科学通报 2015 年 第 60 卷 第 5-6 期：558 ~ 565
专辑：大数据研究与应用 评 述 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

药物分子设计中的大数据问题

严鑫[①], 丁鹏[①], 刘志红[①], 王领[①], 廖晨钟[②], 顾琼[①], 徐峻[①②*]

① 中山大学药学院药物分子设计与生物超算中心, 广州 510006;
② 合肥工业大学医学工程学院, 合肥 230009
* 联系人, E-mail: junxu@biochemomes.com

2014-11-11 收稿, 2014-12-01 接受, 2015-01-09 网络版发表
国家自然科学基金(81173470)、国家高技术研究发展计划(2012AA020307)、广东省引进创新科研团队专项计划(2009010058)、广州超级计算应用研发与扶持专项(2012Y2-00048)和中央高校基本科研业务费专项(2013HGCH0015)资助

摘要 药物创新领域的大数据主要来源于高通量实验、高效能模拟计算、信息化、科技出版物和专利文献4个方面. 这些大数据使我们有可能在系统层面上看到药物分子与许多靶标相互作用的新现象、新规律, 提高药物创新的效率, 也带来新的挑战, 如存储、标引/标注和质控、可视化、数据挖掘和计算复杂度等问题. 这些问题可以通过在超算和云服务技术的支持下发展并行计算方法而逐渐得到解决. 从离散、不完备且信噪比低的大数据中难以找到物质活性与结构之间的连续函数关系, 贝叶斯学习机及其与支持向量机、决策树技术的组合是大数据挖掘的发展方向. 大数据既是科学实验通量化和社会信息化的结果又是原因, 正确解决大数据挖掘问题是提高药物创新效率的核心.

关键词
大数据
药物设计
生物信息学
化学信息学
高性能计算

1 用于生物医药研究的大数据来源

大数据(big data)近年来引起生物医药研究人员的广泛关注[1], 大数据的特点已有很多综述和讨论[2~4]. 药物分子设计领域涉及的大数据因为与生物大分子及小分子结构与性质相关, 其来源、数据挖掘的需求都有自己的特殊性, 需要特殊技术进行处理.

生物医药创新活动涉及设计、制备/提取、筛选/测试各种生物分子材料. 为了设计、制备和测试药物分子, 需要采集和分析各种相关数据, 包括来自仪器设备、模拟计算的数据, 科技出版物和医药卫生服务信息化所带来的数据.

1.1 高通量科学实验产生的大数据

20世纪下半叶人类的3个发明: DNA的体外扩增技术(PCR, 聚合酶链反应)[5,6]、高通量分子制造技术(combinatorial chemistry, CC, 即组合化学)[7]、高通量筛选技术(high throughput screening, HTS)[8], 在生命科学领域引发了以“高通量”为主要特征的科技革命, 它对药物创新的影响主要表现在下述几个方面.

(ⅰ) 在靶标研究方面. DNA的体外扩增技术使人类基因组计划提前完成, 触发了各种组学研究, 据报道, 以“组学”冠名的各种研究已达3000多种[9], 每种组学研究都涉及大量的数据. 例如, 基因组(genome)依赖于DNA 测序技术; 转录组(transcriptome)依赖于微芯技术; 蛋白组(proteome)和代谢组(metabolome)依赖于高分辨率质谱技术; 表型组(phenome)依赖于细胞生物学技术等. 因此, 关于药物靶标方面的数据发生了爆炸式的增长, 我们需考虑的问题是, 如何利用此类大数据加快药物创新的进程?

(ⅱ) 化合物创新方面. 高通量化学合成技术使大规模地探索化学结构多样性与材料性质的关系成为可能, 新物质制造能力空前提高. “点击化学”(click chemistry, 可视为第二波的组合化学)将使人类根据性能要求设计组装小分子的能力更上一层楼[10]. 大

引用格式: 严鑫, 丁鹏, 刘志红, 等. 药物分子设计中的大数据问题. 科学通报, 2015, 60: 558–565
Yan X, Ding P, Liu Z H, et al. Big data in drug design (in Chinese). Chin Sci Bull, 2015, 60: 558–565, doi: 10.1360/N972014-01144

规模的小分子设计与制造产生大量的数据，如：红外、紫外、核磁共振、质谱、色谱、晶体结构等实验数据. 关于物质结构的实验数据分辨率越来越高、维数也越来越高，占有的存储空间越来越大，分析方法越来越复杂，涉及的数据格式越来越多. 在制造新物质的过程中，需要大量的、品种繁多的生物和化学试剂. 这些生物化学制剂对存储条件有极高的要求，用量少，价格贵(从每毫克几百元到几万元)，为了避免造成物资积压和浪费，需要智能仓储系统动态地管理这些资源.

