支持药物虚拟筛选的数据管理工具

刘光远　南凯
中国科学院计算机网络信息中心，北京 100190

摘　要： 虚拟筛选利用计算机技术模拟药物筛选的过程，极大地提高了新药研发的效率，是新药研发的重要环节。虚拟筛选过程将产生大量复杂的中间数据，对这些数据的有效管理是下阶段新药研发的基础。本文介绍了一种药物虚拟筛选数据管理工具，对虚拟筛选过程产生的中间数据进行录入、分析、抽取、计算、持久化和展现，协助科研人员进行有效的科研决策。目前该工具已经在药物虚拟筛选中投入使用。

关键词： 药物发现；虚拟筛选；数据管理

A Data Management Tool for Drug Virtual Screening

Liu Guangyuan, Nan Kai
Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Virtual screening is an important method in drug discovery. Virtual screening models the process of drug screening by computer simulation technology, which will improve the efficiency of drug discovery. A large amount of complex data will be produced in virtual screening and the effective management of these data is the basis of next research in drug discovery. Described in this paper is a data management tool used in drug virtual screening. This tool fulfills entry, analysis, extraction, calculation, persistence and visualization of the data produced in virtual screening which assists experts and researchers to conduct more effective research decision. This data management tool is currently being used in drug virtual screening.

Keywords: Drug discovery; Virtual screening; Data management

1. 引言

药物筛选指通过规范化的实验手段从大量化合物中选择对某一特定作用靶点具有较高活性的化合物的过程，它是现代药物开发中检验和获取具有特定生理活性化合物的重要环节。组合化学的发展使人们形成了大规模合成和分离化合物的能力，也促使药物筛选成为发现先导化合物的主要手段。

虚拟筛选是药物筛选技术发展的一个重要方向。虚拟筛选是在计算机上模拟药物的筛选过程，对化合物的活性做出预测，进而对较有可能成为药物的化合物进行有针对性的实体筛选，这种方法将显著降低药物开发成本。据统计，虚拟筛选的介入，使新药研发的平均周期缩短了0.9年，直接研发费用平均降低了1.3亿美元[1]。

分子对接技术是实现药物虚拟筛选的主要方法[2]。分子对接是指配体分子与受体分子通过几何和能量匹配相互识别的过程。在虚拟筛选中，使用GASDOCK、DOCK、AUTODOCK等分子对接软件能有效地提高分子对接的效率[3]。分子对接软件基于对接分子相关数据及对接参数，利用高性能计算机进行虚拟筛选[4]。虚拟筛选中涉及大量复杂的科学数据，它们是开展药物虚拟筛选的基础，数据管理不当将对药物虚拟筛选乃至整个药物研发流程产生不利影响。

本文介绍一种支持药物虚拟筛选的数据管理工具，以有效地管理药物虚拟筛选中的大量复杂数据，保证药物研发工作的顺利进行。

2. 药物虚拟筛选的数据管理特点

一般来说，每次虚拟筛选都是针对特定的靶标分子，从数以百万计的化合物中筛选出与靶标分子结合最好的候选化合物进行实体筛选。整个虚拟筛选过程需要用到数百万个配体分子的结构和能量数据，每次对接过程都需要一组特定的对接参数，每次对接过程完成后都会产生对接日志数据，记录此次对接的详细信息，选择哪些候选化合物进行实体筛选正是基于这些日志数据。上述数据大多来自特定化合物数据库或由科研人员直接提供，这些数据（特别是对接日志数据）的管理和利用对虚拟筛选有十分重要的意义。

虽然对接日志数据记录了每次对接的详细信息，但科研人员通常只需要其中一部分关键信息，例如配体分子与靶标分子在各结合位点的几何和能量匹配情况，这些关键信息只占日志数据的小部分。如果一次虚拟筛选进行了100万次分子对接，那么就会产生100万个日志文件，而在一般情况下，存放日志数据的日志文件大小都在1~5MB，由科研人员从中逐一查找所需关键信息是不现实的。如果将科研人员需要的数据抽取出来，进行统一的组织，并提供相应的查询接口，将可使科研人员方便地获得需要的对接信息。不仅如此，如果进一步对抽取出的数据进行分析和计算（例如寻找对接能量匹配最好的一组结合位点），将对整个科研工作带来极大的便利。

