药物研发领域中的开放访问数据库

刘志海, 李婕, 韩莉, 李嫣, 刘杰, 聂伟, 刘宇晨, 王任小*

生命有机化学国家重点实验室; 中国科学院上海有机化学研究所, 上海 200032

*通讯作者, E-mail: wangrx@mail.sioc.ac.cn

收稿日期: 2015-01-05; 接受日期: 2015-02-01; 网络版发表日期: 2015-03-27

doi: 10.1360/N032015-00001

摘要 开放访问数据库的发展极大推动了各领域的科学研究. 本文以药物研发的几个代表性阶段为主线, 介绍了该领域涉及的一批主流开放数据库, 旨在为读者描绘药物研发领域值得关注的信息资源. 另外, 本文介绍了本课题组开发的 PDBbind-CN 数据库的构建过程和基本内容, 这是一个致力于收集生物大分子复合物结构和亲合性数据的开放访问数据库, 在药物分子设计领域中得到了国内外同行的广泛应用.

关键词

药物研发
开放访问数据库
PDBbind-CN
亲合性数据

1 引言

信息技术和互联网的飞速发展提高了人们获取知识的能力, 显著提高了工作效率. 经过几十年的发展和积累, 互联网中的信息量呈爆炸式增长, 在此过程中出现了一大批可以通过网络访问的、服务于自然科学研究的高质量数据库. 科学数据库存在“开放访问”和“免费访问”两种常见服务模式. 开放访问通常意味着可以在公共互联网上面免费获取, 允许任何用户读取、下载、复制、打印、搜索或者建立指向原始资源的链接, 抓取进行索引, 作为数据传递给软件或者用于任何其他法律允许的范畴. 而免费访问的概念主要体现在移除了访问资源时的价格障碍, 是开放访问模式的内涵子集^[1]. 开放访问和免费访问两种模式都极大促进了科学的发展进程, 为人类知识、信息和数据的共享繁荣做出了重要贡献. 免费访问模式已经能够满足大多数人的需求, 因此本文所提及的开放访问概念主要内涵为免费访问, 不涉及特别的额外权限(如商业方面的用途和重新发行内容等).

化学和生物学都是较早应用大规模网络数据库的学科. 例如, 早在 20 世纪 70 年代, Margaret Day-

hoff 教授在蛋白进化研究领域中建立了 Protein Sequence Database 数据库(现名 International Protein Information Resource, PIR); 在 20 世纪 80 年代早期, Amos Bairoch 教授开发了 Swiss-Prot 蛋白序列数据库, 因而促进了 GenBank、EMBL 和 DDBJ 世界三大核酸序列数据库的发展融合, 并以此应对全世界范围快速增长的基因序列数据^[2]. 在生物学领域, 数据库的免费开放传统一直较好地被传承. 相比较而言, 化学领域中与化合物相关的数据库大多与商业利益挂钩, 因此早期的化学信息数据库大多采用付费模式, 公众无法免费使用和访问. 值得欣慰的是, 目前化学类数据库的开放程度得到了很大改善. 近 20 年来, 自然科学研究越加明显地呈现出“大科学”的趋势^[3], 各种学科知识相互交叉渗透, 形成了很多前沿交叉学科. 例如, 随着生物学研究深入到分子水平, 其采用的研究手段越来越多地借助化学分子探针和相关技术, 而传统的化学学科也越来越多地选择生物分子体系作为研究对象. 这些新变化催生了“化学生物学”这样的交叉学科. 化学生物学以化学小分子为工具, 通过干扰/调节正常的生命过程来了解蛋白功能, 或者解决相关的生物学问题. 所使用的手段包括高通量及高灵敏度的生物筛选、计算生物学、组合

化学、生物芯片等^[4,5]。对于这样的交叉学科,充分利用互联网访问的便利、提供丰富链接的综合性数据库必然大有用武之地。

药物研发是化学生物学基础研究的主要应用领域。现代药物研发的完整流程包括有效靶标的确认、先导化合物的发现和优化、临床前试验和临床试验等诸多环节,整个研发周期非常长。由于人们对药物的安全性和疗效要求越来越高,导致药物开发的成功率相当低。据统计,平均每研究 10000 个化合物,才能从中筛选出 1 个可能成功上市的药物^[6,7]。根据 2010 年的统计数据,西方发达国家平均每个上市新药研发耗时约 13 年,研发投入约 18 亿美元^[8]。2014 年的最新统计数据显示,每个上市新药的整体投入达到惊人的 26 亿美元^[9]。因此,尽管药物成功上市之后的回报丰厚,但是药物研发仍然是高投入、高风险的行业。提高药物研发的成功率是解决目前新药研发困境的根本出路。为了做到这一点,一方面需要在药物研发过程的各个环节广泛采用新技术新方法;另一方面,也需要注意知识的积累,建立各种汇集大量实验结果的数据库,并进行知识的合理挖掘和分析,为在药物研发的各个阶段做出关键决策提供依据。

本文将按照药物研发的各个主要阶段,介绍所涉及的代表性开放访问网络数据库,为药物研发领域科研工作者描绘实用的免费资源图谱。本文最后一部分将特别介绍本课题组近年来发展的生物大分子复合物亲和性数据库 PDBbind-CN,作为药物研发领域开放访问数据库的一个范例。

2 药物研发领域中的开放访问数据库

根据其在药物研发各个阶段中所起的作用,数据库大致可以分为以下几类: (1) 提供基因组、核酸或蛋白序列信息的数据库,如 GenBank、EMBL 和 DDBJ 等; (2) 提供生物大分子结构与分类信息的数据库,如 PDB, PDBsum, PDBREPORT, SCOP 和 CATH 等; (3) 提供生物大分子结构信息和分析工具的数据库,如 Relibase、PDBeMotif 和 Ligand-Expo 等; (4) 提供活性化合物信息及其生物活性测试结果的数据,如 PubChem, ChEMBL, ZINC, BindingDB 和 PDBbind-CN 等; (5) 提供化合物药代动力学、代谢性质和毒性数据的数据库,如 DrugBank, HMDB,

