· 论著 ·

网上免费医学生物学数据库指南的建立

景 霞△,张其鹏,国强华,卢 铭,朱晓华,石 磊,芮 伟,尚 彤
(北京大学心血管研究所医学生物信息室,北京 100083)

[关键词] 医学数据库;生物学数据库;数据库指南;医学生物信息学

[摘要] 目的:建立一个网上免费的医学、生物学数据库的查询指南(directory of biomedical databases,DBD),方便研究者查询相关数据库。改变目前医学、生物学数据库查询不便的现状,推动医学、生物学的发展。方法:以PubMed和Google作为主要检索工具,制定检索策略,根据收录的标准筛选找到的数据库,入选项管理平台采用Microsoft SQL-Server 2000,利用Macromedia Dreamwaver MX设计DBD的Web发布页面,ASP(active server pages)技术处理用户提交的关键词和评分值。结果:目前收集到医学、生物学数据库66类1258个,中英文版同时发布(http://cmbi.bjmu.edu.cn/DBList/index.htm,http://cmbi.bjmu.edu.cn/DBList/index_en.htm),用户可以通过分类、关键词和入选项名称字顺查询相关数据库,并建立基于内容的数据库评分系统和数据库点击记录。使页面简洁易用。结论:DBD为建立网上核心医学、生物学数据库的评价体系奠定了基础,DBD将有助于数据库在未来医学、生物学研究中发挥强大的作用。

[中图分类号] G353.21:R [文献标识码] A [文章编号] 1671-167X(2004)03-0322-05

Construction of directory for biomedical databases on INTERNET

JING Xia△, ZHANG Qi-peng, GUO Qiang-hua, LU Ming, ZHU Xiao-hua, SHI Lei, RUI wei, SHANG Tong
(Biomedical Informatics Laboratory, Peking University Institute of Cardiovascular Sciences, Beijing 100083, China)

KEY WORDS Medical databases; Biological databases; Directory of databases; Biomedical informatics

SUMMARY **Objective**: To construct a global directory of biomedical databases(DBD), which can be used free of charge on INTERNET. It will be convenient for researchers to find out related databases quickly, easily and accurately by using DBD since there are not enough useful tools for database retrieval. Biomedical databases will be an accelerator in development of biomedicine with the help of DBD. **Methods**: PubMed and Google were main tools for searching related databases. Proper search strategy with rigorous indexing rules helped us to filter databases. The database management system was Microsoft SQL-Server 2000. The web pages of DBD were designed with Macromedia Dreamwaver MX. ASP (active server pages) technology was used to deal with the key words and scores sent by users. **Results**: There were 66 subjects and 1 258 databases in DBD at this time. We released the Chinese and English versions of DBD on the INTERNET at the same time(http://cmbi.bjmu.edu.cn/DBList/index.htm,http://cmbi.bjmu.edu.cn/DBList/index_en.htm). Score system was also established to evaluate the content of the indexed databases. Users can search DBD by subjects,key words and alphabetic databases' names easily. **Conclusion**: DBD has laid the primary foundation for further core biomedical database evaluation system. DBD, as a useful tool for biomedical database retrieval, will be of great aid to users since databases have played a more and more important role in the biomedical research.
(*J Peking Univ [Health Sci]*, 2004,36:322-326)

随着人类基因组测序工作的完成[1],医学、生物学的发展进入了一个崭新的时代。越来越多的医学、生物学研究将以序列数据为基础展开。如此丰富的数据源对于发现一些疾病的新的诊断、治疗、预防方法,认识一些疾病的机制和各种生命现象的生物学基础至关重要,而由此带给临床研究和实践的变化也将是前所未有的[2,3]。高效地管理、发布,方便地查询这些天文数字级的序列数据及近百年来医学研究积累的大量宝贵资料,只能通过数据库来完成,各种生物学数据库正在成为生物学家日常研究的必备工具[4]。随着生命科学工作者的工作重心从了解序列结构转向对序列结构功能、生命意义的阐释[5],数据库(尤其是专业数据库)将是整个医学界和生命研究领域的加速器。目前数目众多的与医

