协同与分布式数据库技术在高通量组学研究中的应用

崔球 徐健
中国科学院青岛生物能源与过程研究所，青岛 266101

摘 要 本文通过已开展或正在开展的应用实例来着重说明协同与分布式数据库技术在组学研究领域中的重点应用，具体阐述了在远程跨地域实验室信息管理、组学数据的协同注释和分布式计算分析、分布式数据整合和挖掘等方面的应用，向读者展示了信息化技术可以为生物学研究提供的各种便利和必不可少的分析工具。

关键词： 系统生物学；代谢组学；分布式数据库；协同注释

The Application of Concurrent and Distributed Database Technique to High-Throughput “Omics” Research

Cui Qiu, Xu Jian
Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao China, 266101

Abstract: This paper presents the application of concurrent and distributed database technique to “omics” research by case studies. To demonstrate the essential power that informatics techniques brought in, we specifically focused on the systems designed for: remote lab informatics management, concurrent annotation of “omics” data and distribution of computational tasks, information integration and data mining from distributed, semantically heterogeneous data sources.

Keywords: System biology; Metabolomics; Distributed database; Concurrent annotation

1. 引言

现代生物技术研究正处于一个大规模变革的时期。组学学科的技术进步，例如基因组测序技术的进步，得到一个物种的全基因组序列不再是一个大的限速步骤，加上Solexa测序技术应用于转录组学，多维色谱－质谱联用技术应用于蛋白质组和代谢物组等方法学上的进步促使从系统生物学角度研究生物体系成为一个必然趋势。组学学科已经渗透到生物学的各个角落，成为现代生物学研究的常规首选手段。

由于组学学科本身的特点，和信息化技术的结合非常深入紧密。组学学科具有内在的高通量、海量数据处理特性。要求通过信息化手段，建立共享服务，需要在组学技术各个环节实现协同和整合。信息化技术对于组学学科的工作进程管理、多用户协同注释、网格计算、数据整合服务等环节都非常重要。如同没有人会质疑基因组学在现代生物学中的作用一样，也没有人会质疑信息化技术如GenBank数据库在基因组学中的作用和地位，由此不难发现信息化技术将成为现代高通量组学学科的必备武器装备。国际同行在部署大项目时，把对信息化技术的需求摆在整个业务规划中的非常重要位置，而且也对信息技术应用提出了很高的期望。比如现在大的基因组中心或结构基因组实验室，几乎所有的实验相关过程都是建立在电脑信息化系统上，包括仪器预约、样品准备、数据采集存储、数据分析等一系列过程。如果这个信息系统出现了问题，整个中心的运作都会瘫痪。然而，目前我国信息化技术在生物学研究过程中的应用远远滞后于国际同行，至今为止，在系统生物学的各个组成学科中，几乎没有一个主流数据库位于我国境内，例如从基因组学的GenBank到代谢物组学的KEGG、BioCyc等国际一流组学数据库中，看不到中国的身影，让人感到非常遗憾，也和我国的大国地位极不相称。

从为科研提供服务角度来看，信息化技术中强调了科研活动在各个环节里需要协同和整合，实现自动化、标准化和信息共享，某些应用还需要利用现代计算机系统强大的并行运算能力。采取实验室信息管理、工作流控制、数据整合等方式，帮助提高效率和服务水平，促进不同合作机构之间的纵向和横向交流。下面就笔者所在的实验室和科研经历，通过已开展或正在开展的应用实例来着重说明协同与分布式数据库系统在组学研究领域中的一些应用。

2. 提供远程跨地域实验室信息管理

一个项目通常涉及很多分工合作步骤，大型合作项目甚至要求跨地域、多学科的协作。在高通量组学研究项目中，大量的样品可能需要依次在不同实验室流通，分别进行不同的分析实验。因此能够提供远程跨地域协同工作环境的实验室信息管理平台是非常重要的，这个平台首先需要协调各实验室，各合作者间的活动，例如提供样品处理状态查询、仪器预约等功能，方便不同实验室间流通及协作。由于项目分工通常在不同阶段涉及不同的研究人员，每个研究人员或研究组对项目的视角和数据要求是不一样的，因此还要求信息管理平台应该对不同用户作出不同的反应，以最适合用户需要的方式呈递数据，提供不同的数据视角；信息管理平台还需要跟踪系统历史，如样品来源、处理历史、存放位置、中间数据等，有利于流程标准化和质量控制。另外还要求实现地域位置无关地输入及查询数据，真正实现异地多用户多实验室之间的无缝协作。

