VLSI IMPLEMENTATION OF ALGORITHMS FOR BIOINFORMATICS APPLICATIONS

T ANURADHA
P G SCHOLAR
RMK ENGINEERING COLLEGE
KAVARAIPETTAI
telagam.anuradha@gmail.com

DR. T. M. INBAMALAR M.E.,Ph.D
ASSOCIATE PROFESSOR
RMK ENGINEERING COLLEGE
KAVARAIPETTAI
tmi.ece@rmkec.ac.in

Abstract- BIOINFORMATICS is the science that gathers, stores, analyze and integrate biological and genetic information. The set of all genes in an organism is called GENOME .Computational genomics is the domain that helps in understanding biological properties of an organism. Advances in human genome research make it almost certain that genomics will become more important. Digital Signal Processing applications in genomic sequence analysis received great attention in recent years. Many methods are being developed to analyze Deoxyribonucleic acid (DNA) sequences. In this project, we present Technique for the implementation of DNA. Next generation sequencing technologies generate a large number of short fragments of fixed length reads of 16 bit data assembly. These reads of 16 bit data is given to 16 bit LUT. Its output is mapped with 16 bit FFT to produce DNA sequencing vectors. Thus, it produces 16 bit biological genome for the corresponding input. The proposed systems provides analysis of organization of DNA and protein sequence predicted from DNA which facilitates protein properties, structure and function as well as identification of mutations that lead to disease.

1. INTRODUCTION

The field of Bioinformatics and Next Generation sequencing (**NGS**) is one of constant change. The rapid growth and development depends on the innovation and creativity of scientists and their ability to see the potential in a basic molecular technique. This opened up many doors in genetic research, including a means of DNA analysis and identification of different genes based on their DNA sequences.

By Numerical Mapping Technique ―DNA can be implemented using FFT. Spectral analysis is necessary. Genome sequence is computed.DNA sequencing is the process of determining the order of nucleotides within a DNA molecule. It includes technology that is used to determine the order of the four bases adenine, guanine, cytosine, and thymine in DNA. The knowledge of DNA sequences has become vital for biological research, diagnostics, biotechnology, forensic biology, and biological systems. DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions, full chromosomes or entire genomes. Sequencing provides the order of individual nucleotides in DNA or RNA. This is useful for bioinformatics. This helps in design of drugs,to calculate the conformation of drug molecule with the protein molecule. To understand the important functions of body and how they react to external objects and external conditions. Genome sequence and powerful computational resources for gene finding has been redefined for computational problem by bioinformatics.

Genomics is a discipline in genetics that applies DNA sequencing methods and bioinformatics to sequence, assemble, and analyze the function and structure of genomes. The genome is the entire DNA content that is present within one cell of an organism. Genomic information is a very real sense; it is represented in the form of sequences of which each element can be one out of a finite number of entities. Rapid advances in human genome research make it almost certain that genomics will become more important for health improvement. A single strand of DNA is a biomolecule consisting of many linked, smaller components called nucleotides. Each nucleotide is one of four possible types adenine (**A**), cytosine (**C**), guanine (**G**), and thymine (**T**) and has two distinct ends, the 5′ end and the 3′ end, so that the 5′ end of a nucleotide is linked to the 3′ end of another nucleotide by a strong chemical bond . Therefore, each DNA single strand is mathematically represented by a character string.

5′ - C-A-T-T-G-C-C-A-G-T - 3′
3′ - G-T-A-A-C-G-G-T-C-A – 5’

In the process of DNA **replication** or **synthesis,** a single exact duplicate copy of a

genome is created, prior to the cell division. The two resulting double strands are identical, and each of them consists of one original and one newly synthesized strand. A cell mechanism recognizes the beginning of a gene. The DNA is then converted into **RNA**, the process is known as **transcription**. Many eukaryotic genes are composed of alternating segments called **introns** and **exons**. Only exons code for proteins

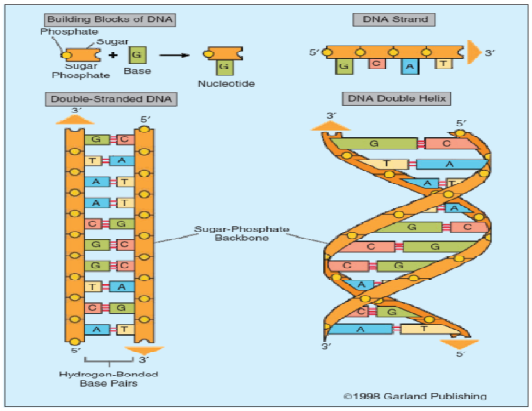

Figure 1.1 **DNA** and its building blocks

. Introns are removed in a process called **splicing**. Finally the translation process translates the mRNA codons into protein, a chain of amino acids. The complete process of moving from DNA to proteins is known as the central dogma of molecular biology. The observation that each gene is responsible for the creation of a protein is often expressed as

