ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Computational Gene Analysis and Mutation Processing Using Hadoop

Arathi, Ankitha K

4th Sem MTech, Dept. of CSE., SCEM, Visvesvaraya Technological University, Adyar, Mangaluru, India

Assistant Professor, Dept. of CSE., SCEM, Visvesvaraya Technological University, Adyar, Mangaluru, India

ABSTRACT: Genes are the most important molecular unit of living organisms. It's the basic of life. The knowledge of their functions and annotations are essential in understanding physiological and pathological processes. It's the most essential component in the process of drugsand therapies development. But discovery of these are often time consuming, expensiveand mostly inaccurate mostly since these are not often revisited before their publications.Genes are in-fact collections of DNA's. These DNA are in turn constituted of Adenine(A),Thymine(T), Guanine(G), Cytosine(C).Collection of genes form chromosome and collectionof chromosomes in turn form a protein and its proteins that infer characteristics of an organism. And mutation is the process wherein there's change in the usual constituents of a gene.The proposed system aims at the computational analysis of genes. The system uses HADOOP technology for data storage that enables better search, transfer and store of data. Also provides details on the possible mutations. The systemproviding and efficient methods to extract and predict reliable gene functions and possiblemutations.

KEYWORDS: gene analysis, mutation, Hadoop, cloudera

I. INTRODUCTION

Genes are the basic unit of life. Like bits are for computers so are genes for a living organism. If we consider an algorithm, it basically has constituent functions and the functions have identifiers and these identifiers are in-turn formed from collection of bits. [1] Similarly if we consider a particular character of a living being (say the color of eye) it's basically derived from its constituent proteins. These proteins are formed from collection of chromosomes and the chromosomes in-turn are formed from collection of genes. Collection of Genes in-fact forms a DNA. These DNA are in turn constituted of Adenine (A), Thymine (T), Guanine (G), Cytosine(C), Figure 1.1 portrays this.

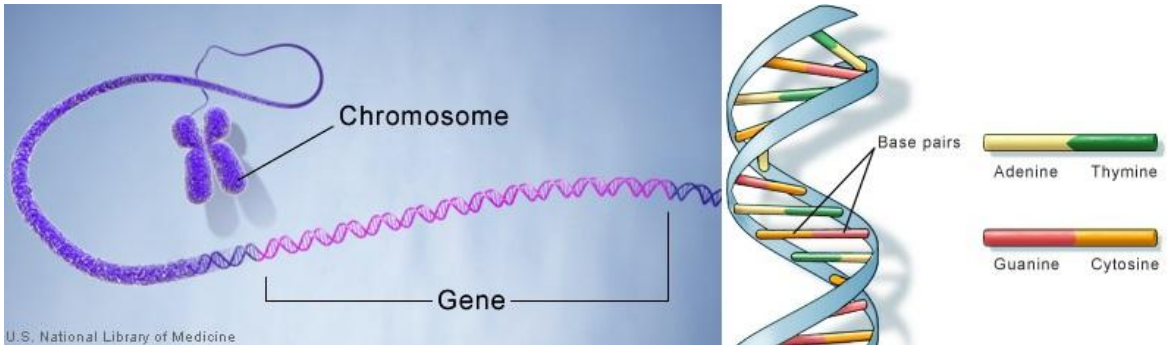


Fig 1: Genes

The nucleotide A can pair with T and G can pair with C [2]. The corresponding mRNA have the same combination but with one variation that A pairs with U. When these pair, three consecutive combinations form the three codon that yields to a protein formation. In this project we such for a particular characteristic that exist in a particular organism. For this this provide the specific gene annotation and search for its presence in the entire source DNA data of that organism. This is a complex task since the source data is very huge. Even a small bacteria can have a DNA dataset of near about 1TB i.e. a "Big data" [3]. This makes the data analysis and study very cubersome and error prone. Also the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

data storage is yet another hurdle. To tackle these two problems HADOOP and HIVE technologies are being made use of in this project. **HADOOP** provides a platform to handle the Big-data storage. It splits the data into several sub jobs and executes it in parallel providing a reliable data storage and access.

Mutation is basically in short something that is different from "normal" terms. Genes have a specific order of pairing and formation (A-T, A-U, and G-C)[4] [12]. In mutation this is not followed. Mutation can be of three types: Insertion, Deletion and Substitution [14]. Insertion is when the DNA and mRNA's bond and an extra nucleotide comes in between the strand. Deletion is when a gap is left between the strand and substitution is when a constituent nucleotide is substituted by some other nucleotide. In this project we account for all possible mutation for a specific DNA annotation.

