轻工学报 2016 年 7 月 第 31 卷 第 4 期
JOURNAL OF LIGHT INDUSTRY Vol. 31 No. 4 Jul. 2016

引用格式:郑乃仁,邓玉林,Venkat Mathura,等. 基于众包的天然产物数据库及知识发现系统[J]. 轻工学报,2016,31(4):102 – 108.
中图分类号:R857.3 **文献标识码**:A
DOI:10.3969/j.issn.2096 – 1553.2016.4.015
文章编号:2096 – 1553(2016)04 – 0102 – 07

基于众包的天然产物数据库及知识发现系统

Crowdsourcing-based natural products database and knowledge discovery system

郑乃仁[1],邓玉林[1],Venkat Mathura[2],Fiona Crawford[2]
ZHENG Nai-ren[1],DENG Yu-lin[1],MATHURA Venkat[2],CRAWFORD Fiona[2]

1. 北京理工大学 生命学院, 北京 100081;
2. Roskamp 研究所, 佛罗里达 34243
1. *School of Life Science*, *Beijing Institute of Technology*, *Beijing* 100081, *China*;
2. *Roskamp Institute*, *FL* 34243, *United States*

关键词:
众包;天然产物数据库;知识发现系统

Key words:
crowdsourcing; natural product database; knowledge discovery system

摘要:针对目前天然产物数据库数据更新不及时、数据量不够大等问题,开发了基于众包的天然产物数据库及知识发现系统. 该系统利用众包技术构建一个天然产物数据库,使用分子指纹对分子结构进行编码,并采用 Tanimoto 系数计算相似度实现天然产物和相关文献的检索,可以实时扩充,并为生物学家了解当前研究热点,确定进一步研究方向提供参考.

收稿日期:2016 – 05 – 03
基金项目:国家重大科学仪器设备开发专项课题(2012YQ04014005)
作者简介:郑乃仁(1984—),男,河北省保定市人,北京理工大学博士研究生,主要研究方向为生物信息学.

Abstract: In view that the present natural products database data update was not in time, the data quantity was not big enough, the natural products database and knowledge discovery system was developed based on crowdsourcing. The system used the crowdsourcing technology to build a natural products database, used molecular fingerprint to encode molecular structure, and used the Tanimoto coefficient to calculate similarity to realize natural products and related literature retrieval, and could expand real-time for biologists to understand the current research hotspot and to offer reference to determine further research directions.

0 引言

所谓天然产物,广义上指自然界中的各种生物所产生的化学物质;狭义上,如具体到药物化学领域,则主要指生物的次级代谢产物. 与直接参与生物的生长、发育与繁殖的初级代谢产物不同,次级代谢产物并不是生物生存的必需品,而是帮助生物抵御天敌或在同类竞争中胜出的一项进化优势. 经过长期的自然选择,次级代谢产物常具有一定的细胞毒性,并拥有良好的生物活性,可直接用于治疗疾病. 所谓"凡毒蛇出没之处,七步内必有解救蛇毒之药",虽是小说家言,却不失为这一原理的通俗解释. 我国拥有庞大的传统中医药宝库,若能对历史上流传下来的诸多验方中的天然产物进行深入研究,必将大有收获. 屠呦呦先生因发现青蒿素而荣获2015年诺贝尔生理学或医学奖,就是传统与现代相结合的一项重大成就.

1950—2014年,全世界抗癌药物中有75%直接或间接来源于天然产物[1],表明天然产物已成为新药发现的重要基础资源. 因此,建立天然产物数据库,将为科研人员提供重要的帮助. 目前,国内外已经建成了一些天然产物的相关数据库. 例如:基于可扩展标记语言XML技术构建的海洋天然产物数据库[2]、喀麦隆天然产物数据库[3]和巴西天然产物数据库[4]等. 针对数据更新不及时、数据量不够大等问题,本文拟开发基于众包[5]的天然产物数据库及知识发现系统,利用互联网时代流行的众包形式和知识发现技术,实时更新与数据库中的天然产物有关的最新发表的文献信息,并按照用户指定的疾病名称列表进行分类筛选,以期帮助生物学家快速确定有价值的研究方向,并提供一种新的天然产物数据库的构建思路.

1 系统功能模块

笔者旨在建立一个包含720种天然产物及其主要信息的数据库,使用本系统的用户可以在http://npdb.aboutproteomics.com访问该系统,自由地注册账号. 系统主要包括如下功能模块.

