植物遗传资源学报 2012,13(4):684-687
Journal of Plant Genetic Resources

基于本体论的植物遗传资源数据整合研究

朱天生,杨　华,罗利军,余舜武,詹向东,龙　萍
(上海市农业生物基因中心,上海 201106)

摘要: 随着现代生物学的发展,全球范围内建立了大量的生物学数据共享中心,同时,在生物学发展的带动下,植物遗传资源数据变得更为复杂、异构化和海量。本文在分析国内外几大著名的数据整合共享中心的基础上,简要介绍了本体论的概念及其在生物学领域中的研究现状,提出了基于生物本体论将植物遗传数据、数据挖掘工具、科技文献和科技交流进行整合的设想,并对数据整合需要考虑的几个问题进行了讨论。

关键词: 本体论; 数据整合; 数据共享; 植物遗传资源

Plant Genetic Resources Data Integration Based on Ontology

ZHU Tian-sheng, Yang Hua, LUO Li-jun, YU Shun-wu, Zhan Xiang-dong, LONG Ping
(*Shanghai Agrobiological Gene Center, Shanghai 201106*)

Abstract: With the development of modern biology, many biological data sharing centers have been set up worldwide. Meanwhile, plant genetic resource data is becoming increasingly complex, heterogeneous and huge in size. This paper introduced types and characteristics of the several main biological databases in the world as well as the concept and status of ontology in biology research. The strategies on establishing integrated biological system of multiple data sources, including data mining tools, scientific literature and academic exchange were also proposed. Several issues of scientific data Integration were further discussed at the end of this paper.

Key words: Ontology; Data integration; Date share; Plant genetic resources

植物遗传资源是生物多样性的重要组成部分,是地球上极为重要的财富,是人类赖以生存和发展的重要物质基础。自 20 世纪初期,我国便开始了植物遗传资源的收集保存和评价鉴定研究,20 世纪 50 年代以后,加强了遗传资源的创新和利用研究。在基础研究方面,进行了物种起源演变与遗传多样性研究和重要特征特性鉴定与重要基因发掘,积累了大量数据资料。近年来,随着生物学及数据处理技术的迅猛发展,生物学数据也以每 18 个月翻一番的速度在增长。如何有效地整合这些数据,实现智能化多重交叉检索和建立共享服务体系已经成为目前生物科学数据整合的重点问题。

国内外已有多家研究机构从事植物遗传资源数据的整合研究,其中访问量最多的有美国种质资源信息网(http://www.ars-grin.gov/)、日本国家农业科学机构(http://www.nias.affrc.go.jp)和我国的农业科学数据共享中心(www.agridata.cn/)。这 3 家网站收集了世界或本国内的大部分植物遗传资源数据,元数据和数据都较为规范,并且很多遗传资源都有实物保存。但这些数据缺少科学有效的整合,大多只是一种普通的存储,虽然能进行简单的检索,但功能不够,难以进行数据分析,为资源的深入研究带来了困难。植物遗传资源数据整合的科学方法使得植物遗传数据得到智能化检索和更有效的数据共享是目前植物遗传数据整合中亟待解决的新难点。

收稿日期: 2011-09-10　　**修回日期:** 2011-12-05
基金项目: 科委农转专项(093919n0300);科委重大基础专项(2009DJ1400500)
作者简介: 朱天生,硕士,主要从事植物遗传资源数据共享与分析研究。E-mail: zts@sagc.org.cn
杨华,硕士,与第一作者同等贡献
通讯作者: 龙萍,副研究员,主要从事植物遗传资源数据共享与分析研究。E-mail: lp@sagc.org.cn

本文设想利用本体论在异构数据语法、语义上通用的概念模型[1]来整合植物遗传资源数据,使各异构的数据类型成为一个有序的统一整体,便于查询和数据挖掘。

1　植物遗传资源数据整合概况

1.1　植物遗传资源的描述规范和数据标准的制定

植物遗传资源是极其重要的自然资源,是生命科学研究和新品种选育的物质基础。到目前为止,全世界已经保存了 740 多万份粮食和农业植物遗传资源,为全球的农业可持续发展和粮食安全提供了保障[2]。作物遗传资源是植物遗传资源中最重要的部分。我国作物遗传资源按照农艺学和用途分为 8 类,即粮食作物、经济作物、蔬菜作物、果树作物、饲用和绿肥作物、花卉作物、药用作物和林木作物[3]。国内的种质资源数据库大多也是采用这种分类进行数据展示。

