Computational Molecular Biology Assignment

Sijie Chen 2016310721

1. Apply dynamic programming to align the two amino acid sequences:

AGWGHEE

AWHEA

Use BLOSUM62 scoring matrix (see course slides), gap penalty -8. Please give results for global and local alignment respectively.

The alignment result generated by my program is listed as follow. See the source code for more detail of the algorithm in the Attachment1.

For global alignment, the score is 11 and the sequence is aligned together is this way:

```
AGWGHEE
A-W-HEA
```

```
*------------------------------------------------------------------------
* Computational Molecular Biology Assignment III: Sequence Alignment
* Author: Sijie Chen (2016310721)
*------------------------------------------------------------------------
Enter Sequence 1:
>>>AGWGHEE
Enter Sequence 2:
>>>AWHEA
Global Alignment Result:
11
AGWGHEE
A-W-HEA
Local Alignment Result:
16
WGHE
W-HE
Continue?(y/n)
>>>
```

2. Align the sequence1.fa against the NCBI Genomes database through the NCBI blast web server and answer the questions below:

(https://blast.ncbi.nlm.nih.gov/Blast.cgi)

1) Which species does the sequence belong to?

SARS coronavirus TOR2

2) List at least 3 kinds of species whose genome is similar with this sequence.

Description	Max score	Total score	Query cover	E value	Ident	Accession
SARS coronavirus, complete genome	54940	54940	100%	0.0	100%	NC_004718.3
Bat coronavirus BM48-31/BGR/2008, complete genome	15389	20411	85%	0.0	83%	NC_014470.1
Bat Hp-betacoronavirus/Zhejiang2013, complete genome	1507	1507	17%	0.0	72%	NC_025217.1
Bat coronavirus HKU5-1, complete genome	398	398	6%	2e-104	71%	NC_009020.1

Three kindred species are found searching against NCBI Genomes database:

Bat coronavirus BM48-31/BGR/2008, Bat Hp-betacoronavirus/Zhejiang2013, and Bat coronavirus HKU5-1.

3) How many protein-coding genes are there in this genome?

14 annotated CDS are listed in the GenBank: AY274119.3

4) What is the percentage of the non-coding sequence in this genome?

Percentage of Non-coding Sequence: 4.93% . Here is the detailed procedure:

Length of CDS:

Coding Seuqences	CDS Length
CDS join(265..13392,13392..21485)	21220
CDS 21492..25259	3767
CDS 25268..26092 *	824
CDS 25689..26153 *	464
CDS 26117..26347 *	230
CDS 26398..27063	665
CDS 27074..27265	191
CDS 27273..27641 *	368
CDS 27638..27772 *	134
CDS 27779..27898 *	119
CDS 27864..28118 *	254
CDS 28120..29388 *	1268
CDS 28130..28426 *	296
CDS 28583..28795	212

Length of the complete genome: 29751

Note that asteroid rows share overlaps.

Sequences covered by CDS are:

(265..21485) with length 21220, (21492..25259) with length 3767,

(25268..26347) with length 1079,(26398..27063) with length 665,

(27074..27265) with length 191, (27273..28118) with length 845,

(28120..28426) with length 306, (28583..28795) with length 212.

The lengths of CDS sum up to 28285. Hence the percentage of Non-coding sequence is: 1-28285/29751≈0.0493=4.93%

2. Please align the sequences in sequence2.fa and sequence2.fastq against human genome using local version of blast and bowtie on our cluster respectively.

1) Account for the cluster:

`Host name: 166.111.5.240  User name: CMB  Password: a123456`

2) `data: ~/homework4/sequence2.fastq  ~/homework4/sequence2.fa`

4) Installed packages:

`blastn  /data/software/blast/blast-2.2.27/bin/blastn -help`

`bowtie  /data/software/bowtie/bowtie-0.12.7/bin/bowtie`

3) Pre-built index and database files:

`bowtie index: ~/homework4/bowtie_index_hg18/hg18`

`blast database: ~/homework4/blastdb/hg18`

Two tasks are submitted on the 166.111.5.240 clusters with jobID=o39923 and o399225.

BLASTn is run with evalue=1e-8 and all other parameters are default value.

Bowtie is run with all parameter set to default value.

The result is shown in the attachment2.