经验交流

基于 PC/Linux 的核酸序列分析系统的构建及其应用*

张成岗 欧阳曙光 张绍文 瞿祥虎 鱼咏涛 周钢桥 吴松锋 贺福初**
(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)

摘要 基于 PC 机和 Linux 操作系统, 利用 Phred/Phrap/Consed 软件和 Blast 软件, 构建了核酸序列大规模自动分析系统. 该套系统可自动完成从测序峰图向核酸序列的转化、载体序列去除、序列自动拼接、重复序列鉴定以及序列的相似性分析, 可加速对大规模测序数据的分析和利用.

关键词 PC 机, Linux 操作系统, Phred/Phrap/Consed 软件, Blast 软件, 生物信息学, 大规模测序
学科分类号 Q754

随着人类基因组计划的深入进行[1], 表达序列标签 (expressed sequence tag, EST)、cDNA 序列和基因组序列在不同实验室中被大量获得, 从而使大规模的序列相似性分析显得日益重要. 美国华盛顿大学基因组中心 (University of Washington Genome Center, UWGC) (http://www.genome.washington.edu/uwgc/) 所开发的 Phred/Phrap/Consed 系列软件可以完成从测序峰图 (如 377 型、3700 型等测序仪的测序结果) 到核酸序列的转换以及序列拼接[2], 而美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 的 Blast 软件则可对本地的数据库进行序列相似性分析[3]. 它们均可运行于 Linux 操作系统中. 常用的核酸序列数据库 (如 GenBank、EMBL) 和蛋白质数据库 (如 SWISS-PROT) 等均可本地化安装和使用. 故而, 本文拟介绍如何以低价位的 PC 机和免费的 Linux 操作系统为基础, 借助这些软件资源和相关数据库构建核酸序列大规模自动分析系统.

1 核酸序列大规模系统的构建过程

1.1 在微机上安装 Linux 操作系统

作为推荐, 硬件可选用 Pentium ㊣微机/Intel CPU 550 MHz/内存 128 MB/18 GB SCSI 硬盘. 操作系统可选用 RedHat Linux 6.1 或者以上版本. 对硬盘进行分区时可参考以下设置: Linux native 分区> 2 GB, Linux Swap 分区= 127 MB, 数据区> 12 GB, 工作区> 4 GB.

1.2 Phred/Phrap/Consed 系列软件的获得及安装

Phred/Phrap/Consed 系列软件由 UWGC 开发, 其组成见表 1. 这些软件均可通过 E-mail 获得. 有关信息查询可访问网址 "http://bozeman.mbt.washington.edu/index.html".

1.3 Blast 软件的获得及安装

Blast 软件是对两条核酸序列之间或者两条蛋白质序列之间进行序列相似性分析的有力工具, 可直接从 NCBI 的主页下载 (ftp://ncbi.nlm.nih.gov/blast/executables/blast.linux.tar.Z), 建议保存目录为 "/usr/blast", 释放后即可使用[4].

* 军事医学科学院科技创新研究启动基金 (9905105), 国家 "863" 项目 (863-102-10-04-04)、国家杰出青年科学基金 (39620514) 与国家自然科学基金重点项目 (39730310) 部分资助.

** 通讯联系人.
Tel: 010-66931216, E-mail: hefc@nic.bmi.ac.cn

收稿日期: 2000-03-06, 接受日期: 2000-03-31

Table 1 Source and functions of Phred/Phrap/Consed software package

软件名称及压缩包内容	联系作者及地址	建议保存目录	功能描述
phred: phred-dist-acd.tar.Z	Brent Ewing: bge@u.washington.edu	/usr/wu/phred	将测序峰图文件转换为核酸序列并生成质量控制文件
phd2fasta: phd2fasta-acd-dist.tar.Z	Brent Ewing: bge@u.washington.edu	/usr/wu/phd	将质量控制文件转换为FASTA格式的序列, 与phred过程配合使用
Phrap/cross_match/swat: distrib.tar.Z	Phil Green: phg@u.washington.edu	/usr/wu/distr	去除载体序列、完成序列拼接等
Consed	David Gordon: gordon@genome.washington.edu	/usr/wu/consed	序列拼接分析
RepeatMasker: RepeatMasker050599.tar.gz	Arian Smit: asmit@nootka.mbt.washington.edu	/usr/wu/repm	去除重复序列如Alu等. 基于Perl 5.0以上环境

安装时需要从压缩包中释放, 在相应子目录进行编译后即可使用.