KEGG 和 admetSAR 等。

在药物研发流程的第一个阶段,研究人员往往需要分析与疾病相关的基因组信息,了解涉及的核酸序列和蛋白序列信息,由此确认与疾病相关的靶标。国际核苷酸序列数据库合作组织(the International Nucleotide Sequence Database Collaboration)由 3 个数据库组成,分别为日本的 DNA Data Bank of Japan (DDBJ)数据库^[10]、欧洲分子生物学实验室的 EMBL ENA 数据库^[11]以及美国 NCBI 的 GenBank 数据库^[12]。这 3 个数据库每天进行数据交换,提供所有公共可得 DNA 序列的注释信息。The Universal Protein Resource (UniProt)^[13]是一个提供蛋白序列和注释信息的综合性数据库,由多家单位共同维护。其系列数据库组成部分中的 UniProt Knowledgebase (UniProtKB)最为人所知,其包含一个人工注释和核查的 Swiss-Prot 数据库和一个自动注释而未核查的 TrEMBL 数据库(Translated EMBL Nucleotide Sequence Data Library)。创建 TrEMBL 数据库的原因是待处理序列数据的增长速度远大于 Swiss-Prot 的处理能力。来自印度的 HPRD 数据库(the Human Protein Reference Database)^[14]可视化地描绘和整合了人类基因组中的功能结构域、后翻译修饰、相互作用网络及每个蛋白与疾病的相关信息。HPRD 数据库中的所有信息均由人工从科学文献中抽取和校对。PRINTS 数据库^[15]主要提供蛋白指纹,指纹被定义为一组描述蛋白家族的保守性 motif,指纹相比单个 motif 可以更好地编码蛋白折叠和功能信息,通过迭代扫描 Swiss-Prot/TrEMBL 数据库能增强其诊断能力。本节相关数据库信息见表 1。

在药物研发流程的第二个阶段,研究人员通常需要了解靶标分子的三维结构和功能。wwPDB 数据库(the Worldwide Protein Data Bank)^[16]包含 PDB 数据存储、数据处理、发布中心等组织。成员包括美国的 RCSB PDB、欧洲的 PDBe、日本的 PDBj 和美国的 BMRB。wwPDB 数据库的使命是为全球科研团体维护一个统一的免费开放大分子结构 PDB 数据库。wwPDB 数据库中包含蛋白质、核酸类的大分子结构以及及各种小分子形成的复合物等,其结构类型涵盖范围丰富,小到只包含若干残基的寡肽分子和寡聚核苷酸,大到核糖体这样复杂的分子机器。了解靶标的三维结构,对于研究靶标与药物或其他蛋白/核酸的相互作用有直接指导意义。由于 wwPDB 数据库

表 1 提供基因组、核酸或蛋白序列信息的数据库举例

数据库	内容和数据量	最后更新	官方网址
NCBI GenBank	1816 亿个核苷酸; 1.783 亿条核酸序列	2014/10	http://www.ncbi.nlm.nih.gov/Genbank/
EMBL ENA	10784 亿个核苷酸; 4.933 亿条核酸序列	2014/11	http://www.ebi.ac.uk/embl/ http://www.ebi.ac.uk/ena/
DDBJ	1660 亿个核苷酸; 1.744 亿条核酸序列	2014/09 Release 98	http://www.ddbj.nig.ac.jp/
UniProtKB (Swiss-Prot/TrEMBL/ PIR-PSD)	Swiss-Prot 包含 546790 条蛋白序列, 1.946 亿个氨基酸残基; TrEMBL 包含 86536393 条蛋白序列, 273.9 亿个氨基酸残基	2014/10	http://www.uniprot.org/ http://pir.georgetown.edu/
HPRD	Release 9 版本包含 30047 条蛋白记录, 41327 条蛋白-蛋白相互作用记录	2010/04	http://www.hprd.org/
PRINTS	Release 42.0 版本包含 2156 条记录, 编码 12444 个 single motifs	2012/02	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/

主要收集直接来源于实验的原始数据, 这些数据只经过简单的归类、整理和注释, 因此, 针对不同的研究需要, 现在有很多二级数据库对 PDB 数据库进行进一步检查、信息加工和分类整理. 为了检查 PDB 数据库中可能存在的错误, PDBREPORT 数据库^[17]通过程序自动为每个 PDB 结构产生结构问题报告, 目前其数据库中大约有 2000 万条诊断信息. 从 1995 年起, SCOP 数据库 (the Structural Classification of Proteins)^[18]开始为 PDB 数据库中的蛋白结构提供结构分类和注释信息. 经过近 20 年的发展, 他们对算法和架构进行了改进, 推出了后续的 SCOP2 版本, 与 SCOP 的简单树状分层架构不同, SCOP2 采用复杂的网络节点架构来描述蛋白之间的相互关系. CATH 数据库 (Class, Architecture, Topology, Homology)^[19]基于人工核查的方式, 辅助以各种分类和预测算法, 提供蛋白结构域的层状分类信息. 每一个蛋白结构都被仔细检查确认, 符合挑选标准之后才被拆分为组成链, 这些链可以被分为更多的功能域, 然后根据结构和功能进行分类. MMDB 数据库 (Molecular Modeling Database)^[20]提供来自 PDB 数据库中的蛋白和核酸结构, 以及一些附加的增值特征, 如显式的化学结构图形、通过计算识别出的三维功能域、文献链接、相似序列、与大分子结合的化合物信息等. PDBsum 数据库^[21]为 PDB 数据库中的结构提供了直观的图形化解析方式, PDB 结构中的组成部分如蛋白链、核酸、配体、金属原子等都被分别标示, 各种相互作用也被以图形化的方式列举出来. 本节相关数据库信息见表 2.

