ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

Map Reduce Model Using FPGA Acceleration for DNA Sequence Mapping

[1] Yeruva Sai Bhanu, [2] Dr Narendra.C.P

[1] Department of ECE, Bangalore Institute of Technology, Karnataka, India.
[2] Associate Professor, Department of ECE, Bangalore Institute of Technology, Karnataka, India.
Corresponding Author Email: [1] Sairadha4346@gmail.com, [2] Narendra.nandini@gmail.com

Abstract*— In this paper we are going to propose a new approach to map the DNA sequences using Smith waterman algorithm with the Gotoh algorithm. In gene Biomedical applications sequence mapping plays a major reinforce identifying various issues regarding diseases, mutations etc. can be analyzed and identified. As the mapping process is the most time-consuming process here we are going to resolve it by introducing g divided and conquer-based approach implementing parallelism for mapping the sequences and we can enhance the performance of the mapping sequence process. The proposed implementation is suitable for FPGA hardware utilization which is the prime factor of our suggested implementation. The synthesis and simulation of the proposed implementation can be done using CAD tools & MATLAB for extracting the DNA data.

Keywords: DNA sequence, Sequence Alignment, Smith waterman algorithm, Gotoh FPGA, D & C principle.

I. INTRODUCTION

In bioscience studies, DNA plays a vital compound to retrieve data about its specifications and species. This particular DNA sequence describes the functionality and features etc., information related to a particular creature is known and the data related will be comprised in it. This sequence mainly consists of four different nucleotide structures that can be defined as A, T, G and C. these are n further abbreviated as A for Aden and T for thymine, C for cytosine and G for guanine. Aligning them all together in various forms constitutes the DNA structure of a particular species. The alignment of the resultant sequence of DNA will be in a structure called helical that is utilized for getting the features and characteristics of that unique species. For a wide of applications, this DNA matching is widely employed to analyse and extract the data which makes use of a method called mapping among two different sequences.

Scientists use two methods to solve the mapping problem: heuristic and exact methods. Heuristic methods solve the mapping problem more rapidly than accurate methods do. However, such methods like BLAST and FASTA, suffer from accuracy. Therefore, bioinformatics researchers use dynamic programming, which is more efficient; an algorithm called smith waterman is a very widely employed algorithm for the determination of correlation between two different data sequences

Sequence:

The four major elements of DNA with a wide range of combinations among these nucleotides can be simply defined as a sequence.

Sequencing:

Sequencing is a method or approach to determining the fragments of DNA.

Sequence alignment:

The comparability lies in structuring the collection of nucleotide sequences for detecting homology sections amongst those fragments. These can be utilised for examining and evaluating whether there exists any relationship between the query (which is our input sequence Thread needle to analyse d) and the database (simply it can be a known sequence for reference from the cloud or any database system). To determine the resemblance between the known and unknown fragments of DNA data of sequence, here the methodologies are a dynamic way of programming which includes the break-down of the whole sequence of nucleotides into simpler, non-interdependent subsets of sequences. That gets the scoring for every search of resemblance to find the congruent with furthermore realistically.

Sequence Alignment methods:

The congruence can be classified into majorly in two ways:

Global Alignment:

Composition wise similar strings of DNA possessing identic numbers of nucleotides can be considered the as best suitable sequences for this type of alignment/congruency. This way of alignment can be processed from start to end by identifying the appropriate and most practically suitable alignment match of nucleotide.

Local Alignment:

Local matching of data is mostly applied for getting a comparison of data which assumes that may be similar or

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

unique data. This uncovers local systems with a significant degree of correlation.

Separate algorithms describe these two alignment approaches, which employ scoring matrices to align the two different series of letters or patterns (sequences). For aligning two distinct sequences, the two alternative alignment algorithms are primarily described by the Dynamic programming methodology.

II. RELATED WORKS

BLASTN:

The most basic and widely utilized data sequence correlation finder in the field of molecular biology is BLATN which is specifically for DNA data. It is employed to efficiently locate biologically significant comparable areas between two sequences. As the amount of genome sequences grows, so does the time required for BLAST to do a thorough genomic database search. As a result, it is necessary to expedite the search process. This method is employed in various scenarios such as when there is a need for determining the correlation amongst the huge amount of genomic datasets concerning the query (unknown/input) sequence. BLASTN method is the extraction from BLAST technique but complies lately dedicated to DNA data. This method is major consisting of various stages such as the matching of words where the determination of matching can be analysed followed by extensions of both un-gapped and gap the ed as next following stage.

