基于 Linux＋Apache＋MySQL＋PHP 的棉花分子生物学数据库系统构建

张荣志，王省芬，马峙英*，张桂寅，迟吉娜
（河北省作物种质资源重点实验室/河北农业大学，河北 保定 071001）

摘要：利用课题组积累的数据信息和网上公用数据库中有关棉花分子生物学数据，采用 MySQL 为后台数据库，以 Apache＋PHP 为环境，在 Linux 操作系统下实现了棉花分子生物学数据库系统的构建和 wwwBLAST 的本地化服务。在脱机状态下使用 blast 进行同源性比对，使得数据查询和序列比对更加方便、快捷、稳定和安全，而且增强了针对性。数据库最终以局域网的形式发布，通过授予权限可向同行提供利用。本课题组在棉纤维品质、黄萎病抗性等相关基因的克隆研究中使该数据库得到初步应用。

关键词：棉花；分子生物学数据库；wwwBLAST 本地化；生物信息学
中图分类号：S562.035.3 **文献标识码**：A
文章编号：1002-7807(2008)05-0399-03

The Construction of Cotton Molecular Biology Database System Based on Linux＋Apache＋MySQL＋PHP

ZHANG Rong-zhi, WANG Xing-fen, MA Zhi-ying*, ZHANG Gui-yin, CHI Ji-na
(*Key Laboratory of Crop Germplasm Resources of Hebei; Agricultural University of Hebei, Baoding, Hebei* 071001, *China*)

Abstract: With the rapid development of biotechnology and the exponent increase of cotton molecular biology data, in order to further dispose, analyze and use these data, it is necessary to establish the corresponding bioinformatics databases. Based on the research data of our laboratory and of public cotton molecular biology, we constructed a cotton molecular biology database system in the Linux operating system. In the database system, MySQL database was used as a background, and the open environment depended on Apache＋PHP. In the Linux operating system, the standalone of wwwBLAST alignment system services was accomplished based on Apache environment. The blast could be operated in the off-line state. The service could make up the blast deficiencies of public databases which were complicated and of weak pertinence. Since the local data were added into databases, the analysis and comparison of data became safer, quicker and more detailed. The database was put out by local area network and could be accessed at author's permission. This database, in our laboratory, has been put into application in related gene cloning of fiber development and *Verticillium* wilt resistance.
Key words: cotton; molecular biology database; standalone wwwBLAST; bioinformatics

随着基因组研究的不断深入，越来越多基因的结构和功能得到阐明，已有大量的公共数据库系统可供研究者使用。但在针对特定物种的生物信息学分类和分析方面仍有待进一步开发，这些公共数据库在内容、数据综合和检索途径上不一定能满足实际研究的需要，因此建立二级数据库已成为研究热点之一。二级数据库是通过搜索、查询已知数据库的信息，进行加工、整理和系统化，构建专用数据库，如拟南芥信息资源 TAIR[1,]、水稻基因组序列数据库 RGP[2,]、玉米基

收稿日期：2007-07-23 **作者简介**：张荣志(1981-)，女，在读硕士研究生；* 通讯作者，mzhy@hebau.edu.cn
基金项目：河北省自然科学基金重点项目(C2006001034)

因组数据库 MaizeGDB[3]、禾本科比较基因组数据库 Gramene[4]等。

棉花专门的基因组数据库正在构建之中，如 CottonDB 数据库和 CMD 数据库，但这些数据库有些查询条目不够细化，查询方式和查询结果有时不能满足试验需要，而且本地数据不能被添加到数据库中进行分析比较。序列比对是生物信息学的基础，两个序列的比对有比对软件包——BLAST[5]和 FASTA[6]可供使用，但公共数据库的比对缺乏针对性，使得比对结果非常复杂。本地化的 wwwBLAST 可使复杂的比对结果变得整齐一致，而且数据安全、比对速度快。

本研究借助生物信息学手段，利用课题组积累的数据信息以及网上公用数据库中有关棉花分子生物学数据，基于 Linux＋Apache＋PHP＋MySQL 组合构建棉花分子生物学数据库系统，基于 apache 环境实现 wwwBLAST 本地化服务，为棉花分子生物学研究提供良好的工具，为实验室生物信息学分析平台的构建提供参考。

1 数据库开发环境

利用 Linux 操作系统建立 Web 站点，选用 Apache 服务器和 MySQL 数据库，主要采用 PHP 实现动态数据交换。在动态网站的构建中，数据库通过 Apache 和 PHP 将用户输入的数据进行处理，并将处理后的数据以网页的形式反馈给用户(图 1)。

图 1 基于 Web 的数据库结构工作模型

Fig. 1 The model of database structure base on the web

基于 Linux＋Apache＋PHP＋MySQL 组合构建动态网站的软硬环境如下：

硬件环境：

PC 机(联想，P4/2.93 G CPU，512 M 内存)作为 Web 服务器和开发机。

软件环境：

操作系统：Red Hat Linux 9.0 (http://www.redhat.com.cn/)；

Web 服务器：apache-2.2.0.tar.gz (http://www.zaccum.com/)；

PHP：php-4.4.2.tar.gz (http://www.php.net/)；

MySQL：mysql-standard-5.0.18-linux-i686-glibc23.tar.gz (http://www.mysql.com/)；

wwwBLAST：wwwblast-2.2.9-ia32-linux.tar.gz (ftp://ftp.ncbi.nih.gov/blast/executables/release/)；