1.1 新化合物添加模块

众包就是把工作外包给大众的一种生产组织形式. 网站上线后,允许用户自由在网站上添加内容,使得数据每时每刻都可能得以更新,克服了网站维护者时间或精力有限而无法及时更新的弊端. 这实质上是众包的一种形式.

所有注册用户,均可自行添加化合物,通过输入化合物的分子式等信息,即可实现化合物的添加(见图1). 系统会自动生成新添加化合物的分子结构图,使得数据库不断扩充.

1.2 天然产物检索模块

为了同时满足便于用户输入和计算机处理两方面的需求,输入系统采用简化分子输入行条目系统SMILES(Simplified Molecular Input Line Entry System)[6]. 该系统使用直观的方式对天然产物的分子式进行编码,可以用一个一维线性的有限长度字符串来表示任意分子结构式. 如图2所示的天然产物戊酸分子结构式,可以表示为CCCCC(O)=O. 通过以SMILES编码的形式输入天然产物的分子式,可以对数据库

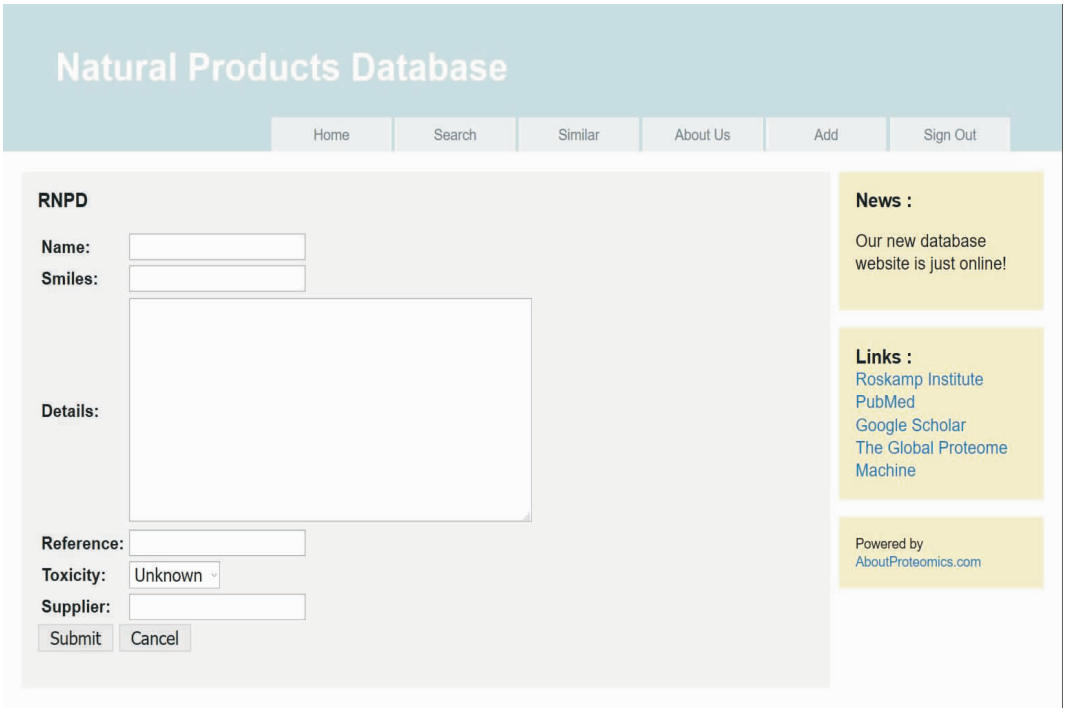

图 1　注册用户自行添加化合物页面

Fig. 1　Page of adding compound by registered users themselves

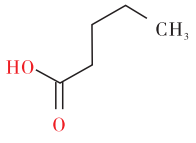

图 2　戊酸分子结构式

Fig. 2　Molecular structural formule of valeric acid

中的已有数据进行检索，找出与所输入的天然产物结构相似的天然产物，并按相似度进行排序. 分子指纹(Molecular Fingerprint)是一个二进制串，用来对分子的结构进行编码[7]. 在对天然产物结构进行相似度计算时，首先使用 Fingerprint 2(FP2)分子指纹算法对天然产物的分子进行编码，然后使用 Tanimoto 系数对不同天然产物之间的相似度进行计算[8]. 在对天然产物进行检索时，用户还可输入自己感兴趣的疾病名称列表，并用竖线分隔不同的疾病名称，这样，系统在获取最新发表的文献时，就会按用户指定的疾病名称来分类列出. 系统对某天然产物的检索结果如图 3 所示.

