生物多样性　第 8 卷,第 4 期,2000 年 11 月
CHINESE BIODIVERSITY　8 (4):397 ~ 404, November,2000

生物多样性信息学:一个正在兴起的新方向及其关键技术*

钟　扬[1]　张　亮[2]　任文伟[1]　陈家宽[1]
1(复旦大学生命科学学院、生物多样性与生态工程教育部重点实验室,　上海　200433)
2(复旦大学计算机科学系,　上海　200433)

摘　要　生物多样性科学和生物信息学是生命科学中两个极为重要也是十分活跃的交叉学科,生物多样性信息学则是目前正在兴起的一个新方向,其发展必将进一步深化信息技术在生物多样性研究中的应用。本文简要介绍了国内外该领域的主要目标与进展,讨论了有关的关键技术(如数据库间的互操作与数字图书馆),并列出了两个原型系统(Species 2000 和 GBIF)和其他相关系统的网址。

关键词　生物多样性,生物信息学,生物多样性信息学,互操作性,数字图书馆

Biodiversity Informatics: a new direction of bioinformatics and biodiversity science and related key techniques/ZHONG Yang[1)], ZHANG Liang[2)], REN Wen_Wei[1)], CHEN Jia_Kuan[2)]

Abstract　Biodiversity science and bioinformatics are two of the most important and active fields in today's life sciences. Currently, biodiversity informatics, a new interacting direction of the two fields, has risen. Its development will deepen the application of information technology in biodiversity studies. This paper introduces the major objectives and advances of biodiversity informatics as well as related key techniques, such as database interoperability and digital library. In addition, two prototype systems, i. e., Species 2000 and GBIF are also introduced briefly and the URLs of other related systems are listed.

Key words　biodiversity, bioinformatics, biodiversity informatics, interoperability, digital library

Author's address　1) Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai　200433
2) Department of Computer Science, Fudan University, Shanghai　200433

1　引言

建立生物多样性信息系统一直是生物多样性研究的一个重要组成部分,其作用早已为国内外大多数研究者和决策者所认识(Bisby et al.,1993;"中国生物多样性保护行动计划"总报告编写组,1994;May,1995;Burley et al.,1997)。然而,在现阶段生物多样性信息系统的开发过程中,仍然面临如下几个方面的问题:1)如何从浩若烟海的信息资源中快速有效地发掘出生物多样性研究信息?2)已找到的信息可能来源不同,它们之间是否存在可比性?3)生物多样性信息系统应如何整合其他系统(如分子生物学数据库和非生物数据库)中的有用信息?解决这些问题,必须依靠生物多样性科学家、生物信息学家以及

* 国家自然科学基金重大项目"中国关键地区生物多样性保育的研究"(批准号:39893360)资助
收稿日期:2000 - 10 - 15;修改稿收到日期:2000 - 10 - 25
钟扬 e_mail 地址:yangzhong@fudan.edu.cn

计算机科学家的通力协作,将大量新的信息技术应用于生物多样性信息管理,并致力于发展新的交叉学科生长点。生物多样性信息学(biodiversity informatics)就是目前正在兴起的一个新方向。

2000 年 9 月 29 日出版的 *Science* 杂志在以" Bioinformatics for Biodiversity"为题的专栏中发表了一组重要文章与报道,包括:" 全球生物多样性图谱 "(Wilson,2000)、" 生物多样性数字化 "(Sugden & Pennisi,2000)、" 分类学的复苏 "(Pennisi,2000)、" 静悄悄的革命:生物多样性信息学与 Internet "(Bisby,2000)、" 生物多样性数据库间的互操作性 "(Edwards et al.,2000) 和" 化石数据库上网 "(Kaiser,2000)等。该专栏全面而系统地评述了全球范围内生物多样性信息学的进展,并展望了该方向的发展前景。例如,Wilson(2000)认为构建全球生物多样性图谱(Global biodiversity map)的工作应当象" 人类基因组计划(HGP)"所构建的人类基因图谱一样,受到国际社会的普遍关注与支持。如果可能的话,项目应在 10 ~ 20 年内完成,而经费资助额应达 50 亿美元。本文作者之一在即将付印的《生物信息学概论》一书中也撰写了" 整合生物学信息系统 "的章节,其中涉及生物多样性信息学的内容①。本文将根据上述文献及其他资料,简要介绍国内外该领域的主要目标与进展,讨论有关的关键技术,例如数据库间的互操作性(database interoperability)与数字图书馆(digital library)。此外,介绍两个原型系统并列出了部分相关系统(网站)的网址,供读者参考。

