REDUCING THE SEARCH SPACE AND TIME COMPLEXITY OF NEEDLEMAN-WUNSCH ALGORITHM (GLOBAL ALIGNMENT) AND SMITH-WATERMAN ALGORITHM (LOCAL ALGORITHM) FOR DNA SEQUENCE ALIGNMENT

by

FATIMAH NONI BINTI MUHAMAD

(0530210062)

A thesis submitted in fulfillment of the requirements for the degree of
Master of Science (Computer and Communication Engineering)

SCHOOL OF COMPUTER AND COMMUNICATION ENGINEERING

UNIVERSITI MALAYSIA PERLIS

2011

Acknowledgements

First and foremost, I am greatly thankful to Allah s.w.t with the completed of this thesis, Alhamdulillah....

I would like to express my sincere gratitude to my thesis advisor, Assoc. Prof. Dr. R. Badlishah bin Ahmad, Assoc. Prof. Dr. Puteh binti Saad, and Puan Salina binti Mohd. Asi, this thesis could not have been completed without their help and guidance.

In regards to the writing of this thesis, a major thanks goes to my dearest husband, Muhamad Nasir bin Murad, and my beloved son, Muhammad Umar Mursyid bin Muhamad Nasir, without their support and happiness, I would not be where I am today.

I also would like to acknowledge the constant support and help from my family and friends at Embedded Cluster, without which, none of this would have been possible. Appreciation also goes to MOSTI (Ministry of Science, Technology & Innovation, Malaysia) for the 2 years scholarship.

Finally, I would like to gratefully acknowledge the panels for the viva voce, Dr. Ong Bi Lynn and Assoc. Prof. Dr. Norita Mohd. Norwawi, and also the staff at the Postgraduate Studies, UNIMAP.

Thank you very much.

TABLE OF CONTENTS

CHAPTER 3 METHODOLOGY

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

Mengurangkan Kompleksiti Ruang Carian Dan Masa Bagi Algoritma Needleman-Wunsch (Penjajaran Global) Dan Algoritma Smith-Waterman (Penjajaran Tempatan) Untuk Penjajaran Urutan DNA

Abstrak

Prosedur asas bagi menganalisis kandungan urutan ialah melalui perbandingan urutan. Perbandingan urutan boleh ditakrifkan sebagai permasalahan untuk mencari bahagian mana yang sama dan bahagian mana yang berbeza, iaitu dengan membandingkan dua urutan dan mengenalpasti persamaan serta perbezaan di antara mereka. Pendekatan khas untuk menyelesaikan masalah ini adalah dengan mencari kesejajaran yang baik dan wajar antara dua susunan. Penyelidikan utama dalam projek ini adalah untuk menyesuaikan susunan DNA dengan menggunakan algoritma Needleman-Wunsch bagi penjajaran global dan algoritma Smith-Waterman untuk penjajaran tempatan dengan berpandukan kepada algoritma Pengaturcaraan Dinamik. Algoritma Pengaturcaraan Dinamik adalah dijamin dalam mencari keselarasan yang optimum dengan mengeksplorasi semua kemungkinan penjajaran dan memilih yang terbaik melalui teknik *scoring* dan *traceback.* Algoritma yang dicadangkan dan dinilai adalah untuk mengurangkan *gap* dalam menyelaraskan susunan mahupun panjang urutan tanpa mengorbankan kualiti atau kesahihan hasilnya. Bagi projek ini, satu kajian tentang bagaimana mengaplikasikan kekuatan pengkomputeran bagi Pengkomputeran Selari untuk mempercepatkan proses perbandingan yang panjang tanpa harus berkompromi dengan hasil yang optimum. Keselarian tersebut hanya diterapkan pada algoritma Needleman-Wunsch. Untuk mengesahkan ketepatan dan konsistensi pengukuran adalah diperolehi dari data yang dibandingkan diantara algoritma Needleman-Wunsch dan Smith-Waterman dengan Emboss (*global*) dan Emboss (*local*) bagi 600 data uji. Hasil kajian juga menunjukkan bahawa program *Needle* dan *Smith* dapat mengurangkan *gap* dan *mismatch*, tetapi tidak menjejaskan ketepatannya. Manakala hasil bagi keselarian algoritma Needleman-Wunsch menunjukkan bahawa keselarian hanya berkesan untuk urutan DNA yang panjangnya 3000 dan keatas, tetapi tidak menunjukkan sebarang peningkatan pada urutan DNA yang panjangnya kurang dari 500 walaupun menggunakan platform teras berganda.