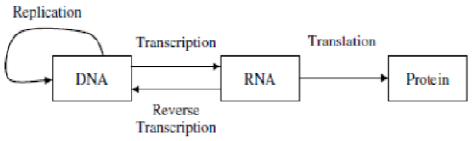

Figure 1.2: Central dogma of molecular biology

The spliced mRNA is divided into groups of three adjacent bases. Each triplet is called a **codon.** Evidently there are 64 possible codons. Thus the mRNA is nothing but a sequence of codons. Each codon instructs the cell machinery to synthesize an amino acid. The codon sequence therefore uniquely identifies an amino acid sequence which defines a protein. This mapping is called the genetic code. The **translation** from mRNA to protein is aided by adaptor molecules called the transfer RNA (tRNA) molecules. In some sense the tRNA molecules store the genetic code.

Most of the identified genomic data is publicly available over the Web at various places worldwide, one of which is the Entrez search and retrieval system of the National Center for Biotechnology Information (**NCBI**) at the National Institutes of Health (**NIH**). The NIH nucleotide sequence database is called **GenBank** and contains all publicly available DNA sequences.

The contributions of this work include the efficient and reliable technique for the analysis of biological sequences. Reads are generated by NGS as 16 bit data assembly. It is encrypted and produced same sequence of bits with different variables. Then, this encrypted data is fed as input to FFT for transformation. Finally, this complex values produce DNA sequence for the biological genome. In **Section II** the closely related prior work is discussed and **Section III** describes proposed techniques used in computation of DNA. **Section IV** deals with implementation and simulation Results. Finally, conclusion and Future work is mentioned in **Section V.**

	U	C	A	G
U	UUU = phe UUC = phe UUA = leu UUG = leu	UCU = ser UCC = ser UCA = ser UCG = ser	UAU = tyr UAC = tyr UAA = stop UAG = stop	UGU = cys UGC = cys UGA = stop UGG = trp
C	CUU = leu CUC = leu CUA = leu CUG = leu	CCU = pro CCC = pro CCA = pro CCG = pro	CAU = his CAC = his CAA = gln CAG = gln	CGU = arg CGC = arg CGA = arg CGG = arg
A	AUU = ile AUC = ile AUA = ile AUG = met	ACU = thr ACC = thr ACA = thr ACG = thr	AAU = asn AAC = asn AAA = lys AAG = lys	AGU = ser AGC = ser AGA = arg AGG = arg
G	GUU = val GUC = val GUA = val GUG = val	GCU = ala GCC = ala GCA = ala GCG = ala	GAU = asp GAC = asp GAA = glu GAG = glu	GGU = gly GGC = gly GGA = gly GGG = gly

Figure1.3: Genetic code Diagram

2. BACKGROUND

Several methods exist for the computation and development of genome assembly methods. Various aspects in genomics are briefly discussed with respect to the contributions made by various authors. B.Sharath Chandra Varma et al proposed HEBs to accelerate the genome assembly process using FPGAs. There is 20% increase in speedups with HEBs over FPGA implementation. The drawback of this is it is very complex process and takes more time to execute. A biological sample is biologically processed and given to a sequencing machine, which generates small fragments of 25, 300 base pairs called “reads”. These sequence fragments overlap each other and are a Part of longer genome sequence. These reads

have to be assembled to construct the genome. In this method further speedups by designing HEBS customized to accelerate de novo assembly using FPGAs. Velvet is the software used for performing genome assembly. Multi objective evolutionary optimization is used to discover optimal motifs in DNA sequences has been proposed. W. Zhang et al demonstrated the advent of next-generation sequencing technologies is accompanied with the development of many whole-genome sequence assembly methods .This provides the information of assembly accuracy and integrity which indicates that string-based assemblers, overlap-layout-consensus (OLC) assemblers are well-suited for very short reads and longer reads of small genomes respectively. Genome assemblers take a file of short sequence reads and a file of quality- values as input. Since the quality-value file for the high throughput short reads is usually highly memory-intensive. For the sake of computational memory saving and convenience of data inquiry, high-throughput short reads data is always initially formatted to specific data structure. Poor knowledge and less efficiency regarding applicability and less accuracy.

3. PROPOSED WORK

The theory and methods of Fast Fourier Transform are becoming increasingly important in molecular biology. New methods are being developed to analyze Deoxyribonucleic acid sequences. Specifically, FFT techniques are developed for the analysis of deoxyribonucleic acid. Reads which are parts of the genome sequence being assembled are. Reads of 16 bit data is given to generate NGS. Its output is mapped with transformation input. The fast fourier transform (FFT) is performed to produce DNA sequencing vectors. Thus, it produces 16 bit biological genome for their corresponding input.