2 生物多样性信息学的主要目标

目前,生物多样性的科学研究、知识普及和政策制定等等都与现代信息技术的发展密切相关。这是因为:1)生物多样性工作者遍布全球,几乎每个国家甚至岛屿都在进行生物多样性研究。科学家们需要快速、准确地交换数据资料;2)有关生物多样性的全球性事件与我们每个人相关。制定生物多样性保护的优先策略更需要综合整体的信息并考虑全人类共同关心的问题;3)现有的信息系统尚不能满足日益增长的需求,特别需要针对生物多样性研究中产生的一些重大问题,开发有效的工具和技术支撑条件。例如,涉及生物群落区(biomes)、生态系统(ecosystem)、" 热点(hotspot) "和遗传侵蚀(genetic erosion)的研究迫切需要整合来自各相关区域的研究者、工作组或研究所收集的信息。

因而,生物多样性信息学的中心目标是:开发具互操作性与知识综合能力的信息系统,使得广泛分布的独立系统能嵌入全球生物多样性知识结构体系之中(Bisby,2000)。Species 2000 和全球生物多样性信息系统(GBIF)就是两个正在逐步发展的原型系统。

3 数据库间的互操作

随着计算机技术的普及与发展,各种信息电子化程度迅速增加,信息系统层出不穷。人们已普遍感到实现不同数据库间的互操作(interoperation)的必要性。例如,通过 Internet 对不同国家、地区和部门间的生物多样性数据库进行互操作,或者更广泛地,对不同类型的信息系统(如分子生物学数据库和 GIS 系统等)进行互操作,可以极大地提高生物多

① 钟扬,徐安龙,赵斌(主编),2000. 生物信息学概论. 北京:科学出版社(印刷中)

样性科学家的工作效率。事实上,生物信息系统间的互操作性问题已经引起了生物学家和计算机科学家的共同兴趣(Blake et al.,1994;Davidson et al.,1995;Gingras et al.,1997;Zhong et al.,1999)。

目前,实现异构数据库之间互操作的技术途径主要有四种:1)超文本漫游(hypertext navigation);2)数据仓库(data warehouse);3)多库查询(multi_database queries);4)联邦数据库(federated databases)(Karp,1995;Karp & Paley,1996)。这些技术各有特点,但没有一种可以完全适合各种用途。例如,在生物分类信息系统中,建立联邦数据库是一个极其自然的考虑。"北美植物志"和"中国植物志"项目均采用联邦数据库来管理标本和相关的形态学及地理分布等方面的信息。然而,欲联合更多的分类数据库,联邦数据库途径主要存在两个困难:1)无法为数量未知的信息源建立合适的数据模型;2)缺乏集成一般性信息资源(如 WWW 信息源)的能力。此外,开发费用也是一个值得考虑的问题。相比之下,多库查询则是一个简便易行的方法。我们开发的 Magnolia 2000 (http://www.ibsfu.fudan.edu.cn/english/magnolia/magnolia.htm) 即采用这一方法①。

在相对复杂的生物多样性信息系统中实现数据库间的互操作,最早为 ERIN(Environmental Resources Information Network,现为 Environment Australia Online)的 GIS 数据分析与建模工作。欧洲生物标本信息服务(BioCISE)则应用广泛的元数据体系集中管理各个库的内容与位置信息,利用智能化软件为用户提供一致的界面。美国堪萨斯大学正在开发 Species Analyst System,应用 XML 语言实现多个动植物标本馆数据库间的互操作以获取较为全面的物种多样性信息。澳大利亚的 Taxa Server 和英国自然历史博物馆的 ENHSIN 也在努力实现类似的目标。