Reducing The Search Space And Time Complexity Of Needleman-Wunsch Algorithm (Global Alignment) And Smith-Waterman Algorithm (Local Algorithm) For DNA Sequence Alignment

Abstract

The fundamental procedure of analyzing sequence content is sequence comparison. Sequence comparison can be defined as the problem of finding which parts of the sequences are similar and which parts are different, namely comparing two sequences to identify similarities and differences between them. A typical approach to solve this problem is to find a good and reasonable alignment between the two sequences. The main research in this project is to align the DNA sequences by using the Needleman-Wunsch algorithm for global alignment and Smith-Waterman algorithm for local alignment based on the Dynamic Programming algorithm. The Dynamic Programming Algorithm is guaranteed to find optimal alignment by exploring all possible alignments and choosing the best through the *scoring* and *traceback* techniques. The algorithms proposed and evaluated are to reduce the *gaps* in aligning sequences as well as the length of the sequences aligned without compromising the quality or correctness of results. In this project, a study on how to apply the computational power of Parallel Computing to speed up the lengthy process of comparing sequences without having to compromise on the optimal results. Parallelization is only applied to the Needleman-Wunsch algorithm. In order to verify the accuracy and consistency of measurements obtained in Needleman-Wunsch and Smith-Waterman algorithms the data is compared with *Emboss (global)* and *Emboss (local)* with 600 strands test data. Results show that the *Needle* and *Smith* programs are reduced *gaps* and *mismatch*, but do not affect the accuracy. Results on the parallelization of Needleman-Wunsch algorithm shows that the parallelization is only efficient for 3000 length of DNA sequences and above, but does not show any improvement for less than 500 lengths of DNA sequences although using multiple core platforms.

CHAPTER I

INTRODUCTION

1.1 Introduction

Nowadays, the current efforts in molecular biology are producing an abundance of biological data stored in numerous databases that spread all over the world. This wealth of data and the complexity of biological processes provided exciting opportunities for new knowledge discovery especially in Bioinformatics. Bioinformatics or Computational Biology refers to an emerging, interdisciplinary field in computer technology, including software, hardware and algorithm, are applied to solve problems arising in biology study. One research area of particular interest in the field is to develop tools for processing biomolecular data (C.Y. Chang, J.T.L Wang & R.K Chang, 1998).

Bioinformatics is the application of information technology to the field of molecular biology. Basically, it uses techniques and concepts from informatics, statistics, applied

mathematics, chemistry, biochemistry, physics, and linguistics to solve biological problems, typically through computer programs and mathematical models.

Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. The major efforts in the field include sequence analysis, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein interactions and the modeling of evolution. Algorithmic development is an important part of bioinformatics, because, techniques and algorithms were specifically developed for the analysis of biological data, for example, the Dynamic Programming algorithm for sequence alignment, and the Machine Learning algorithm for data mining from micro arrays.

The biological or biomolecular data include *protein*, DNA (*deoxyribonucleic acid*), and RNA (*ribonucleic acid*). However, the data representation in this project is DNA sequences. DNA in a genome is not neatly arranged and stores an organism's genetic information in the form of a long sequence of molecules. Specifically, the information is encoded using four key chemicals, *adenine*, *thymine*, *guanine* and *cytosine* (abbreviated as A, T, G and C) (S.K. Moore, 2000).

An example of a genome sequence is shown in Figure 1.1. The DNA sequences of hundreds of organisms have been decoded and stored in databases. This biological sequence data can be obtained from variety of public and private databases.

```
CCTTCATCTAGGAGTTGAGAAGGGTAGATAAGATTCTTGGATACTAGGTATTTAAGAACTTTCTCAG
ATGAAAGGAAGCTGGGAACAAAGTAAGAAAGAATACCTTTTAGGATTCACAAAATTATGAGAAGTCA
GCCACATACGGTAGGTCAGCTTTTTAATGTATTTGCTCCTTTTCTTATTCATTGACCTGTGAAAGGA
AAATATCTTGGACCTCCAAAATCACTAAGAAAACTCAAGCTGGATATTGCTTAGGGCAAACCTGCCT
ACCATTCTATTCAAAGTCACAGAGCATTACAAGAAGCGTATTATTATTATGCTTAAGTCAGAATATT
```

Figure 1.1 : Sub-sequences of DNA for Western Gorilla.

