基于 unix 的 cDNA 序列自动分析系统的构建和应用*

符志彦, 卢 阳, 叶兰汀, 何智良, 徐安龙
(中山大学生命科学学院, 广东 广州 510275)

摘 要: 基于 unix/ linux 操作系统和 mysql 数据库, 利用 phred/ phrap, stackpack, blast 软件, 对 cDNA 和 EST 序列进行大规模自动分析。它可以完成从测序峰图文件向核酸序列的转化, 去除载体污染和重复序列, 序列聚类, 拼接, 分析可变剪切, 数据库搜索进行相似性分析。该系统可以加速大规模 EST 测序的分析速度。

关键词: cDNA 分析系统; 大规模测序; unix/ linux 操作系统; phred/ phrap 软件; stackpack 软件; blast 软件
中图分类号: Q754 文献标识码: A 文章编号: 0529-6579 (2002) 05-0060-04

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) 和 cDNA 序列在不同实验室中大量产生, 美国国家生物技术信息中心 (National Center for biotech information, NCBI) dbEST 数据库中, EST 的含量已达 2 690 828 条。它们携带着完整基因序列某些片断的信息。如今大规模的序列分析已成为瓶颈。许多生物信息研究中心开发了一系列基于 unix/ linux 操作系统的软件来满足这一需要, 如美国华盛顿大学基因组中心 (University of Washington, UWGC) 开发的 phred/phrap, 可完成从测序峰图文件向核酸序列的转化及序列拼接[2,3]。由 Electric Genetic PTY 公司开发的 stackpack 软件包是专门用于 cDNA 序列和 EST 大规模分析的软件包。由美国国家生物技术信息中心 (National Center for biotech information, NCBI) 开发的 blast 软件则可对本地数据库进行快速比对[4]。再加上 internet 的发展使得在公共数据库 (如 Genbank, EMBL, SWISS-PROT) 中的大量有价值的数据都可被下载, 使得本地化的大规模数据分析成为可能。

因此, 本文介绍一种基于 unix/ linux 操作系统, 利用上述免费软件和资源, 构建大规模 cDNA 序列分析系统。该系统完成以下功能: 测序峰图文件的转化——去除载体污染和重复序列——序列聚类——拼接——分析可变剪切——数据库搜索进行相似性分析。

1 CDNA 序列自动分析系统的构建

1.1 系统及 mysql 数据库的安装

硬件为 sgi2400, 操作系统为 IRIX6.5, 安装的 Mysql 数据库的版本为 3.23.32。Apache 版本为 1.3.14。如果用 PC 机和 Linux 操作系统, 其配置建议为 pentium III/ cpu 800 MHz/ 内存 512 M 以上, Linux 操作系统是 RedHat 6.5 以上。

1.2 phred/ phrap, stackpack, blast 软件的获得和安装

phred/phrap 软件有 UWGC 开发, 具体信息见表 1。有关信息查询可访问网页 http://bozeman.mbt.washington.edu/index.html。该软件通过 E-mail 即可获得, 解压后在相应目录下编译即可使用。

Stackpack 软件包是 Est 分析的有力工具, 它包括以下一些程序: d2cluster[6], craw[7] 由休斯敦大学 (university of Houston) 开发, 有关信息可访问网页 http://www.egenetics.com。stackpack 软件包通过

表 1 Phred/ Phrap 软件包的来源及功能

Tab. 1 Souce and Function of Phred/ Phrap software package

软件名称	联系作者和方式	功能
Phred/ phd2fasta	Brent Ewing bge@u.washington.edu	将测序峰图文件转化为核酸序列文件并生成对应的质量控制文件
Phrap/ crossmatch/ swat	Phill Green phg@u.Washington.edu	去除载体序列完成序列拼接
RepeatMasker	Arian Smit asmit@nootka.mbt.washington.edu	去除重复序列[5]

* 收稿日期: 2002-03-01
基金项目: 国家自然科学基金资助项目 (39800073); 国家自然科学基金重点资助项目 (69935020)
作者简介: 符志彦 (1978 年生), 男, 硕士研究生; 通讯联系人: 徐安龙; E-mail: ls36@zsu.edu.cn

email 即可获得。在 stackpack 安装前先要安装好以下软件：crossmatch/phrap, RepeatMasker[6], Mysql (http://www.mysql.com), Apache (http://www.apache.org), pyron (http://www.pyron.org)。然后解压，安装，在安装过程中根据提示输入相应的目录路径，安装必须有超级用户权限。blast 是做序列比对和数据库搜索的最常用工具，可直接从 NCBI 主页上下载 (ftp://ncbi.nlm.nih.gov/tools/blast/executables/)。解压后即可使用。