(ⅲ) 小分子性能测试方面. 高通量筛选技术使人类可以在短期内测试百万种分子的各种性质. 刚获得2014年诺贝尔化学奖的超高分辨率显微镜技术[11]使人类超越了光学显微成像极限，在0.2 μm以下的尺度追踪单个分子在生物系统中的行为. 高内涵筛选技术(high content screening, HCS)[12]使我们能够同时观察到细胞在外界分子作用下的各种行为的改变，产生了巨量的基于芯片的测试数据和图像数据，对这些数据的正确处理决定了药物发现的成败.

1.2 高效能计算模拟科学实验产生的大数据

近年来，高效能计算(high performance computing, HPC)成为国际上技术竞争的制高点之一. 我国在2011年(天河1号)和2013年(天河2号)分别在这个制高点上登顶[13]. 2013年的诺贝尔化学奖授予3名计算化学家，表彰他们在发展复杂化学系统的多尺度模型方面的杰出贡献. 2013年10月，*Science*以封面文章报道Atul Butte仅凭计算方法发现了抗代谢疾病和抗肺癌新药[14]. 通过理论计算，可能探索更大的化学多样性空间. 2012年，瑞士的Reymond课题组[15]枚举出含有1660亿个有机小分子的化合物库GDB-17. 这些数据极大地拓宽了药物筛选的化学空间，为发现新的药物化学骨架提供新的机会.

计算机辅助药物设计(computed-aided drug discovery, CADD)包括分子动力学(molecular dynamics, MD)模拟已经成为当代药物创新的主要工具之一[16]. CADD与HPC的结合使高通量、高命中率的虚拟筛选(victual screening, VS)成为可能. MD模拟实验过程中，平均产生约2 GB/ns以上的数据，如果要模拟微秒时间范围的生物大分子与小分子相互作用的动力学行为，将产生约2 TB(1 TB=1024 GB)数据，约100万帧生物大分子构象. 如果用基于MD模拟的药物虚拟筛选[17]，将会产生更大的数据.

1.3 科技出版物、专利文献和医药卫生服务信息化产生的大数据

(ⅰ) 科技出版物方面. 截止2014年10月26日，PubMed收录的与生物医药有关的科技文章超过2400万篇；美国化学文摘社收录9000多万条小分子化合物数据(包括化学结构、预测的或者测定的性质及其谱图数据)、7510万化学反应和合成制备方法、超过6578万个生物大分子序列、来自880多家制造商的6700多万个化合物产品、来自全球专利文献的105万多个Markush通式结构.

(ⅱ) 专利文献方面. 世界知识产权组织(WIPO)的数据库PatetnScope收录1400多万专利文献和21多万公开的PCT申请，但是其中的化学结构均不能检索. 德温特世界专利索引(Derwent World Patents Index, DWPI)收录4250万专利文献涵盖47个专利授权机构，每2周更新1次，1992年以来授权的化学专利涉及的化学结构式都可以检索[18]. PubMed数据库收集超过1970年以来的医学专利文献，涵盖100多个国家，涉及13700多个专家.

(ⅲ) 医药卫生服务信息化方面. 全球医药卫生服务的持续信息化将产生史无前例的大数据. 我国有13.3亿(2010年第6次人口普查数据)人口，随着中国采用与国际标准兼容的医药卫生服务代码体系(International Statistical Classification of Diseases and Related Health Problems, ICD-9)[19]，居民接受医药卫生服务的数据(病历数据)将会成为世界上最大的生物大数据来源. 除了涉及个人隐私的数据之外，文字性数据(含有病史、治疗史、家族史、卫生习惯、疗效、不良反应等的数据)和诊断性数据(含有关于器官或体液的测定及影像、基因测序[20]等数据)对药物创新研究极为重要.