为了研究靶标与配体分子相互作用的细节, 一系列开放数据库提供了分析蛋白结合口袋、界面性

质、与小分子的相互作用等细节的工具 (表 3). 英国剑桥结晶学数据中心 (CCDC) 的 Relibase 数据库^[22]是商业版 Relibase+ 的一个免费子集功能版本, 但是仍然提供整个 PDB 数据库结构的浏览和搜索功能, 对于学术用户是免费开放和继续维护的. Relibase 提供了多种多样的功能和整合数据, 直观灵活的界面工具使得它成为药物研发过程中很有价值的数据挖掘工具. Ligand Expo 数据库 (原名 Ligand Depot)^[23]是 RCSB PDB 下属的一个数据库, 提供了 PDB 数据库结构中所包含的小分子结构信息, 包括小分子和 PDB code 之间的相互对应关系, 以及 PDB/SDF/MOL/SMILES/InChI String/mmCIF 等各种格式的结构文件下载、记录的浏览和搜索功能等. 这个数据库对于检索 PDB 中的小分子结构信息非常有用. PDBeMotif 数据库^[24]提供了一套非常快速和强大的搜索工具来研究 PDB 中的结构数据, 可以在单一查询中组合使用蛋白序列、化学结构及 3D 数据. PDBeMotif 可以用于检查单一蛋白或某类蛋白如激酶的结合口袋特征, 可以提取其中的保守性结构特征. sc-PDB 数据库^[25]致力于提供 PDB 数据库中具有成药潜力的蛋白结合口袋注释信息. 首先从 PDB 数据库中挑选出高分辨率的蛋白晶体复合物结构, 其中应包含有效的蛋白结合口袋和小分子化合物. 更重要的是, 小分子应当满足一系列的成药性质, 而不是仅从结构角度来考虑. sc-PDB 数据库中的数据均通过 PDB、UniProt 及 GO 等一系列数据库进行了详细注释. Het-PDB Navi 数据库^[26]提供对 PDB 数据库中蛋白-小分子相互作用的分类和检索, 不过这个数据库主要集中在小分子, 而没有处理蛋白质和核酸大分子的相关信息.

表 2 提供生物大分子结构与分类信息的数据库举例

数据库	内容和数据量	最后更新	官方网址
Protein Data Bank (PDB)	104866 个生物大分子三维结构	2014/11	http://www.pdb.org/
PDBREPORT	超过 2000 万条结构诊断信息	2014/11	http://swift.cmbi.ru.nl/gv/pdbreport/
SCOP	2009 年 2 月 23 日的版本包含 38221 个 PDB 三维结构, 110800 个结构域信息. 提供蛋白结构的注释和分类信息. SCOP2 为后续的改进版本	2009/06 2014/01	http://scop.mrc-lmb.cam.ac.uk/scop/ http://scop2.mrc-lmb.cam.ac.uk/
CATH	提供 2600 万个蛋白结构域信息, 归类为 2738 个家族	2013/03	http://www.cathdb.info/
MMDB	提供基于 PDB 数据库的增值服务	2014/11	http://www.ncbi.nlm.nih.gov/structure/
PDBsum	包含 108050 个 PDB 结构, 其中 2142 个为被替换的 PDB code	2014/11	http://www.ebi.ac.uk/pdbsum/

表 3 提供生物大分子结构信息和分析工具的数据库举例

数据库	内容和数据量	最后更新	官方网址
Relibase	包含 103187 个 PDB 结构的解析	2014/11	http://relibase.ccdc.cam.ac.uk/
Ligand-Expo	包含 19202 个来自 PDB 数据库的配体小分子	2014/11	http://ligand-expo.rcsb.org/
PDBeMotif	包含 104594 个 PDB 结构及 16759 个配体小分子	2014/11	http://www.ebi.ac.uk/pdbe-site/pdbemotif/
sc-PDB	包含 9283 个 PDB 结构、3678 个蛋白和 5608 个配体小分子	v2013	http://bioinfo-pharma.u-strasbg.fr/scPDB/
Het-PDB Navi.	包含 104537 个 PDB 结构	2014/11	http://hetpdnavi.nagahama-i-bio.ac.jp/

在药物研发流程的第三阶段, 研究人员需要考虑活性化合物的筛选、先导化合物的设计和改进以及结构-活性关系等方面. 相关数据库种类比较多, 功能也很丰富, 涵盖化合物实体库、药物-靶标亲合性实验数据、药物活性测试实验等方面的信息. 例如, 由美国国立健康研究院(NIH)主持发展的 PubChem 数据库^[27,28]是一个大型的综合数据库, 其包含的 3 个子数据库分别收录 Bioassay 生物活性测试数据、compound 化合物结构数据和 substance 物质数据. 其数据规模庞大, 仅化合物的记录总数就超过 1 亿条, 还提供各种外部数据库的交叉链接信息, 每条记录的注释信息非常丰富, 搜索和分析工具直观, 此外, 用户还可以上传自己的数据供分享. 另外一个有广泛影响的大型综合性数据库是欧洲生物信息学中心(European Bioinformatics Institute)开发的 ChEMBL 数据库^[29]. 该数据库包含大量活性化合物(包括药物和候选药物)的信息, 并提供活性化合物与靶标的亲合性、各种生物功能以及药代动力学性质、毒副作用(absorption, distribution, metabolism, excretion, toxicity, ADMET)等数据. 这些数据绝大部分是通过人工定期从一系列学术期刊中提取出来的, 然后经过核查和规范化再导入 ChEMBL 数据库; 另外, 少部分数据则来自 PubChem 等其他外部数据库以及一些制药公司的内

部数据. ChEMBL 涵盖了化学生物学和药物化学中广泛的数据内容, 对于药物研发具有重要参考意义. 美国加州大学旧金山分校发展的 ZINC 数据库^[30], 汇集了几千万小分子化合物的三维结构信息, 其主要特色是所有化合物都提供了供应商的供货信息, 说明此化合物具有现实的可合成性. 此外, ZINC 还提供了浏览、搜索和下载功能, 方便用户实现各种查询功能和本地应用. 开发者还对 ZINC 数据库中的化合物进行了深度分析和分类, 提供各种类型、服务不同目的的子库. 该数据库也许是学术界最受欢迎的可直接用于虚拟筛选的数据库. 类似的化合物数据库还有英国皇家化学会管理的 ChemSpider 数据库^[31], 为超过 3200 万化合物提供快速文本和结构搜索功能, 这些结构数据来自 100 多家不同的数据源. 由哈佛大学医学院和美国国立癌症研究所(NCI)和美国麻省理工大学(MIT)等几家单位合作开发的 ChemBank 数据库^[32]收录了大量小分子以及相关的生物筛选测试数据, 同时提供了一系列在线分析工具, 可以为生物学、化学、信息分析等领域提供有价值的数据. 此外, 还有一类比较特殊的数据库专门收录基因芯片的测试数据, 如 Connectivity Map 数据库(Cmap)^[33], 该数据库致力于收集具有规范格式的基因表达谱数据, 利用小分子药物处理人类细胞后的基因表达谱的差

异, 建立小分子药物、基因表达和疾病之间的相互关系. 该数据库可以快速利用基因表达谱的数据比出与疾病高关联性的小分子药物, 并能够推测药物分子可能的作用机制.