1.3　知识发现模块

当用户发起对数据库中天然产物的查询时，与数据库耦合在一起的知识发现模块就会自动发起对当前已发表相关文献的实时查询，并展示给用户，用户可以通过单击查看其详细信息，获取与自己感兴趣的化合物结构相似的其他化合物近期的研究成果，随时把握研究前沿，从而给自己进一步的研究提供参考.

2　系统的技术架构

开发数据库驱动的交互式网页应用系统需要一套由目标操作系统、网页服务器、数据库和服务器端编程语言组成的技术栈[9]. 技术栈是指为了支撑某项应用的开发而搭建相应平台的过程中需要的所有软件子系统或组件的集合. 传统的网页应用开发项目通常采用 Linux + Apache + MySQL + PHP，即 LAMP 作为技术栈[10]. 本项目使用了与此略有不同的变种 LAMP 方案，采用 Linux + Apache + MariaDB +

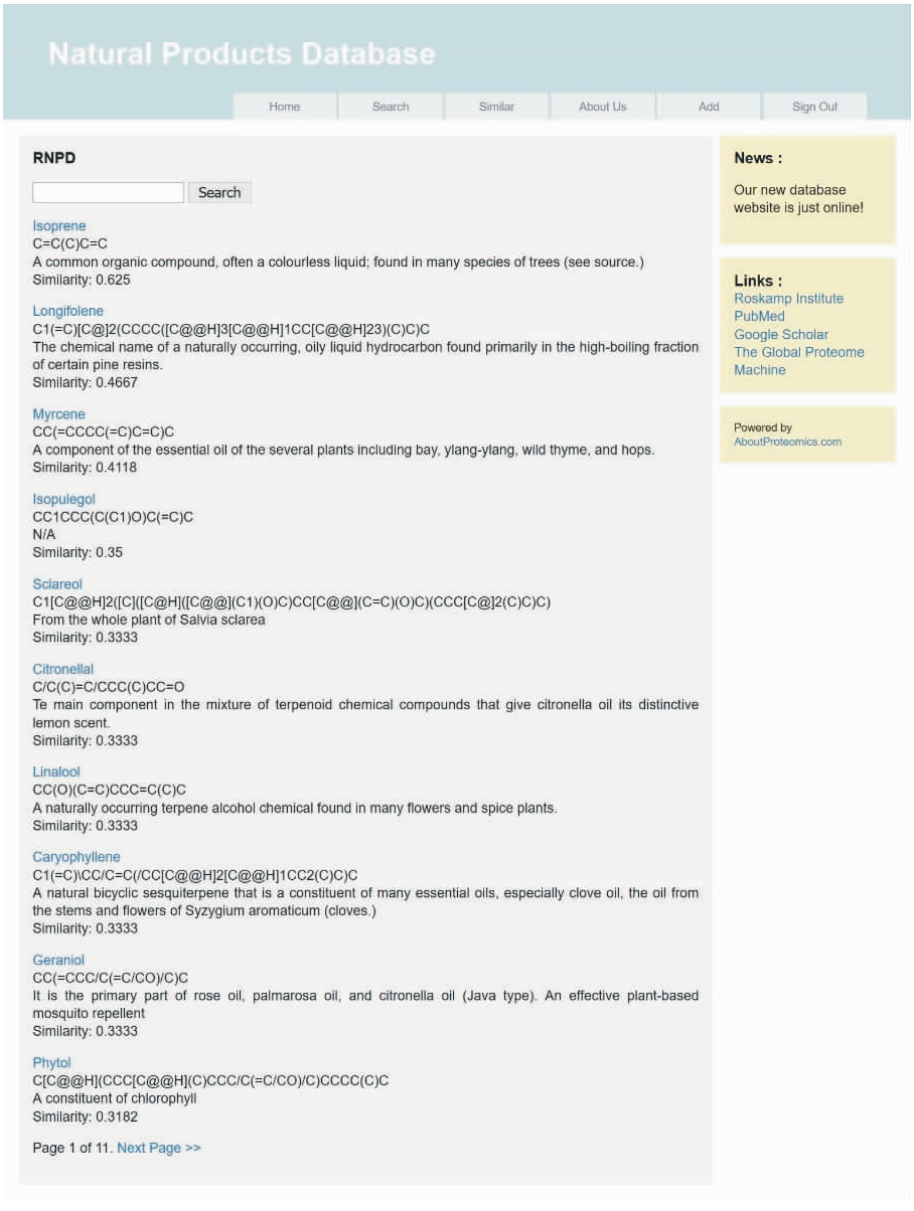

图 3　本系统对某天然产物的检索结果

Fig. 3　The system's query result of natural product

Python 作为技术栈.

CentOS 是基于红帽公司商业版 Linux (RHEL)开发的发行版,拥有大量社区技术支持,与新硬件兼容性好且运行稳定. 因此,本系统选择了 CentOS Linux 发行版作为本项目的操作系统. 而在超文本传输协议(HTTP)服务器的选择上,依然沿用经典的 Apache 套件. MySQL 一直以来都是开源数据库的首选,MariaDB 完全复用了 MySQL 的代码,并在其基础上作了改进,因此,本系统的数据库选用 MariaDB. 由于 PHP 语言和 HTML 语言很容易整合在一起,网站系统常选用 PHP 语言作为服务器端脚本语言. 然而,由于本系统的特殊性,需要对化合物进行相似性比对,因此选择了编程方式更为灵活、可以更方便调用相关化学库 OpenBabel 的 Python 语言.

OpenBabel 是一套使用 C ++ 语言开发的开源软件工具箱,可以很方便地对化学数据进行

各种读、写和格式转换等操作[11]. SWIG 是一项用来自动将 C ++ 语言编写的程序包装成其他语言可以调用的函数的技术[12]. Pybel 则是在此两项技术基础之上开发的调用 OpenBabel 功能的 Python 接口库[13].