运用协同注释分布式数据库系统可以很方便地实现上述的功能，其技术关键是在同一界面实现多用户同时编辑、输入、分析跨地域分布的异质数据网络。我们通过扩展和改良在本实验室一直沿用的sesame实验室信息管理系统来帮助克服上述高通量组学研究中的各实验室/合作者之间的协调难题[1]。整个系统基于JAVA，以CORBA为中间件，使用Oracle作为后台数据库管理系统（RDBMS）。系统设计为管理及联接复杂项目中的各个有机组成部分，采集尽可能完整的中间数

据，包括实验操作方案、标准步骤、背景信息、实验数据等，允许数据跟踪及条形码读取，用来组织和协调各实验室、各合作者间的活动。以代谢物组学模块为例，我们的协同信息管理系统可以提供标准代谢物样品、质谱样品、核磁样品、核磁实验、软件、厂家、详细实验操作步骤、具体实验条件等方面的详细信息，样品可以打上条形码标签，登记到系统，自动跟踪样品信息，例如可以追溯样品的来源、处理步骤和历史、当前位置等，以方便不同实验室间流通及协作。用户可以在任何时候，任何地点通过网络进行信息输入及数据处理，真正实现了多用户多实验室之间的无缝协作。同时提供各种工具来便利数据采集，编辑，处理及分析，从而以最适合用户需要的方式呈现与分析数据，它可以提供各步骤进程和人员需要的不同横向数据视角及纵向的项目进度视角，还可以提供方便的自动报表生成。

以仪器预约模块为例，大型珍贵仪器费用高昂，往往需要多家单位合用一台仪器，例如核磁共振仪。因此需要一种协调机制，能够预约及统筹安排机时，避免撞车行为；方便管理及收费，提高工作效率和减少管理人员的工作强度。图1是仪器预约模块的屏幕截图，该模块实现了管理自动化，提供24小时不间断在线服务，用户可以在任意时刻、任意地点通过网络进行仪器预约，查看本月仪器使用情况，以及各种费用和使用状况统计分析报表，极大地方便了用户，提高了仪器的使用效率，减轻了管理人员的负担，通过规范化和标准化流程，同时还减少了管理上的出错几率。

3. 实现组学数据的协同注释和分布式计算分析

当实验测定了组学学科数据之后，需要建立相应的注释数据库以方便研究者的查询和使用。传统上，数据库尤其是模式生物的注释数据库，如酵母SGD，老鼠MGI，果蝇Fly等，都是遵循专家构建、专家管理的模式。这样的模式对保证数据库的权威性和准确性起到了重要的作用。随着组学实验手段的发展，只依赖专业数据库来更新、编辑已有

