UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

ANDERSON SOUZA ROCHA

IMPLEMENTAÇÃO DE ALGORITMO PARA CÁLCULO DE SIMILARIDADE MOLECULAR NO SISTEMA NATPRODB

FEIRA DE SANTANA
2016

ANDERSON SOUZA ROCHA

IMPLEMENTAÇÃO DE ALGORITMO PARA CÁLCULO DE SIMILARIDADE MOLECULAR NO SISTEMA NATPRODB

Trabalho de Conclusão de Curso apresentado ao Colegiado do curso de Engenharia de Computação como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação pela Universidade Estadual de Feira de Santana.

Orientador: Prof. Dr. Ângelo Amâncio Duarte

Co-Orientador: Prof. Dr. Manoelito C. dos Santos Jr.

FEIRA DE SANTANA
2016

RESUMO

O processo de desenvolvimento de um fármaco é realizado em diversas etapas, que vão desde o projeto e estudo de um possível composto farmacofórico até sua sintetização. Grande parte dos esforços empregados para o desenvolvimento de um novo medicamento são realizados de forma assistida em computadores. As ferramentas computacionais permitem aos pesquisadores tanto catalogar de forma mais eficiente suas estruturas de estudo, quanto avaliar de forma mais rápida características estruturais, bioquímicas, e o comportamento de moléculas ao sofrer variações estruturais. Além disso, algumas dessas ferramentas também realizam triagem em bancos de dados moleculares baseados no conceito de similaridade molecular, que é um conceito chave no qual se baseiam grande parte dos esforços dirigidos ao projeto de um novo fármaco. O presente trabalho descreve a adaptação de um algoritmo apropriado para computação de similaridade molecular em um software desenvolvido para catalogação de moléculas em banco de dados, visando possibilitar ao pesquisador procurar em sua própria base de dados, estruturas similares a moléculas de seu interesse de estudo. Um diferencial deste trabalho é que essa ferramenta será disponibilizada livremente para qualquer usuário, e devido às características inerentes à sua implementação, em contrapartida à outras ferramentas disponíveis para o mesmo fim, este software não necessitará de muitos recursos de hardware para sua execução, ou seja, é compatível com computadores comuns, e também com qualquer sistema operacional utilizado pelo pesquisador. De forma resumida, o sistema se comporta da seguinte forma: Uma vez cadastradas as moléculas e seus respectivos SMILES no banco de dados do usuário, o pesquisador poderá consultar em seu banco se há ocorrência de estruturas com um grau de similaridade desejado à molécula inserida para consulta. O sistema converte os SMILES da molécula de entrada, e das presentes no banco em fingerprints para aplicação da métrica de Tversky para obtenção de um coeficiente de similaridade, retornando por fim para o pesquisador todas as moléculas do seu banco que possuem um grau de similaridade igual ou superior ao desejado (em relação à molécula de entrada). O algoritmo adotado para computação de similaridade molecular foi testado através de uma adaptação do conceito de matriz de confusão e teve como entrada 7 moléculas já disponibilizadas como fármacos, e para os testes realizados no sistema foram verificados que a classificação de moléculas com coeficiente de similaridades $\geq$ 0,8 no sistema se aproximou do comportamento de uma classificação ideal.

Palavras-chave: Similaridade Molecular. Banco de Dados Moleculares. Química Computacional.

ABSTRACT

The developing process of a medicine has many stages. Since the project and study of a possible pharmacophore compound until its synthesizing. Most efforts on the development of a new medicine are aided by computers. The computational tools allow the researchers both cataloguing in a more efficient way their study structures and evaluate in a faster way biochemical, structural features, and the molecules behavior during structural variations.Moreover, some of these tools also helps in the search for similar structures, that is a key concept in which most efforts towards to develop a new medicine are based. This work describes the adaptation of appropriate algorithms for computing molecular similarities in a software developed for molecules tabulation in a database enabling the researcher to screen similar structures to the ones that are in his study concerns at his own database. A differential in this work lies in the fact that this tool is made freely available to any user, and due the inner characteristics of its implementation, face to others tools that exist with the same purpose, this software does not need much hardware resources to run, in other words, it is compatible with common computers, and any operational system used by the researcher. In resume, the system works according to these steps: once the molecules are registered in the user database, the researcher is able to verify in his base if there are structures with a desired degree of similarity to the molecule queried. So, the system converts the inserted molecule, and those ones in the database in fingerprints format in order to use the Tversky's metrics obtaining a table of similarity, and finally retrieving to the researcher all molecules in his database whose similarity coefficient it is equal or superior to the one desired (regarded to the queried molecule).

Keywords: Molecular Similarity. Molecular Databases. Computational Chemistry.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

NatProDB	Natural Products Data Bank
SMILES	Simplified Molecular Input Line Entry System
NCBI	National Center for Biotechnology Information
FPR	porcentagem de amostras erroneamente classificadas como positivas dentre todas as verdadeiramente negativas
LACAD	Laboratório de Computação de Alto Desempenho
LMM	Laboratório de Modelagem Molecular

SUMÁRIO

1 INTRODUÇÃO

O processo de desenvolvimento de um novo fármaco envolve diversas etapas que englobam desde a pesquisa de um determinado alvo biológico, até a descoberta de compostos com atividades biológicas desejadas e com potencial para se tornarem medicamentos a serem comercializados. Durante esse processo, inúmeras ferramentas e abordagens computacionais podem ser aplicadas visando auxiliar o pesquisador no estudo dos compostos, e também acelerar o desenvolvimento do fármaco. Nos últimos anos, devido a introdução de abordagens computacionais, principalmente nas fases inicias do processo de desenvolvimento de um fármaco, onde o foco do pesquisador é o estudo de um determinado alvo biológico, tem difundido técnicas de desenvolvimento baseadas no ligante (Ex: Similaridade molecular, modelo farmacofórico) possibilitado assim a identificação de moléculas protótipos para ensaios biológicos (RODRIGUES *et al*., 2012). Dentre essas técnicas baseadas no ligante, um conceito já popularizado na comunidade científica é o conceito de similaridade molecular, o qual preconiza que, moléculas que possuem estruturas similares, provavelmente compartilhem propriedades físico-químicas, e atividades biológicas semelhantes (SINGH, 2004). Dessa maneira, o princípio do processo de desenvolvimento de um fármaco resume-se ao estudo de um determinado alvo biológico para o desenvolvimento de um composto ligante capaz de interagir com o alvo obtendo uma atividade biológica desejada, e em alguns casos realizar uma triagem em bancos de dados moleculares buscando por compostos similares ao ligante em questão (Virtual Screening).

Nessa perspectiva, os esforços para o desenvolvimento e sintetização de um composto ligante podem ser realizados através de sistemas computacionais que aplicam o conceito de similaridade molecular em três grandes eixos segundo Kubinyi *et al.* (2008): a) Exploração computacional e bioquímica de moléculas com estruturas conhecidas (sintetizadas ou não); b) Desenvolvimento de modelos computacionais para estudo de como variações na estrutura molecular afetam a atividade molecular ou as propriedades da molécula; c) Exame de bancos de dados moleculares visando obtenção de um composto similar à estrutura do ligante projetado pelo pesquisador.

Algumas ferramentas já tem auxiliado pesquisadores nesse sentido como por exemplo o ZINC (IRWIN; SHOICHET, 2005) e PUBCHEM (LI *et al*., 2010), ambas ferramentas web que disponibilizam bancos de dados com uma diversidade de moléculas, e também implementam algoritmos para computação de similaridade molecular, permitindo ao pesquisador realizar uma triagem em suas bases de

dados a procura de um composto com determinado grau de similaridade a uma determinada estrutura molecular de interesse. Apesar dessas ferramentas auxiliarem pesquisadores a realizarem seus estudos, ainda sofrem de limitação de representação do espaço químico, onde apesar do grande número de moléculas já catalogadas em seus bancos de dados moleculares, o usuário tem seu universo de pesquisa limitado às estruturas disponíveis nesses bancos de dados. Outro problema relacionado a utilização dessas ferramentas já disponíveis é que nenhuma delas fornece ao pesquisador a possibilidade de criação de base de dados com moléculas de sua propriedade, e em alguns casos o usuário acaba disponibilizando suas estruturas em bancos colaborativos para que possam utilizar essas ferramentas para desenvolvimento de suas pesquisas, correndo riscos inclusive de perda de seus direitos autorais sobre os seus compostos.

