HYBRID OPTIMAL PATH TRACEBACK SYSTEM DEVELOPMENT FOR OPTIMIZING DNA SEQUENCES ALIGNMENT

NUR DALILAH BINTI AHMAD SABRI

Faculty of Electrical Engineering

Universiti Teknologi MARA

40450 Shah Alam, Selangor

MALAYSIA

TABLE OF CONTENT

ABSTRACT

This paper present the new hybrid optimal path trace back system development for solving optimal path trace back complexity issue. The objective of this paper is to study several optimization techniques and determine the best optimization technique for optimizing the optimal path trace back system. Therefore, two designs are proposed and analyzed for in this study.

The project is divided in two stages which are theoretical and experimental design. In theoretical design, the proved of concept for the proposed design is calculated based on mathematical equation which defined by Smith-Waterman algorithm. Design construction which covers code development, compilation and simulation is carried out under experimental design for both of the proposed designs using Altera Quartus II version 9.0 EDA tools and targeted to Cyclone II EP2C35 FPGA at 100MHz clock cycle. As a result, the second design required three times design area as compare to the first design for determine the optimal path for the same matrix size. Therefore, the first design is the best hybrid approach for the optimal path trace back size since theoretical result has shown both of the design has the same output but the second design suffers in terms of large design area.

Keywords DNA sequence alignment, dynamic programming, Smith-Waterman algorithm, optimal path trace back.

ACKNOWLEDGEMENT

بسماﷲارحمنا رحيم

Assalamualaikum w.b.t

Alhamdulillah and thank God as His willingness and blessings that brought upon completing my final year project thesis which entitled Hybrid Optimal Path Traceback System Development for Optimizing DNA Sequences Alignment. First of all, I would like to express my grateful thanks to my supervisor, Mr. Syed Abd Mutalib Al Junid for all the information and guidance regarding to my project. He has being enormously helpful in giving suggestions, improving this project. Without his guidance I won’t think that I am able to complete this project on time.

Furthermore, I would like to extend my acknowledgement to other lecturers for their thought and ideas even though they did not involve directly into this project report. Then, a lot of thanks to my entire fellow friends which has also give many ideas and share their suggestion for my project. It is always a pleasure working with them. The information and the brilliant knowledge that I would gained from this project can give me benefit in future. Finally, words cannot adequately express my gratitude colleagues as well who gave full support and commitment while doing this project and help me when I encounter problems.

Last but not least, I wish to record this appreciation to my family members for their moral support.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND OF STUDY

Bioinformatics is the combination of biology and information technology which consist of the mixes of informatics, statistics, mathematics, physics and biological sciences for the analysis of biochemical genetic and other things that related with the biological data. DNA sequence alignment is the most highlighted problem in bioinformatics. Furthermore, it is needed to compare the target DNA sequences with the entire source of DNA sequences in the database. The complexity of the process and the need of accurate result in sequence alignment have contributed to rapid improvement of the bioinformatics tools. However, high memory usage due to the large number of sequence in database, algorithm complexity and the high power usage are the contributor factor that cannot be solved until now. Smith Waterman is used in the sequence alignment but the high sensitivity and accuracy remains the Smith Waterman algorithm as the main algorithm for DNA sequence alignment.