

Esta avaliação possibilitou a criação de critérios para identificar o comportamento das regiões da cavidade alvo que estavam fragmentadas. Desta forma, para restringir o conjunto de conformações a ser utilizado no desenvolvimento da função heurística a fim de filtrar os ligantes deve-se descartar as conformações que apresentam a fragmentação da cavidade alvo. A regra define que todas as cavidades que possuem um volume inferior a 290 Å^3 ou que possuem seu escore abaixo de 18 pontos tem sua cavidade alvo fragmentada, devendo ser descartadas. Ao final, foram descartadas 1.461 conformações, restando 1.639 conformações (equivalente a 52,87%). Esta representação foi chamada de modelo do receptor completamente flexível.

4.4 Problemas enfrentados

Além dos problemas já citados no decorrer deste trabalho, este sub-capítulo explica um problema pontual enfrentado no desenvolvimento dos algoritmos deste capítulo. Por ser um problema paralelo as funções desenvolvidas referentes ao foco principal, o problema é descrito nesta seção para não prejudicar a fluência dos passos.

4.4.1 Realinhamento das informações geradas pelo PTRAJ

Como visto anteriormente, o CASTp realiza a leitura de arquivos no formato PDB e variações deste formato (sem ser no cabeçalho) podem gerar distorções nos resultados.

Os arquivos submetidos ao CASTp foram gerados pelo PTRAJ, software do pacote AMBER responsável por gerar os arquivos PDB da simulação por DM.

Analisando os arquivos de saída do PTRAJ, percebe-se que os nomes dos átomos dos receptores foram mantidos idênticos aos contidos na estrutura cristalográfica. Entretanto, os nomes dos átomos do NADH ficaram diferentes, de forma que o CASTp não identificou estas estruturas, gerando diversos erros.

Para solucionar este problema foi feita uma sobreposição da estrutura cristalográfica com uma conformação gerada pelo PTRAJ para identificar os átomos correspondentes (Figura 15). O resultado desta sobreposição resolveu o problema, sendo criado um programa para realizar a conversão em todos os arquivos.

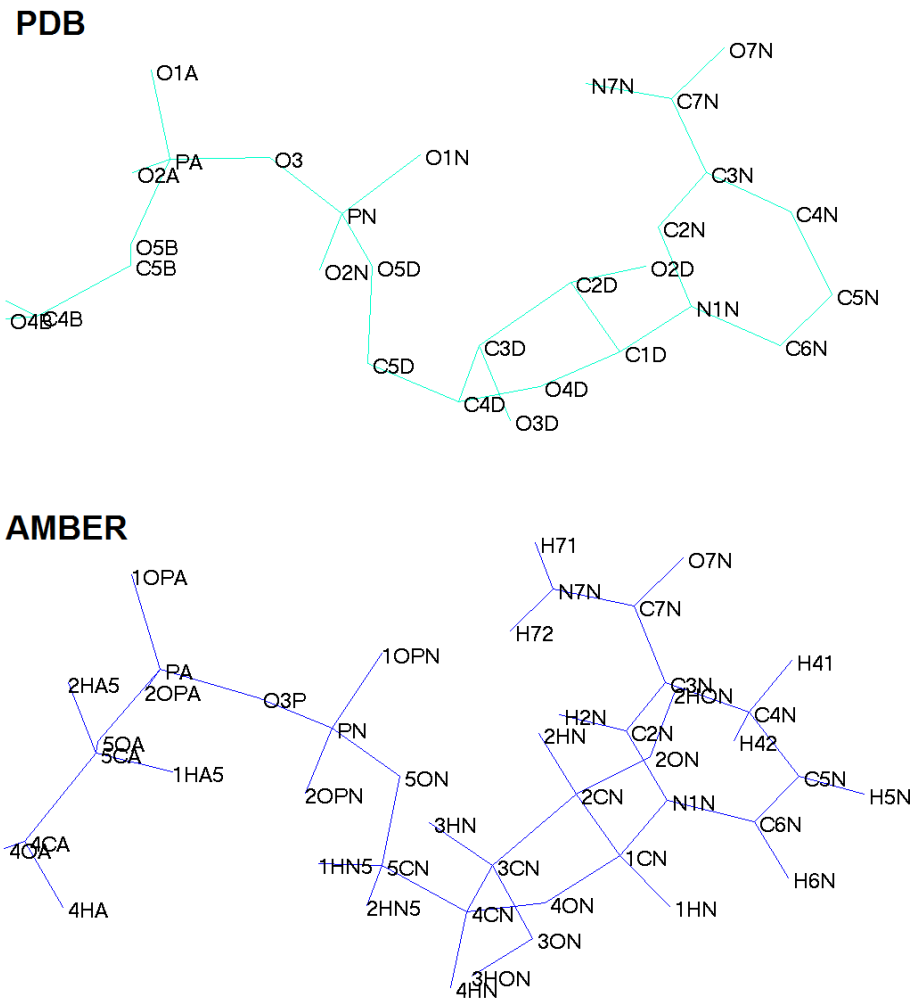


Figura 15 – Visualização de parte da estrutura 3D do NADH exibindo a nomenclatura dos átomos no formato PDB da estrutura cristalográfica (1ENY) e do arquivo gerado pelo PTRAJ. A sobreposição destas estruturas possibilitou a associação dos átomos correspondentes, corrigindo as nomenclaturas para o entendimento do CASTp.

4.5 Considerações finais

Este capítulo descreveu em detalhes a cavidade alvo deste trabalho, demonstrando esta região de interesse através de imagens da cavidade do análogo do substrato da estrutura cristalográfica (1BVR). Também apresentou o funcionamento do algoritmo desenvolvido para a submissão dos arquivos da simulação por DM, bem como a recuperação destas informações no sítio do programa CASTp. Para determinar a cavidade alvo durante as alterações causadas pela flexibilidade foi desenvolvida uma função heurística baseada na atribuição de pesos aos átomos pertencentes aos resíduos que delimitaram a cavidade do análogo do substrato na estrutura cristalina (1BVR) e aos átomos pertencentes ao anel da nicotinamida do NADH. Os resultados apresentaram cavidades fragmentadas e para obter informações mais promissoras, foram alterados os valores do raio de prova informado ao CASTp.

O CASTp define um raio de prova padrão de 1,40 Å, cuja medida é referente ao raio de uma molécula de água. No entanto, a área de superfície acessível da cavidade do substrato está sendo subestimada ao se utilizar o raio padrão, visto que os átomos possuem raios menores. Portanto, neste trabalho foi utilizado o raio de prova correspondente ao menor raio de Van der Waals (1,09 Å) e assim encontrou-se a maior área de superfície acessível por átomos [61,62]. Esta solução reduziu a ocorrência de cavidades alvo fragmentadas, entretanto algumas conformações ainda apresentavam a cavidade fragmentada. Para retirar estas conformações, além do raio de prova, foi desenvolvida uma função heurística para selecionar a cavidade alvo ao longo da trajetória. Para definir esta função foi feita uma avaliação manual de 300 conformações, induzindo a um limiar de volumes acima de 290,00 $Å^3$ e um escore igual ou superior a 18 pontos. Assim, aplicando esta regra na trajetória do receptor, foram selecionadas 1.639 conformações, o que representa aproximadamente 53% da amostra total.

Ao final deste capítulo, as cavidades alvo do modelo flexível encontram-se mapeadas e em todos os arquivos PDB existem os registros dos átomos que determinam a estrutura da cavidade alvo, a área acessível e o volume da cavidade.

5- FUNÇÃO HEURÍSTICA PARA A FILTRAGEM DE LIGANTES

Este capítulo aborda os métodos desenvolvidos para gerar a função heurística capaz de filtrar os ligantes depositados em BD utilizando os critérios geométricos de uma determinada cavidade alvo do receptor.

O desenvolvimento desta heurística foi criado com base em um conjunto de esferas concêntricas para analisar a distribuição dos átomos avaliando as propriedades geométricas da cavidade alvo e dos ligantes. O cruzamento das informações entre os volumes ocupados pelo ligante e pelo receptor em cada esfera possibilita a identificação de critérios, descrevendo para cada ligante a possibilidade de haver encaixe na cavidade alvo.

5.1 Método de esferas concêntricas

Para extrair fatores capazes de identificar as regiões de conflito entre os átomos do receptor com os átomos do ligante foi necessário criar uma forma de gerenciar este espaço 3D. Assim, foi desenvolvido um conjunto de esferas concêntricas para controlar a distribuição do volume ocupado pelos átomos. Estas esferas são compostas por um centro geométrico (CG) comum, mantendo uma distância fixa entre as bordas de cada esfera. Desta maneira, a variação do raio destas esferas obedece aos valores de uma progressão aritmética com uma razão de 0,20 Å. A utilização de uma distância fixa entre as esferas tem como principal objetivo transformar os valores obtidos das coordenadas atômicas em um conjunto finito, reduzindo o processamento e, por consequência, melhorando o tempo de resposta. Uma representação visual deste modelo de esferas concêntricas pode ser acompanhada na Figura 16.

Com o espaçamento padrão entre as esferas é possível criar uma relação utilizando o volume disponível e o volume ocupado em cada faixa das esferas. Esta relação é a base utilizada na função heurística, demonstrada nas sessões seguintes. A Tabela 11 mostra os volumes disponíveis em cada faixa, sendo obtido através do volume total da esfera menos o volume da esfera anterior. Evidentemente, o volume disponível aumenta conforme o crescimento das esferas começa a distanciar-se do CG das esferas.

