福建农业学报 25(3):363~368,2010
Fujian Journal of Agricultural Sciences
文章编号:1008-0384(2010)03-363-06

特色农业微生物基因组研究系统的建立

郑斯平[1],陈 彬[1],宋亚娜[1],关 雄[2],郑伟文[1,2]

(1.福建省农业科学院生物技术研究所,福建 福州 350003;
2.福建农林大学教育部生物农药与化学生物学重点实验室,福建 福州 350002)

摘 要:以Web形式在LAMP(Linux+Apache+MySQL+PHP)环境下,建立以特色细菌菌株为主体的农业微生物基因组数据库和计算机工作平台,充分利用业已公布的生物基因组序列信息和技术平台,揭示特色农业微生物基因的结构功能,提供有自主知识产权意义的DNA序列等生物信息。

关键词:生物信息学;微生物;基因组;数据库
中图分类号:Q 93 文献标识码:A

Establishment of a database system on special microbial genomes

ZNEGN Si-ping[1], CHEN Bin[1], SONG Ya-na[1], GUAN Xiong[2], ZHENG Wei-wen[1,2]

(1. *Biotechnology Institute of Fujian Academy of Agricultural Sciences, Fuzhou, Fujian 350003, China;*
2. *Key Laboratory of Biopesticide and Chemical Biology, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China*)

Abstract: A genome database of agricultural microorganisms with an emphasis on some special bacteria and a platform for analysis was developed using LAMP(Linux, Apache, MySQL, PHP) environment based on the Web. The structures and the functions of the specific genes could be readily accessed. Bioinformation, such as DNA sequences, for intellectual property rights search could also be obtained through worldwide published genome sequences and techniques.

Key words: bioinformatics; microorganisms; genome; database

从事分子生物学的研究,核酸和蛋白质序列的相似性比对及其相关分析是必不可少的。通过互联网联至美国国家生物技术信息中心(National Center for Biotechnology Information NCBI)网站,使用blast软件进行在线比对是常用的分析方法,但是这种分析只能对NCBI所提供的数据进行分析。截至2008年,NCBI的GenBank数据库已有超过950亿碱基数据,可以向全世界提供所有已知的核酸及蛋白质序列[1],但研究人员所需要的仅是某一种或几种基因序列的比对或分析。面对GenBank数据库中浩如烟海的数据,势必因较大的工作量而使比对分析显得繁琐。随着信息技术的迅猛发展以及网络设施的不断完善,许多国家、地区纷纷建立各种类型的微生物菌种数据库。国内已有多个实验室以及生物信息研究中心建立了自己的中小型数据库或开发了一系列软件,以更好地进行数据信息的处理与分析[2-5]。

特色农业微生物基因组研究系统建立的初衷是通过福建省农业科学院生物技术研究所与福建农林大学生物技术中心的密切合作,共同构建以福建省特色细菌菌株为主体的基因组数据库和计算机工作平台,引进相关分析软件,重点研究开发独具特色、有多年研究基础的细菌菌株,并充分利用业已公布的生物基因组序列信息和技术平台,为揭示特色农业微生物基因的结构功能,提供有自主知识产权意义的DNA序列等生物信息,使之成为福建省生物信息研究及其产品研发的平台。

该系统于2005年初建成使用,2007年根据实际应用及可拓展性的需要重新设计。在程序设计上,构建了多个自定义的功能模块,使程序代码更简洁与直观,为系统的维护和升级提供便利。系统采用自建的VBB代码等方法并结合Linux系统自

收稿日期:2010-03-11初稿;2010-04-21修改稿
作者简介:郑斯平(1976-),男,助理研究员,主要从事分子生物学、生物信息学研究(E-mail:ebulin@163.com)
通讯作者:郑伟文(1942-),男,研究员,主要从事分子微生物研究(E-mail:bcfaas01@hotmail.com)
基金项目:福建省科技计划项目(2001Z026)

