Design and Analysis of Multiple Paths Trace Back and Reconstruction Module for DNA Sequence Alignment Accelerator using ASIC Design Flow

Nurul Ain Bt Husaini, Siti Lailatul Mohd Hassan and Syazilawati Mohamed
Universiti Teknologi MARA,
40450, Shah Alam, Selangor, Malaysia.
nurulainhusaini@gmail.com

Abstract - Bioinformatics is the analysis of biological information using computers and statistical techniques. Smith Waterman (S-W) algorithm for sequence alignment is one of the main tools of bioinformatics. It is used for searches and alignment of similarity sequence. This paper presents a novel approach and Analysis of Multiple Paths Trace Back and Reconstructions Module for DNA sequence alignment accelerator using ASIC design flow. The first objective is to construct the trace back and reconstruction module of the S-W algorithm with the multiple blocks and the functionality for each block. Second objective is to perform the timing analysis and third objective to implement the design using ASIC flow. The design was developed in VerilogHDL coding, simulated and synthesized using Xilinx ISE 12 and then re-implemented using Synopsys ASIC Tools implies the timing diagram and analyzes using the Design Compiler and Integrated circuit compiler to produce the layout. Resulted from Xilinx simulator and VCS expressed the output produced in single clock cycle for each blocks. As the conclusion the design is actually fully function for each block of Trace Back and Reconstruction.

Keywords - Bioinformatics, Sequence Alignment, Smith-Waterman (SW) Algorithm, trace back, reconstruction.

I. INTRODUCTION

All the cells of an organism consist of some kind of genetic information and carried by a chemical known as the deoxyribonucleic acid (DNA) in the nucleus of the cell [1,2,3]. DNA is stored as a code made up of four chemical bases, which are adenine (A), guanine (G), cytosine (C) and thymine (T). Variations in our DNA sequences give us individual fingerprints, useful for identification and for the establishment of relationships. The use of DNA analysis as evidence in criminal trials is now well-established [4].

Genetics databases hold extremely large amounts of raw data throughout the years. A human genome contains approximately 3 billion DNA base pairs [4]. Bioinformatics is become most important field of research. In GenBank (National Human Genome Research Institute (NIH) genetic sequence database) an annotated collection of all publicly available DNA sequences. The Genome is doubling every six months. Referred by NIH it contributes the general problem statement, for example increasing the biological sequence has reduces the efficiency of general purposes microprocessor based on software analysis approach [5].

Cluster technique is one of the techniques for running sequencing analysis on a general purpose microprocessor, but it required more than one general purpose microprocessor [2]. By improving the cluster technique, the new parallelism with divide and conquer technique for Field Programmable Logic Array (FPGA) implementation was introduced [2].

The new parallelism will able to reduce the complexity of the Smith Waterman algorithm and improve the sensitivity and minimize the cost of the general purpose of microprocessor. In addition, the potential of massive parallelism existing in this particular application method can achieve very high utilization of the parallelism and obtain great performance gain [3]. By using the parallelism it can give an advantage of the system for example improve the performance of the system because it can run simultaneously.

The methods of pair wise sequence alignment have two types global and local method [6]. The Global method consists of dotplot method as visual approach and Needleman-Wuncsh method. Hence, the local method is consisting of Smith Waterman, (Fast Alignment Search Tools-All) FASTA and (Basic Local Alignment Search Tool) BLAST [6].Global methods attempt to match as many characters as possible, from end to end, whereas a local method focuses on regions of similarity in parts of the sequence only [7]. On the other hand, these algorithms are time consuming and memory intensive. To improve the time consuming dynamic programming is reliable for DNA sequencing alignment. In 1970, Needleman and Wunsch and Sellers [2] proposed alignment based on dynamic programming. Later, Temple Smith and Mike Waterman introduced the Smith-Waterman algorithm which is a modification to Needleman and Wunsch that uses local alignment in 1981[2].

Smith Waterman Algorithm has widely been referred in the domain of Bioinformatics for sequence matching to detect such similarities for local alignment between two DNA [1]. Moreover, SW is implements in this project and the significance of study is to optimize the performance of the trace back and reconstruction module. Besides, improve the performance DNA sequence alignment, this design can make the comparison of the highest score, the optimum path and produced new sample and new target DNA sequence for more sensitivity.

In this paper, the conventional Smith Waterman algorithms are introduced in Part A. The proposed technique and block diagrams of the system are shown in Section II. The results are discussed in Section III and concluded in Section V.

A. SMITH WATERMAN ALGORITHM

The Smith-Waterman algorithm is based on dynamic programming. It is versatile and well known for high sensitivity. Moreover, The Smith-Waterman algorithm is used to search for homology by comparing two sequences. The sequences are compared using local alignments and total number of alignments can be considerable and identification. The best alignments are importance which is the reliability and relevance of the data obtained. In SW algorithm there are three steps involved which are matrix filling, find the maximum value (score) in the matrix and trace back the path that leads to the maximal score to find the optimal local alignment [7].