The proposed system provides details on the intended gene annotation and its possible mutations. Every organisms have various characteristics and the proposed system helps in the data analysis of whether a specific character's gene annotation is present or not. It helps in determining whether a particular annotation is present or not. A gene annotation is given as input and searched up in a source file. The single stranded DNA data is used as input and thereby only forward and reverse search is sufficient rather than diagonal matrix searches as in case of multi-stranded DNA. Access and storage of genomic data is handled in an efficient manner. The data that we are dealing with ranges from several GB's to TB's making it a Big Data [5]. HADOOP system is used for data access and storage making it very reliable and less time consuming. The system provides basis for various research and medical fields like gene therapy(handling disease with the help of gene sequence substitution etc. rather than treatment through drugs) and gene consultations(helps for patients who have genetic disorders) etc. where we require proper analysis and details of a specific gene annotation. Genetic research is a field that is undergoing varied experiments. Genetic research deals with projects like the 1000 Genome Project where the entire gene sequence of humans where predicted with the help of just basic initial data's.

In the proposed system the source data is of AT [7] for the initial system testing. As experiments in new drug designs are tested upon rats similarly the molecular biological experiments are tested on a basic plant AT [16]. Later on various other organisms can be added as per the need be.

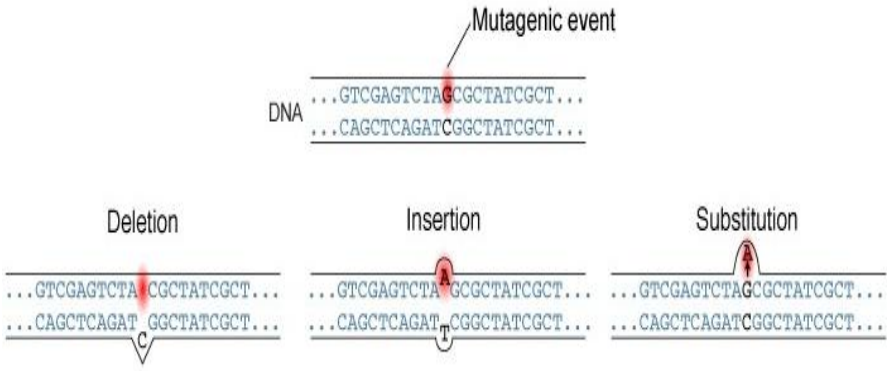


Fig 2: Mutation

Mutation [11] [15] is an equally important field as is gene analysis. Mutations can be a boon when in terms of gene therapy where we introduce new gene sequence into a patient that can provide immunity rather than diagnosis through drugs. But mutation is a bane when it's in cases like genetic disorders, Alzheimer's disease etc. The proposed system can predict the possible mutations. Medical field is yet another area where the proposed system finds its relevance. Gene consultation can access this system to predict and tell the patient as of which gene annotation is responsible for what specific disorder that the patient is facing. And also in gene therapy where we can treat a disease through gene insertion or substitution. node [1].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

II. RELATED WORK

Hadoop provides a dependable, efficient, and robust platform to handle huge amounts of data where it's divided into parallel blocks and then processed[10]. It provides user ease of usage and maintenance.Y. Sun et.al explains a system that uses an ad-hoc grid solution, with a backup task system inspired by MapReduce [6] [13] [17]. It's a programming platform developed by Google. It processes large data in parallel. Here the programmer is free to code in either java or any other language that has the Hadoop Streaming API. Map Reduce is comprised of two steps: **Map**: wherein a master node is present that breaks the input into sub problems and assigns it to slave nodes. In case the sub problem is large then it's broken down again into lower level worker nodes. These does the processing and returns the result back to the master node. **Reduce**: Here; the master node fetches the results from worker nodes and combines them in an order providing the necessary output the original problem.

BLAST is a sequence alignment tool [18]. It reads a sequence at a time, computes its alignment with the input query and gives a tab-separated results line. GSEA is a statistical analysis paradigm for DNA microarray datasets [19]. This tool tests whole gene sets for a measure of the degree to which the gene set is represented at the top or bottom of a ranked gene list. GRAMMAR is a fast association analysis method. It deals with a particular phenotypic trait (e.g. body mass index) [20].

Crossbow an open source software tool that works on Hadoop [21]. Short DNA sequences and human genome are connected. And then distributes and computes the consensus sequence through the comparison with the reference genome.

Cloud-GSQCT (Cloud Gene Sequence Quality Control Tool) is a parallel approach,to screen gene sequence data for phylogenetic analysis [16]. Screening data for phylogenetic analysis from large datasets is a known computational problem of data-intensive application. TheMapReduce paradigm is used to parallelize the solution and to manage its execution. Theparallel approach using Hadoop are implemented and the evaluation of the approach shows that it is a better platform.