以上压缩文件均在 Linux 操作系统下的/usr/local 目录下，用 gzip 或 tar 命令进行解压，安装好 Linux、MySQL、Apache 后，再安装 PHP，将它安装成以 Apache 为服务器的动态模块，并集成对 MySQL 的支持。

2 棉花分子生物学数据库系统的构建

棉花分子生物学数据库系统包括用户登录系统、数据库查询系统和数据库管理系统(图 2)。

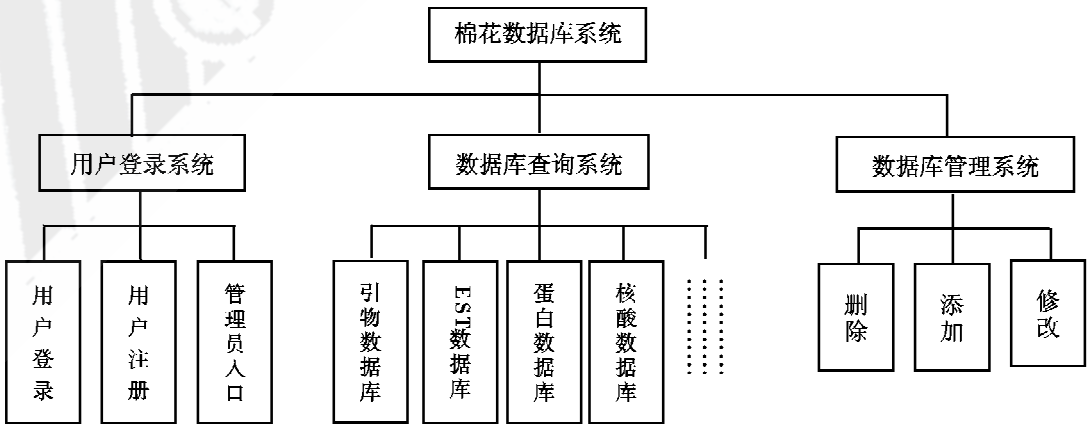

图 2 棉花分子生物学数据库系统

Fig. 2 Cotton molecular biology database system

图 3　EST 数据库查询系统

Fig. 3　EST database query system

下面以 EST 数据库为例说明数据库查询系统的构建过程。首先，用 Dreamweaver 构建网页的基本轮廓(图 3)。

构建好网页的基本轮廓后，用 PHP 调用数据库。

用户登录系统和数据库管理系统也是通过 PHP 对数据库的操作来实现，用户注册是向数据库写入记录，用户登录是验证用户所输入的账号和密码是否与数据库中已有的账号和密码一致，对上述代码进行简单的修改就可以实现。

3　wwwBLAST 本地化的实现

将 wwwBLAST 软件在 Linux 系统下解压，然后将文件 blast. html 中关于使用数据库列表的内容改为棉花的数据库名，代码如下：

```
<a href="docs/blast_databases. html">Database</a>
<select name = "DATALIB">
    <option VALUE = "Gossypium_EST_db"> Gossypium_EST_db
    <option VALUE = "Gossypium_Protein_db"> Gossypium_Protein_db
    <option VALUE = "Gossypium_contig"> Gossypium_arboreum_contig
    </select>
```

将棉花数据库中的序列用 formatdb 命令对其进行格式化，然后将格式化后的数据文件放在 blast 文件下的 db 目录下。

另外还须将 apache 文件下的 httpd. conf 文件进行修改，因为 wwwBLAST 软件包中的一些文件是用 CGI 编写的，因此需在 apache 环境中添加能够执行 CGI 的语句。

配置好 apache 后，在浏览器中输入 http://*. *. *. */blast/blast. html(*. *. *. * 为本机的 IP 地址)，就可以进入和 NCBI 的 blast 程序几乎一致的检索界面。

4　小结

本研究构建的棉花专门分子生物学数据库，可避免在大型公共数据库查询时的复杂操作，能获得更加准确的数据信息。在获取信息方面不受互联网的限制，检索速度明显提高。wwwBLAST 本地化的实现提高了比对速度、结果的准确性和数据的安全性。数据库最终以局域网的形式发布，通过授予权限可向棉花同行提供利用。课题组在棉纤维品质、黄萎病抗性等相关基因的克隆研究中使该数据库得到初步应用。但该数据库还有许多内容需要深入研究，数据库的功能还需不断完善和扩展。随着实验室数据分析要求的不断提高，查询界面应该包含更多的信息，查询条目要进一步细化，并且要加强库与库之间的联系，形成多层次、系统的数据库查询系统。

参考文献：

[1] HUALA E, Dickerman A W, Garcia-Hernandez M, et al. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant [J]. Necleic Acids Res, 2001, 29: 102-105.

[2] SAKATA K, Antonio B A, Mukai Y, et al. INE: a rice genome database with an integrated map view [J]. Nucleic Acids Res, 2000, 28: 97-101.

[3] LAWRENCE C J, Seigfried T E, Brendel V. The maize genetics and genomics database. The community resource for access to diverse maize data [J]. Plant Physiol, 2005, 138: 55-58.

[4] WARE D H, Jaiswal P, Ni J, et al. Gramene, a tool for grass genomics [J]. Plant Physiol, 2002, 130: 1606-1613.

[5] ALTSCHUL S F, Madden T L, Schaffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Res, 1997, 25:3398-3402.

[6] PEARSON W R. Rapid and sensitive sequence comparison with FASTP and FASTA[J]. Methods Enzymol, 1990, 183: 63-98.