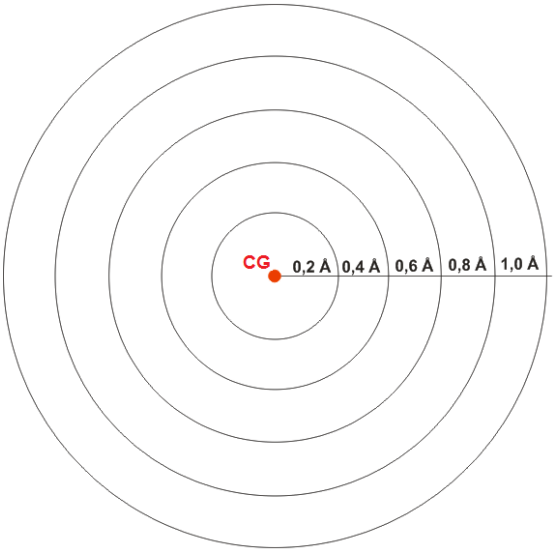


Figura 16 - Corte transverssal no modelo de esferas concêntricas criadas para controlar o volume livre e o volume ocupado pelo átomos do receptor e do ligante. Os raios de cada esfera obedecem aos valores de uma progressão aritmética com razão de 0,2 Å.

Tabela 11 - Distribuição do volume máximo ocupado nas faixas das esferas .

Distância do CG (Å)	**0,2**	**0,4**	**0,6**	**0,8**	**1,0**	**1,2**	**1,4**	**1,6**	**1,8**	**2,0**
Volume máximo da faixa ($Å^3$)	0,0335	0,2346	0,6702	1,4745	2,7143	4,5239	6,9701	10,1871	14,2419	19,2684

O processamento das informações do receptor e dos ligantes ocorre de forma semelhante, sendo utilizado o modelo de esferas para armazenar as informações dos volumes ocupados dos átomos de cada conformação do receptor e de cada ligante. Este modelo de esferas concêntricas, aplicado tanto ao receptor quanto aos ligantes, mantém sempre a mesma escala visando posteriormente o cruzamento das informações. Entretanto, existe uma diferença na determinação onde o ponto CG será definido. Os detalhes de cada implementação está descrito na seção a seguir.

5.2 Definindo o ponto correspondente ao CG do modelo de esferas concêntricas para o ligante e para o receptor

A definição do CG é um passo importante, visto que dependendo da estratégia adotada o ligante pode apresentar ou não colisão com os átomos do receptor, ou seja, a escolha deste ponto pode ser a diferença entre o sucesso e o fracasso da avaliação. O

objetivo é situar o centro geométrico dos ligantes no ponto central dentro da cavidade, sendo este o mais distante possível dos átomos. A definição dos pontos desta maneira pretende maximizar as chances de acerto quando o ligante não possui um volume que possa permitir a sua docagem na cavidade.

No ligante, o ponto do CG das esferas concêntricas corresponde exatamente ao centro geométrico dos átomos do ligante. Assim, foram computadas todas as coordenadas atômicas e divididas pelo número de átomos. A Figura 17 apresenta um exemplo do centro geométrico calculado para o ligante TCL.

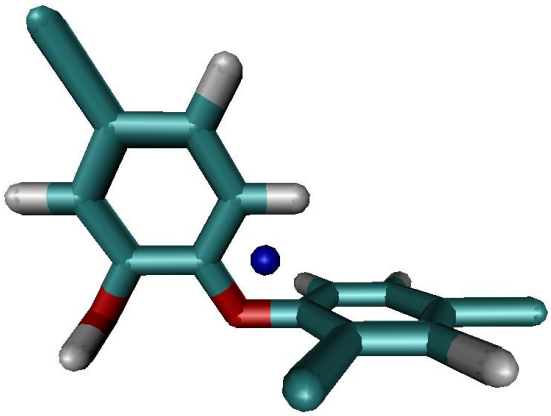


Figura 17 - Ligante TCL e o ponto (azul) definido como CG do modelo de esferas concêntricas .

Como visto, para definir-se o centro geométrico do ligante não há grandes dificuldades. No entanto, no receptor o ponto definido como o CG corresponde ao centro da cavidade alvo, sendo considerada uma tarefa computacional complexa para a descoberta em uma conformação. Cabe ressaltar que ao aplicar esta estratégia em uma conformação não pode ser algo demorado, já que este problema envolve o tratamento de todo o modelo flexível do receptor. Portanto, para o receptor foi analisado cada um dos 1.639 arquivos que compõem o modelo flexível e utilizando as informações dos átomos que delimitam a cavidade alvo (fornecidos pelo CASTp) como referência para determinar o centro da cavidade alvo. A forma mais precisa para identificar este ponto seria aplicar técnicas baseadas em filtros de médias, mas devido às dimensões 3D o custo computacional, o uso desta solução tornou-se inviável.

Para contornar este problema foi desenvolvido um método que seleciona pares de átomos que determinam a cavidade alvo, identificando o ponto médio entre estes pontos. Deste ponto médio são calculadas as distâncias euclidianas para todos os átomos que

delimitam a cavidade, sendo selecionada a distância para o átomo mais próximo. Portanto, para cada ponto médio escolhido há uma relação com os átomos mais próximos. Como buscamos o centro da cavidade, o ponto médio escolhido é aquele que possui a maior distância do átomo mais próximo. A Figura 18 mostra uma conformação do modelo flexível e o ponto determinado como o CG do modelo de esferas concêntricas.

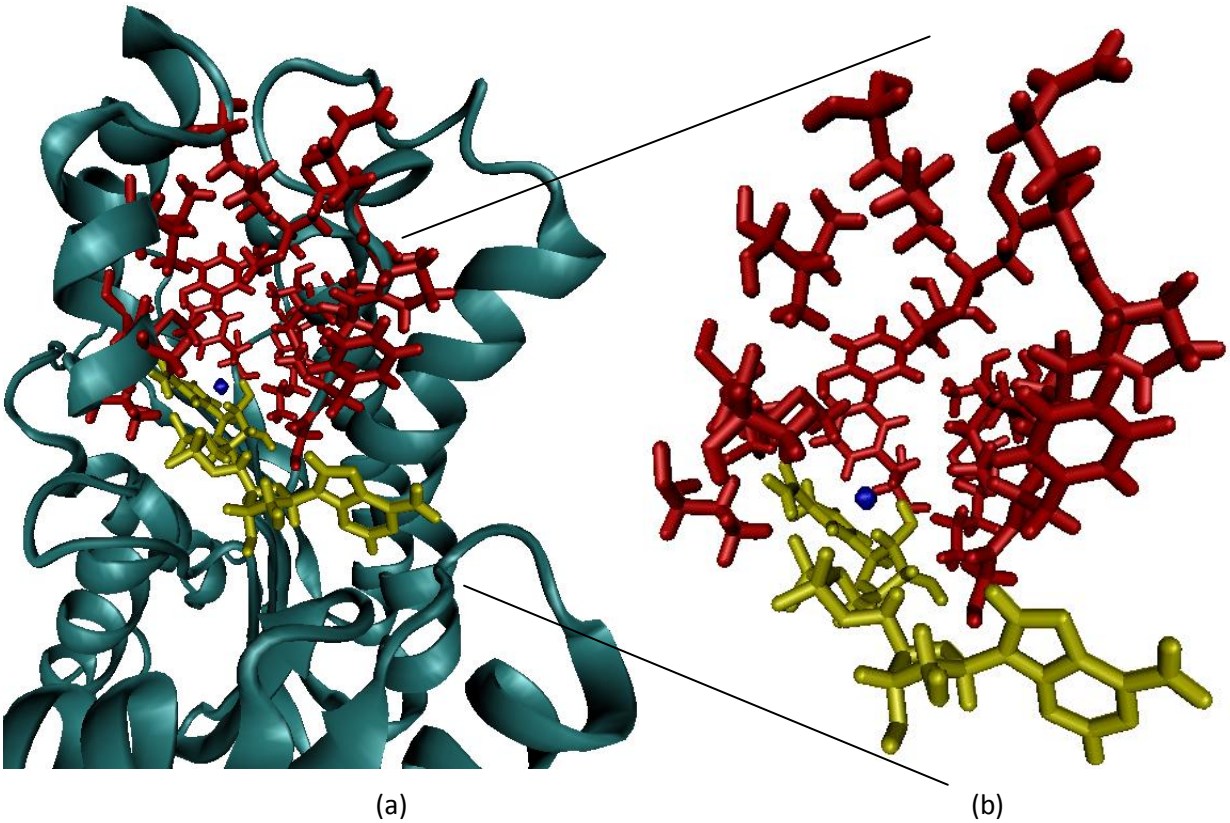


Figura 18 – Esfera (azul) definida como CG do modelo de esferas concêntricas para a conformação 10 do modelo flexível. (a) visão completa da estrutura, mostrando a enzima InhA (verde) representada no modelo de fitas. (b) destaque da região da cavidade definida pelo CASTp (vermelho) e a co-enzima NADH (amarelo), ambas representadas no modelo de palitos.

A solução encontrada apresentou resultados satisfatórios e embora apresente uma complexidade cúbica (O^3), o tempo necessário para calcular cada conformação foi de aproximadamente 3 segundos.

5.3 Distribuição do volume dos átomos nas esferas concêntricas

Com os CG definidos para os ligantes e para receptores, o próximo passo constitui em distribuir o volume dos átomos dentro das esferas concêntricas. Embora possuam CG com coordenadas distintas, a distribuição das posições dos átomos do

receptor e dos ligantes ocorre de forma idêntica, sendo calculada a distância do centro do átomo até o CG e assim definindo a faixa que o centro deste átomo pertence. Desta forma, esta avaliação independe do sistema de referência das coordenadas atômicas, que os átomos estão.

Inicialmente, é feita a identificação da esfera em que o centro do átomo pertence, sendo esta posição determinada a partir do cálculo da distância euclidiana das coordenadas atômicas dos átomos até o CG. O valor obtido é quantificado pelo método de arredondamento, identificando qual o raio da esfera concêntrica se assemelha.

Para identificar a colisão de átomos do receptor com os do ligante é necessário considerar, além do centro dos átomos, o volume de cada átomo. Para determinar os volumes, foram utilizados os raios de Van der Waals [61,62]. Ao avaliar os limites de colisão das extremidades dos átomos aumenta expressivamente o volume ocupado. Por exemplo, o átomo com o menor raio pertence ao hidrogênio, com 1,09 Å. Mesmo sendo o menor raio, o volume deste átomo ocupa até 11 faixas diferentes.