1.3　公共数据库的获得和格式化

从NCBI 网站上根据需要下载 nr(氨基酸序列包括非冗余的 GenBank CDS 翻译序列, PDB, SwissProt, PIR, PRF), nt(核苷酸序列包括所有的 GenBank, EMBL, DDBJ, PDB 序列但没有 EST, STS, GSS, HTGS 序列) 数据库 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) 或 dbEST 数据库 (ftp://ftp.ncbi.nlm.nih.gov/genbank/)。

Owl 数据库 (http://bmbsgi11.leeds.ac.uk/bmb5dp/owl.html) 是综合性蛋白数据库，它包括以下数据库：SWISS－PROT, NBRF PIR1, PIR2, PIR3, PIR4, Genbank, NRL 3D。数据库下载后先解压，然后用 blast 软件中的 formatdb 命令可以将 FASTA 格式的序列格式化。这样数据库就可以用于 blast。

1.4　phred 软件的使用方法

在运行 phred 软件之前先设好环境变量，如果用 C shell，则用以下命令：setenv PHRED_PARAMETER_FILE /usr/local/etc/PhredPar/phredpar.dat 然后运行：$path/phred -id chrom_dir -sa seq.fasta.qa seq.fasta.qual -trim_cutoff 0.032 -trim_alt "" -trim_out；其中 chrom_dir 是存放所有峰图文件的目录。seq.fasta 是生成的 FASTA 格式的序列文件。seq.fasta.qual 是该序列文件对应的质量控制文件。后面参数的意思是截去质量小于 15 的两端序列。也即是使生成的序列只包含质量较好的部分，丢掉测序不准的序列。

1.5　Stackpack 软件包的使用方法及功能

FASTA 格式的序列运行 stackpack 软件包，有两种运行方式：第一种是通过 web 界面，从本地上传序列到服务器，然后通过在网页上点击运行。这种方式操作比较简单，但不适合大数据量的运算。第二种方式是通过终端，以命令行的方式运行。用这种方式可以得到更强大的功能，而且可以和其它软件结合使用。

具体方法如下：

(1) 首先设置环境变量，如果用 C shell，则用以下命令：

setenv PATH ${PATH}:/usr/local/stackpack/bin

setenv LD_LIBRARY_PATH /usr/local/stackpack/lib:/usr/local/stackpack/lib.ext

(2) 创建项目，命令形式为：

stack ProjectManager -create <Project> <Project info> <Project owner>。

其中 Project 是项目名称，Project info 使该项目的描述，Project owner 是项目所有者。

(3) 项目创建好后，将序列导入数据库

它支持 genbank 格式和 FASTA 格式，有关两种格式的描述请参见文档。

倒入 FASTA 格式的命令为：

stack ImportFasta [Project] [source file]

其中 source file 为 FASTA 格式的文件。

(4) 去除载体和重复序列

可以选择去除载体和重复序列，提供相关的载体序列库和重复序列库，这些序列库在 crossmatch 和 RepeatMasker 软件中都有。执行命令：

stack Mask [Project] [Repeat File]

会自动调用软件 crossmatch 或 RepeatMasker。Repeat File 为指定的序列库。这样做是去除载体序列和重复序列。

因为重复序列如 Alu 等在序列聚类和相似性比对中可能引起干扰，冲淡结果。去除后有助得到更准确的分析结果[7]。

(5) cDNA 和 EST 的聚类

stackpack 使用了 d2_cluster 软件进行聚类，该软件是专门设计用来对 EST 和全长 cDNA 进行聚类，采用贪婪聚类算法，能够把任何有相似性的序列聚在一起。该算法在 2000 年做了改进，大大加快了运算速度。命令形式为：

stack Cluster [Project]。

(6) 序列拼接

用 d2_cluster 聚类完后在一个类 (cluster) 里会有一些相关性较小的序列，这需要进一步把它们分出来，stackpack 采用 phrap 软件和自带程序把 cluster 细分成 contig，去掉一些相关性较小的序列。这一步在聚类的基础上做大大加快了速度。命令形式为：stack Assemble [Project]