1.4 EMBL和SWISS-PROT数据库的获得与预处理

EMBL核酸序列数据库[5]和SWISS-PROT蛋白质序列数据库[6]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/). 使用前需要进行一系列预处理, 包括:

a. 数据库的解压缩: 将各个压缩包中的文件释放, 还原为EMBL格式的序列文件. 命令为: gunzip * ; tar xfv *

b. 数据库的格式转换: 将EMBL格式的序列文件转换为具有FASTA格式的序列文件. 命令为: se2f file_in_EMBL_format file_out_fasta_format 其中参数"file_in_EMBL_format"是具有EMBL格式的输入文件, 参数"file_out_fasta_format"是具有FASTA格式的输出文件. se2f是自行设计的序列格式转换程序.

c. 数据库的格式化: 用Blast软件中提供的formatdb程序对FASTA格式的序列文件进行格式化, 以便使用Blast软件进行序列相似性分析. 命令为: formatdb-p F-i /mnt/scsi/db/nr/hum1-o T

其中, 参数"-p"后的"F"表示需要格式化的序列是核酸序列, 而选择"T"则表示需要格式化的序列是蛋白质序列; 参数"-i"后为需要格式化的文件名; 参数"-o"后方的"T"则表示可使用fastacmd软件对格式化好的数据库中的序列根据序列的序列接受号进行检索.

d. 数据库的序列相似性分析过程: 用Blast软件中提供的blastall程序进行本地化的序列相似性分析, 具体用法见后.

e. 数据库中已知序列的检索: 用Blast软件中提供的fastacmd程序可从已完成格式化的数据库中按照序列接受号快速检索所需的序列(具有FASTA格式). 如果需要检索某序列的全部信息(EMBL格式), 可自行编程. 例如: fastacmd-d database_name (已完成格式化)-s accession_number 可用于从已完成格式化的数据库中检索一个指定序列(FASTA格式). fastacmd-d database_name (已完成格式化)-i accession_number_list_file 可用于从已完成格式化的数据库中检索多个指定序列(具有FASTA格式), 其序列接受号保存在文件"accession_number_list_file"中. get_seq database_name (未格式化) accession_number 可从指定的未格式化的数据库中检索指定序列的所有信息(EMBL格式), 而不仅仅是FASTA格式的序列. 其中get_seq是自行设计的序列检索程序.

1.5 Phred/Phrap/Consed系列软件和Blast软件的运行设置

为便于描述, 特设置Linux下的环境变量为"set ifn= $ 1; set path_wu= /usr/wu".

Phred/Phrap/Consed系列软件主要用于从测序峰图中获得高质量的核酸序列, 主要过程包括:

a. 测序峰图文件中序列的判读. 可采用以下命令:

$ path_wu/phred/phred-id chrom_dir-pd phd_dir-qa $ ifn.qual;

$ path_wu/phd/phd2fasta-id phd_dir-os $ ifn-oq $ ifn.screen.qual

其中, 目录chrom_dir下为每次测序所获得的

峰图文件，目录 phd_dir 下为 phred 软件对峰图文件进行核酸序列判读时所生成的质量控制文件.

b. 载体序列的去除. 可采用以下命令:

$ path_wu/distr/cross_match $ ifn $ path_wu/vector.seq -minmatch 12 -minscore 20 -screen > screen.out

上述过程可根据文件$ path_wu/vector.seq 中所收录的载体序列去除已生成的核酸序列中的载体序列.

c. 重复序列片段的去除. 可采用以下命令:

perl $ path_wu/repm/RepeatMasker050599/RepeatMasker $ ifn.screen

上述过程可将核酸序列中的各种短片段重复序列去除[7]. 去重后的序列在用 Blast 软件进行序列相似性分析时可以更加有效地获得分析结果，以免被“Alu”等信息所冲淡.

d. 序列的自动拼接. 可采用以下命令:

$ path_wu/distr/phrap $ ifn.screen-ace-view > phrap.out

按照上述方式运行，phrap 软件可以输出自动拼接好的序列，其扩展名为“*.contigs”，而未能参与拼接的序列则在“*.problems”中存在. Phrapview 软件和 Consed 软件可对拼接好的序列进行观察和校正.

e. 序列的相似性分析. 可采用以下命令:

blastall -p program -d “MyLib1 MyLib2” -i MySeq.fasta -o MySeq.Out -v 100 -b 10 -e 0.01

其中，参数“-p”后的参数“program”可为“blastn、blastp、blastx、tblastn、tblastx”之一. 具体而言，blastn 指用核酸序列检索核酸序列数据库、blastp 指用蛋白质序列检索蛋白质序列数据库、blastx 指用核酸序列检索蛋白质序列数据库(基于所有可能的六个不同相位编码为桥梁)、tblastn 指用蛋白质序列检索核酸序列数据库(基于所有可能的六个不同相位编码为桥梁)、tblastx 指用核酸序列检索核酸序列数据库(基于所有可能的六个不同相位编码为桥梁). 参数“-d”后的“MyLib1 MyLib2”表示可列出一个或多个待检索的数据库. 参数“-i”后的“MySeq.fasta”是需要查询的具有 FASTA 格式的核酸序列或者蛋白质序列文件. 参数“-o”后的“MySeq.Out”为结果输出文件. 参数“-v”后的整数用于限制相似序列的显示数量; 参数“-b”后的整数用于限制相似序列与检索序列对齐的显示数量; 参数“-e”后的数值用于设定序列之间相似的阈值.

