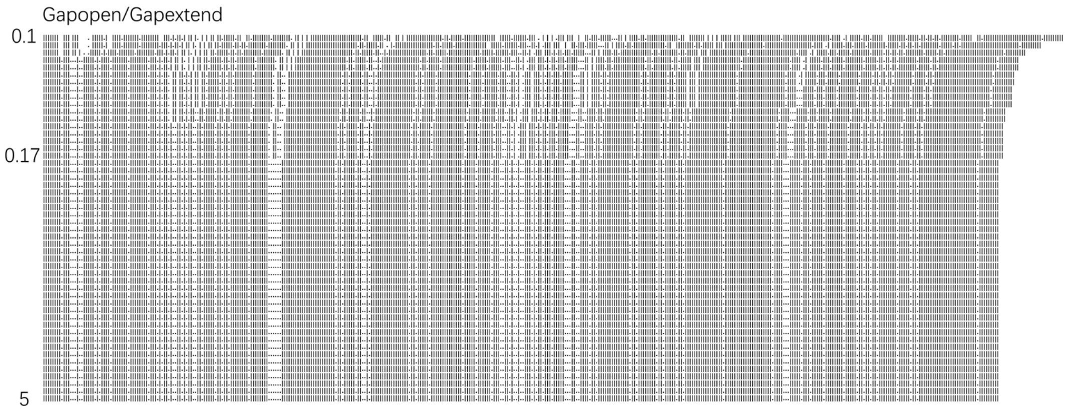
Practical 5 blast

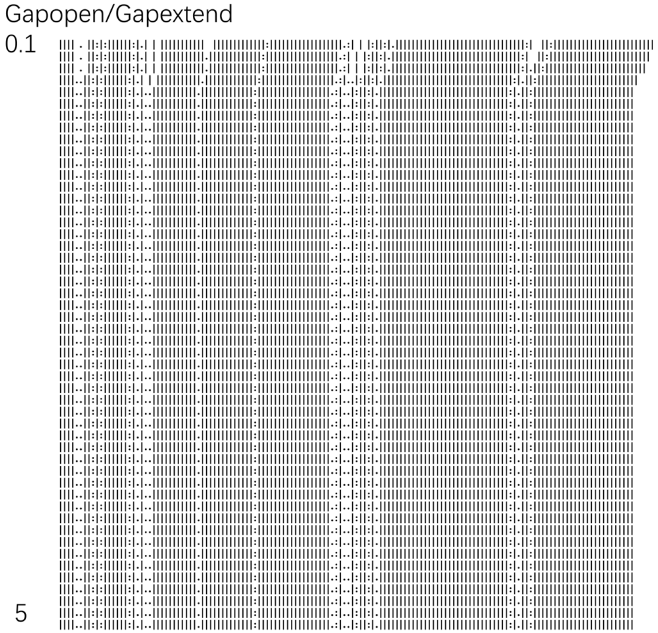
唐凯临

2021.4

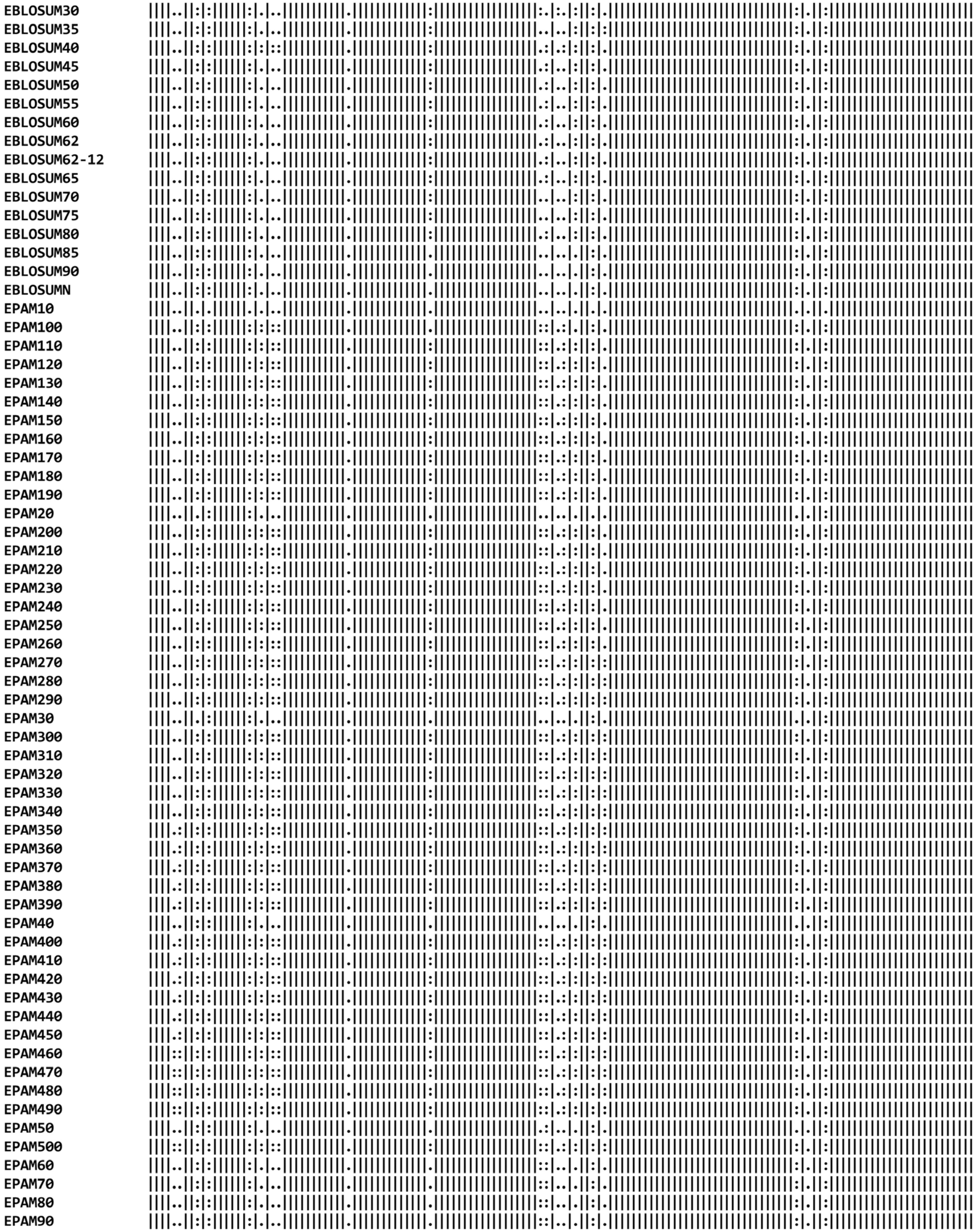
REVIEW

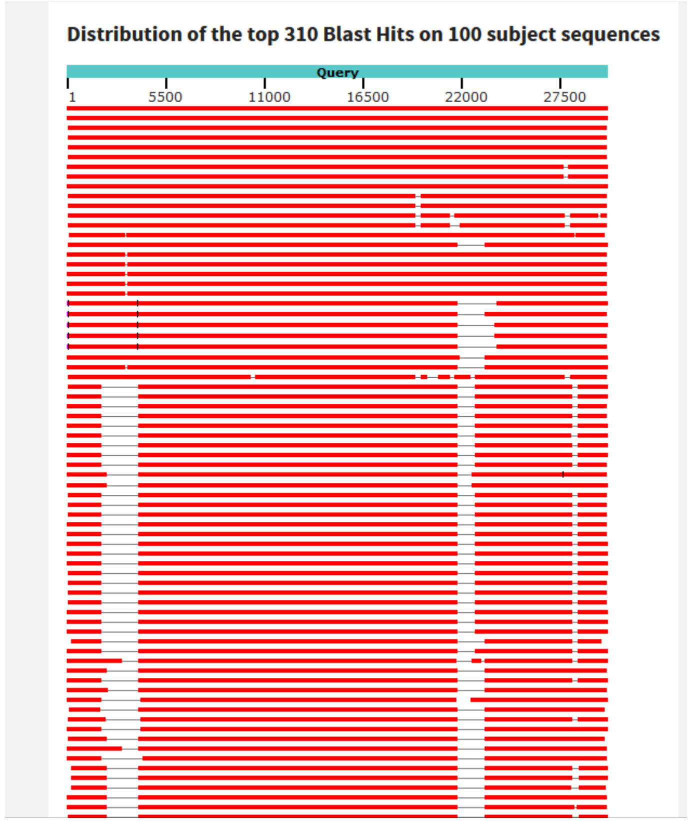
- nuclear & protein sequence



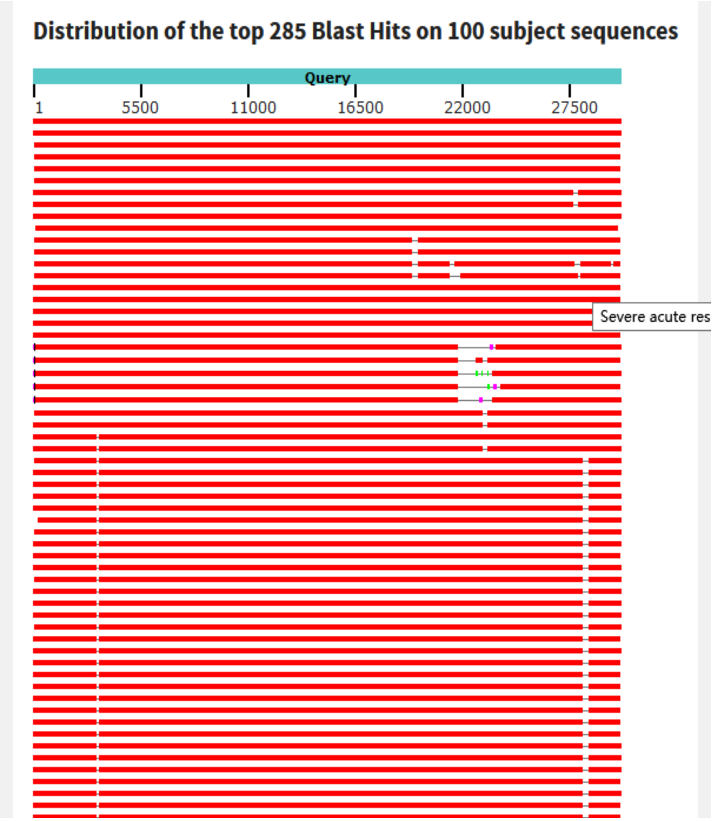


Score Matrix

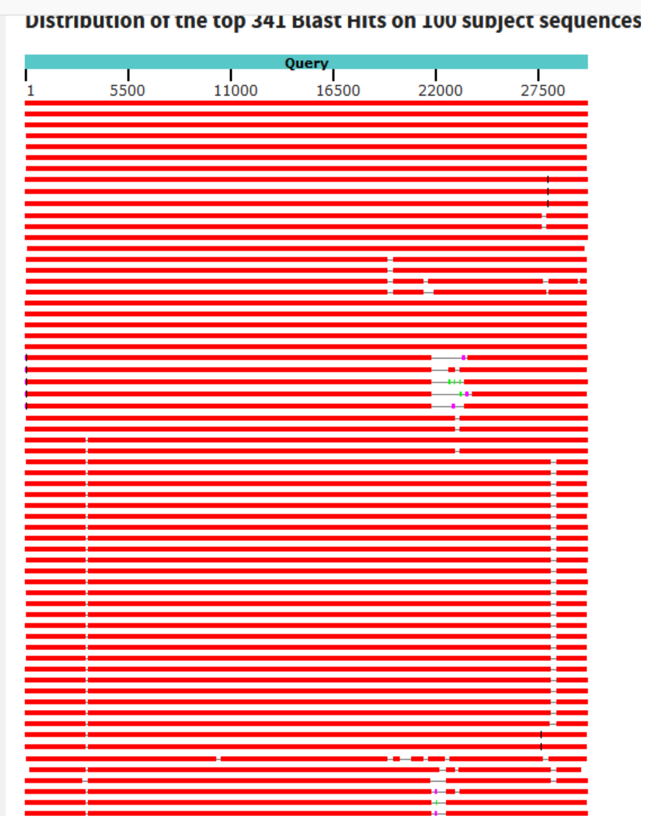




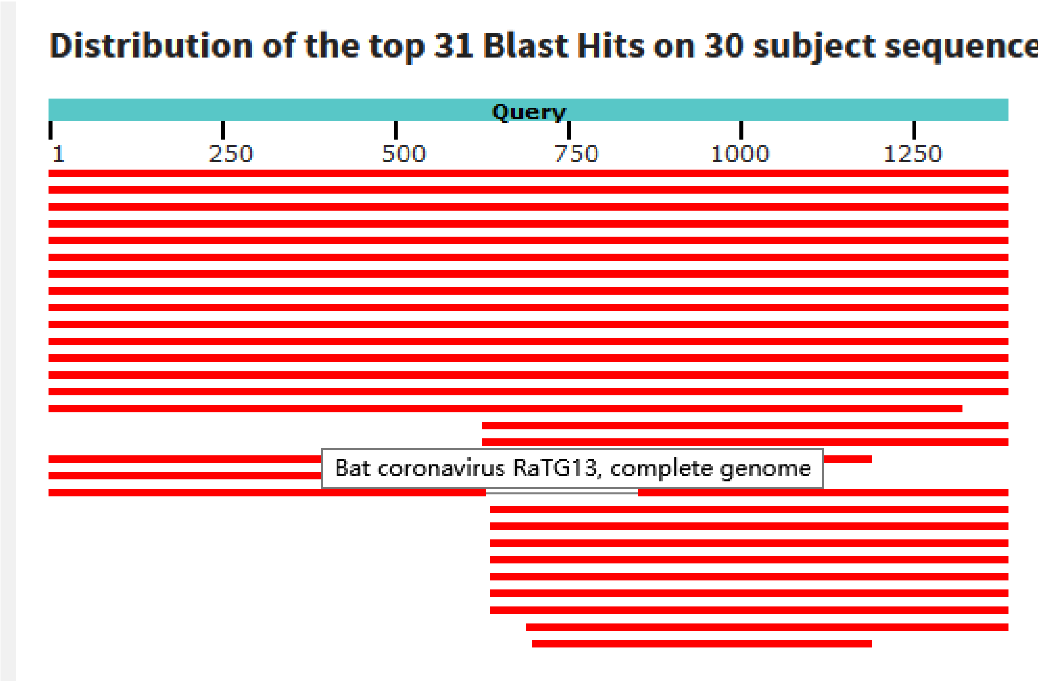
Distribution of the top 310 Blast Hits on 100 subject sequences



Distribution of the top 285 Blast Hits on 100 subject sequences



Distribution of the top 341 Blast Hits on 100 subject sequences



Distribution of the top 31 Blast Hits on 30 subject sequence

REVIEW

Query Sequence	Search Database	BLAST Program	Sequence Comparison	BLAST output
DNA	nucleotide	**blastn**	compare query nucleotide against nucleotide db	Nucleotide
DNA	protein	**blastx**	translate query seq in all reading frames into amino acids, then compare with protein db	Amino acid
DNA	nucleotide	**tblastx**	translate both query & db seq in all reading frames, then compare between protein seqs	Amino acid
Protein	protein	**blastp**	compare query protein against protein db	Amino acid
Protein	nucleotide	**tblastn**	translate db nucleotide seq in all reading frames, then compare between protein seqs	Amino acid

PREVIEW

- Which is better if you want to find representative sequences from many species?

- Which is better if you want to find all related sequences from a single species?

REVIEW

- **nr/nt** database contains **ALL known sequences reported at NCBI**

- NCBI created two databases called **RefSeq_Protein** and **RefSeq_Genomic**, designed to **reduce duplication in nr/nt** by selecting unique representative sequences for each locus

- **Swissprot or Uniprot** is a database of **highly curated protein sequences** , representing an effort to annotate/enrich all the protein sequence records in **nr**