笔者使用基于 Python 语言的 Django 框架开发了整套网站系统. 在对分子相似度进行比对和生成分子图像时,使用 Pybel 调用了通过 SWIG 技术封装起来的 OpenBabel 项目. 目前,通过对化合物分子的结构特征进行编码,生成相对应的分子指纹,然后对不同化合物的分子指纹进行对比,是一种很重要的化合物分子相似度计算方法. 在使用 OpenBabel 这一化学工具库进行分子相似度比对时,本系统使用了 FP2 指纹编码方案. FP2 是一种分子中长度从 1 个到 7 个原子长度的各种不同排列编码为二进制串的算法. 之后,采用 Tanimoto 系数来进行相似度的计算. Tanimoto 系数又称为广义 Jaccard 系数,其基本原理为:将一个二进制串看成由其每一位组成的集合,两个集合交集的元素个数,再除以两个集合并集的元素个数,即为这两个集合所代表的化合物的相似程度[14].

在获取文章时本系统使用了美国国家生物技术信息中心(NCBI)官方提供的 E-utilities 可编程接口,为用户提供了形如 http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db = <database> &term = <term> 的应用程序开发接口,并实时返回相应的 XML. 系统使用 Python 标准库中的 xml.etree.ElementTree 模块对 NCBI 返回的 XML 进行解析,以得到所需的文献信息.

本系统的技术架构如图 4 所示.

3　系统验证

为了对本系统的应用效果进行考察,用丹参素进行验证. 丹参是我国的传统中药,其有效成分是丹参素,分子式的 SMILES 编码形式为 OC(CC1 = CC = C(O)C(= C1)O)C(O) = O,将该 SMILES 分子式输入系统,同时使用默认的疾病列表参数. 检索发现,数据库中与之相似度最高的是 3,4 - Dihydroxy - L - phenylalanine,相似度达到 75%. 查看系统自动获取的与其相关的近期文献,发现:与高血压有关的 0 篇,与糖尿病有关的 0 篇,与阿尔茨海默症有关的 1 篇,与帕金森症有关的 17 篇(见图 5);丹参素的相似天然产物与帕金森症相关的文献远多于其他疾病,经过查阅证实确有文献报道丹参素与帕金森症相关[15]. 由此可见,本系统对

