PDBbind数据库在计算机辅助药物设计中的应用

刘志海　程铁军　王任小
中国科学院上海有机化学研究所，上海　200032

摘　要　本文首先回顾了e-Science的产生背景与内涵，分析了信息化技术在现代计算机辅助药物设计中的重要作用与意义。然后介绍本课题组科研工作的主要内容与框架，以PDBbind数据库的构建和开发为案例，阐明本课题组的信息化应用架构，介绍了一些基于PDBbind数据库开发的软件方法以及研究实例。最后对e-Science在计算机辅助药物设计领域的发展方向做了一些思考和展望。

关键词： 科研信息化；信息化平台；PDBbind数据库；计算机辅助药物设计

Application of the PDBbind Database in Computer-Aided Drug Design

Liu Zhihai, Cheng Tiejun, Wang Renxiao
Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

Abstract: In this paper, we firstly review the background and essence of e-Science, and analyze its impact on computer-aided drug design. We then introduce the on-going research in our group, in particular the construction and implementation of the PDBbind database, reveal the architecture of our information platform, and list some representative research achievements in computer-aided drug design based on the PDBbind database. Finally, the future development of e-Science in computer-aided drug design is prospected.

Keywords: e-Science; Information-based platform; PDBbind database; Computer-aided Drug design

1. 引言

众所周知，基于实验或以理论分析为主的传统科学研究方法存在很多不足，如比较封闭、缺乏模拟和仿真的手段、科研周期往往较长或成本很高、对某些研究领域传统方法目前还无能为力等。进入21世纪后，科学研究面临一些新的挑战，科研环境也发生了巨大的变化。首先，科学研究的问题空前复杂化，其研究的对象往往涉及众多科学领域；其次，科研过程中信息和数据的及时获取和处理越来越重要，仿真和大规模的计算逐渐成为科学研究过程中分析、发现和预测的主要手段之一；最后，科研活动中的合作与交流日益频繁和深入。跨单位、跨地域甚至是跨国家、跨学科的合作和交流，使得学科之间的交叉与融合越来越普遍。在这种背景下，e-Science作为一种科学研究新平台和新环境应运而生[1]。

e-Science是“在重要的科学领域中的全球性合作，以及使得这种合作成为可能的下一代基础设施”[2]。e-Science在网格（grid）技术基础上直接参与科研工作，它具有三个方面的要素：一是随处可得的计算资源；二是海量的数据存储和处理能力；三是人员交流和无缝协作能力。这些正是当前科研工作向更深层次、更大范围进行拓展所亟需的。在中国，国家自然科学基金重大研究计划《以网络为基础的科学活动环境研究》2003年批准的项目总经费2510万元，2005年批准的项目总经费为1350万元，2009年新批准项目经费已达820万元。在其资助的研究方向中就包括“生物信息学示范应用”和“计算化学网格示范应用”等[3]。毋容置疑，e-Science在这些领域的建设成果将为计算机辅助药物设计方面的研究提供强大支持。

2. e-Science与计算机辅助药物设计

新药创制是一项耗资巨大且效率低下的工作。根据国际上的统计，每成功研制一种上市新药，平均需要花费约10-15年的时间，耗资超过10亿美元。新药创制的前期工作主要存在两个瓶颈：一是靶标生物大分子的确定及验证；二是具有生物活性的小分子药物的发现和设计。

在今天，应用各种理论计算方法和分子图形模拟技术，进行计算机辅助药物设计（Computer-Aided Drug Design, CADD），已成为国际上十分活跃的科学研究领域。将CADD应用到新药研究的工作流程中，能够有效缩短研发周期、节约研发费用、提高新药筛选的成功率。据统计[4]，CADD辅助设计方法平均可为每个药物的研发节约开发成本1.3亿美元，缩短开发周期0.9年。目前已经有由CADD方法参与而获得成功的药物上市。

