A Survey of DNA Mapping Technique

Prashant Chaturvedi

Computer Science & Engineering, (High Performance Computing Solution)
Vel-Tech University, Chennai, Tamil Nadu, India
C-DAC, Pune, Maharashtra, India

Abstract— DNA Sequencing is a challenging process where we determine and identify every single DNA base and element that is in the genome of an individual. There are six billion of those in every normal cell in every person..In this paper, we discuss impediments and future works about Hadoop in bioinformatics. We study the MapReduce algorithm from algorithm lay by point and demonstrate the appropriates of our approach by tracing and analyzing efficient MapReduce algorithms for sorting and simulation problem of parallel algorithms specified with the help of pigeonhole principle.Approach: 1, 2.we can get general statistical analysis through R language. After studying the survey paper various approaches of using GPU and MapReduce,we adopted the best solution to use R with MapReduce. An R package is created to shift a set of critical R functions on GPU card. It allows users to run R code with GPU spread that enable much faster large data set of computation.

Keywords— Big Data, Approximation Algorithm, Pigeonhole Principle, NVIDIA card.

I. INTRODUCTION

Bioinformatics department is help of solve the biologists computational problems on purpose confront large amount of data. Recently, computing and sequencing ability has improved throughout the processor.

DNA is the hereditary building in all organisms. DNA resides in every cell in the body of organism. The double helix looks an immensely long step twisted into a helix.

The sides of the step are formed by a backbone of sugar and join many phosphate molecules and the pole consist of nucleotide bases join in the middle by the hydrogen bonds. Thisis four types of nucleotides. Each nucleotide contains a base:

- ✓ Adenine (A)
- ✓ Guanine (G)
- ✓ Cytosine (C)
- ✓ Thymine (T)

Base pairing form naturally between A and T and between G and C at normal data pattern therefore that basic sequence isa single strand of DNA each parts of DNA which can be simplydeduced from that of its partner strand.

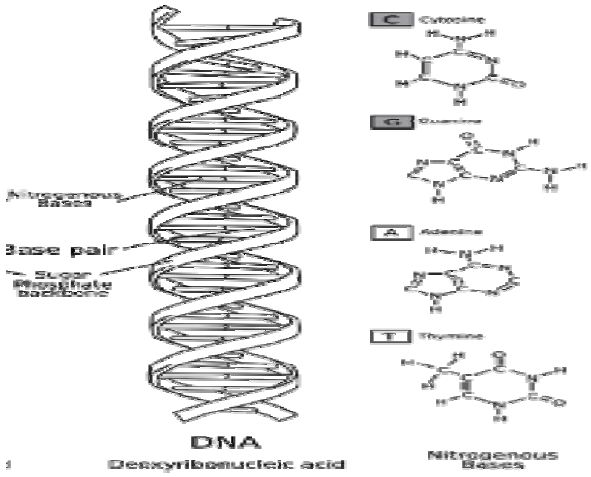

Fig. 1: DNA Structure with Molecular Structure (source: google Image)

DNA is the chemical liquid which carrying instructions to cell. When those instructions have mistaken that cells function normal or not. These cancer causing changesin DNA sequence. It is called mutations. It can cause cells to produce the wrong amount of a certain protein. In cancer many tissues are damaged.

If mutation is changed in our DNA building. A gene mutation is a change in gene or damage to a gene. These changes in your DNA can result in genetic disorders.

Mutations can lead to missing or deformed proteins and that can lead to disease.

Types of mutations:

- Deletion
- Substitution
- Inversion
- Insertion

In this disease this caused changes the human DNA structure and mutated in an individual's DNA sequence.These mutations can create some subsequence an error in DNA replication due to environmental factors such as cigarette smoke, alcohol and divestment to radiation which cause changes in the DNA sequence. Our DNA provides the code for making proteins.