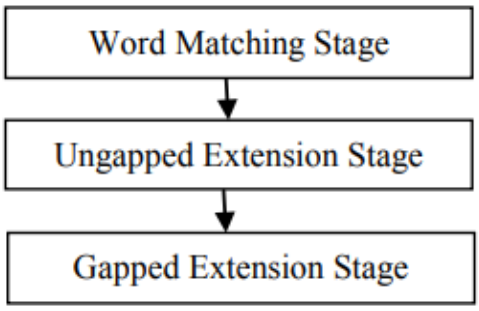


Fig.1: basic steps or stages involved in BLAST

The input for the Word Matching step is a string of DNA bases that commonly comprises A, C, G, and T. The database is searched for brief precise matches of a specific length between the query and the topic sequence. This brief precise match is referred to as the seed. Each seed is extended in both directions during the Ungapped Extension step, permitting replacement. When two sequences of equal length are split and their alignment scores surpass the threshold value, such pairs are referred to be high scoring pairs (HSP). In the end, such HSPs are obtained. Dynamic programming algorithms are utilised to expand the HSPs during the Gapped Extension step. It supports insertions and removals. The fundamental concept of involvement in this particular blast method is the concept of filtering. Though every stage of this method of pipelining results in increasingly complex, it is critical to limit the quantity of data that must be processed owing to the exponential rise in data volume. Filtration is a process that discards unimportant fractions as soon as feasible, reducing total computing time. Each level of BLASTN needs a varying amount of processing time. After analysing each stage, it becomes clear that the word-matching tags are the most time intensive. As a result, to improve BLASTN's overall performance, the calculation time of this step should be sped up.

Dynamic Programming:

The Needleman-Wunsch and Smith-Waterman data of DNA correlation methods come under this dynamic programming approach. In 1981 scientists named Temple F. Smith and Michael S. Waterman are the persons to propose a sequence alignment algorithm known as Smith-Waterman. It gives conserved regions of correlation amongst the sequences and may line up 2 partially overlapping fragments of data sequences. the framed sub-segments of aligned fragments can also be applied for checking the similarity between the known and unknown sequences. this S-W varied from the N-W as it will have any negative scoring only for un-matched conditions any value is negative n the matrix fill it will mark to zero (consider max value in comparison with '0').

III. PROPOSED METHOD

SYSTOLIC ARRAY ALIGNMENT:

The Systolic Arrays can be determined by whether they are included in local data transmission and parallel processing. Processing Elements can be placed in a 2-d format, which is as follows Fig. 3.1(a), linear, as shown in Fig. 3.1(b), or some other random shapes structures like hexagonal, as shown in Fig. 3.1(c), which is here implemented as a circuit with data flowing in 3 various directions. in the very conventional or traditional process[7], it can be designed by starting with the problems in the flow and after that following with these stages: 1) as a first step/stage DG which means a dependency graph which is the generation of flow in a graphical way including nodes that reciprocates the evaluation part and the edges represent the dependency of data that is provided; 2) the next thing is the generation of SFG which is a signal flow graph which is the reduced one axis of DG.

Subsequently, an SFG has a dimension less than the correspondent DG; this is because in a DG each node represents one simple computation, while in the SFG a node is a processing unit, that should be reused in successive time steps, and for this reason nodes of a line in a DG can be mapped to a single node in the SFG; and 3) array processor design, that consists in designing the internal structure of each PE.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

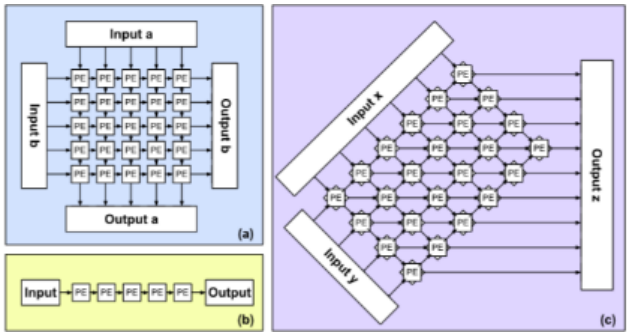


Figure 3.1 Systolic Array

SEQUENCE ALIGNMENT

Sequence comparison is a fundamental operation that allows biologists to find any functional correlations. Sequence alignment is the most computationally intensive task during comparison and providing alignment score that represents similarity index amongst unknown which is query can be defined as Qry and known sequence which is obtained from the database here referred as Subj.

If there is a match or a substitution the value of the scoring must be updated using the value coming from a substitutional matrix: given the two AAs Qry(i) and Sbj(j), this matrix returns a value s(Qry(i), Sbj(j)) that represents the probability that an AA(amino acids) is substituted with another during evolution [21]. When a deletion or an insertion occurs, the alignment score must be updated using instead a gap penalty.

The most used penalty is the affine one (γaff) in which two different values, called gap_open d and gap_extension e, are used to encourage one large gap rather than many small ones since the former condition is more likely. In the following equation, g represents the length of the gap [22]:

$\gamma aff = -\ d-(g-1)\cdot e.$ (1)

for the analysis and extraction of correlated sequences, the widely utilized method of alignment tool is "S-W" [23].