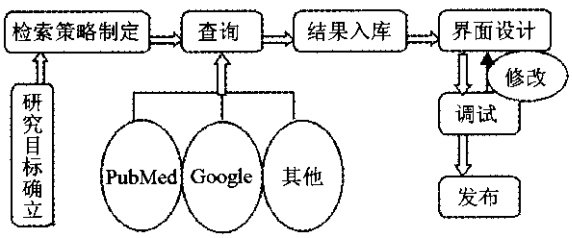
基金项目:国家重点基础研究发展规划项目基金(G2000056907,G1998051015)资助 Supported by the Major State Basic Research Development Program of People's Republic of China(G2000056907,G1998051015)

△Corresponding author email, xiajing@bjmu.edu.cn

学、生物学有关的数据库散布在世界的各个角落，不便于查找，至今未见到查询医学、生物学数据库的专用工具。我们的研究旨在将收集到的常用的、网上免费的医学和生物学数据库按学科分类，并且提供基于关键词的查询，制作一个关于网上免费的医学、生物学数据库的指南（Directory of Biomedical Databases，DBD），以方便相关学科科研工作者迅速，准确地找到与自己专业密切相关的数据库。我们的工作发布在中国医学生物信息网（CMBI）－数据库－医学生物学数据库指南（http://cmbi.bjmu.edu.cn/DBList/index.htm），并进行网上的实时更新。

1　材料与方法

1.1　研究流程



1.2　检索方法

数据库的检索主要依据 PuMed（目前公认的生物医学文献收录最全、最具影响力的文献数据库 http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html）。PubMed 中的检索策略为（"databases"[MeSH Major Topic] AND ("2001"[PDat]: "2004/02/28"[PDat])) NOT ("Nucleic Acids Res"[Journal] AND ("2001"[PDat]: "2004/02/28"[PDat]))。限制发表文献在 2001 年至今，选那些将 database 作为主 MeSH 词的文献，而且去除发表在 *Nucleic Acids Research*（《核酸研究》）上的分子生物学数据库。查到相应的数据库后，或直接将网址记录，或根据正式名称查找相应的数据库，浏览后决定收录否。将检索策略存入 Cubby（http://www.nlm.nih.gov/bsd/pubmed_tutorial/m4101.html），随时更新数据。

考虑到权威性和商业因素的影响（http://www.google.com/technology/whyuse.html），选用 Google 作为原始搜索引擎，分别采用分类查询、高级查询，级联放大方式搜索（http://www.google.com/technology/index.html）。进入 Google 分类页面，选 Health 类下的 40 个子类首先运用 allintitle: database(data base, data bank) 找到核心数据库，遴选后入选。然后在各子类以 database(data base, data bank) 为关键词搜索，所得结果如超过 500 条则只在前 500 条中遴选（因 Google 是根据网页的重要性和浏览量的多少排列结果的）。再选 Science 类下的 Biology, Chemistry, Method & Technology, Biophysics, Crystallography, Medical Physics 子类，同样方法搜索后遴选。

在 Google 的高级搜索页面下键入关键词通过组配，限制日期（过去 6 个月），限制是否出现在题目或页面的 URL，选中 Filter using 框进行查询。参照医学名词审定委员会审定的《医学名词》、《生理学名词》、《生物化学与生物物理学名词》、《细胞生物学名词》选定约 150 个关键词，涉及医学、生物学的各方面（包括各科主要病名、主要生理病理机能名称）。

其他方法包括平时积累的、专业人员推荐的数据库遴选后入选，还有一些针对医学引擎的深层次搜索，如：俄勒冈医科大学的 CliniWeb International、爱荷华大学的 Hardin Meta Directory of Internet Health Source、麦吉尔大学健康图书馆的数据库和电子资源列表、HealthWeb、BioMedNet（Elsevier's Portal to life Science）和 MedExplorer 等，通过医学引擎查询得到的数据库和网络资源中列出的数据库浏览后根据入选标准摘录出来，这是另外一种非主要的方式。记录入选数据库的名称、URL 地址、简介，根据数据库的内容给每个数据库 3～5 个关键词。