Como visto, o volume do átomo está distribuído em mais de uma esfera. Para atribuir o volume específico de um átomo em cada esfera é necessário o uso de uma série de fórmulas matemáticas, que serão detalhadas a seguir. Primeiramente é necessário descobrir o volume do átomo, que possui uma estrutura semelhante a uma esfera. Portanto, a equação 1 é a fórmula utilizada para este cálculo.

$$V = \frac{4}{3}\pi r^3 \qquad (1)$$

A partir da determinação do volume total do átomo, a etapa seguinte tem o objetivo de identificar o volume a ser atribuído para cada esfera concêntrica. Para isso foi calculado o volume de interseção entre duas esferas utilizando a equação 2 (Figura 19). Embora este método tenha um alto fator de precisão, seu cálculo envolve muitas variáveis. A Figura 20 mostra como ocorre a interseção das esferas concêntricas com o volume dos átomos.

$$V = \frac{1}{12d}\pi(R + r - d)^2(d^2 + 2dr - 3r^2 + 2dR + 6rR - 3R^2) \qquad (2)$$

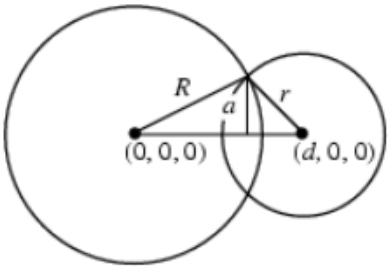


Figura 19 – Representação da interseção de duas esferas demonstrando os parâmetros da equação 3.

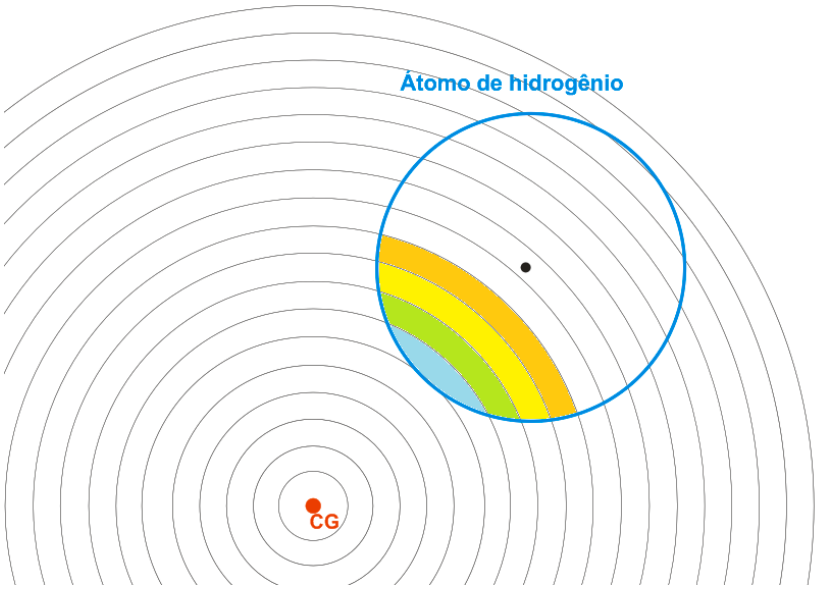


Figura 20 - Distribuição do volume do átomo de hidrogênio de acordo com a interseção com cada esfera concêntrica.

Para atribuir o volume de cada faixa, é necessário descobrir o volume da interseção entre as esferas e posteriormente subtrair do volume já calculado quando houver outras faixas presentes nesta mesma interseção. Cada faixa do átomo pode atribuir volumes diferentes às esferas concêntricas. Além disso, conforme a distância do CG estes volumes atribuídos a cada faixa também variam devido a inclinação das circunferências que cortam os átomos. Quanto mais distante do CG for a interseção a circunferências tendem a apresentar linhas mais paralelas.

Devido a grande quantidade de átomos a serem tratados, realizar todos estes cálculos para atribuir o volume as esferas concêntricas é um processo oneroso. Com o intuito de melhorar a performance, foi criado um programa capaz de gerar um arquivo contabilizando os volumes atribuídos para cada faixa com todas as distância possíveis e

com todos os átomos utilizados. Assim, para atribuir o volume de cada faixa é necessário apenas identificar a distância do CG e o átomo envolvido.

Após ser feita a distribuição do volume de todos os átomos de forma proporcional às esferas, é necessário identificar a ocorrência da sobreposição dos volumes de átomos próximos. A Figura 21 mostra parte de uma conformação da enzima InhA representada por esferas de van der Waals.

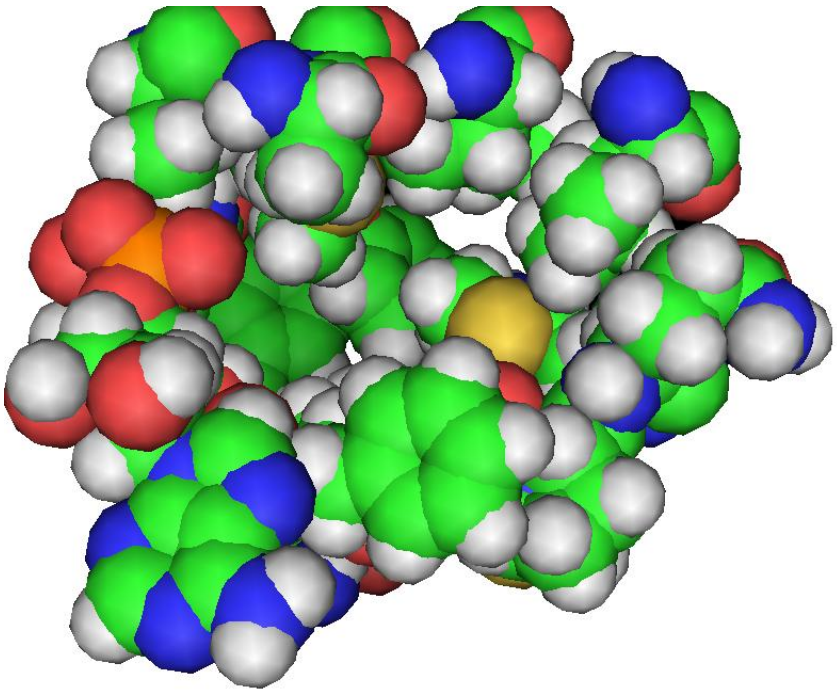


Figura 21 – Visualização de parte da enzima InhA representada por esferas de van der Waals. Nesta representação é possível obsevar a ocorrência da sobreposição do volume dos átomos.

Como observado na Figura 21, todos os átomos visíveis na imagem apresentaram alguma sobreposição em seu volume. Caso a função heurística deste trabalho contabilize mais de uma vez a mesma região ocupada por outro átomo, os resultados obtidos não seriam confiáveis. Portanto, para resolver este problema deve-se desconsiderar o volume contabilizado em excesso para cada uma das esferas.

Identificada a ocorrência da interseção entre dois átomos é necessário descontar este volume sobreposto das esferas concêntricas. Com a descoberta de átomos sobrepostos, calcula-se o volume da interseção entre estes átomos e deve-se retirar este volume excedente das esferas concêntricas. Para isso é aplicada a equação 3, que identifica o volume da interseção de três esferas [63].

$$V_{ABC} = \frac{w}{6} - \frac{a}{2}\left[\beta^2 + \gamma^2 - a^2\left(\frac{1}{6} - \frac{\epsilon_1^2}{2}\right)\right]\tan^{-1}\frac{2w}{q_1} - \frac{b}{2}\left[\gamma^2 + \alpha^2 - b^2\left(\frac{1}{6} - \frac{\epsilon_2^2}{2}\right)\right]\tan^{-1}\frac{2w}{q_2} -$$
$$\frac{c}{2}\left[\alpha^2 + \beta^2 - c^2\left(\frac{1}{6} - \frac{\epsilon_3^2}{2}\right)\right]\tan^{-1}\frac{2w}{q_3} + \frac{2\alpha^3}{3}\left[\tan^{-1}\left\{\frac{bw}{\alpha q_2}(1-\epsilon_2)\right\} + \tan^{-1}\left\{\frac{cw}{\alpha q_3}(1+\epsilon_3)\right\}\right] +$$
$$\frac{2\beta^3}{3}\left[\tan^{-1}\left\{\frac{cw}{\beta q_3}(1-\epsilon_3)\right\} + \tan^{-1}\left\{\frac{aw}{\beta q_1}(1+\epsilon_1)\right\}\right] + \frac{2\gamma^3}{3}\left[\tan^{-1}\left\{\frac{aw}{\gamma q_1}(1-\epsilon_1)\right\} + \tan^{-1}\left\{\frac{bw}{\gamma q_2}(1+\epsilon_2)\right\}\right]$$
$$(0 \le \tan^{-1} \le \pi) \qquad (3)$$

onde:

$$\epsilon_1 = (\beta^2 - \gamma^2)/a^2$$
$$\epsilon_2 = (\gamma^2 - \alpha^2)/b^2$$
$$\epsilon_3 = (\alpha^2 - \beta^2)/c^2$$
$$q_1 = a[b^2 + c^2 - a^2 + \beta^2 + \gamma^2 - 2\,\alpha^2 + \epsilon_1(b^2 - c^2)]$$
$$q_2 = b[c^2 + a^2 - b^2 + \gamma^2 + \alpha^2 - 2\,\beta^2 + \epsilon_2(c^2 - a^2)]$$
$$q_3 = c[a^2 + b^2 - c^2 + \alpha^2 + \beta^2 - 2\,\gamma^2 + \epsilon_3(a^2 - b^2)]$$
$$w^2 = (\alpha^2a^2 + \beta^2b^2 + \gamma^2c^2)(a^2 + b^2 + c^2) - 2(\alpha^2a^4 + \beta^2b^4 + \gamma^2c^4) + a^2b^2c^2(\epsilon_1\epsilon_2 + \epsilon_2\epsilon_3 + \epsilon_3\epsilon_1 - 1)$$

Com a aplicação da equação 3 o volume de cada interseção é distribuído para cada esfera concêntrica. Entretanto, os átomos podem conter 3 ou mais interseções e, para solucionar os demais conflitos, utilizou-se algumas funções do programa ARVO. Este programa é capaz de calcular o volume de três ou mais esferas e seu desenvolvimento também é baseado na equação 3. O ARVO está disponível na linguagem de programação Fortran 77 e para ser utilizado neste trabalho foi necessário um trabalho de conversão da linguagem de programação, sendo este programa reescrito em linguagem C[64].