药物分子与其靶标分子之间形成特异性的相互作用是药物发挥其作用的基础. 因此, 收集和整理药物-靶标亲合性实验数据对于药物研发具有特殊的重要意义, 也是一批开放数据库的工作重点. Binding Database (Binding DB) 数据库^[34]最早由美国马里兰大学建立, 并于 2002 年上线提供公开访问服务, 致力于提供大分子复合物的等温量热滴定(ITC)以及酶活性抑制数据, 通过不断拓展数据来源, 近年来其收录的亲合性数据已经达到百万条的规模. Binding MOAD (Mother of All Databases) 数据库^[35]由美国密西根大学建立, 并于 2005 年上线公布, 其目标在于提供 PDB 数据库中包含的高质量蛋白-配体复合物数据集. 通过人工阅读每个 PDB 结构文件相关的主引文献并核对配体分子结构并搜索亲合性数据, 即使没有找到亲合性数据, 该复合物也会被收录进来. 本课题组开发的 PDBbind-CN 数据库^[36]则致力于涵盖 PDB 数据库中所有的蛋白-小分子、蛋白-蛋白、蛋白-核酸、核酸-小分子复合物, 并通过人工检阅相关文献来获取这些复合物的亲合性实验数据. 目前其数据的规模、质量和更新频率在同类数据库中都处在领先地位, 后文中将专门介绍这项工作. 本节相关数据

库信息见表 4.

在药物研发后期, 需要考虑候选药物在各种动物模型以及临床实验中的表现. 根据一份对美国和欧洲 10 家大型制药公司的研究统计结果, 1991~2000 年, 从首次在人体开始试验到药物登记注册阶段的临床前药物平均成功上市率约为 11%, 这其中约有 1/2 的原因可以归于 ADMET 性质不佳^[37]. 如果在药物设计的早期阶段适当关注这些问题, 就能够给后续研究和试验减少大量不必要的损失. 总体来说, 有如下几类方法达到这个目的: (1) 查阅已有的实验数据, 找到相同或者相近的化学结构, 以此推测可能的 ADMET 性质; (2) 通过软件和算法来预测新颖骨架化合物可能具有的 ADMET 性质; (3) 了解疾病相关的生物通路, 查看蛋白作用网络中有哪些额外的靶标可能导致药物的副作用; (4) 了解药物的代谢通路, 估计代谢过程中可能发生何种反应.

由于这些信息对于选择候选药物具有关键的指导作用, 该领域中存在一批商业数据库(如 Thomas Pharma 数据库)提供这些信息. 免费开放式数据库发挥的作用同样不可小觑. 例如, 日本京都大学化学研究所 Kanehisa 课题组^[38]开发的 KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库旨在从基因和分子水平的数据信息来了解生物体系的高级功能, 如细胞、组织、生态系统. KEGG 数据库包含多达 17 个子数据库, 涵盖生物通路、基因组、蛋白、

表 4 提供活性化合物信息及其生物活性测试结果的数据库举例

数据库	内容和数据量	最后更新	官方网址
PubChem	6309.5 万个化合物; 111.2 万条生物测活数据; 1.8 亿个化学物质	2014/11	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL	10579 个蛋白靶标; 1638394 个化合物; 1411786 个非冗余的化合物; 12843338 条活性数据	2014/11	https://www.ebi.ac.uk/chembl/
ZINC	超过 3500 万个可购买到的化合物	v2012	http://zinc.docking.org/
ChemSpider	超过 3200 万个化合物	2014/11	http://www.chemspider.com/
ChemBank	1236663 个化合物; 4767 条生物测活数据; 1277 个蛋白	v2.1.8	http://chembank.broadinstitute.org/
Connectivity Map	超过 7000 个基因表达谱数据, 涉及 1309 个小分子化合物	Build 02	http://www.connectivitymap.org/cmap/
BindingDB	1058945 条活性数据; 6997 个蛋白靶标; 453657 个配体小分子	2014/10	http://www.bindingdb.org/
Binding MOAD	23269 个蛋白-配体小分子复合物; 8156 条亲合性数据; 11173 个不同的配体小分子	v2013	http://www.bindingmoad.org/
PDBbind-CN	12995 个 PDB 大分子复合物与相应的亲合性数据, 其中, 蛋白-配体小分子复合物 10656 个, 核酸-配体小分子复合物 87 个, 蛋白-核酸复合物 660 个, 蛋白-蛋白复合物 1592 个	2014/09	http://www.pdbbind-cn.org/