1.3　分类方法

收集到的数据库的学科分类基本以《中国图书馆分类法（第四版）》（简称《中图法》）为依据，但根据搜集到的医学生物学数据库的实际分布做了较大的改动。主要分为九大部分：基础医学、预防医学与卫生学、药学、内科学、外科学、其他临床各科、肿瘤学、中医学、其他。与生物医学有关的其他学科（如伦理学、法律等）收入"其他"部分。并收录一些重要网站。

1.4　DBD 的管理和发布

DBD 的管理平台：Microsoft SQL-Server 2000。DBD Web 发布设计环境：Macromedia Dreamwaver MX，Microsoft Internet Information Service（IIS）。DBD 包括 3 个表：主表、学科表和评分表。主表的属性包括数据库的名称、URL 地址、数据库的简介、关键词、数据库的学科分类归属，唯一标识码和点击计数。学科表包括所有分类用的中、英文学科名称、学科标识码和关于学科的描述。评分表包括所收集的数据库的标识码和预设评分值。运用 ASP 技术处理

用户提交的关键词和评分值。Web 页面提供分类查询、关键词查询和入选项名称字顺查询三种入口。

1.5 标准

入选标准：(1)内容准确、可信、有权威性；(2)近一年有更新的、知名大学或研究所创办的、严格意义上的数据库；(3)专业性强；(4)有 Public access；(5)易用且界面友好[6](Criteria for assessing the quality of health information on the Internet. http://hitiweb.mitretek.org/docs/criteria.pdf)。不收录标准：科普类和面向患者服务的数据库；无 Public access 只供内部使用的数据库；商业性数据库。

评分是为了今后更进一步完善 DBD 而设立的，旨在通过用户对所查询到的数据库评分(5 分制)而使评分成为进一步同行评议该数据库的指标。我们的最终目标是建立一个真正的、有同行评议结果的、权威的数据库指南。评分标准：5 分，内容可信、正确，学术性强，利用价值高，近半年有更新；4 分，内容可信、正确，学术性较强，有利用价值，近一年有更新；3 分，内容基本正确，有学术性，近两年有更新；2 分，内容有明显纰漏，学术水平差；1 分，无参考价值，应该从学术数据库收集中被剔除。

此外还建立了基于每个数据库的点击记录，可以配合评分记录成为进一步评价数据库的依据。

2 结果

到目前为止，共收录医学生物学数据库 1 258 个，归入 9 大类 66 小类(学科分类)中。其中基础医学相关 517 个，卫生学相关数据库 83 个，药学相关 89 个，临床医学相关 303 个，肿瘤学 39 个，中医学相关 16 个，其他 211 个(包括重要网站 43 个)。

我们采用中英文双语方式在网络上同时发布，中文版：http://cmbi.bjmu.edu.cn/DBList/index.htm；英文版：http://cmbi.bjmu.edu.cn/DBList/index_en.htm。Web 发布方式如图 1 所示，将分类查询、关键词查询和入选项名称字顺查询列在首页，直接选择学科、输入关键词或选择相应数据库名称首字母查询即可，页面简洁易用。查询结果以表格形式显示(图 2)，直接点击查询到的数据库名称即可链接到相应数据库。名称后有一个评分值，直接单击想给的评分即可。结果表下方有关于评分方法和标准的备注。

基础医学中以生物化学分子生物学数据库数量最突出，有 315 个。此外还有《核酸研究》[7]收录的 548 个分子生物学数据库(http://nar.oupjournals.org/cgi/content-nw/full/32/suppl__1/D3/GKH143TB1)。这是《核酸研究》第 11 次关于生物学数据库的特辑，收录全球范围的最有影响力的分子生物学数据库。EMBL-EBI(European Bioinformatics Institute，全球最有影响的生物信息学研究所，维护着全球最重要的核酸和蛋白质数据库)的 Service Overview 页面(http://www.ebi.ac.uk/services/)将 EBI 的所有数据库、工具包按类列出。其中的 SRS(Sequence Retrieval System，http://srs.ebi.ac.uk/)包括约 480 个序列及文献数据库，是全球运用最广的序列数据库查询系统，北京大学有其镜像站点(http://srs.pku.edu.cn/)，目前收录了 138 个数据库。随着序列数据的大量出现，相应的分析工具也迅速地发展起来，我们收录了 42 个关于序列和蛋白质结构的分析、预测工具。