国际植物遗传资源研究所(International Plant Genetic Resources Institute,IPGRI)和国际水稻研究所(International Rice Research Institute,IRRI)制定了作物种质资源描述规范和数据标准。21 世纪初,我国编制出版了《农作物种质资源技术规范丛书》,制定了 120 种作物种质资源的描述规范、数据标准和数据质量控制规范,用于指导作物遗传资源的表现型鉴定与描述。同时,完成了 15.2 万份种质资源的整合、标准化整理、编目和数字化表达[4]。这些描述规范和数据标准的制定都为数据整合奠定了基础。

1.2　数据整合的概况

随着综合性科学和交叉科学的发展,数据资源在数据库数量方面和数据存储量方面的不断增大,数据整合问题越来越成为现代科学技术数据领域的核心问题。如何建设开放共享的科学数据共享平台让数据交换模式更适合于当前目标定位,刘润达等[5]提出了科学数据联盟模式,由此构建的科学数据共享平台可灵活地解决数据资源建设和整合问题,从而提高数据的利用率,提升服务质量。但这种方式对海量异构的生物学数据却也难以发挥作用。

目前生命科学领域中三大科学数据库: 美国国立生物技术信息中心 NCBI(http://www.ncbi.nlm.nib.gov)、日本 DNA 数据库 DDBJ(www.ddbj.nig.ac.jp/)以及欧洲生物信息学研究所 EMBL(http://www.ebi.ac.uk),在国际上都享有盛名,使用频率也非常高。这些数据库存放的主要是基因组、蛋白组等的数据,实质上大多数的生物学数据库是针对这些海量数据而发展起来的。资源学领域所说的护照信息、形态信息等涉及较少。

我国生物科学数据共享工程已经建设了多个科学数据中心,研发了科学数据共享系统,建立了以国家科学数据共享工程(http://www.sciencedata.cn)、西部数据中心[6]和上海生命科学数据中心[7]为代表的生物科学数据库,初步实现了科学数据共享服务。但是,国家科学数据中心和西部数据中心是分类存储数据,查询检索不太方便,尤其是交叉检索无法实现,比如查询所有耐旱的农作物品种,就只能一个个作物分别查询。上海生命科学数据中心主要参照 NCBI 做的,大部分数据也是从 NCBI 上下载得来,采用文献数据库为枢纽对数据集成虽然有了很大的提高,但是数据量不大,尤其植物遗传资源数据更少。相对来说我国的科技基础资源数据的集成和共享水平还有待提高。

2　本体论在生物学数据整合中的应用

2.1　本体论的概念

本体论一词原本是哲学用语,它是研究存在本质的哲学问题。但近几十年里,这个词被应用到计算机界,并在人工智能、计算机语言以及数据库理论中扮演着越来越重要的作用。本体论是概念化的详细说明,一个 ontology 往往就是一个正式的词汇表,其核心作用就在于定义某一领域或领域内的专业词汇以及他们之间的关系。在这一系列概念的支持下,知识的搜索、积累和共享的效率将大大提高。应用在生物学上是研究生物数据的语义、定义等问题,建立一套精确定义、通用可控制性词汇,使之成为生物学概念的模型。

