



**UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA**  
**BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO**

**ANDERSON SOUZA ROCHA**

**IMPLEMENTAÇÃO DE ALGORITMOS PARA CÁLCULO DE SIMILARIDADE NO  
SISTEMA NATPRODB**

**FEIRA DE SANTANA**  
**2016**

**ANDERSON SOUZA ROCHA**

**IMPLEMENTAÇÃO DE ALGORITMOS PARA CÁLCULO DE SIMILARIDADE 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

Texto da dedicatória.Texto da  
dedicatória.Texto da dedicatória.Texto  
da dedicatória.Texto da dedicatória.Texto  
da dedicatória.Texto da dedicatória.Texto  
da dedicatória.Texto da dedicatória.

## **AGRADECIMENTOS**

Texto dos agradecimentos.

## 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, essas ferramentas também auxiliam na busca por estruturas similares, que é um conceito em que se baseia grande parte dos esforços dirigidos ao projeto de um novo fármaco. O presente trabalho descreve a adaptação de algoritmos apropriados 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 às 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 comportará da seguinte forma: Uma vez cadastradas as moléculas 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 a molécula de entrada, e as presentes no banco em fingerprints para aplicação da métrica de Tanimoto para obtenção de um índice 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).

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

## ABSTRACT

The developing process of a medicine is made of different stages, that come from the project and study of a possible pharmacophore compound until its synthesizing. Most efforts on the development of a new medicine are made in assisted way through 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 characteristics, and the molecules behavior during structural variations. Besides this, the tools also help in the search for similar structures, that is a concept in most efforts in a new medicine project are based. This work describes the adaptation of appropriate algorithms for computing molecular similarities in a software developed for the tabulation of molecules in a database, aiming to allow the researcher to browse, in his own database, similar structures to the ones that are in his study concerns. A differential in this work is that this tool will be made freely available to any user, and due the inner characteristics of its implementation, face to other tools that exist with the same purpose, this software will not need much hardware resources for its execution, in other words, it is compatible with common computers, and also with any operational system used by the researcher. Briefly, the system will work in such manner: once the molecules are registered in the user database, the researcher will be able to verify in his base if there are structures with a desired degree of similarity to the molecule inserted for consultation. The system converts the inserted molecule, and the presents in the database, in fingerprints for the application of Tanimoto's metric, to obtain a table of similarity, returning finally to the researcher all the molecules in his database that have a degree of similarity equal or superior to the one desired (relating to the entrance molecule).

**Keywords:** Molecular Similarity. Molecular Databases. Computational Chemistry.

## LISTA DE FIGURAS

Figura 1	Exemplo de Molécula representada em SMILE	19
Figura 2	Cálculo do Coeficiente de Tanimoto sobre duas Fingerprints	22

## LISTA DE TABELAS

Tabela 1	Matriz de Confusão	29
Tabela 2	Classificação de moléculas similares ao Acetaminophen	31
Tabela 3	Resultado Obtido da Busca de Estruturas $\geq 0.8$ Similares a Fluoxetina	31



## LISTA DE SÍMBOLOS

## LISTA DE SIGLAS

## SUMÁRIO

<b>1 INTRODUÇÃO.....</b>	<b>10</b>
1.1 OBJETIVOS .....	12
1.1.1 OBJETIVO GERAL .....	12
1.1.2 OBJETIVOS ESPECÍFICOS .....	12
<b>2 FUNDAMENTAÇÃO TEÓRICA .....</b>	<b>13</b>
2.1 FERRAMENTAS PARA COMPUTAÇÃO DE SIMILARIDADE MOLECULAR .	14
2.2 DESCRITORES MOLECULARES .....	17
2.2.1 FINGERPRINTS .....	17
2.2.2 DESCRITORES FARMACOFÓRICOS 3D .....	18
2.2.3 SMILES (SIMPLIFIED MOLECULAR-INPUT LINE-ENTRY SYSTEM) ....	19
2.3 ALGORÍTMOS E MÉTRICAS PARA COMPUTAÇÃO DE SIMILARIDADE MOLECULAR .....	20
2.3.1 COEFICIENTE DE TANIMOTO .....	20
2.3.1.1 APLICAÇÃO DO COEFICIENTE DE TANIMOTO NA COMPARAÇÃO DE FINGERPRINTS .....	21
2.3.2 COEFICIENTE DE TVERSKY .....	22
2.3.3 LINGO .....	23
2.3.4 LOCALIZED CO-OCCURRENCE MODEL (LCM) .....	24
<b>3 METODOLOGIA.....</b>	<b>26</b>
3.1 FERRAMENTAS UTILIZADAS PARA IMPLEMENTAÇÃO .....	26
3.2 ALGORITMO E MÉTRICA PARA CÁLCULO DE SIMILARIDADE NO NATPRODB .....	26
3.3 TESTES .....	28
<b>4 RESULTADOS E DISCUSSÕES.....</b>	<b>31</b>
<b>5 CONSIDERAÇÕES FINAIS.....</b>	<b>33</b>
<b>REFERÊNCIAS.....</b>	<b>34</b>

## 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 iniciais 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.

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 grande 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 computação 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 disponíveis no ZINC, e foram enchertadas nesse banco moléculas com grau de similaridade superior a 80% a um conjunto de moléculas 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 as moléculas como 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 computação de similaridade molecular em bancos de dados moleculares.