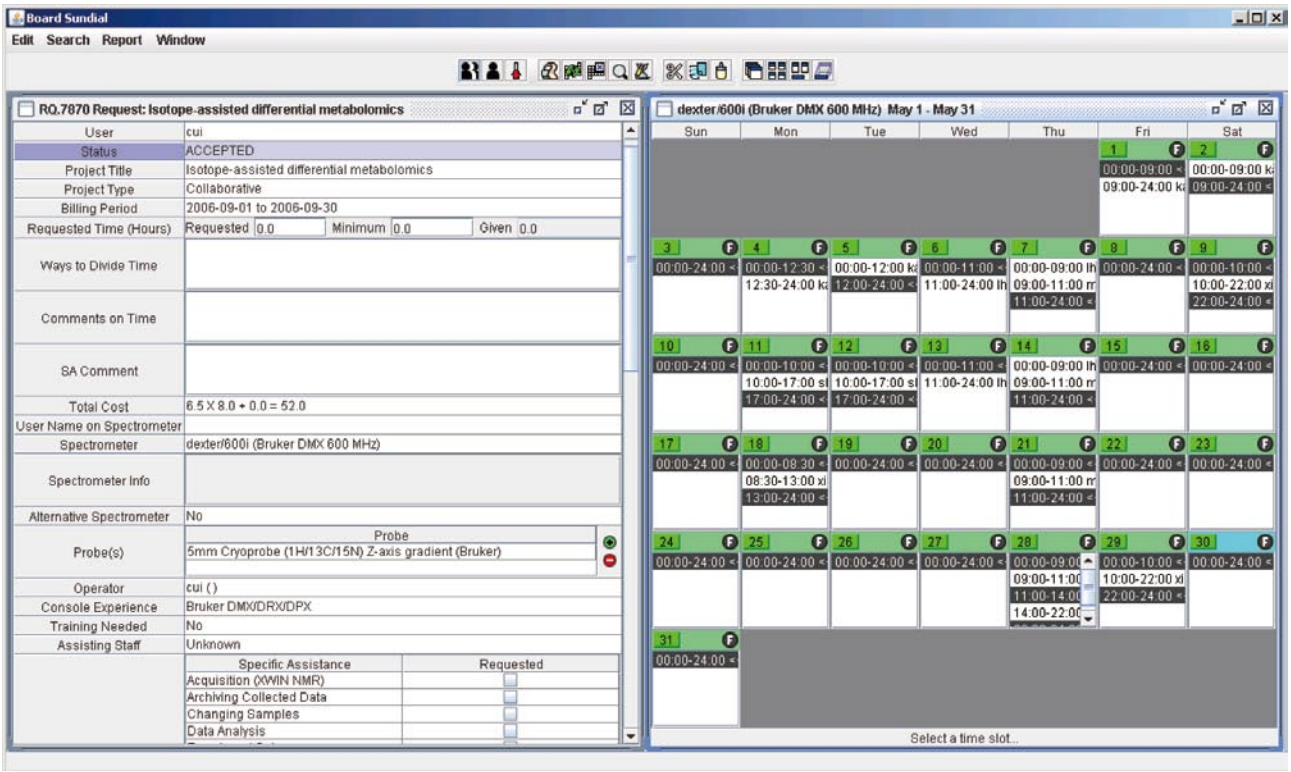


图1 Sesame的仪器预约模块的用户界面

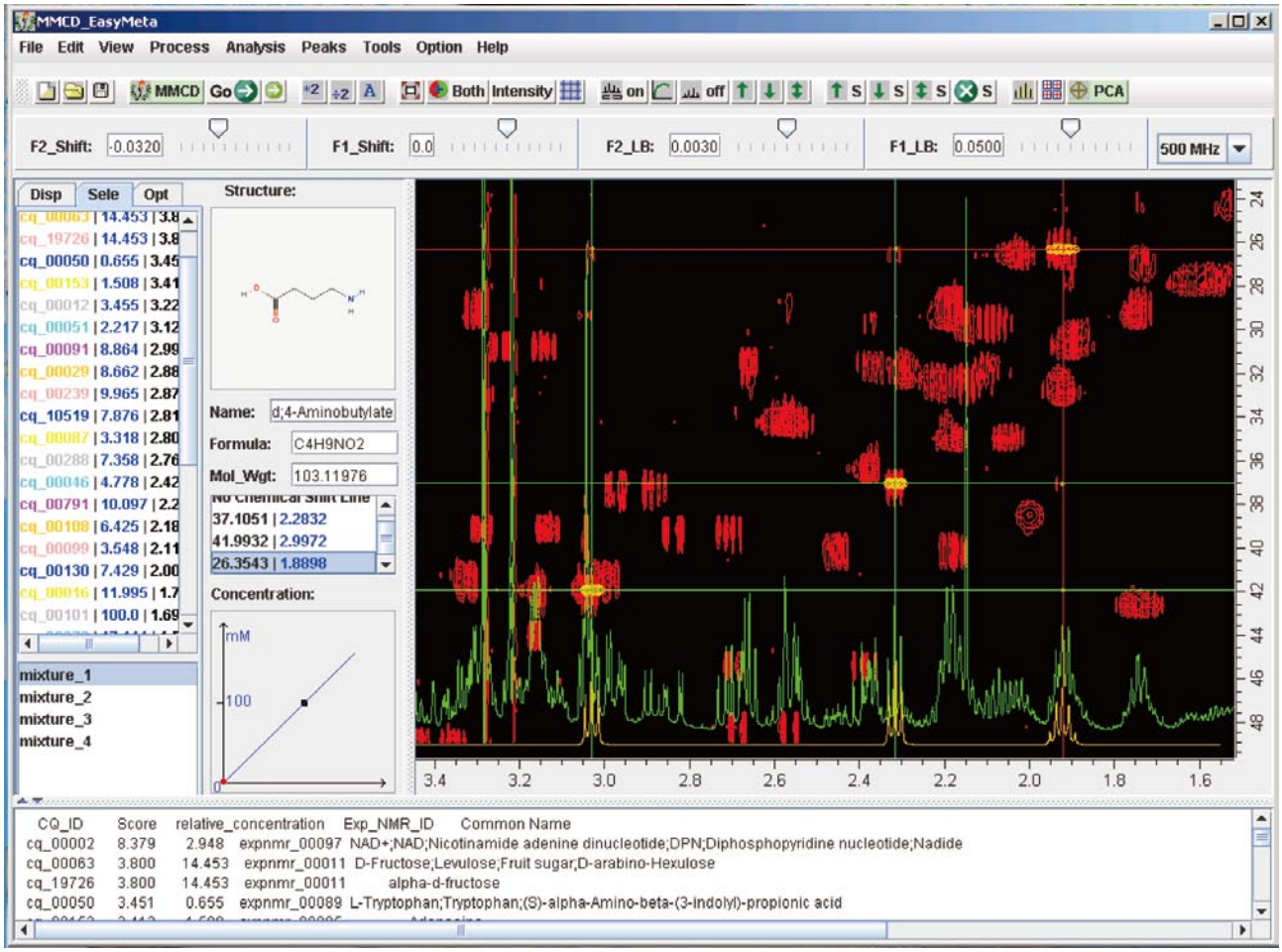


图2 easyMETA分布式代谢物组学分析软件的用户界面

条目及创新新条目已跟不上组学数据扩增的速度。协同注释是因特网普及后的一个新现象。它给予所有上网的人自由编辑和创新的权力。这一新型注释模式的引入极大地促进了数据库的扩增和更新。作为协同注释最有名的例子，维基百科在2005年就增至四百万个题目。有趣的是，在极度扩增的同时，维基百科内容的准确程度并没有显著降低，其准确度甚至近似于大英百科全书[2]。这说明，协同注释并不必然导致数据库质量的下降。现在，协同注释的模式也被越来越多的应用于处理和整合各种组学数据，如GeneWiki[3]，WikiProteins[4]及miRDB微RNA芯片靶基因数据库[5]等协同数据库。我们正在开发一个涵盖各种组学数据类型和数据间相互关系，允许多用户同时分析与注释，能够自动实时更新数据，具有智能化、跨数据类型、跨地域的组学整合信息管理平台（及其用户界面），最终目的是在这一平台上通过全球组学工作者的通力协作，得到具有全球公信力的全面组学注释数据，例如基因功能注释数据、蛋白质相互作用网络等需要大规模协作的组学分析数据。在软件架构上，该整合信息管理平台拟采用三层式Java2/CORBA客户/服务器架构，由客户层、服务器层及数据库层组成。