Neste trabalho será descrita a implementação de um algoritmo para realização de triagem virtual baseada no conceito de similaridade molecular no sistema de banco de dados de moléculas, oriundas de fontes naturais endêmicas do bioma semiárido, denominado Natural Products Data Bank (NatProDB), de domínio público, para utilização em modo local (não conectado à Internet), visando: 1) Facilitar a usuários não especialistas em computação, a catalogação de moléculas e manutenção de bancos de dados moleculares, sem a necessidade de uso de bases de dados na Internet; 2) Prover mecanismos para cálculo de similaridade entre moléculas de interesse frente as moléculas depositadas no banco. A computação de similaridade implementada neste sistema é realizada através da métrica de Tversky (coeficiente de similaridade), aplicada sobre a representação computacional de moléculas através de *fingerprints*. A implementação deste método é realizada por uma biblioteca livre denominada Indigo toolkit, que realiza manipulação de moléculas e sub-estruturas (PAVLOV *et al.*, 2011). Para os testes foi criado um banco de dados para o sistema com um conjunto de moléculas de um banco colaborativo disponibilizado pelo ZINC, e foram enxertadas nesse banco moléculas com grau de similaridade superior a 80% a um conjunto de moléculas de entrada já testados e utilizados pela indústria farmacêutica. A avaliação do sistema foi realizada verificando os resultados obtidos pelo NatProDB, e avaliando através de matrizes de confusão a capacidade do sistema de classificar corretamente as moléculas 0.8 ou 80% similares ou não. Os resultados destes testes, assim como os detalhes da implementação deste sistema, serão detalhados nas próximas seções deste trabalho.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Implementar algoritmo para permitir o processo de triagem virtual de moléculas baseado no conceito de similaridade molecular no sistema NatProDb

1.1.2 Objetivos Específicos

- Implementar um algoritmo para cálculo de similaridade no sistema NatProDB.
- Avaliar a eficiência da métrica de similaridade 2D no NatProDB

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será discutido na seção 2.1, algumas das ferramentas que já se propõe a realizar a comparação de moléculas, assim como alguns dos algoritmos que podem ser utilizados na análise de similaridade molecular que serão apresentados na seção 2.3, sendo que parte desses algoritmos já estão adaptados em algumas das ferramentas a serem apresentadas. Também será exposto na seção 2.2 algumas formas de representação de uma molécula em um ambiente computacional, e sua importância para escolha de qual método de comparação de moléculas deverá ser utilizado.

2.1 FERRAMENTAS PARA COMPUTAÇÃO DE SIMILARIDADE MOLECULAR

Devido a importância das ferramentas computacionais no processo de descoberta de fármacos, há diversos esforços dedicados ao desenvolvimento destas ferramentas, tornando-as cada vez mais eficientes e eficazes. Como resultado desses esforços, já se encontram disponíveis alguns sistemas computacionais que aplicam algoritmos de similaridade, realizam busca em bancos de dados, predição de propriedades físico-químicas e atividades biológicas, entre outras diversas funcionalidades importantes na pesquisa e desenvolvimento de fármacos. Esses softwares diferenciam-se de acordo com o custo de sua licença, funcionalidades implementadas, formatos de moléculas aceitos, representação computacional da molécula, método de computação de similaridade molecular, entre outros aspectos.

Uma das ferramentas populares entre pesquisadores na área de química medicinal é o ZINC. O sistema web ZINC é uma ferramenta livre que fornece para o usuário um banco de dados com mais de 21 milhões de estruturas catalogadas (IRWIN; SHOICHET, 2005). Além disso, o sistema armazena diversas informações sobre cada molécula como: Massa molecular, centros quirais, coeficiente de partição água-etanol calculado (cLogP). O ZINC armazena estruturas em formatos 2D, tal como o formato Simplified Molecular Input Line Entry System (SMILES), que são representações lineares que utilizam uma sequência de caracteres para representação da estrutura molecular (KUMAR, 2012). Além do formato SMILES, o ZINC também aceita como formatos de entradas, Structure Data Format (SDF) e MOL2. O ZINC implementa rotinas para triagem virtual em sua base de dados utilizando os conceitos de similaridade molecular, permitindo ao pesquisador selecionar o grau de similaridade mínimo desejado, e retornando todas as moléculas

em seu catálogo (banco de dados) com grau de similaridade igual ou superior ao selecionado pelo usuário (IRWIN; SHOICHET, 2005).

Outra ferramenta que pode ser utilizada para a comparação de similaridade entre moléculas é o software ANACONDA (DEVILLERS, 1996). Tal sistema realiza a computação da similaridade entre duas moléculas através da comparação de propriedades de suas respectivas superfícies moleculares utilizando projeção gnomônica . O emprego deste método permite a comparação de forma interativa de dois componentes, podendo sugerir modos de superposição entre as moléculas, e assim uma possível geração de um modelo famacofórico (DEVILLERS, 1996). Esta ferramenta apesar de apresentar bons resultados na comparação de similaridade molecular, não permite a comparação de mais de duas moléculas por consulta, e além disso também não fornece ao pesquisador qualquer ferramenta para construção, ou manipulação de um banco de dados para catalogação das suas moléculas de estudo, exigindo assim um esforço para o mesmo catalogar suas moléculas de interesse, e realizar uma computação praticamente serial de similaridade entre moléculas que deseja-se estudar

Utilizando uma abordagem para computação de similaridade molecular baseada na comparação de campos elestrostáticos e de campos de volume esférico, Mestres, Rohrer & Maggiora (1997) descrevem um software denominado MIMIC, que implementa rotinas para realização de triagem virtual baseada na comparação de similaridade molecular. Este software permite ao pesquisador obter um índice de similaridade de alta precisão, pois consegue levar em consideração, além da estrutura da molécula, a contribuição de cada átomo na computação da similaridade molecular. Assim como descrito por Devillers (1996), o MIMIC não realiza manipulação de banco de dados, por tanto necessita que o pesquisador insira as moléculas que deseja comparar na entrada do sistema de forma serial. Dessa maneira, para criar um banco de dados para catálogo de moléculas, o usuário necessita utilizar uma outra ferramenta para manipular bancos de dados, o que nem sempre é conveniente para um pesquisador que possua muitas moléculas para estudo, e que por muitas vezes não possui conhecimentos avançados de computação para fazer tal catalogação por conta própria.

Os esforços para implementação de ferramentas computacionais para cálculo de similaridade molecular não se restringem a programas de computador e/ou web sites. Alguns pesquisadores já tem desenvolvido bibliotecas multi-plataformas, capazes de interagir com várias linguagens de programação, e que implementam não

somente rotinas para manipulação de moléculas em geral, mas também implementam algoritmos para computação de similaridade molecular. Nessa perspectiva, a Indigo Toolkit surge como uma das bibliotecas de acesso livre mais completas. Desenvolvida pela GGASoftware, e atualmente mantida pelo epam lifescience, esta ferramenta é capaz de manipular os principais formatos disponíveis para representação de moléculas como: SMILES, SDF, Molfile, entre outros (PAVLOV *et al.*, 2011). O conjunto de ferramentas disponibilizado por essa biblioteca, permite ao usuário manipular moléculas, computar similaridade, buscar sub-estruturas e reações. A API indigo é capaz de utilizar diversos tipos de descritores moleculares, desde SMILES, até fingerprints (sequência de bits que representa presença ou ausência de uma determinada característica estrututral) e comparando moléculas através de diversas métricas como: coeficiente de Tanimoto, métrica euclidiana, e métrica de Tversky. Para sua utilização é necessário que esta biblioteca seja incorporada a um programa, que pode ser escrito em C/C++, java ou python. Esta biblioteca é utilizada na implementação do NatProDB, conforme será discutido posteriormente.

Utilizando uma abordagem de busca farmacofórica, Pharmer é considerada uma ferramenta bastante robusta, e já tem sido utilizada até mesmo pelo Zinc no processo de triagem de moléculas similares em seu banco de dados (KOES; CAMACHO, 2011). O grande diferencial dessa ferramenta é a velocidade de processamento e computação de similaridade molecular em grandes bases de dados, chegando a ser uma ordem de magnitude mais rápido que as ferramentas computacionais já existentes (KOES; CAMACHO, 2011). O motivo para tal desempenho reside no fato de que em detrimento das demais ferramentas para triagem em bancos de dados moleculares através de modelos famacofóricos (que normalmente realizam a comparação serial de todas as moléculas de sua base de dados), o Pharmer utiliza uma estrutura de organização dados adaptada, denominada Pharmer KDB-tree data structure, e implementa um método de compração baseado em técnicas de computação visual: hashing geométrico e transformada generalizada de Hough. Através da utilização de tais tecnicas, o sistema não somente reduz o custo computacional para comparação de moléculas (grau de similaridade de modelos farmacofóricos), mas também reduz o tempo necessário para consulta no banco uma vez que a estrutura organizacional dos dados em sua base permite um direcionamento do sistema para alvos com maior probabilidade de serem similares.

