Computational Biology

Team

Bitiutskii Artem
Fomin Dmitrii
Semenenko Aleksandr

Supervisor

Clovis Galiez

October 2022

Abstract. A violent bacterial outbreak is currently happening, killing tons of people, and the usual antibiotics have absolutely no effect. This has already happened in the past, something known as AMR, but up to now, the pharama labs managed to circumvent the resistance. Biologist Emmanuelle Charpentier managed to isolate two different strains of the bacteria: the resistant strain to the tetracycline and a wild type. Taking into account a request from TATFAR, current research aims to develop computational tools to determine the origin of the problem, and possibly give biologists some insights on what could be done.

1. Introduction

One of the possible methods to identify mutations that can cause genetic disease is DNA sequencing. The latter can be considered by cutting DNA uniformly random into small pieces the length of r that are usually called reads. Denoting the size of the genom by G, we assume that $r \ll G$. Such an estimation and further discourse are particularly based on two remarks. First, extracting DNA from a cell might be accompanied by some information loss. Second, the mentioned sequencing procedure has errors.

Initially, we want to find a length of k of k-mer.This hyperparameter will play a significant role in our further analysis, since at the beginning of searching for mutation, it will be necessary to set this value in advance. We show histograms‡ in Fig.1 each of which corresponds to $k \in \{15, 21, 25, 30\}$. Notice that there are some picks, most of which are on the right side and consistent with apparent outliers. In a small neighbourhood of zero (Fig.7), in turn, we see quite high bars, which are actually caused by sequencing mistakes, since they are uniformly distributed. After many of these kinds of experiments, we conclude that it would be better to set $k = 21$.

‡ For extra plots see Appendix 4

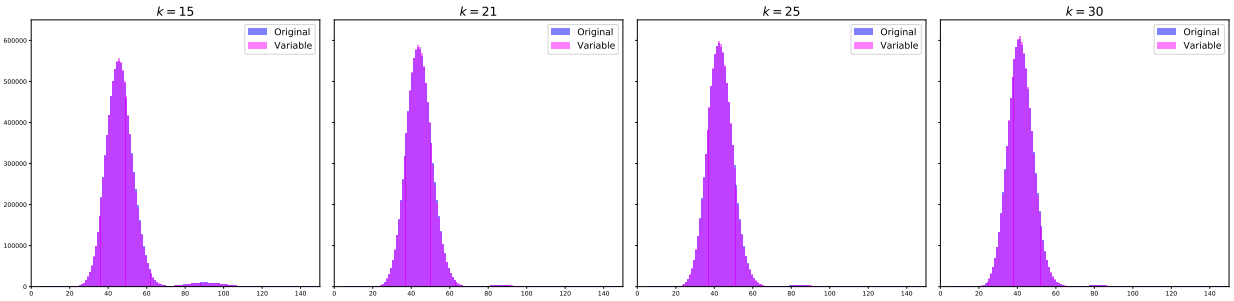

Figure 1. Histograms of original and variable genome for different k's. The higher bar is, the frequenter some length is.

It is obvious that the more unique k-mers we have, the more difficult it is to pick out mutation bars, because the latter will be hidden beyond noise. Conversely, the higher k is, the longer unique k-mers are, which eventually make it difficult to recognise mutation as well. Thus, we have to find optimal k and, as it follows either from Fig.2, it might be $k = 21$.