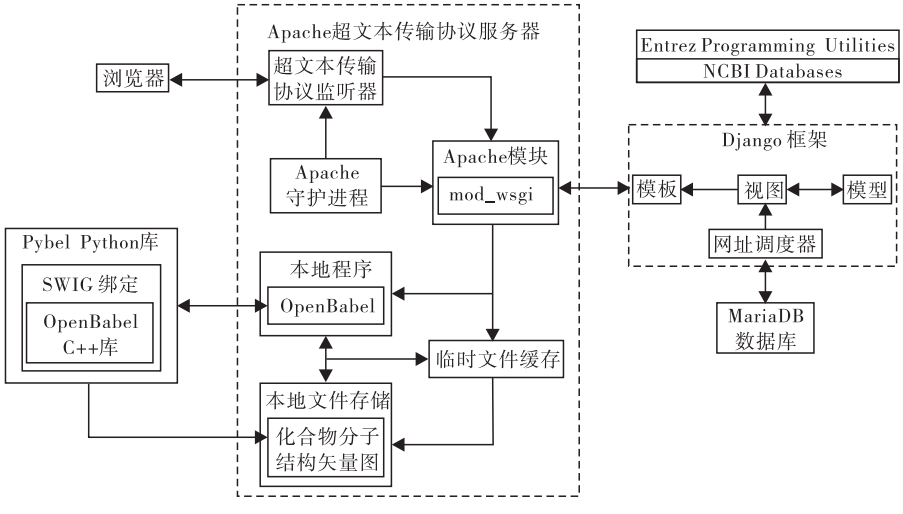

图 4　系统技术架构

Fig. 4　Technological Architecture of System

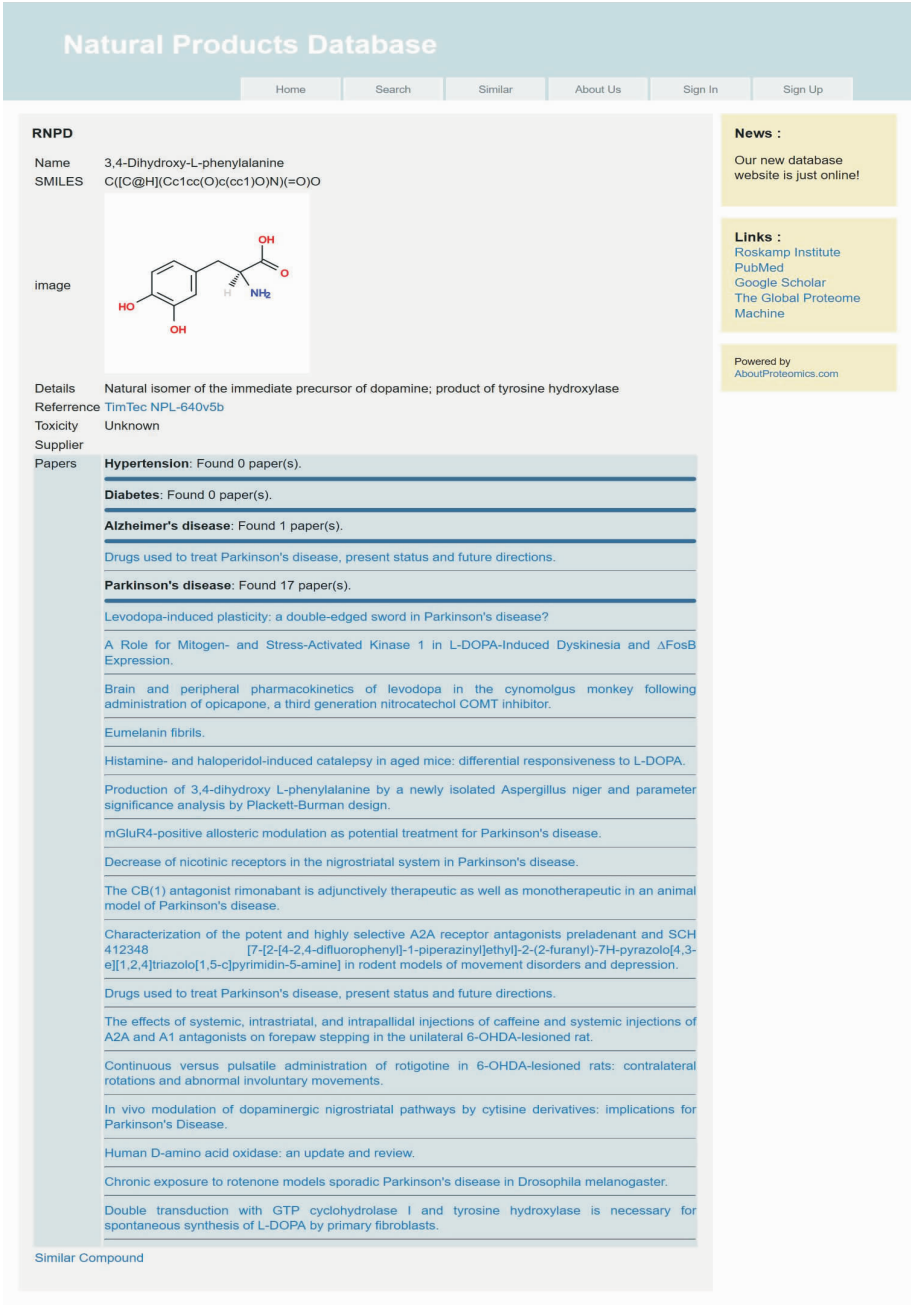

图 5 本系统以丹参素为例的知识发现

Fig. 5 The system's knowledge discovery with Danshensu as an example

天然产物的研究方向的确定,具有一定的参考与辅助作用.

4 结语

本文开发了基于众包的天然产物数据库及知识发现系统. 该系统通过采用众包的形式,利用知识发现技术可以实时获取最近发表的文献,使得数据库的内容扩充得以保证,并拥有良好的时效性. 本数据库在后台预留了用户权限分组功能,后续开发中,将完善用户权限管理模块,通过对自主注册的用户进行权限划分,以及对新添加化合物进行人工审核等手段来确保新增数据的正确性和有效性. 目前的文章获取,只是针对用户指定的特定疾病,获取相关领域与所关注化合物相关的文章列表. 由于 Python 语

言良好的可扩展性,未来可通过进一步开发人工智能算法进行数据挖掘,获得相关天然产物的靶标蛋白等信息.

参考文献:

[1] NEWMAN D J, CRAGG G M. Natural Products as sources of new drugs from 1981 to 2014[J]. Journal of Natural Products, 2016, 79(3): 629.

[2] 雷静,周家驹. 海洋天然产物数据库的设计与建立[J]. 化学通报, 2002, 65(5): 353.

[3] NTIE-KANG F, MBAH J A, MBAZE L M, et al. CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening[J]. BMC Complementary and Alternative Medicine, 2013, 13(1): 88.

[4] VALLI M, DOS SANTOS R N, FIGUEIRA L D, et al. Development of a natural products database from the biodiversity of Brazil[J]. Journal of Natural Products, 2013, 76(3): 439.

[5] HOWE J. The rise of crowdsourcing[J]. Wired Magazine, 2006, 14(6): 1.