化合物、生化反应、疾病、药物等众多领域. KEGG 数据库是生物系统的计算机表现形式, 其数据信息可以归为 4 大类: 系统信息、基因信息、化学信息和健康信息. 加拿大阿尔伯达大学 Wishart 课题组^[39]开发的 DrugBank 数据库是一个综合性的生物信息学与化学信息学数据库, 它提供了各种药物的化学、药理、药剂详细信息, 以及药物靶标的序列、结构和生物通路信息. DrugBank 中目前收录了 7740 种药物的数据信息, 其中 1584 个为美国食品及药物管理局 (FDA) 批准的小分子药物, 157 种为蛋白和多肽类药物, 89 种为营养药物, 其余约 6000 种为试验阶段药物. 每种药物的详情里包含超过 200 种数据字段信息, 其中与药物分子和靶标相关的信息各占 1/2. 同样, 由 Wishart 课题组^[40]开发的 HMDB (the Human Metabolome Database) 是一个收录人体内小分子代谢物信息的免费数据库, 它可以用于代谢组学研究、临床化学、生物标志物研究以及常规教育. HMDB 数据库中包含 3 类数据信息: 化学数据、临床数据和分子生物学/生物化学数据. 该数据库同时收录水溶性和脂溶性的代谢物, 既收录含量丰富的代谢物(>1 $\mu\text{mol/L}$), 也收录含量较少的代谢物(<1 nmol/L). METLIN (the METLIN Metabolite Database) 数据库^[41]由美国 Scripps 代谢组学研究中心开发, 其收录了代谢物信息以及相应的串联质谱数据. 该数据库可以帮助进行代谢组学实验, 已知的代谢物被加上分子量、化学式、结构等注释信息, 而且还能链接到一些外部资源数据库如 KEGG 等. admetSAR 数据库^[42]由华东理工大学开发, 提供人工收集整理的、超过 95000 种化合物的 ADMET 性质数据. admetSAR 还提供了一种内含高质量 QSAR 模型的预测工具 ADMET-Simulator, 能够预测多达 50 种与 ADMET 性质相关的属性. 该数据库对基于计算机筛选候选药物与环境化合物的

ADMET 性质非常有帮助. 本节相关数据库的信息见表 5.

3 联系生物大分子复合物亲合性数据和三维结构的开放型数据库: PDBbind-CN

药物分子与其靶标分子之间形成特异性的相互作用是药物发挥其作用的基础. 因此, 发展可靠的理论模型和计算方法来预测药物-靶标相互作用, 一直是药物分子设计的一个关键问题. 例如, 在药物分子设计中得到广泛应用的分子对接(molecular docking)方法, 即对药物分子与靶标各种可能的结合模式进行采样, 然后依靠打分函数(scoring function)挑选出最优的解. 发展这一类理论计算方法显然需要高质量的复合物三维结构及可靠的亲合性实验数据作为知识基础. 但是, 当分子对接等一批方法在 20 世纪 90 年代初崭露头角之时, 这样的可用数据曾经十分匮乏, 形成了发展此类理论计算方法的一个瓶颈. 当时造成这一现象的主要原因有 2 个: (1) 虽然 PDB 数据库汇集了三维结构信息, 其中不乏各类生物大分子与小分子配体形成的复合物, 但是 PDB 本身没有对其包含的样本进行系统分析和分类、鉴别出对药物研发有借鉴意义的各类复合物; (2) 由各种实验手段测试获得的药物-靶标亲合性实验数据散布在海量的科学文献中, 无法方便地加以利用. 后来虽然有一些数据库(如 Binding DB)开始致力于系统收集亲合性实验数据, 但是由于它们的目标过于宽泛, 不注重三维结构信息与亲合性数据之间的联系. 这样的数据库对药物化学家理解生物活性分子的结构-活性关系更有意义, 但是对药物-靶标作用理论计算方法的帮助则十分有限.

本课题组发展的 PDBbind-CN 数据库即针对本

表 5 提供化合物药代动力学、代谢性质和毒性数据的数据库举例

数据库	内容和数据量	最后更新	官方网址
KEGG	包含 17 个子数据库, 涉及系统信息、基因组信息、化学信息与健康信息几大类	2014/10 (Release 72.0)	http://www.kegg.jp/
DrugBank	7740 个药物的相关数据, 4282 条非冗余蛋白序列	2014, v4.1	http://www.drugbank.ca/
HMDB	41808 个代谢物的相关数据, 5688 条蛋白序列	2014, v3.6	http://www.hmdb.ca/
METLIN	240528 个代谢物的相关数据, 62020 个串联质谱数据, 184129 个预测结构	2014	http://metlin.scripps.edu/
admetSAR	95629 个化合物的基础性性质数据, 提供多达 26 个 ADMET 性质预测模型	2012	http://lmmd.ecust.edu.cn:8000/

领域中长期存在的这一问题, 为发展针对药物-靶标相互作用的理论计算方法“量身定制”而创建. 该数据库的前身为王任小博士在美国密西根大学工作期间创建的 PDBbind 数据库. 该数据库系统地分析 PDB 数据库中各类蛋白与小分子配体形成的复合物, 并在文献中搜集与之匹配的亲和性实验数据, 并于 2004 年首次上线公开发布(<http://www.pdbbind.org>). 该数据库一经公开就引起了广泛关注, 曾得到美国化学学会出版的 *Chemical and Engineering News* 的专文评述^[43]. 王任小博士回国加入中国科学院上海有机化学研究所之后, 领导本课题组独立负责该数据库的后续开发和全部维护工作, 并将该数据库更名为 PDBbind-CN (<http://www.pdbbind-cn.org>).

对于一个开放式的数据库, 只有实现内容的及时更新才具有长久的生命力. 自 PDBbind 数据库创立之初, 就规划了完整的更新计划, 并且十余年来坚持完成了对该数据库的逐年更新. PDBbind-CN 数据库所收录的数据规模也随着 PDB 数据库的增长而增长, 最新的 2014 版中已经包含了 PDB 中约 13000 个各类复合物的亲和性实验数据(表 6). 除常规的内容更新外, 每一个发布的新版本都在算法上有所进步.

表 6 PDBbind-CN 数据库历年发布版本的基本情况^{a)}

版本	PDB 数据库中收录的结构总数	PDBbind 识别的复合物总数	已知亲和性实验数据的复合物总数
2002	19621	5671	1446
2003	23790	5897	1763
2004	28991	6847	2276
2005	34338	9775	2756
2006	34338	9775	2632
2007	40876	11822	3124
2008	48092	18211	4300
2009	55118	23284	5678
2010	62387	26434	6772
2011	70224	30259	7986
2012	78235	34180	9308
2013	87085	38918	10776
2014	96952	44569	12995

a) 2002 版是 PDBbind 数据库 2004 年公开上线以来发布的第一个版本; 由于改变了版本的命名规则, 2006 版实为 2005 版的修订版本, 基于同样的 PDB 数据库内容; 2007 版是 PDBbind 数据库完全转移至国内之后第一个公开发表的版本

例如, PDBbind 数据库的早期版本只考虑 PDB 中由小分子配体与蛋白形成的复合物. 从 2008 年开始, 除了小分子配体-蛋白复合物, 还开始把小分子配体与核酸形成的复合物, 以及蛋白-蛋白复合物、蛋白-核酸复合物等生物大分子复合物包括在内, 从而形成了对 PDB 内容的完整覆盖.

