EXON PREDICTION IN GENOME WITH FREQ ANALYSIS ON FPGA

Ramu Boyapeti¹, Tarannum Sultan²
¹PG Scholar, ²Associate Professor, br.boyapati@gmail.com¹
Department of ECE, SSJ Engineering College, Hyderabad

Abstract: The emergence of a completely digitized version of the human genome has opened the door to a new way of discovering genes using computer algorithms which can alleviate some of the aforementioned issues. This is known as gene prediction. Human genes consist of numerous components including a promoter, splice sites, introns and exons. Exons, also known as coding regions, are areas where the nucleotide code for amino acids which are found at the base of proteins. The prediction of exons could therefore lead to the prediction of genes.

In this project, we propose a pipelined hardware acceleration module for the Goertzel algorithm in the context of exon detection. To perform frequency analysis in the context of exon prediction, the Goertzel algorithm is preferred to the FFT as there is only one frequency of interest. The Goertzel algorithm was designed to calculate the frequency content of a signal at the normalized frequency ωn. This is done in two steps: calculating the different values of y[n] and, from those values, calculating the power of the frequency component at ωn.

Modelsim Xilinx Edition (MXE) and Xilinx ISE will be used simulation and synthesis respectively. The Xilinx Chipscope tool will be used to test the FPGA inside results while the logic running on FPGA. The Xilinx Spartan 3E Family FPGA development board will be used for this project.

Keywords: MXE, FPGA, Goertzel algorithm, FFT

I. INTRODUCTION

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. DNA sequencing is the process used to map out the sequence of the nucleotides that comprise a strand of DNA. It includes any method or technology that is used to determine the order of the four bases Adenine, Guanine, Cytosine, and Thymine in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. These nucleotides also known as base pairs (bps), can be of four types: adenine (A), guanine (G), cytosine (C) and thymine (T). The human genome is divided into chromosomes each of which contains number genes. These genes serve a major role in protein synthesis which is a crucial process in the human body acids which are found at the base of proteins. The prediction of Exons could therefore lead to the prediction of genes. DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions, full chromosomes or entire genomes.

There are basic methods for generating for DNA sequences Maxim-Gilbert sequencing and Chain-termination methods. And also some advanced methods for Large-scale sequencing often aims at sequencing very long DNA pieces, such as whole chromosomes, although large-scale sequencing can also be used to generate very large numbers of short sequences, The process of gene discovery in a genetics laboratory is a very Long and expensive process.

While this previously proposed system was functional and provided good results, the algorithm required that the same data be sent multiple times in order to operate properly. This was a major drawback as it slowed down the processing speed by about thirty times, according to the tests performed in the laboratory. A pipelined hardware acceleration module for the Goertzel algorithm in the context of Exon detection. The proposed system has an architecture that allows for the data to be sent only once, thus accelerating the overall process.

A) Proposed Solution

- Doesn’t require multiple retransmissions.
- Effective implementation using Goertzel algorithm.
- Low area solution in comparison with conventional FFT approach.
- Pipelined Structure for Exon prediction.
B) DNA

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. DNA sequencing is the process used to map out the sequence of the nucleotides that comprise a strand of DNA. It includes any method or technology that is used to determine the order of the four bases Adenine, Guanine, Cytosine, and Thymine in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. These nucleotides also known as base pairs (bps) can be of four types: adenine (A), guanine (G), cytosine (C) and thymine (T). The human genome is divided into chromosomes each of which contains number genes [1].

There are basic methods for generating for DNA sequences Maxims-Gilbert sequencing and Chain-termination methods. And also some advanced methods for Large-scale sequencing often aims at sequencing very long DNA pieces, such as whole chromosomes, although large-scale sequencing can also be used to generate very large numbers of short sequences.

The process of gene discovery in a genetics laboratory is a very Long and expensive process. The emergence of a completely digitized version of the human genome has opened the door to a new way of discovering genes using computer algorithms which can alleviate some of the aforementioned issues: this is known as gene prediction.

