科学通报 2015 年 第 60 卷 第 8 期: 689 ~ 693
专题: 大数据与化学 进 展 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

大数据时代的药物设计与药物信息

孙潭霖[①], 裴剑锋[①②*]

① 北京大学前沿交叉学科研究院定量生物学中心, 北京 100871;
② 北京大学天然药物及仿生药物国家重点实验室, 北京 100191
* 联系人, E-mail: jfpei@pku.edu.cn

2014-10-23 收稿, 2014-11-28 接受, 2015-02-06 网络版发表
国家高技术研究发展计划(2012AA020301)和天然药物及仿生药物国家重点实验室开放基金资助

摘要 大数据时代, 以数据驱动的药物研发(data-driven drug research and development)方式有望显著提高药物研发成功率、缩短药物研发周期以及降低药物研发成本. 本文简短综述了近年来药物设计和药物信息相关数据整合和数据挖掘的最新研究概况, 并对大数据时代的药物设计与药物信息研究提出了展望.

关键词
大数据
药物设计
药物信息
语义网络
深度学习

大数据时代以数据量持续增加(volume)、分析速度持续增长(velocity)和数据形式的多样性(variety)著称, 大数据还包括数据的真实性(veracity)、数据的价值化(value)和数据的复杂性(complexity)等特征. 大数据的战略意义不在于海量信息的积累, 而在于运用专业的数据存储、整合及挖掘手段对数据进行处理, 解决产业瓶颈. 众所周知, 研发新药的道路已经变得越来越艰难. 从寻找新的备选化合物, 到层层实验审批, 往往要花去十几年时间和大量的金钱, 这其中还有很多努力最终会以失败告终. 根据塔夫茨药物研发中心(Tufts center for the study of drug development, CSDD)最近的报告, 现在平均要花25亿美元以上才能得到1个成功上市的新药, 与2003年的数据相比上涨了145%. 大数据时代带来以数据驱动方式的药物研发(data-driven drug research and development), 能否由此显著提高药物研发的成功率以及降低药物研发的周期和成本, 是一个值得期待的话题.

随着高通量筛选、深度基因组测序、临床实验等生物、化学、医药研究数据急速积累, 如何整合庞大的数据日益成为一个关键问题; 另外, 由于大数据的高度复杂性, 如何对这些数据很好地进行挖掘也是一个关键问题. 本文将主要对这2个问题进行讨论, 并展望大数据时代对药物设计与药物信息研究的影响.

1 数据整合

传统的基于单一靶标的药物研发过程因效率低、开支高而难以满足市场需求. “基于系统的药物设计”将药物分子信息与疾病调控网络、基因组、蛋白质组、代谢组等各类数据信息进行综合利用, 是未来的药物设计方向之一[1]. 目前的各类药物信息数据多用“语义网络”(semantic web)实现整合. 语义网络概念及技术应用统一的逻辑架构对各领域数据进行高效整合, 实现信息的共享交互, 同时对信息进行提取和辨识, 其中Resource Description Framework (RDF), Web Ontology Language(OWL)和Simple Protocol and RDF Query Language(SPARQL)是实现标准语言组织联系、注释及检索的3大核心. RDF是一种简单的语言形式, 通常以“三联体”的形式出现(名词-动词-名词), 它在生物信息学中的应用实现了横跨生物、化学、药学等数据库中的各种实体的联系; OWL对RDF信息进行注释, 连接各数据库中的术语; SPARQL对RDF进行跨库注释搜索[2]. Linked data是

引用格式: 孙潭霖, 裴剑锋. 大数据时代的药物设计与药物信息. 科学通报, 2015, 60: 689–693
Sun T L, Pei J F. Drug design and drug information in the big data era (in Chinese). Chin Sci Bull, 2015, 60: 689–693, doi: 10.1360/N972014-01114

“语义网落”实现的关键一环, 用于实现不同数据源之间的语义关联.