虚拟筛选用到的分子描述数据是以二维形式记录的原子坐标、部分电荷和原子类型等信息。虽然这些数据是分子对接软件运行的基础，但科研人员并不能从从中获得对分子对接结果的直观认识。为了更好地支持领域专家进行科研决策，有必要对虚拟筛选涉及的数据进行有效的展现。通常仅靠现有数据无法实现直观的展现，需要科研人员提供分子三维结构图、分子式、NCI数据库编号、分子类药性等更多的参考数据。将已有数据和额外参考数据以有效的方式展现出来，并提供方便的查询接口，将使科

研人员对虚拟筛选有更直观的认识，进而提高药物研发的质量和进度。

在虚拟筛选中，即使某个配体分子对当前靶标的活性不高，但并不意味着对其他靶标没有活性。因此，配体分子的数据和参数可以重用于其他虚拟筛选研究，这就需要将这些数据存储起来供其他筛选使用。同时，科研人员需要对虚拟筛选的每次对接结果进行比较，选出最优的一组候选化合物进行实体筛选，有时也需要查看某个配体分子在不同虚拟筛选中的对接情况。这就有必要将所有对接日志数据持久化到数据库。

本文介绍的支持药物虚拟筛选的数据管理工具可以很好地满足数据抽取、数据分析、数据展现和数据存储的要求，实现对药物虚拟筛选数据的管理，提高药物虚拟筛选的效率。

3. 药物虚拟筛选数据管理工具的设计

药物虚拟筛选数据管理工具是B/S架构的应用。整个应用以虚拟筛选过程的相关数据为线索，所有数据都要经过数据录入、数据分析、数据抽取、基于数据的计算、数据持久化和数据展现等数据管理单元的处理。每个管理单元的输入数据都是前一个管理单元的产出数据，数据管理单元之间环环相扣，构成一套完整的科学数据管理流程。

药物虚拟筛选数据管理工具包括数据处理、数据持久化和数据展现三个主要模块。

数据处理模块负责对不同类型的数据进行分析，根据用户指定的模式对数据进行抽取和整理，然后基于特定的算法对抽取出的数据进行计算，将结果数据交给数据持久化模块。数据处理模块包括数据录入、数据分析抽取和数据计算处理三个子模块。

数据录入模块接受用户输入的多种类型的数据，不仅包括用户通过填写表单提交的字段，还包括用户上传的EXCEL、ZIP、JPG、PDBQT、GPF、DPF等格式的文件。数据录入模块对这些复杂数据进行预处理，以虚拟筛选实验为线索，将每次虚拟筛选涉及的分子信息、参数信息及所有对接日志组织为一个逻辑单位，交给数据分析抽取模块做进一步处理。

