Journal of Advanced Computing Research Vol. 3, Issue 2 (2018) 1-4

Application of Parallel Processing - A Case Study on DNA Sequence Analysis

Azera Zaid [#1], Mohamed Faidz Mohamed Said [#2]

[#] *Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA*
70300 Seremban, Negeri Sembilan, MALAYSIA

[1] azerazaid@gmail.com

[2] faidzms@ieee.org

***Abstract*—A large portion of the analyses for DNA successions is based on the atomic biologists' requirement over discovering the likenesses between different particles or in one particle. In order to apply method of DNA sequence analysis, a program in parallel processing algorithms has to be written. It generally deals with the present issues of a medium-sized PC framework. A significant number for data that are not promptly accessible need to be prescribed with new strategies to discover homology and request over successions. For analysis of DNA sequences, there are an efficient design of computing system of a distributed bioinformatics computing system by using OPTSDNA algorithm. It is used to detect disease as well as criminal case. This algorithm stores several sizes of DNA sequence into database. There are also new methods in developing and analysing DNA by using DSP techniques. To use this technique, DNA sequences must be converted into numeric sequences. In conclusion, there are a lot of methods in designing computing system to analyse DNA sequence analysis.**

Keywords: **homology, bioinformatics, computing system, OPTSDNA algorithm, DSP techniques**

I. Introduction

In the past few years, the amount for the existing living information about sequencing and genomes requires exponential development. Previously, those amount incorporated in the GenBank database from claiming nucleotides require multiplied harshly each 18 month, making a size of 100×109 build pairs (bp) [1].

Generally, starting with analysing living sequences, parallel strategies are made to diminish the time needed on attaining two computationally escalated consideration analyses which are homologous sequence searching and multiple sequence alignment. Those parallel seeking system cuts those recovery durations by nearly a factor of N, the place N will be the number about processors [2].

To recognize those issues that have a chance to be solved, the sort of information must be distinguished in preparing the DNA sequencing and how it is processed. Those DNA holds on some chromosomes, which need aid in length successions for nucleotide bases. Those four nucleotide bases need aid named similarly as A, C, T and G. The human DNA comprises of 3 billion of nucleotide base pairs. Those bases encode those genes of a human organism, and the request about the individuals' bases is the data searched by the DNA sequencing process. To provide human DNA, this procedure cannot attain and clinch alongside one shot. Finally, it is the differences in the sequenced DNA contrasted with those in the reference grouping because they can show a disease-causing mutation.

Next-generation about DNA and RNA sequencing will settle on and should take a gander at independent genomes, as well as should think about hereditary successions and additionally various genomes. These techniques also could be used to figure out the varieties done genomes and gene transcripts starting transcripts from person to person, around populations, and the middle of ordinary and pathologic cells in cancer. While next-generation sequencing advances have made it practical to prepare high-positioning genomic information a great part more effectively, scientists in different fields have transformed and built the next-generation sequencing to find particular features of the genome that impact specific phenotypes [3].

Molecular biologists fundamentally attempt to discover the likenesses between different molecules and in one molecule. There are some different methods that have been applied to solve this problem. The most exciting methods are those which have the capacity to guarantee and determine the best comparability between successions. To determine those likenesses in DNA sequences, it can explore the applicability of an 1CL Distributed Array Processor (DAP) to the general problem [4].

DNA sequence analysis is currently a main research topic amongst mathematicians, computer scientists and physicists. Digital Signal Processing (DSP) could provide a possibility as an important field of science and engineering that has been improved as a result of the constant evolution of computer science and technology [5].

To accomplish computing DNA atoms may be used. The DNA registering data is spared over DNA atoms. An original set of molecules usually contains all correct and wrong solutions. For an example, each vessel includes around 1020 DNA strings. DNA computing might act as a set of processing phases on DNA molecules for solving a specific problem according to an exactly defined procedure [6], [7].

Furthermore, in using heuristics, there are a few works proposed which are without match-miss for approximate pattern searching [8] and they should gap suffix tree of the DNA sequence and each sample sub-tree might be constructed to real memory best.

A. Definition

DNA (Deoxyribonucleic Acid) is characterized as the fundamental unit of an organism. Its period arrays also start with a couple hundred to a few billions for nucleotides for different species. Therefore, to obtain the similitude and the degrees of distinction, successions may be a confounded assignment. It should be obvious that the DNA succession alignment, particularly a few arrangement alignments, will be a profitable DNA exploration subject in terms of the bioinformatics. The outcomes for different arrangements can be used for other complicated genetic research. For example, the outcomes of numerous succession arrangements can be utilized to evaluate those degrees of alignment and to figure out the degrees of similarity of different species as well as to perceive the evolutionary historical backdrop for species. This alternately put a recently discovered species under a group that might have a place with [8].