### 1.1.2 Objetivos Específicos

- Implementar um algoritmo para cálculo de similaridade no sistema NatProDB.
- Implementar rotina para realização de triagem de moléculas com um determinado grau de similaridade no banco de dados do sistema.

## **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 descobertas 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 utilizando 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. 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 farmacofórico (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 nenhuma 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 eletrostáticos 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 catalogo 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 estrutural) 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, O 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 farmacofóricos (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 aceitar 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. Nesta seção será apresentado alguns descritores utilizados por sistemas que calculam similaridade entre moléculas(TODESCHINI *et al.*, 2008 apud KOES; CAMACHO, 2011).

### 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 intrínsecas 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 características, distancias, e/ou propriedades podem ser mapeadas através de algoritmos de *Hashing* para fornecer assim padrões 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 fingerprints para representação computacional de uma molécula reside na simplicidade de aplicação de métodos e métricas para comparação moléculas baseado no conceito de similaridade molecular (XUE; BAJORATH, 2000). Nesta perspectiva, devido as características intrínsecas 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 obtém um determinado coeficiente de similaridade entre duas estruturas moleculares através da realização de um cálculo estatístico que leva em consideração as estruturas e características compartilhadas e as singulares das moléculas comparadas. No sistema NatProDB, as fingerprints possuem um papel fundamental no processo de triagem de moléculas 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 ocorrência dessa ligação. Caso contrário, através de técnicas de triagem virtual (*virtual screening*) pesquisadores podem identificar moléculas que contenham esses grupos funcionais/características, 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 *fingerprint* (Sequência de bits única para uma molécula), onde bits

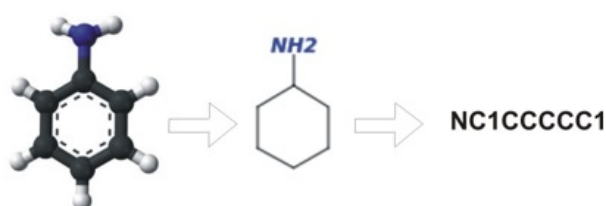
definidos como '1' simbolizam a presença de uma determinada característica ou grupo funcional, e bits definidos como '0' a ausência de tal estrutura na molécula. Sendo assim, ferramentas computacionais podem calcular a similaridade entre duas moléculas gerando fingerprints para cada um dos descritores farmacofóricos 3D e aplicando alguma das métricas já conhecidas para cálculo de similaridade entre fingerprints, 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 (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 ALGORÍTMOS 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 algoritmos 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 molecular, a comparação de similaridade pode ser realizada através de operações booleanas em cadeias binárias para contagem do número de características 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 características presentes em A (bits iguais a 1 em A),  $N_B$  como o número de características presentes em B (bits iguais a 1 em B),  $N_{AB}$  como o número de características 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(11)

$$T_C = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (11)$$

Nota-se em na equação (11), 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).

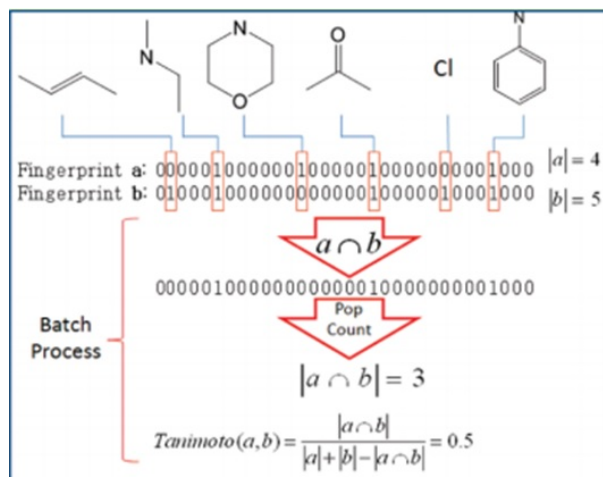
### 2.3.1.1 Aplicação do Coeficiente de Tanimoto na Comparação de Fingerprints

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 (11) 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 ( $a \cap b$ ), 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 metríca 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 (11) através da introdução dos parâmetros  $\alpha, \beta$ :

$$S_{\alpha,\beta}(A,B) = \frac{N_{AB}}{\alpha * N_A + \beta * N_B + (1 - \alpha - \beta) * N_{AB}} \quad (12)$$