2 大数据给药物分子设计带来的机遇和挑战

除了具有一般意义上的大数据所涵盖的特征(即量大、类型多、噪声高、增长速度快和处理等待时间短等)[21]以外，用于药物分子设计的大数据因为与生物大分子及小分子结构、性质及其相互作用密切相关，它具有不同于其他领域大数据所具备的特点——离散数学特点. 分子是有结构的，分子结构的数学描

述是一张拓扑图(topological graph), 图上的节点是原子, 边是化学键. 生物大分子是具有重复单元的结构图(结构子图), 为了便于存储和运算, 生物大分子表达成用子图串联而成的图, 表示分子图最常用的数据结构是连接表[22]. 分子因为有可旋转的化学键而具有柔性, 结果造成了一个分子在不同条件下有非常多的三维空间形状, 称为构象(conformation). 药物分子设计的关键点之一就是要确定配体/药物(ligand/drug)与靶标(target)互相作用时所采取的构象(称为活性构象). 而配体或靶标都可能有成千上万种构象, 而它们的复合物所需的"活性构象"就如沧海一粟那样难以找寻.

在药物分子设计研究中, 用化学信息学处理小分子数据(如研究结构与活性的关系), 用生物信息学处理生物大分子数据(如生物大分子的序列比对)[23]. 药物创新研究中出现的大数据问题是超算和高通量实体实验技术发展的必然结果, 也带来新的机遇[24]: (1) 高通量实体实验产生的大数据使我们有可能在系统层面上看到药物分子与许多靶标相互作用的新现象、新规律; (2) 超算能力使并行地针对多靶标进行千万级数量小分子的虚拟筛选成为可能[25]; (3) 超算能力还使在更长的时间尺度(0.01~10 μs)上模拟药物分子与靶标结合的动力学行为成为可能[26]; (4) 通过对上述虚拟和实体实验产生的大数据进行挖掘, 将提高对化合物和靶标活性构象预测的准确性、开发针对特定人群的特异性药物.

大数据带来的挑战主要有存储与分享、获取与标注、检索、数据格式异质化以及可视化与数据挖掘问题.

(ⅰ) 存储与分享问题(storage and sharing). 因为数据量大(以TB计), 在本地服务器上存储和分享就不现实了, 一般用"云服务"来解决这个问题, 不过, 药物分子的知识产权容易受到侵犯, 科研机构对此持高度保留态度.

(ⅱ) 获取与标注(capture and curation)问题. 传统的生物医药数据获取和标注需要人工操作. 面对大数据, 人工操作几乎不可能. 而自动化的获取与标注技术精准度差, 质量控制成为大问题. 解决这个问题的第一步就是要建立合理的质量控制标准[27], 令人欣慰的是一些研究组正在努力建立这些标准[28]. 对应基因标注、专利标引、活性位点预测、受体-配体结合模式等生物大数据标注问题需要建立标准的流程, 许多相关的算法还有待于开发.

(ⅲ) 检索(search)问题. 与分子有关的大数据检索引擎分为拓扑检索(topological search)[29,30]、相似度检索(similarity search)[31]和语义检索(semantic search)[32] 3类. 对大分子序列的检索需要运用序列比对算法(sequence alignment algorithms)技术[33]. 小分子的拓扑结构检索又分为结构检索、子结构检索和超结构检索(Markush结构检索)[34]. 结构和子结构检索问题虽然仍然属于NP-完全性问题[35], 但是有许多方法可以降低其计算复杂度. 大部分有机分子的节点度数小于5, 加上可以用各种筛法尽可能地减少"原子对原子"(atom-by-atom)的匹配调用, 在实际应用中, 算法的效率还是很高的. Markush结构检索问题因为其通式表达的不确定性和递归性, 在大数据时代, 它的问题可能变得更加难解. 结构的相似度检索问题因为要给每一对分子算出相似度值, 不能采用筛法加速. 这个问题可以通过并行化算法来解决. 语义检索主要用于科技文献的全文检索, 在药物分子设计领域用来自动化地标引靶标、配体的生物学功能. 面向大数据的语义检索算法因为涉及大量统计学计算, 必须获得高效能计算的支持.