DNA sequencing can be defined as the process of determining the specific order of nucleotides in a DNA molecule. It contains some methods that are used to define the order of the four bases - adenine, guanine, cytosine and thymine [9]. The introduction of this DNA sequencing methods has accelerated biological and medical study and detection [10].

DNA atoms might have a chance to be characterized as a linear polymer built with chemical bonds from four building blocks. There are nucleotides denoted by symbols A, C, G and T. Since DNA polymers are committed from claiming four nucleotides, they represent chains of symbols over 4 alphabets. Therefore, DNA computing may be satisfactory for transforming images as well as legitimate structures [11].

The GenBank succession database could be characterized as open access, annotated collection of each publicly existing nucleotide sequences and their protein conversions. This database is shaped and looked by the National Centre for Biotechnology Information (NCBI). Similarly, as a component of the International Nucleotide Sequence Database Collaboration (INSDC). The National Centre for Biotechnology Information is a part of the National Institutes of Health in the United States. GenBank and its collaborators accept sequences produced in laboratories throughout the world from more than 100,000 different organisms. In more than 30 years since its formation, GenBank needs to be the most significant and most powerful database for research in almost all biological fields, whose data are retrieved and cited by millions of researchers around the world. GenBank continues to grow at an exponential rate, doubling every 18 months [12].

II. Background

The sequencing of DNA molecules started in the 1970s with the development of the Maxam Gilbert method, and next continues with the Sanger Method. Mostly Sanger method is used in medical and research laboratories. In the beginning, DNA sequencing was first developed by Frederick Sanger in 1975. Consequently, in 1953, DNA double helix was discovered by James Watson and Francis Crick. In 1965, the first nucleic acid molecule that has been sequenced by Holly is Escherichia coli alanine tRNA. While in 1970 Hamilton Smith discovered type II restriction enzymes. Then it is continued in 1977 by Maxam-Gilbert on his chemical degradation and Frederick Sanger on his dideoxy termination. In 1983, Polymerase chain reaction (PCR) developed by Kary B. Mulis is revolutionary method that allows scientists to rapidly amplify DNA [13].

III. Paper Review

There are several paper reviews discussing about DNA sequence analysis:

A. Processing DNA Tokens in Parallel Computing

Reference [6] displays another technique for sending information between sub-atomic processors. Sub-atomic processor is a preparing information unit. The calculation should be sent to different units by utilizing the type of data to the messages which is known as tokens. Fundamental trials were refined. All operations were connected in DNA. Consequently, this strategy sends information to more than one processor. This is called communicate sending, that is to all processors, and multicast sending, that is to many, yet not all processors.

B. Study of DNA Sequence Analysis using DSP Techniques

DSP strategy or it is called as Digital Signal Processing (DSP) is an application in genomic succession investigation. The DNA groupings must be changed into numeric successions when investigating DNA arrangements. At that point, the DSP calculations are utilized as a part of DNA examination. There are a portion of the imperative DSP ideas utilized for DNA examination which are Discrete Fourier Transform (DFT), Discrete Wavelet change (DWT), advanced sifting, Parametric displaying and entropy. This technique is utilized to get data from genomic and proteomic information to build models of atomic natural frameworks. It additionally helps in growing new analytic instruments, remedial methods and pharmacological medications for applications like malignancy characterization [5].

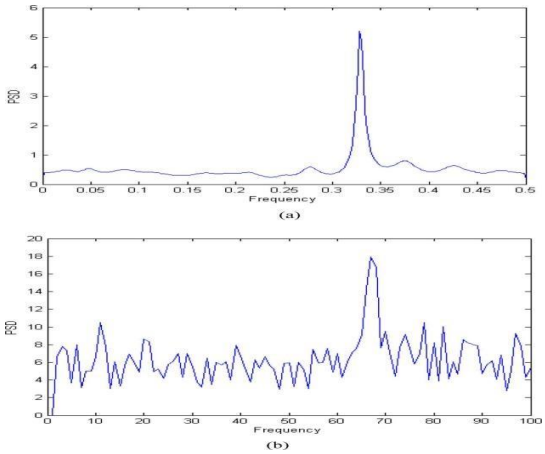

Fig. 1. Parametric model [5]

