Programming Assignment 4 - Viral Quasispecies Assembly

BIOINFO M260

Due: March 8th at 11:59 pm

This programming assignment is designed around the mathematics of quasispecies assembly.

Overview

In this assignment, you are given single-end reads from 4 known donor sequences. These reads are formatted as a single 50 bp-long end. You are also given as reference data 4 sequences from which the single end reads were generated; unlike the previous assignments, there is no "noise" data inserted.

Starter Code

Useful starter code is available under PA 2 (the hashing aligner) and the Jupyter Notebook in CCLE, which can be adapted.

Your main task will be to quantify the number of copies of each strain; the simplest way to do this is to identify the positions where there are SNPs in each strain, and to quantify the number of times each SNP occurs. Then, based on whether the SNP

$$\begin{aligned} X &= \text{Strain SNPs} \\ f &= \text{Strain Frequencies} \\ b &= \text{SNP Frequencies} \\ Xf &= b \end{aligned}$$

Note that X is not an invertible matrix; however, this system can still be solved using what's referred to as the (Moore-Penrose) pseudoinverse, $(X^TX)^{-1}X^T$ to both sides of the equation above:

$$\begin{aligned} (X^TX)^{-1}X^TXf &= (X^TX)^{-1}X^Tb \\ f &= (X^TX)^{-1}X^Tb \end{aligned}$$

This is referred to as the "least-squares" solution to the problem, and is implemented in Python's `numpy` package.

I/O Details

Examples of the output format are in CCLE.

```
>assignment_name
<frequency1>,<sequence1>
<frequency2>,<sequence2>
...
```

Your frequency should be reported as a floating point number between 0 and 1. All of the strains have nonzero frequencies.

Scoring

Because it is exceptionally easy to guess-and-check the solution to this assignment, you will be given a correct example to train your algorithm on, and you will submit your file

This assignment will be scored based on the sum of the squared differences between your quantification of the frequency of each strain, and the actual frequency of the strain itself.

Submit your response under Week 8 in CCLE; I'll post a scoreboard on the cm124 Site.