		A	**T**	**C**	**T**	**C**	**G**	**T**	**A**	**T**
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	**0**	0	0	2	1	0	0
T	0	0	2	1	**2**	1	1	4	3	2
C	0	0	1	4	3	**4**	**3**	3	3	2
T	0	0	2	3	6	5	4	**5**	4	5
A	0	2	2	2	5	5	4	4	**7**	6
T	0	1	4	3	4	4	4	6	5	**9**
C	0	0	3	6	5	6	5	5	5	8
A	0	2	2	5	5	5	5	4	7	7
C	0	1	1	4	4	7	6	5	6	6

Table 1 : Alignment of 9 based-pair DNA sequence using Smith Waterman algorithm

Table 1 [3] above demonstrates the alignment of DNA sequences by using Smith Waterman algorithm. The matrix filled using the match, mismatch and penalty. It shows the trace back process which is in the grey colored. It selects the highest score in the matrix filling as the starting point up and the movement will only going up or left until meet the first of zero. The priority of the movement is in diagonal because it shows that both sequences are met. If the sequence is going up or left it is indicates gap.

II. METHODOLOGY

A. Design Specification

As mentioned in introduction the algorithm consists of three steps. In this research is concentrated to design the Trace Back and Reconstruction module with optimize the functionality using Smith Waterman Algorithm. Firstly, competent to coded the design using the Verilog language to meet the specification. The designed is simulated and synthesized using the Xilinx ISE 12 tools and Synopsys ASIC design tools. It used to analyze the output to produce with 1clock cycle and the functionality of each blocks. Furthermore, the design is implementing in Synopsis ASIC design flow.

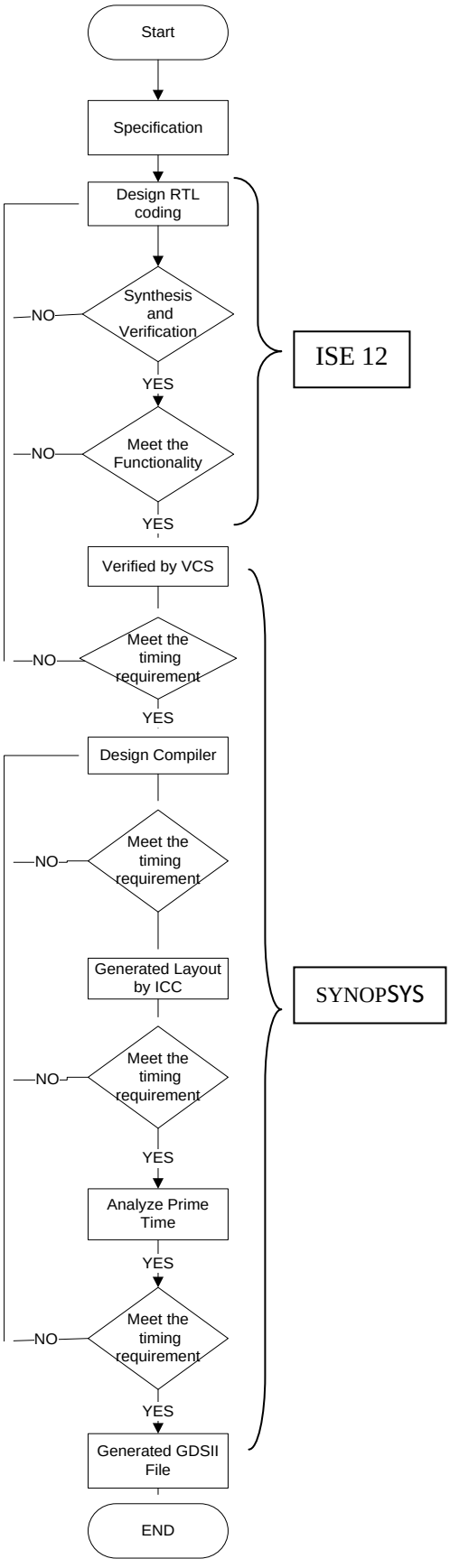


Figure 1 : Flow chart of Synopsys ASIC design tools

Figure 1 presents a simplified flow of using Xilinx ISE 12 and Synopsys ASIC design tools. All the design needs to meet the specification and then synthesized and simulated completely using Xilinx ISE12 tools. Then, the design is re-implemented using Synopsys ASIC design tools. Firstly, RTL coding was written in verilog using Xilinx. After completed, synthesize and verified the coding either meets the requirement from Xilinx.

Upon completion, the coding was verified and synthesizes using Xilinx. Next, the designed was synthesizing using VCS. VCS compiles the verilog source and links the object files to simulation engine to timing waveform. Furthermore, the design constraints such as clock period, input delay, output delay, uncertainty were setup before the design was synthesize by DC. In Design Compiler, DC performs the timing analysis and optimizes the result. All the result is report of setup time and hold time, dynamic and leakage power for the design can be obtained.

After synthesize process, the VLSI layout of the design ware floor planned, performed cell placement and routed using Integrated Circuit Compiler, ICC. After completing the ICC, the timing analysis and post layout netlist of the design is produce by Prime Time, PT to make sure the design have no violation of setup time and hole time. Finally the GDSII file for the front end design is generated through the ICC tool.