Após resolver os problemas das sobreposições dos átomos, as informações dos volumes ocupados em cada faixa são armazenadas em vetores, onde cada posição deste vetor representa o volume ocupado pelos átomos. Estes dados são tratados de forma independente, portanto cada ligante terá um vetor para armazenar suas informações. O mesmo ocorre com o receptor, onde são gerados 1.639 vetores criando-se uma matriz para armazenar as informações do modelo flexível, sendo as linhas atribuídas à conformação e as colunas as faixas desta conformação.

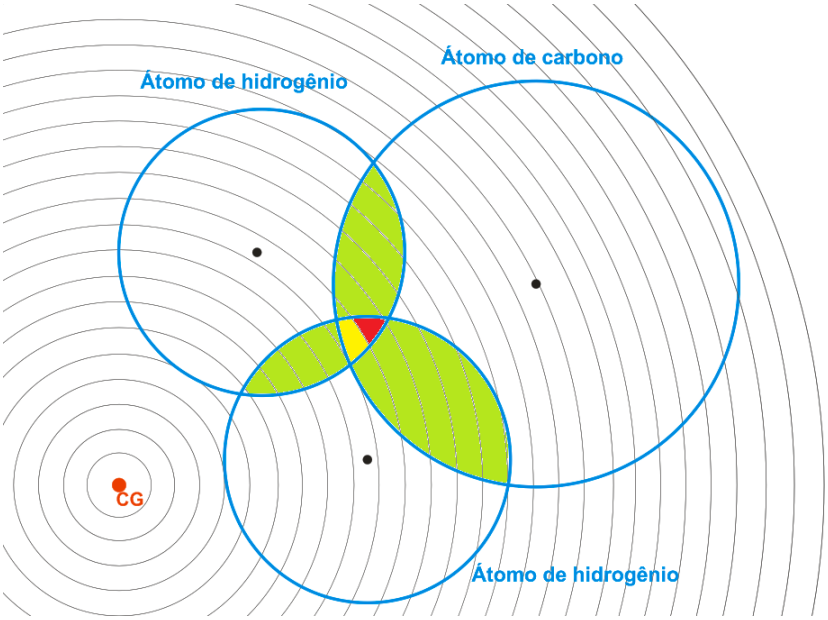


Figura 22 - Interseção de quatro esferas utilizada no desenvolvimento da função heurítica. Nesta conformação foram calculados os volumes de duas interseções entre quatro esferas (amarelo e vermelho).

5.4 Função heurística: cruzamento das informações dos receptores e dos ligantes

Esta função heurística visa identificar os ligantes que possuem seu volume compatível com a estrutura da cavidade alvo. Para isso, as conformações dos receptores e dos ligantes têm suas informações do volume descritas em um modelo de esferas concêntricas. Considerando que os ligantes e os receptores têm seus CG sobrepostos, é feita uma série de testes verificando a possibilidade ou não de encaixe dos ligantes em cada conformação.

Como visto anteriormente, os vetores gerados apresentam as informações dos volumes ocupados distribuídos em cada faixa. Para identificar a possibilidade de colisão entre os volumes é necessário encontrar o volume ocupado acumulado das esferas concêntricas do receptor e do ligante. Assim, a contabilização dos volumes começa a partir do CG e segue acumulando conforme os volumes encontrados na direção das esferas mais distantes. Além disso, a visualização do gráfico (Figura 23) fica mais clara quando é apresentado o volume permitido dentro da cavidade alvo ao invés de colocar no gráfico o volume ocupado. Este volume é obtido quando se retira o volume ocupado do receptor do volume máximo da esfera concêntrica. A Figura 23 apresenta algumas curvas obtidas com os volumes da esfera total e algumas conformações do receptor.

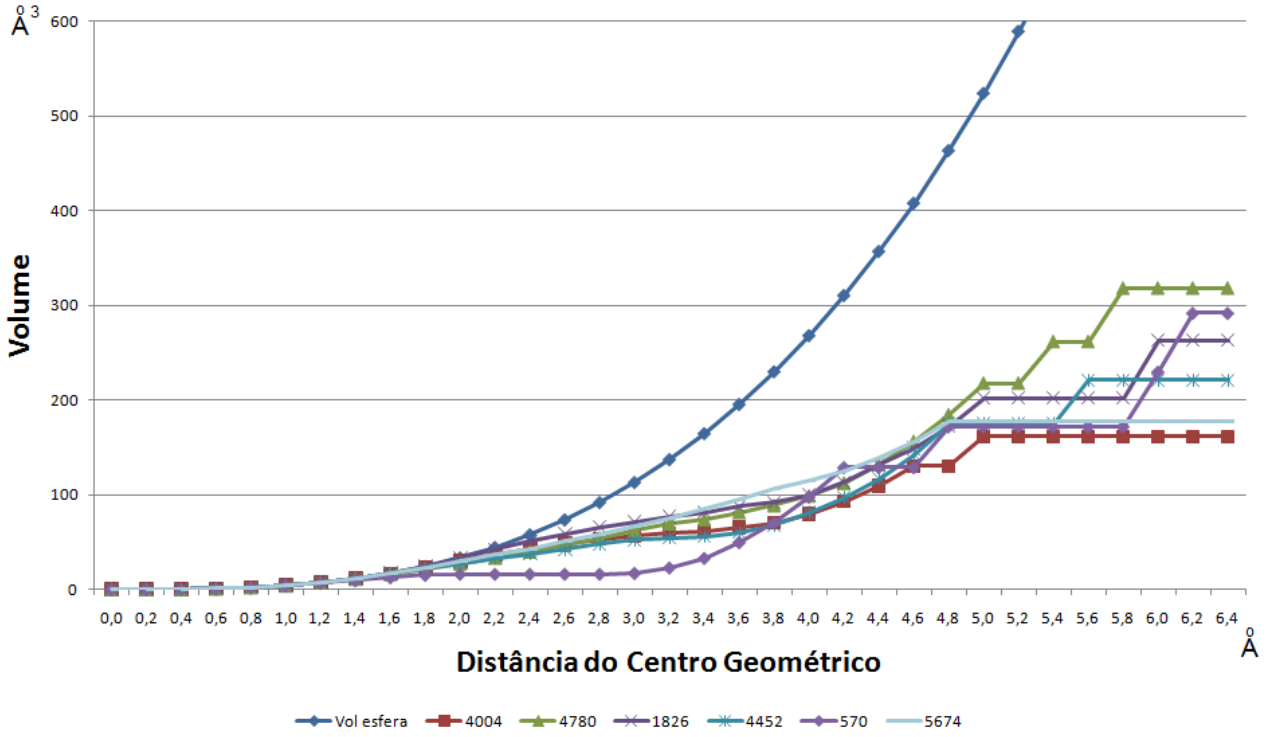


Figura 23 – Gráfico mostrando as curvas que apresentam os volumes livres da cavidade alvo a partir do CG. Analisando a conformação 570, nota-se que a cavidade apresenta um volume livre inicial muito baixo, indicando ser uma cavidade pequena com forte estreitamento até encontrar outro segmento da cavidade.

Como visto na Figura 23 é possível constatar que as curvas do receptor, devido à sua cavidade, apresentam em sua maioria um comportamento com os átomos iniciais distantes do CG. Assim, quanto mais os átomos estiverem distantes do centro da cavidade, maior será o valor inicial apresentado pela curva, aproximando-se da curva apresentada pelo volume total da esfera concêntrica. Por consequência, as conformações que apresentam uma cavidade mais fechada possuem curvas com volumes baixos, demonstrando uma maior dificuldade/impossibilidade do ligante docar nesta região.

As informações dos ligantes a serem representadas no gráfico são referentes ao volume ocupado acumulado em cada faixa. Estes volumes acumulados são a base para acompanhar o crescimento dos limites da estrutura. A Figura 24 apresenta alguns resultados de um grupo de ligantes selecionados do BD ZINC. Inicialmente, as curvas tendem a acompanhar o crescimento da curva do volume total da esfera. Isto ocorre devido a estratégia utilizada, baseada no Centro Geométrico do ligante. Um comportamento natural desta estratégia é que ocorrem curvas elevadas nas primeiras distâncias visto que o ligante tem seus átomos próximos do CG.

