UNIVERSITI TEKNOLOGI MARA

DESIGN OF DNA FRAGMENT ASSEMBLY USING IWP METHOD ON APPLICATION-SPECIFIC-INTEGRATED-CIRCUIT

HASNILIATI BT HASSAN

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

Faculty of Electrical Engineering

October 2017

ABSTRACT

Nowadays, DNA sequence assembly is needed due to limitation with current sequence technology. Depending on the technology used, it can only read small pieces of base between 20 to 1000 bases. The short read fragments generated by DNA sequencing technology offer challenges in the assembly process because some of them contain sequencong errors, coming from various sizes, and contain lots of repeats and overlaps. Therefore, computation capability and requirement is increased in order to assemble large amounts of short read data that are generated by sequencing technology. In recent years, many software programs have been developed in order to improve the assembly process. The main focus is how to efficiently reconstruct full strands of DNA based on the pieces of data they are able to record at the lowest cost possible. Most assemblers nowadays are software-based which take a significant amount of time to execute. Hence, in this research a method is proposed to design the DNA sequence assembly accelerator. The method proposed is to design DNA assembly algorithms in Verilog HDL and implemented it in ASIC design flow. The key innovation is to implement DNA sequence assembly algorithms on ASIC to make it a synthesisable IP block. After reviewing various algorithms, it was decided to apply de Bruijn graphs and Eulerian circuits that are based on graph theory. A de Bruijn graph is a compact representation based of short words (k-mers). While the Eulerian paths that existing in the de Bruijn graph are represented as DNA sequence assembly outputs. This combination of both de Bruijn graph and Eulerian circuit is called the Idury Waterman Pevzner (IWP) Method. The IWP module was modelled and designed in Verilog HDL using Xilinx ISE Design Suite version 14.2. The implementation of design in ASIC was then done using Synopsys EDA tools. In this work, the main focus is to seek optimal solutions for DNA fragment assembly problems in terms of assembly accuracy. Simulation results showed that the IWP module in Verilog HDL can assemble short reads data efficiently same as in theory, besides eliminating repeats. Further analysis has been conducted on the IWP module in ASIC design flow in terms of assembly running time, power consumption and total area consumed. The results obtained are LVC clean, no DRC error, positive slack for both setup and hold time where 83.541ns and 0.097ns respectively. All these analyses were performed using industrial standards via Synopsys EDA tools which are VCS, DC and Astro.

ACKNOWLEDGEMENT

Firstly, I would like to convey an abundance syukur to Allah swt for His guidance and Help. All knowledge in the world belongs to Him and I was nothing if He did not shed some light for me to proceed and completing this long and challenging journey. I would like to express an enormous gratitude to my project supervisor, PM Zulkifli b Abd. Majid, for giving the door of opportunity for me to embark on my Master in Research titled DNA Fragment Assembly Design. It was a great honour to work under his supervision as his opinions and views inspire me to explore more about DNA fragment Assembly Technology in IC Design. This explorative journey led me to explore various digital design techniques. The guidance and support were precious throughout this study.

A special thanks to my co-supervisor, Encik Abdul Karimi b. Halim, who had a lot of patient to guided me throughout my study. Moreover, research knowledge regarding DNA Technology design and his experience in IC Design skill and techniques had major impact to the proposed IWP Verilog Module design.

Finally, I would like to credit my family especially my husband, Qamarul Bahrain b. Mohd. Badiuzzaman for motivating me when I was lost and supporting me against all odds. Thanks to my mother Rohani bt. Chik, my kids; Aisyah Sofiyyah, Isa Nu'man, Muhammad Ridhuan as well as Ibrahim Auf, my other family members and friends for your patience and perseverance to continue to be by my side during good and bad times.

TABLE OF CONTENTS

CHAPTER ONE
THESIS INTRODUCTION

1.1 BACKGROUND STUDY

Deoxyribonucleic acid (DNA) sequencing is the most utilised approach for genome mapping [1]. It is a process used to determine the protein synthesis of biological entities in order for scientists to study the structures, characteristics and functioning of various organisms. Information about this process has become crucial and utilised in so many fields, for instance in diagnostic, biotechnology, forensic biology and biological systems [2]. Due to this, DNA sequence technology has significantly accelerated biological process and discovery [3].

However, since genomes vary widely in size from the smallest bacteria to the larger human form, it can cause complications in DNA sequence technology. This is because the process to decode DNA base pairs involves from 600,000 DNA base pairs of smallest bacteria to 3 million of base pairs contained in human genomes[4]. The current limitation in the sequencing technology is that it cannot read whole genomes at one go but only in small pieces of between 20 to 1,000 bases [5][6][7]. Furthermore, most readings resulting from sequencing machines are with errors, which may be caused from the machine itself or the researcher during laboratory work. The DNA sequence structure itself is also another issue due to repeats that occur in some regions of the genomes [8]. Other than that, imperfect data sets as well as DNA sequences that come in various sizes correspondingly contribute to complicate the DNA assembly process [2][9][10]. All of these issues mean that researchers have to keep studying to improve current DNA fragment assembly algorithms in order to reconstruct full strands of DNA [9][11].

The "Overlap-Layout-Consensus" (OLC) approach was the most applied method in DNA fragment assembly algorithm for the past 25 years [12][13]. This approach basically arranges all DNA fragments into their relative positions and orientations by referring to their overlap information. Assembly is then achieved by constructing all multiple alignments to achieve a consensus sequence [2]. However, with the big gap between short reads produced and higher coverage needed, not only