After completing the design, DNA nucleotides were assigned in three bit size to DNA sequence character. It is expressed in Table 2 below. It is implementing in Test Bench.

Name	Character	Data
Adenine	**a**	**000**
Cytosine	**c**	**001**
Guanine	**g**	**010**
Thymine	**t**	**100**
Gap	**-**	**101**

Table 2: DNA Sequence Character with Reduction Data Assignment

B. Architectural Design

At the beginning of this project, the design of trace back for z1 until z16 is construct in one module. The design did not give the output due to the complexity of the programming design. The design was reconstruct and divided into sub-modules of each trace back and reconstruction blocks. The design is constructing into small partition to achieve the functionality of each block of Trace Back and Reconstruction Blocks.

The revised design consists of three main modules. There are Comparator1 module, Trace Back and Reconstruction module and Comparator2 module as shown in Figure 2. The new design is make debugging process easier and faster if any error occurs.

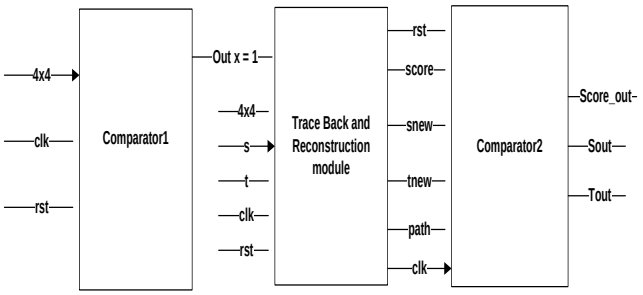


Figure 2: Block diagram of Top Module

The actual architecture of Comparator1 is shown as below in Figure 3. The module consists of input value (score) of matrixes 4x4 (z1-z16), clock and reset. The outputs of the comparator are out x, (x equals to 1 to 16). The first module, Comparator1 capable compares the highest score in 4x4 matrixes. The out x behave as enable, then it will sent enable to the sub-modules of Trace back and reconstruction will appear in z1 until z16.

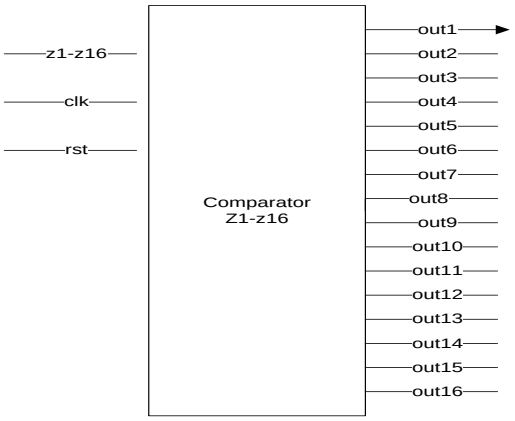


Figure 3 : Comparator 1

Furthermore, the highest value carries out to the next module Trace Back and Reconstruction from the Comparator1. Figure 4 illustrates The Trace Back and Reconstruction which consists of 16 sub-modules (z1-z16). Based on the figure 3, the input of each sub-module consists of s, t, rst, clk, value z1-z16 and out1 until out16 for module z1 until z16 respectively. In addition, the output of each module consists of score x, snew x, tnew x, and path x instantiate the each sub-module of z1 until z16 respectively. Next, the out x will enable the sub-module with the highest score of z value. Moreover inside the sub-module will calculate the score for all possible paths. All calculated scores will then representing the optimal path and will make a reconstruction and come out with total score, snew, tnew and path.

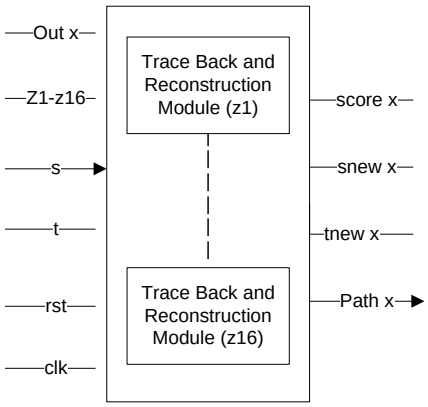


Figure 4: The Trace Back and Reconstruction consist of 16 sub-modules (z1-z16).

Figure 5 demonstrate the Comparator2 that may have the all output of previous modules. Moreover, all the output of previous modules takes as the input of comparator2. This comparator will search for the optimal path and will produce in output of comparator 2 known as Score_out the total score of the optimal path and reconstructed the new of sample/target with gaps of DNA sequences and known as Sout and Tout.