(ⅳ) 数据格式异质化(heterogeneous data)问题. 大数据带来的分子数据格式异质化问题主要表现在分子的结构图以许多不同格式存在, 例如, 一个分子可以有许多不同类型的连接表存储在格式文件中[36](如: SDF, MOE, MOL2, PDB等)、或嵌入在其他图像文件里(如: JPG, PDF, DOCX, PPTX等)、或以线性编码[37](如: SMILES[38], InChI Keys[39], CAS登记号, IUPAC系统命名法、商品名、俗名等)的形式嵌入在一段文章里. 这要求分子结构检索引擎能够自动识别分子结构数据的存在状态, 并能够正确地译成分子结构连接表以完成检索任务.

(ⅴ) 可视化与数据挖掘(visualization and data mining)问题. 很多数据挖掘问题可归结为分类. 药物分子设计方法学的任务是找到一种模式将分子多样性空间划分成有活性的和无活性的2大类, 从而降低制造和测试分子实体的成本, 提高药物创新的效率. 数据的分类往往从数据的可视化开始. 大规模的化合物库(compound library)数据以连接表的形式存储, 只有通过数学变换才能够被可视化. 一般过程是: 将库中的每个分子连接表变换为一组结构描述符(理想的结构描述符组, 其成员之间彼此不相关,

而每个成员都与要考虑的分子性质高度相关). 如果用*n*个描述符表示1个分子, 则库中的每1个化合物被表示成*n*-维广义空间的1个点. 采用广义空间降维技术[40], 如主成分分析(principal component analysis, PCA)[41]、多维标度变换(multidimensional scaling, MDS)[42]或自组织图(self-organization map)[43]. 药物设计领域常采用的高维数据分类方法主要有簇分析(clustering)[44]、机器学习(machine learning)[45]、决策树(decision tree)[46], 贝叶斯方法(Bayesian learning)及它们的组合[47]等. 分类的数学本质是将含有*m*个成员的集合里分成*n*个子集合的问题, 如果*n*已知, 则为有监督的学习(supervised learning), 否则为无监督的学习. 当集合是大数据时, 这是个严重的组合爆炸的问题, 会因为计算复杂度(computing complexity)太高而无法在合理时间内给出结果. 在分类算法中, 还涉及计算2个分子图的相似度[48]或广义距离问题, 这更增加了分类问题的计算难度.

3 药物分子设计需要的大数据处理新工具

近年来, 药物分子设计的大数据处理技术已经有了许多进展. 一批传统的药物设计程序有了并行算法的版本从而与现代超算技术成就同步. 2011年, Collignon等人[25]发布了并行版的Autodock4.lga.MPI, 该版本利用消息传递接口(message passing interface, MPI)实现分子对接的并行计算, 使Autodock 4 程序能在超算服务器上通过调用上千个(最多可达8192个)CPU同时将许多化合物对接到靶标分子的活性位点. 对接考虑了范德华、静电和溶剂化作用. 在24 h内, 调用8192 高效能CPU, 该程序可以将30万柔性化合物或1100万刚性化合物与1个刚性蛋白对接.

对药物分子的虚拟筛选, 仅仅实现分子对接是不够的. 配体作用于受体是分子识别的过程, 它是动态的过程. 在溶液环境, 配体向药物靶标接近, 彼此都要改变自己的构象才能互相识别、互相适配, 需要用MD模拟来预测受体-配体复合物的动力学行为[49]. 然而, MD模拟实验计算量大、代价高. 近年来, GPU技术的发展大大地降低了MD并行计算的代价. 所以, 很多MD模拟程序[50~53](如AMBER, NAMD, GROMACS)都有了GPU版本.

在天河-2号超级计算机的支持下, 中山大学药物分子设计研究中心的超算团队将分子动力学模拟技术应用于药物分子虚拟筛选研究中(图1)[17].