Linked Open Drug Data(LODD)以RDF的形式关联了有关药物的不同方面信息, 例如药物对基因表达的影响、药物全面的靶标信息等, 方便研究者通过检索一个关键词(例如“阿尔茨海默综合症”)而得到与其相关的所有动态、可视化的信息(疾病特征、相关基因、药物等)[3]. 目前, LODD已经整合了多个开源数据库信息, 包括DrugBank, ClinicalTrials.gov, DailyMed和Diseasome等. “语义挖掘”(text mining)能够从非结构化的文本信息中提取未知信息. 通常分为2步: 第1步是信息检索(information retrieval), 第2步是信息提取(information extraction). Downing等人[4]发展的SPECTRa-T能够从电子文档中获取化学分子信息并存储为RDF格式, 方便之后继续利用. 一种文献检索工具LSGraph 能够基于关键词和基因注释提取文献摘要、常见基因和蛋白数据库中的关键信息, 并对信息进行扩展、聚类最终得到与疾病相关的关键靶标[5]. Percha等人[6]用文献检索工具构建了基因-药物关系, 并以此训练出一种搜索文献中Drug-drug Interaction(DDI)的工具, 提示药物副作用的机理. 图论(graph theory)也可以应用到网络、实体关系构建中. 通过检索关键词, 图论相关算法可以找到检索实体之间网络最短路径或有意义的关联信息. Zhu等人[7]发展的Chemogenomic Explorer基于RDF链状链接分析药物小分子与其相关基因的联系并预测相关的疾病. 另外, RDF也大量应用于quantitative structure activity relationships(QSAR)对分子的描述中[8].

大数据时代也需要各领域研究实体交互、合作、信息共享. Hohman等人[9]发展了一种汇集了科研团队、数据库信息、基于web生化分析软件的Collaborative Drug Discovery(CDD)研究社区公共平台. 在平台上, 研究者可以检索信息、存储和发布结果、共享数据、寻找合作者支持, 在“隐私”和“公开”之间切换, 他们还列举了3个CDD如何协助世界各地的研究小组合作研发新药的例子.

传统药物信息数据的存储和整合方式势必不能满足大数据时代数据量急速膨胀的需求. 新型大数据存储技术和解决方案正处于快速发展的时代, 如Hadoop 是一款开源软件, 能够处理海量的各种结构(包括无结构)的数据, 经济有效、灵活且具有一定的容错能力. 与传统的关系型数据库不同, 大数据时代还要求数据存储方面要有庞大的水平扩展性, 而NoSQL正是致力于改变这一现状的非关系型的数据存储技术. 这些方兴未艾的技术都可以应用于药物信息大数据的存储和整合.

2 数据挖掘

数据分析是大数据处理流程里的核心部分, 大数据分析即是从大量的、不完全的、有噪声的、模糊的数据集中识别有效的、新颖的、潜在有用的, 以及最终可以理解的模式的过程. 大数据带来了思维方式的转变: 全样本代替随机抽取、相关关系取代因果关系以及对不精确的容忍度增加. 传统的药物设计与药物信息应用中数据挖掘方法可以分为“基于相似性”的方法(相似性搜索、k-means算法等)和“非基于相似性”的机器学习方法(支持向量机、神经网络算法等). 根据Kimelford 和Wahba[10]的理论, 机器学习的实质也基于相似性, 不同在于它们对选取的特征进行了进一步的抽象, 根据抽象特征的相似性重构数据的组织形式, 这更加符合人类的认知过程.

芯片实验数据较早应用于疾病(尤其是癌症)靶标的发现中, 根据基因表达模式的相似性, 应用有监督/无监督的聚类过程能够区分不同组织来源、癌症进展时期的样本, 识别异常表达的基因图谱用于后续研究[11]. Liu等人[12]利用小分子在基因组空间而非化学空间的相似性, 以简单的分子印记(fingerprint)为特征对实验集中的每个小分子构建了靶标谱(target profile), 并用Drugbank和Therapeutic Target Database测试集中的小分子进行了验证. 他们的研究为老药新用、药物副作用的研究提供了手段. Ding等人[13]比较了当前应用于相似性比较的8种流行算法预测药物-靶标相互作用的能力, 发现算法的优劣与实验条件的设定有很强的关系.

在机器学习方面, Li和Lai [14]以蛋白序列性质(氨基酸的种类、疏水性、极性等)作为特征输入, 用支持向量机(SVM)预测蛋白是否为药物靶标, 并预测了Swiss-Prot中的潜在蛋白靶标, 一些阳性结果已经被证实并得到了关注. 同样利用SVM, Sugaya和Ikeda[15]以结构特征、物理化学性质为输入预测了1295个蛋白相互作用界面(protein-protein interface)的成药性, 综合准确度达到81%. Xuan等人[16]采用SVM等方法建立了4个HIV-1整合酶ST过程抑制剂抑制活性的定量预测模型. 这些模型包含了551个使用放射性标记方法

测试得到的HIV-1整合酶ST过程抑制剂, 其中SVM模型的预测结果的复相关系数R超过0.90, 均方根误差(RMSE)低于0.41. Heikamp和Bajorath[17]综述了SVM在药物发现领域的广泛的应用, 包括化合物的分类、寻找活性化合物以及性质预测等.