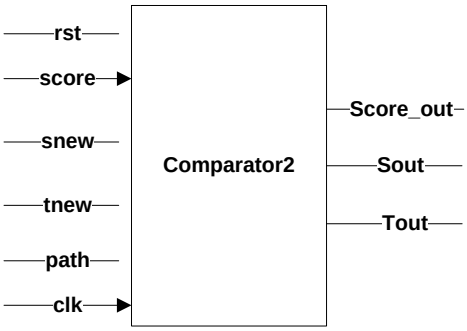


Figure 5 : Comparator 2

III. Results And Discussions

A. *Synthesis Blocks*

Figure 6 shows the Top Module of RTL schematic diagram which is synthesize from the Xilinx ISE. Hence the Figure 7 and 8 shows the sub-module inside the Top Module is synthesize by using Xilinx ISE 12 and VCS respectively. The top modules consist of 18 blocks which are comparator 1, 16 blocks of Trace Back and Reconstruction of z1 until z16 and comparator 2. The design is huge because consists a lot of comparator. By using the comparator the design is actually have more sensitivity. Moreover, by partitioning the design may debug of each the functionality easier.

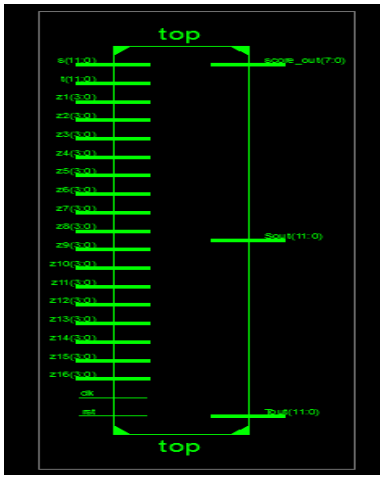


Figure 6: Block Diagram of Top Module

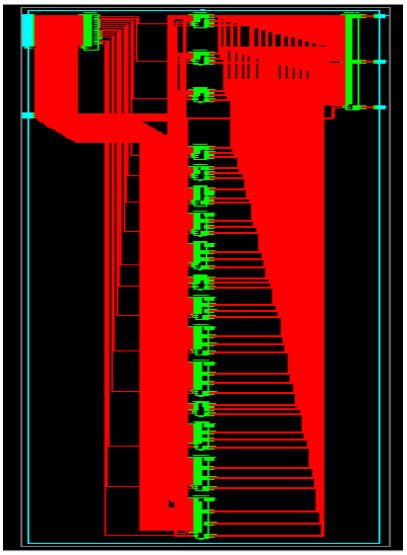


Figure 7 : RTL Schematics Diagram from Xilinx ISE 12 inside the Top Module

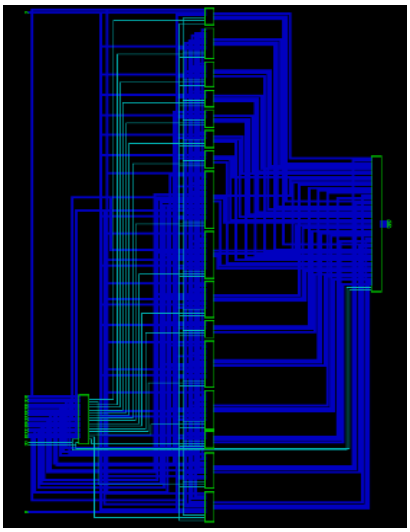


Figure 8: RTL Schematic Diagram from VCS

B. Verification Design

To ensure the design achieved the functionality each block needs to be test. The actual input of the DNA sequence is inserting to produce the accurate result. The timing diagrams are shown the functionality of the block. The output of block z1 until z16 will produce after one clock cycle.

Figure 9 : Timing diagram of comparator 1 form Isim simulator

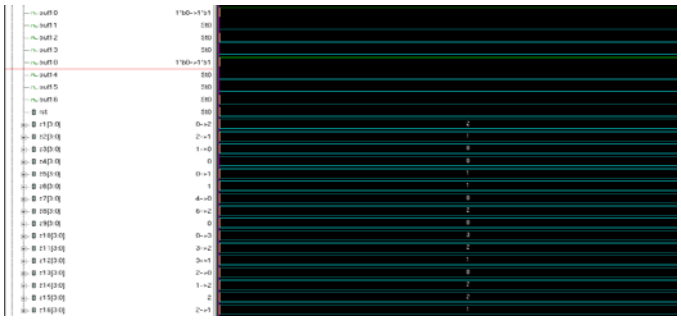


Figure 10: Timing Diagram of Comparator 1 using VCS

Figure 9 and 10 above, demonstrated the timing diagram of Comparator1. The inputs of z1 until z16 are entering to Comparator1. The highest value of z1 until z16 will be chosen. Then, the output from comparator1 will be the input for the Trace back and Reconstruction block.

