doi:10.3969/j.issn.1672-5565.2014.04.11

计算机辅助药物筛选平台及应用

宋新蕊,李　达,陈　洁,赵　勇*
(北京市计算中心,北京 100094)

摘　要:先导化合物发现是创新药物研发的最重要环节之一。针对目前海量功能不明确的小分子化合物,本文构建了一个用来实现快速发现先导化合物,有效降低药物研发成本的计算机辅助药物筛选平台。该平台采用分布式架构思想,集成了AutoDock Vina和多个小分子库,具有数据安全、计算与存储的负载均衡以及实时监控的特点。应用平台进行先导化合物筛选,在较短时间发现了有针对性的活性小分子化合物,命中率高,大大缩短先导化合物发现周期。该平台具有很好的实用性和良好的扩展性。

关键词:虚拟筛选;超算平台;AutoDock Vina;化合物库;先导化合物发现
中图分类号:R91　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-04-300-05

Computer aided drug screening platform and its application

SONG Xinrui, LI Da, CHEN Jie, ZHAO Yong*
(*Beijing Computing Center*, *Beijing* 100094, *China*)

Abstract: Lead compound discovery is the key step of innovative drug research and development. For large numbers of small molecules whose function are not clear, we have established a computer aided drug screening platform (www.vslead.com), which has been considered as a way for quick lead compound discovery and costs reduce. The platform adopted distributed architecture and integrated AutoDock Vina and a number of small molecule libraries. It is featured with data security, load balancing about computation and storage, and real time monitoring. Utilizing the facility provided by the platform, we found an active compound tested by wet lab experiment quickly. Therefore, users can have higher chances to find active compounds and the time of lead compound discovery can be shortened. The platform has very good practicability and extendibility.

Keywords: Virtual screening; Super computing platform; Compound library; AutoDock Vina; Lead compound discovery

创新药物研发具有周期长、费用高的特征。一般而言,从疾病靶标的确定、先导化合物的发现、临床前药理学和药物代谢动力学及安全性评价研究到药物成功上市,一般需要花费10~15年时间。过去4年,研发一种新药的平均费用为13亿美元[1]。对于企业而言,提高研发效率、降低研发成本、缩短新药发现和早期开发时间是保持市场竞争实力的关键[2]。

先导化合物的发现是创新药物研究的关键环节之一。先导化合物通常是通过合理药物设计、组合化学、高通量筛选、虚拟筛选的方法来发现,或是药物化学专家将来源于天然产物或微生物代谢物中的化学成分提取出来,应用各种动物模型进行筛选,从中发现新的功能性化合物。由于组合化学技术的发展,化合物的合成速度显著提高,能够更多更快的发现先导化合物或有功能的化合物,对于创新药物研究是一项挑战,也是缩短新药发现时间的关键。高通量筛选技术采用自动化的操作系统,可以进行大规模的化合物筛选,但是这种方法筛选设备复杂,需要培养大量的靶酶或靶细胞,阳性率低,并且需要大量资金支持,因此仅仅采用高通量筛选的方法进行先导化合物发现成本高效率低。随着计算机技术的

收稿日期:2014-08-15;**修回日期**:2014-09-22.
基金项目:北京市科学技术研究院创新团队计划(IG201305C1)支持。
作者简介:宋新蕊,女,硕士,工程师,研究方向:小分子化合物数据库构建与虚拟筛选;E-mail:songxr@bcc.ac.cn.
*通信作者:赵勇,男,博士,北京市特聘专家,研究方向:生物信息学与创新药物研究;E-mail:zhaoyong@bcc.ac.cn.

