ECOGEO 'Omics Training: 4.2 Annotation Version 2

Rekha Seshadri

Abstract

Introduction to functional annotation and Integrated Microbial Genomes (IMG) at the Joint Genome Institute (JGI).

Open this protocol inside the virtual machine (details in 'Start Instructions') for easy copy, paste of commands into the command line terminal window.

Citation: Rekha Seshadri ECOGEO 'Omics Training: 4.2 Annotation. **protocols.io**
dx.doi.org/10.17504/protocols.io.fi8bkhw
Published: 18 Aug 2016

Guidelines

BLAST has multiple output options:

-outfmt
<String>
alignment view options:
0 = pairwise,
1 = query-anchored showing identities,
2 = query-anchored no identities,
3 = flat query-anchored, show identities,
4 = flat query-anchored, no identities,
5 = XML Blast output,
6 = tabular,
7 = tabular with comment lines,
8 = Text ASN.1,
9 = Binary ASN.1,
10 = Comma-seperated values,
11 = BLAST archive format (ASN.1)

BLAST - tabular output (fmt = 6)

Lots of custom format options for formats 6, 7, 10

qseqid means Query Seq-id
qgi means Query GI
qacc means Query accesion
qaccver means Query accesion.version
sallgi means All subject GIs
sacc means Subject accession
saccver means Subject accession.version
sallacc means All subject accessions

qlen means Query sequence length

sseqid means Subject Seq-id

sallseqid means All subject Seq-id(s), separated by a ';'

sgi means Subject GI

slen means Subject sequence length

qstart means Start of alignment in query

qend means End of alignment in query

sstart means Start of alignment in subject

send means End of alignment in subject

Before start

Before starting, please visit the ECOGEO website for more information on this "Introduction to Environmental 'Omics" training series. The site contains a pre-packaged virtual machine that can be downloaded and used to run all of the protocols in this protocols.io collection. In addition to the VM, the website contains video and presentations from our initial "Intro to Env 'Omics" workshop held at the Univ. of Hawai'i at Manoa on 25-26 Jul 2016.

Please email ‘ecogeo-join@earthcube.org’ to join the ECOGEO listserv for future updates.

Protocol

Local BLAST Database

Step 1.

BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences (Wikipedia).

The following hands-on exercises utilize a genomic bin from the TARA Ocean Project data set from the Mediterranean Sea. Collection contains marker genes for carbon fixation:

CBB, Wood-Ljundahl, reductive TCA, 3-hydroxypropionate, 3-hydroxypropionate/4-hydroxybutyrate

As a cyanobacteria, which carbon fixation pathway is being used?

Putative taxonomy → Cyanobacteria

35 contigs

1,585 putative CDS (as determined by Prodigal)

Approx. 64.64% complete (1.29% redudncancy)

tara_med_examplegenome.fasta & orfs.faa

Local BLAST Database

Step 2.

Create a BLAST index of 'subject' sequences.collection of carbon fixation related genes

cmd COMMAND

```
$ makeblastdb -in carbonfixation_markergenes.faa -dbtype prot
```

Creates 3 index files that end in *phr, *pin, *psq

Local BLAST Data

Step 3.

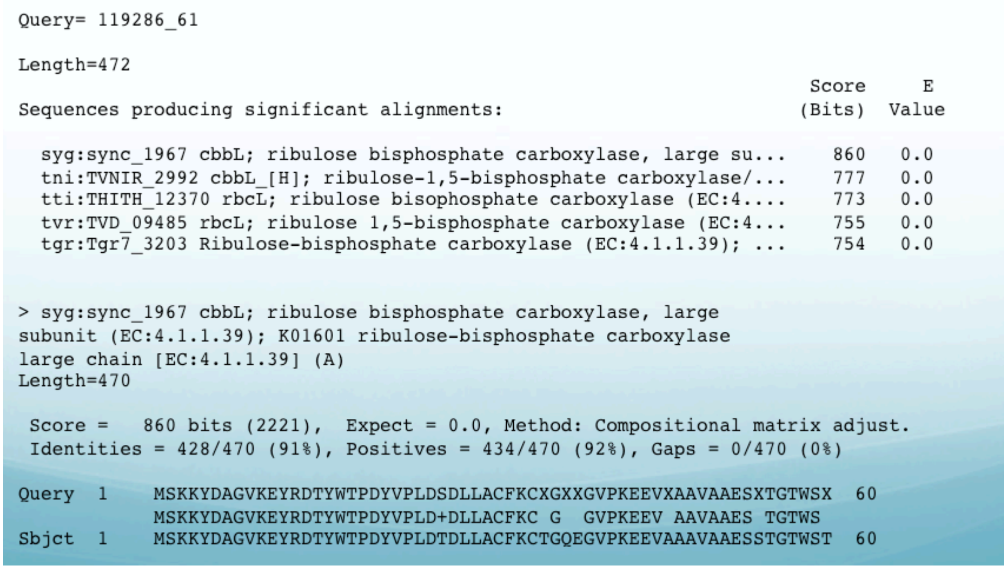
BLAST - standard output format:

cmd COMMAND

```
$ blastp -query tara_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -
out temp_output_file -evalue 1e-20 -num_descriptions 5 -num_alignments 5
```

Sets minimum limit of E-value match and maximum limit for number of print matches and alignments