Conforme podemos observar na equação (12), Quando  $\alpha = \beta = 1$ , esta equação se reduz a equação (11). 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  $\alpha$  e  $\beta$  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 atribuído a  $\alpha$  implica no direcionamento do cálculo de similaridade para superestruturas da molécula A (SWAMIDASS; BALDI, 2007). Por outro lado, um maior valor de  $\beta$ , direciona o cálculo de similaridade para obtenção 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 (13)

$$T_C = \frac{\sum_{i=1}^l 1 - \frac{|N_{A,i} - N_{B,i}|}{N_{A,i} + N_{B,i}}}{l} \quad (13)$$

Observando a equação (13) 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 ( $A \cup B$ ). 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 Structures em 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 será apresentado os métodos e ferramentas aplicados no desenvolvimento e implementação do algoritmo a ser adaptado no NatProDB. Também será apresentado a metodologia dos testes realizados sobre o algoritmo escolhido para adaptação ao NatProDB.

#### 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 foram 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, onde 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 obtidas 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 das Fingerprints para comparação de similaridade molecular, reduz a computação de similaridade entre duas moléculas a contagem e comparação de bits '1' presentes em cada uma das estruturas. 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 algoritmo implementado no sistema NatProDB segue o método descrito na 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 (11) 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  $\alpha$  e  $\beta$  na equação (12). Através do aumento e diminuição dos parâmetros  $\alpha$  e  $\beta$ , 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 TESTES

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 47770 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 0.8 ou 80% às moléculas-alvo, foram realizadas consultas nesse banco colaborativo através do ZINC buscando estruturas com grau de similaridade superior a 80% ou 0.8 para cada uma das moléculas selecionadas para teste. A escolha do coeficiente de similaridade de 0.8 ou 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 encherados nesse banco local: 50 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  ao acetaminophen, 50 moléculas com grau de

similaridade  $\geq 0.8$  ou  $\geq 80\%$  à fluoxetina, 50 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  ao Glyburide, 50 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  ao Imatinib, 34 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  ao Isosorbid, 50 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  ao Vinblastine, 50 moléculas com grau de similaridade  $\geq 0.8$  ou  $\geq 80\%$  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% ou 0.8 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; GOADRIC, 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% ou 0.8 similares a cada uma das moléculas-alvos utilizadas como entradas para o sistema. Dessa maneira as moléculas preditas  $\geq 0.8$  similares e  $\leq 0.8$  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 (15), Sensibilidade(16), Especificidade(17), Precisão

(18), F-Medida (19)

$$Acurácia = \frac{Total\ de\ Acertos}{Total\ de\ Dados\ no\ Conjunto} = \frac{V_P + V_N}{V_N + V_P + F_P + F_V} \quad (15)$$

$$Sensibilidade = \frac{Acertos\ Positivos}{Total\ de\ Positivos} = \frac{V_P}{V_P + F_N} \quad (16)$$

$$Especificidade = \frac{Acertos\ Negativos}{Total\ Negativos} = \frac{V_N}{V_N + F_P} \quad (17)$$

$$Precisão = \frac{Acertos\ Positivos}{Total\ Classificado\ como\ Positivos} = \frac{V_P}{V_P + F_P} \quad (18)$$

$$F - Medida = \frac{2 * Precisão * Sensibilidade}{Precisão + Sensibilidade} \quad (19)$$



## 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 0.8 ou 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  $\geq 0.8$  ou  $\geq 80\%$  similares (extraídas do PUBCHEM) inseridas no banco do NatProDB, o algoritmo implementado no sistema recuperou corretamente 45 delas com coeficiente de similaridade superior a 0.8 (Verdadeiro Positivo), e 5 delas foram classificadas pelo NatProDB com similaridade inferior a 0.8 (Falso Negativo) como podemos verificar na tabela (2)

**Tabela 2: Classificação de moléculas similares ao Acetaminophen**

	Similaridade $\geq 0.8$	Similaridade $\leq 0.8$
Moléculas classificadas c/ similaridade $\geq 0.8$	45	0
Moléculas classificadas c/ similaridade $\leq 0.8$	5	47742

**Fonte:** Próprio Autor.

A partir da matriz de confusão (2) podemos observar para esse teste uma Acurácia = 0.99 (15), Sensibilidade = 0,9 (16), Medida F-Score = 0.947 (19), Precisão = 1 (18), Especificidade = 1 (17)

Para o segundo teste realizado, tendo como alvo a fluoxetina, e utilizando o mesmo limiar de similaridade aplicado anteriormente, das 50 moléculas  $\geq 0.8$  ou  $\geq 80\%$  similares a fluoxetina enxertada no NatProDB (extraídas do PUBCHEM), o algoritmo implementado no NatProDB recuperou corretamente 48 dessas estruturas (Verdadeiros Positivos), e apenas duas delas foram classificadas com coeficiente de similaridade inferior a 0.8 (Falso Negativo) conforme podemos observar na tabela (3).