客户层将能够在任意可以运行Java Web Start的计算机上运行，或者在安装了Java插件的任意网络浏览器上运行。客户层和服务器层使用对象请求代理（Object Request Broker (ORB)）进行通讯，对于不同ORB间的通讯则使用互联网ORB间协议（Internet Inter-ORB Protocol (IIOP)）。服务器层维持与数据库层的联络、组装，执行SQL语句，并对返回结果进行处理。数据库编程使用Java2 JDBC API，使得数据库管理和客户层隔离，

数据库管理对用户是屏蔽的，增加了系统的安全性和可靠性。

由于组学数据通常情况下比较庞大，用户的测量数据通过互联网上传到专业分析服务器上进行分析往往不具可行性。而且随着计算复杂度和同时在线用户数目的增加，服务器迟早会超负荷，而分布式计算分析模式则可以满足这些要求。以代谢物组学分析为例，用户测量的色质联用数据或NMR谱图数据通常较大，直接传输到专业分析服务器上需要较长时间，传统的应对方法是仅仅传输文本形式的峰列表，而丢弃了有价值的峰型等信息。我们通过开发一个easyMETA分布式代谢物组学分析软件来有效解决这个难题（见图2）。easyMETA使用Java Web Start技术，将数据预处理和核心分析隔离，软件经由web start模式运行在用户的计算机上，使用用户本地计算资源来分析本地数据，只有在必要时才向服务器发出协助分析请求。通过在服务器和用户计算机之间自动协调分配计算任务，极大地降低了服务器的负担，使得服务器的运行性能并不随着同时在线用户的增加而明显下降。同时还减少了网络传输的数据量。