In this work the DNA is implemented using the 16 bit FFT. Reads of 16 bit data is given to generate NGS. Its output is mapped with transformation input. The fast fourier transform (FFT) is performed to produce DNA sequencing vectors. Thus, it produces 16 bit biological genome for their corresponding input. The reads are generated by NGS as 16 bit data assembly. Its output is encrypted and produces same sequence of bits with different variables. Then, this encrypted data is fed as input to FFT for transformation. Finally, this complex values produce DNA sequence for the biological genome. Described through the block diagram as in fig.3.1

The LUT is given the 16 bit input as the next generation sequence. Then the 16 bit FFT sequence is computed. This 16 bit LUT and 16 bit FFT implements the output as DNA.

The implementation of DNA sequence through the flowchart representation as

The Techniques that are used for the computation of DNA by FFT are the Compression and Encryption. LUT is taken to reduce the memory space .We store all the data in look up table. To decrease the processing time for the functions which are complex we are creating the LUT. Then the Compression reducing time and delay. Finally encryption takes place to remove the unwanted errors and for achieving data security.

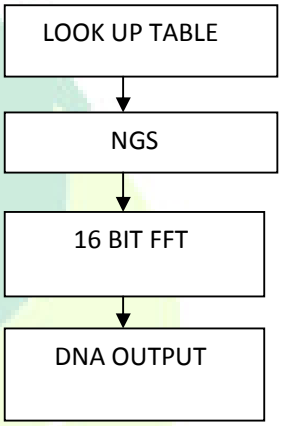

Figure 3.1: Block diagram

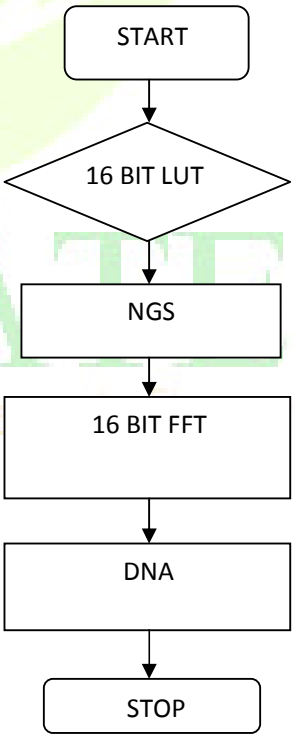

Figure 3.2: Flowchart of DNA Simulation

Fast Fourier transform (FFT) algorithm computes the discrete Fourier transform of a sequence, or it's inverse. Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The main function of FFT is to accomplish the task in more efficient manner. Through the use of techniques descripted above the computation of DNA takes place.DNA is calculated using 1D-FFT. FFT for protein through 3D-FFT.Single dimension fourier transform in each of 3D matrix calculate 3D-FFT. The DNA simulation results are seen in Section IV.

4. RESULTS

We have implemented on XILINX ISE using the Modelsim software. ModelSim may be used to perform the following types of simulations like Logical verification, to ensure the module produces expected results and Behavioral verification, to verify logical and timing. The hardware required for this software environment is Complex programmable logic device called XC9572XL. The hardware must support the software by having enough memory space in the installed drive. The XC9572XL is a 3.3V CPLD targeted for high-performance, low-voltage applications in leading-edge communications and computing systems. The simulation of DNA sequence is described as shown in fig 4.4 by 16 bit LUT Module that represents Next Generation Sequences in fig 4.2 and 16 bit FFT in fig 4.3

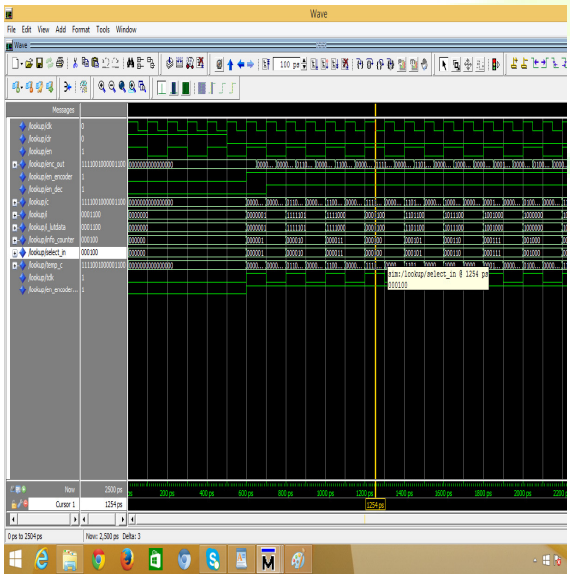

Figure 4.2: **16 Bit LUT** Module

5. CONCLUSION

The Advantages are analysis of organization of genes and genomes and their evolution, protein sequence can be predicted from DNA which facilitates protein properties, structure and function, identification of mutations that lead to disease. The Applications are Molecular Docking, Process in which two molecules fit together in 3D space. In design of drugs, this is used to calculate the conformation of drug molecule with the protein molecule. Molecular Dynamics Simulations , To understand the important functions of body and how they react to external objects and external conditions. The main aim of this paper is to compute the DNA through 16 bit FFT and 16 bit LUT. Time taken for computation is less. The structure of protein can also be determined. The future work includes the optimization of algorithm through DWT for further improvements in gene by incorporating matching of DNA sequences using cross correlation.

