Computational biology project:
Salmonella outbreak

MARKOVIĆ Milena, VONDRAČEK Dušan

October, 2020

Table of contents

Problem analysis

A violent bacterial outbreak is currently happening, killing tons of people, and the usual antibiotics have absolutely no effect.

- TATFAR makes a call for developing tools that answer the current crisis but that can be used for further events. The specifications are:
  one tool that takes two simple FASTA sequencing files (Illumina reads) and outputs a list of SNPs
- One tool that takes a multiple sequence alignment in standard FASTA file format and outputs a protein structure.
- Make a proof of concept of the tools used for the tasks above on the current AMR crisis (antimicrobial resistance) by understanding what is the difference between the strains, what gene(s) is involved, build a model of the protein structure associated to the gene and give a possible explanation.

But for now, what we need to do is make an estimation of the costs (DNA sequencing and also our workload) justified by a basic description of the piece of software we plan to develop.

Task description

The problem can be broken down into several steps.

First off, Illumina is the main technology used nowadays for sequencing. It produces short reads of around 150-250 base pairs and the produced data is reliable with less than 1% of sequencing errors. Those errors can be considered as uniformly distributed, and containing only mutations, meaning substituting a letter by another one.

The starting point of the analysis are two FASTA files containing genetic sequences of two salmonella strains - one that is resistant to tetracycline, one that is not. The first piece of software should compare these two strains and find all single-nucleotide polymorphisms in order to find mutations. To do so, the sequences first need to have the sequencing errors removed in each FASTA file. Only after that can we compare the two genomes in order to find SNP-s.

Costs

According to the National Human Genome Research Institute, the cost of the first DNA sequencing (at the beginning of this century) was estimated at a whopping $300 million!
Based on data collected by NHGRI from the Institute's funded genome-sequencing groups, the cost to generate a high-quality 'draft' human genome sequence had dropped to ~$14 million by 2006.
Luckily for us (and everyone else), we are not in 2000 nor 2006 anymore! Nowadays, the cost to sequence a human genome is less than $1000!

Since we are responsible citizens and trusted collaborators, before getting the call outcome, we have already sent two strains of the DNA for sequencing and are waiting for Illumina to finish their part of the work.

So, if our calculation is right, the cost of DNA sequencing will be around $2000 (to simplify, we will say the cost is the same in euros - 2000€).

As for our salaries: Glasdoor.fr indicates that an average monthly salary for a Data Scientist Intern in Paris is 1414€. Since there are two of us in the team (instead of the usual 3-person team), we would demand a slightly higher than average salary, let’s say 2000€ (as we value our effort and time invested). The length of the project is around a month, so the total cost for our workload would be around 4000€. A bonus after a job well done would be highly appreciated.

In the end, the first estimation of the costs is around 6000€ (plus the bonus). Quite cheap, eh?