Input		Expected output
z1	**2**	
z2	**1**	**Sout = 12'h54**
z3	**0**	**12'b000_001_010_100**
z4	**0**	
z5	**1**	**Tout = 12'h54**
z6	**4**	**12'b000_001_010_100**
z7	**3**	
z8	**2**	
z9	**0**	**Score_out = 20**
z10	**3**	
z11	**6**	
z12	**5**	
z13	**0**	
z14	**2**	
z15	**5**	
z16	**8**	
S 12'b000_001_010_100		
T=12'b000_001_010_100		

Table 3: Expected output of Block z16

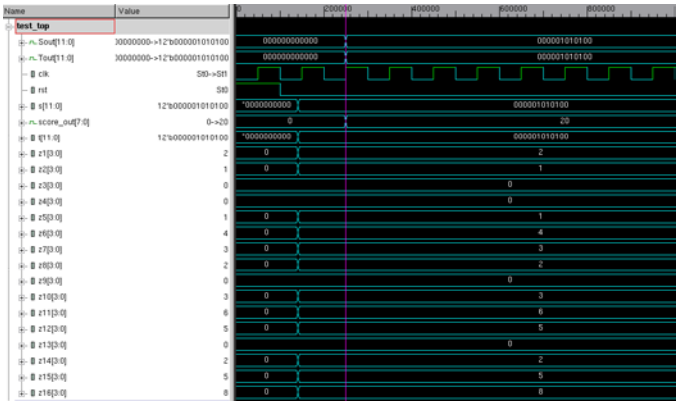


Figure 11: Timing Diagram of Block Trace and Reconstruction of z16

Table 3 presented the input of z1 until z16, s, t and expected output. The output from the table can be illustrated in the Figure 11. By verifying using the VCS shown that the output of the timing diagram will be given the same output as an expected output. It shows that the Block z16 is fully functional. As the previous discussion, each value has been interpreted in binary as in Table 2.

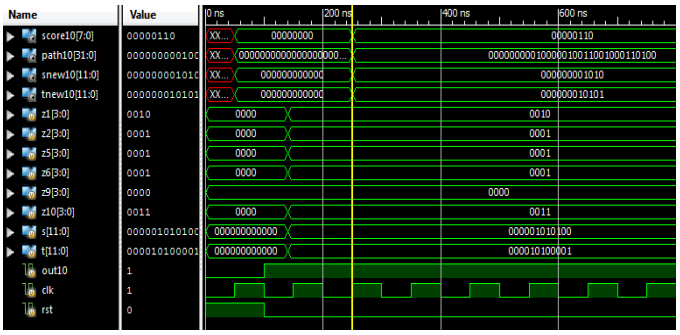


Figure 12 : Timing diagram of block Trace Back and Reconstruction of z10 from Isim Simulator

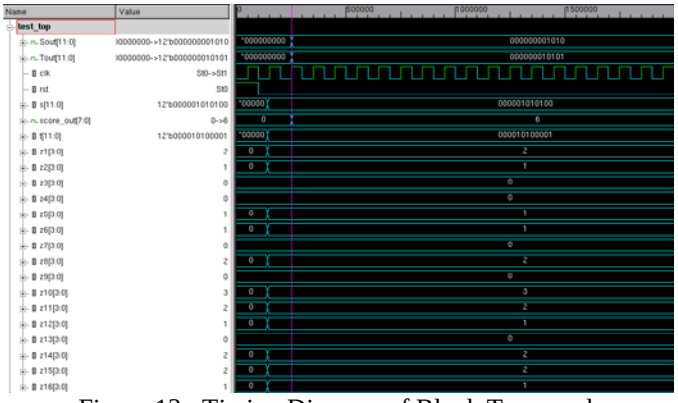


Figure 13 : Timing Diagram of Block Trace and Reconstruction of z10 from VCS

Input		Expected output
z1	2	
z2	1	Sout = 12’h0A
z3	0	12’b000_000_001_010
z4	0	
z5	1	Tout = 12’h15
z6	1	12’b000_000_010_101
z7	0	
z8	2	
z9	0	Score_out = 6
z10	3	
z11	2	
z12	1	
z13	0	
z14	2	
z15	2	
z16	1	
s=12’b000_001_010_100		
t=12’b000_010_100_001		

Table 4 : Expected output of Block z10

Table 4 illustrated the input and the expected output. The Figure 12 and 13 produce the same as the expected output and it is similar to Table 4. It shows blocks z10 can be working as needed.

Input		Expected output
z1	0	
z2	2	Sout = 12’h0A
z3	1	12’b000_000_001_010
z4	0	
z5	2	Tout = 12’h22
z6	1	12’b000_000_100_010
z7	1	
z8	0	Score_out = 6
z9	1	
z10	1	
z11	0	
z12	3	
z13	0	
z14	0	
z15	0	
z16	2	
s=12’b100_000_001_010		
t=12’b000_100_010_010		

Table 5: Expected output of Block z12

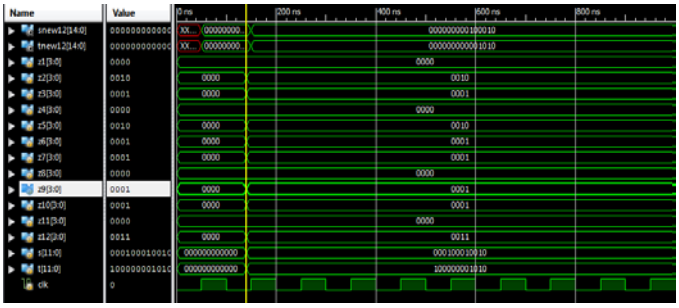


Figure 14: Timing diagram of block trace back and reconstruction of z12 from Isim Simulator

