**Blast本地化：构建自己的搜索数据库**

Posted on [**2009 年 7 月 9 日**](http://boyun.sh.cn/bio/?p=1473) by [**boyun**](http://boyun.sh.cn/bio/?author=1)

blast数据库指包含fasta格式的核苷酸或者蛋白质序列的文件，有三种方式得到fasta格式的序列。

**1 从NCBI的FTP下载相关的数据库文件**

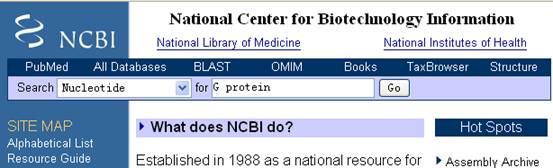
FTP地址：<ftp://ftp.ncbi.nih.gov/blast/db/>，如果你仅是一个测试，请下载一个较小的文件作为例子，例如下载大肠杆菌的核酸序列库——ecoli.nt.Z (1349 Kb)。FTP上的数据库文件都经过了压缩，有关每个文件的含义，请看该目录下说明文件（<ftp://ftp.ncbi.nih.gov/blast/db/blastdb.html> ），下表FASTA目录下文件的说明：

|  |  |  |
| --- | --- | --- |
| 文件名 | 大小（kb） | 内容 |
| alu.a.gz | 90 | alu重复序列的转录的蛋白质 |
| alu.n.gz | 24 | alu重复序列 |
| drosoph.aa.gz | 4183 | 果蝇的蛋白质序列 |
| drosoph.nt.gz | 36059 | 果蝇基因组组序列 |
| env\_nr.gz | 806964 | 环境蛋白质序列 |
| env\_nt.gz | 2298189 | 环境核酸序列 |
| est\_human.gz | 1395676 | 人的EST序列 |
| est\_mouse.gz | 774668 | 老鼠的EST序列 |
| est\_others.gz | 7360044 | 除人和老鼠以外的其他物种的EST序列 |
| gss.gz | 4772615 | 来源于GenBank, EMBL, and DDBJ的GSS序列 |
| htg.gz | 6148203 | 来源于GenBank, EMBL, and DDBJ的高通量基因组测序序列 |
| human\_genomic.gz | 2509222 | 人类基因组序列(NC\_######) |
| igSeqNt.gz | 18900 | 人和老鼠的免疫球蛋白可变区的核酸序列 |
| igSeqProt.gz | 3072 | 人和老鼠的免疫球蛋白可变区的蛋白质序列 |
| mito.aa.gz | 316 | 线粒体蛋白质序列 |
| mito.nt.gz | 935 | 线粒体基因组 |
| month.aa.gz | 50009 | 每月更新的蛋白质序列 |
| month.est\_human.gz | 1 | 每月更新的人的EST序列 |
| month.est\_mouse.gz | 1 | 每月更新的老鼠的EST序列 |
| month.est\_others.gz | 237790 | 每月更新的其他物种的EST序列 |
| month.gss.gz | 71872 | 每月更新的GSS序列 |
| month.htgs.gz | 238921 | 每月更新的htgs序列 |
| month.nt.gz | 54904 | 每月更新的核酸序列 |
| nr.gz | 1512614 | 来源于GenPept, Swissprot, PIR, PDF, PDB, and RefSeq的非冗余蛋白质序列 |
| nt.gz | 6619394 | 除wgs, gss, sts, pat, est, htg以外的核酸序列，注意不是非冗余的 |
| other\_genomic.gz | 1066358 | 人以外的其他物种的染色体序列(NC\_######) |
| pataa.gz | 65344 | 专利的蛋白质序列 |
| patnt.gz | 890761 | 专利的核酸序列 |
| pdbaa.gz | 7541 | 来源于pdb蛋白质结构数据库中的蛋白质序列 |
| pdbnt.gz | 269 | 来源于pdb的核苷酸序列 |
| sts.gz | 159256 | STS序列 |
| swissprot.gz | 86939 | swiss-prot 蛋白质序列 |
| vector.gz | 812 | 载体序列 |
| wgs.gz | 31769282 | 全基因组shotgun拼接序列 |
| yeast.aa.gz | 1906 | 酵母蛋白质序列 |
| yeast.nt.gz | 3645 | 酵母基因组序列 |

注：更新时间为2008-5-26

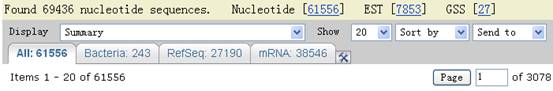
**2 从搜索结果构建数据库**

通过NCBI提供的数据库搜索引擎，选择相关数据库进行搜索，将得到的结果保存为fasta格式，作为blast比对用的数据库。以建立一个G蛋白基因核苷酸数据库为示：  
1）打开<http://www.ncbi.nlm.nih.gov/> 主页，选择核苷酸数据库（Nucleotide），输入“G protein”，点击“Go”，开始搜索



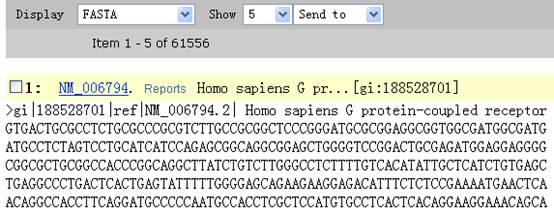
在NCBI执行搜索

2）搜索到结果页面中，在Display的下拉框中选择FASTA



切换为fasta格式

3）在Send to的下拉菜单，选择file



选择file进行保存

4）将结果保存到磁盘



保存到磁盘

**3 由测序结果直接构建自己的数据库**

测序结果，经过处理（比如去除载体、重复序列、拼接等）后，然后将结果转换为fasta格式。自己构建的数据库，数据库中fasta格式序列命名有以下三种方式：

* a） gnl|database|identifier
* b） lcl|identifier
* c） identifier

identifier为你指定的序列名称，dababase为你要指定的数据库的名称（[详细的说明参见fasta序列格式](http://boyun.sh.cn/blog/?p=1192)）