MDVS的分子动力学模拟实验用到的初始构象常常来自实验数据(来自X射线衍射实验的晶体结构数据, 或来自多维核磁共振实验的数据). 这些实验数据代表不同的配体-受体结合模式. 为了提高虚拟筛选的准确性, 各种代表性的配体-受体结合模式都应该作为MDVS的初始构象[54]. 正则模式分析(all-atom normal mode analysis, NMA)是预测受体分子构象和自由能的另一种方法, 采用GPU的并行处理技术对此方法实现了加速[55], 这项工作对进一步提高基于结构的虚拟筛选的效率很有意义. 在此基础上, 绘制配体从各个方向接近受体所产生的反应势能面的自由能全景图(free energy landscape, FEL)[56]是评价受体-配体结合难易程度和复合物稳定性的主要判据[57], 随着超算性价比的提高, 自由能全景分析将成

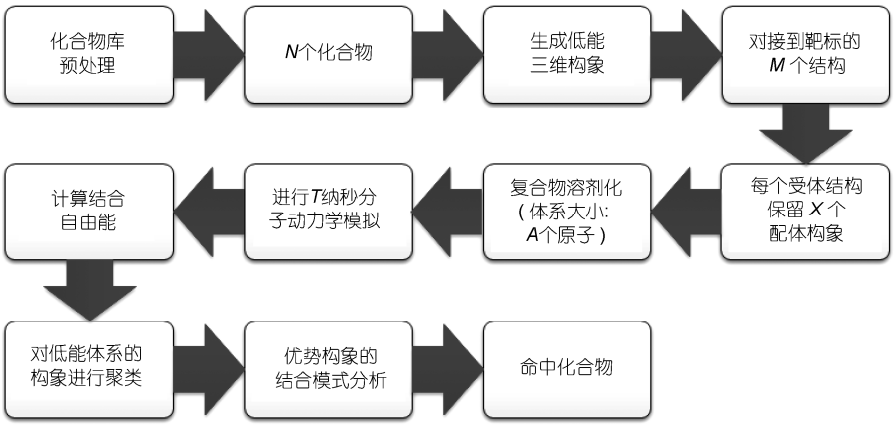

图 1 超算支持的基于分子动力学模拟的虚拟筛选(MDVS)流程

Figure 1 Flowchart of HPC-supported molecular dynamics simulation based virtual screening (MDVS)

为虚拟筛选新工具的组成部分.

基于配体的虚拟筛选也可以通过计算化合物与活性配体的三维结构相似度来实现[58]. 柔性小分子可以有百万以上的构象, 化合物与活性配体的三维构象叠合具有很高的计算复杂度, 通过GPU加速可以很好地提高基于配体的虚拟筛选的效率[59].

由于大数据的离散性、低信噪比和不完备性, 蕴含其中的物质活性与结构之间的关系不是传统的连续函数关系, 超算支持的贝叶斯学习方法在药物设计中得到了应用, 例如基于配体的虚拟筛选[60]、化合物库的化学稳定性预测[61], 用贝叶斯网络研究分子信号转导网络[62], 并且与其他机器学习技术相结合以解决生物医药领域出现的大数据问题, 例如, 与贝叶斯结合的支持向量机[63]和决策树技术[64].

4 展望

大数据的涌现是客观事实, 但数据不一定越大越好[65]. 我们的目标是从数据中提炼出有用的知识. 但是, 当数据变成大数据时, 可能对科学研究产生灾难性干扰: (1) 由于数据文件太大, 很多科学计算可能由于软件或硬件的限制而无法进行; (2) 大数据的信噪比低可能使传统的数据挖掘程序崩溃, 或因为计算复杂度问题而无法完成; (3) 大数据也可能仅仅因为内存不够、或通讯系统的带宽太窄、或读写系统的超高稳定性要求不能满足而导致无法重现的程序崩溃; (4) 大数据生产代价高昂、难以重复, 存在质量隐患, 这样的数据因为占用大量存储空间而维护成本极高, 却难以完成分析而带来效益, 拖累了科技项目的执行; (5) 大数据虽然可能让科技工作者在大系统层面发现新的规律, 但由于缺少数据挖掘工具、或不正确地使用数据挖掘工具而得到误导性结论, 这样的结论往往因为证伪困难而令人困扰.