值得一提的是,作为生物多样性信息系统基础的分类数据模型仍然受到重视。现阶段分类信息系统的每一个基本子系统,一般只能建立在某种现行分类(系统)的基础上,该分类又需在基本子系统建立之前通过专家选定,很可能因人而异,这就使得不同分类信息系统间的数据很难进行联接、转换和比较。因此,运用新的数据结构和比较模型来实现基于多分类的数据库互操作,可以有效地避免分类学信息的损失,还可以同时处理不同研究工作所获得的结果(Beach et al.,1993;Berendsohn,1995;钟扬,1995;Zhong et al.,1996;Zhong et al.,1997;钟扬,洪亚平,1997;Zhong et al.,1999;Pullan et al.,2000)。在前述的 *Science* 专栏文章中,Bisby (2000)专门评述了密西根州立大学 HICLAS 组(Beach et al.,1993;Zhong et al.,1996)、德国柏林大学及 IOPI 组(Berendsohn,1995)和爱丁堡皇家植物园等(Pullan et al.,2000)在这一领域所取得的成果,并指出今后还需致力于开发通用的软件系统。

4　数字图书馆

20 世纪 90 年代起,大量的数字化媒体数据(如数字化图像、音频、视频、图形、动画等)迅速增长,并通过网络(特别是 Internet)迅速蔓延到我们生活的各个角落。数字图书馆就是一种对数字化资源存储、管理和利用的新技术(Borgman,1999;Wilensky,2000;

① 参见 *The Newsletter of Society for Conservation Biology*, 2000(5) 的介绍

Greene et al. ,2000 ;Kogalovsky & Novikov ,2000)。它与传统图书馆有着完全不同的内涵。

1)数字图书馆是一种基于计算机网络(Internet)的数字资源管理系统,它维护分布式、大规模且有组织的数据库和知识库,保护信息资源的安全和知识产权,支持本地和远程用户借助计算机网络对系统内的数据库和知识库进行一致性的访问,传送和表现用户所需的信息,实现资源共享。

2)数字图书馆是以用户为中心的、由分布式数据和服务组成的信息空间。它必须具备从异构的信息源中发现相关资源的资源发现能力、从确定的信息源中查询多媒体信息的信息检索能力、为检索结果产生有益解释的信息选择能力、汇集和保存选择的信息维护能力以及与他人共享信息的信息交流能力。

3)数字图书馆的典型特征是 :数字化各种媒体承载的信息,通过多媒体技术将它们有机结合在一起进行存储和管理 ;信息的组织形式为超链接的网状组织方式,便于构造相互关联的知识体系 ;信息的网络传输使数字图书馆超越时空观念,跨越馆藏信息的地域界限,加快信息交流与反馈的速度 ;包括友好的人机界面与信息空间导航功能、内容的快速传递功能、强有力的快速检索工具和先进的信息处理与分析工具、随时可用的方法指导、非定点全天无间断的信息资源检索、处理和传递服务等。

4)数字图书馆的关键技术包括 :数字式资源的采集技术,即完成直接的数字化资源创建或传统媒介的数字化转换,也包含来源于图书馆自动化系统 MARC 格式的馆藏目录数据库及一些专题数据库 ;数字化资源的存储与管理技术,以支持对分布式资源的一致性访问 ;信息访问与查询技术,包含对数字化资源和多媒体的访问技术 ;数字化资源的传送与信息发布技术,重点关注图像、音频和视频等多媒体信息的传输、同步和服务质量控制 ;数字式化资源的权限管理方法,为开放的网络环境中的用户提供一致性的信息共享。综合各项专门技术,以互操作技术和多媒体与超媒体技术为代表的技术体系构成数字图书馆的重要基础。

数字图书馆已被认为是下一代 Internet 网上信息资源的管理模式,是信息基础设施的核心,也是国家信息管理技术水平的重要体现。1995 年的美国政府蓝皮书就国家信息基础设施(NII)列出了九项国家级挑战,数字图书馆被列为挑战技术之首。1997 年的美国政府蓝皮书中,数字图书馆被列为有效技术,1998 年被列为首要研究发展重点。1999 与 2000 年,再次被纳入新的国家级研究项目,作为新世纪网络基础应用的具体目标。

数字图书馆技术业已引起生物多样性科学家和生物信息学家的高度重视,并被视为生物多样性信息系统的主要发展方向。例如,Bisby(2000)认为 Species 2000 的目标是为世界上已知的物种构建一个统一的合法索引,而这个索引的一个重要用途就是作为世界范围内的物种数字图书馆的重要组成部分,提供生物物种与相关的保护、分子、种质资源和生态方面的链接。GBIF 的长期目标也是开发一个有关生物多样性知识的数字图书馆(Edwards et al. ,2000)。然而,由于技术上的复杂性,已报道的生物多样性数字图书馆还很少。Schnase 等(1997)设计与建立北美植物志(FNA)数字图书馆的工作是一个良好的开端。该数字图书馆建于美国密苏里植物园,包含约 20000 个维管束与苔藓植物物种的基本信息(FNA 数据库),以及相关的文件、地图、图片、图象和计算工具等。