Another type of genes maintains the integrity of genes and provides the accuracy in the information transfer from one

gene to another. XRCC3 gene being a DNA mismatch repair gene responsible for skin cancer

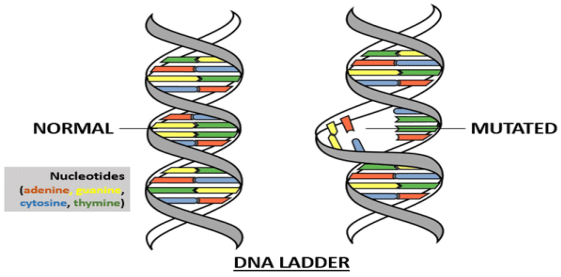

Fig.2: compares normal DNA structure to mutation DNA structure (source: google Image)

II. RELATED WORK

In the past many algorithm have been applied for merging DNA sequencing like an exact algorithm Approximate, Inference Algorithms, Longest common substring(LCS), Longest common subsequence(LCSS), Mappers Algorithm, Greedy algorithm, Construction, Best CODON algorithm, Genetic Algorithm (GA) etc. *Et al Ravichandran* said that the mapping is using waveforms, such as Gaussian functions, with unique sequence representations in the time-frequency plane. This proposed alignment method employs a robust querying algorithm that utilizes a time-frequency signal expansion whose basis function is matched to the basic waveform in the mapping sequences. *Edward B. Fernandez* said that this is using a FPGA Hardware Accelerated Sequencing-matching Tool for find DNA sequencing mapping. All this method discussed below

An exact algorithm- Using these ideas of processing fragments in order, we can also design an exact algorithm that computes an optimal orientation.

Approximate Inference Algorithms-A numbers of general purpose algorithms have been developed in the past. General purpose refers to the fact that these algorithms are not tailored to a specific probabilistic model. Instead, they can be adapted to different models, and graphical models provide a formalism facilitating these adoptions, inference algorithms can be classified into exact and approximate algorithms. Exact are all algorithms that always deliver the correct solution.

Greedy algorithm-This algorithm considered above is both fairly complicated and time consuming because they require the calculation of maximum and minimum weighted matching.

Best CODON algorithm-In the work by Ravi Vijaya Satya et al they have discussed about a pattern matching algorithm for Codon Optimization. Here in this paper, we have only concentrated on the part to find out the best triplet codon for codon optimization. Given a DNA sequence as input, we are finding all possible codons from the sequence. From the codons, we can find the one which has the highest frequency or maximum occurrence. We call those codons as the 'Best Codon'. The algorithm is as following:

Input: DNA sequence (S) ∑N = {A, C, G, T} Dictionary D = {codon, freq}

Begin:

```
fori in Sequence S:
        codon = S[i], S[i+1], S[i+2]
if codon not in D:
        D[codon] = 1
 else:
      D[codon] = D[codon] + 1
i = i + 1
      Sort dictionary D w.r.t freq
End
```

Output: Top element in dictionary D

In this paper,we have tried best and fast sequencing, readable and fast moving data. For this we propose to useHadoop platform for DNA sequence.

Comparative analysis is shown in Table 1

S.N no.	Referred paper	Author	Explanation	Conclusion
1.	Accurate Sequence Alignment using Distributed Filtering on GPU Clusters-2011	Reza Farivar, Shivaram Venkataraman, Yanen Li, Ellick Chan, Abhishek Verma, and Roy H. Campbell	The read length we use in our implementation is 30 base pairs. A next-generation sequencing machine can create up to 6 GB or around 180 million reads per day as of 2010.	With the growing importance of Next Generation Sequencing technologies will fast sequence query systems are necessary to handle the growing volume of information collected.
2.	Heterogeneous Cloud Framework for Big Data Genome Sequencing JANUARY/F	Chao Wang, Xi Li, Peng Chen, Aili Wang,	There are quite a lot of successful short read mapping software	This approach could bring significant speedup for

	EBRUARY 2015	Xuehai Zhou, and Hong Yu	tools that address the problem of processing the enormous amount of data produced by the next-generation sequencing machines.	genome sequencing alignment process. We have presented results both from a theoretical analysis and real hardware platform on Xilinx FPGA development board
3.	Combining Hadoop and GPU to Preprocess Large Affymetrix Microarray Data 2014	Sufeng Niu1, Guangyu Yang2, Nilim Sarma1 , Pengfei Xuan2 , Melissa C. Smith1 , Pradip Srimani2 , Feng Luo2	We tested our tools using four microarray datasets from the Gene Expression Omnibus (GEO) database.	We developed a new toolset that combined the Hadoop with GPGPU/ CUDA for the microarray data quality assessment and pre-processing.
4.	Fast CPU-Based DNA Exact Sequence Aligner	Aryan Arbabi, MiladGholami, Mojtaba Varmazyar, ShervinDaneshpajouh,	This content was to develop a fast design for a part of DNA sequence alignment problem within a time span of one month. For a given human reference genome and millions of short read sequences.	In this contest, the input data are assumed to be in binary format. Our hash based method benefits from this assumption as reading binary file is faster than text file and 64 bit comparisons is faster than character by character comparisons.