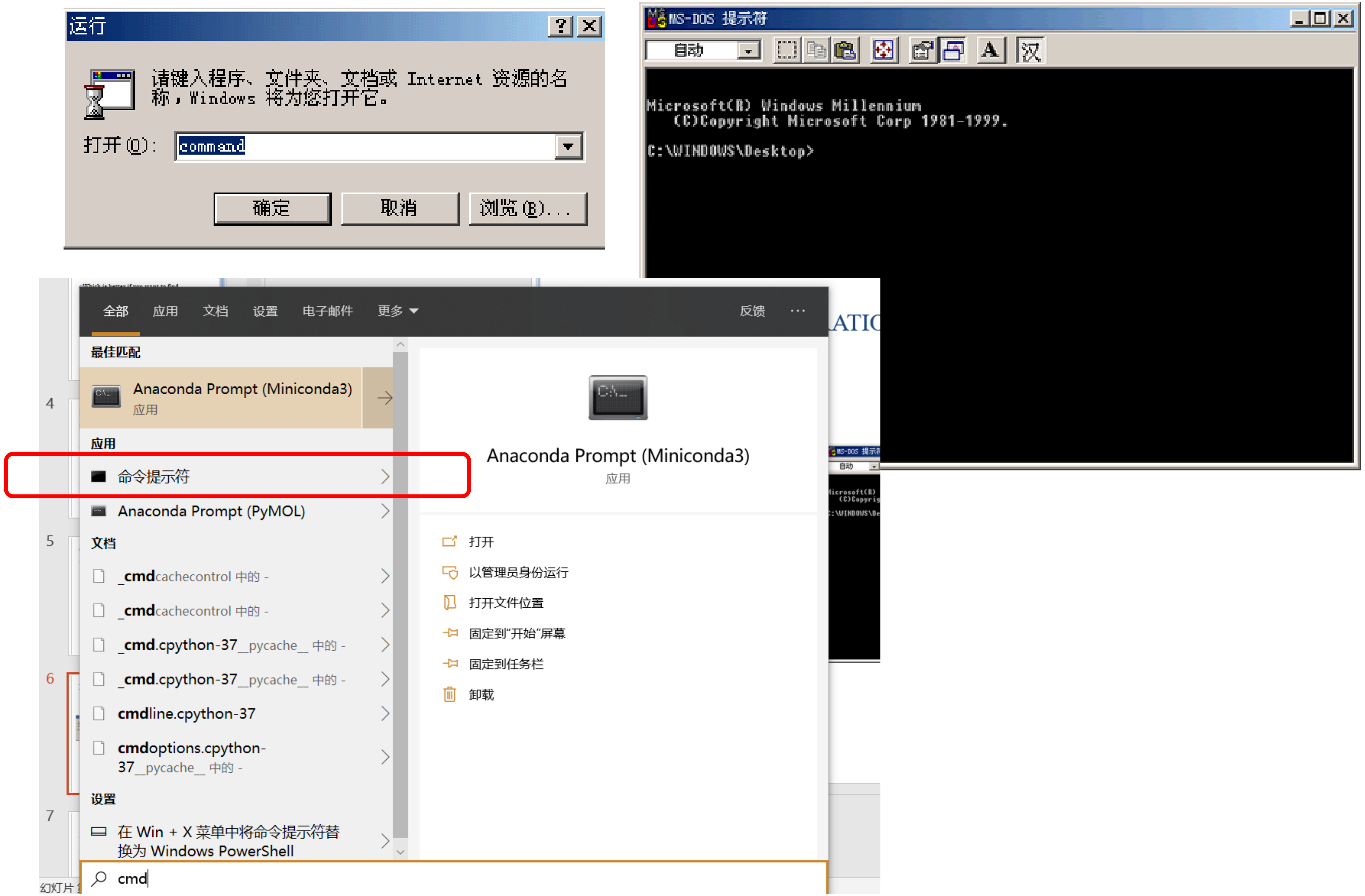
Customizing BLAST to your needs

- Build your own searchable database for a customized dataset and perform BLAST search against it

  - a set of your own sequences that are not available in the public databases.
    - eg, novel sequences from sequencing projects by your lab or collaborators.

  - a set of sequences that are available in public databases, but have been processed/organized/grouped according to your liking.
    - eg, protein sequences grouped according to their function.

DOS——Disk Operation System



运行

请键入程序、文件夹、文档或 Internet 资源的名称，Windows 将为您打开它。

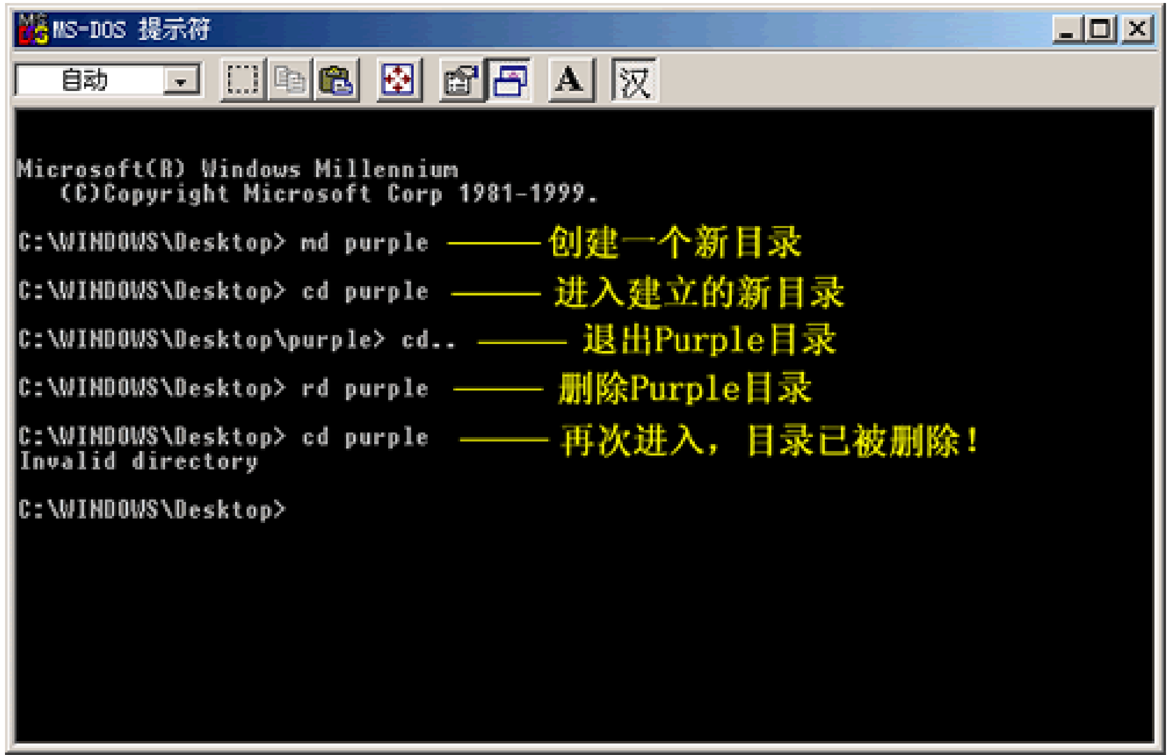
打开(O): command

```
Microsoft(R) Windows Millennium
   (C)Copyright Microsoft Corp 1981-1999.

C:\WINDOWS\Desktop>
```

cmd

常用命令

- dir [C:][path][filename][.ext][/o][/s][/p][/w][/a]
- md [C:]path
- cd [C:][path]， cd .. 返回上层目录
- rd [d:]path 不能删除非空目录；不能删除当前目录。



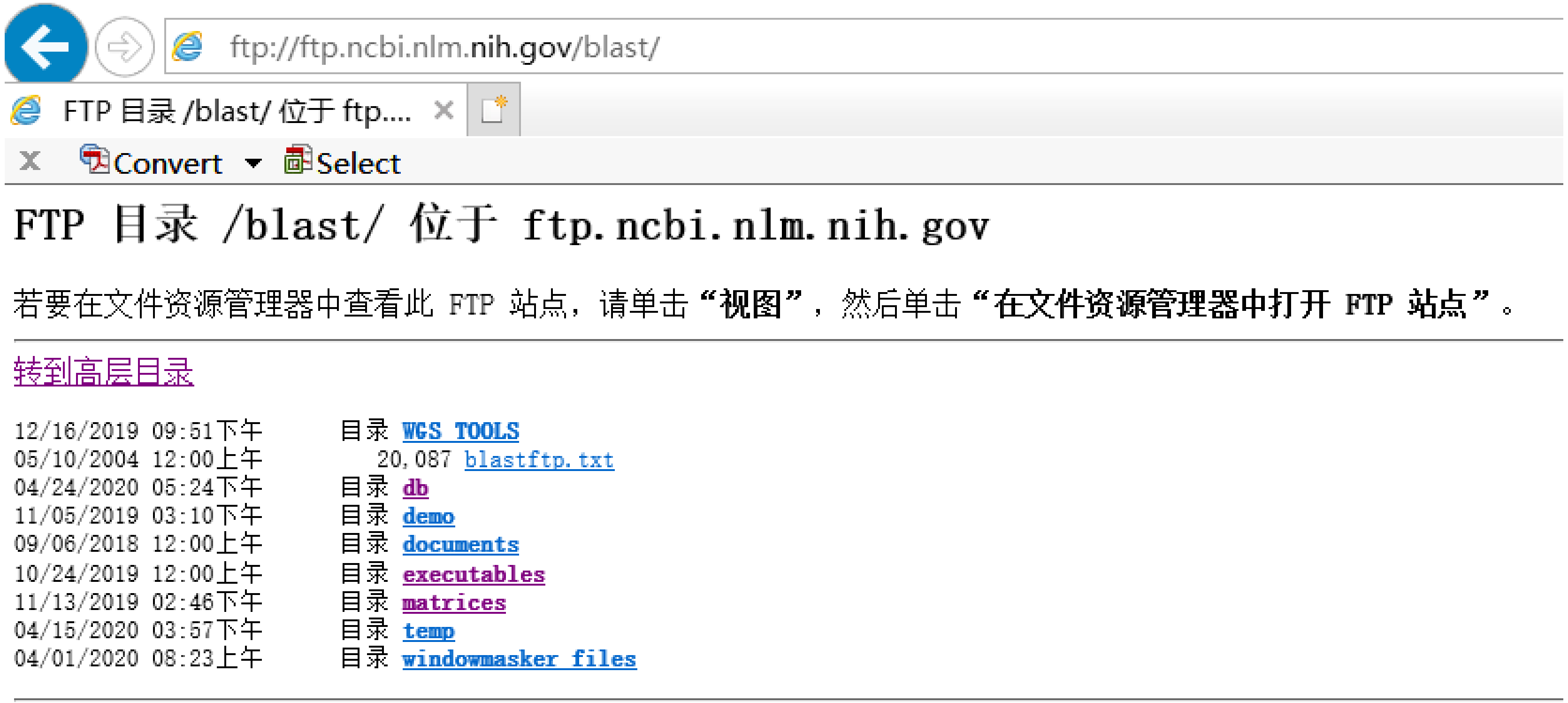
```
Microsoft(R) Windows Millennium
   (C)Copyright Microsoft Corp 1981-1999.

C:\WINDOWS\Desktop> md purple ——— 创建一个新目录

C:\WINDOWS\Desktop> cd purple ——— 进入建立的新目录

C:\WINDOWS\Desktop\purple> cd.. ——— 退出Purple目录

C:\WINDOWS\Desktop> rd purple ——— 删除Purple目录

C:\WINDOWS\Desktop> cd purple ——— 再次进入，目录已被删除！
Invalid directory

C:\WINDOWS\Desktop>
```

