Computation Biology Project One
Salmonella outbreak

Name: Peng-Jhen,Lee; Tsai-Yi, Chen; Jiaying TU

Code available:
https://drive.google.com/drive/folders/1MHiSckwX1qp3Qm6Ko_kzikC4oCt3V7p6?usp=sharing

Introduction

Salmonella outbreak, which is a serious bacteria outbreak, causes numerous people to die from this disease. Moreover, violent Salmonellosis could not be cured by normal antibiotics. However, an outstanding scientist Emmanuelle Charpentier isolated the strains that resist the tetracycline from wild stain successfully.

TATFAR makes a call for developing tools that could distinguish the different points between resistant strain and wild strain, and thus it can be used for further events.

Data

The sequence data which are used in the project are performed by Illumina. This machine would make testing sequences into fragments randomly. Thus, the result of the sequencing will be stored as fragmental information of nucleotide sequences. This tool is able to input two different FASTA files and to output the single-nucleotide polymorphism (SNPs), which is the a single mutation of the nucleotide in the genome. FASTA format is the text-based format and stores the results of the sequencing. Thus, the data would be presented as a single letter representing the nucleotides or amino acids.Two input files are sequencing data of Salmonella. One is the wild type which the phenotypes are typical and could be found in nature, and the other one contains variants that is resistant to normal antibiotics.Each read in these two files is 250bp which only gives the local view of the genome. However, even though this technique is widely applied, this sequencing method generates a 1% error during the sequence. Thus, we introduce the method to detect errors before finding the SNPs.

Methods and Model

There are several notations that would be introduced for further presentation and they were used frequently in this project.

L: the length of each read

G: the dimension of the genome size

N: the total number of reads

k-mer

K-mer is a method that could help the researcher to understand the whole pattern of the genome and also help us to reconstruct the genome afterward. Since the data from highthrough sequencing is usually accompanied by numerous reads. Without reference sequence, the genome survey is needed to manipulate the genome's heterogeneity or the proportion of the duplication. K-mer could substring the k length of reads, so each read could generate L-k+1 number of k-mers (Figure 1).

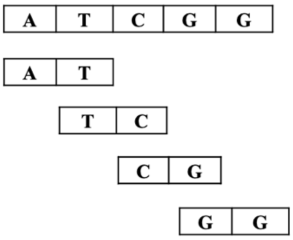


Figure 1. Example of the K-mer (k=2)

As reads are randomly distributed along the genome, the probability P that the read overlapped the position on the genome is

$$P = \frac{L}{G - L}$$

so the reads overlapping position i, noted as X_i. , follow the Binomial distribution.

$$X \sim Bin(N, \frac{L}{G - L})$$

K-mers are the substring of the reads, so the probability p that the reads to overlap any k-mer is

$$p = \frac{L - k}{G - L}$$

so the read overlapping any k-mer x_i also follow the Binomial distribution with the parameters N and p,

$$x \sim Bin(N, \frac{L - k}{G - L})$$

As the number of N and the probability p is small, the binomial distribution can approximate the Poisson distribution, so the number of the reads overlaps the given k-mer follows the Poisson distribution.$w \sim Poi(\lambda = N * \frac{L-k}{G-L})$

Sequencing Error

As we mentioned above this sequencing method would generate a 1% error, the probability p_{ef} for a k-mer from the genome to be error free in a read is

$$p_{ef} = \frac{L - k}{G - L} * (1 - error)^k$$

So the number of k-mers y_i that are error-free in reads follows the binomial distribution with parameters N and p_{ef}, and it will approximate to the Poisson distribution with $\lambda_{ef} = N *$

p_{ef}. Moreover, we need to introduce the cutoff f_0, which is a struggle to correct the errors in the data, since it is the balance between the number of k-mers containing errors and without errors[1].

Reconstructing the Sequence

After deleting the errors that might exist in the sequencing, we try to reconstruct the sequence by concatenating the k-mers to rebuild the section of genome. First, each k-mer would compare and match each other to find the same amino acids of the first or last k-1 of k-mer, so there will be a new amino acid that differs from the other one. Thus, the new character would be added to that certain k-mer, so the new sequence is generated. Secondly, the added sequence will compare to another k-mer. If the new sequence has identical amino acids of the first or last k length, the sequence will add the new character again.

Therefore, if the concatenation is operated successfully, there would be a new character added to the string. However, if the concatenation could not be done, it would restart a new concatenation progress from the remaining k-mers. This program could help us to reestablish the pieces of the genome.