更新以及大数据技术的发展,应用虚拟筛选策略发现先导化合物逐渐成为主流。这种策略通过各种算法对大量化合物库进行搜索来获得有功能的化合物分子,其中应用分子对接方法进行药物或功能化合物发现是一项有效的筛选技术,该技术通过计算的方法将靶蛋白和小分子化合物进行一对一的对接,从大量的化合物库中筛选出与靶蛋白有作用的小分子,从而发现先导化合物。与高通量筛选相比,虚拟筛选方法可以富集活性化合物,降低筛选成本,提高药物筛选的可行性,因此应用虚拟筛选技术进行药物发现已成为新药发现的重要方法。

然而,进行虚拟筛选也存在一些问题,目前 pubchem 数据库中有 4 900 多万个化合物,面对目前如此庞大的化合物空间,如何在较短的时间快速从海量的小分子化合物数据库中筛选出针对靶标有活性的小分子,在现有的计算资源条件下仍是一个需要考虑的问题[3]。目前能够进行大规模虚拟筛选的自动化操作平台比较少,而且平台操作比较复杂,如:Shoichet 实验室开发的 DOCK Blaster[4],主要应用 UCSFDOCK 进行分子对接,对于给定受体结构筛选 ZINC 数据库来寻找潜在的活性小分子,分成 6 个步骤进行,参数设置比较麻烦,而且参数的设置对结果的可靠性影响较大。台湾 YC 实验室开发的 Iscreen[5],可以在线对传统中药进行虚拟筛选,但是小分子化合物数目较少,只针对传统中药,规模较小,多样性不高,筛选时间比较长。国内用于大规模虚拟筛选的服务平台鲜为少见,由于计算资源的限制,各科研院所也只是内部小规模应用单机或服务器进行筛选。当前我国药物研发水平与世界领先水平差距较大,但我国要上市世界级甚至首创药物还是有机会的,也只是个时间问题,因此,建立一个大规模、简单易用的计算机辅助药物筛选平台来缩短先导化合物发现周期、加快我国新药研发进程、增强我国在市场中的竞争力是非常迫切的。针对这一需求,构建了计算机辅助药物筛选平台,实现了大规模虚拟筛选、并在较短时间发现了针对耐药菌有活性的先导化合物,该平台的广泛应用将有助于推动我国创新药物研发。

1　平台总体设计

根据药物作用的靶标结构是否已知将虚拟筛选方法分为基于结构的虚拟筛选和基于配体的虚拟筛选。基于结构的虚拟筛选基本途径是对蛋白质靶标结构进行分析,选定适当的药物作用靶点,针对该靶点利用现有的小分子库进行一对一的模拟分子对接,然后预测小分子的构象和结合亲和力。一般相信基于结构的方法可以提供便宜、合理、快速找到先导化合物的方法。计算机辅助药物筛选平台是采用基于蛋白质靶标的虚拟筛选方法构建的分布式虚拟药物筛选平台,平台主要由虚拟筛选对接工具的整合、小分子化合物数据库构建以及分布式虚拟筛选架构设计等部分构成。

1.1　虚拟筛选对接工具选择

在过去的几十年里开发出了很多种对接软件,应用比较广泛的对接软件有 FlexX[6]、AutoDock[7]、AutoDock Vina[8]、ICM[9]以及 GOLD[10]等。不同的对接软件应用的算法不同,基于分布式虚拟筛选架构的思想,结合高性能计算集群优势,选择 AutoDock Vina 作为平台实现虚拟筛选的对接工具。AutoDock Vina 是 AutoDock 的另外一个版本,可以实现多核运行,相比于 AutoDock 它的速度更快,准确率高,易应用。因此,AutoDock Vina 能够充分利用分布式计算的优势,对于进行大规模分子对接比较适用。

1.2　小分子化合物数据库构建

合理的小分子化合物数据库是进行虚拟药物筛选的前提和基础。由于现存的小分子化合物数据库种类繁多、分类比较杂乱、小分子结构转化容易出错、大多数小分子化合物难以购买到,针对这些不足,构建了格式准确、有注释信息、与生物医学领域密切相关、可购买、适用于虚拟筛选的三维小分子化合物数据库。数据库构建流程如图 1 所示:

