International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 16 (2017) pp. 5844-5848

ASIC Implementation of High Throughput Low Power DNA Protein Sequence Computation using FSM

Sancarapu Nagaraju.

Department of ECE,
Jawaharlal Nehru Technological University,
Ananthapuramu, Andhra Pradesh, India.

Orcid : 0000-0003-0682-1371

Dr. Penubolu Sudhakara Reddy

Department of ECE,
Sri Kalahasteeswara Institute of Technology,
Srikalahasthi, Andhra Pradesh, India.

Orcid: 0000-0002-2153-0487

Abstract

It is always very difficult to identify cancer diseases at initial stages. In recent years DNA Protein comparison is getting popular and it is rapidly outpacing the rate at which it has to be processed to match correlation between different species genetic patterns in order to find various genetic diseases and also diseases like cancers. Traditional software implication based sequence alignment methods can't meet the actual data rate requirements. Hardware based approach will give high scalability and one can process parallel task with a large number of new databases. In this paper, we explain FSM (Finite State Machine) based core processing element to classify the protein sequence. In addition, we analyze the performance of bit based sequence alignment algorithms and present the inner stage pipelined digital architecture for sequence alignment implementations. Synchronized controllers are used to carry parallel sequence alignment. The complete architecture is designed to carry out parallel processing in hardware, with FSM based bit wised pattern comparison with scalability as well as with a minimum number of computations. Finally, the data rate of 12.8 Gbps is achieved and well proved using ASIC advanced design compiler tool.

Keywords: DNA, Protein sequence, Low power, ASIC, FSM, Smith-Waterman algorithm etc.

INTRODUCTION

Bioinformatics local alignment analysis between two biological sequences is carried out with different computational methods to accommodate all possible genetic composition in organisms [1]. Sequence alignment is most widely adopted technique in alignment process [2]. Here the correlation and similarity between different biological sequences are analyzed to find the genetic relationship between sequences. The sequence matching process is different from matching process carried out in NIDS system [3]. Since biological sequences have complicated relationships [4].

LITERATURE REVIEW

Many works have been published [5], [6] for hardware based digitized bio sequence alignment to achieve better performance over software based approaches. However, hardware based sequence alignment requires many core inner processing elements (PEs), logic number of operations are required depending on the DNA length consequences which could affect overall system throughput rate. In [7] pipeline basic processing, elements are used to accomplish sequence alignment in which the Needleman–Wunsch algorithm is used for Global alignment and the Smith–Waterman algorithm is incorporated for local alignment. Processing element length is optimized based on a worst-case pattern range matched. In [8] fine-grained, parallel computation is performed on multiple DNA sequences optimally based S-W driven local alignment. Smith-Waterman (S-W) algorithm [9] is a more advanced matching process for its capability to carry out matching among multiple subsequences with reduced routing delay and it is a basis for optimization done during the mapping process. In some cases, sequence comparison is done based on sequence alignments in quadratic space and time. DIALIGN alignment [10], without making any gaps during matching process total alignment is accomplished as a chain of fragments. But as like S-W algorithm, dynamic programming scores are calculated and matrices are framed accordingly for an efficient hardware implementation.

To solve the problem of S-W algorithm in its tradeoff between computational complexities over complex biological sequences, Xian yang Jiang et al. presented a fine-grain parallel PE (Processing Element) array architecture [11]. But for next-generation hardware DNA sequences, traditional solutions like systolic array-based parallel computations can't meet the basic requirements. The basic requirements for analyzing DNA sequences are growing dramatically with a large number of new databases [12]. Sequence alignment shares some similarities with NIDS system since in both the cases data sequence to get match is American Standard Code for Information Interchange (ASCII) characters. Currently, several popular sequence alignments tools such as Probcons

[13], Primal's [14] etc. are available for DNA analysis. In most cases, ASCII based comparison is carried out in basic core processing. So it is essential to improve the performance of basic core PE used to accelerate pair-wise sequence alignment over maximum data rate. For large-scale deployment of a highly complex and computational intensive task, scalability and power compatibility are of major concerns, and therefore, hardware implementation for on-chip DNA pattern matching needs to be optimized.