带的防火墙以及对服务器软件 apache、数据库管理软件 MySQL 的设置，有效提升了系统的安全性。该系统还预留了部分接口，方便日后与其他管理系统的充分整合。

1 数据组成与开发环境

1.1 数据组成

数据来源分为国内数据和国外数据两大部分。截止目前，本数据库已收录福建省农业科学院生物技术研究所、福建农林大学生物技术中心、福建林业科学研究院保存的蓝细菌(*cyanobacterium*)、真细菌(*Eubacterium*)，包括大肠杆菌(*E. coli*)、嗜水气单胞菌(*Aeromonas hydrophylla*)、苏云金杆菌(*Bacillus thuringiensis*)、粪产碱菌(*Alcaligenes faecalis*)、重氮醋杆菌(*Acetobacter*)和放线菌(Actinomyces)等约3200多种菌株的基因序列、核苷酸序列、DNA 指纹以及生化反应、药敏反应特性的数据(或图像)等基础信息。国外数据中主要有 *Synechocystis* sp. PCC 6803, *Nostoc* sp. PCC 7120, *Anabaena variabilis* ATCC 29413, *Synechococcus* sp. WH8102, *E. coli* K12, *E. coli* O157:H7 等模式菌基因组全序列。

1.2 开发环境

DELL Dimension 4300; CPU: Intel Pentium IV 1.8 GHz; 内存: 1 G; 硬盘: 60 G; Web 服务器程序: Apache 1.3.39; 数据库管理系统: MySQL 5.0.45; 脚本语言: PHP5.2.4; 其他应用软件: phpMyadmin, Adobe Photoshop7.2, UltraEditor11.20b。

2 系统架构设计

系统包括数据库、数据分析环境、用户界面(Web)三部分，形成两大平台：即特色微生物遗传背景与技术数据的资源信息平台和运用国际互联网资源及相关软件进行数据分析处理的技术平台。数据信息包括特色真细菌、古细菌、蓝细菌、真菌等数据子库以及相应的数据管理功能(图1)。数据分析主要包含核苷酸、蛋白质、基因等数据分析以及序列格式转换功能(图2)，从程序模块上分为登录、系统菜单、用户管理、数据管理、数据分析和用户操作6个程序模块。