Outra ferramenta pública que também implementa filtros baseados no conceito de similaridade molecular para triagem virtual de moléculas em bancos de dados

(*Virtual Screening*) é o Pubchem (LI *et al.*, 2010). Este sistema possui mais de 25 milhões de estruturas catalogadas em seu banco de dados, e é mantido pelo *National Center for Biotechnology Information* (NCBI). O Pubchem é um dos sistemas bastantes difundidos na comunidade acadêmica devido ao fato de manipular moléculas sob diversos formatos como (SDF, Mol, SMILES). Além disso, assim como o ZINC (IRWIN; SHOICHET, 2005) por se tratar de uma fonte de pesquisa pública, e de acesso via web, pesquisadores podem utilizar seus recursos computacionais independente de sua localização geográfica, auxiliando pesquisadores em diversos lugares no processo de desenvolvimento de ferramentas para modular processos biológicos e também na identificação de compostos com probabilidade de se tornarem medicamentos utilizados em tratamentos de doenças.

2.2 DESCRITORES MOLECULARES

O descritor molecular pode ser considerado como o resultado da aplicação de procedimentos lógicos e matemáticos que transformam uma representação química codificada, em uma representação simbólica de uma molécula, em um formato padrão ou resultado de algum experimento padronizado, de forma a facilitar a manipulação dessas estruturas (TODESCHINI *et al.*, 2008). Para o presente trabalho, o descritor molecular consiste no formato sobre o qual uma molécula é representada computacionalmente, e tal formato é importante para a computação de similaridade, pois o descritor molecular é considerado um fator determinante da métrica a ser aplicada para cálculo da similaridade estrutural (TODESCHINI *et al.*, 2008). Nesta seção será apresentado alguns descritores utilizados por sistemas que calculam similaridade entre moléculas.

2.2.1 FingerPrints

As Fingerprints podem ser consideradas como um descritor molecular complexo. Nesta forma representação computacional de uma molécula, todas as características e grupos funcionais presentes em um dado composto, são codificados em *bitstreams* (Sequencia de bits 0 e 1) únicos para cada estrutura (XUE; BAJORATH, 2000). Além dos grupos funcionais, e características inerentes a uma determinada molécula, as fringerprints também são capazes de armazenar em sua estrutura as distancias entre estruturas do composto, caminhos conexos dessa através de toda a estrutura molecular, ou diferentes tipos de farmacóforos de interesse do pesquisador (TODESCHINI *et al.*, 2008).

Uma das vantagens na utilização de fingerprints na química computacional, é a capacidade de armazenamento de características intrísecas de um composto em um formato de relativamente fácil manipulação por sistemas computacionais. Por exemplo, dentre os modelos de fingerprints mais utilizados na química-computacional, podemos citar a Daylight fingerprint, que utiliza em torno de 2048 bits para armazenamento de propriedades, características, e distâncias moleculares de um único composto, sendo que para esse tipo de descritor não ocorre uma relação direta entre um bit e um determinado grupo funcional/característica (como em modelos de figerprints mais simples), na verdade determinadas caracteristicas, distancias, e/ou propriedades podem ser mapeadas através de algorítmos de *Hashing* para fornecer assim padões de bits cada vez mais específicos, permitindo maior fidelidade na representação das propriedades de uma molécula em um sistema computacional (XUE; BAJORATH, 2000).

Outra importante vantagem da utilização de figerprints para representação computacional de uma molécula reside na simplicidade de aplicação de métodos e métricas para comparacação moléculas baseado no conceito de similaridade molecular (XUE; BAJORATH, 2000). Nesta perspectiva, devido as características intrísicas desse descritor molecular, o processo de triagem virtual em um banco de dados molecular, pode ser realizado através da geração de fingerprints para cada uma das moléculas do banco e da molécula de consulta, e da aplicação de métricas de similaridade molecular de relativamente fácil implementação em sistemas computacionais como: Coeficiente de Tanimoto, Coeficiente de Tversky, Dice entre outras, que basicamente obtem um determinado coeficiente de similaridade entre duas estruturas moleculares através da realização de um cálculo estatístico que leva em consideração as estruturas e caracteristicas compartilhadas e as singulares das moléculas comparadas. No sistema NatProDB, as fingerprints possuem um papel fundamental no processo de triagem de moleculas de seu banco de dados, uma vez que tal descritor é utilizado para comparação de moléculas conforme será descrito mais aprofundadamente no próximo capítulo.

2.2.2 Descritores Farmacofóricos 3D

Os descritores farmacofóricos tridimensionais quantificam propriedades e as distâncias entre farmacóforos biológicos importantes para a interação entre o ligante e receptor (BAJORATH, 2004). Dentre essas propriedades de interesse de pesquisadores podemos citar: grupos funcionais ou características chaves em

determinada orientação, doadores/receptores de ligações de hidrogênio, partes de uma molécula, entre outras. Normalmente, um descritor farmacofórico 3D são compostos basicamente por 3 ou 4 características, e de 3 a 6 distancias entre elas. Quando a conformação biológica entre o ligante e receptor é conhecida, é possível identificar que características/grupos funcionais são cruciais para ocorrencia dessa ligação. Caso contrário, através de técnicas de triagem virtual (*virtual screening*) pesquisadores podem identificar moléculas que contenham esses grupos funcionais/caracteristicas, ou até mesmo projetar um descritor farmacofórico 3D (BAJORATH, 2004). A aplicação de ferramentas computacionais na comparação de moléculas representadas como descritores farmacofóricos 3D pode ser realizada através do conceito de similaridade molecular aplicando uma transformação nesse descritor em uma *figerprint* (Sequência de bits única para uma molécula), onde bits definidos como ’1’ simbolizam a presença de uma determinada caracteristica ou grupo funcional, e bits definidos como ’0’ a ausência de tal estrutura na molécula. Sendo assim, ferramentas computacionais podem calcular a similaridade entre duas moléculas gerando figerprints para cada um dos descritores farmacofóricos 3D e aplicando alguma das métricas já conhecidas para calculo de similaridade entre figerprints, como por exemplo coeficiente de Tanimoto, e coeficiente de Tversky, ambos serão descritos com maiores detalhes em sessões posteriores.

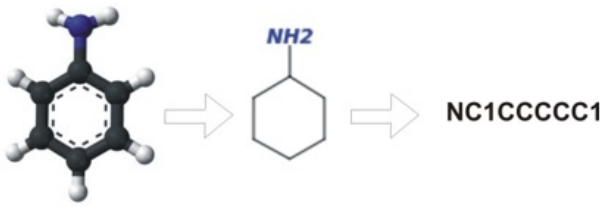
2.2.3 SMILES (Simplified Molecular-Input Line-Entry System)

O SMILES surgiu como descritor molecular para resolver não somente o problema de armazenamento de estruturas moleculares em um computador, mas também facilitar a manipulação dessas estruturas moleculares por softwares, e também facilitar buscas por estruturas moleculares em base de dados na internet internet (KUMAR, 2012). Ainda segundo (KUMAR, 2012) é demonstrado que pode-se considerar o SMILE como uma linguagem que permite ao pesquisador representar uma molécula através de uma notação linear padronizada, que muito se assemelha à notação utilizada comumente na literatura.

Em uma representação de molécula em formato de SMILES cada átomo é representado por seu símbolo atômico, onde os hidrogênios são omitidos (ANDERSON; VEITH; WEININGER, 1987). Os átomos vizinhos, são representados próximos entre si, e em caso de ligação dupla ou tripla, entre os átomos unidos por tais ligações aparecerão respectivamente os símbolos = e #. É importante salientar também que ramificações são representadas por parêntesis, e os anéis aromáticos

pela alocação de dígitos entre os dois átomos que fecham o anel. A figura 2 exemplifica a conversão de uma estrutura tridimensional em uma notação SMILE, e como tal representação se assemelha a forma como a literatura

Figura 1: Exemplo da representação de uma estrutura 3D em SMILES.