数据分析抽取模块按照用户指定的模式，利用正则表达式对结构化的文本数据实施分析和抽取。抽取出的数据一部分直接进

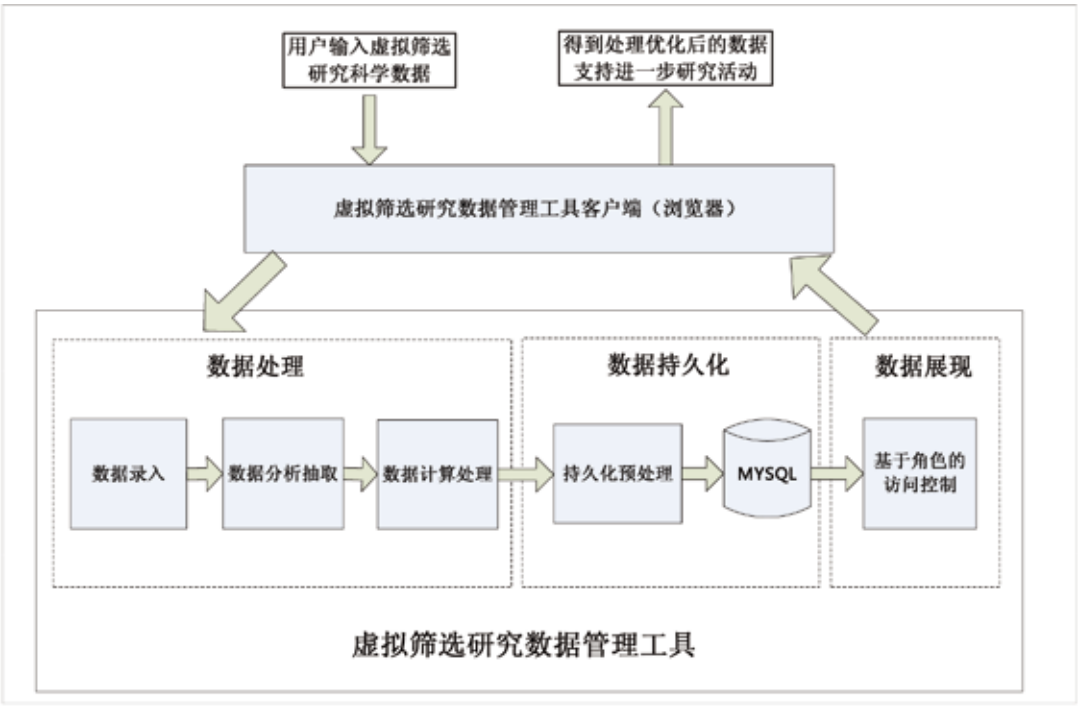

图1 药物虚拟筛选数据管理工具系统架构

行持久化，另一部分交给数据计算处理模块做进一步的计算。

数据计算处理模块基于抽取出的信息（如配体与受体间匹配程度的评价），根据特定搜索，排序算法，计算出每次对接的最佳匹配信息，将其存储到数据库系统。

数据持久化模块将数据分析抽取模块和数据计算处理模块产生的数据进行持久化预处理，并存储到数据库系统，使之更好地支持数据查询。

数据展现模块以Web方式使用基于角色的访问控制策略，实现数据展现和数据共享，协助科研人员开展进一步的工作。

4. 实现与性能评价

药物虚拟筛选数据管理工具的实现使用了多种框架和工具，极大地提高了数据处理的性能，成为有别于同类数据管理工具的技术亮点。以下着重在数据抽取和数据持久化两个方面，对药物虚拟筛选数据管理工具的性能做出评价。

4.1 数据抽取

虚拟筛选过程的相关数据多为结构化的文本数据，采用正则表达式可以提高数据抽取的效率。数据抽取采用Apache-Jakarta-ORO[5]，这是一套处理文本数据的正则表达式工具，提供文本匹配、替换、分割等功能，也是基于Java语言最全面和最优化的正则表达式API之一。数据抽取依据的正则表达式保存在配置文件中，用户可以根据需要改变抽取规则。药物虚拟筛选数据管理工具实现了高效灵活的数据抽取机制。

表1列出了使用Apache-Jakarta-ORO对不同数量不同大小的文件进行抽取所用平均时间的对比。

一般情况下，药物虚拟筛选研究用到的数据文件最大不会超过5MB。如表1所示，对1000个大小为5MB的数据文件进行抽取，平均时间仅为2秒。事实证明，采用Apache-Jakarta-ORO可以有效地提高数据抽取的效率。

表1 数据抽取实验结果

文件大小 /kB	文件数量	平均抽取时间 /ms
50	1000	47
500	1000	266
5000	1000	2063

一部分虚拟筛选数据经过抽取和整理后直接进行持久化，并通过数据展现模块向用户提供查询功能。图2是分子对接日志数据的抽取结果，科研人员可以从中清楚地了解每个分子对接流程的结果，远比直接查看对接日志方便直观。

4.2 数据持久化

经过前期处理的虚拟筛选数据需要高效稳定的支持数据库存储和查询功能，为此数据管理工具采用Hibernate[6]实现数据持久化。Hibernate是一款优秀的对象持久层框架，它通过对JDBC的封