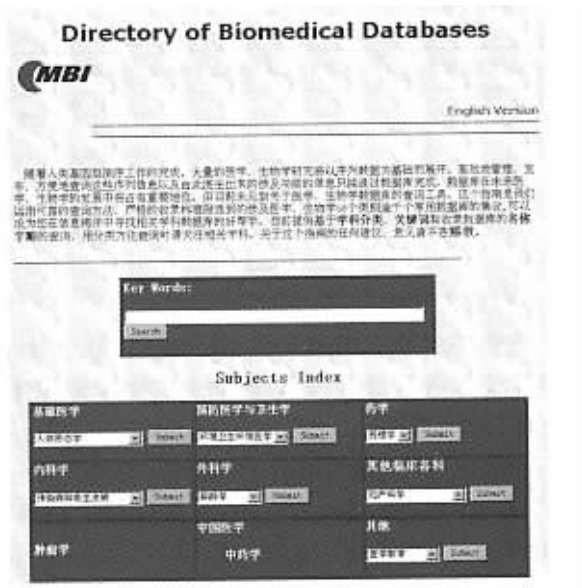


图 1 DBD 用户界面首页

Figure 1 Homepage of Directory of Biomedical Databases

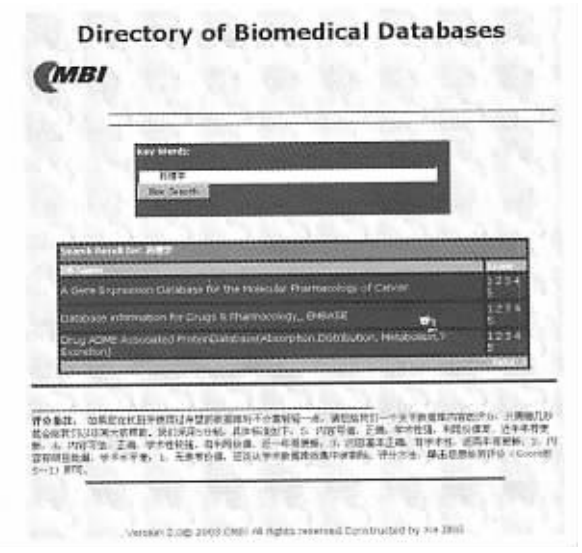


图 2 DBD 查询结果页面

Figure 2 Search result interface of the DBD

临床医学数据库中放射医学和病理学的影像数据库较有特色,这一特点与影像和图像在学科中的重要地位密不可分,它们多有根据系统、症状和疾病名称编排的索引或检索入口,便于查找、交流,而且直观。医学教育部分较有特色的是加拿大麦吉尔大学的虚拟听诊项目(McGill Virtual Stethoscope Project,http://sprojects.mmi.mcgill.ca/mvs/),在这个项目中有关于听诊(正常和病理状态)的教育资源和自测习题集,对于实习资源有限的医学生而言,这种网上的实践与测试机会实用而且珍贵。

临床医学数据库中的 DocDat(http://www.lshtm.ac.uk/docdat/page.php?t=index)[8],收集了英国的百余个临床数据库,对其收录的每一个数据库均有简短的描述和评价。用户可以通过系统、发病机制、治疗方法、年龄组和国家的限制缩小查询范围。此外 DocDat 还提供关于主题数据库的查询、浏览所有入选数据库的 List 的服务。

美国亚利桑那大学健康图书馆的数据库列表(http://www.ahsl.arizona.edu/databases/)列出了共约 330 个数据库,80%以上可以利用公共接口访问,但所收录的不全是严格意义上的数据库,有分类(104 类)和全部列出两种方式发布,每个数据库有简短介绍和来源,并标识有否公共访问入口。