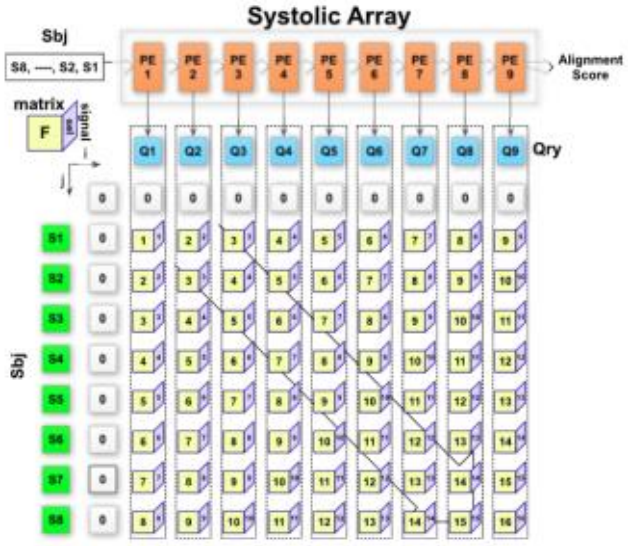


Figure 3.2 DGS_S-W algorithm matrix alignment

DGS_S-W algorithm

S-W is a dynamic programming approach that uses a score matrix F(i, j) to record correlation values at every moment, as seen in the lower section of Fig. 3.2. The scope of this study does not allow for a full discussion of the S-W algorithm. Here, we describe briefly its architecture and to do so, we introduce its working mechanism at a glance. It is important to highlight that adopting the gap affine model, each cell is required to evaluate three different values.

Recently, this algorithm has been optimized for the SA implementation allowing faster and lighter computations. The main idea of this optimization is to use a cell signal that can be either 0 or 1 to choose on-the-fly between gap_open and gap_extension and for this reason, it is called dynamic gap selector (DGS): this substitutes the usage of two gap matrices that are provided by the original algorithm, that required a higher computation effort and an increased complexity. The s-W algorithm is divided into three steps. 1) Initialization: first row and first column of the matrix, respectively, F(i,0) and F(0, j), are initialized to 0. 2) Score Matrix filling: each cell is filled with a value of F(i, j); in the case of DGS_S-W, this value is evaluated.

Traceback: the maximum score represents the starting point for the best local alignment that is found tracing back till the first 0 is found.

Interleaving in SAs

In this process of sequencing correlation, systolic array structure has been involved. Here we are going to elaborate on assigning the interleaving to the PE level and the extension of it to the entire SA. These can be further categorized as 1) with WIL and 2) without having the internal loop; this thing is again extended to the segmentation of subdivisions as 1) results storing in cells and 2) those with the extension of results over cells to get WIL-PT. The DGS S-W SA falls within this category.

The major benefits of bit based matching process are as follows:

- In hardware, RNA rules and DNA patterns are stored in nonvolatile memory. In bit-based matching, each FSM state holds only one bit. Significant speed improvements and area reduction is obtained using FFSM-based state transitions.
- Parallel computation is accomplished with multiple sub-patterns that are stored in different FSM machines and synchronized through PMV vectors.
- Bit wise FSM state transitions fail even with a one-bit mismatch during the string matching process and successive payloads are bypassed without making any further FSM state transition and matching process.

And the key issues related to the DNA biological sequence matching include lack of parallel processing during similarity measure and scoring, selection of core processing elements for lesser power and to exploits high scalability.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

Therefore, all the factors discussed so far gives a complete scenario of the factors that are required to improve the throughput rate of pattern-matching systems for RNA systems and biological pattern analyses. The various factors that motivated this research are:

- The data pattern correlation algorithm leads to a high amount of propagation time that is affected to a greater extent by a wide range of nucleotides together as a pattern along with its complexity. In this case, longer incursions should be broken into many sub-patterns and matched in parallel.
- Scalability and reconfigurable nature of the memory elements used to store the patterns for matching units is essential to update and include the new patterns and also for realizing the parallelism.
- The only possible way to reduce power consumption in any RNA system is dynamic power management (DPM) through a data transition reduction controller.
- FPGA-based systems provide higher flexibility and reconfigurability compared to ASIC-based solutions with satisfactory processing throughput without causing latency problems.
- Integration of the entire patterns matching system using ADPLL will provide a fully synchronous matching process for RNA for a different kind of network traffic rate.

Bit-Based String Matching Methodologies

The statistical nature of RNA rule sets and matrix evaluation of dynamic programming in DNA sequence alignment bit-based string matching methodology is proposed as shown in Fig m which works across RNA patterns with different traffic rates, correlation and gab penalty measures of various DNA sequences. Here, the idea is to process string as a sequence of bits and to handle rule set updating and database changes using reconfigurable hardware devices. In this, a particular number of fragments are required for finding the correlation among the complete data sequence each having its storage cell of memory which is a vital thing to consider here.

PAIR-WISE DNA SEQUENCE ALIGNMENT

Sequence alignment by pairs Dynamic programming is always more accurate than heuristic programming at sequence alignment. However, it has the disadvantage of being computationally complicated and time-consuming. As a result, because DP techniques are exhaustive in nature, optimising them to enhance speed is challenging

DGS_SW is well known Dynamic programming method to explore the correlation level between the sequences through the score matrix generated dynamically furthering the matching process. If a gap is arise during the matching process in the matrix, dynamic programming is used to align the gaps by assigning some unified score values.