C. *Solving DNA Sequence Assembly using Particle Swarm Optimization with Inertia Weight and Constriction Factor*

A powerful method to clarify the DNA succession issue is by utilizing a fluctuation of the standard Particle Swarm Optimization (PSO) called the Constriction Particle Swarm Optimization (CPSO) that was presented in [14]. To take care of the DNA grouping gathering issue, it affects the molecule swarm utilizing a latency weight and a narrowing variable with Smallest Position Value (SPV). The narrowing variables proposed in this work is to ensure the exactness of union of the molecule swarm calculation and fine tune the pursuit. The DNA grouping gathering issue measured in this paper is to build the covering score. It has been uncovered through different trials that the CPSO technique beats the other known strategies as well as PSO based strategy, in terms of uncommon quality arrangement, consistency, speedier joining and precision.

D. *Distributed Bioinformatics Computing System for DNA Sequence Analysis*

Reference [14] gives a productive plan of registering arrangement of a dispersed bioinformatics figuring framework for investigation of DNA successions by utilizing OPTSDNA calculation. It can be utilized for distinguishing ailment, quality forecast, criminal legal investigation, hereditary framework and protein examination. This calculation is used for gathering diverse sizes of DNA succession into database. These information DNA successions are different in size from little to vast. Henceforth, it can be presumed that the intricacy of the calculation rises the response time. The calculation for the Pattern Identification was the extremely complex one and the calculation for the example pursuit was the slightest complex.

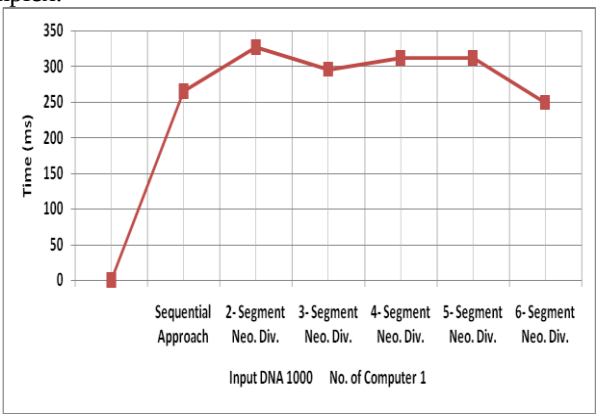

Fig. 2. Effect of data size on using single computer [14]

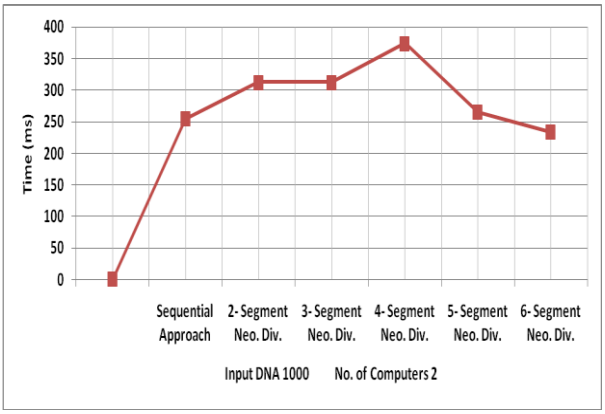

Fig. 3. Effect of data size on using two computers [14]

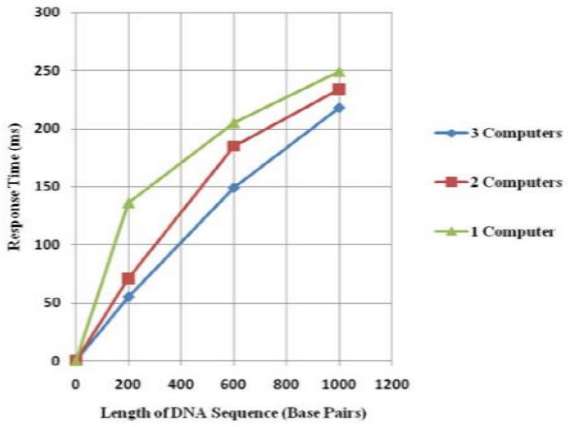

Fig. 4. Effect of data size on computation time [14]

E. *A Parallel Programming Framework for Multi-Core DNA Sequence Alignment*

Reference [1] demonstrates an alternate parallel programming system for DNA arrangement in homogeneous multi-centre processor models. This structure is established on the use of tiling techniques that splits the question arrangement and every database grouping in littler lumps that are simultaneously handled by a few CPUs. The led trial techniques have confirmed that significant increasing velocities of the considered succession arrangement calculations can be proficient. For the ideal SW calculation, the trial speedup is straight with the quantity of existing preparing centres and near the hypothetical most extreme. For the quicker and problematic strategy, the proposed stage has ended up being able to further lessen of the natural handling time.