With the growing amount of data, it became impractical to analyze DNA sequences manually, so faster algorithms and tools are needed. Sequence analysis is the process used to find information about a *nucleotide* or *amino acid* sequence using computational methods (T.J. Vision & A. McLysaght, 2003). Common tasks in sequence analysis are comparing of sequences in order to find similar and dissimilar in compared sequences (sequence alignment), identifying gene-structures, determining the similarity of two genes, determining the protein coded by a gene, and determining the function of a gene by finding a similar gene in another organism with a known function. This work concentrates on efficient developing and evaluating algorithms for determining the similarity of two genes.

The fundamental procedure of analyzing sequence content is sequence comparison. Sequence comparison is the cornerstone of Bioinformatics. Sequence comparison is regarded as one of the most fundamental problems of computational biology, which is usually solved with a technique known as sequence alignment. Sequence alignment can be defined as the problem of finding which parts of the sequences are similar and which parts are different. Generally, it is the process of comparing two sequences to identify similarities and differences between them. So, a measure of how similar they are is also desirable, then,

a typical approach to solve this problem is to find a good and plausible alignment between the two sequences.

Dynamic Programming algorithm for sequence alignment is used (Needleman & Wunsch, 1970). Dynamic Programming algorithm is one of the major strategies for designing algorithms. There are many algorithms written that use the approach of Dynamic Programming. However, Needleman-Wunsch algorithm was the first to introduce Dynamic Programming to compare biological sequences for finding the global alignment between two sequences (Needleman & Wunsch, 1970). Later, the improvement from Needleman-Wunsch algorithm proposed Smith-Waterman algorithm to find the best local alignment between two sequences (Smith & Waterman, 1981). The Needleman-Wunsch and Smith-Waterman algorithms developed are to solve the sequence alignment problems that have computational complexities. In this project, the study on how to analyze large sequences and to reduce the search space and time complexity without compromising the accuracy and efficiency is presented. This is by evaluating the performance of Needleman-Wunsch and Smith-Waterman algorithms in finding the optimal alignment between a pair of DNA sequences.

Since the nature of Dynamic Programming requires to store and compute an *mxn* matrix, many researchers spending most of the effort to reduce calculation time. So, Parallel Computers are also used to reduce the computation time (S. Rajasekaran , V. Thapar, H. Dave & C. H. Huang, 2008). This project also applied the power of parallel computers to speed up the process of comparing sequences without having to compromise the optimal

result. However, this works only focused on the parallel process for Needleman-Wunsch algorithm.

As the database of proteins grows larger, faster algorithms become more important to be able to quickly compare a given sequence to the entire database. The concern on both computational speed and memory management had become a great concern in searching, matching and analyzing. High performance computing such as parallel computers are used, in order to speedup the processes. Researchers also worked on issues in memory management.

1.2 Problem Statement

There has been a tremendous thrust in Bioinformatics research over the last decade. Being a relatively new discipline, Bioinformatics offers a large number of challenging problems. Following are the problems arising from Bioinformatics :

i. Many available algorithms and techniques in solving the problems of sequence alignment.

There are many algorithms that maximize speed and do not concern with the accuracy of the result alignment. And also, there are many algorithms that maximize accuracy and do not concern with the speed. Most current sequence comparison methods used in practice, such as, BLAST (S.F. Altschul, T.L. Madden, J. Zhang,

et. al, 1997) and FASTA (D.J Lipman. & W.R Pearson, 1985) are based on Heuristics (Michalewicz & Fogel, 2000) which are much faster, but do not provide optimal results. Although sequence comparison algorithms based on the Dynamic Programming method, such as Needleman-Wunsch and Smith-Waterman, provide optimal solutions, but tend to have very high computational complexities.

ii. The problem of finding optimal alignment.

Another problem in the field of Bioinformatics for sequence alignment is the comparison of two sequences of biological data, such as DNA sequences or proteins, and finding the optimal alignment for the sequences compared. Moreover, the comparison of a sequence against a large set of sequences, such as a large database and effort to search for the most similar sequences in the set is important. In order to apply Dynamic Programming to the optimal alignment problem, the problem structure must exhibit the properties, which the problem can be broken into subproblems and the optimal solution of the problem can be efficiently computed from the optimal solution of the subproblems. The goal is to produce the best alignment for a pair of DNA or protein sequences (represented as strings of characters). A good optimal alignment has zero or more gaps / mismatch inserted into the sequences to maximize the number of positions in the aligned strings that match. For example, consider aligning the sequences “ATTGGC” and “AGGAC”. By inserting gaps (“-”) in the appropriate place, the number of positions where the two sequences agree can be maximized: “ATTGG-C” and “A–GGAC” (K. Charter, J. Schaeffer & D. Szafron, 2000).

iii. The computational and space complexity of the sequence alignment algorithms.

