UNIVERSITI TEKNOLOGI MARA

DEOXYRIBONUCLIEC ACID (DNA) FRAGMENTATION ACCELERATION TECHNIQUES USING LINEAR DIMENSIONAL ARRAY (1D) ON SMITH WATERMAN AFFINE GAP PENALTY (SWAGP) AND VITERBI PAIR HIDDEN MARKOV MODELS (VPAIRHMMS)

NUR FARAH AIN BINTI SALIMAN

Thesis submitted in fulfillment
of the requirements for the degree of
Master of Science

Faculty of Electrical Engineering

June 2018

ABSTRACT

Computational biology is rapidly pushing the advancement bioinformatics field which involves a wide range of areas, including the Deoxyribonucleic Acid (DNA) sequence alignment similarity region. The technique that involve in similarity region is sequence alignment for finding the similarity region between sequences. This technique the of the DNA sequence alignment gaining close attention due great impact through the various area such as in biological data assembly, comparative biological data, drug design, finding homology and specific sequences and evolutionary biology trees. With the increased number of DNA database, it causes the finding for a similar sequence over large biological database is unable to produce results within a realistic time. Hence, for acceleration improvement over acceleration technique that take account the accuracy, sensitivity, speed and size of architecture. These studies proposed an acceleration technique over the Smith-Waterman Affine Gap Penalty (SWAGP) and Viterbi pair Hidden Markov Models (VpairHMMs) based on Field Programmable Gate Array (FPGA). In order to facilitate clearer of the designs, SWAGP and VpairHMMs development is divided into four stages. First, the SWAGP and VpairHMMs algorithm are started with the theoretical verification. Secondly, the development based on Software Version (SV) on FPGA for both algorithms. Next, optimization has been implemented for accelerate both algorithms by upgrade the SV into Custom Instruction (CI). Finally, the algorithms are designed and implemented on hardware accelerator into Linear Dimensional Array (1D) acceleration technique. Twelve (12) tests for each design (SV, CI and 1D) with ranged from 1 until 2048 length base-pair was conducted. The cores has been designed using Verilog HDL and implemented on DE2-115 FPGA (EP4CE115F29C7). The results show that the proposed structural design archived runtime performances of between 8 until 51 percent (SWAGP) and about 7 percent (VpairHMMS) for SV against CI. Next, SV against 1D showed up that the average speed up results range is between 11 until 76 percent (SWAGP) and 10 until 11 percent (VpairHMMS). The CI case against 1D showed that the speed up results range is between 4 until 52 percent (SWAGP) and 3 until 11 percent (VpairHMMS). In conclusion, the runtime depends on the computational method that being used such as software, hardware and system design task implementation.

ACKNOWLEDGEMENT

All praises due to ALLAH, the most gracious, the most merciful. I would like to express my sincere thanks to those who have involved and supported me towards the successful completion of my Master study.

First and foremost, my deepest gratitude goes to may supervisor, Associate Professor Zulkifli Abd. Majid and co-supervisor Dr. Syed Abdul Mutalib Al-Junid of the Faculty Electrical Engineering at Universiti Teknologi MARA (UiTM). Thank you very much indeed, the door to Associate Professor Zulkifli Abd. Majid and Dr. Syed Abdul Mutalib Al-Junid office that was always open whenever I ran into a trouble spot, had a question about my research or writing, for suggestion, criticism and guidance. This invaluable efforts, patient and kindness to increasing my knowledge especially in hardware design, technical writing and more.

Secondly, this appreciation goes to my labmate or teammate Mrs. Nur Dalilah Ahmad Sabri who has studied together about the fundamental parts of bioinformatics field, sequence alignment, hardware description language, and hardware debugging skills, C programming and software design task debugging skills. Without her passionate participation, the research project could not have been successfully done. In addition, this appreciation also goes Mr. Christopher Lee as the reader of this thesis which I am really gratefully indebted to his very valuable comments on this thesis.

Last but not least, I must express my very profound gratitude to my husband Mr. Luqman Hakim Mat Sout, to my lovely son Mr. Muhammad Ar Rafif Luqman Hakim, to my parents (Mr. Saliman Abu Denan and [redacted]) and sibling (Miss Nurul Natasha Saliman and Mr. Muhammad Afieq Saliman) for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

TABLE OF THE CONTENTS