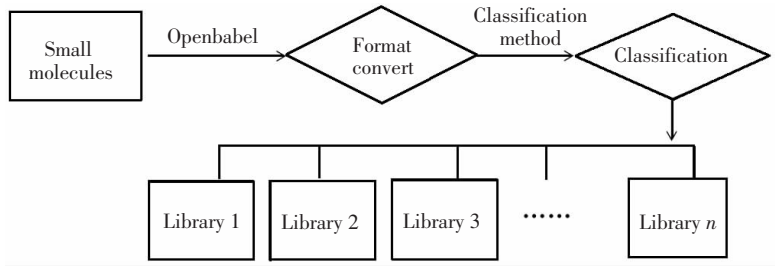

图 1　小分子数据库构建流程

Fig.1　Construction process of small molecule library

首先对于收集到的每个小分子应用 openbabel 进行格式转换[11]，将 mol2 或 sdf 格式的小分子转换成 vina 识别的 pdbqt 格式。然后对收集到的小分子按照一定原则进行分类。尽管当前计算能力大大提高，但盲目的对所有的小分子进行筛选会浪费大量时间和计算资源，对后续的小分子筛选也会造成很大负担，因此有必要对小分子进行分类处理。根据小分子的分子量、氢键供体、氢键受体、可旋转键数目、cLogP 等性质将小分子进行个性化分类[12]，将分类的小分子存放于分布式系统中，提供平台进行大规模筛选。

目前，平台构建的小分子化合物数据库包含了 2 000 多万个小分子，涵盖了先导化合物、类药化合物、中草药成分库、多样性化合物库、上市药物库、片段库、农药库、天然产物库等多种类别的小分子化合物库，这些小分子来源于 ZINC 数据库、中医药资料库等多个数据库。

1.3 分布式虚拟筛选应用架构设计与系统测试

传统的虚拟筛选以及网格环境下的虚拟筛选都需要化学家手动上传小分子文件和收集结果数据，无法实现自动对接以及结果文件的自动收集，其中涉及到的操作误差、繁杂工作流程都会给药物开发带来麻烦。分布式虚拟筛选系统收到用户提交的任务后由后台的调度系统将任务派送给底层分布式集群，由分布式集群进行高性能计算，后台调度系统会对任务进行实时监控，以保证各项任务的协调进行。这种分布式设计思想符合三层架构模式，将应用架构分为表现层、业务逻辑层和数据访问层，在层次上实现了“高内聚，低耦合”的目标。此外，平台的分布式架构具有数据安全、存储负载均衡、实时监控、弹性扩增、快速部署等特点，为正常运行虚拟筛选提供了保证。

（1）虚拟筛选的分布式处理

为了将众多小分子文件在集群内不同节点进行分布式筛选，首先把整体的小文件集按照集群节点的实际情况和就近原则，切分成固定大小的数据分片，并将用户提交的任务发送到不同节点上。然后使用对接工具 AutoDock Vina 进行靶标和小分子一对一的对接计算，将计算任务最密集的部分分布在多节点上分布式执行，以达到高效运算的目的。最后将各个节点计算结果进行排序，提供用户需要的排名靠前的 N 个筛选结果。

（2）后台调度系统

用户通过网站提交数据来与后台进行交互，后台调度系统的主要作用是将上层网站与底层分布式集群连接在一起。调度系统不断对任务列表进行扫描，检测系统中是否有新任务提交，如果有新任务提交则开始对新提交任务进行处理，否则检测列表中正在运行的任务，如果有正在运行的任务，则对运行的任务进行跟踪，否则过一段时间进行下轮扫描。通过对众多任务启动、监控的调度，保证业务的顺利进行。

（3）性能测试

对于平台的运算性能，采用单节点测试和多节点的分布式集群测试来检测平台对于大规模筛选的筛选效率和稳定性。测试结果见表 1。

表 1 单节点和多节点虚拟筛选时间比较

Table 1 Comparison of virtual screening time between single node and multi node

No.	Nodes	Cpus	Small molecules	Time
1	1	24	6 000	23 h
2	3	72	6 000	5 h 18 min 3 sec
3	3	72	6 000	5 h 13 min 50 sec
4	13	232	6 000	56 min 15 sec