(7) 用 craw 软件分析结果

为了将聚类产生的错误减到最低，stackpack 用 craw 软件对 cluster 进行检查，并区分可变剪切。它尽可能延长一致序列 (consensus sequence)，最后得到最佳的一致序列。它将可能的可变剪切形式从

cluster 中分开。运行命令为：

stack Analysis [Project]

stackpack 运行完毕后，可通过网页的形式查看聚类结果，还可以看到拼接的具体情况（图 1）。

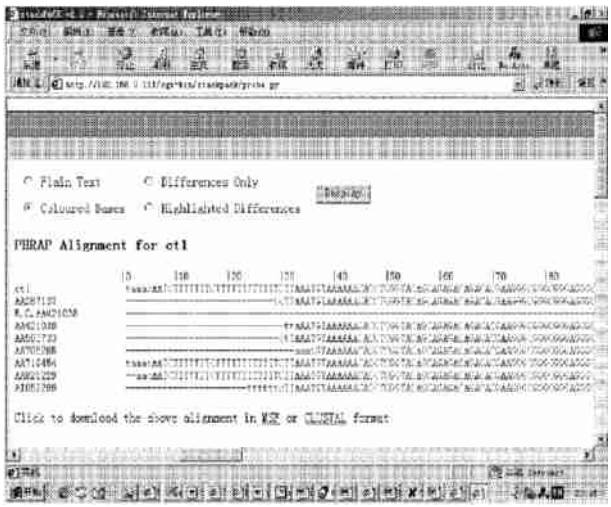

图 1 stackpack 软件显示的拼接结果

Fig. 1 Alignment showed by stackpack software

为了方便下一步的分析，stackpack 提供用网页或命令行的形式以 FASTA 格式输出所有的一致序列。这些一致序列可以用于 blast 比对或其它进一步分析。

为了检验该软件包，从 NCBI 的 unigene 数据库中（http：//www. ncbi. nlm. nih. gov/UniGene/）下载了 unigene cluster hs. 181357。该 cluster 由 2820 条 cDNA 序列和 EST 组成，属于 LAMR1（laminin receptor 1）家族。将这个 cluster 经过 stackpack 软件的处理后得到的结果如下：其中的 2811 条聚类成一个 cluster，这个 cluster 由分成 59 个 contig，有些 contig 有多于一种的剪切方式，总共得到 336 条一致序列（consensus sequence）。另外有 9 条序列因为相似程度很低，被 stackpack 软件将其从 cluster 分出来。将一致序列与 2002 年 1 月下载的 nr 库做 blastx 发现 laminin 受体中不同的抗原决定簇基本被分为不同的 contig，如 ribosomal protein SA 和 2H5 epitope。相同的 laminin 受体在一个 contig 中，如 GenBank 编号为 NM_002295，BC010418，BC005391 它们都是 ribosomal protein SA 的 mRNA 序列。结果中还提供了大量预测的可变剪切形式。Unigene 把 2820 条序列聚在一起，而作者却对它进行了更细致的聚类分析，提供了不少有价值的线索。

1.6 序列相似性分析

用 blast[8] 程序在公共数据库中搜索相似的 cDNA，EST 和蛋白序列可以帮助预测该序列的功能和其进化关系。其命令为：

blastall -p program -d database -i seq. fasta -o blast. out

其中 program 可为“blastn，blastx，blastp，tblastn，tblastx”之一。database 为要搜索的数据库。seq. fasta 为要查询的 FASTA 格式的序列文件。blast. out 为输出的结果文件。

1.7 系统的整体运行方式

我们用 perl 语言实现整个系统的整合。整个过程可以用一个程序整合在一起，自动完成从峰图文件的转化到序列聚类，再到相似性检索。各阶段的结果都可以分别保存到 mysql 数据库中，再以 web 页面的形式将分析结果显示出来。这个流程还可以继续往下扩展以满足不同的需要。

2 应用实例及结果

将本实验室测得的 187 条赤[illegible] cDNA 序列用本系统处理分析后，得到 3 个 cluster，其中两个 cluster 分别只有一个 contig。另一个 cluster 有 23 个 contig，这 23 个 contig 又由 128 条一致序列（consensus sequence）组成。这些一致序列（consensus sequence）基本上是由同一个克隆的正反测序的序列拼接而成，代表非冗余的基因。许多基因片段经过拼接得到全长的 cDNA 序列。