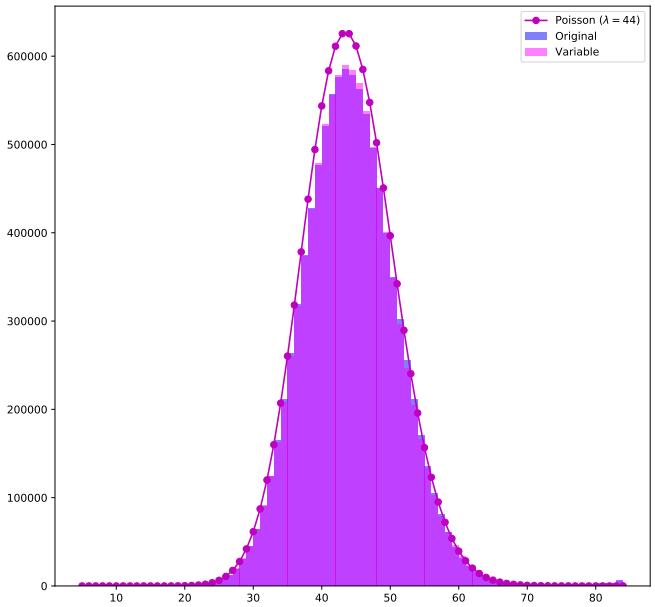

Figure 2. The Poisson distribution approximates the histogram corresponding to k=21.

On the other hand, the probability of that k-mer is contained by read with length r is $p = \dfrac{r-k}{G-r} \ll 1$. Since sequencing produces reads independently (i.i.d. more

precisely), we come to the conclusion that the number X of reads overlapping a given k-mer is a binomially distributed random variable, i.e. $X \sim \text{Bin}(N, p)$, where N denotes the total number of reads. Now is the time to recall that $N \gg 10^3$, so the Poisson distribution with parameter $\lambda = Np \approx 44$ can be used as an approximation to $\text{Bin}(N, p)$ as long as the binomial distribution converges towards the Poisson distribution as $N \to \infty$. Consequently, $X \sim \text{Pois}(Np)$ as it is showed in Fig.2.

2. Genetic mutation

Once we chose k, we create the list of k-mers which belong to the original genome but not to the variable genome. On the contrary, we act with the second list, which, in turn, contains variable genome k-mers but not the original.

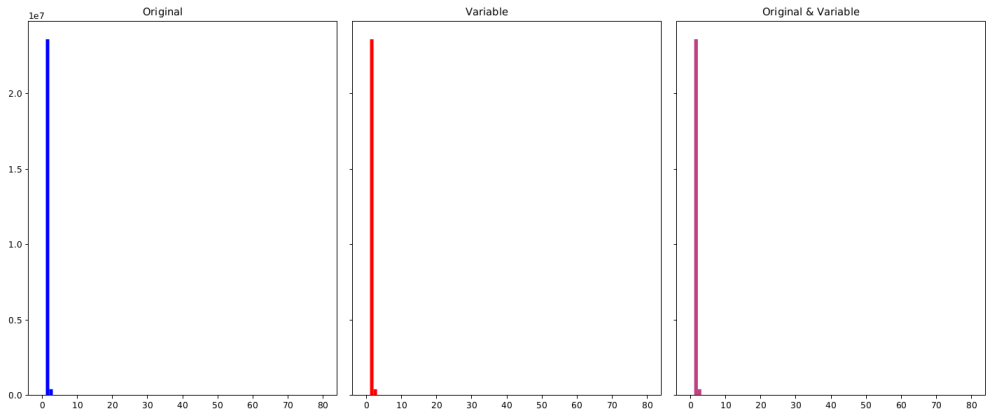

Figure 3. From left to right, histograms for original, variable genome frequencies and their union.