测试中使用了单节点、3 个计算节点和 13 个计算节点分别对 6 000 个小分子进行筛选，随着节点数增加，总的筛选时间显著减少，对于大规模的筛选，平台能够显著减少虚拟筛选所用时间，提高筛选效率。此外，相同的节点数进行筛选所用的时间很接近，说明平台具有良好的稳定性。

2 计算机辅助药物筛选平台的应用

2.1 计算机辅助药物筛选平台的应用方法

计算机辅助药物筛选平台实现了自动化的虚拟筛选，首先需要登录网站 www.vslead.com 进行注册。平台使用流程（见图 2）。

完成注册后根据收到的确认邮件登录平台。平台构建了多种小分子化合物数据库，可以勾选一个或多个小分子库进行筛选。更重要的，用户需要上传蛋白质结构文件并填写靶点的空间位置信息。最后，输入需要返回的结果数目，提交筛选任务，等待下载筛选结果。

平台返回的筛选结果包括小分子与靶蛋白对接的构象文件、log 文件和排名靠前的 N 个结果排序文件。通过查看筛选结果选择感性趣的小分子进行后续分析。

2.2 应用计算机辅助药物筛选平台发现先导化合物

目前有成千上万的小分子化合物功能不明确，这些功能性的小分子化合物的应用比较广泛，如在临床医学[13]、食品添加剂[14]、化妆品添加剂[15]等

应用方面都可以发挥一定的功能作用。平台构建的庞大的小分子化合物库中,很多小分子化合物都有可能发挥一定的功能作用。针对庞大的潜在功能性化合物资源库,北京市计算中心赵勇博士针对耐药菌相关蛋白利用计算机辅助药物筛选平台做了大规模的虚拟筛选,发现 2 个针对耐药菌有活性的小分子,经过多次生物活性实验验证后发现这两个小分子化合物确实有抗耐药菌功能[16]。

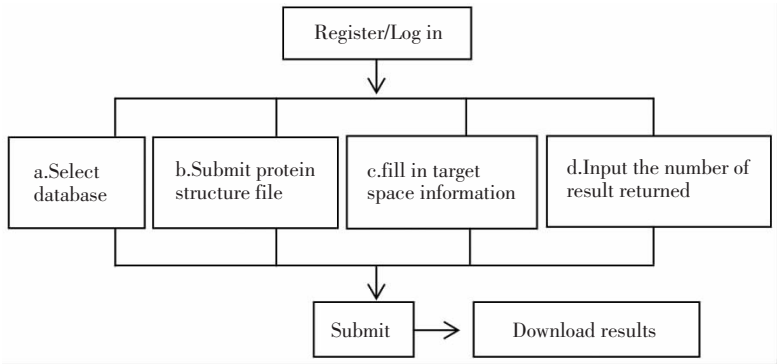

图 2　计算机辅助药物筛选平台使用流程

Fig.2　Application of computer aided drug screening platform

3　结　论

基于构建计算机辅助药物筛选平台的目的,该平台基本实现了以下功能:第一,简化操作流程。用户只需要选择感兴趣的小分子化合物库,上传一个蛋白质靶标结构文件,填写靶点的空间位置信息,即可实现大规模的虚拟筛选,省去了繁杂的软件操作以及构建化合物数据库的过程。第二,方便快捷。平台以庞大的计算资源为基础,采用分布式架构思想,极大的缩短了虚拟筛选的时间。第三,有效筛选。平台构建了丰富的化合物数据库,用户对靶点位置分析合理,基本可以保证发现有活性的小分子化合物。

同时,平台的计算和存储能力可以进行扩增,因此可以根据自身的需求灵活的选择计算或存储资源,避免资源浪费。平台具有良好的扩展性,设计了多个接口,方便添加更多的功能。然而,平台也存在一些不足,例如,目前的化合物数据库是固定的,用户还不能根据自己的筛选条件自动生成个性化的数据库去做筛选,在后续的功能完善和升级中会将该功能进行补充和完善。随着我国药物研发事业的蓬勃发展以及各种需求的不断增加,计算机辅助药物筛选平台也会不断增加新的功能,相信该平台的不断完善以及广泛应用会提供广大药物研发人员简单快捷的方法,共同加速我国创新药物研究的进程。

参考文献(References)

[1] MULLARD A. 2013 FDA drug approvals[J]. Nature Reviews Drug Discovery, 2014, 13(2): 85-89.