EXPECTED RESULTS



```
Query= 119286_61

Length=472
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  syg:sync_1967 cbbL; ribulose bisphosphate carboxylase, large su...    860   0.0
  tni:TVNIR_2992 cbbL_[H]; ribulose-1,5-bisphosphate carboxylase/...    777   0.0
  tti:THITH_12370 rbcL; ribulose bisophosphate carboxylase (EC:4....    773   0.0
  tvr:TVD_09485 rbcL; ribulose 1,5-bisphosphate carboxylase (EC:4...    755   0.0
  tgr:Tgr7_3203 Ribulose-bisphosphate carboxylase (EC:4.1.1.39); ...    754   0.0


> syg:sync_1967 cbbL; ribulose bisphosphate carboxylase, large
subunit (EC:4.1.1.39); K01601 ribulose-bisphosphate carboxylase
large chain [EC:4.1.1.39] (A)
Length=470

 Score =   860 bits (2221),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 428/470 (91%), Positives = 434/470 (92%), Gaps = 0/470 (0%)

Query  1     MSKKYDAGVKEYRDTYWTPDYVPLDSDLLACFKCXGXXGVPKEEVXAAVAAESXTGTWSX  60
             MSKKYDAGVKEYRDTYWTPDYVPLD+DLLACFKC G  GVPKEEV AAVAAES TGTWS
Sbjct  1     MSKKYDAGVKEYRDTYWTPDYVPLDTDLLACFKCTGQEGVPKEEVAAAVAAESSTGTWST  60
```

NOTES

Rebecca Stevick 26 Jul 2016

Typo in the command - missing an 'a' in tara.

Correction: blastp -query tara_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -out BLAST_output_Cfixation_fmt0 -evalue 1e-20 -num_descriptions 5 -num_alignments 5

Xiang Liu 26 Jul 2016

The query is "tara_med_examplegenome.orfs.faa"

Local BLAST Database

Step 4.

BLAST - tabular output (fmt = 6):

cmd COMMAND

```
$ blastp -query tara_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -
out BLAST_output.tab -evalue 1e-20 -max_target_seqs 10 -
outfmt '6 qseqid qstart qend sseqid slen sstart send bitscore pident evalue'

$ less BLAST_output.tab
```

Query ID, Query Start, Query End, Subject ID, Subject Length, Subject Start, Subject End, Bit Score, Percent Identity, E-value

NOTES

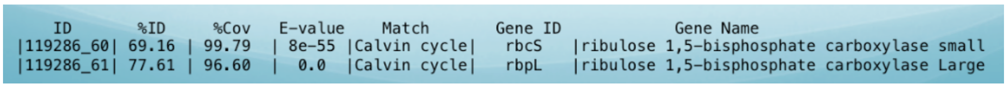
Xiang Liu 26 Jul 2016

The query is "tara_med_examplegenome.orfs.faa"

Local BLAST Database

Step 5.

Use a filter to find 'real' matches.



```
     ID          %ID      %Cov    E-value     Match         Gene ID              Gene Name
|119286_60| 69.16 | 99.79  | 8e-55 |Calvin cycle|   rbcS    |ribulose 1,5-bisphosphate carboxylase small
|119286_61| 77.61 | 96.60  |  0.0  |Calvin cycle|   rbpL    |ribulose 1,5-bisphosphate carboxylase Large
```

cmd COMMAND

```
$ awk '{if ($9>=50) print }' BLAST_output.tab
$ awk '{if ($9>=50) print }' BLAST_output.tab | sort -nrk 9,9
$ grep 'syg:sync_1967' carbonfixation_markergenes.faa
```

Cutoff of 50% sequence identity and sorted by column 9 (% identity)

NOTES

Ken Youens-Clark 27 Jul 2016

awk '$9 > 50 { print }' BLAST_output_Cfixation_fmt6

Elisha Wood-Charlson 10 Aug 2016

Can also use $ awk '$9 > 50 { print }' BLAST_output.tab

ANNOTATIONS

Xiang Liu 26 Jul 2016

Cutoff 50% sequence identity:

$ awk '{if ($9>=50) print }' BLAST_output_Cfixation_fmt6

With result sorting

$ awk '{if ($9>=50) print }' BLAST_output_Cfixation_fmt6 **| sort -nrk 9,9**

Find what we are looking for:

$ grep 'syg:sync_1967' carbonfixation_markergenes.faa

Local HMM Database

Step 6.

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

Part of the tool HMMER perform searches, and also builds new HMM models.

cmd COMMAND

```
$ hmmbuild --amino -informat afa <HMM OUTFILE NAME> <ALIGNMENT FILE>
```

Example command

Local HMM Database

Step 7.

Search tara_med_examplegenome using an HMM database for the 16 ribosomal marker proteins used to construct Hug et al (2016) Tree of Life. Utilizes a mixture of Pfam and TIGRfam models to identify targets in a genome.

cmd COMMAND

```
$ hmmsearch --tblout HMM_output.tab --cut_tc --
notextw hug_ribosomalmarkers.hmm tara_med_examplegenome.orfs.faa
$ less HMM_output.tab
```

--cut_tc = controls the threshold of match "trusted cutoff" --notextw = formatting option HMM = hug_ribosomalmarkers.hmm