2018-2019 FINAL YEAR PROJECT

Computational Problems for Bioinformatics

Supervisor: Dr. S.M. Yiu
Name: Ting Ming Yin
UID: 3035271256

Contents

1. Introduction

1.1 Project Background

In 1860s, Johann Friedrich Miescher, a Swiss chemist, identified the first DNA molecule when doing a research about the white blood cell [2]. However, due to the limitation of technology, the structure of DNA could not be solved. Until 1960s, with Rosalind Franklin, Francis Crick, James Watson, and Maurice Wilkins contributions, the DNA structure was finally solved, and sequencing DNA became a possible but difficult task. Nowadays, DNA help scientists understand lots of unsolved problems before, like the cause of cancer, the origin of humanity or crime solving. It becomes an important field that scientists put lots of effort to solve the mystery inside the DNA sequence.

However, DNA sequence is impossible to be solved by human hands because the information stored in a cell is too much. The DNA sequence in one human cell can be 2 meters long and more than three billion nucleotides need to be mapped [1]. Therefore, scientists use computer and software tools to help them analysing DNA sequence and this is called bioinformatics. Although computers enhance the analysis greatly, it is still a difficult task as the information stored in one cell is about 1.5 Gb large which means handling big data is one of the issues in this project. This project aims to solve the computational problems and reconstruct a complete genome using the appropriate tools.

1.2 Motivation

Bioinformatics is a combination of Biology and Computer Science to construct and analyse DNA sequence. I am interested in this field because of one important project, human genome project, which was initiated in 1990 and aims to map all the human genome and understand the hidden message between the sequences [3]. This project has made lots of contribution to human society and inspired me that how important DNA is. This final year project can help me practise the basic techniques and algorithms used in genome mapping, which may be beneficial to my future career. Moreover, solving computational problems can enhance my problem-solving skills related to real world problems, for example, running time of programs handling big data. As a result, I choose this topic to be my final year project.

2. Objective

The objective of this project is reconstructing an unknown genome of a species, probably a marine creature. Due to the limitation of technology, a complete genome cannot be extracted directly from the creature but only small pieces which may contain repeated or polluted genome. Therefore, the challenge of this project is how to use these short pieces of genome to reconstruct a complete genome of the species and analyse the correctness of the DNA sequence.

3. Methodology

The operating system of the software tools will be Linux which is commonly used by bioinformaticians to handle and analyse the DNA sequence. It provides lots of convenient tools such as command line and file managing system which make the process simple and fast. For example, the “grep” function can search a certain pattern in a file with one command line. It can also perform different behaviours by simply adding extra symbols to the line such as “-o”, “-n” which shows the exact wording and line position of the input pattern. The features of this function make it significant when the position of a genome pattern need to be found within millions of lines.

Beside the convenient functions, c++/c language will also be used in this project to do some works automatically as the DNA sequence is too long for human works. Different algorithms will be used to increase the correctness of the final delivery, such as checking the correctness of connected genome, detecting and deleting overlapped genome, determining the position of new genome. These algorithms have been created and used by scientists and need to be combined and used properly in this project to make a complete and correct genome.

Extracting the genome from the samples will be done in the labs but it is not included in this final year project so no methodology about the extraction will be mentioned in this project.

4. Schedule and Milestone

Date	Deliverable/Milestone
30/09/2018	Project Plan Project Website
Oct	Algorithms studies Software testing
Nov - Dec	30% genome constructed
7-11 January 2019	First presentation
20 January 2019	60% genome constructed Detailed interim report
14 April 2019	Complete genome Final report
15-19 April 2019	Final presentation
29 April 2019	Project exhibition

5. Reference

1. Science Focus. Retrieved from: https://www.sciencefocus.com/the-human-body/how-long-is-your-dna/
2. Your Genome (2018, February 26). Retrieved from: https://www.yourgenome.org/stories/the-discovery-of-dna
3. National Human Genome Research Institute (2016, May 11). Retrieved from: https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/
4. Ian Murnaghan BSc (2018, April 28). Retrieved from: http://www.exploredna.co.uk/the-importance-dna.html