序列同源性分析软件 Blast 的 WEB 界面构建及其应用*

张成岗[1)] 张利达[2)] 欧阳曙光[1)] 张启发[2)] 贺福初[1)**]

([1)]军事医学科学院放射医学研究所基因组学与蛋白质组学研究室，北京 100850;
[2)]华中农业大学作物遗传改良国家重点实验室，武汉 430070)

摘要 基于局域网 (Intranet) 内的 PC/Linux 服务器，构建了序列同源性分析软件 Blast 的 WEB 界面. 局域网内的所有计算机均可通过 WEB 方式访问该服务器进行公共数据库和自建数据库的查询，具有保密、高效、免费的优点，能够满足实验室和研究院所的大规模、快速数据分析任务.

关键词 微机，Linux 操作系统，局域网，Blast 软件，WEB 界面，生物信息学

学科分类号 Q754

核酸和蛋白质序列的相似性分析是几乎每个分子生物学工作者经常要做的工作之一. 联网到美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 使用 Blast 软件 (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) 进行核酸和蛋白质序列对库检索是最常用的分析方法[1]. 但是，这种方式只能对 NCBI 所提供的数据库进行序列相似性分析. 国外和国内已有多家实验室从事大规模测序过程并建立了自己的核酸和蛋白质序列数据库. 通常方式下，对这些自行建立的数据库的相似性分析一般使用 Telnet 程序登录到相应的服务器后在局域网环境下使用 Blast 等软件进行对库检索，使用十分不方便. 事实上，NCBI 同时提供了能够通过 WEB 界面进行数据分析的 Blast 程序. 本文报道以 PC 机和 Linux 操作系统为基础构建 Blast 分析系统 WEB 界面的方法及其应用.

1 构建 Blast 软件的 WEB 界面

具体的实现过程包括几个方面：安装并配置基于 PC/Linux 操作系统的局域网络；下载 WEB 界面的 Blast 软件；下载有关数据库 (如 EMBL 或 GenBank 核酸序列数据库、SwissProt 蛋白质序列数据库等) 并完成解压缩；自建 FASTA 格式的序列数据库；进行 Blast 分析系统配置和网络化界面调试.

1.1 安装 Linux 操作系统并组建局域网

作为推荐，硬件可选用 Pentium Ⅱ微机/Intel CPU 500 MHz/内存 128MB/18GB SCSI 硬盘. 操作系统可选用 RedHat Linux 6.1 或以上版本. 对硬盘进行分区时可参考以下设置：Linux native 分区> 2GB，Linux Swap 分区= 127MB，数据区> 12GB，工作区> 4GB. 如有可能，可将硬件配置增加到多个 CPU、1GB 以上内存，将大大提高数据分析能力. 随后按照常规方法将实验室的所有计算机构建成一个小型局域网，并将该微机作为服务器使用，由所在单位网络管理中心分配唯一的 IP 地址，保证用户能够对该微机进行访问[2]. 从 http://httpd.apache.org/dist/apache_1.3.19.tar.gz 下载 Apache 软件作为 WEB 服务器，提供一个优良的、完全免费和完全源代码开放的 Httpd 服务器网络环境. 有关安装和配置过程可参考文献 [2].

1.2 安装 WEB 界面的 Blast 软件

提供 WEB 界面的 Blast 软件可从 NCBI 下载 (ftp://ncbi.nlm.nih.gov/blast/server/wwwblast_Nov.6.2000/wwwblast.Linux.tar.gz)[3]. 随后在服务器上由管理员用以下命令释放，并将文件权限赋予相应的用户和用户组：/usr/home] $ gzip -d wwwblast.Linux.tar.gz; tar -xvpf wwwblast.Linux.tar

此后，用户将获得权限进行以下工作：http://my_hostname/blast/blast.html：调用 Blast

* 军事医学科学院科技创新研究启动基金 (9905105)、国家自然科学基金项目 (39900041, 39900074) 与重点项目 (39730310)、国家“863”项目 (863-102-10-04-04) 部分资助.