Fonte: Autor

O formato do SMILE gerado para a molécula não é necessariamente único, existem outras representações possíveis, dependendo da estratégia de notação adotada pelo pesquisador. Dessa forma é muito comum, para facilitar a busca em bancos de dados, a utilização de um formato SMILE canônico, ou seja, que padroniza qualquer SMILE para um formato único, o que simplifica a manipulação de qualquer molécula, sem aumentar a complexidade dessa operação para o pesquisador (KUMAR, 2012).

Este descritor molecular, além de necessitar de recursos computacionais relativamente baixos para representação de uma molécula, devido sua estrutura, é bastante utilizado no desenvolvimento de fármacos, principalmente nas atividades que envolvem a busca de estruturas similares com uma desejada atividade biológica, pois por ser um descritor molecular linear, ele permite a simplificação de uma comparação de moléculas realizadas por exemplo em uma forma tridimensional, em uma comparação linear de cadeias de caracteres, de mais fácil manipulação em um computador.

2.3 ALGORITMOS E MÉTRICAS PARA COMPUTAÇÃO DE SIMILARIDADE MOLECULAR

A busca por similaridade molecular é um ponto importante no projeto de um fármaco, pois tal conceito é um dos norteadores para triagem virtual de moléculas de interesse em bancos de dados. Os algorítmos e métricas utilizados na comparação de moléculas baseados no conceito de similaridade são dependentes do descritor molecular utilizado, logo a orientação do método de comparação selecionado

deve ser direcionada pela escolha da melhor forma para representar a molécula computacionalmente (TODESCHINI; CONSONNI, 2000). Tal afirmação permite inferir que uma escolha não-adequada de descritor molecular para aplicação de um algoritmo pode inviabilizar a aplicação do mesmo, assim como retornar resultados indesejados, ou não fidedignos com a literatura, prejudicando assim a pesquisa.

2.3.1 Coeficiente de Tanimoto

Existem diversas métricas utilizadas para determinação de similaridade entre duas moléculas, em geral elas geram um score indicando o grau de similaridade dos compostos comparados. Algumas dessas métricas utilizam distancias euclidianas para determinar esse score, como no caso das métricas de Hamming, e Euclidiana. É comum também a utilização de coeficientes de associação tais como: Tanimoto, Dice, e coeficientes de cossenos.

O coeficiente de Tanimoto realiza a comparação de moléculas através de uma simples contagem de características compartilhadas (grupos funcionais, propriedades, etc) entre as moléculas submetidas à comparação. Essa contagem gera um determinado valor entre 0 e 1 que indica o grau de similaridade entre as estruturas verificadas (DOGRA, 2007).

O coeficiente de Tanimoto pode ser utilizado na comparação de descritores moleculares 2D. Nessa perspectiva, este coeficiente se mostra bastante efetivo na comparação de similaridade entre moléculas representadas computacionalmente como *fingerprints*, pois devido a estrutura desse descritor molécular, a comparação de similaridade pode ser realizada através de op erações booleanas em cadeias binárias para contagem do número de caracteristicas em comum entre cada par de fingerprints (WILLETT, 2003). Dessa maneira, a aplicação do coeficiente de Tanimoto para medir similaridade entre duas fingerprints A e B pode ser realizada tomando N_A como o número de caracteríticas presentes em A (bits iguais a 1 em A), N_B como o número de caracteríticas presentes em B (bits iguais a 1 em B), N_{AB} como o número de caracteríticas compartilhadas por A e B (bits iguais a 1 em A e B), e finalmente o coeficiente de Tanimoto T_C pode ser obtido aplicando a equação(8)

$$T_C = \frac{N_{AB}}{N_A + N_B - N_{AB}} \tag{8}$$

Nota-se em na equação (8), que somente são considerados na computação do coeficiente de Tanimoto a presença das estruturas (características), a ausência não interfere neste processo, o que otimiza a comparação dessas estrutura (DOGRA,

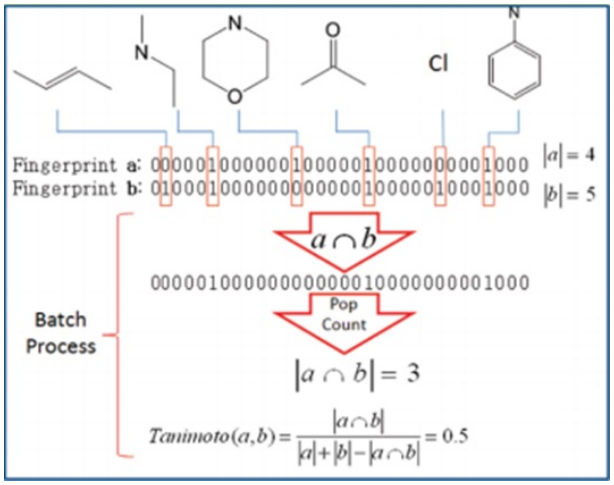
2007).

2.3.1.1 Aplicação do Coeficiente de Tanimoto na Comparação de Fingerprints

Seguindo a descrição do coeficiente de Tanimoto apresentada por (DOGRA, 2007), o trabalho de (MA; WANG; XIE, 2011) descreve a utilização dessa métrica para comparação de fingerprints. Em resumo, cada fingerprint presente no banco de dados do autor é representada como vetores de bits com 992 posições, ou seja, para cada molécula são verificadas a presença ou ausência de 992 estruturas (características) (DOGRA, 2007).

Após verificada essa presença ou ausência dessas estruturas nas moléculas, ou seja gerada a fingerprint, é aplicada a equação (8) sobre as moléculas que se deseja comparar. Assim é obtido o score de similaridade entre as estruturas, e verifica-se se há similaridade ou não entre as moléculas submetidas, conforme podemos observar na figura (2).

Figura 2: Aplicação do coeficiente de Tanimoto para Computar Similaridade entre Fingerprints.



Fonte: (MA; WANG; XIE, 2011)

Apesar foco central do trabalho apresentado por (MA; WANG; XIE, 2011), não ser a comparação de similaridade molecular, mas sim a paralelização do processo de comparação de similaridade entre bancos de dados moleculares, a figura 2 extraída de sua obra, ilustra o procedimento de comparação entre fingerprints, exemplificando desde a contagem de estruturas (características) na fingerprint *a* , e na fingerprint *b*, assim como a intersecção entre as características de ambas estruturas ($a \bigcap b$), e por fim a computação do coeficiente de Tanimoto. De acordo com (DOGRA, 2007), o

resultado do coeficiente de Tanimoto (0,5) obtido na figura (2) indica baixa similaridade entre as moléculas a e b.

2.3.2 Coeficiente de Tversky

O coeficiente de Tversky pode ser considerado como uma metrica generalizada do coeficiente de Tanimoto (SWAMIDASS; BALDI, 2007). Assim como o coeficiente de Tanimoto, o coeficiente de Tversky é uma métrica bastante utilizada para obtenção de índice de similaridade entre moléculas representadas sob descritores moleculares 2D (BAJORATH, 2004).

Enquanto o coeficiente de Tanimoto calcula a similaridade entre duas moléculas através da taxa de bits comuns (configurados como 1) entre suas respectivas fingerprints e o número total de bits configurados como 1 em cada molécula como foi observado na figura(2), por sua vez o coeficiente de Tversky generaliza a equação (8) através da introdução dos parametros α, β:

$$S_{\alpha,\beta}(A,B) = \frac{N_{AB}}{\alpha * N_A + \beta * N_B + (1-\alpha-\beta) * N_{AB}} \tag{9}$$

Conforme podemos observar na equação (9), Quando α = β = 1, esta equação se reduz a equação (8). A introdução desses parâmetros garante ao coeficiente de Tversky uma maior flexibilidade na comparação de similaridade entre duas moléculas uma vez que, os valores de α e β permite a aplicação determinadas ponderações na comparação de fingerprints, atribuindo um maior peso a determinadas características presentes na molécula A em detrimento da molécula B, ou vice-versa. Tal característica é importante para buscas de superestruturas e subestruturas de uma molécula, uma vez que um maior peso atribuido a α implica no direcionamento do cálculo de similaridade para superestruturas da molécula A (SWAMIDASS; BALDI, 2007). Por outro lado, um maior valor de β, direciona o cálculo de similaridade para obteção de subestruturas de A, considerando que A é a molécula de entrada no sistema para a qual serão triadas no banco de moléculas as estruturas similares.