生物本体论主要分为以下几类: 植物性状本体(Trait Ontology,TO),描述植物可区分的特征、特性、质量或表型,比如胶质胚乳、抗病、株高、光敏感性、雄性不育等; 基因本体论(Gene Ontology,GO)是应用最广泛的生物学本体论,基因本体论开发了 3 类独立的、不重叠、可控制的词汇表: 分子功能、生物学途径和细胞学组件。一个基因产物可能有一个或多个相关分子功能,在一个或多个生物学途径中发挥着作用,也可能和一个或多个细胞学组件相关联[8]。目前基因本体论已经广泛应用于基因功能注释和数据整合; 植物本体论(Plant Ontology,PO),描述植物解剖学和形态学的特征[9],包括植物结构(Plant Structure,PS)和生长阶段(Cereal Plant Growth Stages,GRO)两个分支; 环境本体论(Envi-

ronment Ontology,EO),研究植物在环境影响下的基因表达和表现型特征;物种分类本体论(Taxonomy Ontology,GR_tax),按本体论格式描述物种分类;地理本体论(Gazetteer Ontology,GAZ),描述生物资源的地理位置信息。

2.2 本体论是生物数据整合的重要手段

在生物学中,由于描述生物学的术语存在着语义上的巨大差别,再加上语义关系及语法,如不加以统一或标准化,就难以实现有效的知识共享,从而使研究者浪费大量时间和精力在搜寻信息上。甚至由于不同的生物数据库使用不同的语义,导致查询遗漏。比如查询植物干旱相关的数据,如果只用"干旱"作为检索词来查询就会遗漏掉"节水","耐旱","避旱"等数据信息。

目前已有诸如宾夕法尼亚大学计算机系的 Biokleisli 系统[10] 和 IBM 研究院的 Discoverylink 系统[11] 等对生物学数据进行整合。但这些系统都是基于计算机算法来实现数据集成或使用查询优化来提高检索的准确率,并未用生物学领域专业的语义来考虑数据的整合,因而效果都不是很好,而且几乎没有关于植物遗传资源的数据。而本体论本身的特点将各生物学数据库中生物学数据的术语规范统一,使用标准化词汇整合异构生物学数据,从而克服了上述研究方案的缺陷,更加有效实现生物数据共享。

2.3 基于本体论的数据整合概况

目前本体论在生物学数据整合领域已得到广泛应用,尤其是基因本体论(GO)在生物学中是最有权威性的本体论。例如,GO 在异构生物数据整合、消除歧义、处理遗漏数据清洗脏数据中得到应用[12]。曹顺良等[13] 开发的大型生物学数据仓库系统 BioDW 在生物学数据集成上也采用了基于 GO 的方法。杨文等[14] 通过总结 GO 在生物学基因功能注释、生物学文献主题检索、异构生物信息学数据库整合等方面的应用,在分析 GO 应用实例的基础上,提出了利用 GO 进行生物学科学数据与文献数据整合,利用 GO 分类文献和生物学数据。在植物遗传资源方面,鄂志国等[15] 构建了基于本体论的水稻生物学数据库。国外也有多家研究机构从事生物本体论的研究,如 Gene Ontology Consortium、The Plant Ontology Consortium(POC)和 Gramene 等。

随着生物本体论的普遍应用,国内外研究机构也着手开发基于本体论的数据挖掘工具,GoPubMed 通过概念抽取,使文献摘要和 GO 中的概念形成映射,对文献摘要进行 GO 标识分类,从而达到利用 GO 控制检索结果的目的[16];AnnotQTL 将 GO 用于基因和 QTL 功能注释[17]。Conesa 等[18] 开发了未知序列的 GO 功能注释软件 Blast2GO。国内也开发了整合 BLAST 搜索与 GO 注释的软件 GoBlast[19]。这些工具的开发与使用更加速了生物本体论的发展应用。

2.4 利用本体论整合植物遗传资源数据探析

整合植物遗传资源数据的目的主要是为了存储保存这些有价值的数据,为深度数据挖掘提供原材料。植物遗传资源数据包括遗传资源分子生物学数据、环境地理数据、形态学解剖学数据、资源分类数据和资源性状数据。通过生物本体论通用可控的词汇、语义,将植物遗传资源数据、科技文献、科技交流信息和数据挖掘工具映射到本体论,利用本体论的词汇对研究所得的信息进行分类,整合后的逻辑结构如图 1。其中科技交流主要是想给研究者提供一个学习交流植物遗传资源研究的平台。数据挖掘工具用来集成或开发研究植物遗传数据的软件工具。文献为研究植物遗传资源或开发工具本体论等的科技文献资料。植物资源信息及性状指植物遗传资源的来源信息及性状表现数据,参照国家制定的数据描述规范和数据标准。通过生物本体论整合后,比如,稻瘟病抗性基因 *Pi-a*[20] 以性状本体论归类为叶瘟病抗性(TO:0000468),以基因本体论归类为对真菌类刺激的反应(GO:0009620)。研究者可以从多角度去查询所需要的信息,比如查询对基因 *Pi-a* 的研究情况,查询与叶瘟病抗性相关的所有植物分子生物学数据等,就可以方便精确无遗漏的得到所需信息。