常用命令

- copy [C:][path][filename.ext] [C:][path]filename.ext
- del [C:][path]filename.ext
- ren [C:][path]filename1[.ext] filename2[.ext]
- cls 清除
- 系统命令 + /?　帮助
- 通配符：＊和？
  - *表示一个字符串
  - ？只代表一个字符

如何本地运行BLAST

- ftp://ftp.ncbi.nlm.nih.gov/blast/



ftp://ftp.ncbi.nlm.nih.gov/blast/
FTP 目录 /blast/ 位于 ftp....
Convert
Select

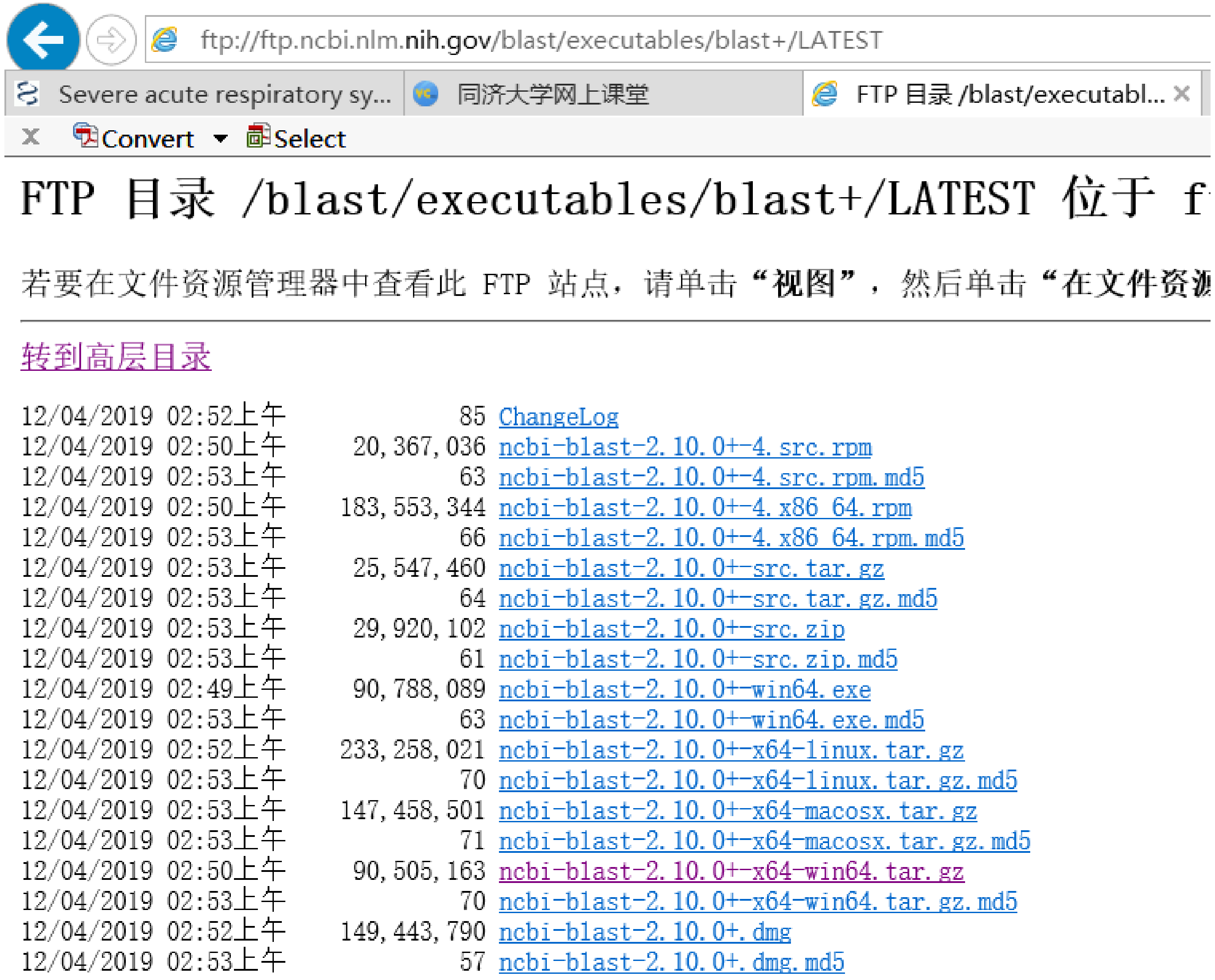
FTP 目录 /blast/ 位于 ftp.ncbi.nlm.nih.gov

若要在文件资源管理器中查看此 FTP 站点，请单击**“视图”**，然后单击**“在文件资源管理器中打开 FTP 站点”**。

转到高层目录

```
12/16/2019 09:51下午       目录 WGS_TOOLS
05/10/2004 12:00上午          20,087 blastftp.txt
04/24/2020 05:24下午       目录 db
11/05/2019 03:10下午       目录 demo
09/06/2018 12:00上午       目录 documents
10/24/2019 12:00上午       目录 executables
11/13/2019 02:46下午       目录 matrices
04/15/2020 03:57下午       目录 temp
04/01/2020 08:23上午       目录 windowmasker_files
```

Download



ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
Severe acute respiratory sy...
同济大学网上课堂
FTP 目录 /blast/executabl...
Convert
Select

FTP 目录 /blast/executables/blast+/LATEST 位于 f

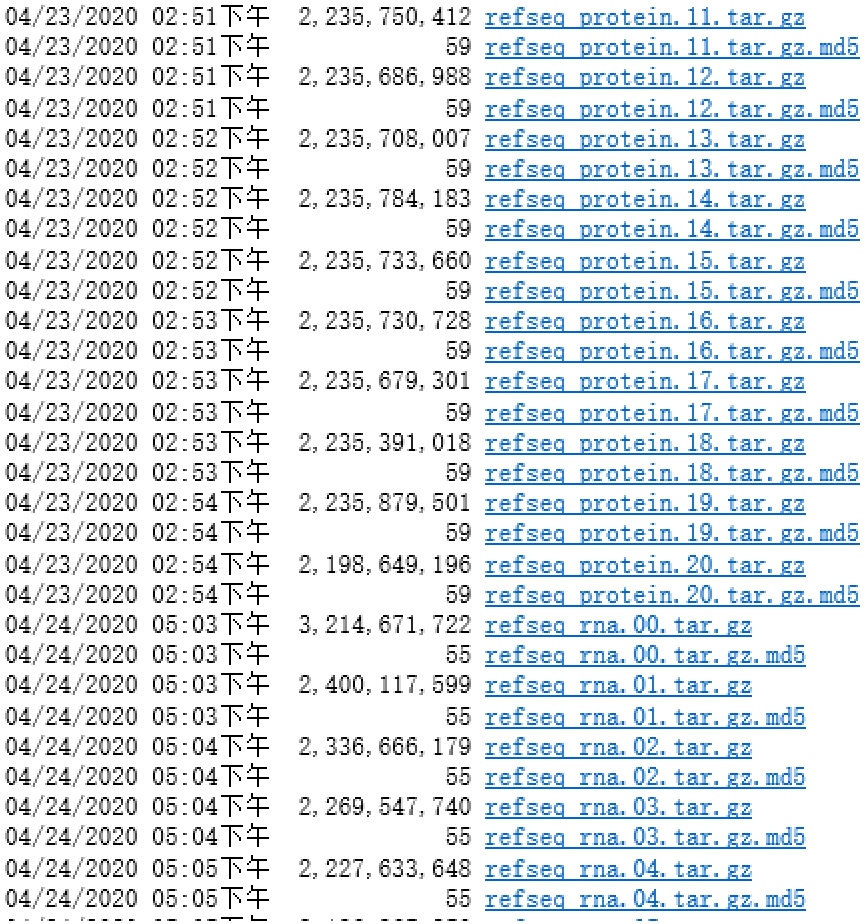
若要在文件资源管理器中查看此 FTP 站点，请单击“**视图**”，然后单击“**在文件资**