RELATED WORK

In this paper, it is intend to apply a technique called inner stage pipelining to increase the throughput of FSM core processing element. Pipeline technique has been widely used in many digital systems especially in the case of pattern matching algorithms. Here, we explore its merits with FSM machine, which is highly complex one. In this work with highly parallel architecture careful retiming of inputs to ensure the synchronization between the cores of FSM machines in which we carry out bit based matching process. In many previous works systolic array based approach used to accomplish this parallel task in a protein alignment algorithm, to get the benefits of a high data rate. Here by utilizing Dynamic Gap Selector Smith-Waterman (DGS_SW) sequence alignment algorithm is used to perform the local alignments of biological sequences. It's been used for protein alignments, DNA and RNA molecules providing maximum exploring of sequence alignment algorithms and also feasible for architectures implementation, in particular, FPGA-based (Field Programmable Gate Array) reconfigurable architectures. The rest of the paper is organized as follows. In section 2, FSM based bitwise matching process is described. In Section 3, a parallel architecture for DGS_SW is explained in detail. Section 4 provides comparison results and finally, the conclusion in Section 5.

PROPOSED METHODOLOGY

A. The smith-waterman algorithm

For DNA sequence matching only the matching results from local alignment are used for making sequence alignment over very divergent sequences and mostly it contain an only small portion of similarity with the database. In pair-wise sequence alignment, two sequences build a matrix based on S-W algorithms as shown in Fig 1. Sequence alignment is also performed on protein sequences by considering its representation by long sequences of amino acids with larger alphabet letters as described in [15]. This Sequence alignment process consists of alignment score, to find the grade of similarity between the input sequences called query and the database sequence called subject [16]. Here the alignment procedure is carried out taking into account only assessing the similarity between two sequences; the number of match points produces a measure of similarity.

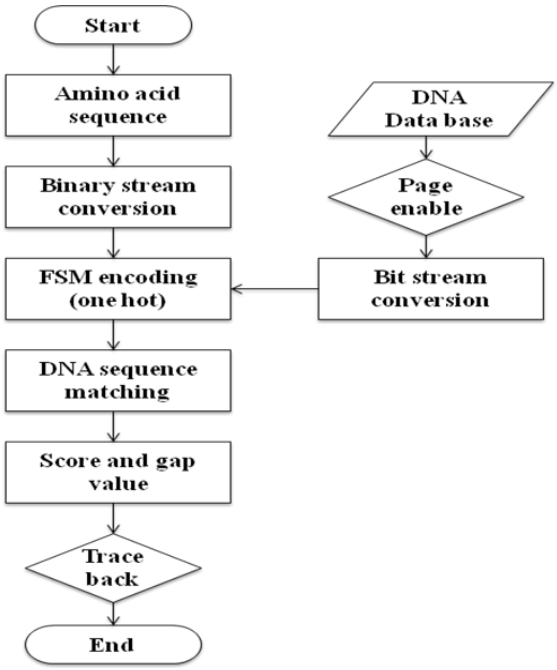

Figure 1: Flow chart for smith-waterman algorithm

B. FSM based bit wised Sequence Alignment Algorithm

1. Conversion of amino acid sequence into binary streams
2. To form systolic array for given DNA sequence of
3. length L
4. Partitioning into L1-L2-L3
5. for (i=1; i<8; i=++)

   if (L1[i]=ARRAY_CELL_1_FSM1[i] && L2[i]

   ARRAY_CELL _1_FSM2[i] && L3[i] =

   ARRAY_CELL _1_FSM3[i])

   assign match score=2;

   else

   assign gap difference=0;

   skip to trace back process step 5.
6. Forward the score values to next by elements in all
7. possible ways (row wise-column wise diagonal-wise)
8. for parallel task.
9. Trace back based on maximal driven array points.
10. Decision making process.

In order to incorporate the basic demand of parallelism in sequence alignment algorithms on FPGA, we designed FSM machines to carry out matching process. Totally here we used 8 FSM machines to accomplish the amino acid sequence (capital letters) comparison. For each matching results, score values will be assigned since no negative values are there in the Smith-Waterman dynamic programming table. The overall system architecture consists of a large number of core FSM machines. Both local transmission of the incoming sequence to successive FSM machines and parallel computation is synchronized with incoming sequence rate as shown in Fig 2.