easyMETA在服务器和用户计算机间协调分配计算任务，谱图处理和分析在本地进行，代谢产物的定性搜索则回传到服务器上执行。

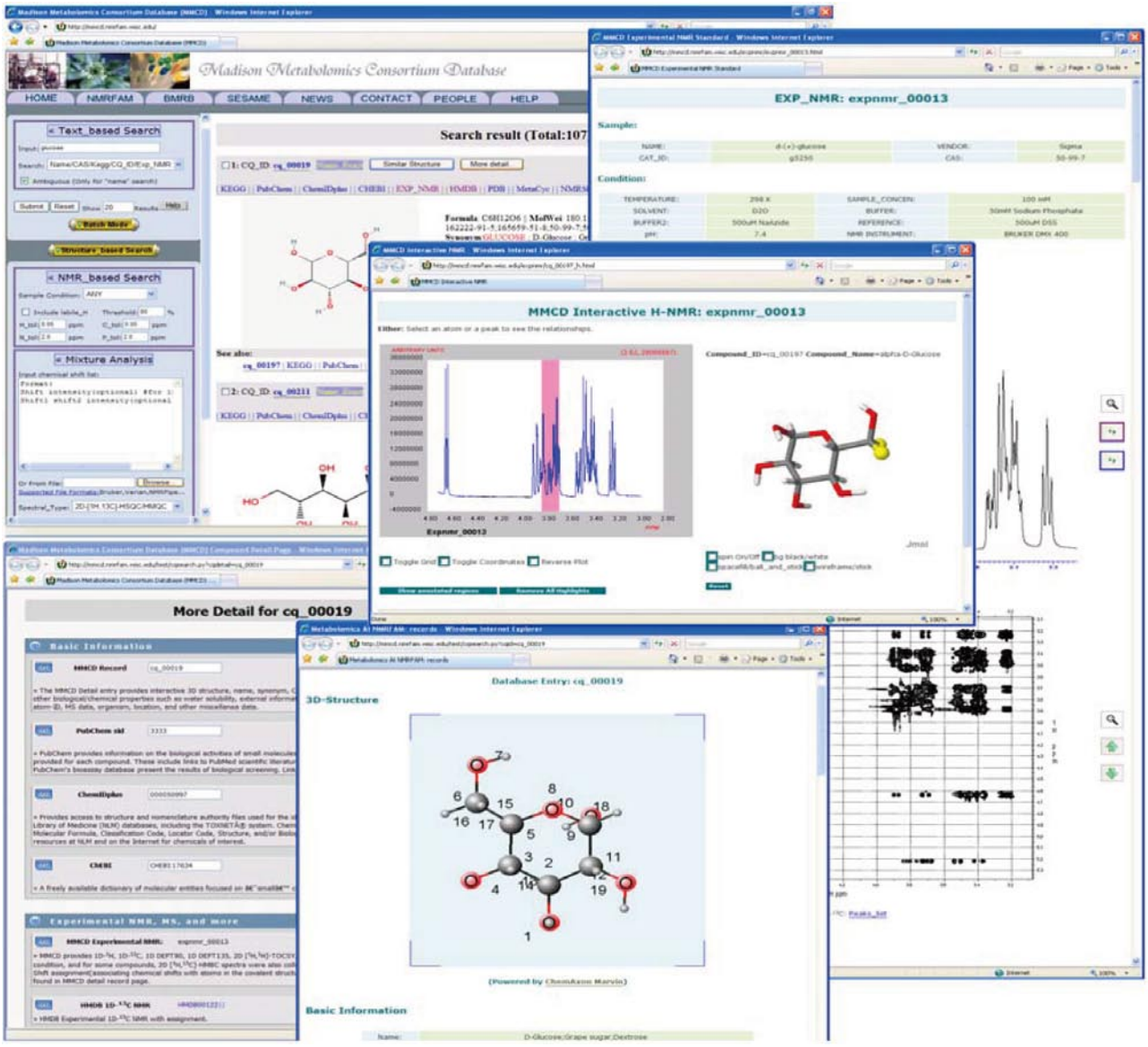


图3 MMCD代谢物组学数据库的各子功能页面

4. 实现分布式数据整合和挖掘

生物学现在成为一个海量信息学科，由基因测序及高通量功能基因组数据所推动的系统生物学方法正在颠覆着生物学研究的模式。随着基因组、转录组、蛋白质组、代谢物组等组学学科的兴起与成熟，带来以前难以想象的细胞各种分子层面上的研究深度和广度，同时也催生了数据整合与挖掘的问题。现代生物学研究极大地依赖于信息的获取及利用效率，多种来源的生物学数据经常需要进行整合，经过系统性的分析来获得对研究体系更全面更深入的认识。通常当数据被组织化存放于专业数据库后，可以通过特定的查询语言，如结构查询语言（对关系数据库而言）或者对象查询语言（对象数据库）取出数据。然而，在现代生物学中，探索一个特定主题的各种不同种类的可利用数据极具挑战性，因为数据分散在互联网上，由大量独立、异质、高度专业化的资源构成。为了实现高效的数据整合与利用，必需充分考虑现代组学学科的一些特殊性质：