人工神经网络算法是机器学习算法之一, 它是模仿动物神经网络行为特征, 进行分布式并行信息处理的算法数学模型. 这种算法依靠系统的复杂程度, 通过调整内部大量节点之间相互连接的关系, 从而达到处理信息的目的. 它有较强的非线性映射能力, 适合解决内部机制复杂的问题. 人工神经网络算法在药物设计研究中广泛使用. 在学习目标方面, 有预测类药性[18]、QSAR分析[19]; 在输入特征方面, 有基本的物理化学性质[20]、分子描述(descriptor)或印迹[19]以及分子结构的矢量描述[21]. 人工神经网络算法中存在的过度拟合问题值得重视, 它源于样本范围狭窄、参数的选取和初始化较为随机, 表现为"机器"对训练集分子预测能力较好、对测试集或其他类别的分子预测能力差, 因而限制了人工神经网络算法的应用范围(泛化能力差, 一些研究用"集成学习"的方式解决过度拟合问题[22]). 再者, 目前应用到药物研究中的神经网络通常只有一层隐含层(hidden layer), 容易造成深度不足, 特征提取能力较差, 需人工参与特征的筛选.

传统的数据挖掘方法在分析大数据时一般会显得能力不足, 近年来出现了一些可针对大数据的分析方法, 其中值得注意的是一种称为深度学习(deep learning)[23]算法. "谷歌大脑"(Google Brain)是由斯坦福大学著名的机器学习教授Andrew Ng与大规模计算机系统方面的世界顶尖专家Jeff Dean共同主导, 2012年6月, 运用16000个CPU 的并行计算平台和深度学习算法, 使机器系统自己发现或者领悟了"猫"的概念. 深度学习算法与传统神经网络算法相比, 包含多层隐层, 并在学习时维持不同抽象层次之间信息传递的双向保真, 能够将输入的诸多特征抽象概括成为高级类别或属性; 同时, 它可以进行"无监督学习", 对数据结构的要求较低, 能够过滤掉诸多噪声, 更加接近人脑的认知模式. 深度学习被成功地用于包括图像识别、语音识别等多个领域, 大幅度提高了预测的准确度. 最近, Google公司使用深度学习方法, 实现了用计算机准确识别图像, 并能用自然语言描述出图像的内容[24].

在药物设计研究中, Lusci等人[21]用2层隐含层人工神经网络算法训练了小分子的水溶性预测模型, 他们的学习算法接近了深度学习算法, 但和传统算法比没有得到特别出色的预测结果, 原因可能是训练样本数据量小, 不能体现出算法对于大数据的优势. 深度学习算法具备的强大特征抽象能力及大数据处理能力事使其在药物设计和药物信息领域具有广泛的应用前景.

3 总结与展望

大数据时代, 基于系统的药物设计与药物信息研究更需要数据的整合, 包括研究手段的整合和不同领域研究实体的合作, 目前主要基于语义网络的公共数据整合提供了大量可供挖掘的数据结构. 尽管如此, 还有大量的数据以及不断快速产生的新数据未被整合, 迫切需要用大数据相关技术联系起这些信息孤岛, 为药物研发提供用于深层信息发掘的大数据. 由于相关数据的整合量还不充分, 目前Hadoop, SPARK, NoSQL等大数据技术还没有应用于药物设计和药物信息处理的报道. 另外, 药物研发的大量珍贵数据处于私有状态, 能公开获取的数据不足, 这也是大数据时代药物设计与药物信息研究面临一个重大挑战. 目前最先可能得到使用的是疾病相关高通量测序基因组大数据, 以这些数据为驱动, 将有可能极大促进个体化药物的研发.

深度学习算法一种非常适合于大数据分析的机器学习算法, 具有"抽象概念"处理能力. 使用深度学习算法, 有望改进以往药物设计与药物信息中已建立的多种机器学习模型; 在药物小分子结构信息处理上, 由于化学分子数量多、结构复杂, 使用传统的算法处理信息时能力常有不足, 而使用深度学习等算法有望改变这一局面, 促进化学信息学的发展. 另外, 大数据分析方法对于组学和系统生物学等复杂数据具有较强的分析能力, 有望促进基于系统的药物设计和药物信息研究的发展, 如药物靶标鉴定和关键靶标的选择和组合等. 以中药信息研究为例, 中药的药理学和毒理学研究的是一个复杂问题, 包括中药的复方、药材、分子成分和含量、分子代谢、对应症、中药分子和靶标之间的复杂的相互作用等, 以上因素之间存在多重关联关系, 这些复杂的动态和非线性特征提示深度学习等大数据分析方法可应用于上述领域.