The other main problem in sequence alignment is making sequences having the same size with the insertion of gaps in locations along the sequences and creating a correspondence between sequences. The sequence alignment problem can be illustrated as, given a scoring function that measures the score of aligning characters at the same position from each sequence, calculate the total score of the alignment by adding the scores of all positions and find the maximum total score of every possible alignments. The effort is still being studied by researchers in finding the best way to align sequences by reducing the gaps and sizes of the sequences.

iv. The effort to determine the degree of similarity for the optimal sequence alignment.

One of the problems in the comparison of sequences of biological data is an effort to determine their degree of similarity. Algorithms, such as the Smith-Waterman algorithm, are used to compare two sequences allowing for genetic mutations such as insertions, deletions, or substitutions. It is also useful to compare multiple sequences, however, the number of operations involved in non-heuristic algorithms is proportional to the product of the lengths of the sequences, and as a result, require a great deal of processing time. It is useful to determine the evolutionary history or phylogenic tree of a given a set of sequences. This problem is also computationally intensive. The non-heuristic algorithms also have an exponential time complexity based on the length of the protein, and as a result, it cannot be solved in a reasonable

amount of time for even medium length proteins. However, the accuracy in sequence alignment is compromised because many algorithms (including the Dynamic Programming) require human intervention while they are optimizing results. This intervention will have to be done by biologists who are very familiar with the data and thus the usage of such an algorithm is limited biologist availability. Often, it is not necessary to find the most accurate alignment among the sequences.

v. Improving the speed for the sequence alignment algorithms.

A search of a biological sequence database can take a great deal of time for two main reasons. First, sequence alignment algorithms have a significant time complexity because they must allow for genetic mutations, such as character insertions, deletions, or substitutions. Second, the size of a sequence data can be very large, some on the order of billions of characters. This biological sequence data can be obtained from a variety of public and private databases.

1.3 Objective

The objectives of this project are :

a) To study and compare techniques in DNA sequence alignment.

b) To develop the optimal sequence alignment algorithm for two sequences of DNA by finding the optimal alignment in much less time and space complexity.

c) To design scoring functions without using the substitution matrices (e.g. PAM or BLOSSUM) in solving the computational complexities efficiently and for improving the performance of space and time complexity.

d) To evaluate the accuracy and the complexity of optimal sequence alignment.

e) To evaluate the effect on parallelization for Needleman-Wunsch algorithm.

1.4 Summary

Bioinformatics derives knowledge from computer analysis of biological data. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. The main kinds of information stored in biological databases are DNA, RNA

and proteins sequences. The data representation in this project is DNA sequences. As biological databases grow in size, faster algorithms and tools are needed. However, the field of bioinformatics consists of many computationally challenging problems, many of which involve very complex system. This work emphasizes the sequence alignment problems, where a variety of computational algorithms have been applied to the sequence alignment problems, including slow but formally optimizing methods like Dynamic Programming and efficient Heuristic or Probabilistic methods designed for large-scale database search.

In order to apply Dynamic Programming to the optimal alignment problem, the problem structure must exhibit the properties, which the problem can be broken into subproblems and the optimal solution of the problem can be efficiently computed from the optimal solution of the subproblems. It can be seen that the problem of finding optimal alignment obeys the properties and hence Dynamic Programming can be applied very fruitfully. Thus, the problem exhibits optimal substructure and Dynamic Programming can be applied by using Needleman-Wunsch algorithm for global alignment, Smith-Waterman algorithm for local alignment and parallelized the Needleman-Wunsch algorithm in finding the optimal alignment in much less time and space complexity.

In Chapter 2, the studies have been carried out in more detail with respect to sequence alignment algorithms and comparing the efficiencies of the various algorithms and see what sacrifices the algorithms make in exchange for speed and in reducing the search space.

CHAPTER II

LITERATURE REVIEW

2.1 Biological Concept

At a microscopic level, living things are all composed of cells. Cell is the basic structure of life and performs essential functions for all living organisms. In biology, an organism is any living system, such as animal, plant, fungus, or micro-organism. If a human look at in a mirror, what they see is about 10 trillion cells divided into about 200 different types of cells. Muscles are made of muscle cells, livers of liver cells, and there are even very specialized types of cells that make the *enamel* for teeth or the clear lenses in eyes. Understanding cell is important to understand how human body works. Everything from reproduction to infections or to repairing a broken bone happens down at the cellular level. It is also important to understand new areas like biotechnology and genetic engineering, by understanding cells. Some of the terms commonly used are Biotechnology, Gene splicing, Human genome, Genetic engineering, Recombinant DNA, Genetic diseases, Gene therapy,