PROTIEN SEQUENCE ALIGNMENT WITH GPU: DATABASE OPTIMIZATION

This thesis is presented in partial fulfillment for the award of the

Bachelor of Engineering (Hons.) in Electronics Engineering

UNIVERSITI TEKNOLOGI MARA

(MARCH 2013 – JANUARY 2014)

AHMAD FAIZ BIN MOHD RAHI
FACULTY OF ELECTRICAL ENGINEERING
UNIVERSITI TEKNOLOGI MARA
40450 SHAH ALAM SELANGOR
MALAYSIA

ACKNOWLEDGEMENT

In the name of Allah, the Most Merciful and the Most Compassionate.

Alhamdulillah and peace upon Prophet Muhammad S.A.W, this final year project completed according to time and objectives required.

Firstly, I would like to pay my gratitude to Allah S.W.T for giving me strength to be able to complete this project. Without His blessing and permission, this project could not have been completed. I would like to give my sincere appreciation to my supervisor Miss Ili Shairah binti Abdul Halim for her concern, advices, supports and encouragement throughout this thesis progress. This project would not have been possible without the assistance and support from my supervisor. My gratitude also goes to my coordinator of Project (EE210) FKE, UiTM Shah Alam, Madam Nor Akmar binti Mohd Yahya for her valuable guidance in the completion of this project.

I would like to thanks to EE210 student and all my friends for academic guidance, support, encouragement and advice that directly or indirectly helped me in during preparing this project.

Finally yet importantly, thoughtful thanks to my parents, who gave me an appreciation of learning and taught me the value of perseverance and resolve. Thanks for inspiring me in such a means that could not be written in words.

ABSTRACT

In the era of advance technology, people desire new and very powerful chip in their graphic hardware. Graphic processing unit are called as GPUs is increasing rapidly in the last few years. The use of graphic processing unit (GPUs) is to accelerate the graphic rendering. Alignment algorithms are used to find similarity between biological sequences, such as DNA and proteins. By aligning a sequence with a database, similar sequences can be found. These can be used to identify the source of a query sequence, to find commonalities between organisms, or to infer an ancestral relation. Various methods of performing biological sequence alignment exist, including dynamic programming and heuristic methods. Dynamic programming methods are guaranteed to find all optimal alignments, but are relatively slow; heuristic methods are faster but less precise.

This thesis investigates the acceleration of one such optimal algorithm, the Smith-Waterman local sequence alignment algorithm, by using graphics processing units (GPUs). A fully functioning GPU-based protein database search tool was designed, implemented and optimized. The optimizations mostly concern the elimination of memory bottlenecks and the conversion of the database to a format well suited for GPU use. The final implementation offers the same features its CPU-based counterparts do, such as user configurable scoring and substitution matrix settings, and includes a web interface for convenient and remote usage.

The performance of the GPU accelerated implementation was evaluated and compared to other solutions. We achieve a 1.9 times faster over the serial method Ssearch. The new implementation improves the performance by reducing the number of memory accesses and optimizing the database organization. The database is organized in equal length sequence sets resulting in an equal workload distribution for all the threads of each multiprocessor on the GPU. In comparison with the state-of-the-art implementation on an NVIDIA Geforce 610M graphics card, our implementation reports a 1.9 times performance improvement in terms of execution time.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 A REVIEW OF MOLECULAR BIOLOGY

What follows is a short review of the basics of molecular biology. Although no in-depth knowledge is required of the chemical processes involved, subjects such as DNA and protein construction are critical to understanding the relevance of sequence alignment, the procedure upon which much of this thesis is based. The information in this section largely comes from [15] and [16].