** 通讯联系人.
Tel: 010-66931246, E-mail: hefc@nic.bmi.ac.cn

收稿日期：2001-03-07，接受日期：2001-05-17

程序对库检索; http://my_hostname/blast/wblast2.html: 两两序列对齐分析; http://my_hostname/blast/psiblast.html: 调用 PSI/PHI Blast; http://my_hostname/blast/rpsblast.html: 调用 RPS Blast.

该软件同时提供了测试用的核酸序列数据库 (test_na_db) 和蛋白质序列数据库 (test_aa_db), 可直接用于对库检索.

1.3 检索用数据库的准备

EMBL 核酸序列数据库[4]和 SWISS-PROT 蛋白质序列数据库[5]均可从欧洲分子生物学实验室 (http://www.embl-heidelberg.de) 或其镜象网址下载 (如北京大学的镜象网站http://www.cbi.pku.edu.cn). 下载后需要进行解压缩, 并需将 EMBL 格式的序列文件转换为具有 FASTA 格式的序列文件. 具体过程参见文献 [6]. 通过大规模测序以及通过其他渠道所获得的序列数据可直接处理成为 FASTA 格式的序列文件. 在进行数据库检索之前, 需要用 Blast 软件所携带的 formatdb 程序对数据库进行格式化, 具体命令参见文献 [6]. 随后需要将上述已格式化的数据库文件 (或其链接关系) 移动到数据库目录"./db"中.

1.4 Blast 软件的配置

在使用之前, 需要对 Blast 程序的运行环境进行配置, 首先考虑向服务器配置文件"blast.rc"中加入数据库名称. 参考表 1 可完成该文件的配置. 数据库列表的数量则没有限制.

Table 1 Modification of the configuration file "blast.rc" of the WEB-blast software

Parameter	Value	Comment
NumCpuToUse	2	Number of processors to use
blastn	nt_0 nt_1 pat gss sts yeast pdb vector est my_db	List of the nucleic acid sequence databases to be used
blastp	aa sp pat pdb yeast	List of the amino acid sequence databases to be used
blastx	aa sp pat pdb yeast	List of the amino acid sequence databases to be used
tblastn	nt_0 nt_1 pat gss sts yeast pdb vector est my_db	List of the nucleic acid sequence databases to be used
tblastx	nt_0 nt_1 pat gss sts yeast pdb vector est my_db	List of the nucleic acid sequence databases to be used

随后对 Web Blast 的输入表格文件"blast.html"进行修改. 找到该文件中关于所使用数据库的列表处, 将拟使用的数据库名称进行添加即可. PSI/PHI Blast 程序所使用表格文件(psiblast.html)中数据库列表的添加方式同上. 但是, 需要注意的是, PSI/PHI Blast 所使用的数据库的 FASTA 格式必须含有以下前缀"> gil ...". 同时, 供 PSI/PHI Blast 程序所使用的数据库在进行格式化时必须使用参数"-o T".

1.5 Blast 分析环境的使用

在局域网中, 用户可通过任何一台计算机通过 WEB 方式联网到该服务器. 假如该服务器的 IP 地址为"202.38.152.252", 用户则可在 Internet Explore 或 Netscape 浏览器的地址栏中输入"http://202.38.152.252/blast/blast.html", 即可进入检索界面. 用户的序列可通过 Windows 的拷贝/粘贴功能输入, 点击按钮 "Search" 即可工作. 分析结果的输出方式也和联网到 NCBI 主页进行分析完全一致. 同理, 在浏览器的地址栏输入"http://202.38.152.252/blast/wblast2.html"则可进行核酸和蛋白质序列对两两对齐分析. 按照以上方式, 用户可将任意序列构建成核酸和蛋白质序列数据库进行检索, 极大地扩大了进行序列同源性分析的自由度.