[6] WEININGER D. SMILES, a chemical language and information system[J]. Journal of Chemical Information and Modeling, 1988, 28(1): 31.

[7] WILLETT P. Searching techniques for databases of two- and three-dimensional chemical structures[J]. Journal of Medicinal Chemistry, 2005, 48(13): 4183.

[8] HOLLIDAY J D, RANADE S S, WILLETT P. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases[J]. Quantitative Structure-Activity Relationships, 1995, 14(6): 501.

[9] Wikipedia. Solution stack[EB/OL]. (2016 - 06 - 08)[2016 - 01 - 02] https://en.wikipedia.org/wiki/Solution_stack.

[10] LEE J, WARE B. Open source development with LAMP: using Linux, Apache, MySQL and PHP[M]. New Jersey: Addison-Wesley Professional, 2002.

[11] OLBOYLE N M, BANCK M, JAMES C A, et al. Open Babel: an open chemical toolbox[J]. J Cheminf, 2011, 3: 33.

[12] COTTOM T L. Using SWIG to bind C ++ to Python[J]. Computing in Science and Engineering, 2003, 5(2): 88.

[13] O'BOYLE N M, MORLEY C, HUTCHISON G R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit[J]. Chem Cent J, 2008, 2(1): 5.

[14] LIPKUS A H. A proof of the triangle inequality for the Tanimoto distance[J]. Journal of Mathematical Chemistry, 1999, 26(1): 263.

[15] 李洪莲. 丹参素对帕金森小鼠运动障碍的影响及机制研究[D]. 烟台: 烟台大学, 2014.