5　若干原型系统和相关系统网址

5.1　原型系统 1:Species 2000

"Species 2000"(网址:http://www.sp2000.org/)是 1994 年 9 月由国际生物科学联合会(IUBS)组织,与国际科技数据委员会(CODATA)和国际微生物学联合会(IUMS)以及其他生物多样性科学组织(如联合国环境项目的生物多样性工作组等)合作实施的一个大型的生物多样性信息网络项目。建立 Species 2000 的主要目的是:

- 为世界范围内的生物多样性编目工作提供电子版的物种名录;
- 为连接世界范围内物种数据库(网络)提供索引;
- 为比较不同编目提供参照系统;
- 为查询世界范围内物种的命名、分类和现状提供综合资料。

为了实现上述目标,该项目计划:

- 建立一个动态查询系统(称为 Species Locator),用户通过 Internet 进入该系统,可用一个物种名找到一系列的在线分类数据库;
- 建立一个相对稳定的物种索引(称为 Species 2000 Annual Checklist)。该索引每年更新一次,通过 Internet 或 CD-ROM 发布;
- 完善现有的分类数据库,并建立新的数据库来填补缺失环节;
- 建立一个连接系统,使得物种数据库能与其他相关数据库(如种质资源、博物馆与标本馆、生态系统等)共享信息。

在技术上,Species 2000 主要采用的方式有:通过联邦数据库途径实现现有分类数据库间的互操作,开发专用的数据维护与更新系统,以及通过与国际生物命名法权威的合作以保证物种名称的可靠性与稳定性等等。

从 1996 年起,第一批加入 Species 2000 项目的数据库包括:病毒、细菌、珊瑚虫、软体动物、甲壳动物、双翅目、姬蜂、蛾与蝴蝶类、象虫类、鱼类、鸟类、哺乳类、菌类、仙人掌类、棕榈类、豆科、伞形科以及化石植物等,其他类群的数据库也在相继进行之中。

5.2　原型系统 2:GBIF

GBIF(http://www.gbif.org/)由德国联邦教育与研究部(BMBF)和斯图加特自然历史博物馆共同建立,旨在存储和提供世界范围内有关生物多样性研究的第一手资料。与其他生物多样性信息系统所不同的是,GBIF 致力于应用新的信息技术对生物多样性信息进行编辑、链接、标准化、数字化和全球传播。

在 GBIF 即将正式发表的 10 年计划中,主要包含以下几个方面的工作内容:

- 投送数据、信息和网络资源;
- 开发新的用户互操作界面;
- 为新的或现存的数据库建立标准,包括协议、有效性、记录与质量控制等;
- 使用户方便地连接新的和现存的数据库;
- 加强与相关机构和项目间的合作;
- 更新高速网络和计算设备;
- 共享计算设备,包括高容量数据储存器;

● 培训研究者、数据管理员和技术员。

图 1 表示了 GBIF 设计的一个有关已知生物名称的电子目录框架，这些生物名称可用于连接其他生物和非生物数据库。这有可能实现目前还难以想象的生物多样性数据挖掘(data mining)工作(Edwards et al. ,2000)。