**Tabela 3: Resultado Obtido da Busca de Estruturas  $\geq 0.8$  Similares a Fluoxetina**

	Similaridade $\geq 0.8$	Similaridade $\leq 0.8$
Moléculas classificadas c/ similaridade $\geq 0.8$	48	0
Moléculas classificadas c/ similaridade $\leq 0.8$	2	47742

**Fonte:** Próprio Autor.

A partir da tabela (3) também podemos observar que para esse teste foram obtidas: Acurácia = 0.99 (15), Sensibilidade = 0,96 (16), Medida F-Score = 0.979 (19), Precisão = 1 (18), Especificidade = 1 (17).

## 5 CONSIDERAÇÕES FINAIS

Espera-se que o uso do estilo de formatação  $\text{\LaTeX}$  adequado às Normas para Elaboração de Trabalhos de Conclusão de Curso dos estudantes de Engenharia de Computação, da UEFS (`abnt-uefs.cls`) facilite a escrita de documentos no âmbito desta instituição e aumente a produtividade de seus autores. Para usuários iniciantes em  $\text{\LaTeX}$ , além da bibliografia especializada já citada, existe ainda uma série de recursos (CTAN, 2009) e fontes de informação (TEX-BR, 2009; WIKIBOOKS, 2009) disponíveis na Internet.

Recomenda-se o editor de textos Kile como ferramenta de composição de documentos em  $\text{\LaTeX}$  para usuários Linux. Para usuários Windows recomenda-se o editor  $\text{\TeX}$ nicCenter (TEXNICCENTER, 2009). O  $\text{\LaTeX}$  normalmente já faz parte da maioria das distribuições Linux, mas no sistema operacional Windows é necessário instalar o software  $\text{\MiKTeX}$  (MIKTEX, 2009).

Além disso, recomenda-se o uso de um gerenciador de referências como o JabRef (JABREF, 2009) ou Mendeley (MENDELEY, 2009) para a catalogação bibliográfica em um arquivo  $\text{\BibTeX}$ , de forma a facilitar citações através do comando `\cite{}` e outros comandos correlatos do pacote  $\text{\abnTeX}$ . A lista de referências deste documento foi gerada automaticamente pelo software  $\text{\LaTeX}$  +  $\text{\BibTeX}$  a partir do arquivo `abnt-uefs.bib`, que por sua vez foi composto com o gerenciador de referências JabRef.

O estilo de formatação  $\text{\LaTeX}$  do curso de Engenharia de Computação da UEFS foi elaborado por João Carlos Nunes Bittencourt (`joaocarlos@ecomp.uefs.br`), e este exemplo de utilização adaptado de Diogo Rosa Kuiaski (`diogo.kuiaski@gmail.com`) e Hugo Vieira Neto (`hvieir@utfpr.edu.br`). Sugestões de melhorias são bem-vindas.

## REFERÊNCIAS

- ABNTEX. **Absurdas normas para T<sub>E</sub>X**. 2009. Disponível em: <<http://sourceforge.net/apps/mediawiki/abntex/index.php>>. Acesso em: 8 nov. 2011.
- 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.
- CTAN. **The comprehensive T<sub>E</sub>X archive network**. 2009. Disponível em: <<http://www.ctan.org>>. Acesso em: 8 nov. de 2011.
- 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.
- JABREF. **JabRef reference manager**. 2009. Disponível em: <<http://jabref.sourceforge.net>>. Acesso em: 8 nov. 2011.
- 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.
- 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.

MENDELEY. **Mendeley**: academic software for research papers. 2009. Disponível em: <<http://www.mendeley.com>>. Acesso em: 8 nov. de 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.

MIKTEX. **The MiKTeX project**. 2009. Disponível em: <<http://www.miktex.org>>. Acesso em: 8 nov. de 2011.

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.

TEX-BR. **Comunidade T<sub>E</sub>X-Br**. 2009. Disponível em: <<http://www.tex-br.org/index.php>>. Acesso em: 8 nov. 2011.

TEXNICCENTER. **T<sub>E</sub>XnicCenter**: the center of your L<sup>A</sup>T<sub>E</sub>X universe. 2009. Disponível em: <<http://www.texniccenter.org>>. Acesso em: 8 nov. 2011.

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.

WIKIBOOKS. **L<sup>A</sup>T<sub>E</sub>X**. 2009. Disponível em: <<http://en.wikibooks.org/wiki/LaTeX>>. Acesso em: 8 nov. 2011.

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.