制药行业研发周期长，投资风险大，迫切需要广泛开展合作与交流，获得各方面的支持。

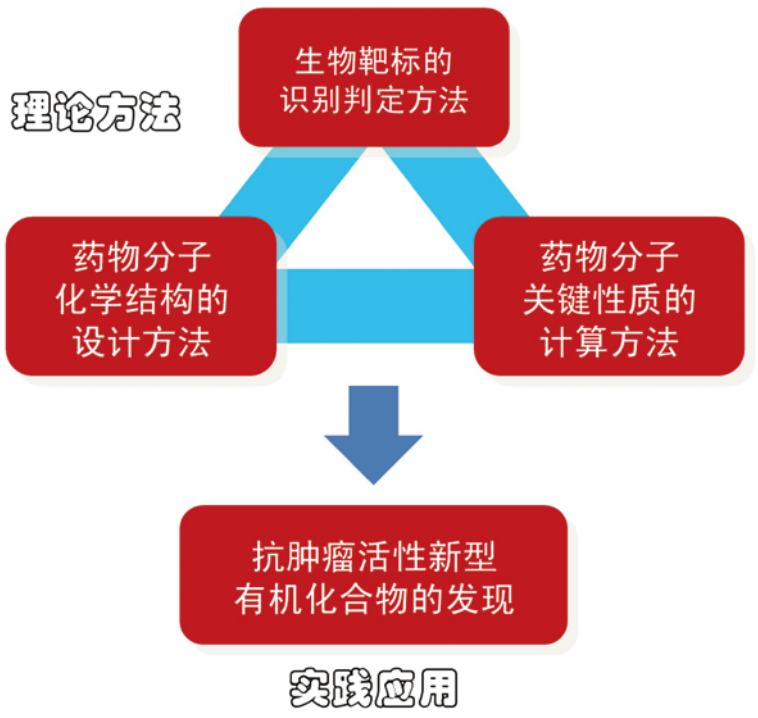


图1 本课题组科研工作的整体规划

图2 本课题组基于Web的网络科研服务平台

因此，制药行业实际上非常适合大规模地应用信息化技术来辅助药物设计，尤其是在药物研究的初期阶段。它的实现依赖于e-Science在各个相关学科领域的进展和成果，其中首要的问题是要坚持开放性，其中包括开放科学数据、开源软件、免费公众数据库等，这些都是信息学研究的基本要求。

本课题组研究内容围绕计算机辅助药物设计中的几个关键问题，致力于药物设计初期阶段的候选化合物设计、筛选、优化与验证工作，其整体研究规划如图1所示。一方面我们以分子模拟、生物信息学和化学信息学为手段，研究有机小分子和生物大分子相互作用的基本理论，发展和完善计算机辅助药物设计的理论方法。另一方面我们针对与重大疾病过程相关的靶标，运用计算机辅助药物设计方法并结合必要的化学和生物实验手段，来寻找和设计有活性的有机小分子化合物。

3. 基于Web网络的科研信息化平台

现代科学研究活动往往需要存储和快速地处理大量的信息和数据，要求信息资源在成员之间便捷地分享和交流，并及时地对外发布所取得的研究成果。因此，基于Web网络的科研信息化平台无疑是我们的最佳选择。本课题组的平台整体架构如图2所示。

3.1 HPC高性能计算平台

计算机辅助药物设计是一门高度交叉的新兴学科，涉及生物化学、合成化学、计算化学和生物信息学等众多领域[5]。其中计算化学是根据物理化学的基本理论以大量的数值运算方式来探讨体系的各种性质。高精度量子化学计算、分子动力学模拟以及统计力学等方法可以用来研究生物大分子结构与功能关系，然而这些应用都是非常耗费计算资源的。在生物信息学方面，随着全球在基因组学和蛋白质组学等领域的研究进展，相关信息出现了爆炸性增长，研究人员经常需要面对海量的生物学数据。大量的核酸、蛋白质序列、基因多态、基因表达谱和蛋白质谱数据的积累