Here to increase the throughput of PE, pipelining was merged into it.

DGS_SW: Dynamically selection of Gap in Smith-Waterman

This is the approach which is most widely utilized for the congruency of data sequences and here we further implemented a technique called pipelining in the systolic array approach.

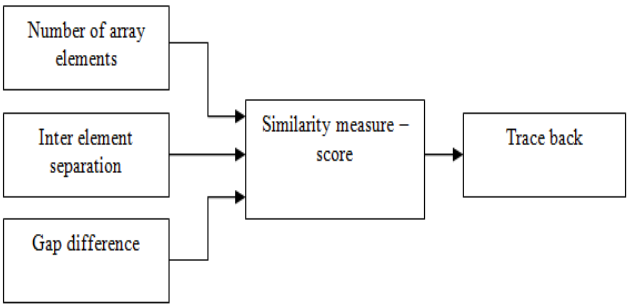


Fig. 3.4. dual data fragments correlation using smith waterman technique

Here two segments of whole data are correlated obtained from the matrix PE for performing the congruency operation as shown in Fig 4. Utilizing SA based pipelining amongst the PE blocks. Finally, score assignments are done for analyzing the amount of correlation and penalty gap between unknown input and reference input. DGS_SW is a dynamic programming-based algorithm that determines a score matrix F (i, j) during the correlation-determining process as shown in Fig 5.

Smith-Waterman Algorithm

1. Converting the raw data to bits of "0" & "1"
2. Generating an L-length systolic matrix array
3. Segmenting the L into subsequences
   Assigning 2 as the match score if the condition is satisfied or else assigning zero as the gap difference
   Once this is satisfied we can directly go to the 5^{th} step.
4. Escalate the score numbers to preceding rows &b columns for every new task of nucleotide matching.
5. By utilizing max driven point sin array we have traced back the sequence of data
6. Finally need to decide on the process

The result indicates that dynamic programming has the efficiency to operate at high frequency and shows a clear improvement in hardware utilization rate with significant hardware complexity reduction.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

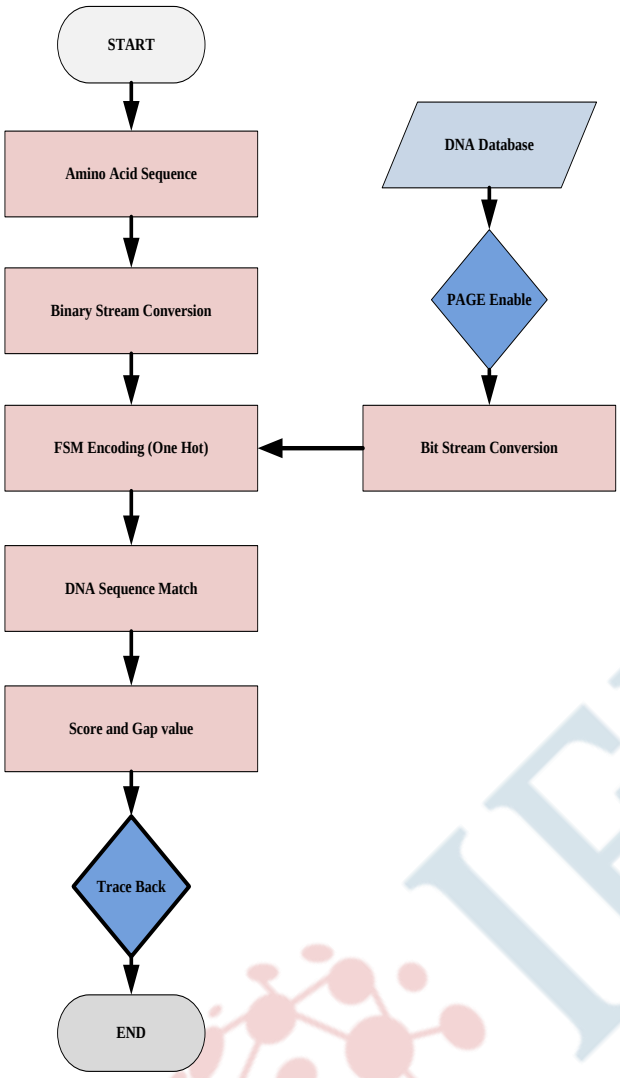


Fig. 3.5. Architecture for sequence alignment using FSM machines

Though the sequence alignment of biological sequence matching is a data dependence algorithm here with FSM-based bit holding method constraint it over a maximum number of PEs required to match the total query sequence in the database and the problem over trade-off complexity over incoming protein sequences lengths is solved.

During the hardware synthesis, it was observed that,

- When compared to an ASCII-based matching procedure, the number of comparisons and bit transitions is significantly reduced.
- The maximum number of PEs in FSM-based architecture is set with pre-defined pattern lengths.

The bitwise FSM state transitions share the same features that of an RNA system hence it is essential for a high-speed DNA matching process. Consequently, all these observations and advantages were analyzed and in the subsequent section, a detailed explanation is given about FSM machines proposed for RNA and DNA biological sequence matching process.