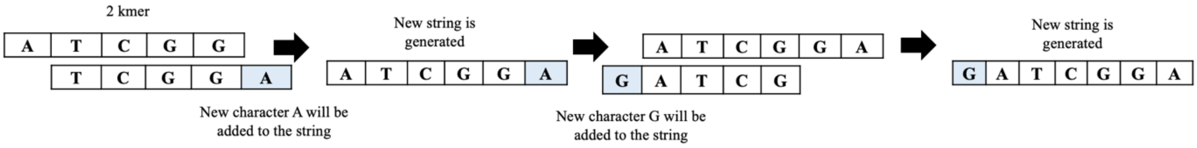


Figure 2. Example of the concatenation

SNPs location

There are different ways to compare two different sequences. However, in this project, we use Levenshtein distance[2], which is the way to calculate the minimum number of single-character edits to change one word into the other. The reason to choose this method but not the Smith-Watman algorithm is that we do not know the space of the substring of the sequence after concatenation, but in the calculation of the Smith-Watman algorithm would be utilized the spaces for different weights of the score. Thus, we decided to use Levenshtein distance that would not involve the space calculation.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j)\,, if\ \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} , otherwise \end{cases}$$

If the distance between two strings is the same, then finding SNPs would be executed directly and print color on the changing places. However, if the distance between lines is different, then we would make the shorter string into the longer one by inserting space or deleting the characters, which would make both strings have the same length for comparison.

Result

To comprehend the sequence's pattern and determine the value of k, we first presented the frequency over the depth with k from 17 to77 (Figure 3).

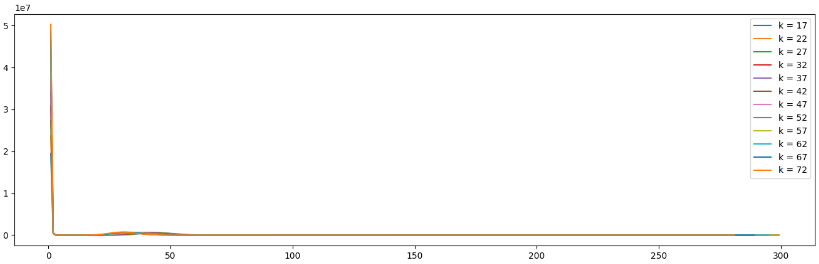


Figure 3 k-mer frequency over the depth (wild type strings with different k and without log-transform on y-scale)

Unfortunately, the selection of the k will influence the accuracy of the position on the genome. For a smaller k, the information of reconstructing the sequence will decrease, but it could help reduce DNA storage. On the other hand, the larger k would increase the DNA storage, but it benefits adjoining the sequence as it helps avoid overlapping with other k-mers and the problem of the small repeat region. For the pattern without log-transforming on the y-scale, there is only one main peak. Nevertheless, if we took log-transforming, there is an increase steeply at the beginning, (Figure 4-1 and Figure 4-2), which may contribute to the error the sequencing, and there is another smaller peak afterward, which may result from the duplication of the genome.

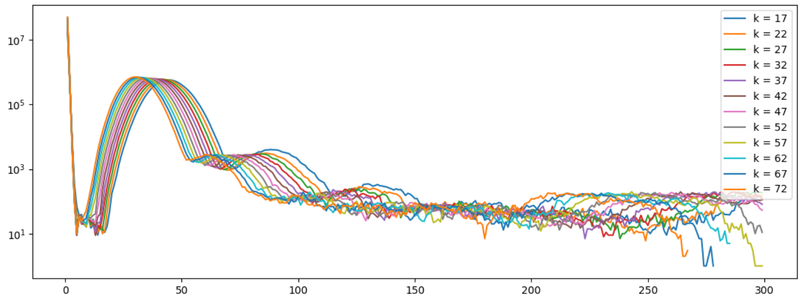


Figure 4-1 k-mer frequency over the depth (wild type strings with different k and log-transform on y-scale)