1.6 该系统完整运行方式

上述有关操作可集成于 Linux Shell 命令中，即将序列判读、载体序列去除、重复序列去除等过程组织在一个过程 A 之中，将 Blast 软件过程组织在一个过程 B 之中，此两个过程被合并在过程 C 之中，则用户只需要通过调用过程 C 即可实现对每次测序数据的自动分析，即获得目标序列及其序列相似性分析的结果(限于篇幅，具体过程从略).

2 讨 论

本文介绍了如何将 Phred/Phrap/Consed 系列软件和 Blast 软件在 PC 机和 Linux 操作系统中配置以构建核酸序列大规模自动分析系统. 采用该套系统，我们已经成功地处理了 16 000 余条 EST 序列，得到了很好的结果[8].

人类基因组计划的发展也导致了相关软件资源的发展. 在序列判读软件方面，美国 PE 公司设计的运行于苹果机的 Sequence analysis 软件是伴随 377 型等测序仪同时向用户提交的软件，但是它只能完成序列的判读. Sequencher™ 软件是美国基因编码公司的商业产品(http://www.genecodes.com/)，也能够完成批量峰图文件中核酸序列的判读、载体序列的去除、序列的拼接等工作. 然而，与本文所介绍的 Phred/Phrap/Consed 系列软件相比，Sequencher™ 软件中没有序列的去重功能. Sequencher™ 软件运行于苹果机中，所以与运行于 Linux 系统中的 Blast 软件之间的衔接还需要进一步的调试. 本文中所采用的序列相似性分析软件是 NCBI 的 Blast 软件，实际上使用美国华盛顿大学(Washington University, WU)的 Blast 软件(WU-BLAST2)(http://dove.embl-heidelberg.de/Blast2/readme.html)以及 EBI 的 FASTA 软件(http://www2.ebi.ac.uk/fasta3/)也可实现，只是在运算速度、参数的使用以及结果输出等方面有一些差异. 同时，选用 Linux 操作系统可充分利用其作为多任务操作系统的优势，例如，可将安装有该套系统的微机作为服务器置于一个局域网中，该局域网中的其他 Windows、Linux 和 Unix 用户则可以通过 Telnet 远程登录使用该套系统，十分方便.

参 考 文 献

1 Collins F S, Patrinos A, Jordan E, *et al*. New Goals for the U. S. Human Genome Project: 1998~2003. Science, 1998, **282**

(5389): 682~ 689
2 Ewing B, Hillier L, Wendl M C, *et al*. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research, 1998, **8** (3): 175~ 185
3 Altschul S F, Madden T L, Sch¾ffer 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
4 张成岗, 张绍文, 贺福初, 等. 同源性分析软件 BLAST 的本地化实现及 VB 接口的编制. 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
5 Baker W, van den Broek A, Camon E, *et al*. The EMBL nucleotide sequence database. Nucleic Acids Res, 2000, **28** (1): 19~ 23
6 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
7 Deininger P L, Batzer M A. Alu repeats and human disease. Mol Genet Metab, 1999, **67** (3) 183~ 193
8 Zhang C G, Yu Y T, He F C, *et al*. Characterization, chromosomal assignment, and tissue expression of a novel human gene belong to ARF GAP family. Genomics, 2000, **63** (3): 400~ 408

Construction and Application of a Large-scale DNA Sequence Analysis System Based on PC/Linux*

ZHANG Cheng-Gang, OUYANG Shu-Guang, ZHANG Shao-Wen, QU Xiang-Hu, YU Yong-Tao, ZHOU Gang-Qiao, WU Song-Feng, HE Fu-Chu**
(Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing 100850, China).

Abstract More and more DNA sequences have been obtained since the start-up of human genome project. Powerful system is badly needed for data mining on these DNA sequences. Based on a personal computer and Linux operating system, the Phred/Phrap/Consed software and Blast software were used to construct a platform for batch analysis of the sequences, including identifying raw DNA sequence from chromatogram file, vector sequence removing, contig analysis (sequence assembly), repeat sequence identifying and sequence similarity analysis. Result demonstrated that this robust platform could accelerate data analysis for large-scale DNA sequencing.

Key words personal computer, Linux operating system, Phred/Phrap/Consed software, Blast software, bioinformatics, large-scale DNA sequencing

* This work was partially supported by Initiative Foundation for Scientific and Technological Innovation of Academic Military Medical Science (9905105), Chinese High-Technology "863" Program of China (863-102-10-04-04) and Chinese National Distinguished Young Scientist Award (39625014) and National Natural Sciences Foundation of China for Key Program (39730310).

** Corresponding author. Tel: 86-10-66931216, E-mail: hefc@nic.bmi.ac.cn

Received: March 6, 2000 Accepted: March 31, 2000