IV. RESULTS AND DISCUSSION

Fig. known and unknown DNA

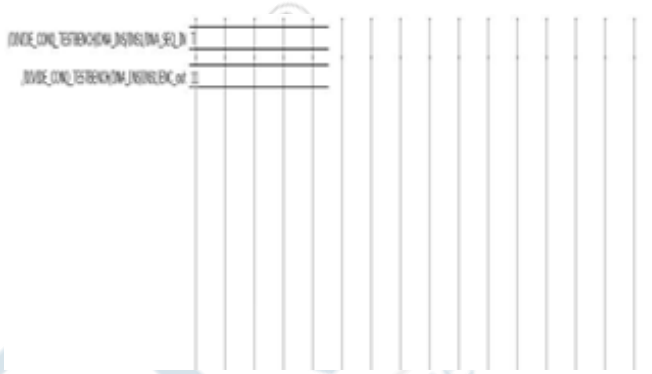


Fig. Encoded output

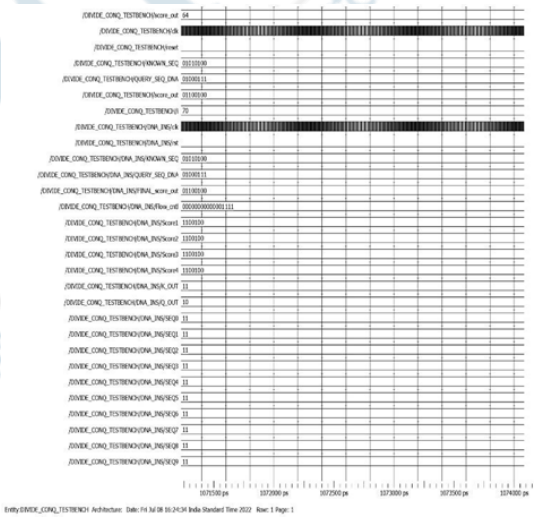
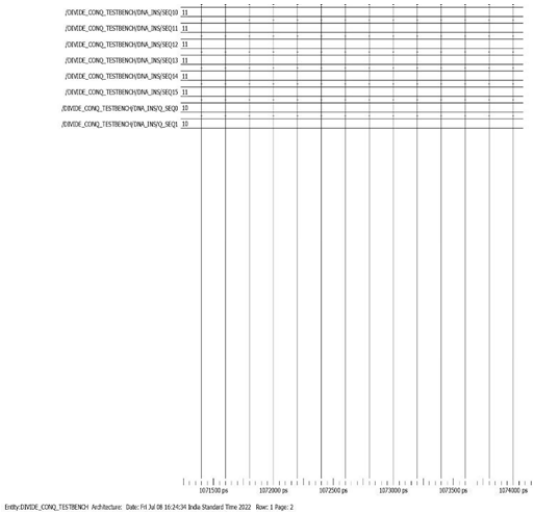


Fig. Score value generation



ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

Date: July 10, 2022 Project: final

Flow Status	Successful - Sun Jul 10 13:28:01 2022
Quartus II Version	9.1 Build 222 10/21/2009 SJ Web Edition
Revision Name	final
Top-level Entity Name	DNA_TOP
Family	Stratix II
Met timing requirements	No
Logic utilization	1 %
Combinational ALUTs	102 / 12,480 (< 1 %)
Dedicated logic registers	136 / 12,480 (1 %)
Total registers	136
Total pins	26 / 343 (8 %)
Total virtual pins	0
Total block memory bits	0 / 419,328 (0 %)
DSP block 9-bit elements	0 / 96 (0 %)
Total PLLs	0 / 6 (0 %)
Total DLLs	0 / 2 (0 %)
Device	EP2S15F484C3
Timing Models	Final

Page 1 of Revision: final

Fig. Logical elements of the proposed system

Fig. FSM State values

Date: July 10, 2022 Project: privious

Flow Status	Successful - Sun Jul 10 13:29:27 2022
Quartus II Version	9.1 Build 222 10/21/2009 SJ Web Edition
Revision Name	privious
Top-level Entity Name	DNA_TOP
Family	Stratix II
Met timing requirements	Yes
Logic utilization	1 %
Combinational ALUTs	93 / 12,480 (< 1 %)
Dedicated logic registers	125 / 12,480 (1 %)
Total registers	125
Total pins	26 / 343 (8 %)
Total virtual pins	0
Total block memory bits	0 / 419,328 (0 %)
DSP block 9-bit elements	0 / 96 (0 %)
Total PLLs	0 / 6 (0 %)
Total DLLs	0 / 2 (0 %)
Device	EP2S15F484C3
Timing Models	Final