III. PROPOSED METHODOLOGY

A. Approximation Algorithm

Steps to perform DNA mapping

- ✓ Input DNA Dataset into system
- ✓ Extract DNA Dataset
- ✓ Generate Block to place the element
- ✓ Apply Pigeonhole Principle to search, insert and delete element in DNA structure to store pairing of element into each block.
- ✓ Through Hamming distance algorithm find mismatched element in dataset.
- ✓ Pattern matching algorithm will be applied to find out various combination of alphabetic
- ✓ Compare DNA structure sequence and pattern with original dataset
- ✓ Calculate percentage of diseases using ClustalW tools
- ✓ Statistical analysis to the professionally.

B. Code Flow

Let ARR is an array with N elements. Num is the searching element and LOC is the location of searching element.

1. *Begin*
2. *Real ARR and NUM*
3. *Set found = 0*
4. *Set i=0, Loc=0*
5. *Repeat slep(6) while i<=N-1*
6. *if ARR[i]= NUM*
   than Set found =1
   and Set loc =i
   and break the loop.
7. *if found =1*

then Write : Element <NUM> found at position <Loc>
else Write : Element <NUM> not present in ARR

8. *return*

C. Flow Chart

To further improve the computation time since the chromosomes are independent of each other all the individual processes are expedited by parallel programming. Flow Chart of approximation algorithm processes is shown in Figure-3

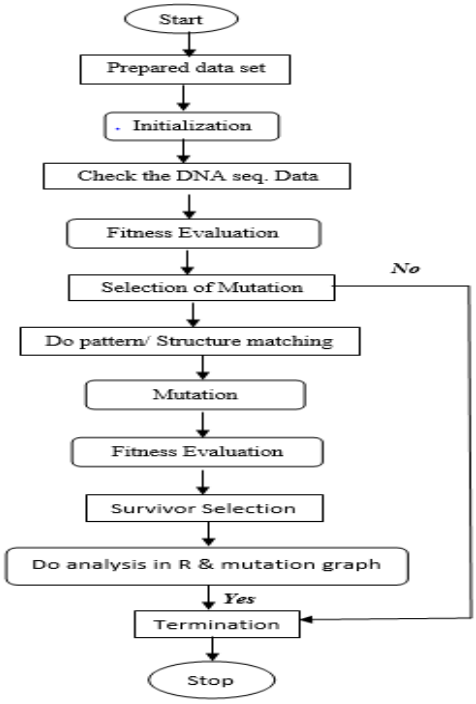

Fig.3 approximation algorithm processes

D. IMPLEMENTATION

The Hadoop used in this research work which is process through Pigeonhole and Hadoop Algorithm (PHA). The DNA data and sequencing is a very large sequencing process. So this process cannot bydone manually. Many processes are existing in research center. But that technology takes more time. Therefore, we use the new Hadoop Technology.The input data will follow the mentioned process. Firstly data will be prepared for the initialization. This initialization process check the DNA sequencing data with C coding in pigeonhole algorithm. Then data fit on block one by one. Then again check the mutation.If mutation available in that data so it is select the mutation data then send data another technology which are map the data through the Mapper tasks are select data to keep keys and values pairs which is defined mapper tasks then sorting the data in pieces in local host. That output is go to shuffle and merge the data means this process is keep separate key and value pairs It is keep number and alphabetic key separate. Then next process will reduce the data in reduce tasks process. The reduce task process is the reduce size of data and delete the unwanted sequence and get output. If there is not any mutation data so process already terminates here and stop the process. Ahead process is not do in pigeonhole algorithm. Then data go to check the DNA sequence pattern and do matching. The mutations keep separate and again fitting in the block. Then do the analysis in R language.