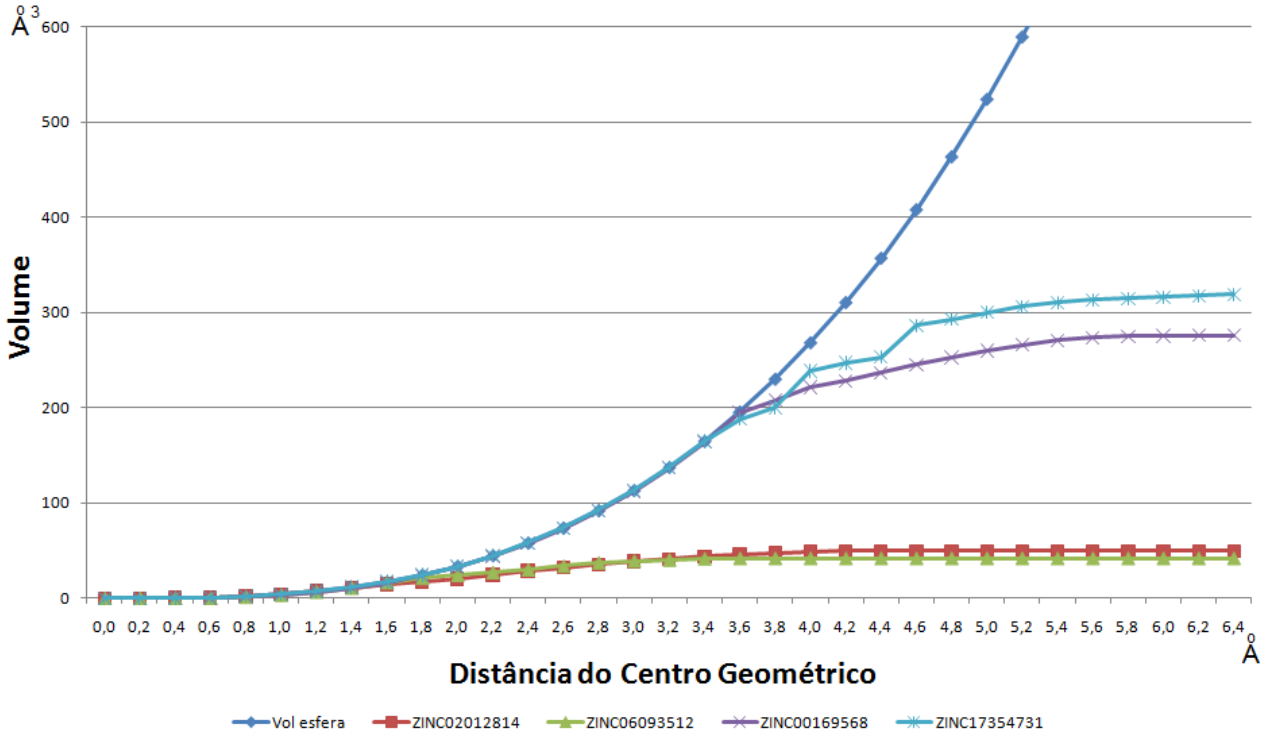


Figura 24 - Gráfico mostrando as curvas dos volumes ocupados acumulados dos ligantes a partir do CG. São apresentados quatro ligantes com dois volumes distintos. Os ligantes ZINC00169568 e ZINC17354731 mostram uma alta concentração de volume ocupado próximo do CG, dificultando a possibilidade de haver encaixe com outra conformação.

Para comparar as informações entre os receptores e os ligantes é feito um cálculo identificando se o volume do ligante atinge um valor superior ao volume disponível na cavidade. Nas faixas em que este volume ocupado pelo ligante for maior que o disponível pela cavidade é feito um somatório deste volume. Assim, para cada ligante é gerada uma lista com o resultado de cada conformação. Todas as conformações que apresentarem o valor do somatório nulo são consideradas conformações que devem ser testadas com a docagem molecular. No próximo capítulo são apresentados dois testes demonstrando com detalhes os testes de validação da função heurística.

Ainda encontra-se em fase de avaliação o valor do limite máximo do volume que pode ser compartilhado entre os átomos do receptor e do ligante quando ocorrem interações válidas. Assim, o limiar para eliminar os testes de docagem das conformações é um valor configurável, dependendo dos volumes envolvidos. Portanto, as conformações que apresentarem um valor baixo do somatório também devem ser testadas, evitando assim falsos positivos.

6- VALIDAÇÃO DA FUNÇÃO HEURÍSTICA PROPOSTA

Neste capítulo foram desenvolvidos dois testes com a finalidade de validar a função heurística proposta. O primeiro teste busca verificar se ligantes com um volume superior a cavidade são comprovadamente descartados pela função heurística desenvolvida. O segundo avalia um conjunto de resultados de docagem molecular e comparando as FEB obtidas de cada conformação com os indicadores dos resultados dos gráficos gerados pela função heurística.

6.1 Teste A

Este teste tem como objetivo identificar um conjunto de ligantes com volumes superiores ao volume de uma conformação da cavidade alvo e esta deve restringir todos os ligantes pesquisados. Caso o gráfico indique que não deva ocorrer o descarte, esta avaliação já estará inferindo a ocorrência de falsos positivos.

O conjunto de ligantes testados foi baixado do BD ZINC e indicava uma quantidade superior a 192.000 ligantes, mas devido algum problema na disponibilização destes dados pelo BD ZINC, apenas 3.745 moléculas não eram duplicadas. Com este conjunto definido, o próximo passo é calcular o volume dos ligantes. Para isso, foi pesquisado na literatura uma série de programas, dentre eles foram encontrados o Molinspiration (ZINC), ChemAxon, V3cavity e o arquivo fonte em linguagem C do programa Mol_volume [65]. Por questões de agilidade no processamento (visto que três eram via internet) e pelo domínio pessoal da linguagem C definiu-se o programa Mol_volume para realizar o cálculo do volume dos ligantes.

As conformações testadas foram selecionadas pelo tamanho do volume da cavidade indicado pelo CASTp. As cavidades com o menor volume do modelo flexível possuem aproximadamente 295 Å³. Dentre estas, foi escolhida a conformação 4.004 ns para exibir maiores detalhes. Na Figura 25 é mostrada parte da estrutura desta conformação, que são os resíduos que determinam a cavidade alvo. É possível constatar uma cavidade bastante fechada na proximidade do CG, entretanto ela possui espaçamentos permitindo a interação do ligante com o restante da cavidade.

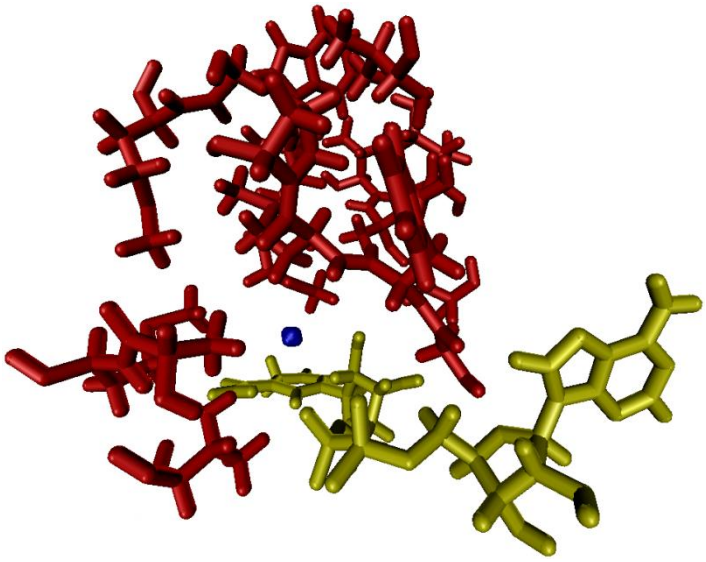


Figura 25 – Estrutura do substrato da conformação 4.004 ps gerada pelo CASTp. A cavidade é formada pela coenzima NADH (amarelo) e por um conjunto de resíduos indicados (vermelho) como delimitadores da cavidade alvo pelo CASTp. A esfera em azul aponta o Centro Geométrico da cavidade.

Para definir o conjunto de ligantes a serem testados para esta conformação, foram pesquisados aqueles ligantes com um volume superior ao volume apresentado pelo receptor. Desta forma, a faixa de volume selecionada para executar o teste são volumes acima de 295 Å³.

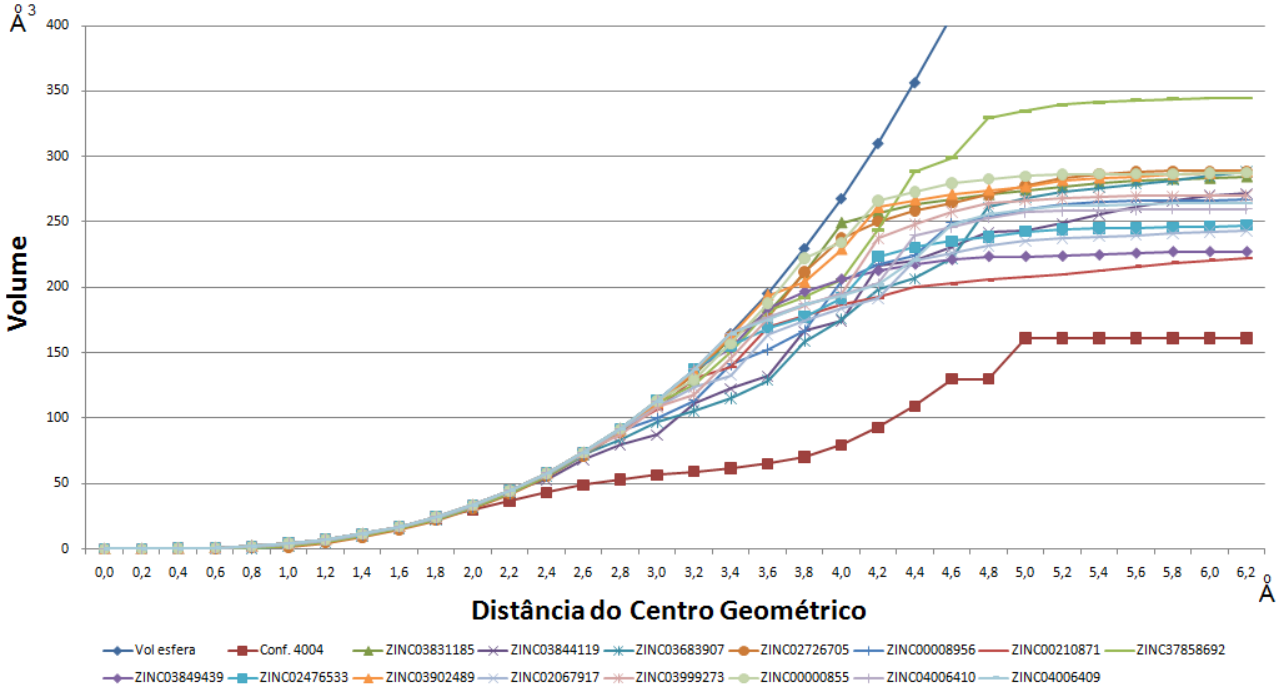


Figura 26 – Comparação dos resultados obtidos pela função heurística para a conformação 4004 e um conjunto de ligantes com volumes superiores a 295 Å³. Percebe-se que os volumes dos ligantes estão acima do limite disponível pela conformação, descartando-se todos os ligantes.