已远远超出了一般实验室的分析计算能力。以上研究内容均要求建立高性能计算平台，使得大规模、快速的数据处理成为可能。

目前我们课题组的HPC高性能计算平台的硬件设备资源有联想Deepcomp 1800集群（16 CPUs）、SGI Origin 300集群（32 CPUs）、宝德Powerleader集群（128 CPUs）。平台内各模块分工明确、各司其职，同时配备有Gaussian、Sybyl、Discovery studio、Schrödinger、MOE、GOLD等种类齐全的量化计算和分子模拟软件。

研究中如果遇到特别大型的作业，还可以依托上海超级计算中心的强大计算资源，该中心在生命科学、汽车制造、土木工程、计算化学、环境模拟、航空航天等众多领域均可提供专业服务。

3.2 基于Web网络的科研协作平台

在高性能Web服务器的支持下，我们以此为沟通枢纽，将课题组的各项科研活动有机联系起来，构建了一个高效、开放、共享的科研协同工作环境。

首先，我们通过用户友好的Web管理系统，对运行于HPC高性能计算平台上的作业进行实时在线地监控，大大降低了Linux集群系统的使用难度。如有需要，我们可与相关单位开展合作，共享计算资源，以提高硬件资源的利用率。

第二，我们设计并实现了在线的PDBbind数据库（稍后详细介绍），免费提供给学术机构和企业单位使用。该数据库具有方便的文本查询、结构查询功能以及直观的三维分子结构显示界面，筛选的结果可以PDF、Excel等格式方便地下载。目前该数据库在全球已经有正式注册用户1300余人，遍及高等院校、科研学术机构和知名制药企业。PDBbind数据库中所包含的高质量数据集是药物设计中不可多得的基础性数据。

第三，我们通过课题组网站（http://www.sioc-ccbg.ac.cn）对外发布由本课题组自主开发的各种软件，这些软件针对学术机构全部免费。目前已经成功推出的包括有分子结构自动感知与格式转换软件I-interpret[6]，化合物脂水分配系数预测软件XLOGP3[7]。这两个软件除提供下载和在线计算服务外，还建设有相关的论坛板块为用户提供技术支持。这种共享、开放的交流形式有助于我们及时从用户处获得更多的反馈意见，从而能够更好地改进我们的程序。

第四，我们通过建立课题组成员在线工作日历的方式，使得个人能合理安排工作进度，增强了团队成员之间的互动了解和沟通，大大提高了工作效率。另外通过TB级的高容量磁盘阵列搭建了FTP文件共享服务器，为课题组成员提供大量的专业学习资料、数据备份及文档交流服务。

第五，我们自主开发了CMS化合物管理系统，用于统一管理实验室自行合成和商业购买的化合物。系统中记录了每个化合物的理化性质、合成路线（包括实验条件）、生物活性测试数据等详尽信息。该系统提供方便的录入、查询、修改、分权限共享等功能，不仅提高了化合物库管理效率，也减少了科研活动中的大量不必要的重复劳动。该化合物管理系统有望在本研究所范围内进行推广。

3.3 辅助的高性能桌面工作站

通过为每位成员配备一台以上的高性能Linux系统桌面工作站，来随时提交各种中小型的计算任务，以及从事分子模拟等可视化研究。这批工作站在课题组内部通过高速以太网互联，可以实现资源共享，减少计算资源的空闲浪费。

4.本课题组基于PDBbind数据库的相关研究工作

PDBbind数据库作为本课题组科研信息化平台的重要内容之一，是对外部用户提供的主要开放服务项目。PDBbind数据库是本课题组与美国密西根大学Shaomeng Wang教授课题组合作发展的项目[8,9]，设计目标是为研究各种生物体系中的分子识别过程提供高质量的素材。PDBbind数据库针对美国RCSB的Protein Data Bank（PDB）数据库进行二次开发，系