转到高层目录

```
12/04/2019 02:52上午          85 ChangeLog
12/04/2019 02:50上午  20,367,036 ncbi-blast-2.10.0+-4.src.rpm
12/04/2019 02:53上午          63 ncbi-blast-2.10.0+-4.src.rpm.md5
12/04/2019 02:50上午 183,553,344 ncbi-blast-2.10.0+-4.x86_64.rpm
12/04/2019 02:53上午          66 ncbi-blast-2.10.0+-4.x86_64.rpm.md5
12/04/2019 02:53上午  25,547,460 ncbi-blast-2.10.0+-src.tar.gz
12/04/2019 02:53上午          64 ncbi-blast-2.10.0+-src.tar.gz.md5
12/04/2019 02:53上午  29,920,102 ncbi-blast-2.10.0+-src.zip
12/04/2019 02:53上午          61 ncbi-blast-2.10.0+-src.zip.md5
12/04/2019 02:49上午  90,788,089 ncbi-blast-2.10.0+-win64.exe
12/04/2019 02:53上午          63 ncbi-blast-2.10.0+-win64.exe.md5
12/04/2019 02:52上午 233,258,021 ncbi-blast-2.10.0+-x64-linux.tar.gz
12/04/2019 02:53上午          70 ncbi-blast-2.10.0+-x64-linux.tar.gz.md5
12/04/2019 02:53上午 147,458,501 ncbi-blast-2.10.0+-x64-macosx.tar.gz
12/04/2019 02:53上午          71 ncbi-blast-2.10.0+-x64-macosx.tar.gz.md5
12/04/2019 02:50上午  90,505,163 ncbi-blast-2.10.0+-x64-win64.tar.gz
12/04/2019 02:53上午          70 ncbi-blast-2.10.0+-x64-win64.tar.gz.md5
12/04/2019 02:52上午 149,443,790 ncbi-blast-2.10.0+.dmg
12/04/2019 02:53上午          57 ncbi-blast-2.10.0+.dmg.md5
```

Download

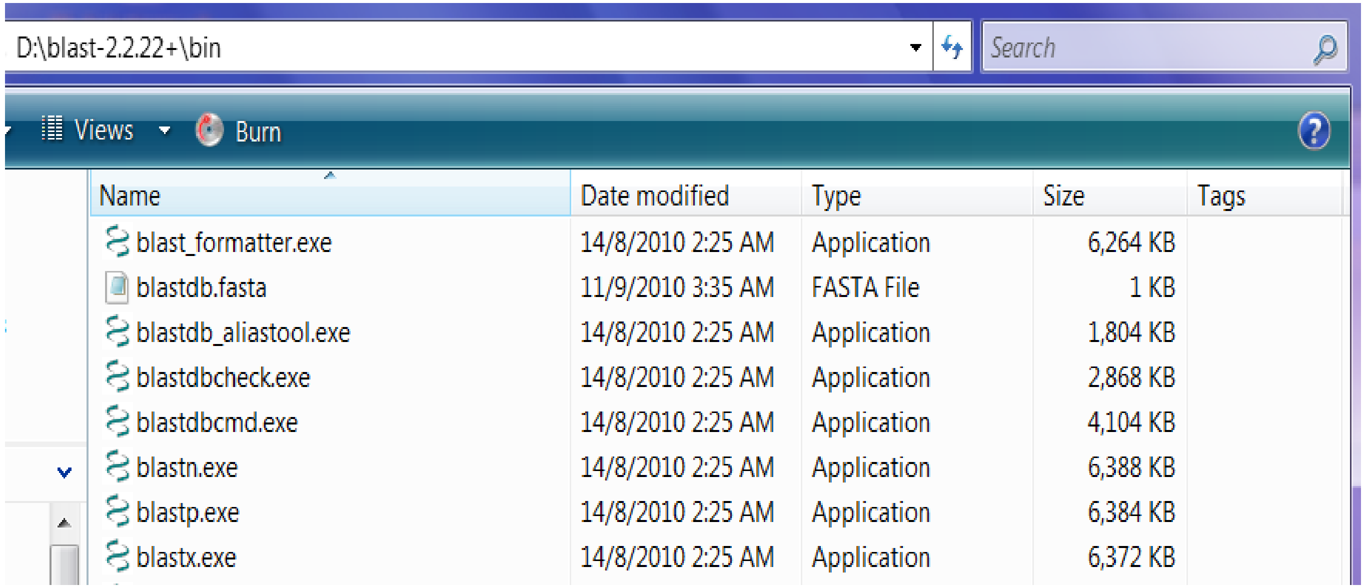
- ftp://ftp.ncbi.nlm.nih.gov/blast/db/



```
04/23/2020  02:51下午   2,235,750,412 refseq_protein.11.tar.gz
04/23/2020  02:51下午              59 refseq_protein.11.tar.gz.md5
04/23/2020  02:51下午   2,235,686,988 refseq_protein.12.tar.gz
04/23/2020  02:51下午              59 refseq_protein.12.tar.gz.md5
04/23/2020  02:52下午   2,235,708,007 refseq_protein.13.tar.gz
04/23/2020  02:52下午              59 refseq_protein.13.tar.gz.md5
04/23/2020  02:52下午   2,235,784,183 refseq_protein.14.tar.gz
04/23/2020  02:52下午              59 refseq_protein.14.tar.gz.md5
04/23/2020  02:52下午   2,235,733,660 refseq_protein.15.tar.gz
04/23/2020  02:52下午              59 refseq_protein.15.tar.gz.md5
04/23/2020  02:53下午   2,235,730,728 refseq_protein.16.tar.gz
04/23/2020  02:53下午              59 refseq_protein.16.tar.gz.md5
04/23/2020  02:53下午   2,235,679,301 refseq_protein.17.tar.gz
04/23/2020  02:53下午              59 refseq_protein.17.tar.gz.md5
04/23/2020  02:53下午   2,235,391,018 refseq_protein.18.tar.gz
04/23/2020  02:53下午              59 refseq_protein.18.tar.gz.md5
04/23/2020  02:54下午   2,235,879,501 refseq_protein.19.tar.gz
04/23/2020  02:54下午              59 refseq_protein.19.tar.gz.md5
04/23/2020  02:54下午   2,198,649,196 refseq_protein.20.tar.gz
04/23/2020  02:54下午              59 refseq_protein.20.tar.gz.md5
04/24/2020  05:03下午   3,214,671,722 refseq_rna.00.tar.gz
04/24/2020  05:03下午              55 refseq_rna.00.tar.gz.md5
04/24/2020  05:03下午   2,400,117,599 refseq_rna.01.tar.gz
04/24/2020  05:03下午              55 refseq_rna.01.tar.gz.md5
04/24/2020  05:04下午   2,336,666,179 refseq_rna.02.tar.gz
04/24/2020  05:04下午              55 refseq_rna.02.tar.gz.md5
04/24/2020  05:04下午   2,269,547,740 refseq_rna.03.tar.gz
04/24/2020  05:04下午              55 refseq_rna.03.tar.gz.md5
04/24/2020  05:05下午   2,227,633,648 refseq_rna.04.tar.gz
04/24/2020  05:05下午              55 refseq_rna.04.tar.gz.md5
```

Run Local BLAST client

- Download the correct version of BLAST for your computer
- Install BLAST;
- copy query & database fasta files into your folder
- Customize BLAST database & run BLAST search through command window