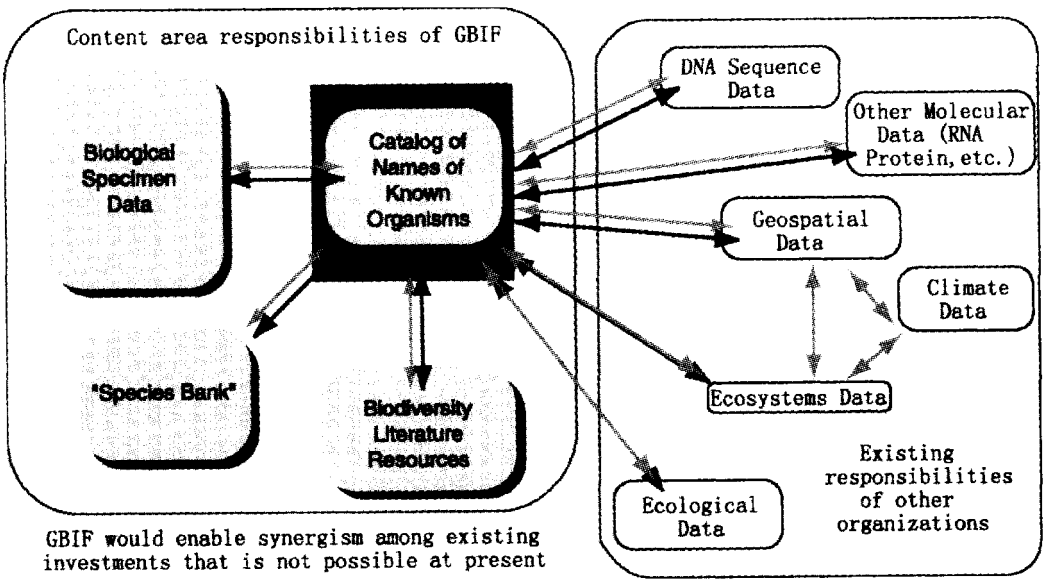

图 1 GBIF 中有关已知生物名称的电子目录框架图(自 Edwards et al. ,2000)

Fig. 1 A design of electronic catalog of the names of known organisms in GBIF (Edwards et al. , 2000)

5.3 其他相关系统网址：

ABIF (http ://www. anbg. gov. au/abrs/abif. htm)

ALICE (http ://dialspace. dial. pipex. com/town/)

BioCISE project (http ://www. bgbm. fu_berlin. de/biocise/default. html#)

Biodiversity and biological corrections web server (http ://www. keil. ukans. edu)

BIOS (http ://www_sp2000ao. nies. go. jp/bios/index. html/)

CephBase (http ://www. cephbase. dal. ca/)

CHM (http ://www. biodiv. org/chm/)

CONABIO (http ://www. conabio. gob. mx)

Darwin Core metadata standard (http ://habanero. nhm. ukans. edu/Z. X/)

DIVERSITAS (http ://www. icsu. org/DIVERSITAS/)

ENHSIN (http ://www. nhm. ac. uk/science/rco/enhsin/)

Environment Australia Online (http ://www. environment. gov. au/search/search. html/)

ERMS (http ://erms. biol. soton. ac. uk/)

FishBase (http ://www. fishbase. org/)

FloraBase (http ://florabase. calm. wa. gov. au/)

HICLAS (http://aims.cse.msu.edu/hiclas/)
ILDIS LegumeWeb (http://www.ildis.org/)
INBio (http://www.inbio.ac.cr/)
Integrated Taxonomic Information System (http://www.itis.usda.gov/)
International Plant Names Index (http://www.ipni.org/)
IOPI Global Plant Checklist (http://bgbm3.bgbm.fu_berlin.de/iopi/gpc/)
IT IS (http://www.it is.usda.gov/)
LITCHI (http://litchi.biol.soton.ac.uk/)
MultiFlora (http://www.cs.man.ac.uk/ai/MultiFlora/)
SINGER (http://www.singer.cgiar.org/)
SPICE (http://www.systematics.reading.ac.uk/spice/)
The Species Analyst (http://habanero.nhm.ukans.edu/)
TreeBase (http://herbaria.harvard.edu/treebase/)
Tree of Life (http://phylogeny.arizona.edu/tree/phylogeny.html/)
URMO (http://www2.eti.uva.nl/database/urmo/default.html)
WORLDMAP (http://www.nhm.ac.uk/science/projects/worldmap/)

致谢　感谢 John H. Beaman 教授、Sakti Pramanik 教授和 Sungwon Jung 博士对我们多年的支持与帮助。赵斌、殷寿华和张晓艳同志对本文的写作提出了宝贵意见。在此一并致谢!