另外，将本实验室测得的 195 条居士鬼[illegible] cDNA 序列用本系统处理分析后，得到 25 个 cluster，他们由 27 条一致序列组成。

这些一致序列与下载的 nr 数据库作 blastx 后，比上许多有意义的蛋白，如青环素，硒蛋白，半胱氨酸蛋白酶抑制剂，IPL 基因等。还有许多是没有比对上的序列，这些序列有可能是新的基因。

3 讨 论

人类基因组计划的发展产生了大量的 cDNA 和 EST 序列，这导致了大规模分析软件的发展。在序列判读方面有许多软件，如 PE 公司设计的用于苹果机的 Sequence Analysis，用于 windows 操作系统的 Chromas 等，但这两个软件只能完成序列判读，而且往往产生判读错误且没有对应的质量控制。Sequencher 是美国基因编码公司的商业产品（http：//www. genecodes. com/），它也能完成大规模的序列判读，载体序列的去除，序列拼接等工作，然而它与 phred/phrap 软件相比没有去重复序列的功能，且是用在苹果机上，与其它软件的衔接有困难。所以 phred/phrap 软件已成为大基因组中心普遍使用的产品，在人类基因组计划中起到重要的作

用。

在 EST 聚类软件方面，Pangean system 公司的商业产品 CAT 也具有与 stackpack 相同的功能，其方法与算法都与 stackpack 类似，并声称 CAT4.0 在时间复杂度和空间复杂度上进行了大改进。但 CAT4.0 价格很高。Stackpack 软件可大规模处理，软件可免费获得，容易使用的 web 界面可从远地访问，自动去冗余及按功能聚类的功能可以监测文库质量，判断文库中基因的数目。从上面的应用实例可以看出它的结果还是较好的，聚类结果即反应了序列之间的关系也体现了它们的区别。在 的 sgi2400 上，stackpack 和 blast 软件都可以很容易的实现并行化，这样可以大大加快运算速度。

可以在这基础上进行进一步分析，比如判断聚类后的一致序列是否全长，帮助克隆到全长的基因。确定 cDNA 读码框，将它翻译为蛋白序列然后寻找 motif，domain，预测蛋白的结构和功能等。

这套系统使用的都是免费、通用的软件，平台简单低廉，移植性很好，便于大、中、小型实验室推广应用。

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[J]. 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[J]. Genome Research, 1998, 8(3): 175—185.

[3] 张成岗，欧阳曙光，贺福初，等. 基于 PC/linux 的核酸序列分析系统的构建及其应[J]. 生物化学与生物物理进展, 2001, 28(2): 263—266.

[4] ALTSSCHUL 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(17): 3389—3402.

[5] SMIT A F. Origin of interspersed repeats in the human genome[J]. Curr Opin Genet Devel, 1996, 6(6), 743—749.

[6] HIDE W, BURKE J, DAVISON D B. Biological evaluation of d2, an algorithm for high—performance sequence comparison [J]. J Comput Biol, 1994, 1(3): 199—215.

[7] CHOU A, BURKE J. CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences[J]. Bioinformatics, 1999, 15, (5): 376—381.

[8] DEININGER P L, BATZER M A. Alu repeats and human disease[J]. Mol Genet Metab, 1999, 67(3): 183—193.

Construction and Application of a Large Scale cDNA Sequences Analysis System Based on Unix

FU Zhi-yan, LU Yang, YE Lan-ting, HE Zhi-liang, XU An-long

(School of Life Sciences, Sun Yat-sen(Zhongshan) University, Guangzhou 510275, China)

Abstract: With increasing huge amount of cDNA sequences have been obtained since the human genome project, a powerful system is urgently needed for data mining these cDNA sequences. Based on unix/linux operating system, phred/phrap, stackpack and blast software have been used to construct a platform for batch analysis of cDNA and EST sequences, including base-calling, vector and repeat sequence removing, sequence clustering, assembling, alternative splicing analysis and sequence alignment. Our results demonstrated that this platform could accelerate data analysis for large scale EST sequencing and suggest some useful clues.

Key words: cDNA analysis system; large-scale DNA sequencing; unix/linux operating system; phred/phrap software; stackpack software; blast software