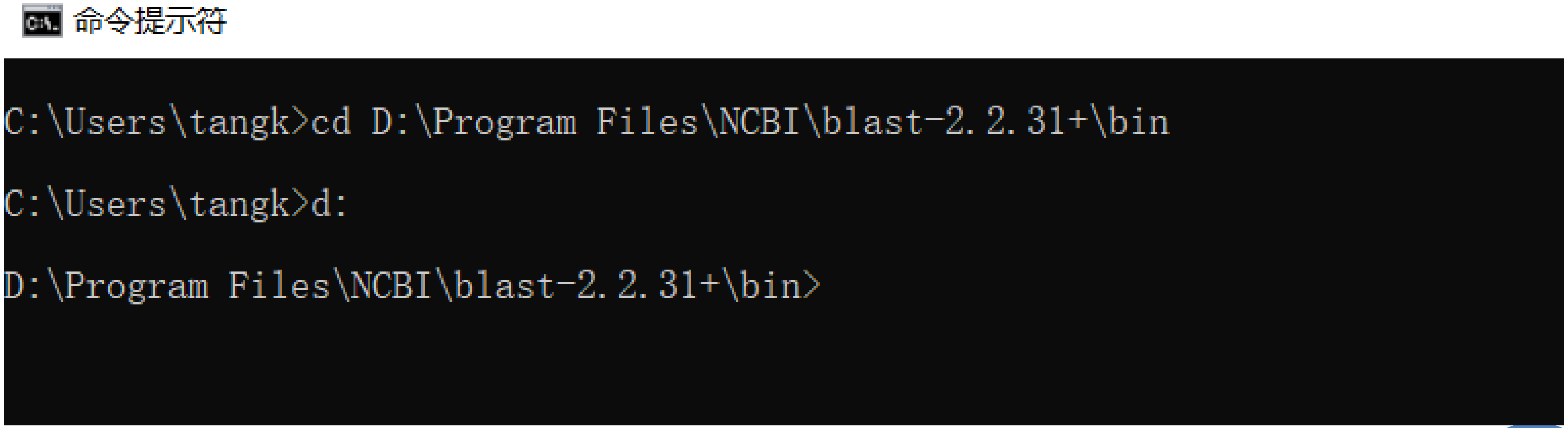
D:\blast-2.2.22+\bin

Name	Date modified	Type	Size	Tags
blast_formatter.exe	14/8/2010 2:25 AM	Application	6,264 KB	
blastdb.fasta	11/9/2010 3:35 AM	FASTA File	1 KB	
blastdb_aliastool.exe	14/8/2010 2:25 AM	Application	1,804 KB	
blastdbcheck.exe	14/8/2010 2:25 AM	Application	2,868 KB	
blastdbcmd.exe	14/8/2010 2:25 AM	Application	4,104 KB	
blastn.exe	14/8/2010 2:25 AM	Application	6,388 KB	
blastp.exe	14/8/2010 2:25 AM	Application	6,384 KB	
blastx.exe	14/8/2010 2:25 AM	Application	6,372 KB	

改变路径

- 改变目录：cd

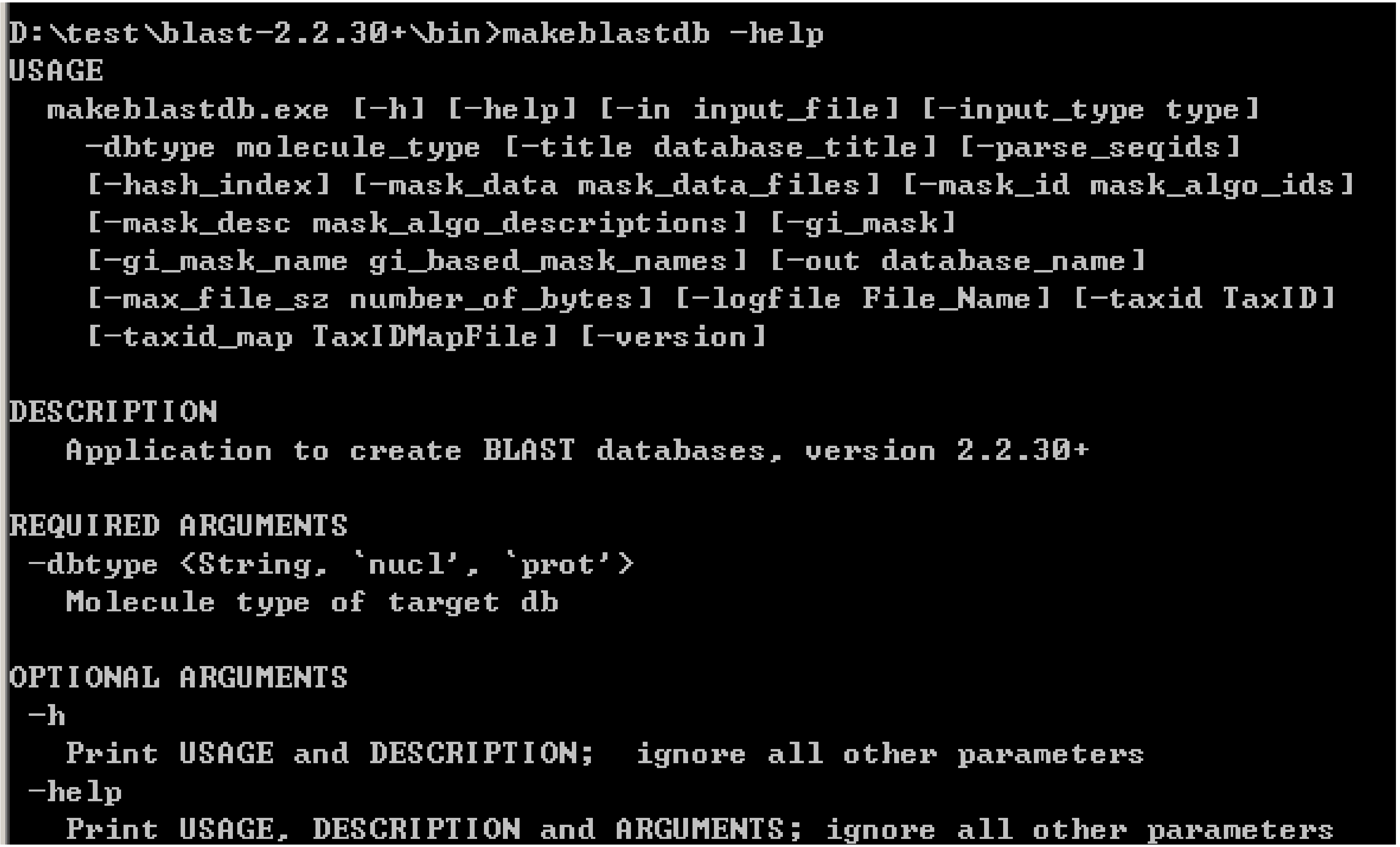
- 改变盘符



```
C:\Users\tangk>cd D:\Program Files\NCBI\blast-2.2.31+\bin

C:\Users\tangk>d:

D:\Program Files\NCBI\blast-2.2.31+\bin>
```

COMMAND

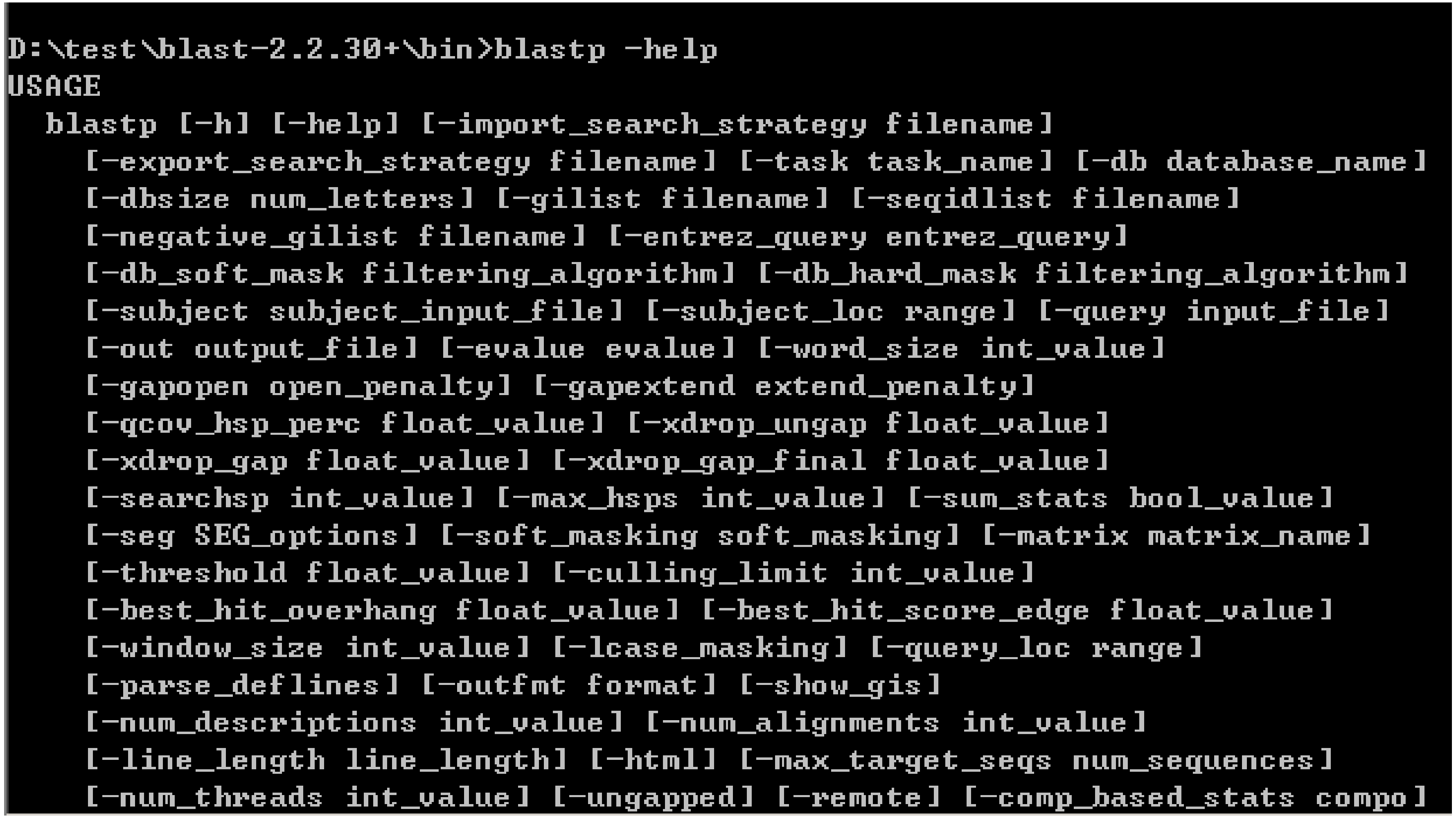
- makeblastdb -help



```
D:\test\blast-2.2.30+\bin>makeblastdb -help
USAGE
  makeblastdb.exe [-h] [-help] [-in input_file] [-input_type type]
    -dbtype molecule_type [-title database_title] [-parse_seqids]
    [-hash_index] [-mask_data mask_data_files] [-mask_id mask_algo_ids]
    [-mask_desc mask_algo_descriptions] [-gi_mask]
    [-gi_mask_name gi_based_mask_names] [-out database_name]
    [-max_file_sz number_of_bytes] [-logfile File_Name] [-taxid TaxID]
    [-taxid_map TaxIDMapFile] [-version]

DESCRIPTION
   Application to create BLAST databases, version 2.2.30+

REQUIRED ARGUMENTS
 -dbtype <String, `nucl', `prot'>
   Molecule type of target db

OPTIONAL ARGUMENTS
 -h
   Print USAGE and DESCRIPTION;  ignore all other parameters
 -help
   Print USAGE, DESCRIPTION and ARGUMENTS; ignore all other parameters
```

