Ain Shams University
Faculty of Computer & Information Sciences
Scientific Computing Department

Developing a Parallel Algorithm for Protein 3D Structure Comparison and Classification

Thesis submitted to the department of Scientific Computing In partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences

By
Nada Mamdouh

BSc in Computer and Information Sciences (2006)

Faculty of Computer and Information Sciences - Ain Shams University - Cairo

Under the supervision of

Prof. Dr. Mostafa Gadal-Haqq M. Mostafa

Professor of Computer Science in Computer Science Department

Faculty of Computer and Information Sciences, Ain Shams University

Prof. Dr. Mahmoud E. Gadallah

Professor of Computer Science and the head of Computer Science Department

Modern Academy for Computer Science and Management Technology

Dr. Hala Moushir Hassan Ebeid

Assistant Professor in Scientific Computing Department

Faculty of Computer and Information Sciences, Ain Shams University

Cairo, 2017

ACKNOWLEDGEMENTS

I want to thank my supervisors Prof. Mostafa Gadal-Haqq, Prof. Mahmoud Gadallah and Dr. Hala Ebeid for being too generous and patient. I appreciate their support and encouragement all the way through my study.

I would like to extend my gratitude to my parents who offered me endless care, love, prayers, and support. Last of all, I would like to show appreciation to my husband Eng. Mohammed Ali for standing by me during these tough times.

ABSTRACT

The importance of pairwise protein three-dimensional (3D) structure comparison process in structural bioinformatics has become vital. However, the complexity of this process is categorized as non-deterministic polynomial-time hard (NP-hard) which forced bioinformaticians to develop different algorithms to overcome the heavy computational execution time. Still, most of these algorithms tend to achieve accurate comparison results regardless of computational execution time.

In this thesis, we propose a parallel algorithm, PTM-MatAlign, which is an enhanced and accelerated version of Matrix Alignment (MatAlign). This proposed algorithm is designed to use Template Modeling Score (TM-score) in the comparison process instead of the MatAlign regular score function. Also PTM-MatAlign provides two parallel paradigms; one is built to run on NVIDIA Graphical Processing Units (GPUs) using Compute Unified Device Architecture (CUDA) programming model and the other one is built to run on multi-core CPUs using Open Multi-Processing (OpenMP). Moreover, the comparison process is based on two-level pairwise alignment and the proposed parallel paradigms parallelize only the first level since the second level is inherently sequential.

The parallel algorithms are implemented using two common APIs for C++ parallel programming, which are OpenMP 2.0 for multi-core CPUs and CUDA 6.5 for multi-core GPUs. To run the CUDA parallel implementation, we used an nVIDIA GeForce GTX 860M series (Maxwell class) graphics card. This GTX 860M has nVIDIA compute capability 5.0 and consists of 5 streaming multiprocessors. Each multiprocessor has 640 cores, shared memory with size 49 KB per block, total constant memory with size 65 KB, 65536 registers, and total global memory with size 2GB. For running the OpenMP parallel implementation, we used a hyper-threaded dual-core 2.5 GHz Intel CPUs which provides at least 8 and up to 16 independent Pthreads. The results show that beside the significant improvement of the parallel implementation over the sequential one, it also shows that the multi-core GPU parallel implementation improves speedup over the multi-core CPU parallel implementation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

1. Nada M. A. Mohammed, Mostafa G. M. Mostafa, Hala M. Ebeid, Mahmoud E. A. Gadallah, “Parallel Protein Structure Alignment: A Comparative Study of Two Parallel Programming Paradigms”, International Journal of Computer Applications (IJCA), Vol. 150, No.7, pp: 43-48, September 2016.

2. Nada M. A. Mohammed, Mostafa G. M. Mostafa, Hala M. Ebeid, Mahmoud E. A. Gadallah, “PTM-MatAlign: A Fast GPU-Based Algorithm for Pairwise Protein Structure Comparison”, Journal of Computational Biology (JCB) – Revised version submitted in December 2016.

LIST OF ABBREVIATIONS

Abbreviation	Explanation
HUGO	Human Genome organization
DNA	deoxyribonucleic acid
RNA	RiboNucleic Acid
CPU	Central Processing Unit
GPGPU	General Purpose Graphical Processing Unit
3D	Three dimension
Cα	Carbon Alpha
SSE	Secondary Structure Element
PDB	Protein Data Bank
NMR	Nuclear Magnetic Resonance
SCOP	Structural Classification of Proteins
SM	Streaming MultiProcesor
CUDA	Compute Unified Device Architecture
SDK	Software Development Kit
SIMT	Single Instruction Multiple Thread
API	Application Programming Interface
SMP	Shared-Memory Parallel
RMSD	Root Mean Square Deviation
AFP	Aligned Fragment Pairs
RTP	Residue Transition Pattern
OpenMP	Open Multi Procesing

CHAPTER 1

INTRODUCTION

CHAPTER 1
INTRODUCTION

Earlier in this century, biologists use developed some sort of technology in their experiments that enables them to collect information faster than they can understand it [1]. As an example, human fingerprints have large amounts of DNA sequence. As a result, how do biologists be sure which parts of that DNA manage the chemical processes? Although some proteins have a clear structure with obvious function, yet it is hard to determine the function of new proteins? And the prediction of how a protein will look like based on it is sequence is still hard.

Since most of the software developers have become possessed a strong knowledge about biology, it is expected to find more computational applications that have great interest in the molecular data. To implement applications to deal with such data, a software developer has to be familiar with not only classical biology, but also statistics and computer science [2]. These intersections of these three disciplines are gathered in new field named as Bioinformatics. So

Bioinformatics is the rescue!

1.1. HISTORY OF BIOINFORMATICS

The Austrian monk Gregor Mendel, who is known as the “Father of Genetics”, had discovered the bioinformatics field over a century ago by

cross-fertilizing different colors of the same species of flowers. He illustrated that his experiments in the agriculture engineering field could be definitely explained if it was controlled by factors transmitted from one generation to another. Since Mendel, bioinformatics and genetic record keeping have come a long way.

Later in 1980s, the first complete genome map was published of bacteria Haemophilus Influenza at the Human Genome organization (HUGO). In the 1990s, the Human Genome Project was started where a total of 1879 human genes had been physically mapped. In the late nineties, the final version of the human genetic map was published to be the end of the first phase of the Human Genome Project [3].

There is no agreed definition for bioinformatics. In general, it is defined as the use of the computational methods for comparative analysis of genome data. Also it is used to refer to the study of informatics processes in biological systems [3]. As such, it deals with methods for storing, retrieving and analyzing biological data, such as nucleic acid (DNA/ RNA) and protein genomic, biological and chemical data to support the biomedical processes. Several people narrow the definition of bioinformatics to include only those areas contending with the genome project sequencing data management. Others understand bioinformatics more broadly and include all areas of computational biology, including population modeling and numerical simulations [3].

Theoretically, bioinformatics is defined as the general use of computer technologies to process biological data. Commonly, a lot of people think that bioinformatics is just a synonym for “Computational Biology” which is the characterization for the molecular components of