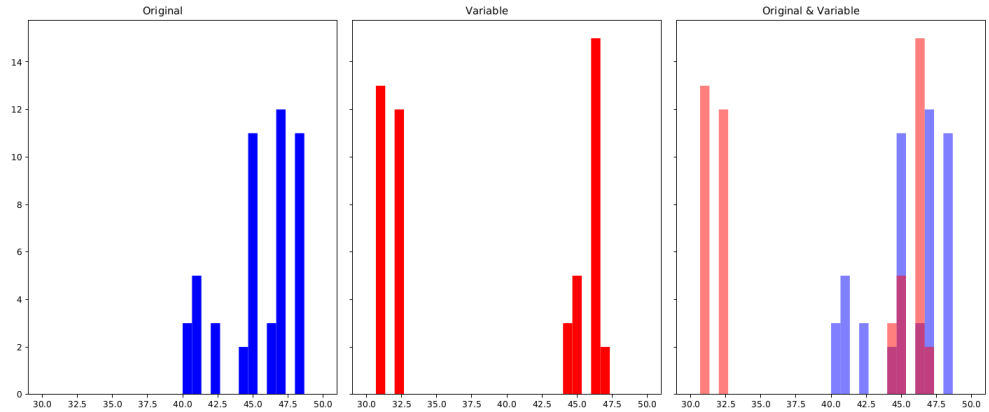

Figure 4. The histograms are from Fig.4 without noise.

On the top three histograms, we observe a large number of k-mers, whose frequencies of appearing are located near zero, which could be a result of sequencing errors. Therefore, we will look at our data without these values. The DNA molecule is a polymer consisting of two polynucleotide chains, so we consider two parts of our histograms separately (left and right) for the original genome and the genome with mutation.

Both the original ($[40, 43]\&[44, 49]$) and variable ($[31, 33]\&[43, 48]$) genomes have two distinct areas. So, first, we compare, or look for mutation in, the results of k-mers concotanation for the original genome and variable, taking into account corresponding "intervals" (i.e., picks at $[40, 43]$ vs picks at $[31, 33]$, and so on). Thus, we come to the following two stings in the case of the right parts:

$$CTACACCTAGCTTCTGGGCGAG\_TTT\_ACGGGTTGT$$

$$CTACACCTAGCTTCTGGGCGAG\_GGG\_ACGGGTTGT$$

Then we repeat the same with the left parts of the histograms and get

$$CAACCCGT\_AAA\_CTCGCCCAGAAGCTAGGTG$$

$$CAACCCGT\_CCC\_CTCGCCCAGAAGCTAGGTG$$

We can see the mutation. Since the real data was given, it is possible to check the result of our algorithm with the Blastx Nucl-Prot database.

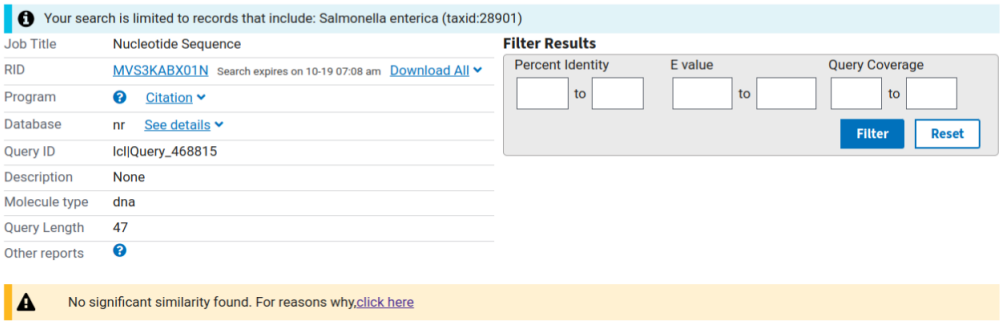

Figure 5. Blastx database ($k = 21$ case).

The result turned out to be not positive and we decided to conduct some more experiments by changing the values of k. When we assume that $k = 25$, we make similar calculations and get for the right parts

$$CTCTACACCTAGCTTCTGGGCGAG\_TTT\_ACGGGTTGTTAAACCTTCGATTCC$$

$$CTCTACACCTAGCTTCTGGGCGAG\_GGG\_ACGGGTTGTTAAACCTTCGATTCC$$

and then for the left parts:

$GGAATCGAAGGTTTAACAACCCGT\_AAA\_CTCGCCCAGAAGCTAGGTGTAGAG$

$GGAATCGAAGGTTTAACAACCCGT\_CCC\_CTCGCCCAGAAGCTAGGTGTAGAG$

We chose $k = 21$ at first, even though $k = 25$ was not worse (see Fig.1). It gives hope to make another attempt at our research, but now for $k = 25$.