COMMAND

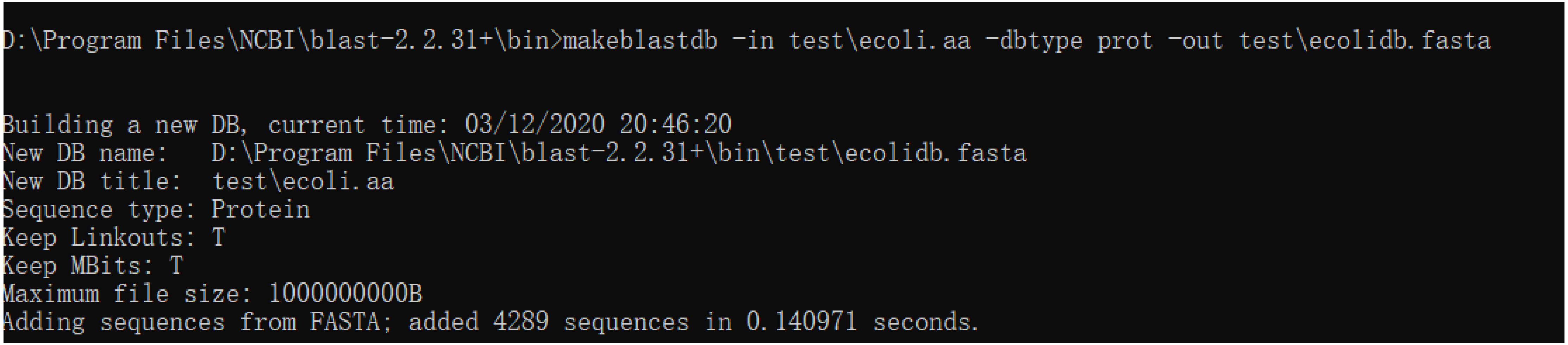
- blastp -help



```
D:\test\blast-2.2.30+\bin>blastp -help
USAGE
  blastp [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-qcov_hsp_perc float_value] [-xdrop_ungap float_value]
    [-xdrop_gap float_value] [-xdrop_gap_final float_value]
    [-searchsp int_value] [-max_hsps int_value] [-sum_stats bool_value]
    [-seg SEG_options] [-soft_masking soft_masking] [-matrix matrix_name]
    [-threshold float_value] [-culling_limit int_value]
    [-best_hit_overhang float_value] [-best_hit_score_edge float_value]
    [-window_size int_value] [-lcase_masking] [-query_loc range]
    [-parse_deflines] [-outfmt format] [-show_gis]
    [-num_descriptions int_value] [-num_alignments int_value]
    [-line_length line_length] [-html] [-max_target_seqs num_sequences]
    [-num_threads int_value] [-ungapped] [-remote] [-comp_based_stats compo]
```

建库

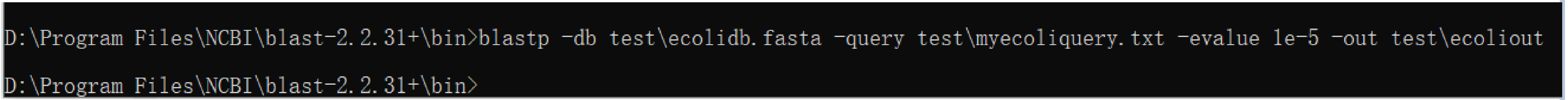
- makeblastdb -in test\ecoli.aa -dbtype prot -out test\ecolidb.fasta



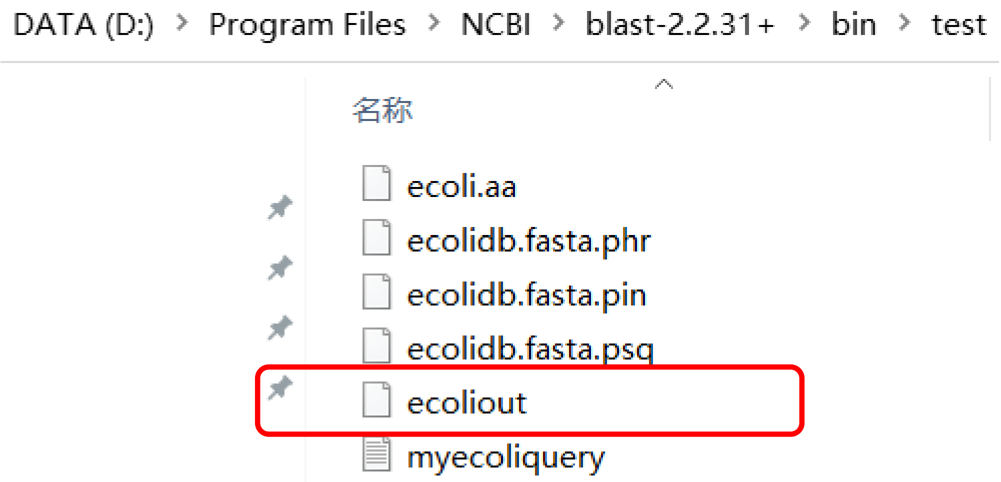
```
D:\Program Files\NCBI\blast-2.2.31+\bin>makeblastdb -in test\ecoli.aa -dbtype prot -out test\ecolidb.fasta


Building a new DB, current time: 03/12/2020 20:46:20
New DB name:   D:\Program Files\NCBI\blast-2.2.31+\bin\test\ecolidb.fasta
New DB title:  test\ecoli.aa
Sequence type: Protein
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 4289 sequences in 0.140971 seconds.
```

比对

- blastp -db test\ecolidb.fasta -query test\myecoliquery.txt -num_alignments 1 -evalue 1e-5 -out test\ecoliout



```
D:\Program Files\NCBI\blast-2. 2. 31+\bin>blastp -db test\ecolidb. fasta -query test\myecoliquery. txt -evalue 1e-5 -out test\ecoliout

D:\Program Files\NCBI\blast-2. 2. 31+\bin>
```