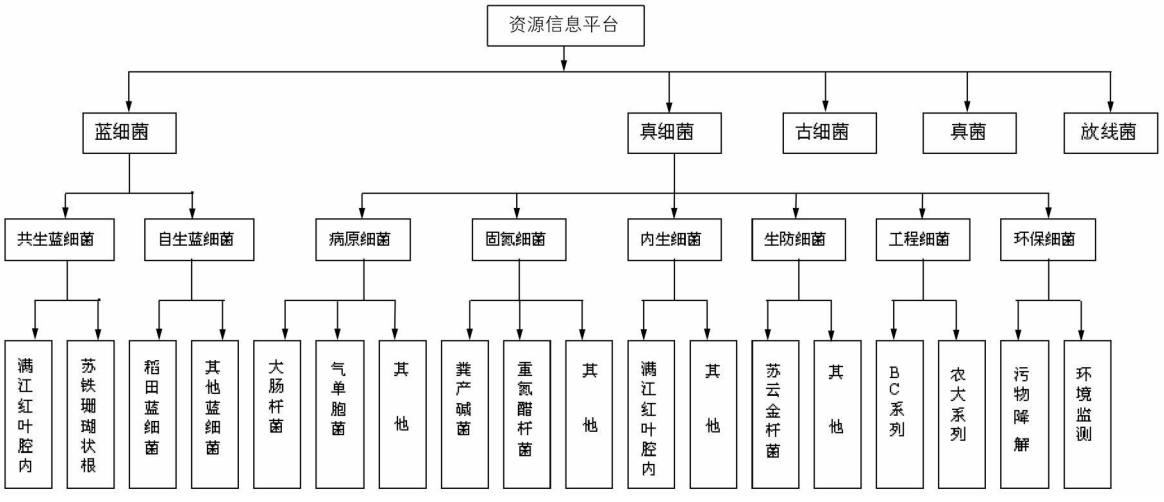

图1 资源信息平台模块图

Fig 1 Components of information platform

3 系统功能设计

3.1 功能设计

系统基本功能是实现基因组信息的存储、查询、编辑、比对、分析，具体功能如下：1) 特色细菌遗传背景和相关信息的登录、查询(包括提供核酸序列登录号等)；2) 基于本数据库的同源性搜索；3) 序列信息的保护功能等；4) 核苷酸数据分析；5) 基因数据分析；6) 蛋白质数据分析；7) 多种序列格式转换。

3.2 程序模块设计

登录模块，执行权限判别以及系统初始化设置；系统菜单模块，由登录模块调用，接收系统参数信息，启动用户管理、数据管理、数据分析、用

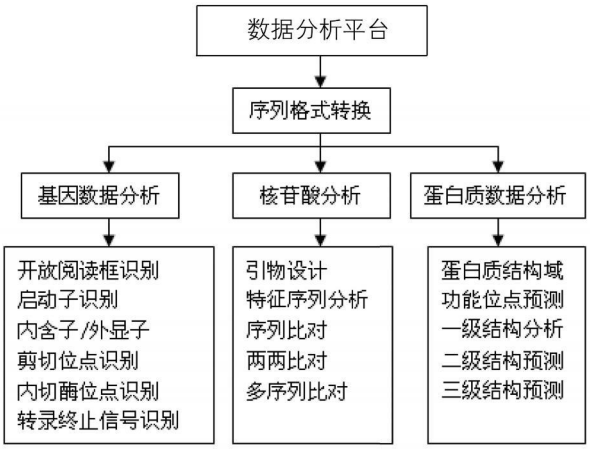

图 2　数据分析平台模块图

Fig 2　Components of data analysis platform

户操作等模块; 用户管理模块, 由系统菜单模块调用, 可根据用户需要调用相应功能函数执行添加、删除、查询用户、设置权限、密码修改、用户列表等功能; 数据管理模块, 由系统菜单模块及用户操作模块调用, 可根据用户需要调用相应功能函数执行添加、删除、编辑、查询、加密、浏览数据等功能; 数据分析模块, 由系统菜单模块及用户操作模块调用, 可根据用户需要调用相应功能函数执行核苷酸、基因、蛋白质序列分析和序列转换等功能; 用户操作模块, 由系统菜单模块调用, 可根据用户需要调用相应功能函数执行数据浏览、查询、管理、分析、打印以及用户密码修改等功能。

3.3　数据库结构设计

在微生物研究中, 根据不同的需要会产生大量的记录数据, 如培养条件、菌落形态或电镜照片等, 为便于这些数据的记录与管理, 在参考 GenBank 数据结构的基础上建立了微生物资源数据库。

以细菌数据库为例: 设计了 3 个基本数据表。基本数据表包括了细菌菌种表、菌种图片表和文献表。细菌菌种表与菌种图片表之间是一对多的关系, 根据图片编号调用菌种图片表中的相应图片; 细菌菌种表与文献表之间是一对一关系, 根据文献编号调用文献表中的相应文献资料。细菌菌种表包括菌名、拉丁名、属名、种名、品系、革兰氏、ACCESSION、相应的 NCBI 链接、采集地、主要特性、分离方法、生理生化、保存条件、培养条件、培养基及其配方、菌落形态、菌体结构、16S 序列及序列注释、主要的基因序列及注释、参考文献、备注、添加人员、添加时间、保密级别等。菌种图片表包括图片编号、图片名、图片大小、图片类型、图片描述、图片链接、添加人员、添加时间、保密级别等。文献表包括文献编号、类目编号、标题、作者、摘要、关键字、文献链接、刊物信息、添加时间、保密级别等。