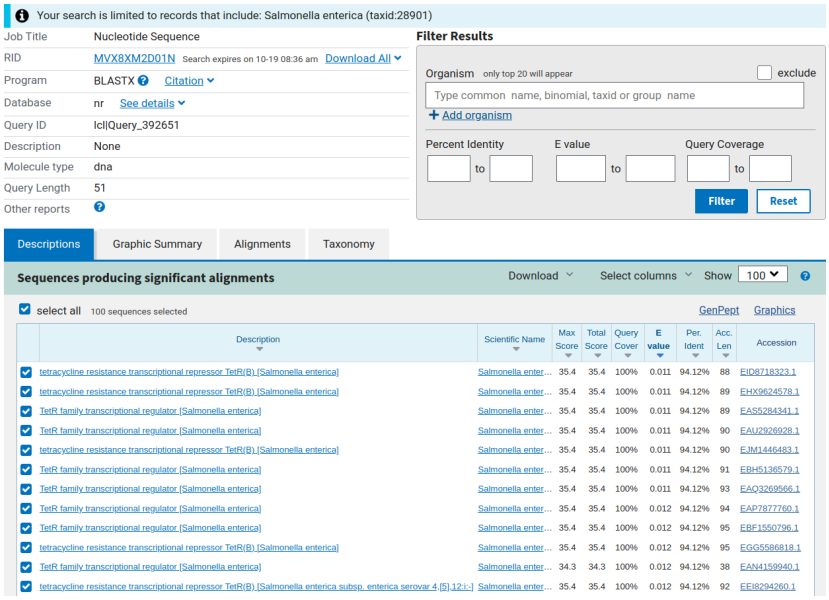

Figure 6. Blastx database ($k = 25$ case).

Finally, for the sequences at $k = 25$ we managed to get the result with the Blastx Nucl-Prot database. Based on the obtained data shown in the Fig.6, we can see that there are correspondences for this mutation.

3. Conclusion

The mutation was found and, as it turned out, the success of achieving this aim is closely connected with the value of k. We proposed some intuitive approaches to finding the letter. Finally, some tools to manage with finding a mutation in the genome were developed.

4. Appendix

The source code with invented tools and details of experiments are available on the GitHub. We presented linear time algorithm $O(n)$.

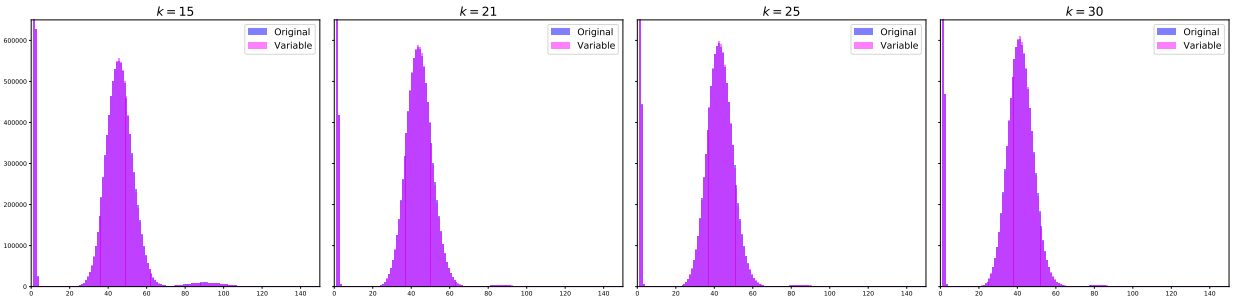

Figure 7. Histograms of original and variable genome for different k's. The higher bar is, the frequenter some length is. Here we show bars in a small neighbourhood of zero.