2.3.3 LINGO

O método proposto por (VIDAL; THORMANN; PONS, 2005) denominado LINGO realiza a computação da similaridade entre moléculas através da conversão dos SMILES de cada molécula (descritor molecular) em um conjunto de substrings sobrepostas de tamanho definido (o autor sugere tamanho 4 para melhores resultados). Desta forma evita-se a manipulação de estruturas tridimensionais, e de

grafos, que são mais complexos para se tratar computacionalmente.

Para gerar cada subestrutura do conjunto de substrings a ser utilizado, é necessário a realização de uma canonização do SMILE, onde converte-se o SMILE de cada molécula para uma forma padronizada, conforme já exposto anteriormente. Uma vez canonizadas, as SMILES de cada molécula são quebradas em n-(q-1) subestruturas, onde n é o número de caracteres do SMILE, e q é o tamanho definido para cada subestrutura. Os conjuntos de subestruturas quebradas a partir dos SMILES das moléculas a serem comparadas, são agrupados em vetores para aplicação da métrica de similaridade (VIDAL; THORMANN; PONS, 2005).

Para computar a similaridade entre os dois vetores (que passaram a representar as moléculas), é utilizado o coeficiente de Tanimoto, mas em um formato modificado como podemos verificar na equação (10)

$$T_C = \frac{\sum_{i=1}^{l} 1 - \frac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l} \tag{10}$$

Observando a equação (10) verifica-se que $N_{A,i}$ corresponde ao número de ocorrências da i-ésima subestrutura na molécula A (vetor de subestruturas de A), e $N_{B,i}$ o número de ocorrências da i-ésima subestrutura na molécula B (vetor de subestruturas de B), e l indica o comprimento de um vetor união ($A \bigcup B$). Dessa forma devido ao somatório presente no numerador, quanto mais idênticas as frequência de aparições de cada subestrutura em ambas moléculas, mais alto será o coeficiente de Tanimoto entre as mesmas, indicando assim uma maior similaridade.

2.3.4 Localized Co-occurrence Model (LCM)

Este algoritmo é utilizado para computação de similaridade sobre moléculas representadas por coordenadas tridimensionais. Segundo (HUANG; SHEN; ZHOU, 2008) cada átomo é representado por uma tripla (x,y,z), que informa suas respectivas coordenadas em um espaço tridimensional, e assim para cada átomo, são buscados seus três outros átomos mais próximos, formando um tetraedro denominado Unit Structure, este procedimento é realizado para todos os átomos de uma molécula.

Para cada Unit Structure é calculado o seu volume, e assim então para cada molécula será obtido um conjunto de Unit Structures que terá associado as mesmas um valor de volume, e então o algoritmo utilizará o menor volume obtido, e o maior volume encontrado para definir os limites de um intervalo de volumes para esta molécula. Além disso são definidos um conjunto de estados que será associado ao

intervalo de volumes, dessa forma o intervalo de volumes será particionado de acordo com a quantidade de estados definidos para a molécula. Cada Unit Structure recebe um valor de estado correspondente à partição de volume que a mesma pertence.

Após definido o estado de cada Unit Structure, para cada par de UnitsStructures o algoritmo faz uma conexão entre elas caso o centro de massa entre as mesmas seja menor que um valor definido, este procedimento é realizado para todos os possíveis pares que podem ser formados dentro de um conjunto de Unit Structures. Uma vez formadas todas as conexões possíveis, o algoritmo verifica a co-ocorrência das relações das Unit Structuresem relação aos seus valores de estado (HUANG; SHEN; ZHOU, 2008).

Por fim este algoritmo computa a probabilidade de co-ocorrência para cada combinação de pares de estados, gerando para cada molécula um LCM. A similaridade entre duas moléculas é determinada pela comparação dos LCM de cada uma delas, e verificando o grau de semelhança entre as representações das moléculas em forma de LCM, quanto mais parecidos os LCMs maior a similaridade entre as moléculas, em contrapartida quanto menor a semelhança entre os mesmos, menor a similaridade entre as moléculas.

Neste capítulo foram apresentadas algumas das ferramentas utilizadas por pesquisadores para comparação de moléculas com determinados graus de similaridade. Foi apresentado também algumas das principais formas de representação de uma molécula computacionalmente (de agora em diante denominada descritores moleculares), e sua influência na escolha do algoritmo para computação da similaridade molecular. Além disso foi discutido alguns dos algoritmos e métricas possíveis de utilização para realizar tal atividade,contando com uma pequena descrição prévia do selecionado para utilização neste trabalho.No próximo capítulo será discutida as ferramentas utilizadas na adaptação do algoritmo selecionado, bem como a metodologia aplicada para avaliar algoritmo adaptado ao sistema NatProDB.

3 METODOLOGIA

Neste capítulo serão apresentados os métodos e ferramentas aplicados no desenvolvimento e implementação do algoritmo que foi adaptado no NatProDB, e a metodologia dos testes realizados sobre o algoritmo adaptado no sistema.

3.1 FERRAMENTAS UTILIZADAS PARA IMPLEMENTAÇÃO

Tanto o módulo de gerenciamento de banco de dados do sistema NatProDB, assim como o método aplicado para realização de triagem de moléculas baseado no coeficiente de similaridade molecular foram desenvolvidos baseados no paradigma de Programação Orientada à Objetos (POO), e utilizando como padrão de projeto o *Model View Controller*, visando assim desacoplar o código do sistema, e também facilitar a manutenção e inclusão de novas funcionalidades no mesmo (PATTERNS, 2003). A implementação do sistema, assim como o desenvolvimento do algoritmo de cálculo de similaridade molecular fora realizados utilizando como linguagem de programação Java devido ao fato dessa linguagem de programação ser multi-plataforma (capaz de ser executada em vários sistemas operacionais), gratuita ,e também por questões de compatibilidade com a biblioteca indigo toolkit e com o próprio módulo de gerenciamento de banco de dados do NatProDB. O ambiente selecionado para o desenvolvimento foi o NetBeans 8.1.

3.2 ALGORITMO E MÉTRICA PARA CÁLCULO DE SIMILARIDADE NO NATPRODB

Conforme já discutido na seção 2.2, a escolha do descritor molecular a ser utilizado é um ponto chave para decisão sobre que algoritmo/métrica melhor se adequa para comparação de similaridade entre moléculas (TODESCHINI *et al.*, 2008). Nesta perspectiva, o sistema NatProDB utiliza dois descritores moleculares para realizar cálculo de similaridade molecular: SMILES e Fingerprint. Os SMILES são utilizados para armazenar no banco de dados do sistema a estrutura física das moléculas devido a sua simplicidade na representação da molécula em formato 2D e também baixo custo computacional para armazenamento das características estruturais dos compostos. Os SMILES do ligante (estrutura de interesse do pesquisador) e das moléculas presentes no banco de dados do NatProDB são utilizadas na fase de pré-processamento das estruturas, na qual as mesmas são preparadas para posterior comparação de similaridade. Nesta fase, devido ao fato de existirem diversos métodos para geração de SMILES para uma molécula

(KUMAR, 2012), visando evitar possíveis distorções nos resultados da comparação de similaridade entre as estruturas, tanto o SMILES do ligante quanto os SMILES das moléculas presentes no banco passam por um processamento de canonização, que basicamente re-organiza o SMILES de forma a garantir que todos os SMILES do sistema sejam obtidos seguindo o mesmo padrão de geração (KUMAR, 2012), obtendo assim estruturas únicas para cada uma das moléculas do sistema. Tal canonização é realizada no sistema pela biblioteca Indigo toolkit, que implementa algoritmo para obtenção de SMILES únicos conforme descrito por Pavlov (2011).

Uma vez obtidos os SMILES canônicos das moléculas, o sistema NatProDB realiza uma conversão desses SMILES para outro descritor molecular, no caso Fingerprints. Apesar da existência de metodologias de comparação de similaridade baseadas nos SMILES, como por exemplo o LINGO (VIDAL; THORMANN; PONS, 2005), as fingerprints se mostram mais populares na literatura para realização de comparação de similaridade entre moléculas (VARNEK, 2011). Primeiramente, devido a forma como a fingerprint codifica a estrutura molecular e as características intrínsecas de cada composto (sequencia binária) favorece a aplicação de métricas de similaridade já difundidas na comunidade científica como: Coeficiente de Tanimoto, e Coeficiente de Tversky (WILLETT, 2003). Além disso, a utilização desse descritor molecular, permite que a comparação de similaridade entre moléculas seja realizada através da contagem e comparação das estruturas/grupos funcionais presentes em suas estruturas (bits da fingerprint definidos como valor 1). Dessa maneira, para realizar a conversão dos SMILES das moléculas para Fingerprint, o sistema NatProDB utiliza a biblioteca Indigo toolkit, que implementa uma rotina para realizar tal conversão, finalizando assim o pré-processamento das estruturas para aplicação de métrica de similaridade adotada para o sistema.