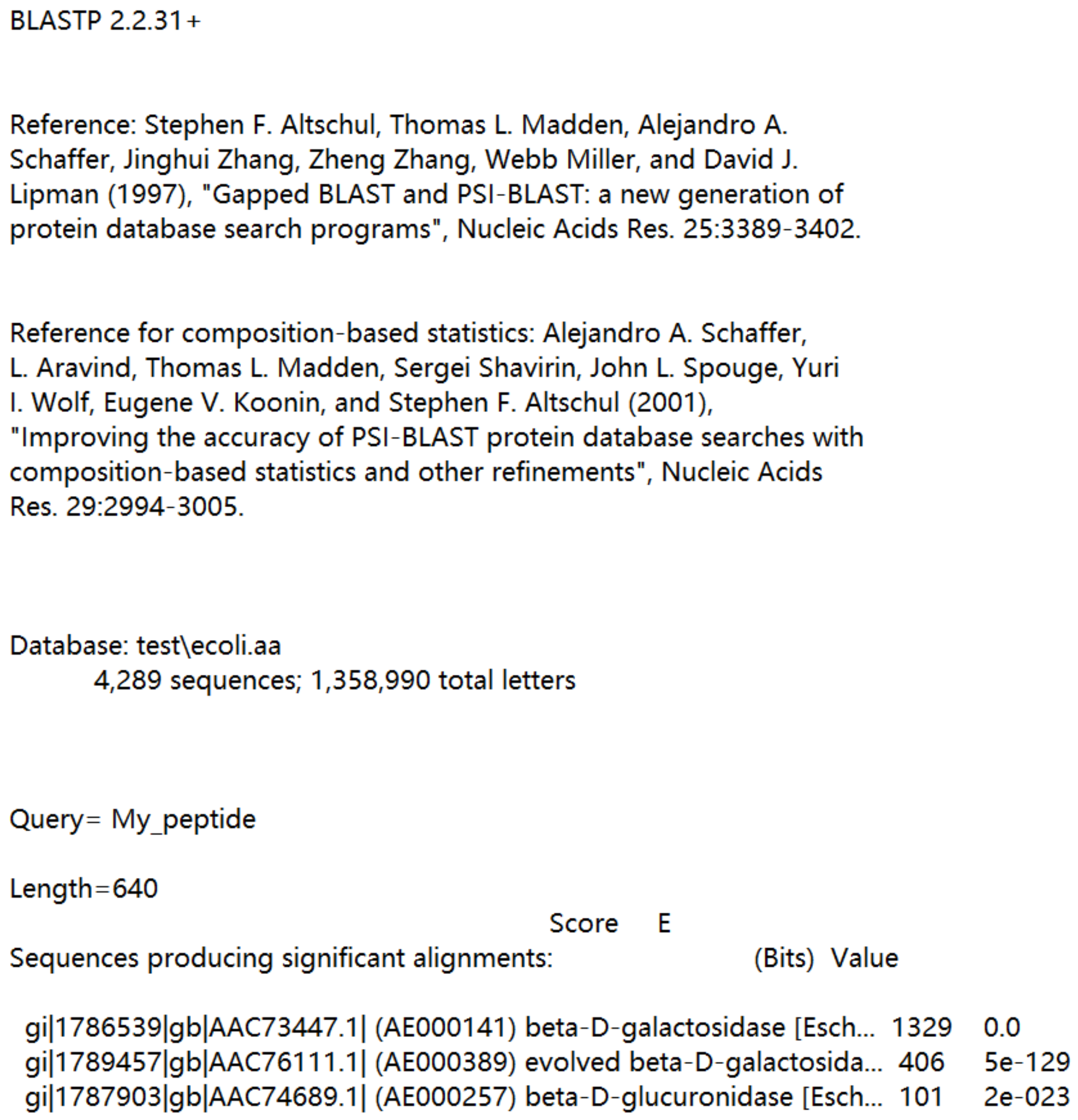


DATA (D:) › Program Files › NCBI › blast-2.2.31+ › bin › test

名称

- ecoli.aa
- ecolidb.fasta.phr
- ecolidb.fasta.pin
- ecolidb.fasta.psq
- ecoliout
- myecoliquery

结果



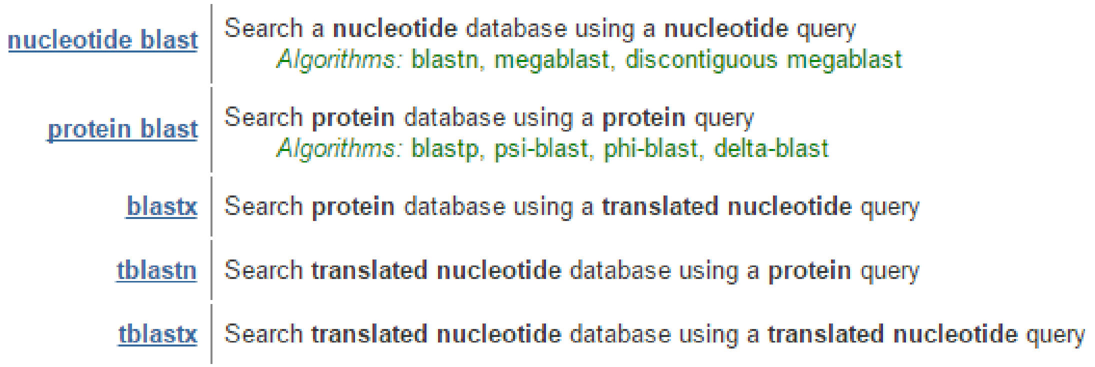
```
BLASTP 2.2.31+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.



Database: test\ecoli.aa
           4,289 sequences; 1,358,990 total letters



Query= My_peptide

Length=640
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  gi|1786539|gb|AAC73447.1| (AE000141) beta-D-galactosidase [Esch...  1329    0.0
  gi|1789457|gb|AAC76111.1| (AE000389) evolved beta-D-galactosida...   406    5e-129
  gi|1787903|gb|AAC74689.1| (AE000257) beta-D-glucuronidase [Esch...   101    2e-023
```

TIPS

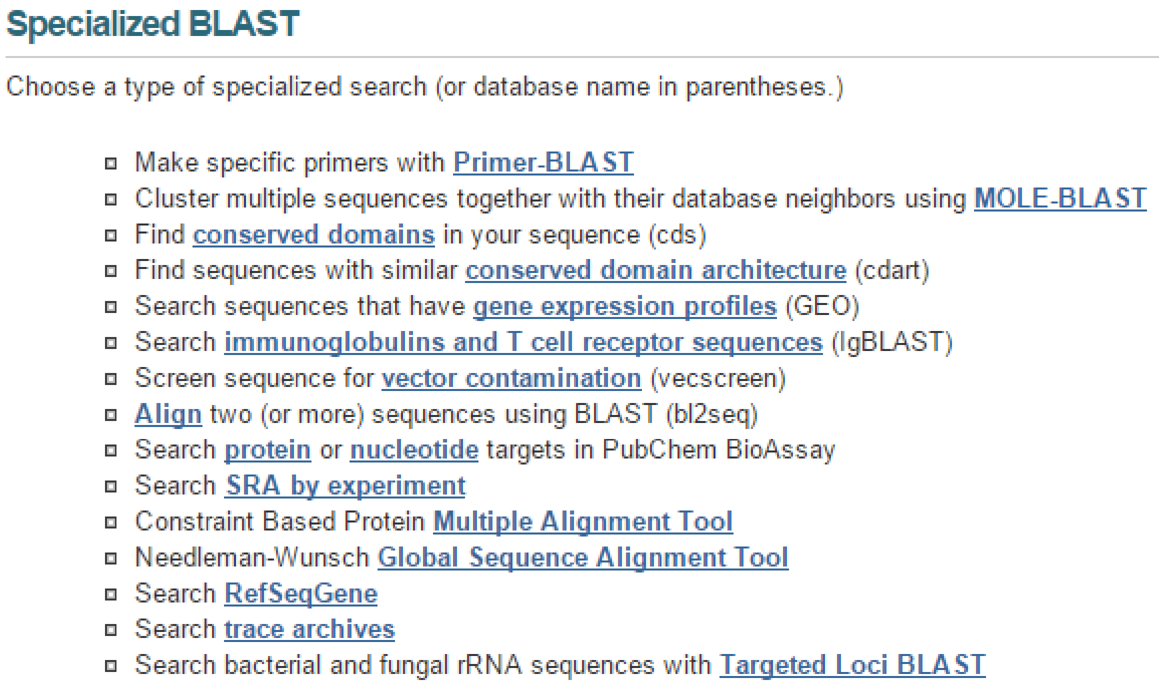
- Too much results
  - Refseq
  - Weight matrix
  - E ↓
  - ……

- Less results
  - Nr/nt
  - E↑
  - Weight matrix
  - Word ↓
  - ……

基本的BLAST与特别的BLAST



nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query



Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Cluster multiple sequences together with their database neighbors using MOLE-BLAST
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA by experiment
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search trace archives
- Search bacterial and fungal rRNA sequences with Targeted Loci BLAST

SUMMARY OF BLAST

- Introduction on BLAST
  - What is BLAST ?
  - BLAST flavors ?
  - BLAST databases ?
  - BLAST Access: Web Blast ? Local BLAST Client?
- Application of BLAST
  - How BLAST can be applied for life science research ?
    - eg, sequence identity verification; conserved domains; similar genes/proteins; distant relatives; homology; etc.
  - Advantages/Disadvantages of BLAST