大数据的这些问题可以采用下述措施而逐步解决: (1) 普及云计算技术. 大数据的存储和分享问题通过“云存储”服务得到解决. 云超算将满足大数据产生的超算需求; (2) 加强面向交叉学科的大数据挖掘技术的开发. 大数据与领域知识密切相关, 大数据的挖掘算法, 如知识富集、分类、可视化, 都必须与领域知识和经验结合起来; (3) 加强大数据处理人才的培养. 今天的数据处理和编程理念和技术与20世纪80年代的完全不同, 软件开发和使用与网络密不可分. 大数据处理人才除了要具备交叉学科的背景之外, 还应该具备数学建模、算法理论、图论知识, 积累主要编程语言(如, C/C++等程序设计语言)的编程技巧. 对药物设计领域而言, 基因序列及其存储格式知识、生物大分子序列比对算法原理、簇分析、分子建模、GMOD或Galaxy(用于基因标注、网络数据库建设的开源代码)以及主要化学信息学知识是必须的. 在操作系统方面, 应该具备Linux, Windows, Mac三大操作系统的经验; 应该熟悉Visual Studio, Eclipse, GCC/GDB这类软件开发环境, 掌握多线程、Cuda等并行程序设计工具, 积累client/server, Java, SQL等数据库开发工具的经验.

我国在传统的药物设计技术研究方面仍然比较落后, 大数据给了我们急起直追的机遇, 因为没有历史负担, 我们有可能从高起点出发, 直接开发面向大数据的新一代生物信息学和化学信息学的工具, 在药物分子设计方法学方面走出一条跨越式发展的道路.

参考文献

1 Costa F F. Big data in biomedicine. Drug Discov Today, 2014,19: 433–440
2 Shen B, Teschendorff A E, Zhi D, et al. Biomedical data integration, modeling, and simulation in the era of big data and translational medicine. Biomed Res Int, 2014, 2014: 731546
3 Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. Yearbook Med Informatics, 2014, 9: 14–20
4 Ndiaye N C. Systems medicine in the era of “big data”: A game-changer for personalized medicine? Drug Metab Drug Interact, 2014, 29: 127
5 Bartlett J S, Stirling D. A Short history of the polymerase chain reaction. In: Bartlett J S, Stirling D, eds. PCR Protocols. New York: Humana Press, 2003. 3–6
6 Mullis K B, Ferré F, Gibbs R A. The Polymerase Chain Reaction. New York: Birkhauser Boston Inc., 1994
7 Merrifield RB. Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. J Am Chem Soc, 1963, 85: 2149–2154
8 Pereira D A, Williams J A. Origin and evolution of high throughput screening. Br J Pharmacol, 2007, 152: 53–61
9 Baker M. Big biology: The ’omes puzzle. Nature, 2013, 494: 416–419

