科技消息

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*

张成岗 张绍文 鱼咏涛 欧阳曙光 周钢桥 罗 凌 贺福初[1)]

(北京放射医学研究所, 北京 100850)

美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 的 Blast 软件是进行核酸序列和蛋白质序列同源性比较的有力工具[1, 2]. 通过访问 NCBI 的主页 (http://www.ncbi.nlm.nih.gov) 进行 Blast 同源性比较是常用的分析方法. 然而在很多场合需要在脱机状态下使用 Blast 进行同源性比较, 所以, Blast 软件的本地化实现有其必要性. 本文介绍实现 Blast 本地化过程的经验, 同时用 Visual Basic 5.0(以下简称 VB) 设计了接口程序"LocalBlast.exe", 可方便地使用 Blast 软件.

1 Blast 软件的本地化实现方法

具体的实现过程包括几个方面: 下载软件、解压缩, 进行系统配置, VB 程序接口设计.

1.1 下载软件

使用浏览器 Netscape 或 Internet Explore 访问以下网址 (表 1), 下载系统配置文件"nentrezz.exe" 和功能文件"Blastz.exe", 并将这两个文件存放到不同的子目录中, 例如: "C:\Blast\Blastz.exe" 和 "C:\nentrez\nentrezz.exe".

表 1 Blast for Windows95 软件的下载网址

软件名称	下载网址(URL)	建议保存目录
Nentrezz.exe	ftp://ncbi.nlm.nih.gov/entrez/network/mswin/win32/nentrezz.exe	C:\nentrez\
Blastz.exe	ftp://ncbi.nlm.nih.gov/Blast/executables/Blastz.exe	C:\Blast\

上述网址对映于 Windows 95 版本的 Blast 软件. 目前 (1999-04-16) 下载可以获得最新的 2.08 版. 以下操作均按照 Blast 2.08 for Windows 95 进行.

1.2 软件解压缩

运行程序: "nentrezz.exe" 和"Blastz.exe", 它们将自行解压缩, 生成以下文件(表 2):

表 2 系统配置文件 nentrezz.exe 释放以后所产生的文件

软件、目录名称	字节数	功能描述
Nentrez.exe	2 801 152	
Netentcf.exe	366 784	配置文件
DATA	目录: 754 290	内含 26 个文件
ASNLOAD	目录: 245 296	内含 36 个文件

表 3 Blastz.exe 释放以后所产生的文件

软件、目录名称	字节数	功能描述
Blastall.exe	820 736	Blast 软件的主体
Blastpgp.exe	864 256	
Formatdb.exe	576 512	对源序列进行格式化
Seedtop.exe	749 568	
DATA	目录: 4 144 650	内含 38 个文件

其中 Netentcf.exe、Formatdb.exe 和 Blastall.exe 是需要运行的文件.

* 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.

1) 通讯联系人.

收稿日期: 1999-01-07, 修回日期: 1999-04-20

1.3 进行系统配置

运行“Netentcf. exe”进行系统配置. 按照屏幕的提示回答有关问题即可.

a. Username: 此处请输入使用者的姓名.

b. Dispatcher Internet Address: 此时关闭选项“Test connection during configuration”继续.

c. Entrez Service Selection: 此时关闭选项“Test connection during configuration”继续.

d. Configuration Complete!: 此时选择“Accept”即可完成系统配置.

配置结束以后, 将会在 Windows 目录下生成一个名为“NCBI. ini”的配置文件. 至此, Blast 软件就可以使用了.

1.4 运行 Blast 软件

例如现有蛋白质序列文件“C: \ Blast \ MySeqLib. txt”和“C: \ Blast \ OneSeq. txt”, 需要对二者进行同源性比较. 那么首先需要使用“Formatdb. exe”程序对文件“C: \ Blast \ MySeqLib. txt”进行格式化. 主要的对话框见表 4.

表 4 Formatdb. exe 运行时的对话内容

标题行	说明	屏幕提示	输入示范	备注
String:	Title for database file	[空]	My Seq File	
File In:	Input file for formatting (this parameter must be set)	[空]	C: \ Blast \ MySeqLib. txt	需要格式化的序列文件名, 文件应为 FASTA 格式, 见图 1
File Out	Logfile name: Type of file1	Formatdb. log T-Protein F-nucleotide	C: \ Blast \ formatdb. log 选择 T	可选

按照程序的提示进行操作, 完成后选择“OK”. 然后运行程序“Blastall. exe”, 通过屏幕的提示在各个输入框中输入相应内容即可进行序列的同源性比较. 主要的对话框见表 5.

```
> gil 532319| pirl TVFV2E| TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRA.....
```