2 讨 论

本文报道了在 PC 机和 Linux 操作系统中实现 Blast 软件的 WEB 界面分析方式, 对于小型实验室和研究院所分析自行构建的核酸和蛋白质序列数据库具有积极意义. NCBI 的 Blast 软件是进行核酸和蛋白质序列相似性分析的有力工具, 可用来对任意核酸和蛋白质序列数据库进行检索. 具有 WEB 界面的 Blast 软件, 则为普通用户基于自行构建的数据库高效地进行分析, 提供了十分重要的分析手段, 用户无需关注该程序如何工作, 而且其数据的输入、输出方式也和联网到 NCBI 进行分析的界面完全一致, 具有简单易用的优点. 该套独立运行的 WEB Blast 软件不支持队列操作, 任务一旦提交将立即运行. 其运行效率主要依赖于用户的硬件设备. 使用其他类型服务器和操作系统也可构建 Blast

软件的 WEB 分析界面, 如 OSF1V4alpha、SGI、SolarisIntel 和 SolarisSPARC 等, 可下载相应版本的 Blast 软件 (ftp://ncbi.nlm.nih.gov/blast/server/). 在国内大量实验室普遍使用 PC 机的情况下, 选用 Linux 操作系统可充分利用其作为免费而又稳定的多任务操作系统的优势, 可以促进生物信息学在我国的推广与应用.

参 考 文 献

1 Altschul S F, Madden T L, Schäffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25** (17): 3389~ 3402

2 施世帆, 吕建毅, 碧 恩. Linux 网络服务器使用手册. 北京: 清华大学出版社, 1999. 31~ 60
Shi S F, LÜ J Y, Bi E. Handbook of Linux network server. Beijing: Qianghua University Press, 1999. 31~ 60

3 张成岗, 张绍文, 贺福初, 等. 序列同源性比较软件 Blast 的本地化实现及 VB 接口程序的编制. 生物化学与生物物理进展, 1999, **26** (5): 516~ 518
Zhang C G, Zhang S W, He F C, *et al*. Prog Biochem Biophys, 1999, **26** (5): 516~ 518

4 Baker W, van den Broek A, Camon E, *et al*. The EMBL Nucleotide Sequence Database. Nucleic Acids Res, 2000, **28** (1): 19~ 23

5 Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res, 1998, **26** (3): 38~ 42

6 张成岗, 欧阳曙光, 贺福初, 等. 基于 PC/Linux 的核酸序列大规模自动分析系统的构建及其应用. 生物化学与生物物理进展, 2001, **28** (2): 263~ 266
Zhang C G, Ouyang S G, He F C, *et al*. Prog Biochem Biphys, 2001, **28** (2): 263~ 266

Construction and Application of the WEB Interface of Blast Package*

ZHANG Cheng-Gang[1], ZHANG Li-Da[2], OUYANG Shu-Guang[1], ZHANG Qi-Fa[2], HE Fu-Chu[1]**

([1] *Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing* 100850, China;

[2] *National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan* 430070, China)

Abstract Based on a microcomputer installed with the Linux operating system and integrated with the intranet environment, the Blast software with the WEB interface was constructed for visualized sequence similarity analyze. All of the users in the intranet could visit the WEB Blast server using WEB-browser to analysis their sequence easily. The databases to be searched might be either the public database such as EMBL or GenBank nucleic acid sequence database and SwissProt amino acid sequence database or the private sequence databases prepared by themselves. Result demonstrated that this intranet WEB Blast environment is very useful in large-scale nucleic acid sequences and amino acid sequences analysis.

Key words microcomputer, Linux operating system, intranet, Blast software, WEB interface, bioinformatics

* This work was partially supported by grants from Initiative Foundation for Scientific and Technological Innovation of Academic Military Medical Science (9905105), Chinese National Natural Science Foundation General Program (39900041, 39900074) and Key Project (39730310), Chinese High-tech Program (863-102-10-04-04).

** Corresponding author. Tel: 86-10-66931246, E-mail: hefc@nic.bmi.ac.cn

Received: March 7, 2001 Accepted: May 17, 2001