10 Massarotti A, Brunco A, Sorba G, et al. ZINClick: A database of 16 million novel, patentable, and readily synthesizable 1,4-disubstituted triazoles. J Chem Inf Model, 2014, 54: 396–406
11 Clery D. Light loophole wins laurels. Science, 2014, 346: 290–291
12 Giuliano K A, Haskins J R, Taylor D L. Advances in high content screening for drug discovery. Assay Drug Develop Technol, 2003, 1: 565–577
13 Zhang X, Yang C, Liu F, et al. Optimizing and scaling HPCG on tianhe-2: Early experience. In: Sun X H, Qu W, Stojmenovic I, et al., eds. Algorithms and Architectures for Parallel Processing. 14th International Conference (ICA3PP 2014). Cham, Switzerland: Springer International Publishing, 2014. 28–41
14 Service RF. Biology's dry future. Science, 2013, 342: 186–189
15 Ruddigkeit L, van Deursen R, Blum L C, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model, 2012, 52: 2864–2875
16 Ieong P U, Sørensen J, Vemu P L, et al. Progress towards automated kepler scientific workflows for computer-aided drug discovery and molecular simulations. Proc Comput Sci, 2014, 29: 1745–1755
17 Ge H, Wang Y, Li C, et al. Molecular dynamics-based virtual screening: Accelerating the drug discovery process by high-performance computing. J Chem Inf Model, 2013, 53: 2757–2764
18 White M J. Chemical patents. In: Currano J, Roth D, eds. Chemical Information for Chemists: A Primer. Cambridge: Royal Society of Chemistry, 2013. 53
19 World Health Organization. International Statistical Classification of Diseases and Related Health Problems Tenth Revision (ICD-10), 2007
20 Genovese G, Handsaker R E, Li H, et al. Using population admixture to help complete maps of the human genome. Nat Genet, 2013, 45: 406–414
21 Feinleib D. The Big Data Landscape. Big Data Bootcamp. New York: Apress, 2014. 15–34
22 Xu J. GMA: A generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. J Chem Inf Comput Sci, 1996, 36: 25–34
23 Degtyarenko K, Hastings J, Matos P, et al. ChEBI: An open bioinformatics and cheminformatics resource. Curr Protoc Bioinf, 2009, 14: 14.9
24 Marx V. Biology: The big challenges of big data. Nature, 2013, 498: 255–260
25 Collignon B, Schulz R, Smith J C, et al. Task-parallel message passing interface implementation of Autodock4 for docking of very large databases of compounds using high-performance super-computers. J Comput Chem, 2011, 32: 1202–1209
26 Shaw D E, Maragakis P, Lindorff-Larsen K, et al. Atomic-level characterization of the structural dynamics of proteins. Science, 2010, 330: 341–346
27 Rutherford K M, Harris M A, Lock A, et al. Canto: An online tool for community literature curation. Bioinformatics, 2014, 30: 1791–1792
28 Shi L, Campbell G, Jones W D, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol, 2010, 28: 827–838
29 Xu J. Two-dimensional structure and substructure searching. In: Gasteiger J, ed. Handbook of Chemoinformatics. Weinheim: Wiley-VCH Verlag GmbH, 2008. 868–884
30 Barnard J M. Substructure searching methods: Old and new. J Chem Inf Comput Sci, 1993, 33: 532–538
31 Zhang L, Zhang Y, Gu X, et al. Scalable similarity search with topology preserving hashing. IEEE Transact Image Proc, 2014, 23: 3025–3039
32 Bontcheva K, Tablan V, Cunningham H. Semantic search over documents and ontologies. In: Ferro N, ed. Bridging Between Information Retrieval and Databases. Berlin: Springer-Verlag, 2014. 31–53
33 Pearson W. BLAST and FASTA similarity searching for multiple sequence alignment. In: Russell D J, ed. Multiple Sequence Alignment Methods. New York: Humana Press, 2014. 75–101
34 Geyer P. Markush structure searching by information professionals in the chemical industry—Our views and expectations. World Patent Inf, 2013, 35:178–182
35 Gasarch W I. The P=?NP poll. SIGACT News, 2002, 33: 34–47
36 Smalter H A, Shan Y, Lushington G, et al. An overview of computational life science databases & exchange formats of relevance to chemical biology research. Comb Chem High Throughput Screen, 2013, 16: 189–198
37 Herndon W C, Bertz S H. Linear notations and molecular graph similarity. J Comput Chem, 1987, 8: 367–374
38 Warr W A. Representation of chemical structures. Wiley Interdiscip Rev Comput Mol Sci, 2011, 1: 557–579
39 Southan C. InChI in the wild: An assessment of InChI Key searching in Google. J Cheminf, 2013, 5: 10
40 Tenenbaum J B, Langford J C, Silva V D. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290: 2319–2323