Page 1 of Revision: privious

Fig. Logical elements of the existing system

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022

Date: July 10, 2022 Project: privious

PowerPlay Power Analyzer Status	Successful - Sun Jul 10 13:32:38 2022
Quartus II Version	9.1 Build 222 10/21/2009 SJ Web Edition
Revision Name	privious
Top-level Entity Name	DNA_TOP
Family	Stratix II
Device	EP2S15F484C3
Power Models	Final
Total Thermal Power Dissipation	324.38 mW
Core Dynamic Thermal Power Dissipatio	0.00 mW
Core Static Thermal Power Dissipation	302.98 mW
I/O Thermal Power Dissipation	21.40 mW
Power Estimation Confidence	Low: user provided insufficient toggle rate data

Fig. Power Analyzer report of the proposed system

Fmax Summary

	Fmax	Restricted Fmax	Clock Name	Note
1	304.69 MHz	304.69 MHz	clk	

Fig. Power Analyzer report of the existing system

Fmax Summary

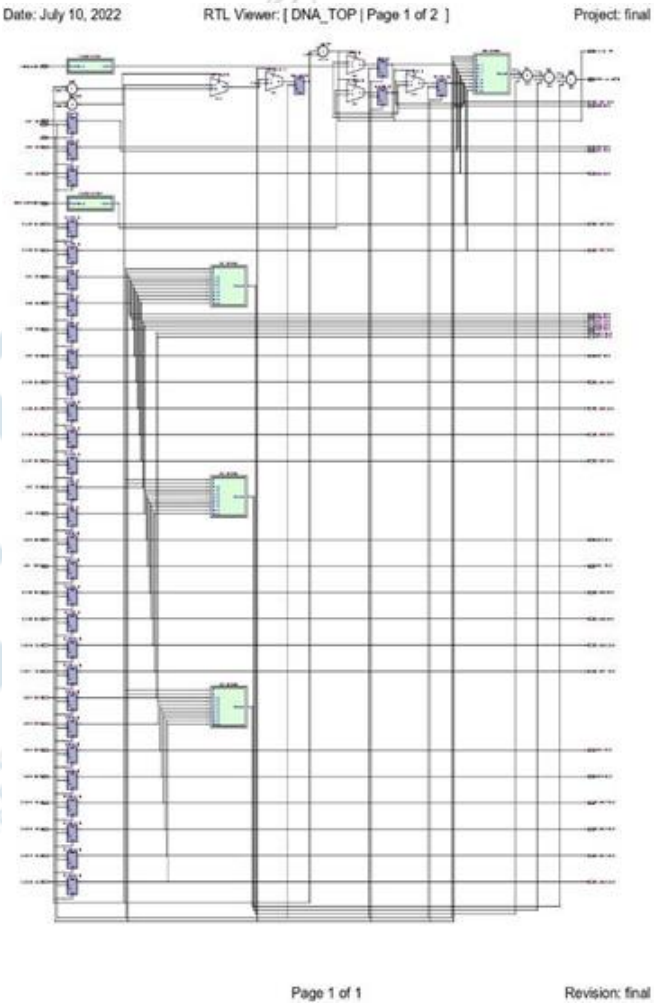
	Fmax	Restricted Fmax	Clock Name	Note
1	306.65 MHz	306.65 MHz	clk	

Fig. Time Analyzer report of the proposed system

Type used	LE's used	Fmax (M Hz)	Power (mW)
Core processing element	93 logical elements	306.65	302.98
FSM-based array element	102 logical elements	302.98	302.98

Here DNA grouping arrangement recreation is done to decide the malignant growth types given succession arrangements on an enormous data set where various examples of different disease types are sorted. Here thorough test seat is made with a succession of different protein arrangements and recognition is likewise conveyed after the match score is assessed.