参 考 文 献

"中国生物多样性保护行动计划"总报告编写组,1994. 中国生物多样性保护行动计划. 中国环境科学出版社

钟扬,1995. 植物分类信息系统概述. 植物学通报(增刊):1~6

钟扬,洪亚平,1997. 交互分类信息系统和电子植物志的设计与实现 I. 应用 UNIC 结构、OMES 模型和关系数据库记录多个交互分类. 见:中国植物学会数量分类学专业委员会(编),数量分类学与微机信息处理研究进展. 云南科技出版社,87~100

Beach J H, Pramanik S, Beaman J H, 1993. Hierarchic taxonomic databases. In: Fortuner R (ed.), *Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision*. Baltimore: John Hopkins University Press, 241~252

Berendsohn W G, 1995. The concept of potential taxa in databases. *Taxon*, **44**(2): 207~212

Bisby F A, 2000. The quiet revolution: biodiversity informatics and the Internet. *Science*, **289**: 2309~2312

Bisby F A, Russell G F, Pankhurst R J, 1993. Designs for a global plant species information system. Oxford University Press

Blake J A, Bult C J, Donoghue M J, Humphries J, Fields C, 1994. Interoperability of biological databases: a meeting report. *Systematic Biology*, **43**: 585~589

Borgman C L, 1999. What are digital libraries? competing visions. *Information Processing and Management*, **35**(3): 227~243

Burley J, Scott P R, Speedy A W, 1997. Biodiversity: the role of information technology (IT) in distributing information. In: Hawksworth D L, Kirk P M Dextre_Clarke S (eds.), *Biodiversity Information: Needs and Options*. Oxford: CAB International, Wallingford, 157~171

Davidson S B, Overton C, Buneman P, 1995. Challenges in integrating biological data sources. *Journal of Computational Biology*, **2**: 557 ~ 572

Edwards J L, Lane M A, Nielsen E S, 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, **289**: 2312 ~ 2314

Gingras F, Lakshmanan L V S, Subramanian I N, Papoulis D, 1997. Languages for multi_database interoperability. In: *Proceedings of ACM SIGMOD International Conference on Management of Data*. Tucson, 536 ~ 538

Greene S, Marchionini G, Plaisant C, Shneiderman B, 2000. Previews and overviews in digital libraries: designing surrogates to support visual information seeking. *Journal of the American Society for Information Science*, **51**(4): 380 ~ 393

Kaiser J, 2000. Fossil databases move to the web. *Science*, **289**: 2307

Karp P D, 1995. A strategy for database interoperation. *Journal of Computational Biology*, **2**: 573 ~ 586

Karp P D, Paley S, 1996. Integrated access to metabolic and genomic data. *Journal of Computational Biology*, **3**: 191 ~ 203

Kogalovsky M R, Novikov B A, 2000. Digital libraries as a new class of information systems. *Programming and Computer Software*, **26**(3): 119 ~ 122

May R M, 1995. Conceptual aspects of the quantification of the extent of biological diversity. In: Hawksworth D F (ed.), *Biodiversity: Measurement and Estimation*. London: Chapman & Hall and The Royal Society, 13 ~ 20

Pennisi E, 2000. Taxonomic revival. *Science*, **289**: 2306 ~ 2308

Pullan M R, Watson M F, Kennedy J B, Raguenaud C, Hyam R, 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon*, **49**(1): 55 ~ 75

Schnase J L, Kama D L, Tomlinson K L, Sanchez J A, Cunnius E L, Morin N R, 1997. The Flora of North America digital library: a case study in biodiversity. *Journal of Network and Computer Applications*, **20**(1): 87 ~ 103

Sugden A, Pennisi E, 2000. Diversity digitized. *Science*, **289**: 2305

Wilensky R, 2000. Digital library resources as a basis for collaborative work. *Journal of the American Society for Information Science*, **51**(3): 228 ~ 245

Wilson E O, 2000. A global biodiversity map. *Science*, **289**: 2279

Zhong Y(钟扬), Jung S, Pramanik S, Beaman J H, 1996. Data model and comparison and query methods for interacting classifications in a taxonomic database. *Taxon*, **5**(2): 223 ~ 241

Zhong Y(钟扬), Meacham C A, Pramanik S, 1997. A general method for tree_comparison based on subtree similarity and its use in a taxonomic database. *BioSystems*, **42**: 1 ~ 2

Zhong Y(钟扬), Luo Y, Pramanik S, Beaman J H, 1999. HICLAS: a taxonomic database system for displaying and comparing biological classification and phylogenetic trees. *Bioinformatics*, **15**(2): 149 ~ 156

(责任编辑:时意专)