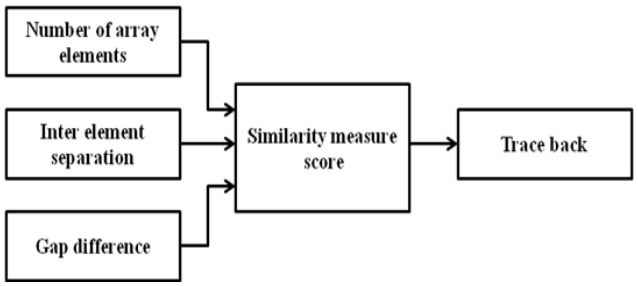

Figure 2: Pair-wise sequence matching based on S-W Algorithms

C. Sequence alignment

DGS_SW is a dynamic programming algorithm, a score matrix F (i, j) is generated during sequence matching, the level of correlation between the sequence is explored. Here one unknown input sequence is matched with many database sequences sequentially by using the maximum number of core FSM processing units as shown in Fig 3. Here we adopt the conventional Systolic array based parallelism which made of a pipelining network in between two processing elements. For each FSM match, corresponding score values will be updated and forwarded.

Here scalability is achieved using PAGE wised FSM matching units each will represent unique bio genome sequence alignment algorithms which take the advantage of systolic array to realize the parallelism. The number of FSM used for bitwise computation will be updated based on new attributes included in the database and contributes linear performance increase. Compare to core processing element based approaches FSM based core processing consume lesser power and exploits high scalability. FSM also has potential merits of high performances and proportional to the higher frequency of operation. There are two main reasons to limit the maximum number of PEs.

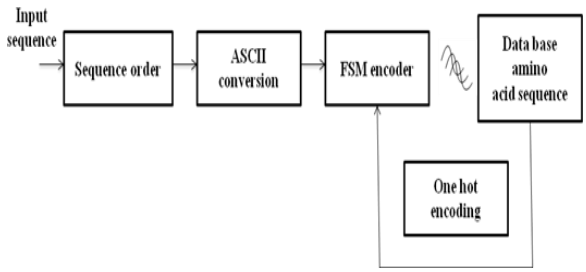

Figure 3: Matrix Architecture for a Systolic array of FSM machines.

As we discussed earlier biological sequence matching is data dependence algorithm. Here with FSM based bit holding states the constraints over a maximum number of PEs required to match the total query sequence in the database and the problem over tradeoff complexity over incoming protein sequences lengths is solved.

D. Trace back

The FSM with maximum score explores the best local alignment that FSM will be considered as a start point tracing back to find similarity match results. Here in this paper, the trace back procedure is accomplished with manual interpretations since here we focused on improvising the performance of sequence alignment using maximized parallelism with low power consumption.

EXPERIMENTAL RESULTS

Here simulation is carried out to determine the cancer types based on sequence alignments carried out on a cosmic database where different samples of various cancer types are categorized. Here exhaustive test bench is created with a sequence of various protein alignments and detection is also carried after the match score is evaluated. To make the performance comparison between FSM based arrays over the conventional systolic array, we carried out ASIC implementation with the same hardware optimization techniques except utilizing improved DGS_SW algorithm for both the cases. After the attribution of FSM used for bitwise computation at the core matching points transformation both delay and power is considerably reduced as shown in Table I. Here inner stage pipelining is used for delay optimized routing architecture in order to moderate throughput rate and it is proved to be several Gbps through Cadence design compiler 15.2 ASIC tool using TSMC 45nm library and its hardware Complexity reduction is also proven. Moreover, the computational accuracy and overall power reduction largely depend on a number of PEs used and the complexity of biological sequence used for matching process. In addition, since our bit grouping algorithm has stable FSM states, it is also suitable for any complex biological sequence with minimum complexity overhead.

E. Results comparison

The propagation delay of sequence matching scheme largely depends on bit-width used to represent the amino acid sequence. In this paper, we compare our FSM model with high speed systolic array mode for delay and power reduction. The proposed architecture is modelled using the Verilog HDL and synthesized using ASIC synthesizer TSMC 45nm technology library file. Detailed comparison results are shown in Table. 1. The primary objective goal of high performance and power reduction with less resource area utilization is proved from the EDA synthesis results as shown in Fig 4. Hardware utilization report is also given in Table I. Here with our proposed model both power and speed metrics are achieved without compromising hardware complexity. As shown in parameters listed below. Proposed periodic FSM based transition control system has shown efficient computational energy with varying the number of PEs.

