DNA Comparison using Smith-Waterman Algorithm

Swaroopa S.Kulkarni[#1], Dr. T.B. Mohite-Patil[*2] Smita A.Patil[#3,] Megharani B. Chougule[#4]

[#]*E&TC Department, D.Y.Patil College of Engg. & Tech,Shivaji University*
KasbaBawda,Kolhapur,Maharashtra,India
[1]swaroopa_kulkarni@yahoo.co.in, [3]*Smital.patil05@gmail.com*, [4]Megha.chougule5991@gmail.com

[*]*Dr. D.Y.Patil pratishthan's college of Engg.,Shivaji University*
Salokhe- Nagar,Kolhapur, Maharashtra,India
[2]tanaji.mpatil@gmail.com

***Abstract*— This paper gives a brief description of various papers regarding concepts of DNA and DNAs comparison. The content of these papers is summarised in the literature review. After that the proposed system to overcome the drawbacks of previous work is explained. The methodology of implementation is also stated. The expected results will be tested by functional simulation. Then the system will be realised using hardware.**

***Keywords*— DNA, gene, strands, sequencing, codes, alignment, FPGA, digital logic.**

I. INTRODUCTION

DNA (Deoxyribonucleic acid) is present in all cells of a body of an individual. It acts as a molecular blueprint for the cell of that individual. DNA is most important in bioinformatics because it stores the genetic information which influences important characteristics of an organism. Hence for various applications in medicine, biotechnology and microbiology it is essential and useful to compare two DNAs. The comparison of two DNA needs the comparison of approximately 3 billion DNA base pairs, which is not an easy task. The DNA sequence can be represented as a sequence of characters from 4 alphabets [2]. The DNA sequences of hundreds of organisms have been decoded and stored in databases .In Biology this information is used to analyze the evolution of the species and to study the biodiversity. In Medicine, DNA analysis is used for many different purposes like finding correlations between a given disease and DNA information to analyze the cancer mutations and studying the evolution of viruses.

.

II. LITERATURE REVIEW:

This is a review paper, in which the views and works of researchers are elaborated. The structure of DNA, the algorithms for comparison, its implementation etc. can be found from following papers.

1. S. L. Wolfe (Ed.), 1993[1]:

In this paper the author has provided thorough information about the DNA structure. It is stated that Deoxyribonucleic acid can be called the molecule of life, because it is the chemical code specifying appearance, lineage and function. It is unique for every individual. But this paper does not give any idea of matching of two DNAs.

2. M. Canella, et.al2003. [2]:

This paper includes DNA comparison on a Bio-inspired Tissue of FPGAs. String comparison is a critical issue in many application domains, including speech recognition, contents search and bioinformatics. The similarity between two strings of lengths N and M can be computed in O (N X M) steps by means of a dynamic programming algorithm developed by Needleman and Wunsh. The algorithm can be effectively mapped onto a systolic array, resulting in a parallel implementation of the Needleman–Wunsch algorithm on a Bio-inspired wall, a giant reconfigurable computing tissue conceived to prototype bio-inspired cellular systems. The drawback is that the Bio-Wall suffers from the typical performance limitations of a large prototyping platform.

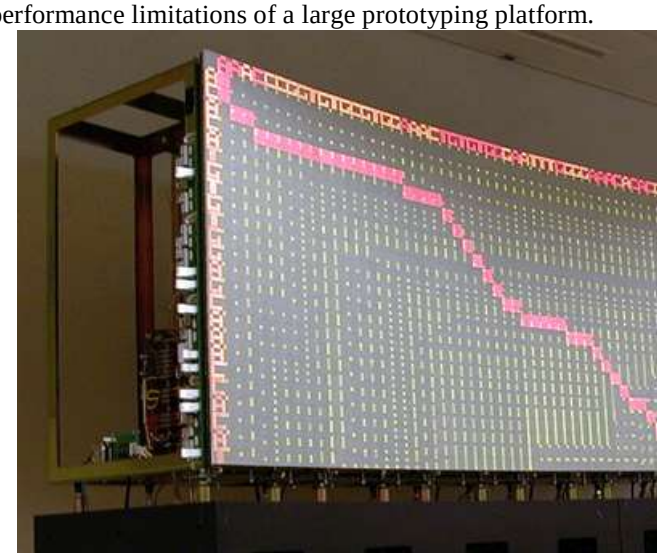

Fig.01 **Frontal view of the BioWall**

3. S. B. Needleman and C. D. Wunsch.1970 [3]:

Needleman & Wunsch **(1970)** first introduced an iterative matrix method of calculation. In this paper a computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant

homology exists between the proteins. The information is used to trace their possible evolutionary development. Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two dimensional array, and all possible comparisons are represented by pathways through the array. For this maximum match only certain of the possible pathways must be evaluated. A numerical value, one in this case is assigned to every cell in the array representing like amino acids. The maximum match is the largest number that would result from summing the values of every pathway. But, hardware implementation is not stated in the paper.