美国种质资源信息网、日本国家农业科学机构和我国的农业科学数据共享中心,美国国立生物技术信息中心 NCBI、日本 DNA 数据库 DDBJ 以及欧洲生物信息学研究所可以按照本体论所限定的语义和语法结构来进行整合。整合的主要步骤如下:(1)按照本体论的语义语法及层次结构建立本体论数据库;(2)从这些网站抽取所需要的数据,并清洗数据以提高数据质量;(3)正对清洗后的数据格式建立相应的数据库,比如基因数据就建立基因库,资源性状数据就建性状库;(4)建立对应关系数据库,即本体论库和其他库之间的关系库,由数据标识号和本体论号的对应关系组成。(5)建立跨库的搜索引擎。对于数据库中的植物遗传资源数据,应该具有一个或多个本体论号,如基因功能 GO 号、植物 PO 号、植物性状 TO、环境本体 EO 号、物种编号GR_TAX、地理位置 GAZ 号,可用来进行数据挖掘的工具,遗传数据以

及所有对这些数据研究的文献和科研交流信息。查询时,如果有科研人员需要查询和水分胁迫相关的研究,则可以根据 TO 的有向无环图(DAG) 找到水分胁迫的 TO(性状分类(TO: 0000387) – 抗性或外源物刺激性状(TO: 0000164) – 非生物胁迫性状(TO: 0000168) – 水胁迫(TO: 0000237)),由此可以查询到所有水分胁迫相关的文献、遗传数据、科研交流等信息。这种查询方式是由用户输入关键词或本体论号→查到本体论库对应的本体论号→检索关系数据库得到相关的数据标识号→利用标识号去检索植物遗传资源相应的数据库获取信息。由此可见这种数据的整合方式对检索是准确的无遗漏的,不会出现由于各个数据库对数据描述的术语不一致而检索不到的现象。

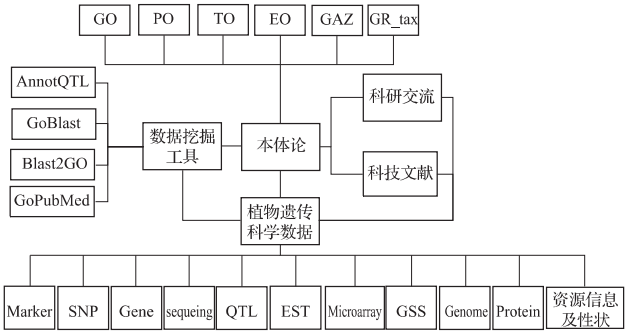

图 1　植物遗传资源数据整合逻辑结构图

Fig. 1　The map of plant genetic resources data omtegration

3　讨论

本文介绍了植物遗传资源科学数据整合和本体论的概念及其研究的概况,并提出了基于本体论将植物遗传数据、数据挖掘工具、科技文献和科技交流进行整合,规范了数据存储,以期达到智能化的交叉检索。但是,首先,从目前实际考虑,由于生物学数据的异构复杂性以及数据之海量,关系型数据库无法很好有效的进行存储。第二,海量的数据以目前的软硬件条件无法集中存储在一起。第三,存在数据安全问题。第四,有关数据共享法律法规。笔者认为,对于异构复杂的生物学数据考虑采用正在发展的对象数据库。对象数据库具有处理海量信息和复杂数据结构的能力[21]。可以考虑采用分布式技术使生物信息资源建设实现高度的自动化、并行化[22]。数据库之间的数据交换采用通用的 XML 格式。数据安全问题的解决需要从多方面的因素考虑,比如人为、网络传输、计算机系统、数据库系统等,各个环节都要严密防范。国外在数据共享的法律法规方面制定了相关法律保护数据共享,我国制定了《国家科技计划项目科学数据汇交暂行办法(草案) 》(2003. 10) ,规范数据共享。