统地收集了PDB数据库中各种类型复合物的三维结构信息以及亲合性实验数据。其构建过程可简述如下：首先通过程序算法识别出PDB数据库中所有的蛋白质-小分子配体复合物，然后通过程序自动过滤并结合人工复核的方式获得每个复合物所对应的参考文献中的亲合性实验数据，包括解离常数Kd、抑制常数Ki和半数抑制浓度IC50值；根据实际科研课题的具体需求，采用更加严格的区分条件和筛选标准从整个PDBbind数据库中衍生出各种子集。

我们为PDBbind数据库配备了专用服务器，并从2007年起以域名http://www.pdbbind.org.cn公开发布，面向国内外学术界和企业用户提供免费服务。用户可以在其Web界面上方便地实现数据检索、结构浏览、批量下载等功能。PDBbind数据库目前拥有来自40余个国家超过1300名的注册用户，网站访问量已经突破30000次，并在持续增长中。本课题组对PDBbind数据库每年定期进行更新以适应PDB数据库内容的扩增。从2008版本开始，我们引入了蛋白质-蛋白质复合物、蛋白质-核酸复合物、核酸-小分子复合物三种新的复合物类型，使得PDBbind数据库不仅包括蛋白-小分子配体相互作用信息，也包括生物大分子之间的相互作用信息，进一步拓宽了PDBbind数据库的应用范围。

下面以本课题组围绕该数据库所开展的相关研究工作（见图3）为例，简要介绍一下PDBbind数据库在计算机辅助药物设计中的几种应用形式。

4.1 打分函数X-Score的开发

基于结构的药物设计依赖于一些计算方法来准确预测候选小分子与靶标大分子的结合能力。在现有方法中，打分函数（Scoring Function）因其快速、准确的优点获得了广泛的应用。不仅如此，打分函数在虚拟筛选、从头设计、基于片段的药物设计，结合位点识别和集中库设计等方面同样有着重要的应用。

X-Score是由本课题组自主开发的经验型打分函数，因其较高的准确性在学术界备受赞誉。我们知道，经验打分函数的准确度依赖于所使用的数据集的质量，包括实验测定的蛋白-配体复合物结构数据和亲合性数据。PDBbind数据库的建立和不断完善，为我们进一步提高X-Score的精度提供了坚实的基础。从表1中可以看出[10,11]，X-Score最新版本（v.2）的准确度进一步提升，且明显优于一些主流大型商业软件中的打分函数。

4.2 打分函数的评估

打分函数发展到今天，已经有几十种可用的程序和方法。各种方法的性能怎样，各自的优缺点是什么，以及如何从中选择最适合当前研究工作的打分函数等问题都是非常值得研究的课题，同时也具有非常重要的现实意义。我们在PDBbind数据库基础之上构建了PDBbind“核心集”，该测试集在保证多样性前提下消除了数据冗余，其质量之高，在前人的工作当中是非常罕见的。因此，可作为评估分子对接程序和打分函数的通用测试集。例如，应用该测试集我们对

图3 本课题组基于PDBbind数据库的研究工作

打分函数	N（样本数）	R（相关系数）	SD（标准偏差）	MUE（平均无符号误差）
X-Score v.2	800	0.628	1.72	1.32
X-Score v.1	800	0.566	1.82	1.42
Sybyl::ChemScore	797	0.499	1.91	1.50
DrugScore	800	0.476	1.94	1.50
Cerius2::PLP	800	0.458	1.96	1.52
GOLD::ChemScore	762	0.449	1.96	1.52

* SD和MUE的单位均为 - logKd

表1 各打分函数对800个蛋白-配体复合物的评分与实验结合常数间的相关性统计

当前主流的16种打分函数，即Discovery Studio中的五个打分函数（LigScore, PLP, PMF, Jain, LUDI），Sybyl中的五个打分函数（D-Score, PMF-Score, G-Score, ChemScore, F-Score），GOLD中的三个打分函数（GoldScore, ChemScore, ASP），Schrödinger中的GlideScore，以及来自学术界的DrugScore和X-Score，从对接能力和打分能力两方面进行了系统的评价[12]。所得结论（图4）不仅对用户在实际工作中选择合适的打分函数具有参考价值，还为打分函数未来的发展趋势提供了前瞻性的思路。