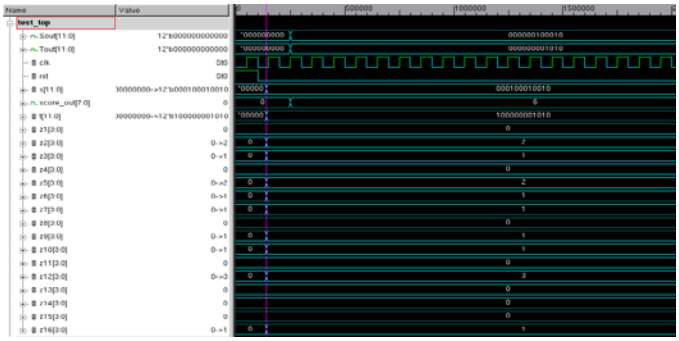


Figure 15: Timing Diagram of Block Trace and Reconstruction of z12 from VCS

Table 5 demonstrated the input of z1 until z16, s, t and expected output for block z12. The expected output from the table is shown the timing diagram in the Figure 14 and 15. The both figure show that the outputs are matching from the table. Moreover, the functionality of block z7 was indicated the timing diagram shows in Figure 16 and 17. Basically, all the blocks of z1 until z16 have given the output same as the expected output from the table 6.

Input		Expected output
z1	0	
z2	2	Sout = 12’h04
z3	1	12’b000_000_000_100
z4	0	
z5	0	Tout = 12’h04
z6	1	12’b000_000_000_100
z7	4	
z8	3	
z9	0	Score_out = 6
z10	0	
z11	3	
z12	3	
z13	2	
z14	1	
z15	2	
z16	2	
s=12’b001_000_100_001		
t=12’b000_100_010_001		

Table 6 : Expected output of Block z7

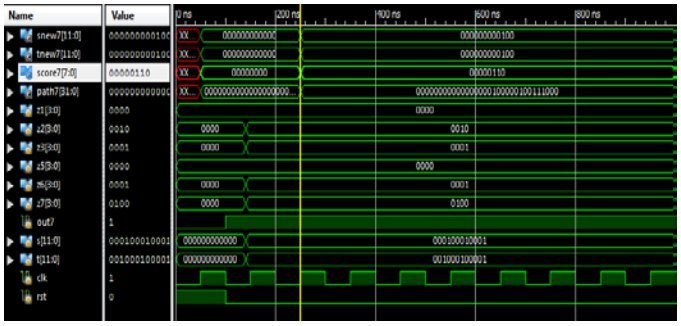


Figure 16 : Timing diagram of block trace back and reconstruction of z7 from Isim Simulator

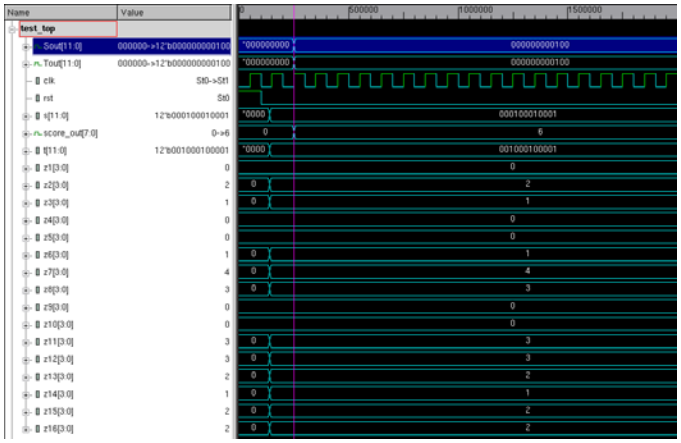


Figure 17 : Timing Diagram of Block Trace and Reconstruction of z7 from VCS

Figure 18 and 19 illustrated the timing diagram for the Comparator 2 and it behaved as enable. It is used to eliminate and produced the accurate value of the reconstruction of the Sample and Target for DNA sequence Alignment. The highest score from the previous block will be the input for the comparator 2. The both Figure demonstrated the highest score from the Block z16 so the output will produce as table 7 below:

Input	Output
Snew=12'b000_001_010_100	Sout=12'b000_001_010_100 / 12'h054
Tnew=12'b000_001_010_100	Tout=12'b000_001_010_100 12'h054
Score= 20	Score_out=20

Table 7 : Output of Comparator 2

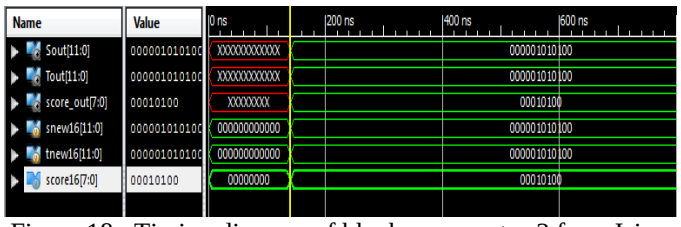


Figure 18 : Timing diagram of block comparator 2 from Isim Simulator

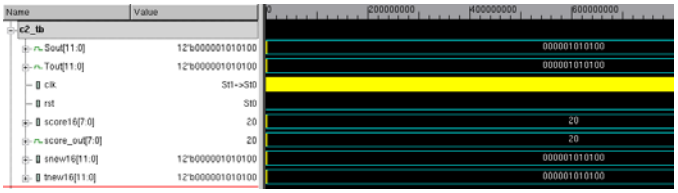


Figure 19 : Timing diagram of block comparator 2 from Isim Simulator

After combine all the 18 blocks of the design, it needs to be naming it as Top Module as Figure 20. From the timing diagram shows the several input are inserting to the top module and the outputs are generated after one clock cycle. The output of Top module have been expressed the same as the expected output and it is already discuss all the functionality of each Block before.