4. T. F. Smith and M. S. Waterman 1981[4]:

In this paper a metric (matrix) represents the minimum number of "mutational events" required to convert one sequence of DNA into another. It is of interest to note that Smith (1980) shown that under some conditions the generalized Sellers metric that is equivalent to the original homology algorithm of Needleman & Wunsch (1970). In this paper the authors extend the above ideas to find a pair of segments, one from each of two long sequences, such that there is no other pair of segments with greater similarity (homology). The similarity measure used here allows for arbitrary length deletions and insertions. For this purpose three basic computations are defined: addition (Insertion), Deletion (Removal), and substitution (Replacement). The mutation distance is thus the number of minimum events required to convert the one sequence into another. Also this paper does not provide the hardware implementation for the algorithm.

5. E. Fernandez, et.al .2012[5]:

In this paper the authors mention that short read alignment is a computationally intensive operation that involves matching millions of short strings (called reads) against a reference genome. At the time of writing, a representative run requires to match tens of millions of reads of length of about 100 symbols against a genome that can consists of a few billion characters. Existing short read aligners are expected to report all the occurrences of each read as well as allow users to control the number of allowed mismatches between reads and reference genome. Popular software implementations can take many hours or days to execute, making the problem an ideal candidate for hardware acceleration. In this paper, they describe FHAST (FPGA Hardware Accelerated Sequencing-matching Tool), a hardware accelerator that acts as a drop-in replacement for short read alignment software. The architecture masks memory latency by executing many concurrent hardware threads accessing memory simultaneously and consists of multiple parallel engines to exploit the parallelism available on an FPGA. The limitation of this method is that for speeding up the execution more than one accelerator engines (FPGAs) are required.

The drawbacks of above work stated in different papers mentioned above are minimized in our proposed work like to find a cost effective assembly for DNA comparison using digital logic which will be easier to use, compared to cost and complexity of supercomputers. It is described in following section.
The block diagram of the proposed work is given below:-

III. BLOCK DIAGRAM:

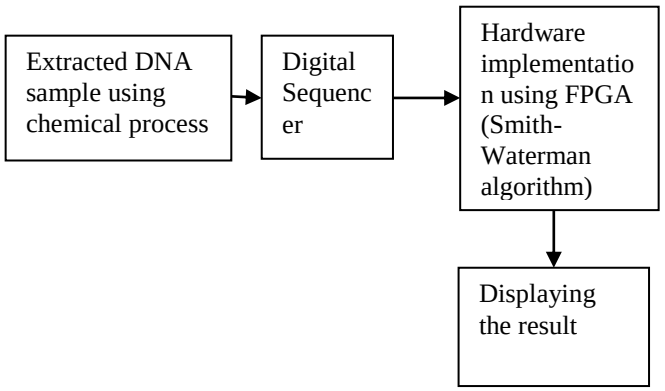

Fig. 02 Proposed system

a. **Extracted DNA sample**- DNA will be extracted by chemical processes such as mixing the grounded sample of material (for example, green peas) with liquid soap, then adding alcohol etc. This process will be carried out under the project.

b. **Digital sequencer**- The DNA sequence of the extracted sample is then generated by the sequencer. Sequencer is available in the Shivaji University, Biotechnology department. It will be used with the prior permission.

c. **Hardware implementation using FPGA**- Smith-Waterman algorithm will be implemented firstly by functional simulation and then on the hardware using FPGA

d. **Displaying the result**- The result of comparison that is, whether the DNAs matched or not will be displayed on LCD display.

The DNA sample is synthesized and the sequence is found with the help of digital sequencer. This sequence is compared with the database by hardware implementation of Smith – Waterman algorithm using FPGA. Finally we get the mutation distance between the two DNAs. And depending on it, DNAs matching results are generated. The mutation distance of the same species is very small compared to DNAs of two different species. The scope of the project will be find the DNAs belong to same species or not.

III.1 FUNCTIONAL SIMULATION:

The designs will be tested for proper functioning and desired working on simulator. The available options will be tested by hardware simulations. Functional simulation is useful because it is fast and it can be done even before synthesizing the project. The hardware delays are accurately included in the Post-Place & Route simulation at the cost of a slower simulation speed.