4.3 分子对接软件的评估

在基于结构的药物设计中，分子对接程序被广泛用于预测给定分子与其靶标的结合模式。目前各类商业与学术软件中可用的分子对接程序很多，因而非常有必要对这些程序的性能进行一次系统、客观的评价。该部分工作仍然建立在PDBbind核心集基础之上，我们系统考察评估了四种流行的分子对接程序，包括Glide, GOLD, LigandFit以及Surflex（相关工作即将发表）。通过调节各种分子对接程序的参数设置，在不同的分子构象采样水平上来考察这些对接程序。考察的标准是看各分子对接程序能否重现复合物晶体结构当中配体分子与蛋白分子的结合模式，以RMSD表示。重点考察了各分子对接程序在整个测试集上的性能以及对接精确度与采样时间的关系。我们在分子对接程序评估工作中所得到的结论揭示了目前主流分子对接程序的一些优缺点，不仅对用户在实际工作中选择合适的对接程序具有参考价值，还对分子

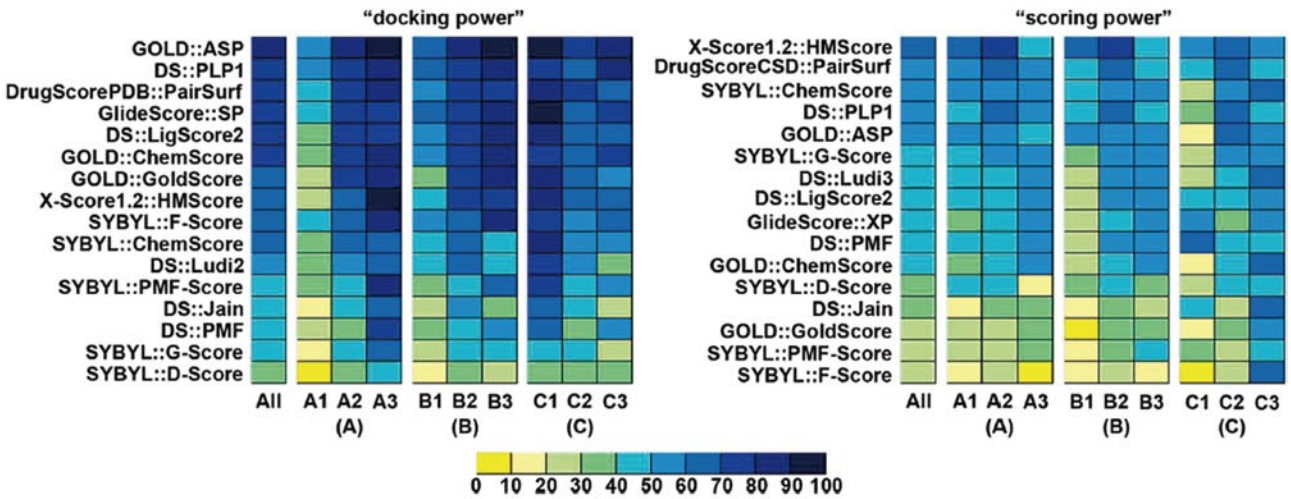


图4 各打分函数在PDBbind核心集及子测试集上的评估结果比较（图中All表示整个测试集；A，B，C分别表示按三种划分标准得到的子测试集，更详细说明请参阅文献[12]