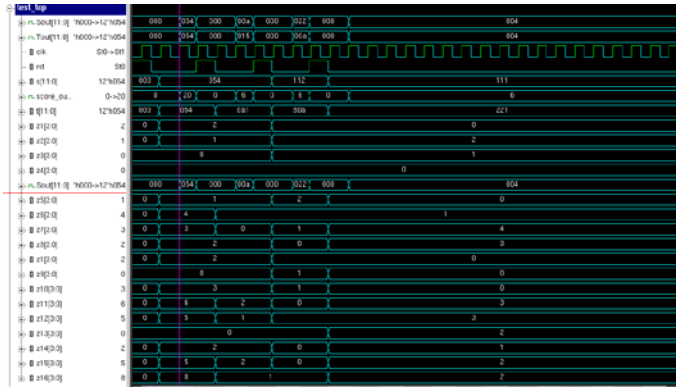


Figure 20 : Timing Diagram of the Top Module from VC

After the entire blocks have been test the functionality of the design, the timing analysis needs to verify. The timing need to be analyzing using the Design Compiler (DC), Integrated Circuit Compiler (ICC) and Prime Time (PT) by varied a constant such as clock period, input delay, output delay and uncertainty. The large area of the design it gives some problem to meet the timing requirement in ICC and PT. Several steps are taking to get the best result. For example varied the Floor planner, insert the several command to give the high effort for the design during placement cell and routing. It is used to tie up the design and the design can be operate in 55 clock period. The design takes 1 hour to complete DC, ICC about 4 to5 hours and PT about 30minutes.

Timing Maximum	DC	ICC	PT
Data Required Time	49.14	9.69	49.70
Data Arrival Time	-22.17	-4.79	-2.62
Slack	26.97	4.91	47.07
Characteristics	Met	Met	Met

Table 8: Comparison of Timing (Maximum) Analysis

Timing Minimum	DC	ICC	PT
Data Required Time	1.42	0.05	0.02
Data Arrival Time	-1.42	0.24	0.03
Slack	0	0.29	0.05
Characteristics	**Met**	**Met**	**Met**

Table 9: Comparison of Timing (Minimum) Analysis

Table 8 and 9 is illustrated maximum and minimum timing analyses are met requirements for three cases DC, ICC and PT. To meet the timing requirement the slack is in positive. If any negative value of slacks that shows the design have some violation. To meet the characteristic, the data requirement must be greater than data arrived.

Area	DC	ICC	Difference
Cell Area	1636632.061630	1673598.345246	36 966
Design Area	1636632.061630	1673598.345246	36 966

Table 10: Design Area

Power	DC	ICC
Dynamic	7.3903Mw	8.4103mW
Leakage	62.322uW	61.362uW

Table 11: Power consumption

Table 11 is indicated the dynamic and the Leakage power of the design. Dynamic power occurs while the IC is operating which are points to switch and the leakage is the power during the IC doesn't have to operate.

IV. Conclusion

The result of this research shows that the design of the Trace Back and Reconstruction fully utilized the Smith Waterman Algorithm. On the other hand, the timing analysis from the Isim simulator and VCS meet the requirement by producing output after one clock cycle of each block. The result shows that the design have a functionality same as the expected output. In addition, the design are implemented using the Synopsys ASIC design tools and the layout as shown in appendix, the design are produce from the GDSII after meet timing perform by DC, ICC and PT. The entire value meet by having the positive slack after comparing time of data requirement is greater than time of data arrived. The design can be operate at 50 clock period. It is because the complexities of programming design. As a conclusion the entire objective, to construct the Trace Back and Reconstruction module, to perform the timing analysis and also to implement the design using the ASIC flow is fully utilized. As a conclusion all the objective of this paper are achieved.

Recommendation

The design can be improve by using Finite State Machine, FSM Method. FSM is the method to improve the complexity of the programming, design area, slack time and power consumptions.

Acknowledgment

I would like to express my deep and sincere gratitude to my supervisor, Siti Lailatul Mohd Hassan, co-supervisor Mr Abdul Karimi Halim, Syazilawati Mohamed and Syed Abdul Mutalib Al Junid. Their wide knowledge and their logical way of thinking have been great value for me. Their understanding, encouraging and personal guidance have provided a good basis for this technical paper. I owe my loving thanks to my parent. Without their encouragement and understanding it would have been impossible for me to finish this work.