系统的其他数据表有用户表和用户组表。用户表包括用户编号、用户名、密码、用户类型、用户组、联系方式、注册时间、登录时间、登录 IP、登录次数等。用户组表包括用户组编号、用户组名、允许登录、允许添加、允许编辑、允许删除、允许加密、允许搜索、允许上传、允许分析、允许导入导出、允许用户管理、允许修改密码等。

4　本系统的应用成效与主要特点

依托本系统, 运用生物信息学方法, 可挖掘特色基因片段 (功能片段), 登录 GenBank, 并通过分子生物学和常规方法进行验证, 对部分特色基因、蛋白质的序列进行预测。如筛选、发现含抗艾滋病病毒 (HIV) CV-N 的蓝细菌菌株 RZHA4 (准备申报专利, 未登录 GenBank); 克隆出 8 条抗病同源基因序列[6], 并登录 GenBank; 发现 3 株含有可降解石油、苯酚等环境污染物的特色基因的菌株[7,8]。同时, 依托数据库和分子标记技术平台, 对从福建省土壤或水体分离的有益或病原细菌、真菌等进行分子鉴定, 其中包括对近 3000 株环境中的有益或病原细菌的特色 (毒素) 基因进行分子鉴定[9-10]; 对蓝细菌-苏铁人工重组体进行了分子鉴定[11]; 对 54 个水稻根内外 *nif* H 基因的阳性克隆进行了 PCR-RFLP 分析[12]; 对 40 株福建稻田固氮菌进行 PCR-DGGE 分析与测序鉴定[13-15]; 对 22 株满江红内生菌进行 16SrRNA-PCR-DGGE 分析与测序鉴定[16]。

本系统实现从最初单纯的数据收集、录入向数据分析、比对转变; 从单一的某项分析向系统分析转变, 为资源数据、试验数据的收集整理、应用分析等方面提供方便, 具体有以下几个主要特点:

4.1　较高的稳定性和安全性

本系统使用的 GNU/Linux 操作系统是一套免费使用和自由传播的类 UNIX 系统, 它拥有性能稳定、速度快、开放源代码、不受任何商品化软件的版权所制约等优势, 我国已把它作为政府的指定网络操作系统[17]。选用的服务程序 Apache, 同样也是开源软件, 较之微软的 IIS 更具安全性[18], 可以支持并稳定运行多种脚本语言 (如: FastCGI、