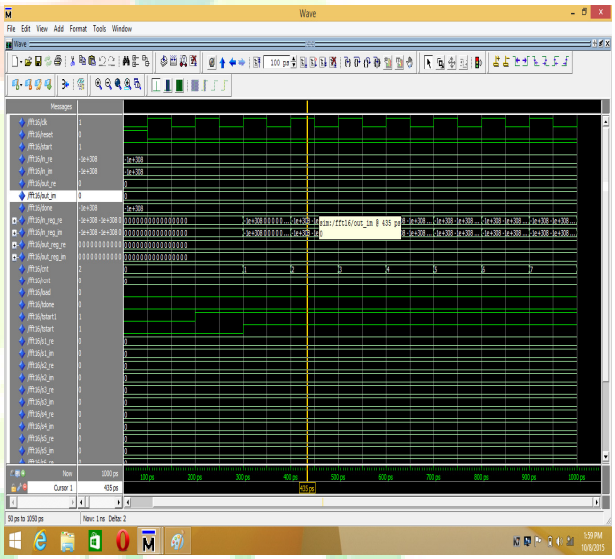

Figure 4.3: **16 Bit FFT** Module

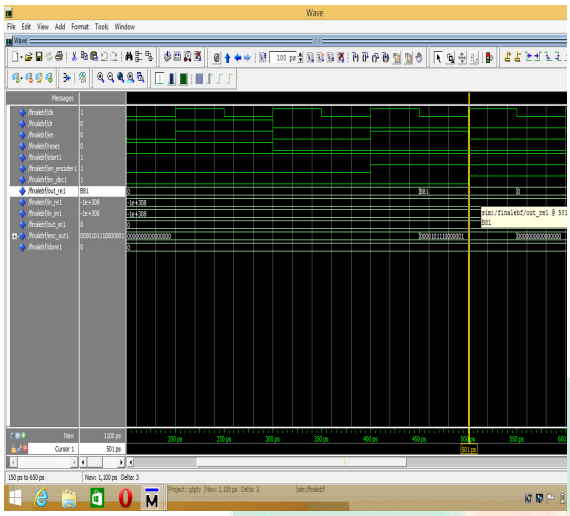

Figure 4.4: **DNA** Simulation Module

REFERENCES

[1] Baoshan Ma, Yi- Sheng Zhu (2007), "An Algorithm for Gene Prediction Based on the Z Curve," Dalian Maritime University, Dalian, China, Conference IEEE.

[2] Chen Y, Souaiaia T, and Chen T (2009), "PerM: Efficient Mapping of Short Sequencing Reads with Periodic full Sensitive Spaced Seeds," Bioinformatics, vol. 25, no. 19, pp. 2514–25.

[3] Chung- wei yeh and chih – ping Chu (2008), "Molecular Verification of Read Based systems Based on DNA Computation," IEEE Transactions on Knowledge and Engineering, Vol 20, No.7.

[4] DNA Microarray Repository: Kent Ridge Bio-Medical Dataset Repository, from the Agency for Science, Technology and Research http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.

[5] Gen Bank, Genetic sequence database. Available: www.ncbi.nlm.nih.gov/genbank.

[6] Jianobo Gao, Yinhe cao ,Yan Qi ,Jing Hu (2005), " Building Innovative Representations of DNA Sequences to facilitate Gene Finding ,"IEEE Intelligent systems.

[7] Kozakov D, Brenke R, Comeau S R, and Vajda S (2006), "PIPER: an FFT based Protein Docking Program with Pairwise Potentials." Proteins.

[8] Langmead B, Trapnell C, Pop M, and Salzberg (2009), "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," Genome Biology, vol. 10.

[9] Mingjum Zhang, Maggie X Cheng (2006) ," A Mathematical Formulation of DNA Computation ," IEEE Transaction on Nano BioScience.

[10] Olson C, Kim M, Clauson C, Kogon B, Ebeling C, Hauck S, and Ruzzo W (2012), "Hardware Acceleration of Short Read Mapping," in IEEE Symposium on FCCM, pp. 161 –168.