Genomic Bioinformatics

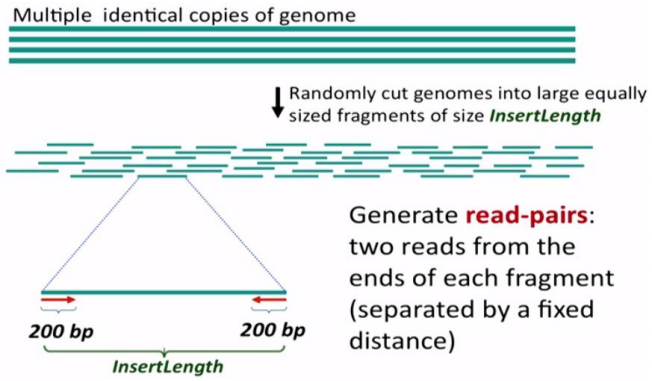
Lab Programming Task Documentation

Task Overview

In this task you should assemble two types of reads using the De Bruijn graph representation and the Eulerian path to obtain the original sequence from that representation.

First type of reads : Short reads with the same size

Second type of reads : Pair-reads with a known distance between them



Task Details

Language : Python
Number of Members in team: 5-7 members
Week : 6

Task Requirements

Input :
Your program should be able to read **a file (for the single or paired reads).**

The first line would be:

- the length of the sequences (for single reads).
- the length of the sequences in each side and the length of the gap (for the paired reads).

Each read will take one.
The pair reads will be separated by "|" .. Example : AGCC|TTAA

Output : The program then outputs **the assembled sequence** to the screen. (or write it in a file)

Example ***for single reads*:**

Sample Input:

```
GAGG
GGGG
GGGA
CAGG
AGGG
GGAG
```

Sample Output:

```
AGG -> GGG
CAG -> AGG
GAG -> AGG
GGA -> GAG
GGG -> GGA,GGG
```

The Eulerian walk for the above case is

1->2
3->1
4->1
5->4
2->5,2
3->1->2->2->5->4->1
CAGGGGAGG

Example ***for paired reads*:**

Sample Input:

```
2
GAGA|TTGA
TCGT|GATG
CGTG|ATGT
TGGT|TGAG
GTGA|TGTT
GTGG|GTGA
TGAG|GTTG
GGTC|GAGA
GTCG|AGAT
```

Sample Output:

```
GTGGTCGTGAGATGTTGA
```

Notes

- Document your code whenever is possible with appropriate comments.
- Use meaningful names for variables and functions.
- Each student should have a version of the code that was written by their team.
- Each student should be able **answer questions** about the code and the biological meaning behind it and **run sample data** on it.
- **The codes will be run through a plagiarism check. Any online/previous year/current year plagiarism cases will be considered zero with no exceptions. Please do your own work.**

Good Luck :)