[2] 袁丽,杨悦. 国际创新药物研发现状及未来发展趋势[J]. 中国新药杂志,2013,22(18):2120-2125.
YUAN Li, YANG Yue. Current situation and future trend of international drug innovation[J]. Chinese Journal of New Drugs, 2013, 22(18): 2120-2125.

[3] 吴可柱, 李昆, 李爱秀. 虚拟筛选技术与新药开发[J]. 武警医学院学报, 2011, 20(5): 415-419.
WU Kezhu, LI Kun, LI Aixiu.Virtual screening and new drug discovery[J]. Acta Academiae Medicinae CPAF, 2011, 20(5): 415-419.

[4] IRWIN J J, SHOICHET B K, MYSINGER M M, et al. Automated docking screens: a feasibility study[J]. Journal of Medicinal Chemistry, 2009, 52(18): 5712-5720.

[5] TSAI T Y,CHANG K W, CHEN C Y C. IScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan[J]. Journal of Computer-Aided Molecular Design, 2011, 25(6): 525-531.

[6] KRAMER B, RAREY M, LENGAUER T. Evaluation of the FLEXX incrementalconstruction algorithm for protein-ligand docking[J]. Proteins: Structure, Function, and Bioinformatics, 1999, 37(2): 228-241.

[7] MORRIS G M, GOODSELL D S, HALLIDAY R S, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function[J]. Journal of Computational Chemistry, 1998, 19(14): 1639-1662.

[8] TROTT O, OLSON A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading[J]. Journal of Computational Chemistry, 2010, 31(2):

455-461.

[9] ABAGYAN R, TOTROV M, KUZNETSOV D. ICM-a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation[J]. Journal of Computational Chemistry, 1994, 15(5): 488-506.

[10] JONES G, WILLETT P, GLEN R C, et al. Development and validation of a genetic algorithm for flexible docking [J]. Journal of Molecular Biology, 1997, 267(3): 727-748.

[11] O'BOYLE N M, BANCK M, JAMES C A, et al. Open Babel: An open chemical toolbox [J]. Journal of Cheminformatics, 2011, 3(1): 1-14.

[12] IRWIN J J, SHOICHET B K. ZINC-a free database of commercially available compounds for virtual screening [J]. Journal of Chemical Information and Modeling, 2005, 45(1): 177-182.

[13] 龚陈媛,陆宾,杨莉, 等. 石斛联苄类化合物抑制血管新生的机制[J]. 药学学报,2013, 48(3):337-342.
GONG Chenyuan, LU Bin, YANG Li, et al. Bibenzyl from Dendrobium inhibits angiogenesis andits underlying mechanism[J]. Acta Pharmaceutica Sinica, 2013, 48(3): 337-342.

[14] 袁鹏,陈莹, 肖发, 等. 姜黄素的生物活性及在食品中的应用[J]. 食品工业科技,2012, 33(14):372-375.
YUAN Peng, CHEN Ying, XIAO Fa, et al. The bioactivities of curcumin and its application in foods[J]. Science and Technology of Food Industry, 2012, 33(14): 372-375.

[15] 张玉杰,徐文清,沈秀. 迷迭香酸的提取分离及药理学新发现[J]. 中国新药杂志,2013,22(4):433-437.
ZHANG Yujie, XU Wenqing, SHEN Xiu.New progress in the research of rosmarinic acid separation,purification and pharmacological actions [J]. Chinese Journal of New Drugs, 2013, 22(4): 433-437.

[16] ZHAO Y, SONG X R. Search of potential inhibitor against antibiotic-resistant bacteria from a serial of small molecular libraries[C] //proceedings of BIT's 4th annual International conference of medichem 2013. Hainan: China Medicinal Biotech Association, Haikou Municipal Government, Information Research Center of International Talents, SAFEA, 2013:93.