2. GOERTZEL ALGORITHM

The article deals with the Goertzel algorithm, used to establish the modulus and phase of harmonic components of a signal. The advantages of the Goertzel approach over the DFT and the FFT in cases of a few harmonics of interest are highlighted. In the case of discrete-time signals, the discrete Fourier transform (DFT) is widely used for spectral analysis. The frequencies of the harmonics in the DFT always depend on the length of the transform, N, and they are integer multiples of the fundamental frequency Δf = f/s N, where f represents the sampling frequency. Thus, Δf gives the frequency resolution of the DFT. In the case, when the transform length N is not a multiple of the signal period, the signal is a sum of harmonic components whose frequencies are not integral multiples of the fundamental frequency.

2.1. Computational and Memory Complexities

The FFT algorithm used with N being a power of two has computational demands proportional to N log₂ N, the absolute number depends on the particular implementation. Usually the number of real-number operations found in the literature is approximately 6N log₂ N (taking one complex multiplication as a combination of four multiplications and two summations). When working with real signals, a number of operations can be avoided; however, it is at the cost of increased complexity of the algorithm, and, it is not true that the demands can be reduced by half, as can be read, for example in . The Goertzel algorithm is implemented in the form of a 2nd-order IIR filter as shown in Figure 2.3.1(a). This filter computes a single DFT output (the kth bin of an N-point DFT) defined by

\[ X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn}/N. \]

Where time index n is an integer in the range 0 ≤ n ≤ N−1. The frequency-domain index k is also an integer in the range 0 ≤ k ≤ N−1.

The Goertzel filter’s z-domain transfer function is

\[ H(z) = \frac{1 - e^{-j2\pi k/N}}{1 - 2 \cos(2\pi k/N)z^{-1} + z^{-2}}. \]

With a single z-domain zero located at \( z = e^{-j2\pi k/N} \) and conjugate poles at \( z = e^{\pm j2\pi k/N} \) as shown in Figure 1(b). The pole/zero pair at \( z = e^{-j2\pi k/N} \) cancel each other.

Fig 2.1. Stable (a) Filter Implementation and (b) Z plane poles and zero

The advantages of the Goertzel algorithm in above Figure (a) are that it is guaranteed-stable, and its exhibits no output error due to a pole or zero being slightly inside the z-plane unit circle.
Now the perceptive reader would factor Equation (4) as

\[ H_{eq,final}(z) = \frac{1 - z^{-1}}{(1 - z^{-2}) (1 - z^{-3})} = \frac{1}{1 - z^{-2}} \]

and redraw above Figure (a) as shown in below Figure (a).

3. FPGA AND TOOL FLOW

3.1 FPGA Design Considerations

FPGA demonstrates good performance and logic capacity by exploiting parallelism. At present single FPGA platform can play multi-functions, including control, filter and system. FPGA design flow is a three-step process consisting of design entry, implementation, and verification stages, as shown in Fig 3.1. The full design flow is an iterative process of entering, implementing, and verifying the design until it is correct and complete.

VHDL

VHDL (VHSIC hardware description language; VHSIC: very-high-speed integrated circuit) is a hardware description language used in electronic design automation to describe digital and mixed-signal systems such as field-programmable gate arrays and integrated circuits. The structural and dataflow descriptions show a concurrent behavior.

That is, all statements are executed concurrently, and the order of the statements is not relevant. On the other hand, behavioral descriptions are executed sequentially in processes, procedures and functions in VHDL. The behavioral descriptions resemble high-level programming languages.