参考文献

1 Pei J, Yin N, Ma X, et al. Systems biology brings new dimensions for structure-based drug design. J Am Chem Soc, 2014, 136: 11556–11565
2 Wild D J, Ding Y, Sheth A P, et al. Systems chemical biology and the semantic web: What they mean for the future of drug discovery research. Drug Discov Today, 2012, 17: 469–474
3 Samwald M, Jentzsch A, Bouton C, et al. Linked open drug data for pharmaceutical research and development. J Cheminf, 2011, 3: 1–6
4 Downing J, Harvey M, Morgan P, et al. Spectra-t: Machine-based data extraction and semantic searching of chemistry e-theses. J Chem Inf Model, 2010, 50: 251–261
5 Pospisil P, Iyer L, Adelstein S, et al. A combined approach to data mining of textual and structured data to identify cancer-related targets. BMC Bioinformatics, 2006, 7: 1–11
6 Percha B, Garten Y, Altman R. Discovery and explanation of drug-drug interactions via text mining. Pac Symp Biocompu, 2012, 410–421
7 Zhu Q, Sun Y, Challa S, et al. Semantic inference using chemogenomics data for drug discovery. BMC Bioinformatics, 2011, 12: 1–12
8 Willighagen E, Alvarsson J, Andersson A, et al. Linking the resource description framework to cheminformatics and proteochemometrics. J Biomed Semant, 2011, 2: S6
9 Hohman M, Gregory K, Chibale K, et al. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. Drug Discov Today, 2009, 14: 261–270
10 Kimeldorf G, Wahba G. Some results on tchebycheffian spline functions. J Math Anal Appl, 1971, 33: 82–95
11 Perry A S, Loftus B, Moroose R, et al. In silico mining identifies igfbp3 as a novel target of methylation in prostate cancer. Br J Cancer, 2007, 96: 1587–1594
12 Liu X, Xu Y, Li S, et al. In silico target fishing: Addressing a “big data” problem by ligand-based similarity rankings with data fusion. J Cheminf, 2014, 6: 33
13 Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug-target interactions: A brief review. Brief Bioinf, 2014, 15: 734–747
14 Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics, 2007, 8: 1–11
15 Sugaya N, Ikeda K. Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinformatics, 2009, 10: 1–13
16 Xuan S, Wu Y, Chen X, et al. Prediction of bioactivity of HIV-1 integrase st inhibitors by multilinear regression analysis and support vector machine. Bioorg Med Chem Lett, 2013, 23: 1648–1655
17 Heikamp K, Bajorath J. Support vector machines for drug discovery. Expert Opin Drug Discov, 2014, 9: 93–104
18 Byvatov E, Fechner U, Sadowski J, et al. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Inf Comput Sci, 2003, 43: 1882–1889
19 Worachartcheewan A, Nantasenamat C, Naenna T, et al. Modeling the activity of furin inhibitors using artificial neural network. Eur J Med Chem, 2009, 44: 1664–1673
20 Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure-activity relationship (QSAR) analysis. J Med Chem, 1990, 33: 2583–2590
21 Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules. J Chem Inf Model, 2013, 53: 1563–1575
22 Varnek A, Baskin I. Machine learning methods for property prediction in chemoinformatics: Quo vadis? J Chem Inf Model, 2012, 52: 1413–1437
23 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
24 Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator. arXiv: 1411.4555v1, 2014

Drug design and drug information in the big data era

SUN TanLin[1] & PEI JianFeng[1,2]

[1] *Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China;*
[2] *State Key Laboratory of Natural and Biomimetic Drugs, Peking University, Beijing 100191, China*

Big data brings new ways of data-driven drug research and development, which are expected to prominently increase the success rate, shorten the time and reduce the cost of drug discovery. In this review, we briefly summarized the development of data integration and data mining in drug design and drug information in recent years. Semantic web with techniques such as resource description framework, web ontology language, simple protocol and RDF query language, and linked data were commonly used for drug related data integration. Machine learning methods such as support vector machine and artificial neural network were widely used for data mining in drug information processing. We also give an outlook on future data integration and data mining for drug design and drug information. The rapidly booming big data integration and processing techniques can be adopted and a new machine learning method, deep learning, is specially recommended for data mining in the field of drug design and drug information.

big data, drug design, drug information, semantic web, deep learning

doi: 10.1360/N972014-01114