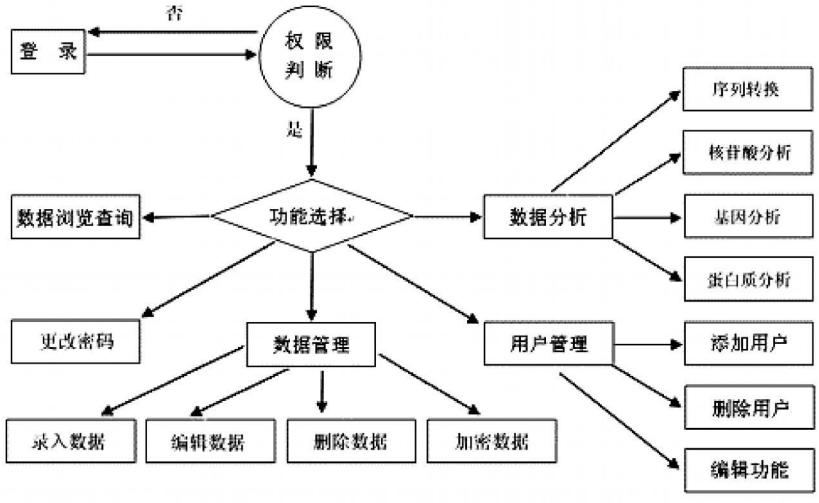

图3 系统处理流程

Fig 3 System flow diagram

图4 系统页面截图

Fig 4 System front page

注:(1) 首页;(2) 管理页面;(3) 分析平台;(4) 记录浏览页面

ASP、PHP、JSP、Perl)。本系统还采用当前世界上最流行的开源数据库软件[19] ——MySQL数据库管理系统。它提供了其他数据库少有的编程工具,是一个多用户、多线程的SQL (Structured Query Language结构化查询语言) 数据库,可同时处理几乎不限数量的用户 (可处理多达5 000万以上的记录),命令执行速度快,并有简单有效的用户特权系统[20]。通过对Linux、Apache和MySQL的设置以及对数据的保密设置,使系统的稳定性和安全性、硬件向下兼容性和运行速度均得到保证。

4.2　使用方便，提高科研效率

面对 GenBank 以指数级增长的数据量，研究人员的在线搜索和比对分析往往因为网络不畅而耗费大量精力和时间。本数据库参照国内多个实验室和生物信息研究中心的做法，实现了 Blast、ClustalW 等软件进行本地数据的比对。由于下载了常用的模式微生物基因组全序列，建立了资源信息的管理和分析平台，从而保证在网络过于拥挤或发生异常的情况下，可随时利用本单位的分析系统。本系统不仅能为普通用户提供特色农业微生物基础信息，而且还可以将相关资源信息、试验数据的动态管理及数据分析充分整合，使常用的研究分析可以直接在本系统内完成。系统的投入使用，使实验室在试验结果的存储、整理、查询和分析等方面的工作效率有了明显提高，简化了资料归纳整理、装订归档等繁琐的工作，同时也简化了资料查询和数据分析。

4.3　成本低、便于维护和升级

在 Linux 环境下，相当数量的软件都是基于 GNU 协议的开源软件，如服务器程序 Apache，数据库管理系统软件 MySQL，脚本语言 PHP 都可以通过 Internet 免费下载使用，与 Windows 相比，可以节省不少成本。Linux 对硬件设备的要求也不高，即使是一台 486 计算机就足以使用。MySQL 使用为雅虎、谷歌、诺基亚、维基百科等大型的公司节省时间和资金[21]。作为基于 Web 设计的数据库系统主要使用脚本语言，增加或更改某项功能只需要调用脚本直接编写代码，然后覆盖源代码文件即可。这种即改即用的方式，省去了基于软件设计所必须的编译、打包等繁琐过程，维护比较方便。2007 年本系统经重新设计，将各功能模块具体细化，以功能函数的形式体现，使每个功能模块的程序代码显得更简洁与直观，同时留有程序接口，为今后系统的数据扩容和升级提供便利。

4.4　使用范围广，用户交互性好

本系统基于 Web 设计，在服务器端建立系统后，用户在客户端上通过网络浏览器（如 Internet Explorer）即可使用，省去了另行安装插件或客户端软件等不必要的麻烦；将系统移植到 Internet 空间，世界范围内的研究者均可使用。本系统在首页的帮助链接中提供详细的用户帮助信息也便于用户使用。

5　结　论

“特色农业微生物基因组数据研究系统”是基于 Linux 平台，采用 B/S 模式、基于 Web 数据库技术构建的，于 2007 年 11 月、2008 年 2 月分别通过福建省计算机软件测试实验室检测和省级专家验收。该系统具有运行稳定、界面友好、操作方便等特点，不仅能提供较详实的细菌菌种信息，而且在实现信息动态管理的同时，也提供核苷酸、基因、蛋白质数据分析和序列转换等多个功能，使常用的研究分析可直接在本系统内完成。该系统的投入使用，明显提高了试验结果的分析工作效率，简化了资料归纳整理、装订归档资料的查询、数据的分析等步骤，具有较强的实用性。在农业微生物基因组的分子鉴定、特色基因的克隆、序列分析、功能诠释、指纹图谱建立等实际应用上已取得成效。目前已开通网站 http://www.e-biol.com/。

参考文献：

[1] DENNIS A B, ILENE KARSCH-MIZRACHI, DAVID J L, et al. GenBank [J]. Nucleic Acids Research, 2009 (37): 46-51.