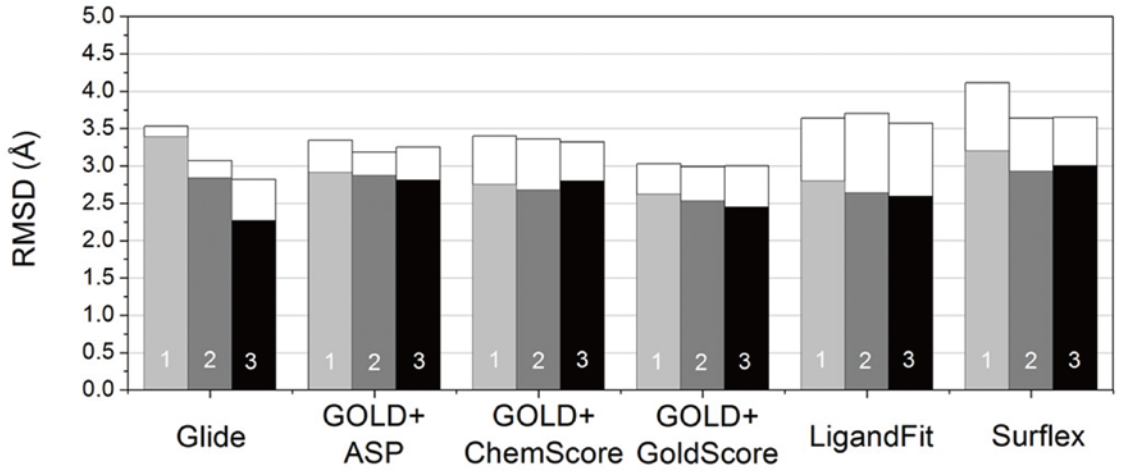


图5 各分子对接程序在不同计算水平（以数字表示）的性能比较（纵坐标表示测试集中各配体得分最高的对接构象与晶体结构构象间的RMSD平均值；灰度和白色填充分别表示采用配体的晶体结构和随机结构作为分子对接的初始构象的结果）

对接程序未来发展具有一定指导作用。

5. e-Science在计算机辅助药物设计领域发展方向的思考

随着生命科学的迅速发展，研究小分子的化学信息学，同研究生物大分子的生物信息学一起发展。在药物发现与设计方面，尤其在药物小分子与生物大分子相互作用方面，将会发挥越来越重要的作用[13]。可以预见，未来信息化领域（如生物信息学与化学信息学）的发展，对计算机辅助药物设计的进步至关重要。与之同步的是，e-Science的发展和完善也将呈现出自身的特点。

5.1 e-Science将朝开放、开源的方向发展

随着互联网与自由开源软件的发展，出现了越来越多的公众开放数据库和免费软件。如免费、开放的PDB数据库，不仅自身收录的生物大分子晶体数据不断增加，而且成为建立其他二级数据库的信息源。而化学信息学中，则有功能强大的开源Open Babel项目等[14]。

在软件的发行方式上，传统的软件收费模式将逐渐被面向服务和需求的收费模式所取代，即用户不再需要为软件付费，而只需要为自己获取的资源与服务付费。很多专业的商业药物设计软件不再靠单纯出售软件一次性收费，转而按年出租使用权License许可，并提供相应的技术支持。

5.2 e-Science将朝大型、广泛合作的方向发展

作为一种新型的科学研究环境，e-Science需要广泛的国际交流与合作。以中国科学院为例，既参加了国际PRAGMA亚太地区的网格合作，也参加了圣地亚哥超级计算中心的合作，还跟国际上其他一些机构建立了联系。中国科学院的e-Science计划在“七五”、“八五”、“九五”、“十五”期间一直大力支持科学数据库的建设，并已取得相当的成果。很多研究所之间发展了一些跨学科领域的合作[15]。

当前，计算机辅助药物设计领域的研究问题复杂，流程较多，数据量大，需要众多科研单位与企业的共同参与及合作。充分利用e-Science的网格计算环境可提高系统的速度以及数据库资源的可用性，通过建立一批示范性的面向大规模信息数据的专业服务网站、系统软件和数据库，能直接为网格中的成员乃至全世界的科研组织开放数据库并共享资源。

