Advanced BLAST

Slides

Topics

- words, scores & expect values
- BLAST programs

Browser

Tips for Web BLAST

BLAST FAQs for avoiding "Error: CPU usage limit was exceeded, resulting in SIGXCPU."

- avoid tblastx
- select a smaller (subset) database
    - Organism limit
    - Entrez query limit
    - GI List demo
- Adjust Expect value threshold
- Adjust word size (nuc-nuc comparisons)
- Adjust filter settings
- give up and use another BLAST approach

Slide (Yes, only one)

Alternatives to web BLAST

1) Blast+ package (= standalone blast = command-line blast = blast+ executables)

2) Local SRA blast (blast_vdb; SRA toolkit)

3) Cloud BLAST

4) URL API

Browser

BLAST Help page

Where to get the blast+ package: "Download BLAST Software and Databases"

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

BLAST FTP site

Blast+ package:

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

- Choose appropriate platform

NCBI blast databases:

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

- CAVEAT: is it a subset database? (pdbaa, pdbnt, swissprot)
- Use update_blastdb.pl to download databases under /blast/db

BLAST FASTA files:

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

- Will need to run makeblastdb on unpacked FASTA

Developer information (URL API and Source code)

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=DeveloperInfo

- REST = URL API. Get documentation and sample code here
- Get source code links here

Command-line

Manual: http://www.ncbi.nlm.nih.gov/books/NBK1763/

Installation help: http://www.ncbi.nlm.nih.gov/books/NBK1763/#CmdLineAppsManual.Installation

Test installation:

$ ./blastn -help (this should list all blastn command line options)

Run a search:

- Download a query file, "10.pfa"

ftp://ftp.ncbi.nlm.nih.gov/pub/education/regional_workshops/2014/NCentral_California_Sept/Day2/AM_Advanced_BLAST/10.pfa

- Make sure file is in /bin/

Send a search to NCBI by using the "-remote" option:

$ ./blastp -query 10.pfa -db pdbaa -remote -evalue 1e-6 -out 10.remote.out &

Note: with "-remote" you can use "-entrez_query" to search a custom database

Run the same search locally

- Install a database

$ mkdir db (same level as /bin)

$ cd db

- retrieve database
  - For databases in /blast/db, use update_blastdb.pl

    $ perl update_blastdb.pl -d -p <database name>

    (-d = decompress; -p = passive)
  - For databases in /blast/db/FASTA

    $ wget "ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/pdbaa.gz"

    or

    $ curl –o pdbaa.gz "ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/pdbaa.gz"

    $ gunzip pdbaa.gz
- Format the database (in /db/)

  $ ../bin/makeblastdb -dbtype prot -in pdbaa -parse_seqids -out pdbaa

Run the local search:

$ ./blastp -query 10.pfa -db ../db/pdbaa -evalue 1e-6 -out 10.local.out &

Command line tips

- use multiple cpu: -num_threads 2
- limit to a taxonomic group
  - gi list
  - blastdb_aliastool
- customize your result format
  - see options under "-outfmt"
  - for scientific or common organism names, need to download taxdb.tar.gz:

    ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz

    Place unpacked files (2) in /bin.

Local SRA BLAST

Beta version (only for Linux; new SRA Toolkit and blast_vdb binaries soon):
ftp://ftp.ncbi.nlm.nih.gov/blast/temp/vdb_test.tar

- Unpacks to: blastn_vdb
- No need to download the databases; the program fetches them on-the-fly

Example search (matches a search in the Gene Expression module):

$ ./blastn_vdb -query nm_001025366 -db "SRR633599 SRR633600 SRR633601 SRR633602" -evalue 1e-7 -word_size 7 -outfmt 7 -max_target_seqs 10000 -out vegf_fots.10K.out -num_threads 4 &

BLAST in the cloud

Browser

BLAST Help page

"BLAST Searches at a Cloud Provider"

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=CloudBlast

http://blast.ncbi.nlm.nih.gov/blast/docs/**BLAST_at_AWS.pdf**

http://blast.ncbi.nlm.nih.gov/blast/docs/aws_blast.pl (script for URL API)

Webinar video: http://www.youtube.com/watch?v=wLm-RfdcvnU

BLAST URL API

Documentation: http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html

Sample Perl script: http://www.ncbi.nlm.nih.gov/blast/docs/web_blast.pl

Basic URL

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?QUERY=AAB94045.1&DATABASE=nr&HITLIST_SIZE=10&EXPECT=1E-6&PROGRAM=blastp&CLIENT=web&CMD=Put

Via Perl script, web_blast.pl

- Use this command to see usage and examples:

$ perl web_blast.pl

usage: web_blast.pl program database query [query]...

where program = megablast, blastn, blastp, rpsblast, blastx, tblastn, tblastx

example: web_blast.pl blastp nr protein.fasta

example: web_blast.pl rpsblast cdd protein.fasta

example: web_blast.pl megablast nt dna1.fasta dna2.fasta

- Need to modify script to change default search settings. Modify line #95,

  $args = "CMD=Put&PROGRAM=$program&DATABASE=$database&QUERY="

  E.g., add &EXPECT=1E-6. Get syntax from the documentation.

On the cloud BLAST instance: Use aws_blast.pl