[2] 马俊才，张荣肖. 中国微生物资源数据库细菌性状子库的数据结构与检索功能设计 [J]. 微生物学通报，1990，17 (4): 210-214.

[3] 禹胄，李涛，蔡涛，等. 微生物基因组注释系统 MGAP [J]. 微生物学报，2003，43 (6): 805-808.

[4] 刘旭光，宋福平，张广杰，等. Bt cry 序列本地数据库的建立及本地 BLAST 的实现 [J]. 中国农学通报，2005，21 (11): 375-378.

[5] 黄金光，朴春根，田国忠，等. 林业微生物菌种资源数据库查询系统构建 [J]. 农业网络信息，2004 (6): 21-24.

[6] 高丽华，周以飞，郑伟文，等. 南瓜 NBS 类抗病基因同源序列的克隆与分析 [J]. 长江蔬菜，2007 (8): 40-43.

[7] 林智敏，宋亚娜，姚梅宾，等. 一株苯酚降解菌（*Alcaligenes faecalis* BC2001）的 PCR 检测 [J]. 福建农业学报，2007，22 (1): 50-53.

[8] 李友发，福建稻田细菌群落 PCR-DGGE 分析和石油烃降解基因克隆 [D]. 福州：福建农林大学，2008.

[9] 黄文文，陈彬，庄一廷，等. 闽江流域福州过境段大肠杆菌毒素基因的定性定量分析与季节性变化 [J]. 福建农业学报，2008，23 (4): 408-414.

[10] 陈彬，庄一廷，黄文文，等. 闽江流域表面水体中大肠杆菌毒素基因的多重 PCR 检测 [J]. 安全与环境学报，2009，9 (1): 112-116.

[11] 陈彬，郑斯平，郑伟文. 蓝细菌与福建苏铁（*Cycad revoluta*）的侵染性重组的研究 [J]. 福建农业学报，2007，22 (4): 350-353.

[12] 陈彬，郑斯平，周莉娟，等. 水稻根际土壤及根组织内外固氮微生物的遗传多样性分析 [J]. 农业生物技术学报，2007，15 (5): 841-846.

[13] 宋兵，李友发，林智敏，等. 稻田固氮细菌分离物的 PCR-DGGE 及其序列同源性分析 [J]. 福建农业学报，2007，22

(4): 346-349.

[14] 罗青, 宋亚娜, 郑伟文. PCR-DGGE法研究福建省稻田土壤微生物地区多态性 [J]. 中国生态农业学报, 2008, 16 (3): 669-674.

[15] 李友发, 宋兵, 宋亚娜, 等. 福建省稻田土壤细菌群落的16S rDNA-PCR-DGGE分析 [J]. 微生物学通报, 2008, 35 (11): 1715-1720.

[16] 郑斯平, 陈彬, 关雄, 等. 小叶满江红内生细菌多样性的PCR—DGGE及电子显微镜分析 [J]. 农业生物技术学报, 2008, 16 (3): 508-514.

[17] 杜顶, 甘仞初, 雷育生. Linux环境下基于Web的信息系统建设研究 [J]. 计算机应用研究, 2004 (6): 53-54, 89.

[18] 林慧琛. 走进Apache世界 [J]. 在线技术, 2004, 8: 66-69.

[19] MICHAEL K. The Definitive Guide to MySQL5, e3 [M]. Berkeley Apress Inc, 2005: 5-7.

[20] 姚继锋, 伊欣, 吴瞻, 等. linux应用实例与技巧 [M]. 北京: 机械工业出版社, 1999 (9): 210-212.

[21] About MySQL [EB/OL]. [2009-11-19]. http://www.mysql.com/about/.

(责任编辑: 柯文辉)