下文中将简要介绍 PDBbind-CN 数据库的基本构建方法和主要内容. PDBbind-CN 数据库是建立在 PDB 数据库基础上的二级数据库(secondary database), 目的在于对 PDB 数据库进行系统分析和整理, 并且提供额外的亲和性实验数据, 建立复合物的三维结构和亲和性数据之间的联系. PDBbind-CN 数据库的构建流程主要包括以下几个关键步骤:

(1) 对 PDB 数据库中的复合物进行识别和分类. 工作流程的第一步是下载整个 PDB 数据库, 包括所有的三维结构及辅助性信息. PDB 数据库本身并不提供关于结构的详细分类信息, 也没有关于大分子复合物的具体结构化组成信息. 因此, 多年来本课题组^[36]开发并不断改进了一套程序算法对 PDB 结构文件进行分析, 从中识别出有效的蛋白-小分子、蛋白-蛋白、蛋白-核酸及核酸-小分子复合物. 这些算法中考虑了复合物组成单元之间的相互作用情况、配体小分子结构特征等众多因素. 整个分类过程完全自动化, 不依赖任何 PDB 结构文件中的文字注释信息, 能以很高的准确度鉴定出有效的复合物. 例如, 在最新公开的 2014 版本中, 考察了截至 2014 年 1 月 1 日 PDB 数据库中收录的 96952 个实验测得的生物大分子三维结构, 本课题的程序一共识别出 44569 个有效的复合物结构, 其中 32926 个为蛋白-小分子复合物, 6474 个为蛋白-蛋白复合物, 4397 个为蛋白-核酸复合物, 772 个为核酸-小分子复合物.

(2) 人工搜集亲和性数据. 工作流程的第二步是从公开发表的文献中搜集第一步中鉴定出的复合物的亲和性实验数据. 这里的首要问题是从哪些文献中搜集数据, 面对海量的学术文献如何找到工作量和效率的平衡点. 本课题组的做法是系统检查由 PDB 提供的每个结构的所谓“主引参考文献”(primary reference), 该文献的主要功能通常是描述三维结构的解析过程和解析结果. 假设某个 PDB 结构是一个复合物, 可以猜想作者的研究兴趣除了解析这个复合物的三维结构, 也许还包括复合物构成单元之间的相互作用. 如果是这样, 则相关的亲和性实验测试

结果就很有可能在同一篇文献中进行报道. 或者, 如果这一复合物是一个热门的研究对象, 其亲合性数据已由前人测定, 那么这一篇文献就很有可能再次提及或引用这样的结果. 基于这样的假设, 对于每个复合物仅需人工查阅一篇文献(必要的时候再查阅其附加材料信息和其他关键参考文献), 就有可能获得需要的数据.

事实证明, 这种策略的命中率和效果令人满意, 极大地节省了人工. 在 PDBbind-CN 近年来的每次更新工作中, 大约要处理 8000 多个新增的 PDB 结构, 相应地人工阅读约 3000 篇公开发表文献. 在此过程中, 本课题组收录 3 种符合剂量-效应类型的亲合性数据, 即解离常数 K_d , 抑制常数 K_i 和半数抑制浓度 IC_{50} 数据. 人工阅读大量文献难免产生错误. 为了保证数据的准确性, 每一篇文献原则上由 2 人独立完成数据查找, 只有获得相同的结果才会收录入数据库. 截至 2014 版本, 已经总计下载处理超过 28000 篇原始全文文献, 涵盖 252 家不同的 SCI 学术期刊. 如此巨大的工作量和文献资源是多年坚持不懈的努力的结果.

(3) 蛋白-小分子配体复合物结构的后处理. 作为一个基于 PDB 的二级数据库, PDBbind-CN 数据库提供了具有附加价值的内容. 除了上文提到的亲合性实验数据之外, 还提供经过处理的复合物结构, 方便用户使用. 由于原始的 PDB 结构文件只包含了分子的基本三维坐标信息, 并没有提供精确的原子互连信息、原子类型和化学键类型信息、电荷信息、质子化信息等, 对于大多数的分子模拟软件来讲, 这样的结构文件无法直接使用, 需要进行一定的预处理. 为此, 本课题组开发了一系列程序来自动拆分、处理和规范化相应的分子结构, 简化了原始 PDB 文件中多余的内容, 修正了一些常见的错误, 并将这些结构文件存储为标准格式(如 PDB 格式), 可以被目前多数分子模拟软件直接利用. 所有处理过的结构文件与亲合性数据一起打包放在 PDBbind-CN 网站上供用户免费下载. 这一工作目前只针对由小分子配体与蛋白形成的复合物, 未来将逐步推广到其他类型复合物(蛋白-蛋白复合物以及蛋白-核酸复合物).