A partir das fingerprints obtidas no pré-processamento das estruturas, o algorítmo implementado no sistema NatProDB segue o método descrito nas seção 2.3.1.1 para comparação de similaridade entre moléculas. Basicamente, é realizada a contagem de características em comum e distintas entre as fingerprints de entrada (ligante) e as fingerprints das moléculas do banco, e então aplicada uma métrica para obtenção do coeficiente de similaridade. Diferentemente do método descrito na seção 2.3.1.1, ao invés de utilizar o coeficiente de Tanimoto descrito na equação (8) como métrica de similaridade para realização da triagem em seu banco de dados, o NatProDB utiliza o coeficiente de Tversky. Tal escolha é orientada pelo fato deste coeficiente de permitir o favorecimento das características presentes na

fingerprint de uma moléculas em detrimento da outra através da introdução dos parâmetros α e β na equação (9). Através do aumento e diminuição dos parâmetros α e β, o sistema NatProDB consegue flexibilizar seu método para comparação de similaridade, permitindo assim a comparação não somente de similaridade molecular, mas também a busca de super-estruturas e/ou subestruturas de acordo com os valores selecionados para esses coeficientes. Assim como na etapa de pré-processamento das moléculas, a computação da similaridade entre estruturas através do coeficiente de Tversky é realizado também pela indigo toolkit, que possui uma rotina chamada *similarity*(), que recebe como parâmetro duas fingerprints, e o tipo de coeficiente a ser aplicado (Coeficiente de Tversky para o caso do NatProDB), e retorna para o sistema o grau de similaridade entre as duas fingerprints. Este método é aplicado serialmente entre a molécula de entrada (ligante) e cada uma das moléculas cadastrada no banco de dados do sistema, retornando para o pesquisador somente as moléculas cujo coeficiente de Tversky encontrado é superior ao grau de similaridade selecionado previamente pelo pesquisador.

3.3 EXPERIMENTOS

Para realização dos testes no sistema, foram selecionados 7 moléculas-alvo com estruturas moleculares diversificadas entre si, sendo elas: Acetaminophen, Fluoxetina, Glyburide, Imatinib, Isosorbid, Vinblastine, Propranolol. Todas moléculas citadas já são distribuídas para o mercado como fármacos, ou seja substâncias que contem estruturas que possibilitam sua absorção e metabolização pelo corpo humano. Visando criar um banco controlado para os testes, foram baixadas 47458 moléculas de um banco molecular colaborativo disponibilizado pelo ZINC (IRWIN; SHOICHET, 2005) para o banco local do NatProDB. Para garantir que não existia moléculas com grau de similaridade superior a 80% às moléculas-alvo, foram realizadas consultas nesse banco colaborativo através do ZINC buscando estruturas com grau de similaridade superior a 80% para cada uma das moléculas selecionadas para teste. A escolha do coeficiente de similaridade de 80% como limiar foi baseada no trabalho de Drogra(2007) que afirma que estruturas com coeficiente de similaridade superior ou igual a 0,85 podem ser consideradas similares. Para cada uma dessas 7 consultas o ZINC não retornou nenhuma molécula 80% similar ou com grau de similaridade superior às moléculas alvo. Após populado o banco de dados do NatProDB, foram enxertados nesse banco local: 50 moléculas com grau de similaridade $\geq$80% ao acetaminophen, 50 moléculas com grau de similaridade $\geq$80% à fluoxetina, 50

moléculas com grau de similaridade $\geq$80% ao Glyburide, 50 moléculas com grau de similaridade $\geq$80% ao Imatinib, 34 moléculas com grau de similaridade $\geq$80% ao Isosorbid, 50 moléculas com grau de similaridade $\geq$80% ao Vinblastine, 50 moléculas com grau de similaridade $\geq$80% ao Propranolol. Todas essas estruturas foram extraídas do PUBCHEM (LI *et al.*, 2010), que além de implementar algoritmo para recuperação de moléculas baseado no conceito de similaridade, permite o download dessas estruturas em formato .smi, que é um formato já aceito pelo NatProDB para importação de moléculas.

Após a população do banco de dados do NatProDB foram repetidas no sistema as mesmas 7 consultas realizadas no ZINC, visando assim validar o método implementado para computação de similaridade molecular verificando se o algoritmo aplicado consegue recuperar para cada uma das moléculas-alvo, as moléculas 80% similares as mesmas dentro do conjunto de moléculas do banco. Os resultados obtidos em cada uma das 7 consultas realizadas no NatProDB foram avaliados através de uma adaptação da matriz de confusão, que apesar de ser mais frequentemente utilizada para medição da qualidade de classificadores (principalmente na inteligência artificial) através da comparação do número de classificações corretas(obtidas por testes no classificador) *versus* classificações preditas de uma determinada classe (obtidas por análise de modelo de referência) (DAVIS; GOADRICH, 2006). Para o contexto dos testes realizados no NatProDB, esse conceito foi adaptado para medir a qualidade do método implementado em classificar as estruturas 80% similares a cada uma das moléculas-alvos utilizadas como entradas para o sistema. Dessa maneira as moléculas preditas $\geq$ 80% similares e $\leq$ 80% similares correspondem respectivamente: moléculas-alvo extraídas do PUBCHEM, e moléculas presentes no banco colaborativo disponibilizado pelo ZINC.

Tabela 1: Exemplo de matriz de confusão para avaliação de classificadores

Classe	**Predita** C_+	**Predita** C_-
Verdadeira C_+	Verdadeiro Positivo (V_P)	Falso Positivo (F_P)
Verdadeira C_-	Falso Negativo (F_N)	Verdadeiro Negativo (V_N)

Fonte: Próprio Autor.

A partir da tabela (1), podemos a partir dos valores de V_P, V_N, F_P, e F_N, derivar algumas métricas comumente utilizadas para medir a qualidade do classificador implementado como: Acurácia (12) que indica a proporção de predições corretas (Verdadeiras e negativas); Sensibilidade(13) que é um indicador da capacidade do

sistema em identificar verdadeiros positivos quando as amostras verdadeiramente atendem a uma dada condição; Especificidade(14) que indica a capacidade do sistema em identificar verdadeiros negativos quando as amostras verdadeiramente não atendem a uma dada condição; Precisão (15) que indica a porcentagem de amostras classificadas corretamente como positivas dentre todas classificadas como positivas; F-Medida (16) que é uma média ponderada entre a sensibilidade e a precisão; Porcentagem de amostras erroneamente classificadas como positivas dentre todas verdadeiramente negativas (FPR) (17) (KOHAVI; PROVOST, 1998).

$$Acurácia = \frac{\textit{Total de Acertos}}{\textit{Total de Dados no Conjunto}} = \frac{V_P + V_N}{V_N + V_P + F_P + F_V} \tag{12}$$

$$Sensibilidade = \frac{\textit{Acertos Positivos}}{\textit{Total de Positivos}} = \frac{V_P}{V_P + F_N} \tag{13}$$

$$Especificidade = \frac{\textit{Acertos Negativos}}{\textit{Total Negativos}} = \frac{V_N}{V_N + F_P} \tag{14}$$

$$Precisão = \frac{\textit{Acertos Positivos}}{\textit{Total Classificado como Positivos}} = \frac{V_P}{V_P + F_P} \tag{15}$$

$$F - Medida = \frac{\textit{2* Precisão*Sensibilidade}}{\textit{Precisão + Sensibilidade}} \tag{16}$$

$$FPR = \frac{\textit{Erros Positivos}}{\textit{Erros Positivos + Acertos Negativos}} = \frac{F_P}{F_P + V_N} \tag{17}$$

4 RESULTADOS E DISCUSSÕES

De acordo com Dogra (2007), a similaridade entre duas moléculas pode ser verificada quando o coeficiente de similaridade resultante da comparação de duas estruturas é superior a 80%. Desta maneira para o primeiro teste realizado no sistema, tendo como alvo o acetaminophen e utilizando o limiar de similaridade citado anteriormente, das 50 moléculas$\geq$80% similares (extraídas do PUBCHEM) inseridas no banco do NatProDB, o algoritmo implementado no sistema recuperou corretamente 45 delas com coeficiente de similaridade superior a 80% (Verdadeiro Positivo V_P), 5 delas foram classificadas pelo NatProDB com similaridade inferior a 80% (Falso Negativo F_P), como podemos verificar na tabela 2

Tabela 2: Classificação de moléculas similares ao Acetaminophen

	Similaridade $\geq$ 80%	**Similaridade $\leq$ 80%**
Moléculas classificadas c/ similaridade $\geq$ 80%	45	0
Moléculas classificadas c/ similaridade $\leq$ 80%	5	47742

Fonte: Próprio Autor.

