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.

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 maisrápida 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 trabalhoé 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 troughthrough 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

LISTA DE TABELAS

LISTA DE SÍMBOLOS

λ	comprimento de onda
v	velocidade
f	frequência

LISTA DE SIGLAS

CCECOMP	Colegiado do Curso de Engenharia de Computação
DAEComp	Diretório Acadêmico de Engenharia de Computação
UEFS	Universidade Estadual de Feira de Santana

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 estrutras 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.

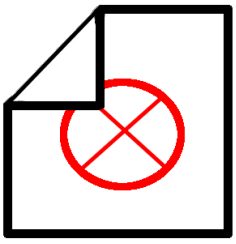
2 DESENVOLVIMENTO

A seguir ilustra-se a forma de incluir figuras, tabelas, equações, siglas e símbolos no documento, obtendo indexação automática em suas respectivas listas. A numeração sequencial de figuras, tabelas e equações ocorre de modo automático. Referências cruzadas são obtidas através dos comandos `\label{}` e `\ref{}`. Por exemplo, não é necessário saber que o número deste capítulo é 2 para colocar o seu número no texto. Isto facilita muito a inserção, remoção ou relocação de elementos numerados no texto (fato corriqueiro na escrita e correção de um documento acadêmico) sem a necessidade de renumerá-los todos.

2.1 FIGURAS

Na figura 2 é apresentado um exemplo de gráfico flutuante. Esta figura aparece automaticamente na lista de figuras. Para uso avançado de gráficos no LaTeX, recomenda-se a consulta de literatura especializada (GOOSSENS *et al.*, 2007).

Figura 1: Exemplo de uma figura onde aparece uma imagem sem nenhum significado especial.



Fonte: ABNTEX, 2009

2.2 TABELAS

Também é apresentado o exemplo da Tabela 1, que aparece automaticamente na lista de tabelas. Informações sobre a construção de tabelas no LaTeX podem ser encontradas na literatura especializada (LAMPORT, 1986; BUERGER, 1989; KOPKA; DALY, 2003; MITTELBACH *et al.*, 2004).

Tabela 1: Exemplo de uma tabela mostrando a correlação entre x e y.

x	**y**
1	2
3	4
5	6
7	8

Fonte: Próprio Autor.

2.3 EQUAÇÕES

A transformada de Laplace é dada na equação (1), enquanto a equação (2) apresenta a formulação da transformada discreta de Fourier bidimensional[1].

$$X(s) = \int_{t=-\infty}^{\infty} x(t)\, \mathrm{e}^{-st}\, dt \tag{1}$$

$$F(u,v) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) \exp\left[-j2\pi \left(\frac{um}{M} + \frac{vn}{N}\right)\right] \tag{2}$$

2.4 SIGLAS E SÍMBOLOS

O pacote abnTEX permite ainda a definição de siglas e símbolos com indexação automática através dos comandos `\sigla{}{}` e `\simbolo{}{}`. Por exemplo, o significado das siglas CCECOMP , DAEComp e UEFS aparecem automaticamente na lista de siglas, bem como o significado dos símbolos λ , v e f aparecem automaticamente na lista de símbolos. Mais detalhes sobre o uso destes e outros comandos do abnTEX são encontrados na sua documentação específica (ABNTEX, 2009).

[1] Deve-se reparar na formatação esteticamente perfeita destas equações!

3 FUNDAMENTAÇÃO TEÓRICA

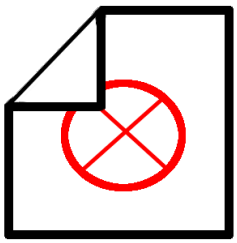
Neste capítulo será discutido na seção 2.1, algumas das ferramentas que já se propõe a realizar a comparação de moléculas, assim como alguns dos algoritmos que podem ser utilizados na análise de similaridade molecular que serão apresentados na seção 2.3, sendo que parte desses algoritmos já estão adaptados 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.

3.1 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).

Figura 2: Exemplo de uma figura onde aparece uma imagem sem nenhum significado especial.



Fonte: ABNTEX, 2009

4 METODOLOGIA

Descrever as principais ações realizadas. É preciso justificar, com base na literatura, a escolha feita pela metodologia, técnicas e instrumentos.

5 RESULTADOS

Apresentar os resultados da sua pesquisa.

6 CONSIDERAÇÕES FINAIS

Espera-se que o uso do estilo de formatação 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 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 LaTeX para usuários Linux. Para usuários Windows recomenda-se o editor TeXnicCenter (TEXNICCENTER, 2009). O LaTeX normalmente já faz parte da maioria das distribuições Linux, mas no sistema operacional Windows é necessário instalar o software 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 BibTeX, de forma a facilitar citações através do comando `\cite{}` e outros comandos correlatos do pacote abnTeX. A lista de referências deste documento foi gerada automaticamente pelo software LaTeX + BibTeX a partir do arquivo `abnt-uefs.bib`, que por sua vez foi composto com o gerenciador de referências JabRef.

O estilo de formatação LaTeX do curso de Engenharia de Computação da UEFS foi elaborados 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 TEX**. 2009. Disponível em: <http://sourceforge.net/apps/ mediawiki/abntex/index.php>. Acesso em: 8 nov. 2011.

BUERGER, D. J. **LATEX for scientists and engineers**. Singapura: McGraw-Hill, 1989.

CTAN. **The comprehensive TEX archive network**. 2009. Disponível em: <http://www.ctan. org>. Acesso em: 8 nov. de 2011.

GOOSSENS, M. *et al.* **The LATEX graphics companion**. 2. ed. Boston: Addison-Wesley, 2007.

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.

KOPKA, H.; DALY, P. W. **Guide to LATEX**. 4. ed. Boston: Addison-Wesley, 2003.

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.

LAMPORT, L. **LATEX:** a document preparation system. Boston: Addison-Wesley, 1986.

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

MENDELEY. **Mendeley:** academic software for research papers. 2009. Disponível em: <http://www.mendeley.com>. Acesso em: 8 nov. de 2011.

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

MITTELBACH, F. *et al.* **The LATEX companion**. 2. ed. Boston: Addison-Wesley, 2004.

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.

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

TEXNICCENTER. **TEXnicCenter:** the center of your LATEX universe. 2009. Disponível em: <http://www.texniccenter.org>. Acesso em: 8 nov. 2011.

WIKIBOOKS. **LATEX**. 2009. Disponível em: <http://en.wikibooks.org/wiki/LaTeX>. Acesso em: 8 nov. 2011.