Avaliando todas as curvas pode-se concluir que a função heurística apontou uma grande sobreposição da conformação com os ligantes testados, resultando então no descarte de todos os ligantes testados.

6.2 Teste B

Para avaliar a função heurística desenvolvida de forma mais refinada foi necessário gerar um processo de docagem molecular da estrutura da enzima InhA contendo a coenzima NADH. O ligante definido para docar dentro da cavidade foi o TCL, que é um ligante com suas cargas já calculadas pelo LABIO. A Figura 27 mostra a caixa criada pelo mkbox (programa que pertence ao pacote do Autodock 3.0.5). Esta caixa, de dimensões 40x50x50 e um grid de 0,375 Å, restringe a região em que o ligante tenta docar. Além da caixa, a Figura 27 também destaca a presença do NADH dentro da estrutura inicial da macromolécula e a posição inicial do ligante.

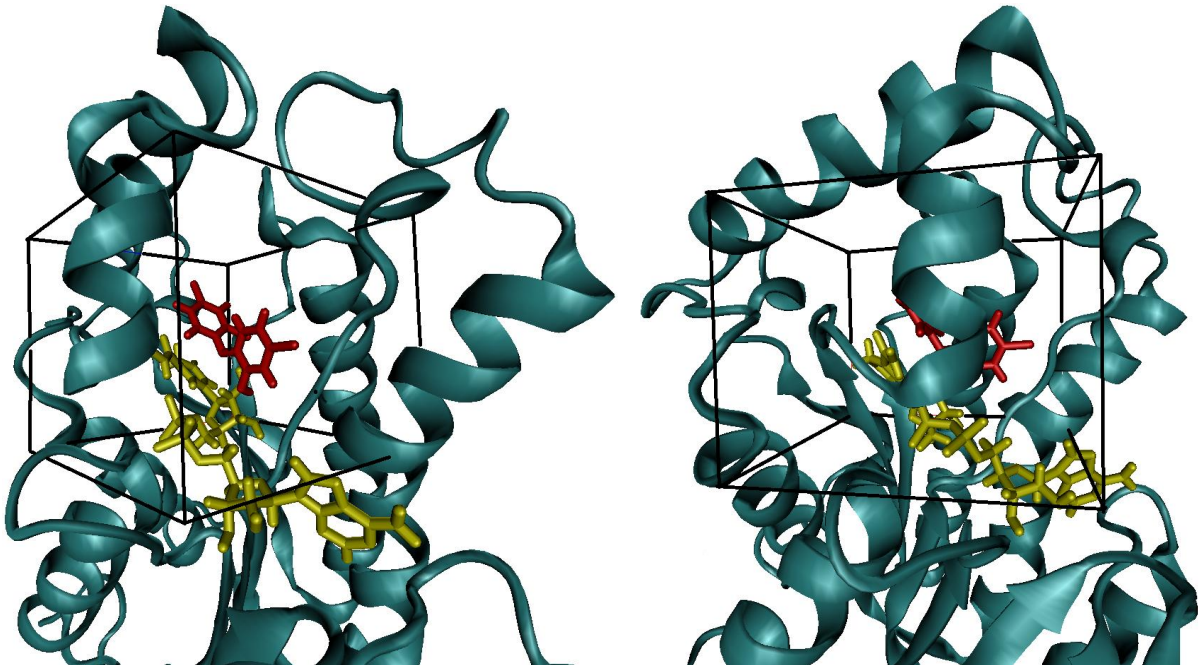


Figura 27 – Posicionamento inicial da molécula InhA-NADH e do ligante TCL antes de começar o processo de docagem. A caixa, com dimensões 40x50x50, define a região em que o ligante pode tentar docar.

Os resultados da docagem molecular são avaliados por métricas capazes de identificar a quantidade de energia liberada. Portanto, são calculados os termos da energia livre de ligação (FEB), que quanto mais negativa, melhor a interação receptor-

ligante. Os resultados apresentando as melhores FEB desta docagem molecular com ligante flexível estão descritas na Tabela 12.

Tabela 12 – Visualização dos melhores resultados da FEB do processo de docagem molecular entre a InhA-NADH com o ligante TCL.

Conformação	*Run*	RMSD	FEB
2	14	5,59	-11,35
13	21	5,78	-11,29
3	24	5,29	-11,09
1	3	5,44	-11,08
9	7	5,29	-10,94
6	20	5,47	-10,82
8	24	6,04	-10,6
21	13	6,49	-10,41
4	24	5,19	-10,37
7	23	5,41	-10,27
5	13	5,13	-9,97
1166	9	6,56	-9,76
10	9	6,23	-9,66
12	19	6,52	-9,66
11	23	6,60	-9,54
19	2	6,37	-9,5
1099	10	7,22	-9,26
1365	18	7,14	-9,24
15	17	6,50	-9,12
459	8	7,44	-8,73
16	23	5,90	-8,69
18	19	6,83	-8,67

Embora tenham ocorrido bons resultados de FEB, uma verificação dos resultados das docagens mostrou que o posicionamento do ligante fora da cavidade do substrato, sendo estes resultados classificados como falsos positivos. Este fato pode ser observado pelo alto valor do RMSD indicado realmente o afastamento do ligante da cavidade alvo. Outros testes foram realizados utilizando dimensões menores para a caixa e os resultados indicaram apenas 10 bons resultados de FEB. Ainda assim, estes resultados apresentaram a posição do ligante afastado da cavidade do substrato.

Esperava-se obter bons resultados de FEB com o ligante posicionado dentro da cavidade do substrato conforme a estrutura cristalina semelhante (código PDB: 1P45). Diante deste fato, foi feita uma comparação entre estas estruturas (Figura 28) mostrando

que a cavidade do substrato no modelo de receptor flexível utilizado neste trabalho é menor que a cavidade da estrutura cristalina, impossibilitando a docagem do ligante TCL na cavidade do substrato.

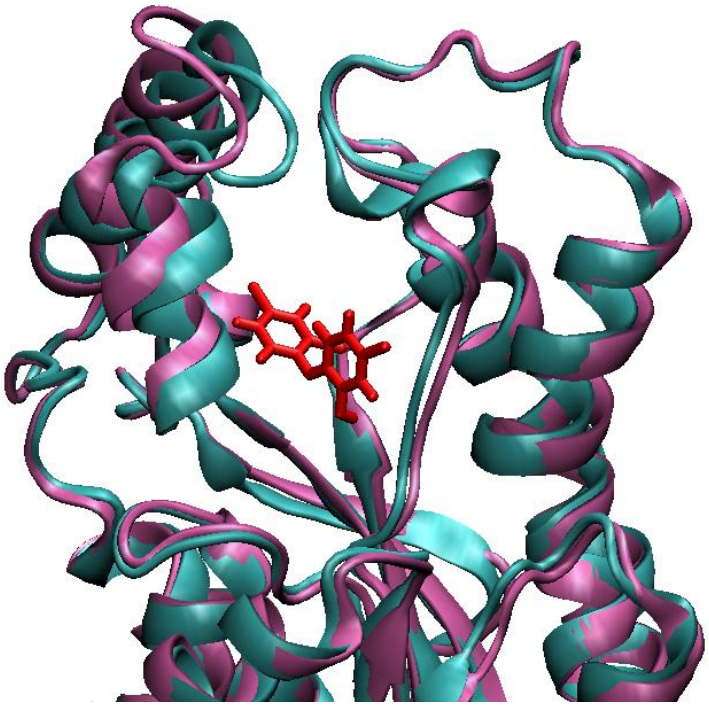


Figura 28 – Comparação entre a estrutura cristalina 1P45 e a conformação inicial do modelo do receptor completamente flexível. Nota-se um estreitamento da cavidade do substrato devido ao fechamento das fitas.

O teste utilizando a função heurística utilizou o modelo do receptor totalmente flexível comparando com o ligante TCL, exatamente como foi feito no processo de docagem anterior. A Tabela 13 aponta as conformações que apresentaram as menores sobreposições do volume do ligante sobre os resultados da estrutura.

Tabela 13 – Visualização das faixas onde ocorreram sobreposições do volume do ligante sobre o volume das conformações e o volume total da sobreposição para as conformações que apresentaram as menores sobreposições.

Conformação	**Volume sobreposto (Å³)**	**Faixas onde o volume do ligante foi maior que o volume da conformação (número da faixa) volume sobreposto (Å³)**									
4708	22,03	(16)2,57	(17)13,31	(18)6,13							
386	32,76	(10)0,04	(15)0,92	(16)8,62	(17)17,18	(18)5,97					
1268	48,14	(9)0,11	(10)0,60	(15)2,43	(16)11,56	(17)21,74	(18)11,67				
4754	67,45	(7)0,02	(8)0,19	(9)0,64	(10)1,60	(12)1,84	(13)2,20	(15)6,31	(16)14,93	(17)25,05	(18)14,6
3066	77,36	(8)0,02	(9)0,24	(10)0,83	(12)0,46	(13)1,00	(15)6,39	(16)16,21	(17)27,41	(18)20,11	(19)2,78
4688	94,63	(9)0,23	(10)0,88	(12)2,23	(13)3,65	(14)2,21	(15)10,45	(16)20,41	(17)30,85	(18)20,89	(19)4,65