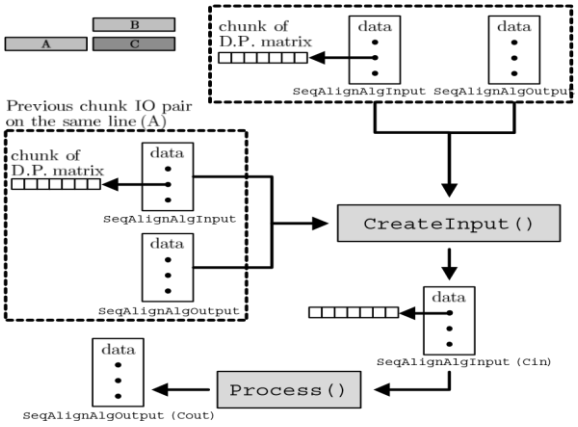

Fig. 5. Chunk *C* processing using the data from neighbours *A*&*B* [1]

IV. Conclusion

By using so many different methods and models emerging from the current research, it can be concluded that DNA computing can be more accurately described as a collection of new computing paradigms rather than a single focus. Each of these different paradigms within the bio-molecular computing can be associated with different potential. The use of DSP techniques is to get data from genomic and proteomic information to construct models of sub-atomic organic frameworks. Henceforth more profound comprehension of the structure and elements of living frameworks will be obtained further. This will assist in growing new indicative apparatuses, remedial strategies and pharmacological medications for applications like disease order and expectation. DAP projects can be connected to this issue. However, the best enthusiasm for this kind of examination is in its application to generally short protein successions, and the DAP approach might be especially profitable if it can be produced to make concurrent inter-comparisons of expansive numbers of successions.

References

[1] T. Almeida and N. Roma, "A Parallel Programming Framework Formulti-Core DNA Sequence Alignment," n.d.

[2] T. K. Yap, O. Frieder, S. Member, IEEE, Robert L. Martino, and Member, "Parallel Computation in Biological Sequence Analysis," *IEEE Transactions on Parallel and Distributed Systems,*, vol. 9, no. 3, 1998.

[3] "Columbia Genome Center website [Online]." https://systemsbiology.columbia.edu/genome-center/sequencing-and-analysis (accessed.

[4] J.F.Collins and A.F.W.Coulson, "Applications of Parallel Processing Algorithms for DNA Sequence Analysis," *Nucleic Acids Research*, vol. 12, no. 1, pp. 181-192, 1984.

[5] I. T. M and S. R, "Study of DNA Sequence Analysis using DSP Techniques," *Journal of Automation and Control Engineering*, vol. 1, no. No 4, 2013, doi: 10.12720/joace.1.4.336-342.

[6] R. Nowak, PiotrWasiewicz, J. J.Mulawka, and A. Płucienniczak, "Processing DNA Tokens in Parallel Computing," 2001.

[7] P. Frisco. "The Collection of Computer Science Bibliographies website [Online]." http://liinwww.ira.uka.de/bibliography/Misc/dna.html (accessed).

[8] K. Ahuja and Kompal, "Analysing Multiple DNA Sequence Alignment Algorithms- Smith Waterman Algorithm and Parallel Smith Waterman Algorithm,"*International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 2, no. IV, 2014.

[9] A. Zaid. "Application of Parallel Processing: A Case Study on DNA Sequence Analysis" https://www.youtube.com/watch?v=UpCk_bPvBCw (accessed.

[10] "The Wikipedia website [Online]." https://en.wikipedia.org/wiki/DNA_sequencing (accessed).

[11] PiotrWasiewicz *et al.*, "DNA computing: implementation of data flow logical operations," *Future Generation Computer Systems*, vol. 17, pp. 361-378, 2001.

[12] "The Wikipedia website [Online]." https://en.wikipedia.org/wiki/GenBank (accessed).

[13] "SlideShare website [Online]." https://www.slideshare.net/UsmanAyub6/lecture-no-19 (accessed.

[14] M. I. Khan, K. Deb, and C. Sheel, "Distributed Bioinformatics Computing System for DNA Sequence Analysis," *Global Journal of Computer Science and Technology: A Hardware & Computation*, vol. 14, no. 1, 2014.