References

[1] C.W. Yu, K.H. Kwong, K.H. Lee and P.H.W. Leong, "A Smith-Waterman Systolic Cell", Hong Kong, 2007

[2] Syed Abdul Mutalib Al Junid, Zulkifli Abd Majid, Abdul Karimi Halim, "Development of DNA Sequencing Accelerator Based on Smith Waterman Algorithm with Heuristic Divide and Conquer Technique for FPGA Implementation', Kuala Lumpur,2008

[3] Hsien-Yu Liao, Meng-Lai Yin, Yi Cheng, "A Parallel Implementation of the Smith-Waterman Algorithm for Massive Sequences Searching", San Francisco, CA, USA , 2004

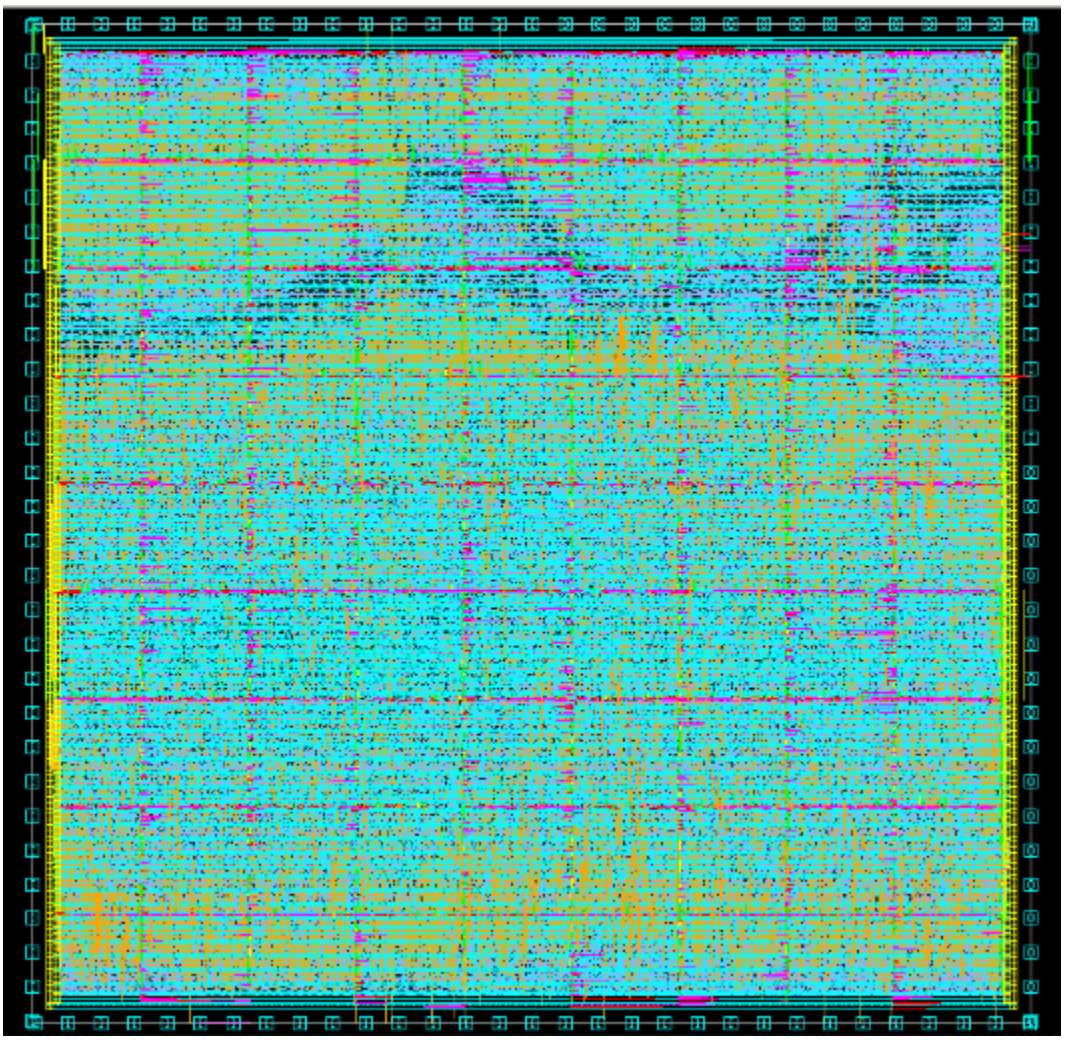
[4] Introduction to Bioinformatics, 2nd Edition, by Arthur M.Lesk, 2005

[5] F. Zhang, X-Z. Qiao, Z-Y. Liu, "A parallel smith-waterman algorithm based on divide and conquer," ICA3PP '02, 2002.

[6] Laiq Hasan,Zaid Al-Ars and Stamatis Vassiliadis, "Hardware Acceleration of Sequence Alignment Algorithms - An Overview", Netherlands,2007

[7] L. Hasan and Z. Al-Ars, "Performance Improvement of the Smith Waterman Algorithm",

Appendix



Layout of 50clock period