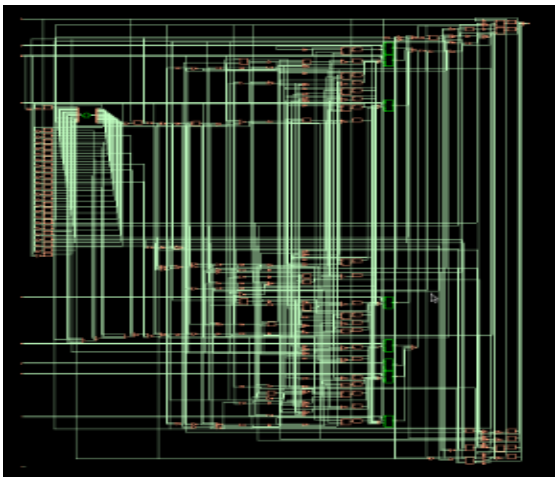

Figure4: ASIC hardware view report

The delay profile is shown in Table I. Consequently, with increasing number of PEs, with increasing system size, the energy gab between ASCII based pattern matching and FSM based bit wise computation is increased hence, the overall workload handled by each FSM state in increase since FSM states are reused.

TABLE I. ENERGY COMPARISON

Number of PE elements	Sarkar et al,.[17] NoC Architecture	FSM based approach
64	2.0e-05	9.58e-06
32	1.0e-05	4.8e-06
16	7.01e-06	2.3e-06
8	5.00e-06	1.09e-06

TABLE II. COMPARATIVE EVALUATION OF FSM BASED APPROACH OVER NOC ARCHITECTURE

Type used	Delay (ps)	Power Report
Sarkar et al,.[17]-NoC architecture	599.8 (1.667GHz)	20uW
FSM based array element	523.11(1.911GHz)	9.58uW

In the FSM approach, the total number of steps during bit wise patern comparison is constant, irrespective of the system size. This is because there is a data exchange between FSM states at every cell. Consequently, the computational energy is also proved to be effeicnt with increase in the system size. However, in ASCII based approach, the computational energy increases with the system size and delay is also get increased with system size and pattaren length. This explains the sudden rise in the total computational energy shown in Fig 5. and comparable delay reduction as shown in table II.

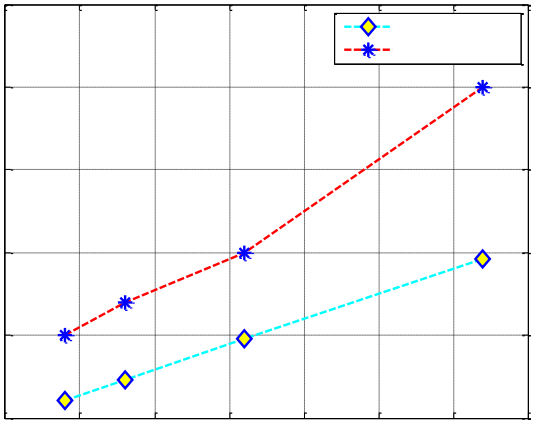

Figure5: Energy profile plot

F. Comparison of Power over Complexity overhead

Though it is well proved that decomposed systolic array approach always provides very high performance, flexibility to optimize the power over DGS_SW algorithm is not possible since multiple sequences are matched in parallel. The only possible way is to address different parts of PEs to approximate matching without using any power reduction techniques. But this will cause significant performance loss. Here through FSM states based PEs leads lesser comparison and bit transition over ASCII based matching process and we compare our FSM implementation with systolic implementations by maximum operating clock speed along with power reduction. The proposed approach reported with 13% delay reduction than conventional ASCII based PEs with

NoC architecture. Moreover, power is reduced more than 50% which is better than the most recent advance DGS_SW algorithm NoC architecture implementation.