另外, 创建 PDBbind-CN 数据库的一个核心目标是服务于分子相互作用的理论方法研究. PDBbind-CN 数据库中收录的复合物的结构和亲合性数据质量有高低之分, 从质量控制的角度考虑, 并不是所有样本都适合用来作为支持理论方法研究的基础. 因此

本课题组设计了一系列规则从 PDBbind-CN 中收录的蛋白-小分子配体复合物提炼出几类高质量的子集. 为示区分, 将 PDBbind-CN 中所有蛋白-小分子配体复合物的集合称之为“普通集”(general set). 通过应用一系列相当严格的筛选规则^[44], 从普通集中挑选出复合物的组成、结构及亲合性数据等各方面都不存在明显问题的样本集合, 称为“精选集”(refined set). 一般而言, 普通集中只有不足 1/3 的样本可以进入精选集. 例如, 在最新的 v2014 版中, 精选集一共包含 3446 个高质量的蛋白-小分子复合物结构, 约占普通集的 26.5%. 精选集更适合作为发展分子对接方法或打分函数的训练集以及测试集, 得到了国际同行的广泛应用. 例如, 华盛顿大学 Marshall 课题组^[45]选择 PDBbind 精选集 2009 版, 用于发展 PHOENIX 打分函数; 四川大学杨胜勇课题组^[46]选择 PDBbind 精选集 2011 版, 用于发展 ID-Score 打分函数; 芝加哥伊利诺大学 Abad-Zapatero 课题组^[47]选择 PDBbind 精选集 2007 版, 用于分析配体分子效率等.

作为本课题组基于 PDBbind-CN 数据库的一项主要应用, 我们发展了评估打分函数性能的方法体系 CASF (comparative assessment of scoring functions)^[48,49], 从 PDBbind 精选集中按照非冗余原则挑选出了 65 类共 195 个蛋白-配体复合物作为测试集, 按照打分函数的典型应用方式设计了“打分能力”、“排序能力”、“对接能力”、“筛选能力”4 项基本指标, 对主流药物分子设计软件中的 20 种打分函数进行了系统的评估. 该评价方法体系将构象采样过程与打分函数分离开来, 因此可以对打分函数的性能做出更准确的评估. 根据本课题对文献的跟踪, 近年来 1/2 以上新发表的打分函数采用了该类 CASF 方法体系.

(4) PDBbind-CN 数据库的 Web 用户界面. 作为工作流程的最后一步, 将 PDBbind-CN 数据库中收录的复合物的结构、亲合性数据以及其他相关信息, 整合到一个可以通过 Web 界面访问的数据库中. 本课题组采用 Microsoft IIS 服务器, ASP.NET 动态脚本语言, Access 数据库和一系列 C++编写的可执行程序来实现 PDBbind 的网络数据库服务. 服务器采用双路四核英特尔至强处理器, 8 GB 内存, 百兆互联网络, 以保证数据库的服务质量. 如图 1 所示, 对于 PDBbind-CN 中的每个复合物, 用户可以获取 PDB 代码、分辨率、发表年份、蛋白和配体名称、酶分类号、原始文献信息和链接、蛋白序列等基本信息, 还能在

图 1 PDBbind-CN 数据库网站用户界面

线查看蛋白、配体、复合物、结合口袋的三维结构。在三维结构图的下方, 用户可以找到与该复合物具有相同配体分子或类似靶标蛋白的其他复合物列表、外部数据库链接等丰富的参考信息。此外, 用户还可以在其他页面按照数据子集、PDB code、蛋白和配体名称、分辨率、酶分类号、亲合性大小等参数进行模糊查询, 匹配的结果可以保存为 Excel 文件用作原始数据格式, 也可以保存为 PDF 文档供阅读。用户还可以对配体分子结构进行精确匹配、相似度搜索和子结构搜索。值得一提的是, 通过 RCSB 的 ligand-expo 数据库, 实现了对整个 PDB 数据库中配体分子的结构搜索功能, 而并不仅限于 PDBbind-CN 中收录的配体分子。最后, 关心特定靶标的用户还可以使用蛋白/核酸序列进行 BLAST 相似度搜索, 每段序列的比对和匹配情况可以在线进行查看。同样, PDBbind-CN 支持对 PDB 中所有的蛋白和核酸序列进行比对, 不仅限于 PDBbind-CN 中收录的部分结构。

目前 PDBbind-CN 数据库可以通过域名 <http://www.pdbbind.org.cn> 和 <http://www.pdbbind-cn.org> 进行访问。最新公开版本为 2014 版, 于 2014 年 9 月发布。该版本中共收录 PDB 数据库中 12995 个生物大

分子复合物结构的亲合性数据, 与 v2013 版相比数据增长约 20%。其中, 10656 个为蛋白-小分子复合物, 1592 个为蛋白-蛋白复合物, 660 个为蛋白-核酸复合物, 87 个为核酸-小分子复合物。亲合性数据类型方面, 5639 个为 K_d 数据, 3496 个为 K_i 数据, 3861 个为 IC_{50} 数据。PDBbind-CN 数据库目前拥有活跃的注册用户超过 2600 人, 分别来自世界 67 个国家和地区。总计页面访问次数已超过两百万次, 数据库工作总结他引次数超过 400 次。PDBbind-CN 数据库目前已经与 PDB, PDBsum, PubChem 等数据库建立了相互的数据链接, 此外还有 sc-PDB、AutoBind 等一系列数据库在其网站提供指向我们原始数据的链接。特别值得一提的是, PDBbind-CN 被 PDB 选为提供亲合性实验数据的数据源之一, 我们提供的亲合性数据被展示在每个相关 PDB 结构的页面中。这一事件也充分说明了 PDBbind-CN 数据库所形成的国际影响力。

4 结论

让更多人能够接触到有用的信息资源是发挥其价值的最佳方式, 开放访问因此是科学数据库发展的一个大趋势。本文探讨了生物和化学领域中开放访问数据库的基本特征以及对于推动科学研究进程的积极意义, 系统介绍了与药物研发各个阶段相关的约 30 种开放数据库的基本功能、特色和运行现状, 为从事药物研发工作的读者提供了互联网上可获得的数据资源的参考信息。

作为一个实例, 本文重点介绍了本课题组开发和维护的生物大分子复合物 PDBbind-CN 数据库, 包括其创建背景、制作流程和方法、最近的更新和典型应用情况。通过这个例子可以看到, 数据库的研发和维护都需要大量人力物力的持续投入, 而这些投入在短期内并不能获得可见的利益回报。但是, 如果能够坚持深入挖掘数据库的科学价值, 最终这些开放访问数据库所能发挥的作用将远超过商业数据库。在此我们呼吁政府有关部门, 重视对基础研究领域中建设开放式数据库的资助, 力争在国际大科学研究领域中逐渐占据制高点。