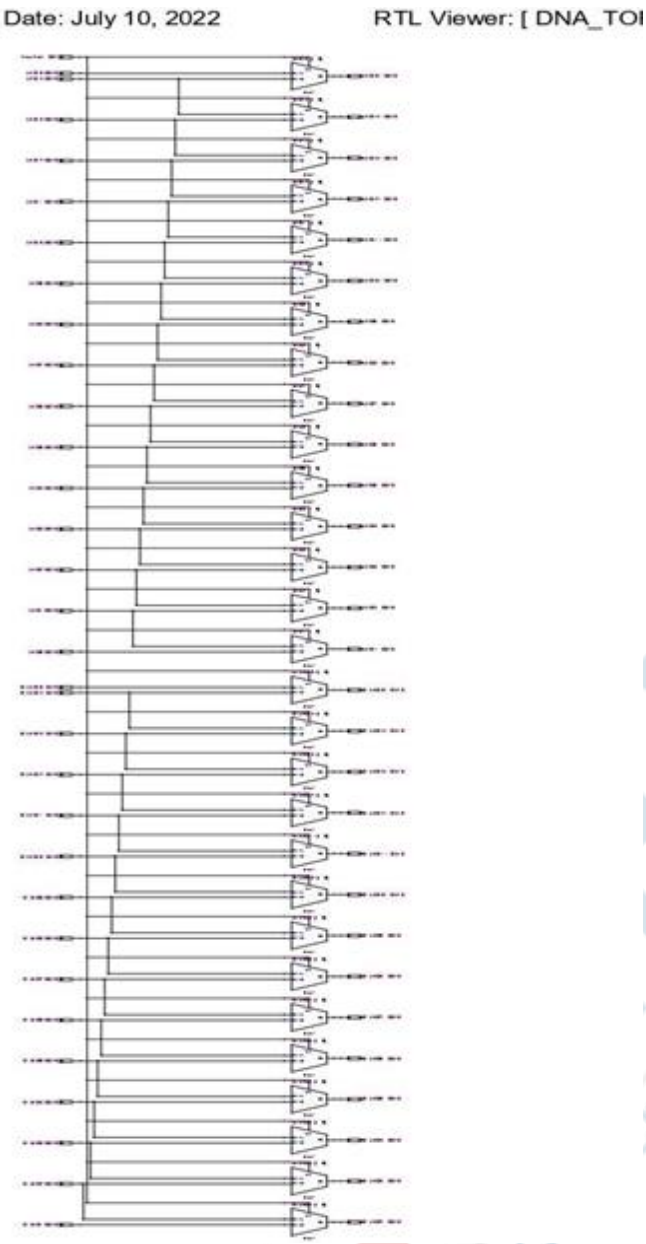
RTL VIEW OF THE COMPLETE SYSTEM

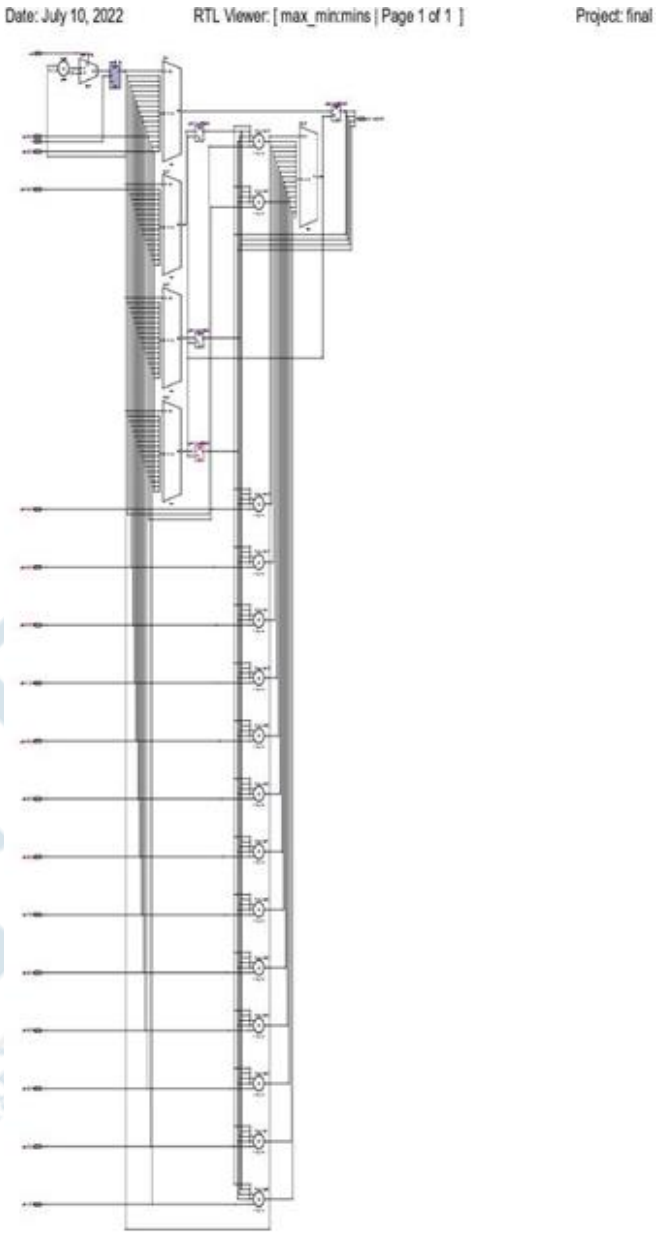


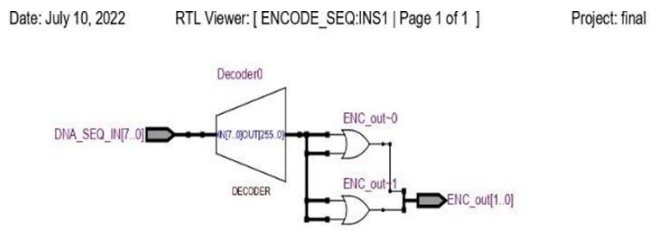
ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022



