Series 5 - BLAST exercise solutions

Genomics and bioinformatics - Week 5

October 18, 2011

WORK IN PROGRESS.....

1 BLAST

1. Do you get any matches to `fragment_007`? Which parameters did you use? Record the alignment statistics for the top hits.

   Using the Nucleotide Collection (nr/nt) Database and optimizing for "more dissimilar sequences" (discontiguous megablast)

```
>emb|CR954246.1|  Pseudoalteromonas haloplanktis str. TAC125 chromosome I, complete
sequence
Length=3214944

 Features in this part of subject sequence:
   Serine protease precursor

 Score = 59.0 bits (64),  Expect = 4e-05
 Identities = 55/70 (79%), Gaps = 0/70 (0%)
 Strand=Plus/Plus

Query  5        GGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGAGTTG  64
                |||||||  ||  |||||||||  ||  ||  ||  ||  ||  |  ||||||||||    ||  ||||
Sbjct  2803550  GGCACGGTACACATGTAGCGGGTACTGTTGCTGCAGTTACTAATAATGGTGAGGGTGTTG  2803609

Query  65       CCGGGGTTGC  74
                 |  |||||||||
Sbjct  2803610  CTGGGGTTGC  2803619


>emb|FP565814.1|  Salinibacter ruber M8 chromosome, complete genome
Length=3619447

 Features in this part of subject sequence:
   peptidase families S8 and S53 domain protein

 Score = 51.8 bits (56),  Expect = 0.005
 Identities = 63/86 (73%), Gaps = 0/86 (0%)
 Strand=Plus/Minus

Query  1        AACGGGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGA  60
                 |||||  |||||||||||||||   |  |||||  |||||  |||||  |   ||  ||  |    ||||
Sbjct  2941385  AACGGTCACGGGACGCATGTGACCGGAACGGTGGCTGCCGTCACCAACAACGCCTTCGGC  2941326

Query  61       GTTGCCGGGGTTGCAGGAGGAAACGG  86
                 ||  ||  ||    |||  ||  |||||  ||
Sbjct  2941325  GTAGCGGGCACTGCCGGTGGAAATGG  2941300
```

2. Extract the sequence of the hit with the highest query coverage (this may not necessarily be the top hit) and perform another nucleotide BLAST, using the same parameters. Record the alignment statistics for the top hits.

```
>emb|CR954246.1| Pseudoalteromonas haloplanktis str. TAC125 chromosome I, complete
sequence
Length=3214944

 Features in this part of subject sequence:
   Serine protease precursor

 Score = 59.0 bits (64),  Expect = 3e-06
 Identities = 55/70 (79%), Gaps = 0/70 (0%)
 Strand=Plus/Plus

Query  5        GGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGAGTTG  64
                |||||||  ||  ||||||||  ||  ||  ||  ||  ||  |  ||||||||||    ||  ||||
Sbjct  2803550  GGCACGGTACACATGTAGCGGGTACTGTTGCTGCAGTTACTAATAATGGTGAGGGTGTTG  2803609

Query  65       CCGGGGTTGC  74
                |  |||||||||
Sbjct  2803610  CTGGGGTTGC  2803619


>emb|FP565814.1| Salinibacter ruber M8 chromosome, complete genome
Length=3619447

 Features in this part of subject sequence:
   peptidase families S8 and S53 domain protein

 Score = 51.8 bits (56),  Expect = 5e-04
 Identities = 63/86 (73%), Gaps = 0/86 (0%)
 Strand=Plus/Minus

Query  1        AACGGGCACGGGACGCATGTAGCTGGAACAGTGGCAGCCGTAAATAATAATGGTATCGGA  60
                |||||  |||||||||||||||  |  |||||  |||||  |||||  |  ||  ||  |  ||||
Sbjct  2941385  AACGGTCACGGGACGCATGTGACCGGAACGGTGGCTGCCGTCACCAACAACGCCTTCGGC  2941326

Query  61       GTTGCCGGGGTTGCAGGAGGAAACGG  86
                ||  ||  ||   |||  ||  |||||  ||
Sbjct  2941325  GTAGCGGGCACTGCCGGTGGAAATGG  2941300
```

3. What changes do you observe in the E-values? To which parameter could you attribute the these changes?

   Improvement in the E-values. Parameter - Query coverage.

4. What is the default threshold for the E-value on NCBI BLAST?

   10

5. Do you have any significant hits suggesting a possible function for `fragment_007`?

   The E-values are not significant.

1.1 Protein BLAST

1. Are any well-known protein domains found?

   Peptidases S8 S53 superfamily

2. Do you get any significant hits? Record the alignment statistics for the top hits.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
ZP_05857871.1	subtilase family domain protein [Prevotella veroralis F0319] >gb|EEX18266.1| subtila	194	194	99%	8e-55
YP_004253567.1	peptidase S8 and S53 subtilisin kexin sedolisin [Odoribacter splanchnicus DSM 20712	192	192	99%	2e-54
ZP_06407789.1	subtilase family domain protein [Prevotella melaninogenica D18] >gb|EFC73751.1| su	187	187	99%	2e-52
YP_003815043.1	peptidase families S8 and S53 [Prevotella melaninogenica ATCC 25845] >gb|ADK974	187	187	99%	2e-52
CBK65311.1	Subtilisin-like serine proteases [Alistipes shahii WAL 8301]	186	186	99%	3e-52
ZP_08836625.1	hypothetical protein HMPREF0666_02801 [Prevotella sp. C561] >gb|EGW49352.1| hy	185	185	99%	9e-52
ZP_08172701.1	peptidase, S8/S53 family [Prevotella denticola CRIS 18C-A] >gb|EGC85848.1| peptid	185	185	100%	1e-51
ZP_08137410.1	subtilase family domain protein [Prevotella multiformis DSM 16608] >gb|EGC18824.1	184	184	100%	2e-51
YP_004329881.1	peptidase, S8/S53 family [Prevotella denticola F0289] >gb|AEA22070.1| peptidase, S	184	184	100%	4e-51

3. What is the possible function of the protein encoded by `fragment_007`?

   `fragment_007` encodes for a subtilase family domain protein. It is a member of the peptidases S8 (subtilisin and kexin) and S53 (sedolisin) family. These include endopeptidases and exopeptidases.

4. Which species is most predominant in your BLAST output?

   Prevotella species

5. Could you have obtained the same results using another BLAST program, without having to translate the nucleotide sequence of `fragment_007`?

   Yes, using blastx

6. How do results from Protein BLAST compare with the results from Nucleotide BLAST?

   Amino acid sequences are more conserved than nucleotide sequences. Often even the highest-scoring subject sequences retrieved using the nucleotide sequence will cover only small regions of the query sequence, while quite often the corresponding sequences retrieved using the amino acid sequence will cover more of the gene.

2 Finding orthologs

Putative ortholog of *Saccharomyces cerevisiae* Pho2p in *Candida glabrata*?
`CAGL0L07436g`
Are the findings of the paper consistent with your observations?
Yes