41 Abdi H, Williams L J. Principal component analysis. Wiley Interdiscip Rev Comput Stat, 2010, 2: 433–459
42 Kruskal J B. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 1964, 29: 115–129
43 Kohonen T. Self-Organization And Associative Memory. 3rd ed. New York: Springer-Verlag, 1989
44 Jain A K, Murty M N, Flynn P J. Data clustering: A review. ACM Comput Surv, 1999, 31: 264–323
45 Warmuth M K, Liao J, Rätsch G, et al. Active learning with support vector machines in the drug discovery process. J Chem Inf Comput Sci, 2003, 43: 667–673
46 Cramer G, Ford R, Hall R. Estimation of toxic hazard—A decision tree approach. Food Cosmet Toxicol, 1976, 16: 255–276
47 Kohavi R. Scaling up the accuracy of naïve-bayes classifiers: A decision-tree hybrid. In: Simoudis E, Han J, Fayyad U, eds. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Menlo Park, CA: AAAI Press, 1996. 202–207
48 Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. Drug Discov Today, 2007, 12: 225–233
49 Durrant J D, McCammon J A. Molecular dynamics simulations and drug discovery. BMC Biol, 2011, 9: 71
50 Götz A W, Williamson M J, Xu D, et al. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. J Chem Theory Comput, 2012, 8: 1542–1555
51 Salomon-Ferrer R, Götz A W, Poole D, et al. Routine microsecond molecular dynamics simulations with Amber on GPUs. 2. Explicit solvent particle mesh Ewald. J Chem Theory Comput, 2013, 9: 3878–3888
52 Stone J E, Hardy D J, Ufimtsev I S, et al. GPU-accelerated molecular modeling coming of age. J Mol Grap, 2010, 29: 116–125
53 Suhartanto H, Yanuar A, Wibisono A. Performance analysis cluster and GPU computing environment on molecular dynamic simulation of BRV-1 and REM2 with GROMACS. Int J Comput Sci Issu, 2011, 8: 131–135
54 Wang L, Gu Q, Zheng X, et al. Discovery of new selective human aldose reductase inhibitors through virtual screening multiple binding pocket conformations. J Chem Inf Model, 2013, 53: 2409–2422
55 Liu L, Liu X, Gong J, et al. Accelerating all-atom normal mode analysis with graphics processing unit. J Chem Theory Comput, 2011, 7: 1595–1603
56 Li H, Xie Y, Liu C, et al. Physicochemical bases for protein folding, dynamics, and protein-ligand binding. Sci China Life Sci, 2014, 57: 287–302
57 Li C, Ge H, Cui L, et al. Molecular mechanism of action of K(D)PT as an IL-1RI antagonist for the treatment of rhinitis. RSC Adv, 2014, 4: 48741–48749
58 Yan X, Li J, Liu Z, et al. Enhancing molecular shape comparison by weighted Gaussian functions. J Chem Inf Model, 2013, 53: 1967–1978
59 Yan X, Li J, Gu Q, et al. gWEGA: GPU-accelerated WEGA for molecular superposition and shape comparison. J Comput Chem, 2014, 35: 1122–1130
60 Zheng M, Liu Z, Yan X, et al. LBVS: An online platform for ligand-based virtual screening using publicly accessible databases. Mol Divers, 2014, 18: 829–840
61 Liu Z, Zheng M, Yan X, et al. ChemStable: A web server for rule-embedded naïve Bayesian learning approach to predict compound stability. J Comput Aided Mol Des, 2014, 28: 941–950
62 Asadi N B. High performance reconfigurable computing for learning bayesian networks with flexible parametrization. Doctor Dissertation. Palo Alto: Stanford University, 2010
63 Fang J, Yang R, Gao L, et al. Predictions of BuChE inhibitors using support vector machine and naive bayesian classification techniques in drug discovery. J Chem Inf Model, 2013, 53: 3009–3020
64 Wang L, Chen L, Liu Z, et al. Predicting mTOR Inhibitors with a classifier using recursive partitioning and naïve bayesian approaches. PLoS One, 2014, 9: e95221
65 Ekins S, Freundlich J S, Reynolds R C. Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for mycobacterium tuberculosis. J Chem Inf Model, 2014, 54: 2157–2165

Big data in drug design

YAN Xin[1], DING Peng[1], LIU ZhiHong[1], WANG Ling[1], LIAO ChenZhong[2], GU Qiong[1] & XU Jun[1,2]

[1] *Research Center for Drug Discovery & HPC for Life Sciences, School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China;*
[2] *School of Medical Engineering, Hefei University of Technology, Hefei 230009, China*

Big data collection in the pharmaceutical research industry has four sources, high-throughput scientific experiments, high-performance computations, automated information acquisition and office automation, and scientific publications and patents. Big data is the product and the promoter of high-performance scientific experiments, therefore the technology for mining big data is the key to future drug discovery. However, big data brings greater challenges, such as, storage, retrieval, curation and quality assurance, sharing/transfer, analysis, visualization, modeling and computing complexities. This review outlines the current progress of processing big data in the drug design field. These problems may be resolved by adopting cloud computing and high-performance computing technologies, and parallelizing existing chemoinformatics and bioinformatics programs. Machine-learning approaches involving Bayesian learning methods and other methods, such as support vector machine and recursive petitioning, can be used for big data mining. Recent progress includes parallelized and GPU-accelerated molecular dynamics simulation technology, enhanced molecular docking technology, new parallelized algorithms for shape-based virtual screening, free-energy landscape calculations, and machine-learning algorithms for big chemical structural data mining. Big data from drug discoveries will increase, so conventional drug design software and methods need to be upgraded. This is a long-term project and we highlight the tasks that need to be accomplished to meet this goal.

big data, drug design, bioinformatics, chemoinformatics, high performance computing

doi: 10.1360/N972014-01144