A partir da matriz de confusão representada na tabela 2 podemos observar para esse teste uma Acurácia = 99,00% equação(12), Sensibilidade = 90,00% equação13, Medida F-Score = 94,70% equação(16), Precisão = 100% equação(15), Especificidade = 100% equação(14)

Para o segundo teste realizado, tendo como alvo a fluoxetina, e utilizando o mesmo limiar de similaridade aplicado anteriormente, das 50 moléculas $\geq$80% similares a fluoxetina enxertada no NatProDB (extraídas do PUBCHEM), o algoritmo implementado no NatProDB recuperou corretamente 48 dessas estruturas (V_P), e apenas duas delas foram classificadas com coeficiente de similaridade inferior a 80%(F_N), e nenhum F_P foi verificado conforme podemos observar na tabela 3.

A partir da tabela 3 também podemos observar que para esse teste foram obtidas: Acurácia = 99,00% equação(12), Sensibilidade = 96,00% equação(13), Medida F-Score = 97,90% equação(16), Precisão = 100% equação(15), Especificidade = 100% equação(14).

O terceiro teste realizado teve como alvo Glyburid. Sob o mesmo limiar de similaridade dos testes anteriores, das 50 moléculas $\geq$80% similares ao Glyburid, inseridas no banco de dados do NatProDB (extraídas do PUBCHEM), o algoritmo

Tabela 3: Resultado Obtido da Busca de Estruturas $\geq$80% Similares a Fluoxetina

	Similaridade $\geq$ 80%	**Similaridade $\leq$ 80%**
Moléculas classificadas c/ similaridade $\geq$ 80%	48	0
Moléculas classificadas c/ similaridade $\leq$ 80%	2	47742

Fonte: Próprio Autor.

implementado recuperou corretamente todas as 50 estruturas (V_P), e por tanto nenhuma delas foram classificadas com coeficiente de similaridade inferior a 80%(F_N), o que gerou uma Acurácia = 100% equação(12), Sensibilidade = 100% equação(13), Medida F-Score = 100% equação(16), Precisão = 100% equação(15), Especificidade = 100% equação(14). conforme podemos observar na tabela 4.

Tabela 4: Resultado Obtido da Busca de Estruturas $\geq$ 80% Similares ao Glyburid

	Similaridade $\geq$ 80%	**Similaridade $\leq$ 0.8**
Moléculas classificadas c/ similaridade $\geq$ 0.8	50	0
Moléculas classificadas c/ similaridade $\leq$ 0.8	0	47742

Fonte: Próprio Autor.

O quarto teste realizado teve como alvo a molécula Imatinib. A triagem de moléculas no banco do NatProDB cujo coeficiente de similaridade $\geq$ 80% com relação a molécula-alvo, retornou 49 das 50 estruturas $\geq$80% similares ao imatinib enxertadas no banco. Dessa maneira foram obtidos 49 V_P, apenas 1 F_P, e novamente nenhum F_N foi verificado. A partir desse resultado foram obtidas: Acurácia = 99,00 equação(12), Sensibilidade = 98,00% equação(13), Medida F-Score = 98,90% (16), Precisão = 100,00% equação(15), Especificidade = 100,00% (14). conforme podemos observar na tabela 5

O quinto teste realizado no sistema teve como alvo o Isosorbid. A consulta realizada no banco do NatProDB por estruturas 80% similares à molécula-alvo, retornou todas as 34 moléculas $\geq$ 80% similares as que foram enxertadas no banco do sistema, não gerando assim nenhum falso-positivo ou falso-negativo. Consequentemente foram obtidas a partir dos resultados apresentados na tabela 6: Acurácia = 100% equação(12), Sensibilidade = 100,00% equação(13), Medida F-Score = 100,00% (16), Precisão = 100,00% equação(15), Especificidade = 100,00%

Tabela 5: Resultado Obtido da Busca de Estruturas ≥80% Similares ao Imatinib

	Similaridade ≥80%	**Similaridade ≤80%**
Moléculas classificadas c/ similaridade ≥ 80%	50	0
Moléculas classificadas c/ similaridade ≤ 80%	0	47742

Fonte: Próprio Autor.

equação(14).

Tabela 6: Resultado Obtido da Busca de Estruturas ≥80% Similares ao Isosorbid

	Similaridade ≥80%	**Similaridade ≤80%**
Moléculas classificadas c/ similaridade ≥80%	34	0
Moléculas classificadas c/ similaridade ≤80%	0	47758

Fonte: Próprio Autor.

Para o sexto teste foi selecionada como molécula-alvo o Vinblastine. A triagem realizada no banco de dados do NatProDB por estruturas ≥ 80% similares a molécula-alvo, recuperou todas as 50 moléculas ≥80% similares ao Vinblastine enxertadas no banco de dados do sistema. Novamente, assim como no teste anterior, não foi verificada a ocorrência de falsos-positivos e/ou falsos-negativos, resultando em medidas de: Acurácia = 100,00% equação(12), Sensibilidade = 100,00% equação(13), Medida F-Score = 100,00% equação(16), Precisão = 100,00% equação(15), Especificidade = 100,00% equação(14). Conforme podemos verificar pela tabela 7

Tabela 7: Resultado Obtido da Busca de Estruturas ≥80% Similares ao Vinblastine

	Similaridade ≥ 80%	**Similaridade ≤ 80%**
Moléculas classificadas c/ similaridade ≥ 80%	50	0
Moléculas classificadas c/ similaridade ≤ 80%	0	47742

Fonte: Próprio Autor.

O sétimo teste realizado teve como molécula-alvo o Propanolol. A consulta realizada no NatProDB buscando as estruturas $\geq$ 80% similares ao Propanolol recuperou 42 das 50 moléculas enxertadas no banco do sistema (Verdeiros positivos V_P), e as 8 moléculas restantes foram classificadas com coeficiente de similaridade inferior a 0.8 (Falso negativos F_P). A partir desse dados foram obtidas a partir da tabela (8): Acurácia = 99,00% equação(12), Sensibilidade = 84,00% equação(13), Medida F-Score = 91,00% equação(16), Precisão = 100,00% equação(15), Especificidade = 100,00% equação(14).

Os resultados desse teste podem ser verificado na matriz de confusão (8)

Tabela 8: Resultado Obtido da Busca de Estruturas $\geq$80% Similares ao Propanolol

	Similaridade $\geq$ 80%	**Similaridade $\leq$ 80%**
Moléculas classificadas c/ similaridade $\geq$ 80%	42	0
Moléculas classificadas c/ similaridade $\leq$ 80%	8	47742

Fonte: Próprio Autor.

Tabela 9: Resumo dos Resultados Obtidos nos Testes Realizados

Teste	**Acurácia**	**Sensibilidade %**	**F-Score %**	**Precisão %**	**Especificidade %**	**FPR %**
1º	99,00	90,00	94,70	100,00	100,00	0,00
2º	99,00	96,00	97,90	100,00	100,00	0,00
3º	100,00	100,00	100,00	100,00	100,00	0,00
4º	99,00	98,00	98,90	100,00	100,00	0,00
5º	100,00	100,00	100,00	100,00	100,00	0,00
6º	100,00	100,00	100,00	100,00	100,00	0,00
7º	99,00	84,00	91,00	100,00	100,00	0,00

Fonte: Próprio Autor.