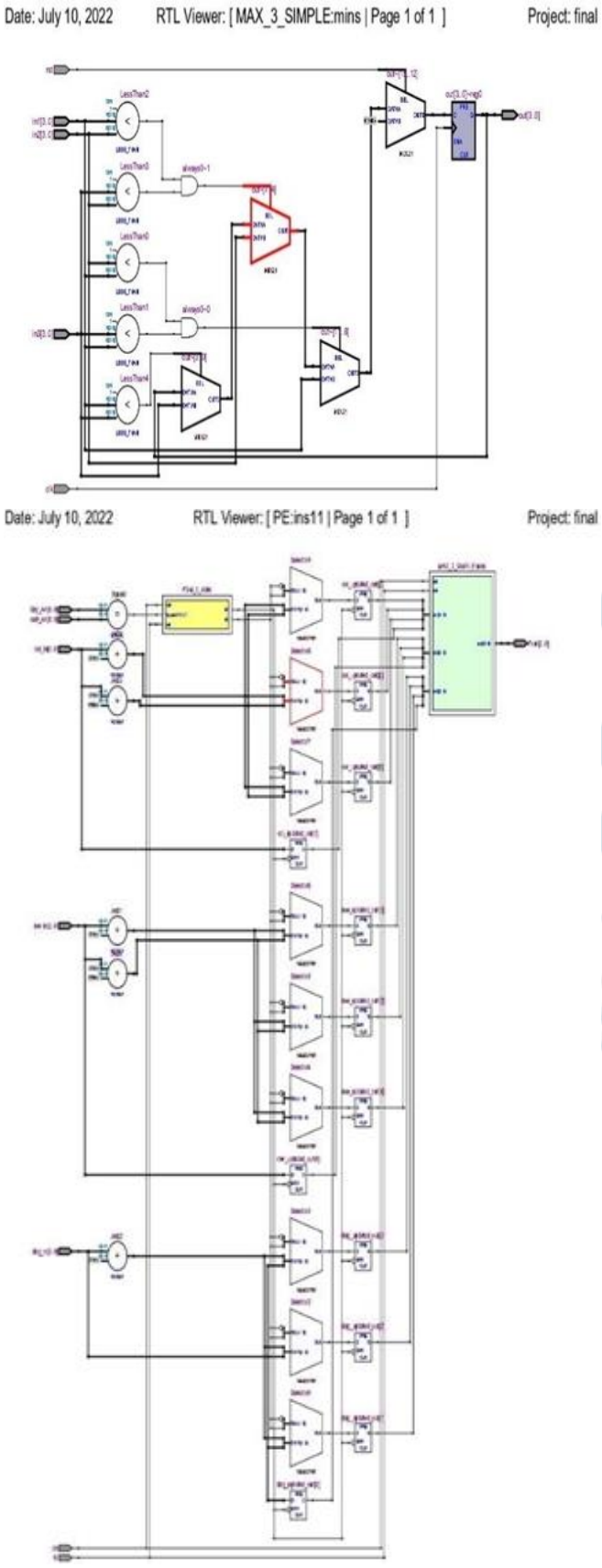




ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)

Vol 9, Issue 11, November 2022



V. CONCLUSION

Because several sequences are matched in parallel, it is demonstrated that the deconstructed systolic array technique always delivers extremely good performance. However, flexibility to optimise the power over the DGS SW algorithm is not available. The only approach to approximate matching without using any power reduction strategies is to address distinct regions of PEs. However, this will have a substantial impact on performance. A systolic array-based matrix calculation is used to calculate the similarity score between DNA sequences to efficiently extract the gap penalty.

REFERENCES

[1] Yupeng Chen, Bertilschmit, Douglas L. Maskel, "Reconfigurable Accelerator for the Word-Matching Stage of BLASTN," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 21, no. 4, pp-659-667, April 2013.

[2] Snort, Ver.2.8, Network Intrusion Detection System, HTTP:// www.snort.org., 2011.

[3] Clam AntiVirus, Ver.0.95.3. http://www.clamav.net., 2011.

[4] C.-H. Lin, Y.-T. Tai, and S.-C. Chang, "Optimization of Pattern Matching Algorithm for Memory Based Architecture," Proc. Third ACM/IEEE Symp. Architecture for Networking and Comm. Systems, pp. 11-16, 2007.

[5] Deterministic Finite-State Machine, http://en.wikipedia.org/ wiki/Deterministic_finite_state_machine, 2011.

[6] H. Kim, H. Hong, H.-S. Kim, and S. Kang, "A Memory-Efficient Parallel String Matching for Intrusion Detection Systems," IEEE Comm. Letters, vol. 13, no. 12, pp. 1004-1006, Dec. 2009.

[7] Virtex-4 FPGA User Guide, http://www.xilinx.com/support/ documentation/user_guides/ug070.pdf., 2011.

[8] F. Yu, Z. Chen, Y. Diao, T.V. Lakshman, and R.H. Katz, "Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection," Proc. Second ACM/IEEE Symp. Architecture for Networking and Comm. Systems, pp. 93-102, 2006.

[9] A.V. Aho and M.J. Corasick, "Efficient String Matching: An Aid to Bibliographic Search," Comm. ACM, vol. 18, no 6, pp. 333-340, 1975.

[10] L. Tan and T. Sherwood, "A High Throughput String Matching Architecture for Intrusion Detection and Prevention," Proc. 32nd IEEE/ACM Int'l Symp. Computer Architecture, pp. 112-122, 2005