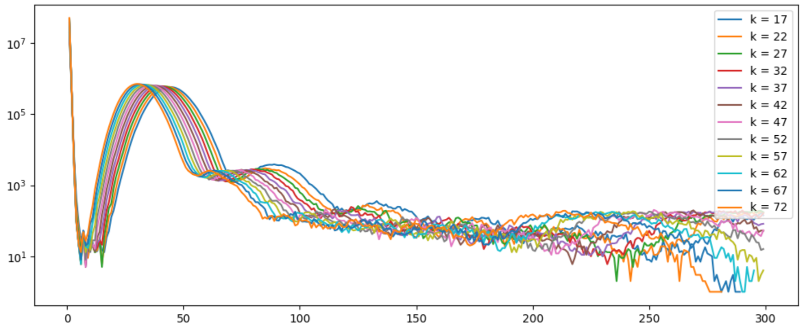


Figure 4-2 k-mer frequency over the depth (variant type strings with different k log-transform on y-scale)

Regardless the value of k, the pattern of different k looks similar, so we chose the k=47 empirically (Figure 5) and $f_0 = 13$ for further analysis.

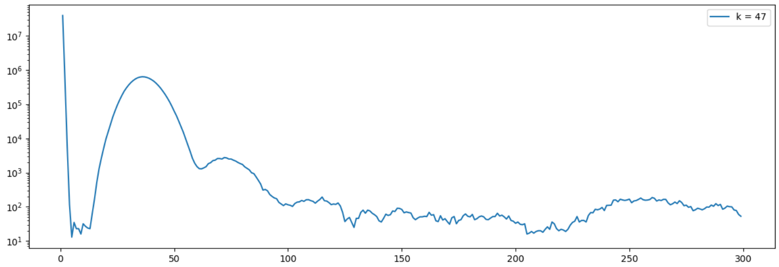
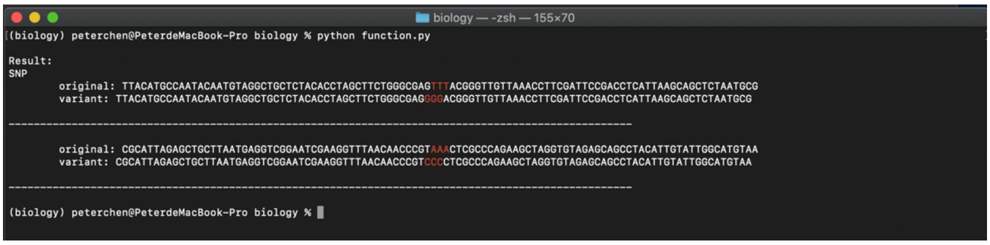


Figure 5. k-mer the frequency over the depth (the wild type with k =47)

There are two parts of mutations of the nucleotides from the variant strings. The first substituted thymine to guanine, and the other one was adenine to cytosine (Figure 6). To know whether SNPs worked on the protein synthesis region, so we put the pieces of the genome into BLAST[3]. After searching different gene databases by BLASTx, which is a searching method that first translated nucleotides into proteins and compared them with other protein sequences. We found that both parts located on the Tet Repressor proteins regulator (TetR), which plays an important role in giving antibiotics resistance to the bacteria.



```
(biology) peterchen@PeterdeMacBook-Pro biology % python function.py

Result:
SNP
        original: TTACATGCCAATACAATGTAGGCTGCTCTACACCTAGCTTCTGGGCGAGTTTACGGGTTGTTAAACCTTCGATTCCGACCTCATTAAGCAGCTCTAATGCG
        variant: TTACATGCCAATACAATGTAGGCTGCTCTACACCTAGCTTCTGGGCGAGGGGACGGGTTGTTAAACCTTCGATTCCGACCTCATTAAGCAGCTCTAATGCG

--------------------------------------------------------------------------------------------------

        original: CGCATTAGAGCTGCTTAATGAGGTCGGAATCGAAGGTTTAACAACCCGTAAACTCGCCCAGAAGCTAGGTGTAGAGCAGCCTACATTGTATTGGCATGTAA
        variant: CGCATTAGAGCTGCTTAATGAGGTCGGAATCGAAGGTTTAACAACCCGTCCCCTCGCCCAGAAGCTAGGTGTAGAGCAGCCTACATTGTATTGGCATGTAA

--------------------------------------------------------------------------------------------------

(biology) peterchen@PeterdeMacBook-Pro biology % 
```

Figure 6. SNPs

Reference

1. Zhao L, Xie J, Bai L, Chen W, Wang M, Zhang Z, Wang Y, Zhao Z, Li J. Mining statistically-solid k-mers for accurate NGS error correction. *BMC Genomics.* 2018;19(10):912.

2. Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady.* 1966;10:707.

3. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-3402.