中国创建的医学、生物学相关数据库目前收录的只有 33 个(其中台湾 4 个,香港 2 个),约占收录总数的 3%,主要集中在中医中药库,有 11 个。其中中国医学生物信息网(CMBI)创建、参与创建、维护的数据库 8 个。在分子生物学领域有一席之地的只有天津大学创办的基因组 Z 曲线数据库(Z Curve Database of Genomes, http://tubic.tju.edu.cn/zcurve/)[9]。

3　讨论

3.1　DBD 建立的意义

21 世纪医学生物信息学将会给整个医学生物学领域带来全新的发展机遇,而数据是医学生物信息学的基础,数据库的建立是核心,而且是未来数字化医学时代的重要基础工程[10]。数据库的高效管理、发布、查询功能是其他工具无法比拟的,也是管理指数级增长数据的理想工具。关于疾病、健康的众多诊、治、防方法及机制研究的突破都将有赖于数据库管理的大量数据。数据库在未来生命科学的发展中将是一个必不可少的基础工具。目前国际上约有 30 万不同学科/领域的数据库和网站,其中医学、生物学数据库约占 3%～5%[10]。如此数目众多的数据库如果利用好将会极大地推动医学、生物学的发展。但是缺乏针对这些分散数据库的专用、有效查询工具,就不可避免地限制了数据库的广泛运用。虽然目前有一些专用的医学引擎,但是我们在测试后发现其查询数据库的能力有限,查询结果中冗余太多。DBD 的建立在一定程度上解决了这个问题,通过 DBD 用户可以方便地找到医学生物学常用数据库。DBD 收录数据库数目较多,学科分类涉及面广,收录标准严格,页面简洁易用。我们最终的目标是将 DBD 建成一个关于全球医学、生物学数据库的权威性指南。目前已经建立了一套基于内容的 5 分制投票体系,并记录每个数据库的被点击情况。下一步的工作将以此为基础,再请相关领域专家评价,把几个指标加权后的得分作为评价数据库影响力的指标,同时推出关于数据库的简介和评论,使用户在选择数据库的时候有更详细的背景资料。DBD 为建立网上核心医学、生物学数据库的评价体系奠定了基础,是医学、生物学研究的好帮手,它将成为更好地利用数据库的一个必要工具,也是医学生物信息学的一个基础工程。

3.2　数据库收录过程中所碰到的问题

数据库收录尚没有一个公知公认的标准可以遵循。采用 PubMed 搜索保证了重要数据库的收录,PubMed 收录期刊标准严格,同时传统媒体的严格同行评议过程也是一个被利用的间接有利条件,但同时也有其局限性,这一切是在假定重要数据库建好后会有论文发表的前提下。这样 Google 就成为 PubMed 搜索后有力的补充。即使是假定,也是很有依据的,那些重要的学术数据库的建立是为了相应领域研究者的使用,对于学术领域而言,得到承认、能有机会被同行了解的最重要、最普遍的方式是在传统媒体发表论文。利用 PubMed 和 Google 可以保证最重要的数据库没有被遗漏。

目前学科分类的依据是《中图法》(第四版),而《中图法》是图书馆、资料馆对印刷版书籍和非书资料进行分类的依据[11],是否完全适合网络上资料的分类还有待商榷。目前对于网络资料没有能被普遍承认的分类依据,《中图法》毕竟是以科学分类和知识分类为基础的,我们的做法是采用《中图法》,并根据实际情况作适当的调整。由于分类依循的标准内容有交叉,所以用户使用分类查询时一定要关注相关学科。

3.3　目前数据库建立中存在的问题和展望

在收录的过程中我们发现目前的医学、生物学数据库主要存在以下问题:(1)分布不均。基础医学相关(包括生物学)数据库的发展遥遥领先,基础医学中又以生物化学分子生物学数据库的发展为甚,从数据库的数量分布就可以清楚地看出这一点;相对而言临