致谢 本工作由国家自然科学基金(81172984, 21072213, 21102168)、国家高技术研究发展计划(2012AA020308)资助, 特此一并致谢。

参考文献

- 1 Williams AJ. A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today*, 2008, 13: 495–501
- 2 Ellis LB, Attwood TK. Molecular biology databases: today and tomorrow. *Drug Discov Today*, 2008, 6: 509–513
- 3 Furner J. Little book, big book: before and after little science, big science: a review article, part I. *J Lib Inf Sci*, 2003, 35: 115–125
- 4 Bucci M, Goodman C, Sheppard TL. A decade of chemical biology. *Nat Chem Biol*, 2010, 6: 847–854
- 5 Rojas-Ruiz FA, Vargas-Méndez LY, Kouznetsov VV. Challenges and perspectives of chemical biology, a successful multidisciplinary field of natural sciences. *Molecules*, 2011, 16: 2672–2687
- 6 DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ*, 2003, 22: 151–185
- 7 Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. *Health Policy*, 2011, 100: 4–17
- 8 Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov*, 2010, 9: 203–214
- 9 Mullard A. New drugs cost US\$2.6 billion to develop. *Nat Rev Drug Discov*, 2014, 13: 877
- 10 Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res*, 2002, 30: 27–30
- 11 Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G. The European nucleotide archive. *Nucleic Acids Res*, 2011, 39: D28–D31
- 12 Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*, 2013, 41: D36–D42
- 13 The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, 2014, 42: D191–D198
- 14 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database—2009 update. *Nucleic Acids Res*, 2009, 37: D767–D772
- 15 Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Romá-Mateo C, Theodosiou A, Mitchell AL. The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. *Database*, 2012, 2012: bas019
- 16 Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 2003, 10: 980
- 17 Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature*, 1996, 381: 272
- 18 Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res*, 2014, 42: D310–D314
- 19 Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton JM, Orengo CA. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*, 2013, 41: D490–D498
- 20 Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH. MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res*, 2014, 42: D297–D303
- 21 Laskowski RA. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res*, 2001, 29: 221–222
- 22 Hendlich M, Bergner A, Günther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 2003, 326: 607–620
- 23 Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, 2004, 20: 2153–2155
- 24 Golovin A, Henrick K. MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, 2008, 9: 312
- 25 Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D. sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model*, 2006, 46: 717–727
- 26 Yamaguchi A, Iida K, Matsui N, Tomoda S, Yura K, Go M. Het-PDB Navi.: a database for protein-small molecule interactions. *J Biochem*, 2004, 135: 79–84
- 27 Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem*, 2008, 4: 217–241
- 28 Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant

- SH. PubChem's BioAssay Database. *Nucleic Acids Res*, 2012, 40: D400–D412
- 29 Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 2012, 40: D1100–D1107
- 30 Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model*, 2012, 52: 1757–1768
- 31 Pence HE, Williams A. ChemSpider: an online chemical information resource. *J Chem Educ*, 2010, 87: 1123–1124
- 32 Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA. ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res*, 2008, 36: D351–D359
- 33 Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 2006, 313: 1929–1935
- 34 Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, 2007, 35: D198–D201
- 35 Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (mother of all databases). *Proteins*, 2005, 60: 333–340
- 36 Liu ZH, Li Y, Han L, Li J, Liu J, Zhao ZX, Nie W, Liu YC, Wang RX. PDB-wide collection of binding data: current status of the PDBbind-CN database. *Bioinformatics*, 2015, 31: 405–412
- 37 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*, 2004, 3: 711–715
- 38 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*, 2014, 42: D199–D205
- 39 Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res*, 2014, 42: D1091–D1097
- 40 Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorn Dahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A. HMDB 3.0—the Human Metabolome Database in 2013. *Nucleic Acids Res*, 2013, 41: D801–D807
- 41 Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol*, 2012, 30: 826–828
- 42 Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, Lee PW, Tang Y. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model*, 2012, 52: 3099–3105
- 43 Wilson EK. New online databases combine structure and binding of protein-ligand complexes. *Chem Eng News*, 2005, 83: 33–34
- 44 Li Y, Liu ZH, Li J, Han L, Liu J, Zhao ZX, Wang RX. Comparative assessment of scoring functions on an updated benchmark: I. Compilation of the test set. *J Chem Inf Model*, 2014, 54: 1700–1716
- 45 Tang YT, Marshall GR. PHOENIX: a scoring function for affinity prediction derived using high-resolution crystal structures and calorimetry measurements. *J Chem Inf Model*, 2011, 51: 214–228
- 46 Li GB, Yang LL, Wang WJ, Li LL, Yang SY. ID-score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J Chem Inf Model*, 2013, 53: 592–600
- 47 Abad-Zapatero C, Perišić O, Wass J, Bento AP, Overington J, Al-Lazikani B, Johnson ME. Ligand efficiency indices for an effective mapping of chemico-biological space: the concept of an atlas-like representation. *Drug Discov Today*, 2010, 15: 804–811
- 48 Cheng TJ, Li X, Li Y, Liu ZH, Wang RX. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*, 2009, 49: 1079–1093
- 49 Li Y, Han L, Liu ZH, Wang RX. Comparative assessment of scoring functions on an updated benchmark: II. Evaluation methods and general results. *J Chem Inf Model*, 2014, 54: 1717–1736

Open access databases in the field of drug discovery

Zhihai Liu, Jie Li, Li Han, Yan Li, Jie Liu, Wei Nie, Yuchen Liu, Renxiao Wang^{*}

State Key Laboratory of Bioorganic Chemistry; Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

^{*}Corresponding author (email: wangrx@mail.sioc.ac.cn)

Abstract: Open access databases have been proved as a great aid to the research in many scientific areas. In this article, we review the main-stream open access databases at each major stage in the drug discovery pipeline. In particular, we describe the PDBbind-CN database developed by our group. It is a database linking the three-dimensional structures and experimental binding data of biomolecular complexes, which has been applied by researchers worldwide in various molecular modeling or bioinformatics/cheminformatics studies.

Keywords: drug discovery, open access database, PDBbind-CN, binding affinity data