VHDL allows a mixture of various levels of design entry abstraction. Precision RTL Synthesis Synthesizes will accept all levels of abstraction, and minimize the amount of logic needed, resulting in a final net list description in the technology of your choice. Top-Down Design Flow is shown in below Figure.
Like MPGAs, FPGAs comprise an array of uncommitted circuit elements, called logic blocks, and interconnect resources, but FPGA configuration is performed through programming by the end user. An illustration of a typical FPGA architecture appears in Figure 3.4. As the only type of FPD that supports very high logic capacity, FPGAs have been responsible for a major shift in the way digital circuits are designed.

**Fig 3.2 Structure of a PAL**

Field-Programmable Gate Arrays (FPGAs).

**Fig 3.3 Structure of an FPGA**

**3.3 FPGA Design Flow**

**Fig 3.6 Xilinx Design Flow**

**3.4 Model Sim Tool Flow**

**Basic Simulation Flow**

The following diagram shows the basic steps for simulating a design in ModelSim.

**Chip Scope**

Chip scope is an embedded, software based logic analyzer. By inserting an “integrated controller core” (icon) and an “integrated logic analyzer” (ILA) into your design and connecting them properly, you can monitor any or all of the signals in your design. Chip scope provides you with a convenient software based interface for controlling the “integrated logic
4. IMPLEMENTATION

4.1. Implementation of Exon Prediction
In this project I am going to implement Goertzel algorithm in context of predicting an Exon. Exons are the coding parts of human genomes which play a vital role to determine the amino acids that are begin produced. An Exon is any nucleotide sequence encoded by a gene that remains present within the final mature RNA product of that gene after Introns have been removed by RNA splicing. In many genes, each of the Exons contain part of the open reading frame (ORF) that codes for a specific portion of the complete protein.

The code defines how sequences of these nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis. With some exceptions, a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid. Because the vast majority of genes are encoded with exactly the same code (see the RNA codon table), this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact some variant codes have evolved. For example, protein synthesis in human mitochondria relies on a genetic code that differs from the standard genetic code.

5. EXPERIMENTAL RESULTS

5.1. Simulation Result

Adder

5.2. Simulation Result

Multiplier:

Fig.5.2 simulation Result Fixed Point Multiplier

Pipo Register:

Fig 5.3 Simulation Result for 8-Bit PIPO (Parallel in Parallel out Shift Register)

6. ADVANTAGES AND APPLICATIONS:

6.1 Advantages:
- We proposed a pipelined hardware accelerator for Exon prediction.
- The current design does not require information to be sent recurrently by keeping them in memory and by shifting them.
- Low area solution in comparison with conventional FFT approach.

6.2. Applications:
It finds its applications in Genome prediction.
Exon prediction have wide applications in the field of Bioinformatics

CONCLUSION
In this paper, I was proposed a pipelined hardware accelerator For Exon prediction. Unlike the previous system, the current Design does not require information to be sent recurrently by keeping them in memory and by shifting them. Using cyclical Data shifting, it is possible to generate a new output on every Clock cycle once the pipeline is full.

While the chosen Platform only operates at a moderate speed, we were able to Show that it is possible to perform frequency analysis on a Window within a single cycle which is roughly thirty times faster than our previous solution. This is significant as gene Prediction necessitates the operation of a large amount of data and this type of accelerated analysis could facilitate gene Discovery

FUTURE SCOPE
The process of gene discovery implemented till now using the Xilink and MATLAB Tools; it can also implement in CAD-Tool to get accurate result. In this way if we provide digital implementation of exon prediction, we can analyze the human genome construction with minutes rather than weeks carried out in laboratories today.

REFERENCES

Author Biography
Ramu Boyapati, PG Scholar in SSJ Engineering College, Hyderabad. He has completed his B.Tech from JNTU Hyderabad. His Area of interest is VLSI, Emedded Systems.

Tarannum Sultan, Associate Professor SSJ Engineering College. M.Tech and B.Tech from JNTU Hyderabad. She has published 1 International conference and journals. Her Area of interest is VLSI and Embedded Systems and Artificial Intelligence. She has more than 8 years of teaching experience. She is a member of ISTE, and IACSIT.