CONCLUSION

This paper concludes the performance enhancement of protein sequence alignment implementation in order to meet power and speed constraints for next generation sequence alignment. The efficiency and optimal nature of DGS_SW sequence alignment algorithms with our FSM based bitwise matching process over score generation and matrix computation, are detailed and compared. Results revealed that the proposed technique performance is several times higher than the original conventional processing core based method and a relevant decrease of power dissipation that results when using bitwise mechanisms. Finally, we proved that scalability applied to extend the query search at a reasonable area and power dissipation cost.

REFERENCES

[1] 1000 Genomes: A Deep Catalog of Human Genetic variation. URL:http://www.1000genomes.org.

[2] Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon et al. "Initial sequencing and analysis of the human genome." Nature 409, no. 6822 (2001): 860-921.

[3] Raghunath, Bane Raman, and Shivsharan Nitin Mahadeo. "Network intrusion detection system (NIDS)." In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on, pp. 1272-1277. IEEE, 2008.

[4] Wieczorek, Dominika, Bożena Małysiak-Mrozek, Stanisław Kozielski, and Dariusz Mrozek. "A method for matching sequences of protein secondary structures." Journal of Medical Informatics & Technologies 16 (2010).

[5] Shah, Hurmat Ali, Laiq Hasan, and Nasir Ahmad. "An optimized and low-cost FPGA-based DNA sequence alignment—A step towards personal genomics." In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp. 2696-2699. IEEE, 2013.

[6] Mahram, Atabak, and Martin C. Herbordt. "FMSA: FPGA-accelerated ClustalW-based multiple sequence alignment through pipelined prefiltering." In Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on, pp. 177-183. IEEE, 2012.

[7] Benkrid, Khaled, Ying Liu, and AbdSamad Benkrid. "A highly parameterized and efficient FPGA-based skeleton for pairwise biological sequence alignment." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 17, no. 4 (2009): 561-570.

[8] Oliver, Tim, Bertil Schmidt, Darran Nathan, Ralf Clemens, and Douglas Maskell. "Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW." Bioinformatics 21, no. 16 (2005): 3431-3432.

[9] Oliver, Timothy F., Bertil Schmidt, Yanto Jakop, and Douglas L. Maskell. "High speed biological sequence analysis with hidden Markov models on reconfigurable platforms." IEEE Transactions on Information Technology in Biomedicine 13, no. 5 (2009): 740-746.

[10] Boukerche, Azzedine, Jan M. Correa, Alba Cristina Magalhaes Melo, and Ricardo P. Jacobi. "A hardware accelerator for the fast retrieval of DIALIGN biological sequence alignments in linear space." IEEE Transactions on Computers 59, no. 6 (2010): 808-821.

[11] Jiang, Xianyang, Xinchun Liu, Lin Xu, Peiheng Zhang, and Ninghui Sun. "A reconfigurable accelerator for smith–waterman algorithm." IEEE Transactions on Circuits and Systems II: Express Briefs 54, no. 12 (2007): 1077-1081.

[12] Pop, Mihai, and Steven L. Salzberg. "Bioinformatics challenges of new sequencing technology." Trends in Genetics 24, no. 3 (2008): 142-149.

[13] Do, Chuong B., Mahathi SP Mahabhashyam, Michael Brudno, and Serafim Batzoglou. "ProbCons: Probabilistic consistency-based multiple sequence alignment." Genome research 15, no. 2 (2005): 330-340.

[14] Pei, Jimin, and Nick V. Grishin. "PROMALS: towards accurate multiple sequence alignments of distantly related proteins." Bioinformatics 23, no. 7 (2007): 802-808.

[15] Yu, Chi Wai, K. H. Kwong, Kin-Hong Lee, and Philip Heng Wai Leong. "A Smith-Waterman systolic cell." In New Algorithms, Architectures and Applications for Reconfigurable Computing, pp. 291-300. Springer US, 2005.

[16] Koonin, Eugene V., and Michael Y. Galperin. "Principles and methods of sequence analysis." In Sequence—Evolution—Function, pp. 111-192. Springer US, 2003.

[17] Sarkar, Souradip, Gaurav Ramesh Kulkarni, Partha Pratim Pande, and Ananth Kalyanaraman. "Network-on-chip hardware accelerators for biological sequence alignment." IEEE Transactions on Computers 59, no. 1 (2010): 29-41.