床数据库数量少,不成规模,而且几乎均以静态数据、以公知公认的知识发布为主,只是改变了传播的媒介;序列数据库多,涉及功能的、系统的数据库少,以序列数据为基础、解决临床实际问题的数据库少;国际上数据库的发展轰轰烈烈,中国医学、生物学数据库所占的比例仅为 3%,国家关于数据库构建的系统工程有待进一步重视加强。实际上只有临床医学才是直接为人类健康服务的,其他一些基础医学、生物学的发展只是临床医学进一步发展可利用的基础和工具,临床数据库将更直接地推动医学的发展。(2)缺乏标准。数据库建立时缺乏统一的数据模式、界面,没有公认的规范、标准可以依循。不同的数据库采用不同的数据库管理系统,而且不提供标准的数据接入方式[4]。分散在各种学术机构、医疗机构、制药、生物技术公司和各种公共数据库的数据由于各自独立,交互性差,无法做到统一查询,致使其利用效率低下,远未达到其应该的程度。由于缺乏标准在技术上就使得整合面临大量的问题,当然整合困难不仅仅体现在技术层面[4]。缺乏公认的规范、标准,数据库的学术水平就难以用一致的标准衡量。建立医学生物学数据库的定量评估与监测系统是一个重要方向。此外在查询收录数据库的过程中,我们发现国内几乎没有关于最新版教科书的数据库,国外有一些但是也不多,最新版教科书对生物医学相关人员的继续教育、工作参考有重要意义,是生物医学工作者毕业后教育的主要组成部分之一,可以成为生物医学相关人员计算机桌面上的常备工具,而且随着网络的进一步普及,网络版教科书查询方便、经济实用的特点是传统印刷版图书无法比拟的。

(本文提到的链接在 2004 年 4 月 2 日测试均可用。)

参考文献

1 Pennisi E. Human Genome: Reaching their goal early, sequencing labs celebrate[J]. Science, 2003, 300:409
2 Baxevanis AD. The molecular biology database collection: 2002 update[J]. Nucleic Res, 2002,30:1-12
3 Nadon R, Sladek R. Bioinformatics in research and clinical practice[J]. Clin Invest Med, 2003,26:75-77
4 Stein LD. Integrating biological databases[J]. Nature reviews/genetics, 2003, 4: 337-345
5 Miller CJ, Attwood TK. Bioinformatics goes back to the future [J]. Nature Reviews Molecular Cell Biology, 2003,4:157-162
6 张咏. 网络信息资源评价的方法及指标[J]. 图书情报工作,2001,45:25-29
7 Galperin MY. The molecular biology database collection:2004 update[J]. Nucleic Res,2004,32:Database issue D3-D22
8 Black N, Payne M. Improving the use of clinical databases[J]. BMJ,2002,324:1194
9 Zhang CT, Zhang R, Ou HY. The Z curve database: a graphic representation of genome sequences[J]. Bioinformatics,2003, 19:593-599
10 尚彤,国强华,景霞. 常用医学生物信息学数据库[M]. 北京:北京大学医学出版社,2003
11 中国图书馆分类法编委会.《中国图书馆分类法》第 4 版使用手册[M]. 北京:北京图书馆出版社, 2002.15

(2003-10-08 收稿)
(本文编辑:赵 波)

· 读者 · 作者 · 编者 ·

《北京大学学报(医学版)》特邀编委王晓东教授当选美国国家科学院院士

据新华网华盛顿 4 月 23 日电,美国国家科学院在本周举行的年度大会上选出了新一届院士,美籍华人科学家王晓东名列其中。41 岁的王晓东是改革开放以来中国大陆 20 多万留美人员中迄今获得美国科学院院士的第一人。

王晓东是美籍华人生命科学家中的佼佼者。目前他在得克萨斯大学西南医学院任教授,并在美国著名的霍华德·休斯研究所任研究员。王晓东现阶段的研究重点是细胞凋亡的生物化学途径和生物化学过程,曾发现了一些在细胞凋亡中起关键作用的蛋白质,他的研究对于寻找癌症和老年性痴呆等疾病的新疗法有望起到重要作用。王晓东教授自 2000 年 12 月起担任本刊第 7 届编辑委员会特邀编委。

(本刊编辑部)