PLPTH813, Homework 3

The following data were provided for this homework.

Reference genome:

MG1655.fasta

Gene DNA sequence:

clpA.DNA.fasta

Gene protein sequence:

clpA.aa.fasta

Illumina data:

DH10B.pair1.fq

DH10B.pair2.fq

Date can be copied from the following directory at Beocat:

/homes/liu3zhen/teaching/datasets/HW03/

1. Use seqtk to tackle these tasks:

1. Convert DH10B.pair1.fq to a fasta file
2. Randomly sample five paired reads from the Illumina data

2. Perform both quality and adaptor trimming using Trimmomatic for the Illumina data. Report the script and the result summary (e.g. log output).

3. Apply the Smith-Waterman algorithm to find the best-scored local alignment between two sequences: ATACTCAT vs.

GTATCAC. Here is the score scheme:

Match +1

Mismatch -1

Gap -2

Fill in score in each cell in the table and draw the path of the best-scored local alignment.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | A | T | A | C | T | C | A | T |
|  |  |  |  |  |  |  |  |  |  |
| G |  |  |  |  |  |  |  |  |  |
| T |  |  |  |  |  |  |  |  |  |
| A |  |  |  |  |  |  |  |  |  |
| T |  |  |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |  |  |
| A |  |  |  |  |  |  |  |  |  |
| C |  |  |  |  |  |  |  |  |  |

Report the best local alignment(s):

e.g. ACTGG

|||

CTG

4. Determine the BWT form of a sequence of “CGCAGT”; report the intermediate steps as well; and illustrate how to find the sequence of “GCA”.

5. Report BLAST commands you use for the following jobs

1. Build the MG1655.fasta database
2. Search clpA.DNA.fasta in the database; explain what the bitscore and e-value mean.
3. Search clpA using the gene protein sequence (clp.aa.fasta) in the database
4. Search clpA.DNA.fasta in NCBI BLAST website. Report bitcore and e-value for the top hit. Compare them to bitscore and e-value using MG1655.fasta as the database. Discuss what is the difference and what causes it.