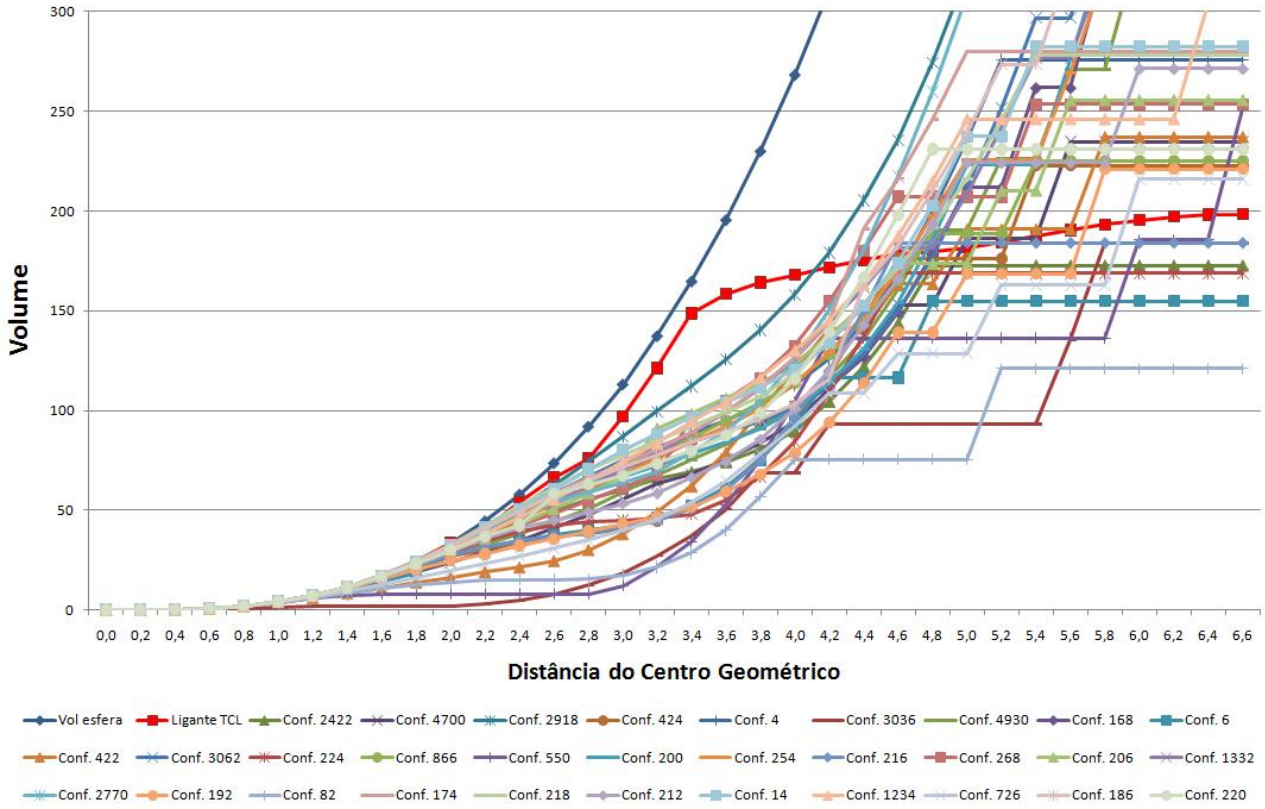


Figura 29 – Avaliação da função heurística com o ligante TCL testado com o modelo do receptor completamente flexível.

Este resultado pode ser melhor acompanhado na Figura 29, comprovando a avaliação da função heurística que indica a impossibilidade de ocorrer uma boa docagem nas conformações para o ligante TCL.

7- CONCLUSÕES

Este trabalho apresentou o desenvolvimento de um novo conceito, gerando um filtro para a seleção de ligantes considerando um modelo de receptor completamente flexível. Este filtro seleciona do Banco de Dados somente os ligantes que possuem probabilidades geométricas capazes de gerar a docagem molecular em uma cavidade específica do receptor. A função heurística desenvolvida para criar este filtro avalia as estruturas geométricas do ligante e da cavidade do receptor, gerando indicadores que permitam afirmar se um ligante possui ou não a possibilidade de docar em uma determinada cavidade.

Considerado como uma etapa de pré-docagem, além de filtrar os ligantes que não possuem características geométricas favoráveis para docar no modelo de receptor completamente flexível, esta heurística também possibilita o descarte das conformações que não irão gerar bons resultados de FEB. Desta forma, os ligantes que têm os menores índices de conflito com os volumes do receptor são os melhores ranqueados.

Embora ainda existam novos experimentos a serem aplicados para melhor validar a função heurística desenvolvida, os testes apresentados neste trabalho demonstraram resultados muito promissores. Além disso, pode-se afirmar que quanto menor for o volume da cavidade alvo, maior é a possibilidade de filtrar os ligantes por suas características geométricas.

Por ser um trabalho inovador, muitos ajustes foram necessários ao longo do desenvolvimento do trabalho. Ainda assim, ocorreram limitações que neste primeiro momento precisaram ser contornadas. A seção 7.2 apresenta algumas limitações deste trabalho e algumas soluções que estão sendo planejadas.

7.1 Publicações

- Pôster no International Society for Computational Biology (ISCB) 2010 Uruguai abordando o estudo comparativo dos BD de ligantes.

- Pôster no Brazilian Symposium on Bioinformatics (BSB) 2010 apresentando o algoritmo de automatização para a submissão de todo o modelo flexível e recuperação das informações da cavidade alvo.

7.2 Trabalhos futuros

Ao longo do desenvolvimento deste trabalho foram surgindo particularidades, sendo algumas adotadas e evoluindo o trabalho e outras, que devido ao tempo, ficaram sugeridas como etapas futuras para serem melhor analisadas:

- Algumas etapas são apenas um aperfeiçoamento visando a redução do tempo de execução que, embora não seja considerado lento, tem o potencial de ser mais rápido. Nesta etapa destaca-se, principalmente, os problemas com relação a etapa mais demorada, que foi a do processamento do dados no CASTp. Cada arquivo submetido e recuperado com os resultados custa aproximadamente 30 segundos. Trabalhando com modelos flexíveis maiores esta busca no CASTp pode se tornar um gargalo a ser evitado no processo.

- Inicialmente, o planejado considerava apenas uma esfera concêntrica para desenvolver o trabalho dentro da cavidade. No entanto, algumas cavidades selecionadas têm apresentado um fechamento da estrutura, mas não a ponto de fragmentar a cavidade. Nestes casos, pode ser interessante gerar mais de uma esfera concêntrica, trabalhando com uma região maior que uma única cavidade.

- Sabe-se que os algoritmos que consideram as ligações flexíveis têm um ganho maior que os trabalhos considerando apenas as moléculas como rígidas, portanto esta é uma limitação importante a ser resolvida nas etapas seguintes. Para contornar este problema, pode ser estudada uma forma de identificar ângulos de rotação de forma automática e simular parte da flexibilidade dos ligantes. Outro fator essencial, além das propriedades geométricas, é conseguir adicionar alguns estudos já existentes que têm

feito o uso de mapas farmacofóricos para filtrar mais ligantes e obter resultados com maiores chances de ser promissor.

- Reavaliar o método escolhido para fazer a seleção das conformações, visto que esta função heurística acabou sendo desenvolvida para uma cavidade muito específica. O objetivo é que este filtro seja de funcionamento independente da molécula alvo.

- Desenvolver um método capaz de fazer o uso do BD para armazenar os ligantes e acessar de forma mais rápida. Armazenar os resultados e destes resultados eliminarem execuções desnecessárias.

- Melhorar o acesso aos programas desenvolvidos, facilitando o entendimento de outros usuários. Além disso, desenvolver novos testes para a avaliação realizando a docagem molecular e comparar com os valores de FEB e RMSD.

REFERÊNCIAS

[1] Caskey, C. T. “The drug development crisis: efficiency and safety”. *Annual Review of Medicine*, vol. 58, 2007, pp. 1–16.

[2] Kapetanovic, I. M. “Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach”. *Chemico-Biological Interactions*, vol. 171-2, 2008, pp. 165-176.

[3] McInnes, C. “Virtual screening strategies in drug discovery”. *Current Opinion in Chemical Biology*, vol. 11-5, 2007, 494-502.

[4] Lyne, P. D. “Structure-based Virtual Screening: an overview”. *Drug Discovery Today*, vol. 7, 2002, pp. 1047-1055.

[5] Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. “Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein–small molecule complexes”. *Chemical Biology & Drug Design*, vol. 67, 2006, pp. 5-12.

[6] ZINC. “The University of California at San Francisco ZINC database”. Capturado em: http://zinc.docking.org/, Janeiro de 2011.

[7] Luscombe, N. M.; Greenbaum, D.; Gerstein, M. “What is Bioinformatics?: a proposed definition and overview of the field”. *Methods of Information in Medicine*, vol. 4, 2001, pp. 346-358.

[8] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. “A geometric approach to macromolecule–ligand interactions”. *Journal of Molecular Biology*, vol. 161, 1982, pp. 269–288.

[9] Amaro, R. E.; Li, W. W. “Emerging Methods for Ensemble-Based Virtual Screening”. *Current Topics in Medicinal Chemistry*, vol. 10-1, 2010, pp. 3-13.

[10] Lipinski, C. A. “Drug-like properties and the causes of poor solubility and poor permeability”. *Journal Pharmacol Toxicol Methods*, vol. 44-1, 2000, pp.235-249.

[11] Kadam, R. U.; Roy, N. “Recent Trends in Drug-Likeness Prediction: A comprehensive review of In silico methods”. *Indian Journal of Pharmaceutical Sciences*, vol. 69-5, 2007, pp. 609-615.

[12] Amaro, R. E.; Baron, R.; McCammon, J. A. “An improved relaxed complex scheme for receptor flexibility in computer-aided drug design”. *Journal of Computer-Aided Molecular Design*, vol. 22-9, 2008, pp. 693–705.

[13] World Health Organization Global Tuberculosis Program. “Global Tuberculosis Control 2010”. Capturado em: http://www.who.int/tb/publications/2011, Março de 2011.

[14] Schroeder, E. K.; Basso, L. A.; Santos, D. S.; Norberto de Souza, O. “Molecular dynamics simulation studies of the Wild-Type, I21V, and I16T Mutants of isoniazida-resistant Mycobacterium tuberculosis Enoyl Reductase (InhA) in complex with NADH: toward the understanding of NADH-InhA different affinities”. *Biophysical Journal*, vol. 89-2, 2005, pp. 876-884.

