Hardware Efficient Exon Prediction in Genome with Frequency Analysis on FPGA

G.VENKATA SUBBARAO1, D.RAMADEVI2
1Research Scholar, ECE Dept, Teegala Krishna Reddy Engineering College, Hyderabad, AP-INDIA. E-mail: venkatsubbu.401@gmail.com,
2Asst Prof, ECE Dept, Teegala Krishna Reddy Engineering College, Hyderabad, AP-INDIA.

Abstract: In this paper, we propose a novel architecture for exon prediction using Goertzel algorithm, the major outcome of this project has been to produce a digitized version of the human genome which consists of a long string of nucleotides (approximately six billion). These nucleotides, also known as base pairs (bps), can be of four types: adenine (A), guanine (G), cytosine (C) and thymine (T). 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. The analysis of an entire human genome is time-consuming so an FPGA acceleration module is used for Exon detection. Hence efficient Goertzel algorithm is implemented to compute the energy at fs/3 and further Exon prediction. Modelsim Xilinx Edition (MXE) and Xilinx ISE will be used for simulation and synthesis respectively. The Xilinx Chip scope tool will be used to test the FPGA inside results while the logic running on FPGA. The Xilinx Spartan 3 Family FPGA development board will be used this project.

Keywords: Carry Skip Adder, RCA, BEC, Power Consumption.

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[1]. DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions, full chromosomes or entire genomes. Depending on the methods used, sequencing may provide the order of nucleotides in DNA or RNA isolated from cells of animals, plants, bacteria, archaic, or virtually any other source of genetic information. The resulting sequences may be used by researchers in molecular biology or genetics to further scientific progress or may be used by medical personnel to make treatment decisions or aid in genetic counseling. Human genome is divided in to chromosomes each of which contains number of Genes. Which consist of Introns and Exons? Exons also know as coding regions. Within a coding region, nucleotides are combined in groups of three and the resulting structure is called codon.

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. 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.

Several techniques have been proposed for the detection of coding regions using digital signal processing (DSP). These techniques include asymmetric codon distribution, frequency analysis, hexamer counting, and autoregressive modeling [2]. Frequency analysis which is known as a technique that provides good results [3]. The use of frequency analysis in Exon prediction has been thoroughly explored in literature [4]. The most popular method is the fast Fourier transform (FFT). However, it has been shown that the Goertzel algorithm is much more computationally efficient when the number of frequencies to be analyzed is small [5]. While the Goertzel algorithm is well known and is easily implementable in software, the analysis of an entire human genome is time consuming as it contains six billion bps. 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.

II. DNA CODON

DNA Codon table due to the biochemical nature of the protein translation process. However, with the rise of computational biology and genomics, proteins have become increasingly studied at a genomic level. As a result, the practice of representing the genetic code as a DNA codon table has become more popular. The fig.1. table shows the 20-Amino acids, their and their corresponding DNA codons. The human genome can be regarded as a string of six billion characters each of which can be A, T, G or C.

Within a coding region, nucleotides are combined in groups of three and the resulting structure is called a codon. Each codon codes for a different amino acid. Knowing that there are 64 different codons and that there are only 20 different amino acids, they are:


It can be shown that different codons will code for the same amino acid. For a given amino acid, certain codons will be more commonly found than others. This creates a non-random nucleotide distribution within coding regions. In fact, it has been well documented that coding regions exhibit a periodicity of three. That is, when frequency analysis is performed on such a region, the frequency spectrum has a more pronounced peak at fs/3, where fs is the sampling frequency (or equivalently, number of samples). It is therefore possible, in many cases, to identify coding regions using frequency analysis and examining the frequency component at fs/3. Frequency analysis on a sequence of nucleotides, it is first required to transform it into a form that can be processed. Separating the nucleotide sequence into four binary vectors, each representing the presence of a given type of nucleotide [7]e. For instance, the vector A would contain ‘1’ whenever the given nucleotide is an adenine otherwise it will be ‘0’.

ATGCAATTG GCC  
Vector A 100011000000 
Vector B 010000110000 
Vector C 001000011000 
Vector D 000100000011 

![Fig.2. Example of nucleotide sequence to vector conversion](image)

After separating the nucleotide sequence into four vectors, frequency analysis is performed on each vector separately to find the frequency content at fs/3. That frequency component is then squared and added to the squared frequency component of the other nucleotides. That is P=P_A+P_T+P_G+P_C.

III.GOERTZEL ALGORITHM

The Goertzel algorithm is a digital signal processing (DSP) technique that provides a means for efficient evaluation of individual terms of the Discrete Fourier Transform (DFT). The Goertzel Algorithm analyses one selectable frequency component from a discrete signal. The Goertzel algorithm is implemented in the form of a 2nd-order IIR filter as shown in

---

Fig.1. Codon Table.

International Journal of Scientific Engineering and Technology Research  
Volume. 02, Issue No. 08, August-2013, Pages: 700-708
Hardware Efficient Exon Prediction in Genome with Frequency Analysis on FPGA

International Journal of Scientific Engineering and Technology Research
Volume. 02,IssueNo.08, August-2013, Pages:700-708

Figure 1(a). This filter computes a single DFT output (the \( k \)th bit of an \( N \)-point DFT) defined by

\[
X(k) = \sum_{n=0}^{N-1} x(n)e^{-2\pi i nk/N}.
\]

(1)

Where time index \( n \) is an integer in the range \( 0 \leq n \leq N-1 \). The frequency-domain index \( k \) is also an integer in the range \( 0 \leq k \leq N-1 \). The filter’s \( y(n) \) output is equal to the DFT output frequency coefficient \( X(k) \) at the time index \( n = N \). We must be mindful that the filter’s \( y(n) \) output is not equal to \( X(k) \) at any time index when \( n \neq N \).