参考文献

[1] Guarino N. Formal Ontology and Information Systems [C] //Proc of the 1st 'Int'l Conf on Formal Ontology in Information Systems. Trento, Itsyl: IOS Press, 1998: 3-15

[2] FAO. The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture [EB/OL]. (2010-10-01) [2011-06-01]. http://www.tsoshop.co.uk/bookstore.sap?FO=1159999&DI=628378

[3] 郑殿升,杨庆文,刘旭. 中国作物种质资源多样性 [J]. 植物遗传资源学报,2011,12(4) : 497-500,506

[4] 王述民,李立会,黎裕,等. 中国粮食和农业植物遗传资源状况报告(Ⅱ) [J]. 植物遗传资源学报,2011,12(2) : 167-177

[5] 刘润达,赵辉,李大玲. 科学数据共享平台之数据联盟模式初探 [J]. 中国基础科学,2010,6: 27-32

[6] 王亮绪,南卓铜,吴立宗,等. 西部数据中心数据集成和共享的回顾与展望 [J]. 中国科技资源导刊,2010,5: 30-36

[7] 许庆炜,曹顺良,李荣,等. 上海生命科学数据中心的设计与实现 [J]. 计算机应用与软件,2008,4: 37-39

[8] Harris M A, Clark J, Ireland A, et al. The Gene Ontology(GO) database and informatics resource [J]. Nucleic Acids Res, 2004, 32: 258-261

[9] Avraham S, Tung C W, Ilic K, et al. The Plant Ontology Database: acommunity resource for plant structure and developmental stages controlled vocabulary and annotations [J]. Nucleic AcidsResearch, 2008, 36(S1) : 449 – 454

[10] Davidson S B, Overton C, Tannen V, et al. BioKleisli: A Digital Libraty for Biomedical Researchers [J]. Int J Digit Libr, 1997, 1: 36-53

[11] Schwarz P, Kodali E, Kotlar J, et al. DiscoveryLink: A system for integrated access to life sciences data sources [J]. IBMSystemsJournal, 2001, 40(2) : 489

[12] 夏燕,张忠平,曹顺良,等. Gene Ontology 在生物数据整合中的应用 [J]. 计算机工程,2005(2) : 57-58

[13] 曹顺良,张忠平,李荣,等. BioDW—一个生物信息学数据集成系统 [J]. 微计算机应用,2005(1) : 59-62

[14] 杨文,孙继林. GO 在生物数据整合中的应用 [J]. 图书情报工作,2008(11) : 124-127

[15] 鄂志国,虞国平,王磊. 水稻生物学本体的构建 [J]. 中国稻米,2009(5) : 52-54

[16] Doms A, Schroeder M. GoPubMed: Exploring PubMed with the Gene Onlology [J]. Nucleic Acids Research, 2005, 33 (2) : 783-786

[17] Lecerf F, Bretaudeau A, Sallou O, et al. AnnotQTL: a new tool to gather functional and comparative information on a genomic region [J]. Nucleic Acids Research, 2011, 39(S2) : 328-333

[18] Conesa A, Götz S, García-Gómez J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research [J]. Bioinformatics. 2005, 21(18) : 3674-3676.

[19] 王成刚,莫志宏. 整合 BLAST 搜索与 GO 注释的软件 GoBlast [J]. 中国生物化学与分子生物学报,2006,22 (12) : 1003-1006

[20] Okuyama Y, Kanzaki H, Abe A, et al. A multifaceted genomics approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes [J]. The Plant Journal, 2011, 66(3) : 467-479

[21] 张婧. 面向对象数据库探析 [J]. 硅谷,2011(5) : 1-1

[22] 范海巍,孙琰,赵饮虹,等. 分布式技术在生物信息资源建设中的应用. 生物信息学,2009(3) : 229-231