III.2 FLOW CHART:

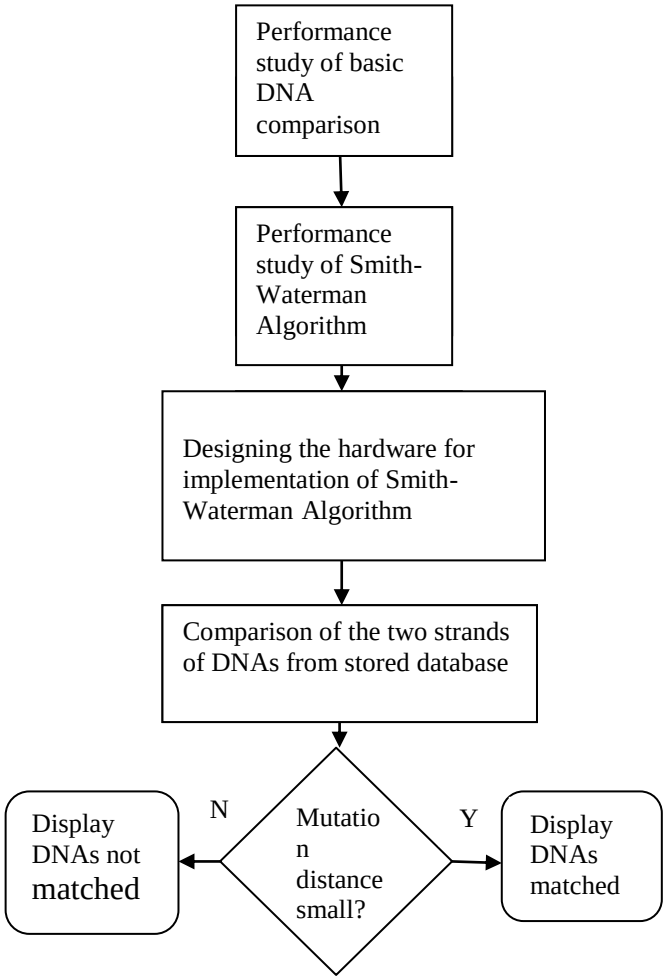

Fig. 03 Flow chart of proposed work

III.3 METHODOLOGY:

Proposed Work for Dissertation deals with following steps:

III.3.a. Performance study of basic DNA comparison
III.3.b. Performance study of Smith-Waterman algorithm
III.3.c. Designing the hardware for implementation of Smith-Waterman algorithm
III.3.d. Designing the basic cell of the matrix: the single logic block in FPGA array
III.3.e. Designing the modules to calculate minimum mutation distance, d, by comparing two DNAs
III.3.f. Functional Simulation: for performance evaluation of proposed work.
III.3.g. Performance validation using hardware
III.3.h. Comparison of two strands of DNAs from stored database.
III.3.i. Displaying the result

IV. CONCLUSION

Such type of work will prove to be beneficial to the society as many genetic diseases which are caused by faulty DNA can be detected in early stages. For example, carcinogenic genes detection at an early stage will be helpful for preventive treatment. For reducing the project cost, inexpensive options like FPGA, PLD and ASICs can be used. So the entire project will be a better option for DNA comparison at a much reduced cost and hence, beneficial for developing countries.

V. ACKNOWLEDGMENT

We offer our profound gratitude to the management of D.Y. Patil college of Engineering and technology for giving us an opportunity of exposure to use our theoretical knowledge with our practical experience, in a professional environment, and allowing me to use their labs.
We are also thankful to Prof. Dr. S. P. Govindwar and Prof. Dr. A.K. Sonawane from Shivaji University, Kolhapur for allowing me to share their labs.

VI. REFERENCES

[1]S. L. Wolfe (Ed.) "**Molecular and cellular biology**." *Wadsworth Publishing Company*, 1993.

[2]M. Canella, F. Miglioli, A. Bogliolo,E. Petraglio, E. Sanchez, "**Performing DNA comparison on a Bio-inspired Tissue of FPGA***"EPFL-LSL*, Lausanne(Switzerland)IEEE 2003.

[3] S. B. Needleman and C. D.Wunsch. "**A general method applicable to the search for similarities in the amino acid sequences of two proteins**",*J. Mol. Biol*.,48:443-453,1970

[4] T. F. Smith and M. S. Waterman,"**Identification of common molecular sequences***",J. Mol. Biol*.,147:195-197,1981

[5] E. Fernandez, W. Najjar, S. Lonardi, J. Villarreal, *"* **Multithreaded FPGA Acceleration of DNA sequence Mapping***", Riverside*, USA,IEEE 2012

