INTRODUCTION TO BASIC GENETICS AND HARDWARE ACCELERATION OF PROTEIN SEQUENCE DATA PROCESSING

S.M. Vidanagamachchi[1], S.D. Dewasurendra[1], R.G. Ragel[1] and M. Niranjan[2]

[1] *Department of Computer Engineering, Faculty of Engineering, University of Peradeniya*

[2] *School of Electronics and Computer Science, University of Southampton, UK*

Introduction

Cell is the basic building block and the functional unit of all living organisms. Most organisms are multi-cellular and some primitive organisms such as bacteria are unicellular. Living content of the cell that is surrounded by plasma membrane is known as protoplasm, which consists of many compounds including water, mineral salts and organic compounds (carbohydrates, fats and proteins). Protoplasm of a cell consists of three main parts; nucleus (containing genetic material), cytoplasm (fluid that fills the cell) and plasma membrane. In eukaryotic organisms genetic information (Deoxyribonucleic acid/ DNA) is stored in the nucleus and in prokaryotes it is stored in the cytoplasm. DNA is organized as chromosomes and genes reside on it.

Main objectives of this paper are to introduce basic genetics, gene sequencing and then a way to use FPGAs to accelerate protein sequence data processing with an automated test bed. They are discussed next.

Genes are working subunits of DNA that carry the genetic information. Procedure of producing proteins involves two main processes called "transcription" and "translation".

Transcription starts at the promoter, which is located in the upstream of a gene. During transcription DNA sequence is read by RNA polymers (enzyme that produces RNA), transcribing the genetic information of DNA to produce a complementary RNA (ribonucleic Acid) strand known as mRNA (messenger RNA), which carries this coding information to places where protein synthesis is performed (ribosomes). Here DNA acts as a template for the synthesis of RNA. Then the produced mRNA is decoded by ribosome into a particular amino acid chain in the cell's cytoplasm. The large and small subunits of the ribosomes are located in the cytoplasm. This process is known as translation and the amino acid chain synthesised is known as a protein.

Genome is the entirety of hereditary information of a given living organism. It includes both coding part of the gene (exon) and non-coding part (intron) of genes. Human genome is stored on 23 chromosome pairs: 22 autosomal chromosome pairs and 1 sex chromosome pair. A chromosome is an organized structure of DNA and proteins. DNA contains all the genetic instructions that are used for the development and functioning of an organism.

DNA sequencing is a process of identifying exact order of bases (adenine, thymine, guanine and cytosine) in 23 chromosomes. Initial method (chemical degradation) was introduced by Frederick Sanger and Wally Gilbert to sequence DNA fragments containing up to 500 nucleotides. Subsequently chain termination methods and Dye terminator sequencing method were introduced. Then high-throughput sequencing (454 pyrosequencing, Solexa, SOLiD) was introduced. DNA sequencing needs improvements in sequencing speed, reliability and costs (Cells, 2010). Gel electrophoresis is used as the standard method of DNA/RNA or protein molecule separation method.

Sequencing process usually uses blood cells from females and sperm cells from males, because they contain all chromosomes necessary for the study. In order to sequence the DNA, first chromosomes must be broken into shorter pieces. Next template preparation is done; which is used to generate a set of fragments using above short pieces (template preparation step and sequencing reaction step). Here these broken pieces act as templates. Then fragments are separated using gel electrophoresis, which enables sorting of molecules according to size and charge (separation step). Next final base at the end of each fragment is identified (base calling step) and the sequence with 'A', 'T', 'C' and 'G' bases for each fragment are created. Finally assembling of generated sequences is done and once finished they are submitted to a public database such as 'GenBank' (Human Genome Programme, 2008).

There are two main protein identification methods: Mass spectrometry and peptide mass fingerprinting. Mass spectrometry can be used as an analytical method to measure the molecular mass of a sample and can be used for understanding the chemical structures of molecules. Mass spectrometry can be of two main types; MALDI peptide spectrum and Tandem peptide spectrum.

Problem Definition

There exist a huge amount of sequenced gene/protein data and therefore there is a pressing need of efficient computational methods to cope with them. However a significant bottleneck exists in the analysis of such data. Many attempts have been made by several research groups to develop efficient algorithms as well as dedicated hardware/software solutions to deal with this data explosion. Here the main objective is to use FPGAs to accelerate pattern matching of proteins.

Methodology

String matching algorithms can be used to identify several patterns that are found within a large string. Aho Corasick algorithm plays a major role in multiple string matching, since it is the best and the widest used algorithm for multi- pattern matching. Yoginder et al. (2008) have described an implementation for hardware acceleration of peptides pattern matching using this algorithm for the 1st chromosome of human genome.

They have used a maximum number of 30 and a minimum of 5 amino acids per peptide for testing. This can be used only with exact string matching applications and cannot be used with approximate string matching. They have used bit split Aho Corasick implementation to reduce storage space. Here each amino acid is represented by 5 bits and then 5 Finite State Machines (FSMs) were generated for the given set of peptides. We replicated this experiment with different data sets on a different FPGA (Altera) from the one they used (Xilinx). Since each FSM contains a large number of states (~250) the VHDL code generation was automated as described in a sequel paper. Resource utilisation and processing times were compared for different peptide lengths. Results for two typical lengths are given in Table 1.

Data collection, pre-processing and results

We have used protein data from GenBank database and to generate peptides we used PeptideMass software which is available in Expasy Proteomics Server.

Table 1. Experimental Results

Min len.	Max len.	*Tile utilis-ation	Mem bits	Total time (s)
3	8	1%	0%	29
5	70	7%	3%	155

* Percentage utilisation of Logic units in a tile of FPGA

Here we have used a minimum of 3 amino acids and a maximum of 70 amino acids per peptide and for the input string length a maximum of 500 amino acids. According to our test results the maximum number of peptides that can be coded into one tile of the Altera Cyclone II FPGA is 33 (when maximum number of states is around 250).

Discussion

According to the results when the length of peptides increase system needs more logic elements, compilation time and memory, it does not depend on the length of the input string.

References

Cells (2010). Retrieved from http://library.thinkquest.org/19347/cells.html.

Human Genome Programme (2008). Facts about Genome Sequencing, Retrieved from http://www.ornl.gov/sci/techresources/Human_Genome/faq/seqfacts.shtml.

Yoginder, S.D., Shane, C.B., Mark, L. and Susan, M.B. (2008). Accelerating String Set Matching in FPGA Hardware for Bioinformatics Research. BMC Bioinformatics 2008 Journal, 9, 197. doi:10.1186/1471-2105-9-197