[15] Fischer, E. “Einfluss der Configuration auf die Wirkung der Enzyme”. *Chemische Berichte*, vol. 27, 1894, pp. 2985–2993.

[16] Guex, N.; Peitsch, M. C. “Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling”. *Electrophoresis*, vol. 18-15, 1997, pp. 2714-2723.

[17] Barreiro, E. J.; Fraga, C. A. “Química Medicinal- as bases moleculares da ação dos fármacos”. Porto Alegre, Brasil: Artmed, 2008.

[18] Kuntz, I. D. “Structure-based Strategies for Drug Design and Discovery”. *Science*, vol. 257, 1992, pp. 1078-1082.

[19] Lengauer, T.; Rarey, M. “Computational methods for biomolecular docking”. *Current Opinion in Structural Biology*, vol. 6, 1996, pp. 402-406.

[20] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. “Do structurally similar molecules have similar biological activity”. *Journal of Medicinal Chemistry*, vol. 45-19, 2002, pp. 4350-4358.

[21] Andrade, C.; Pasqualoto, K.; Zaim, M.; Ferreira, E. “Abordagem racional no planejamento de novos tuberculostáticos: Inibidores da InhA, enoil-ACP redutase do M. tuberculosis”. *Brazilian Journal of Pharmaceutical Sciences*, vol. 44-2, 2008, pp. 167-179.

[22] Bursavich, M. G.; Rich, D. H. “Designing non-peptide peptidomimetics in the 21st century: inhibitors targeting conformational ensembles”. *Journal of Medicinal Chemistry*, vol. 45-3, 2002, 541-558.

[23] Totrov, M.; Abagyan, R. “Flexible ligand docking to multiple receptor conformations: a practical alternative”. *Current Opinion in Structural Biology*, vol. 18, 2008, pp. 178–184.

[24] Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; et al. “Target flexibility: An emerging consideration in drug discovery and design”. *Journal of Medicinal Chemistry*, vol. 51-20, 2008, pp. 6237-6255.

[25] Carlson, H. A.; McCammon, J. A. “Accommodating Protein Flexibility in Computational Drug Design”. *Molecular Pharmacology*, vol. 57, 2000, pp. 213–218.

[26] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; et al. “AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules”. *Computer Physics Communications*, vol. 91, 1995, pp.1-41.

[27] Clarage, J. B.; Romo, T.; Andrews, B. K.; Pettitt, B. M.; Phillips Jr, G. N. “A sampling problem in molecular dynamics simulations of macromolecules”. *Biophysics*, vol. 92, 1995, pp. 3288-3292.

[28] PDB. “RCSB Protein Data Bank”. Capturado em: http://www.rcsb.org/pdb, Janeiro de 2011.

[29] Tripos. “Tripos- A Certara Company”. Capturado em: http://tripos.com/data/support/mol2.pdf, Janeiro de 2011.

[30] Irwin, J. J.; Shoichet, B. K. “ZINC -- a free database of commercially available compounds for virtual screening”. *Journal of Chemical Information and Modeling*, vol. 45-1, 2005, pp. 177-182.

[31] Cochrane, G. R.; Galperin, M. Y. “The 2010 nucleic acids research database issue and online database collection: a community of data resources”. *Nucleic Acids Research*, vol. 38, 2009, pp. D1-D4.

[32] Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; et al. “ChemBank: a small-molecule screening and cheminformatics resource database”. *Nucleic Acids Research*, vol. 36, 2008, pp. D351–D359.

[33] Chen, J. H.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. “ChemDB update: full-text search and virtual chemical space”. *Bioinformatics*, vol. 23-17, 2007, pp. 2348-2351.

[34] Chen, J. H.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. “ChemDB: a public database of small molecules and related chemoinformatics resources”. *Bioinformatics*, vol. 21-22, 2005, pp. 4133-4139.

[35] Masciocchi, J.; Frau, G.; Fanton, M.; Sturlese, M.; Floris, M.; Pireddu, L.; et al. “MMsINC: a large-scale chemoinformatics database”. *Nucleic Acids Research*, vol. 37, 2009, pp. D284-D290.

[36] Ihlenfeldt, W.-D.; Voigt, J. H.; Bienfait, B.; Oellien, F.; Nicklaus, M. C. “Enhanced CACTVS Browser of the Open NCI Database”. *Journal of Chemical Information and Modeling*, vol. 42, 2002, pp. 46-57.

[37] Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. “Molecular biology: NIH Molecular Libraries Initiative”. *Science*, vol. 306-5699, 2004, pp. 1138–1139.

[38] Daylight. “Daylight Chemical Information Systems”. Capturado em: http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html, Janeiro de 2011.

[39] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; et al. “DrugBank: a comprehensive resource for in silico drug discovery and exploration”. *Nucleic Acids Research*, vol. 34, 2006, pp. D668-D672.

[40] OpenEye. “OpenEye Scientific Software”. Capturado em: http://www.eyesopen.com/, Dezembro de 2010.

[41] Talete. “Talete - Dragon 6:”. Capturado em: http://www.talete.mi.it/products/dragon_description.htm, Janeiro de 2011.

[42] Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. “The Design of Leadlike Combinatorial Libraries”. *Angewandte Chemie International Edition*, vol. 38-24, 1999, pp. 3743-3748.

[43] Carr, R. A.; Congreve, M.; Murray, C. W.; Rees, D. C. “Fragment-based lead discovery: leads by design”. *Drug Discovery Today*, vol. 10-14, 2005, pp. 987-992.

[44] Irwin, J. J. “Using ZINC to acquire a virtual screening library”. *Current Protocols in Bioinformatics,* 2008, pp. 14.6.1-14.6.23.

[45] Binkowski, T. A.; Naghibzadeh, S.; Liang, J. “CASTp: Computed Atlas of Surface Topography of proteins”. *Nucleic Acids Research*, vol. 31-13, 2003, pp. 3352–3355.

[46] Quemard, A.; Dessen, A.; Sugantino, M.; Jacobs, W. R.; Sacchettini, J. C.; Blanchard, J. S. “Binding of Catalase-Peroxidase-Activated Isoniazid to Wild-Type and Mutant Mycobacterium tuberculosis Enoyl-ACP Reductases”. *Journal of the American Chemical Society*, vol. 118, 1996, pp. 1561-1562.

[47] Rozwarski, D. A.; Vilchèze, C.; Sugantino, M.; Bittman, R.; Sacchettini, J. C.; “Crystal structure of the Mycobacterium tuberculosis enoyl-ACP reductase, InhA, in complex with NAD+ and a C16 fatty acyl substrate”. *The Journal of Biological Chemistry*, vol. 274, 1999, pp. 15582-15589.

[48] Kleywegt, G. J.; Jones, T. A. “Detection, delineation, measurement and display of cavities in macromolecular structures”. *Acta Crystallographica*, vol. D50, 1994, pp. 178-185.

[49] Peters, K. P.; Fauck, J.; Frömmel, C. “The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria”. *Journal of Molecular Biology*, vol. 256, 1996, pp. 201-213.

[50] Binkowski, T. A.; Freeman, P.; Liang, J. “pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins”. *Nucleic Acids Research*, vol. 32, 2004, pp. W555-W558.

[51] Laskowski, R. A. “SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions”. *Journal of Molecular Graphics*, vol. 13, 1995, pp. 323-330.

[52] Edelsbrunner, H.; Mücke, E. P. “Three-dimensional alpha shapes”. *ACM Transactions Graphics*, vol. 13, 1994, pp. 43-72.

[53] Richards, F. M. “Calculation of molecular volumes and areas for structures of known geometry”. *Methods Enzymol*, vol. 115, 1985, pp. 440-464.

[54] Connolly, M. L. “Analytical molecular surface calculation”. *Journal of Applied Crystallography*, vol. 16, 1983, pp. 548-558.

[55] Voss, N. R.; Gerstein, M. “3V: cavity, channel and cleft volume calculator and extractor”. *Nucleic Acids Research*, vol. 38, 2010, pp. W555–W562.

[56] Ruby. “Ruby - A programmer´s best friend”. Capturado em: http://www.ruby-lang.org, Janeiro de 2011.

[57] Campbell, S. J.; Gold, N. D.; Jackson, R. M.; Westhead, D. R. “Ligand binding: functional site location, similarity and docking”. *Current Opinion in Structural Biology*, vol. 13-3, 2003, pp. 389-395.

[58] Laskowski, R. A.; Luscombe, N. M.; Swindells, M. B. “Protein clefts in molecular recognition and function”. *Protein Science*, vol. 5, 1996, pp. 2438-2452.

[59] Klebe, G.; Sotriffer, C. “Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design”. *Farmaco*, vol. 57-3, 2002, pp. 243-251.

[60] Liang, J.; Edelsbrunner, H.; Woodward, C. “Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design”. *Protein Science*, vol. 7, 1998, pp. 1884-1897.

[61] Bondi, A. “Van der Waals Volumes and Radii”. *Journal of Physical Chemistry*, vol. 68-3, 1964, pp. 441–451.

[62] Rowland, R. S.; Taylor, R. “Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der waals Radii”. *Journal of Physical Chemistry*, vol. 100, 1996, pp. 7384-7391.

[63] Gibson, K. D.; Scheraga, H. A. “Volume of the Intersection of Three Spheres of Unequal Size. A Simplified Formula”. *The Journal of Physical Chemistry*, vol. 91, 1987, pp. 4121-4122.

[64] Buša, J.; Džurina, J.; Hayryan, E.; Hayryan, S.; Hu, C.-K.; Plavka, J.; et al. “ARVO: A Fortran package for computing the solvent accessible surface area and the excluded volume of overlapping spheres via analytic equations”. *Computer Physics Communications*, vol. 165, 2005, pp. 59–96.

[65] Schlick, T.; Skeel, R. D.; Brunger, A. T.; Kalé, L.; Board, J. J.; Hermans, J.; et al. “Algorithmic Challenges in Computational Molecular Biophysics”. *Journal of Computational Physics*, vol. 151, 1999, pp. 9-48.