UNIVERSITI TEKNOLOGI MARA

DESIGN OF DNA SEQUENCE ALIGNMENT ACCELERATED SYSTEM USING 2-DIMENSIONAL ARRAY AND CUSTOM INSTRUCTION ON FPGA

NUR DALILAH BINTI AHMAD SABRI

MSc

June 2018

ABSTRACT

Nowadays, the (Deoxyribonucleic Acid) DNA sequence database has been increased linearly with the time taken to comparing with the search of DNA sequence alignment system. Hence, the requirements of the computational method for comparing DNA sequences of the sequence alignment area are in high demand. Thus, this research concentrates on the optimization techniques based on Custom Instruction (CI) and Rectangular Dimensional Array (2D) that proposed on Smith-Waterman (SW) algorithm. The proposed techniques are used to improve the similarity searching of the DNA sequence alignment system. The system performance is often degraded due to the suffer issues in terms of the time and sensitivity. Other than that, an implementation of a platform called Field Programmable Gate Array (FPGA) has been used in optimizing the 2D Array to perform accelerate sequence alignment activities. The proposed 2D Array on SW algorithm has been developed and designed by using ideas over previous work of CI on SW algorithm. As the result, when comparing the DNA database within 2x2 to 64x64 base pairs with the 2D Array system on FPGA, the acceleration in hardware version gives improvement in the result performance. The highest percentage speedup of the optimization core can goes up to 52% when the proposed technique is applied in aligning two based pair DNA sequences and requires only one bit data storage over matrix cell. In conclusion, the proposed technique shows the result compared to the other algorithms and the results is consistent with the expected theoretical result analysis. Ultimately, the performance time taken to complete the system is reduced proportionally from 1.09us till 0.1617s where it saves the time and aligns the two DNA sequences with 32 sequences in one clock cycle.

ACKNOWLEDGEMENT

First and foremost, I would like to express my greatest gratitude to Allah SWT, the Almighty. Alhamdulillah. I would like to appreciate my supervisor, Associate Professor Zulkifli Abd. Majid and co-supervisor Dr. Syed Abdul Mutalib Al Junid for their support and encouragement towards the successful completing my Masters study. One simply could not wish for better or friendlier supervisors.

I would also like to thank to a great friend of mine Mrs. Nur Farah Ain Saliman, who support me with her patience and knowledge throughout my thesis endlessly. Also, thank you to my beloved mother, Nor Azizan Abdullah, who supported me financially and reads my whole thesis even when she is a very busy woman herself. Without their encouragement I would have not reached the level where I am now.

Last but not least, I would like to thank my dearest father, Ahmad Sabri Harun, who supported me financially, my beloved husband Mohd Shafizan Elias, who helped me tremendously, my lovely son Muhammad Arsyad Mohd Shafizan and the rest of my family and friends who keep encouraging me in whatever I do. Thank you very much.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 BACKGROUND

The term Bioinformatics is defined as the informatics science tools used for storing, organizing and analysing biological data using advanced computing techniques [1][2]. The biological data is known as the cell information that encodes the genetic instruction through Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA) and proteins [3][4]. Under primary data, those biological data is important in many application development such as creating a new vaccine, detect the functioning of the living organism, disease detection, find the curing solution and many other applications [5][6]. However, with a huge amount of biological data from a hundred to several billions of nucleotides related to various living organisms has been collected [3][7] it causes the informatics science tools performance degraded [8]. This happen when the execution time of algorithms increases and the used of large memory space are needed as well [9][10]. Therefore, the demand for high performance computational method for searching and comparing biological sequences are also increased significantly [11].

In Bioinformatics field, there are a variety of major research areas and it included many other tasks like sequence alignment, protein structure, function prediction and phylogenetic analysis [12][13]. With that, it is important and become a relevant field in this thesis to develop an efficient computational techniques based sequence alignment task. The sequence alignment task is widely used in this field to process the biological information by searching the similarity sequences [8]. Additionally, sequence alignment also can be referred as a primary or fundamental tool in molecular biology and is used to discover biological information extracted from the large amounts of sequenced DNA and promises to help understand possible genetically transmitted diseases [14]. Specifically, sequences alignment is the process of comparing two or more sequences by searching for a series of individual characters or characters patterns that are in the same order in the sequences data for detecting region of similarity [15].