5.3 我国的e-Science建设仍处在起步阶段

中国的e-Science网格建设

已经初具一定规模，同时有中国国家网格、中国科学院网格等下属的众多网格项目在运行中，包括计算网格、数据网格和信息服务网格等。但是与e-Science的发源地英国、资助力度巨大的美国和网格应用广泛的欧盟相比，我国仍处于起步发展与壮大阶段。主要存在以下几点不足：自主知识产权的网格软件稳定性和成熟度亟待提高；另外网格尚未在各行各业得到推广，其应用的深度和广度有待提高；最后，网格环境作为联系众多分立节点的共享设施，其运行管理机制有待进一步探索。在这些网格的建设中，我们欣喜地看到新药发现应用网格、生物信息应用网格，高性能计算化学应用系统、药物研发网格等与计算机辅助药物设计相关的内容。

参考文献:

[1] 宋琳琳, e-Science发展情况简介[J].图书馆学研究,2005, 7:21-23.

[2] Taylor, J., National e-Science Centre. http://www.nesc.ac.uk/. July 2006.

[3] 国家自然科学基金委员会. 以网络为基础的科学活动环境研究2005年项目指南. http://www.nsfc.gov.cn/nsfc/cen/02/htmlcreated/2004jh/2005-07-08.htm. July 2006.

[4] Leaute, J.-B., Impact of Genomics and in Silico Related Technologics in the Drug Discovery Process[J]. Chin. J. Chem. 2003, 21:1241-1246.

[5] 徐筱杰，侯廷军，乔学斌等.计算机辅助药物分子设计[M].北京:化学工业出版社,2004年.

[6] Zhao, Y., Cheng, T., Wang, R., Automatic Perception of Organic Molecules Based on Essential Structural Information[J]. J. Chem. Inf. Model. 2007, 47, (4):1379-1385.

[7] Cheng, T., Zhao, Y., Li, X., Lin, F., Xu, Y., Zhang, X., Li, Y., Wang, R., Lai, L., Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge[J]. J. Chem. Inf. Model. 2007, 47, (6):2140-2148.

[8] Wang, R., Fang, X., Lu, Y., Yang, C. Y., Wang, S., The PDBbind Database: Methodologies and Updates[J]. J. Med. Chem. 2005, 48, (12): 4111-4119.

[9] Wang, R., Fang, X., Lu, Y., Wang, S., The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures[J]. J. Med. Chem. 2004, 47, (12):2977-2980.

[10] Wang, R., Lu, Y., Fang, X., Wang, S., An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of

800 Protein-Ligand Complexes[J]. J. Chem. Inf. Comput. Sci. 2004, 44, (6):2114-2125.
[11] Wang, R., Lu, Y., Wang, S., Comparative Evaluation of 11 Scoring Functions for Molecular Docking[J]. J. Med. Chem. 2003, 46, (12):2287-2303.
[12] Cheng, T., Li, X., Li, Y., Liu, Z., Wang, R., Comparative Assessment of Scoring Functions on a Diverse Test Set[J]. J. Chem. Inf. Model 2009, 49, (4):1079-1093.
[13] Boguski, M. S., Bioinformatics[J]. Curr. Opin. Genet. Dev. 1994, 4, (3):383-388.
[14] 乔圆圆,鹿涛,车云霞.化学信息学与生物信息学开放性比较[J]. Prog. Chem. 2007, 19, (4):624-632.
[15] 阎保平. 中国科学院：e-Science的建设与思考[J]. 互联网天地. 2004, 2:59-63.

收稿时间：2009年5月4日

基金项目：863目标导向类项目（2006AA02Z337）

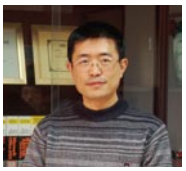
作者信息

刘志海

中国科学院上海有机化学研究所，硕士，研究实习员，研究方向为药物信息学。

程铁军

中国科学院上海有机化学研究所，博士，研究助理，研究方向为计算机、辅助药物设计。



王任小

中国科学院上海有机化学研究所，“百人计划”研究员，博士生导师，研究方向为化学生物学、计算机辅助药物设计。