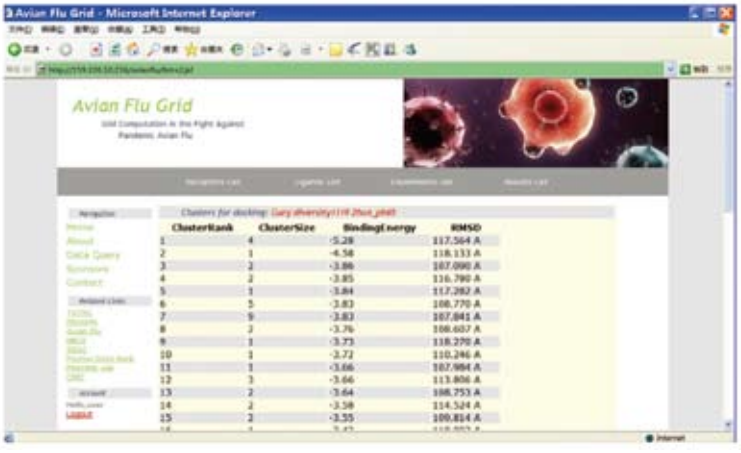

图2 抽取结果的展现

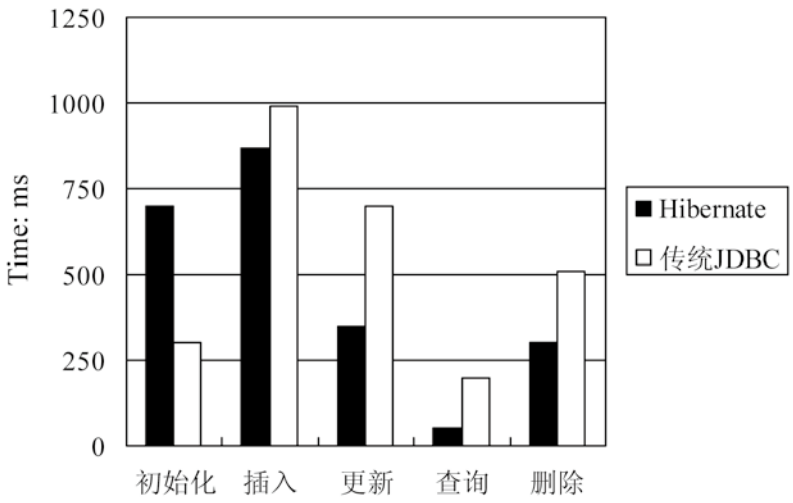

图3 Hibernate与JDBC性能对比

装实现数据库操作的透明。

与单纯使用JDBC相比，Hibernate在性能和可操作性方面均有优势。性能方面，Hibernate使用连接池技术，省去了创建物理连接的过程，提高了数据库操作的速度，同时Hibernate优化了数据库操作接口，保证了数据库操作的稳定性。可操作性方面，Hibernate基于对象关系映射思想，使开发人员对数据库的操作由传统的直接对数据库表的操作变成了对Java对象的操作。如此将使开发人员更加关注业务，降低了代码的复杂度，提高了代码的健壮性。

图3展示了使用Hibernate3与传统的JDBC进行不同数据库操作的平均执行时间的对比。由于Hibernate是对JDBC的封装，所以Hibernate建立数据库连接等初始化操作的时间要略多于传统JDBC，但是进行数据库插入、删除、更新等操作的时间要明显少于传统的JDBC。特别是在药物虚拟筛选中，科研人员85%的数据操作行为是数据查询操作。图中可以看出，使用Hibernate进行数据查询的平均时间是50ms，而使用传统JDBC进行查询需要200ms，前者耗时仅为后者的1/4。

5. 结论

药物虚拟筛选数据管理工具通过对虚拟筛选过程相关数据的录入、分析、抽取、处理、持久化和展现，解决了虚拟筛选中关联数据无组织，数据利用率低，无统一持久化策略，无高效统一的数据查询接口等实际问题，为进一步的实体筛选实验奠定了重要基础，为科研人员的研究工作带来极大便利，提高了药物研发的整体效率。目前，药物虚拟筛选数据管理工具已经在禽流感药物的研究中投入使用，作为构建禽流感药物发现数据网格的基础系统，提高了研究效率并获得科研人员的好评。

药物虚拟筛选数据管理工具现阶段只是实现了基本的功能，满足了科研人员当前的需求，仍然需要在功能性和通用性上进行进一步的改进，以求便更好地为科研活动提供支持。进一步，将在药物虚拟筛选数据管理工具中加入数据挖掘模块，基于持久化的虚拟筛选数据进行知识发现，届时将更有效地支持科研决策，并为新药研发带来切实的效益。

参考文献：

[1] 李洪林, 沈建华，等. 虚拟筛选与新药发现[J]. 生命科学, 2005, 17(2): 125−131.

[2] Campbell McInnes. Virtual screening strategies in drug discovery. Current Opinion in Chemical Biology, 2007, 11: 494−502.

[3] Kitchen D B, Decornez H, Furr J R and Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov, 2004, 3: 935−949.

[4] Warren G L, Andrews C W, Capelli A M, Clarke B, LaLonde J, Lambert M H, Lindvall M, Nevins N, Semus S F and Senger S et al. A Critical Assessment of Docking Programs and Scoring Functions. J Med Chem, 2006, 49: 5912.

[5] Apache. http://jakarta.apache.org/oro/, 2008.

[6] Hibernate. http://www.hibernate.org/, 2008.

收稿时间:2008年10月28日

作者信息

刘光远

中国科学院计算机网络信息中心，硕士研究生，主要研究方向为数据网格、协同计算和e-Science应用。

南凯

中国科学院计算机网络信息中心，研究员，主要研究方向为网格与协同计算、数据网格等。