De acordo com Davis & Goadrich (2006), embora as medidas de acurácia e F-medida forneçam uma medida global da qualidade da classificação, na qual quão mais próximos de 100,00% os valores obtidos para as mesmas melhor a classificação, o autor relata que uma classificação ideal ocorre quando a Sensibilidade (13) $\simeq 100,00\%$, e FPR (17) $\simeq 0\%$. Sob essa perspectiva, tomando como base o pior caso dos testes realizados no sistema, que teve como alvo a molécula de propranolol - que reportou F_N = 8, podemos facilmente identificar a partir tabela de resumo dos testes (9) que foram obtidas uma Sensibilidade = 84%, e FPR = 0%, o que indica que o método aplicado para separar as moléculas $\geq$ 80% similares ao

propranolol obteve uma resposta satisfatória, uma vez que ambas sensibilidade e FPR se aproximaram respectivamente de 100% e 0% (classificador ideal). Estendendo tal verificação para todos os testes realizados, foi observado que não ocorreram F_P nos testes realizados, ou em outras palavras, o método implementado para triagem virtual de moléculas baseado no conceito 80% de similaridade não retornou nenhuma molécula com coeficiente de similaridade inferior a para cada uma das moléculas-alvo testadas. Consequentemente, as sensibilidades verificadas em todos os testes realizados variam dentro de um intervalo de [84,00% - 99,00%], o que de acordo com Davis & Goadrich (2006) pode ser considerado como um comportamento próximo de um classificador ideal.

5 CONSIDERAÇÕES FINAIS

O presente trabalho é fruto da parceria entre o Laboratório de Computação de Alto Desempenho(LACAD) e o Laboratório de Modelagem Molecular (LMM) da UEFS, no qual foi proposto o estudo e implementação de um algoritmo para o sistema NatProDB visando possibilitar ao mesmo a realização de uma triagem virtual de moléculas (Virtual Screening) no seu banco baseada no conceito de similaridade molecular. O sistema juntamente com o algoritmo adaptado neste trabalho permite aos pesquisadores construírem localmente suas próprias bases de dados de moléculas, e selecionar dentro do seu banco moléculas similares aos ligantes de seu interesse para realização de estudos.

A aplicação de uma abordagem de comparação de moléculas 2D proposta neste trabalho simplifica a computação de similaridade molecular, uma vez que permite a utilização de descritores moleculares lineares como as fingerprints, e também a aplicação de técnicas de comparação como o Coeficiente de Tversky ou até mesmo o Coeficiente de Tanimoto, que calculam a similaridade entre estruturas através da contagem de características/grupos funcionais presentes em cada molécula. O método implementado no NatProDB é uma abordagem bastante difundida na literatura, e já vem sendo utilizada inclusive por sistemas consolidados na comunidade científica como o PubChem por exemplo. Além disso, devido a escassez de ferramentas gratuitas que se proponham a fornecer um ambiente computacional onde o pesquisador possa controlar e gerenciar as suas moléculas, assim como realizar triagem em bancos locais através do conceito de similaridade molecular, o sistema desenvolvido surge como uma ferramenta para auxiliar esses pesquisadores desenvolverem seus estudos e até mesmo criarem base de dados com estruturas de sua propriedade para posterior publicação.

Embora o sistema NatProDB já venha sendo utilizado por pesquisadores do LMM, esse projeto evoluiu de forma a abarcar não somente o estudo de estruturas moleculares localmente, mas também permitir a difusão das moléculas oriundas do semi-árido para a comunidade científica em geral através da internet. Nesse sentido, recentemente foi criado um ambiente web para o NatProDB, que fornece a possibilidade de acesso a um catálogo de moléculas extraídas de fontes naturais endêmicas oriundas do bioma semi-árido através da internet.

Devido ao fato do algoritmo implementado no NatProDB realizar uma comparação baseada em contagem de características/grupos funcionais e avaliação estrutural das moléculas, questões como rotações, propriedades intrínsecas de

alguns compostos, e até mesmo átomos quirais, não são tratadas pelo algoritmo implementado no sistema, o que pode ocasionar obtenção de índices de similaridade altos entre duas estruturas que não necessariamente possuem atividades biológicas semelhantes. Dessa maneira, visando solucionar tal problema, será implementado posteriormente no sistema tanto local quanto web, algoritmos que realize a computação de similaridade molecular através de descritores moleculares 3D, visando assim permitir que o sistema não somente compare moléculas baseados nas suas características estruturais, mas que também leve em consideração as propriedades intrínsecas e particularidades dos compostos.

REFERÊNCIAS

ANDERSON, E.; VEITH, G. D.; WEININGER, D. **SMILES, a line notation and computerized interpreter for chemical structures**. [S.l.]: US Environmental Protection Agency, Environmental Research Laboratory, 1987.

BAJORATH, J. **Chemoinformatics: concepts, methods, and tools for drug discovery**. [S.l.]: Springer Science & Business Media, 2004.

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: ACM. **Proceedings of the 23rd international conference on Machine learning**. [S.l.], 2006. p. 233–240.

DEVILLERS, J. **Neural networks in QSAR and drug design**. [S.l.]: Academic Press, 1996.

DOGRA, S. K. **Script for computing Tanimoto coefficient**. jan 2007. Disponível em: <http://www.qsarworld.com/virtual-workshop.php>.

HUANG, Z.; SHEN, H. T.; ZHOU, X. Localized co-occurrence model for fast approximate search in 3d structure databases. **Knowledge and Data Engineering, IEEE Transactions on**, IEEE, v. 20, n. 4, p. 519–531, 2008.

IRWIN, J. J.; SHOICHET, B. K. Zinc-a free database of commercially available compounds for virtual screening. **Journal of chemical information and modeling**, ACS Publications, v. 45, n. 1, p. 177–182, 2005.

KOES, D. R.; CAMACHO, C. J. Pharmer: efficient and exact pharmacophore search. **Journal of chemical information and modeling**, ACS Publications, v. 51, n. 6, p. 1307–1314, 2011.

KOHAVI, R.; PROVOST, F. Glossary of terms. **Machine Learning**, v. 30, n. 2-3, p. 271–274, 1998.

KUBINYI, H. *et al.* **Virtual screening for bioactive molecules**. [S.l.]: John Wiley & Sons, 2008.

KUMAR, A. Applications of chemoinformatics in drug discovery : Substructure/structure search. **Asian Journal of Biochemical and Pharmaceutical Research**, AJBPAD, v. 2, n. 4, p. 135–143, 2012.

LI, Q. *et al.* Pubchem as a public resource for drug discovery. **Drug discovery today**, Elsevier, v. 15, n. 23, p. 1052–1057, 2010.

MA, C.; WANG, L.; XIE, X.-Q. Gpu accelerated chemical similarity calculation for compound library comparison. **Journal of chemical information and modeling**, ACS Publications, v. 51, n. 7, p. 1521–1527, 2011.

MESTRES, J.; ROHRER, D. C.; MAGGIORA, G. M. Mimic: A molecular-field matching program. exploiting applicability of molecular similarity approaches. **Journal of Computational Chemistry**, Wiley Online Library, v. 18, n. 7, p. 934–954, 1997.

PATTERNS, D. **Model-View-Controller**. [S.l.]: Microsoft Patterns & Practices, http://msdn. microsoft. com/practices/type/Patterns/Enterprise/DesMVC, 2003.

PAVLOV, D. *et al.* Indigo: universal cheminformatics api. **J. Cheminformatics**, v. 3, n. S-1, p. 4, 2011.

RODRIGUES, R. P. *et al.* Estratégias de triagem virtual no planejamento de fármacos. **Revista Virtual de Química**, v. 4, n. 6, p. 739–776, 2012.

SINGH, R. Reasoning about molecular similarity and properties. In: IEEE. **Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE**. [S.l.], 2004. p. 266–277.

SWAMIDASS, S. J.; BALDI, P. Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. **Journal of chemical information and modeling**, ACS Publications, v. 47, n. 2, p. 302–317, 2007.

TODESCHINI, R. *et al.* **Handbook of Molecular Descriptors**. Wiley, 2008. (Methods and Principles in Medicinal Chemistry). ISBN 9783527613113. Disponível em: <https://books.google.com.br/books?id=PCXXdjUJiLoC>. Acesso em: 24 Nov. 2016.

VARNEK, A. Fragment descriptors in structure–property modeling and virtual screening. **Chemoinformatics and Computational Chemical Biology**, Springer, p. 213–243, 2011.

VIDAL, D.; THORMANN, M.; PONS, M. Lingo, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. **Journal of chemical information and modeling**, ACS Publications, v. 45, n. 2, p. 386–393, 2005.

WILLETT, P. Similarity-based approaches to virtual screening. **Biochemical Society Transactions**, Portland Press Limited, v. 31, n. 3, p. 603–606, 2003.

XUE, L.; BAJORATH, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. **Combinatorial Chemistry & High Throughput Screening**, Bentham Science Publishers, v. 3, n. 5, p. 363–372, 2000.