IV. CONCLUSION

From the six papers presented for survey, it describes the advanced computational capability achieved through different algorithm. As of future work, we will use approximation algorithm with help of pigeonhole principle. The R analytics software will executive data fast and find DNA disorder into bits. These algorithms will sort extremely complex and unstructured data. Studies many papers presented for survey it describes the advanced computational capability achieved used different algorithm. In this Presentation we presented a string short sequence DNA data will separate alphabetic map the all element and read easy using Hadoop MapReduce algorithm. Approximate algorithm in Pigeonhole Principle to search and mismatch string of nucleotide in DNA structure to store pairing of element into each block. As of future work if molecular science has brought up new study and results about cancer genes for other types of cancer disease. This work can be extended for all types of cancers. So we will create new tools which will do fast alignment sequencing data with searching, insert and delete element which used base on pigeonhole algorithm. These algorithms will sorts extremely complex and unstructured data arrange easily and analysis the percentage of cancer disease.

V. ACKNOWLEDGMENT

I would like to thank my C-DAC and Vel Tech University for providing a good environment and facilities like access to the IEEE access which was very helpful to me in this research.

REFERENCES

[1] Izzat Alsmadi and Maryam Nuser, String Matching Evaluation Methods for DNA Comparison, Vol. 47, October, 2012, International Journal of Advanced Science and Technology

[2] LI Xu-bin, JIANG Wen-rui, JIANG Yi, ZOU Quan*, “Hadoop Applications in Bioinformatics”, 2012 7th Open Cirrus,Summit,978-0-7695-4908-8/12$26.00 © 2012 IEEE DOI 10.1109/OCS.2012.40

[3] Gang Liao, Longfei Ma, Guangming Zang, Lin Tang, Parallel DC3 Algorithm for Suffix Array Construction on Many-core Accelerators

[4] Aryan Arbabi, Milad Gholami, Mojtaba Varmazyar, Shervin Daneshpajouh, Fast CPU-Based DNA Exact Sequence Aligner, 978-1-4673-1313-1/12/$31.00 ©2012 IEEE

[5] Da Li, Michela Becchi, Multiple Pairwise Sequences Alignments with Needleman-Wunsch Algorithm on GPU

[6] Chad Nelson, Kevin Townsend, Bhavani Satyanarayana Rao, Phillip Jones, Joseph Zambreno, Shepard: A Fast Exact Match Short Read Aligner

[7] *SOPHIE SCHBATH,1 *VE ´ RONIQUE MARTIN,1 MATTHIAS ZYTNICKI,2 JULIEN FAYOLLE,1 VALENTIN LOUX,1 and JEAN-FRANC ¸ OIS GIBRAT 1, Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis,JOURNAL OF COMPUTATIONAL BIOLOGY, Volume 19, Number 6,2012# Mary Ann Liebert, Inc. Pp.796–813, DOI: 10.1089/cmb.2012.0022

[8] U. Manber, "Finding similar files in a large file system [C/OL]", In: Proceedings of the Winter USENIX Conference, (1994), pp. 1-10.

[9] Wei Wang, Juan Liu* "Distinguishing Single-Stranded and Double-Stranded DNA binding Proteins Based on Structural Information", 978-1-4799-1310-7/13/$31.00 ©2013 IEEE, 2013 IEEE International Conference on Bioinformatics and Biomedicine.

[10] Ka Kit Lam and Nihar B. Shah, Towards Computation, Space, and Data Efficiency in de novo DNA Assembly: A Novel Algorithmic Framework.

[11] Gang Liao, Qi Sun, Longfei Ma, Zhihui Qin, GPUAccelerated Multiple Deoxyribose Nucleic Acid Sequence Parallel Matching , arXiv:1303.3692v1 [cs.DS] 15 Mar 2013

[12] Wei-Chun Chung∗†‡, Yu-Jung Chang∗, D. T. Lee∗‡§, Jan-Ming Ho∗†, Using Geometric Structures to Improve the Error Correction Algorithm of High-Throughput Sequencing Data on MapReduce Framework, 2014 IEEE International Conference on Big Data, 978-1-4799-5666-1/14/$31.00 ©2014 IEEE

[13] LI Xu-bin, JIANG Wen-rui, JIANG Yi, ZOU Quan*, Hadoop Applications in Bioinformatics, 2012 7th Open Cirrus Summit, IEEE Computer Society, 978-0-7695-4908-8/12 $26.00 © 2012 IEEE, DOI 10.1109/OCS.2012.40

[14] Cancer Genomics: What Does It Mean for You?, The Cancer Genome Atlas (TCGA), NIH Publication no. 10-7556 Printed July-2010

[15] Snehal P. Adey, Dr. Vandana Inamdar, GPU Accelerated Pattern Matching Algorithm for DNA Sequences to Detect Cancer using CUDA,Department of Computer Engineering and Information Technology