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 \( \omega_n \). To perform frequency analysis in the context of Exon prediction, this can be done in two steps:

1. Calculating the different values of \( y[n] \).
2. From those, calculating the power of the frequency component at \( \omega_n \). The calculation of the different \( y[n] \) values is done using the following recursive equation.

\[
y[n] = x[n] + \cos (2 \pi \omega_n) y[n-1] - y[n-2]
\]

The power of the chosen frequency component (\( \omega_n \)) can be calculated using the following equation:

\[
P(N) = y[N]^2 + y[N-1]^2 - \cos (2 \pi \omega_n) y[N]y[N-1].
\]

For Exon prediction, the normalized frequency of interest is \( \omega_n = 1/3 \). When substituted in the above equation becomes

\[
\]

This equation \( Y[n] = x[n] - y[n-1] - y[n-2] \) which shows the Block diagram of Goertzel algorithm.

Fig.3. Block Diagram of the Goertzel algorithm for \( fs/3 \).

Similarly, the calculation of the frequency power for the equation of \( P(N) = y[N]^2 + y[N-1]^2 - \cos (2 \pi \omega_n) y[N]y[N-1] \) is also simplified:

\[
P(N) = y[N]^2 + y[N-1]^2 + y[N-1]
\]

The power spectrum can be calculated with three multipliers.

IV.IMPLEMENTATION

In this project I am going to implement Goertzel algorithm in context of predicting an Exon. Exons are the coding parts 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.

In my project I am going to implement the nucleotide input sequences using two dimensional ROM storing different Nucleotide samples. is going to store 4 Nucleotides namely A, C, T, G. we have chosen 512 samples for each Nucleotide. Whenever there is an input nucleotide corresponding entity is filled with one and remaining with zeros. This process of separating the nucleotide sequences is called as splitting as shown in fig 4. Thus we will get 4 columns for each nucleotide vector. Each of these vectors is sent to the Goertzel block for Frequency analysis. Goertzel algorithm is used for frequency analysis at a particular frequency component. I am going to perform frequency analysis at Fs/3 frequency component. Then the output of this Goertzel block is sent to Magnitude calculator block which calculates the Magnitude of each Goertzel block output.

Fig.4 Proposed method for creating the input vectors.
Then the power is calculated for each corresponding magnitude by squaring each of the output. All the power components for A, C, T, G vector are then added up to get the final power calculated for the total sequences the above total process is explained in fig.6. Thus this forms the Output for Final Exon prediction An Exon is said to be predicted whenever there is a peak at a particular given input samples. We can also find out the Amino acid with which nucleotide sequence it has been formed. The sequence is split into nucleotide vectors that are to be processed separately. The boxes entitled Frequency Analysis and Power of 2 are implemented as a single step using Goertzel algorithm as shown in fig.5

Then By using The GOERTZEL’S algorithm I am finding the power of four sequences of nucleotides, using frequency analysis in Exon detection by Goertzel algorithm.

V.EXPERIMENTAL RESULTS

The different modules which are used in the project:

Fig.5 Block diagram of the FPGA implementation.

Fig.6.internal part of FPGA implementation.

Fig.7.simulation result of Adder.
Hardware Efficient Exon Prediction in Genome with Frequency Analysis on FPGA

Fig. 8. Simulation result of multiplier

Fig. 9. Simulation result of Pipo shift register
G. VENKATA SUBBARAO, D. RAMADEVI

Fig. 10. Simulation result of counter

Fig. 11. Simulation result of top module Exon prediction.
Fig 11 shows simulation result of top module Exon prediction, if we observe carefully there are some peaks are appearing, means codon was detected. According to that codon we can find out corresponding amino acid present at protein base. For example I was got first three peaks correspond to ATG sequence of nucleotide and its corresponding amino acid is methionine (Met) and second three peaks for TGC sequence of nucleotide sequence and their corresponding amino acid is Cysteine (Cys) from the codon table of fig.1.

VI. CHIPSCOPE RESULTS

Fig 12 shows chip scope output at one particular instant.

Fig 13 chip scope output at second instant.
Fig. 14. Chip scope output at third instant

Xilinx Chip scope was used for viewing results from chip. From above three figures of chip scope output if we observe the peaks are different compared with the other one, when a run button was pressed in chip scope tool at particular instant of samples of nucleotide sequences are running.

VII. TOOLS and HARDWARE

- Simulation software - Modelsim Xilinx Edition (MXE)
- Synthesis, P&R - Xilinx ISE
- On chip verification - Xilinx Chip scope
- Hardware kit - Xilinx Spartan 3 Family FPGA board

- The work includes VHDL modeling of various control blocks and sub blocks.
- All the modules will be simulated by Modelsim for functional verification.
- Xilinx ISE will be used for FPGA synthesis, Place & Routing and timing analysis.
- Xilinx Chip scope will be used for viewing results from chip.

VIII. ADVANTAGES AND APPLICATIONS

- 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.
- It finds its applications in Genome prediction.
- Exon prediction have wide applications in the field of Bioinformatics.

IX. 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.
X. REFERENCES


Author’s Profile:

G.Venkata Subbarao (M.Tech),
Teegala Krishna Reddy Engineering College, Hyderabad, AP-INDIA.

D.Ramadevi, Asst.Prof M.E,
Teegala Krishna Reddy Engineering College, Hyderabad, AP-INDIA.