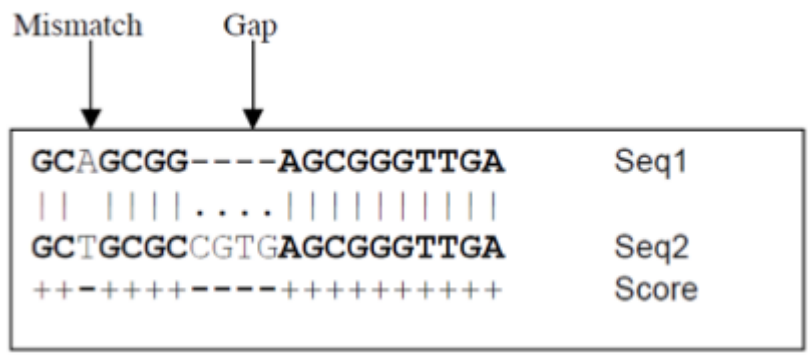
Local BLAST database

Rekha Seshadri and Georgios Pavlopoulos

Abstract

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).



http://www.hypothesisjournal.com/wp-content/uploads/2011/08/boutros-3-1-fig2.png

Compare tara_med_examplegenome protein sequences to a custom collection of carbon fixation related genes (downloaded from KEGG)

Collection contains marker genes for:

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

As a cyanobacteria - which carbon fixation pathway?

Citation: Rekha Seshadri and Georgios Pavlopoulos Local BLAST database. **protocols.io** dx.doi.org/10.17504/protocols.io.fakbicw
Published: 25 Jul 2016

Protocol

Step 1.

Create a BLAST index of 'subject' sequences.

cmd COMMAND

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

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

Step 2.

Compare the tara_med_examplegenome sequences to collection.

Step 3.

BLAST has multiple output options:

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

Step 4.

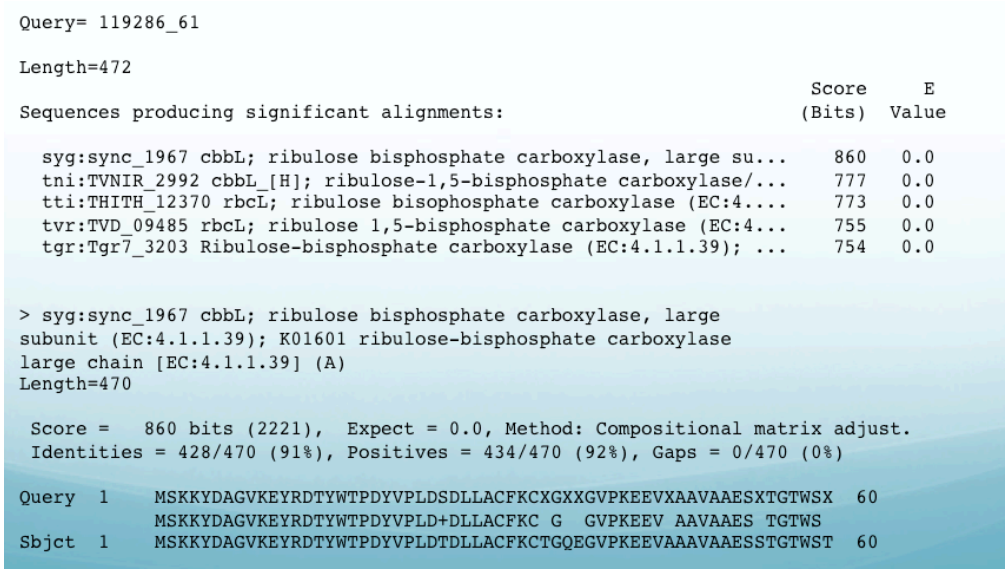
BLAST - standard output format:

cmd COMMAND

```
$ blastp -query tar_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -
out BLAST_output_Cfixation_fmt0 -evalue 1e-20 -num_descriptions 5 -num_alignments 5
```

Set 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
```

ANNOTATIONS

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"

Step 5.

BLAST - tabular output (fmt = 6)

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

qseqid means Query Seq-id	sallgi means All subject GIs
qgi means Query GI	sacc means Subject accession
qacc means Query accesion	saccver means Subject accession.version
qaccver means Query accesion.version	sallacc means All subject accessions
qlen means Query sequence length	slen means Subject sequence length
sseqid means Subject Seq-id	qstart means Start of alignment in query
sallseqid means All subject Seq-id(s), separated by a ';'	qend means End of alignment in query

sgi means Subject GI	sstart means Start of alignment in subject
	send means End of alignment in subject

Step 6.

BLAST - tabular output (fmt = 6):

cmd COMMAND

```
$ blastp -query tar_med_examplegenome.orfs.faa -db carbonfixation_markergenes.faa -
out BLAST_output_Cfixation_fmt6 -evalue 1e-20 -max_target_seqs 10 -
outfmt '6 qseqid qstart qend sseqid slen sstart send bitscore pident evalue'
```

ANNOTATIONS

Xiang Liu 26 Jul 2016

The query is "tara_med_examplegenome.orfs.faa"

Step 7.

View results:

cmd COMMAND

```
$ less BLAST_output_Cfixation_fmt6
```

EXPECTED RESULTS

```
118497_18       7       722     hmo:HM1_1745    962     262     948     388   36.83   2e-122
118497_18       8       714     toc:Toce_0785   683     5       674     380   32.49   3e-122
118497_18       3       713     aar:Acear_2299  901     220     895     383   32.30   3e-121
118497_18       4       716     dal:Dalk_2597   918     241     917     377   33.47   1e-118
118497_18       7       713     tjr:TherJR_0942 894     220     890     374   31.84   1e-117
118497_18       7       711     dto:TOL2_C28850 903     223     891     372   33.14   5e-117
118497_18       7       713     nth:Nther_0100  893     222     890     363   31.31   8e-114
118497_18       7       713     dat:HRM2_16890  921     247     917     363   32.20   1e-113
```

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

Step 8.

Use a filter to find 'real' matches. Cutoff 50% sequence identity and 50% length of subject.

```
    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
```

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

Ken Youens-Clark 27 Jul 2016

awk '$9 > 50 { print }' BLAST_output_Cfixation_fmt6