图 1 具有 FASTA 格式的序列文件

可参考网上文件“http: //www. ncbi. nlm. nih. gov/Blast/fasta. html”的描述

表 5 Blastall. exe 运行时的对话内容

标题行	说明	提示	输入值	输入示范	备注
String:	Program name	[空]	Blastn: 如果进行核酸序列比较 Blastp: 如果进行蛋白质序列比较	Blastp	两者任选其一, 但是必须输入
String:	Database	nr	库序列的文件名	C: \ Blast \ MySeqLib. txt	文件内容为 FASTA 格式, 见
File In:	Query File	stdin	待查询序列的文件名	C: \ Blast \ OneSeq. txt	文件内容为 FASTA 格式
File Out:	Blast report Output file	stdout	结果文件名	C: \ Blast \ Result. txt	输出为文本文件

输入结束后选择“OK”, 即可得到结果“C: \ Blast \ Result. txt”, 随后可用写字板打开此结果文件进行分析.

2 Visual Basic 程序接口设计

虽然 Blast 可以直接在 Windows 95 环境下运行, 然而, 这种方式显得十分不便. 首先, 作为库序列的文件需要使用“Formatdb. exe”程序进行

预处理后才能使用“Blastall. exe”程序进行同源性比较. 而且, 即使在 1024 × 768 的分辨率下, BlastAll 的主画面需要占据 3 个显示屏的宽度. 同时, 在“Blastall. exe”的主画面中, 仅仅提供了用于文件方式的同源性比较, 无法进行类似于 NCBI 主页上通过剪贴板方式快速进行两条序列的同源性比较, 所以最好将“Formatdb. exe”程序和“Blastall. exe”程序组合在一起, 能够实现数据格式化和同源性比较的自动化. 为此, 我们用 Visual Basic 5.0 语言设计了一个接口程序 LocalBlast. exe, 提供了一个简洁、有效的用户界面, 能够方便地使用 Blast 软件进行同源性比较. 接口程序 LocalBlast 的 VB 源程序“LocalBlast. BAS”可向作者索取, 使用 Microsoft Visual BASIC 5.0 版或以上版本编译后即可使用.

接口程序 LocalBlast. exe 中提供了两种使用 Blast 软件的方法. 一种是通过剪贴板进行同源性比较; 一种是通过文件名进行同源性比较, 后者要求文件中的序列具有 FASTA 格式. 具体使用时运行程序“LocalBlast. exe”, 主画面出现后用鼠标器选择相应的按钮, 按照提示, 就可以方便地使用本地化的 Blast 软件了. 同源性比较的结果将在程序的显示区中显示. 在此过程中, 用户不必和 Blast 软件直接打交道, 极大地方便了操作.

3 讨 论

本文较为全面地叙述了同源性比较软件 BLAST 的本地化实现过程, 并设计了相应的接口程序. 通过以上操作, 可以顺利地实现 Blast 软件的本地化快速使用. 没有上网条件的用户可以和本文作者联系, 获得有关程序. 在使用 Blast 软件的过程中出现的问题可以通过 NCBI 的电子信箱“info@ncbi. nlm. nih. gov”进行咨询.

在序列的同源性比较方面, NCBI 的 Blast (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast?Jform=1) 最为流行. 另外, FASTA 软件也较常用 (http://www2.ebi.ac.uk/fasta3), 其余的还有 Oxford Molecular Ltd. 的 Omiga 软件[3] (PC/GENE 软件的 Windows 版本)、MacVector 软件 (苹果机版本) (http://www.oxmol.com/prods/#bio), 以及 Washington University 的 Data Analysis Tools 系列软件 (http://www.genome.washington.edu/UWGC/methods.htm). 国产的软件以军事医学科学院基础医学研究所吴加金等设计的 GoldKey 软件[4] 为主. 比较起来, BLAST 软件以其快速、准确而广为使用, 所以, 本文提供的 BLAST 软件的本地化实现方法有其实用性. 事实上, 我们在对人胎肝进行大规模 cDNA 序列测定过程中, 经常需要对所测出的两条或者多条序列之间进行同源性比较, 以确定是否是同源序列. 此时, 使用本地化的 BLAST 软件就显得十分重要了.

致谢 在实现 BLAST 软件本地化的过程中, 得到了 NCBI 的 Scott McGinnis M S 的大力帮助, 特此致谢.

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 夏 云. Internet 使用技术与生物医学应用. 北京: 军事医学科学出版社 (Xia Y. Internet Technology and Application in Biology, Medicine. Beijing: Millitary Medical Science Press), 1997. 341~343

3 Calvet J P. Software: comprehensive sequence analysis. Science, 1998, **282** (5391): 1057~1058

4 李伍举, 吴加金 (Li W J, Wu J J). 蛋白质功能位点预测. 生物化学与生物物理进展 (Prog Biochem Biophys), 1993, **20** (1): 60~62