1）组学学科的信息容量很大，往往超出现有普通计算工具的合理处理能力。组学数据都很庞大，很难做到轻松传输，尤其是通过网络传输往往不具备实际可行性。需要利用各研究小组固定计算点的资源进行本地分析，所以整合信息处理平台应该包括分布式计算处理。

2）数据异质性阻碍数据整合。由于缺乏统一的标识机制，同一对象在不同资源中往往有不同标识。各种资源互相隔离独立创建，使相互关联和数据整合变得非常困难。对于本地组学分析资源而言，尽管单个步骤可能已经有了相关软件来管理或分析数据，但各软件输入/输出数据的不相容性阻碍了彼此间的交流和数据集成。

3）大量的生物学信息是上下文依赖性的，生物学知识的来源地信息也很重要。而且组学信息的积累及更新速度非常快，具有不确定，不完全，可变等特点，设计数据整合分析系统时必需充分考虑这种可变可扩充特点。

基于以上考虑，我们认为分布式协同数据库系统可以很好承担这种整合任务，我们将在已有MMCD数据库的基础上[6]，建立一个分布式协同数据库入口应用，允许数据查询、可视化，粘合显示，自动化地从各种公共资源进行数据挖掘。在用户层提供一体化组合型查询界面，用户可以在同一界面进行所有组学层次的各种组合查询及数据分析。按照用户需求，自动生成报表。后台程序则提供与各个步骤所需的软件工具配合的能力，链接各种不同的软件的输入输出，在不需要改动原有软件的基础上，使数据在不同阶段不同软件间能进行顺畅交流。数据分析存储采用分布式协同数据库形式，实时分派查询或分析任务到各种支持子数据库系统，使用当前的语义web技术如RDF、OWL、SPARQL来整合、查询和显示多个来源的数据。图3显示了笔者独立开发的MMCD代谢物组学数据库的用户界面，该数据库采用分布式数据整合模式，自动地从多个数据来源收集数据，如基因、蛋白质、代谢网络、生化反应等信息，实现了从代谢物到催化其转化的蛋白质到编码该蛋白的基因整个数据链的无缝链接。

5. 结束语

通过本文介绍的一些信息化应用例子，我们希望向读者展示信息化技术带来的各种便利甚至是必不可少的功能。我们坚信随着信息化技术的进一步成熟，必将对生物学研究带来革命性的促进作用。

参考文献:

[1] Markley, J. L., Anderson, M. E., Cui Q., et al. "New bioinformatics resources for metabolomics." Pac Symp Biocomput, 2007: 157-68.

[2] Giles, J. "Internet encyclopaedias go head to head." Nature, 2005, 438(7070): 900-1.

[3] Huss, J. W., 3rd, Orozco, C., et al. "A gene wiki for community annotation of gene function." PLoS Biol, 2008, 6(7): e175.

[4] Mons, B., Ashburner, M., et al. "Calling on a million minds for community annotation in WikiProteins." Genome Biol, 2008, 9(5): R89.

[5] Wang, X. "miRDB: a microRNA target prediction and functional annotation database with a wiki interface." Rna, 2008, 14(6): 1012-7.

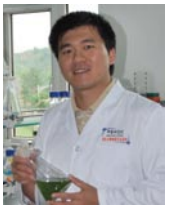
[6] Cui, Q., Lewis, I. A., et al. "Metabolite identification via the Madison Metabolomics Consortium Database." Nature Biotechnology, 2008, 26(2): 162-4.

收稿日期：2009年5月31日

作者信息

崔球

中国科学院青岛生物能源与过程研究所,研究员、博士生导师，研究方向为代谢物组学及蛋白质结构和功能。



徐健

中国科学院青岛生物能源与过程研究所，研究员、博士生导师，研究方向为通过算法与软件开发，进行基因组、转录物组和代谢物组水平上对能源微生物功能及调控机制的认识和模拟，以及基因组解码与分析技术的开发和改进。