III. PROPOSED SYSTEM

The exhaustive search is accomplished by means of forward and reverse search. Since the DNA strand may have the required pattern in any order. In HADOOP the jobs are executed in parallel with the help of MAP-REDUCE. The input query is split into multiple jobs and they are given to the MAPREDUCE functions. The MAP operator tree and REDUCE operator handles it. The data so derived in stored in the HADDOP File System (HDFS) [8]. HDFS stores data in the form of hashtag series i.e. in the NAME and DATA nodes. It has 1: many relationship. From HDFS the data is fed back to the UI. Mutation is generally of 3 types: addition deletion and substitution [9]. Addition introduces the new random modification. Deletion removes a random value of the gene annotation and substitution replaces a value with a random one. Proposed system basically infers the exhaustive search of a necessary gene annotation in a particular organism. It also predicts the possible mutations. The system architecture is as given below

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

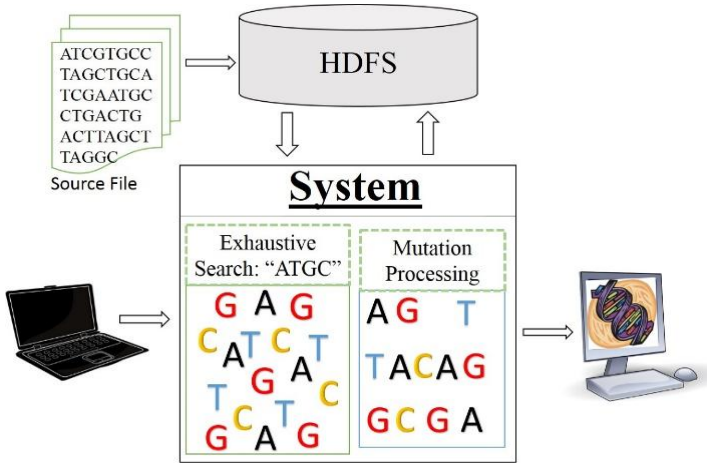


Fig 2: System Architecture

Here, the user enters the DNA annotation to be searched. In the proposed system it performs exhaustive search and derives whether the given annotation is present or not in the selected organism. It can then derive the possible mutations. For these the data access and storage is done through HADOOP technology. Finally a report is generated that describes the entire flow right from the selected organism, entered annotation to the search result and mutation patterns.

To represent the stake holder's interaction with system we infer the Use case Diagram. It represents the functionality provided by system in form of various actors and their dependencies between those use cases[10]. Its simple representation of what all system functions which respective actor uses. Given below is the Use case representation.

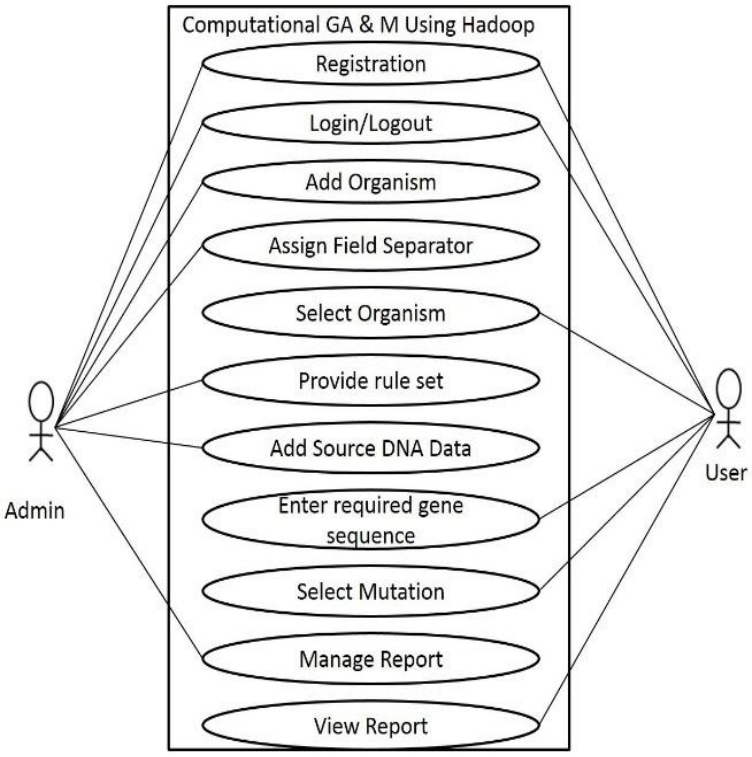


Fig 3: Use case Diagram

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

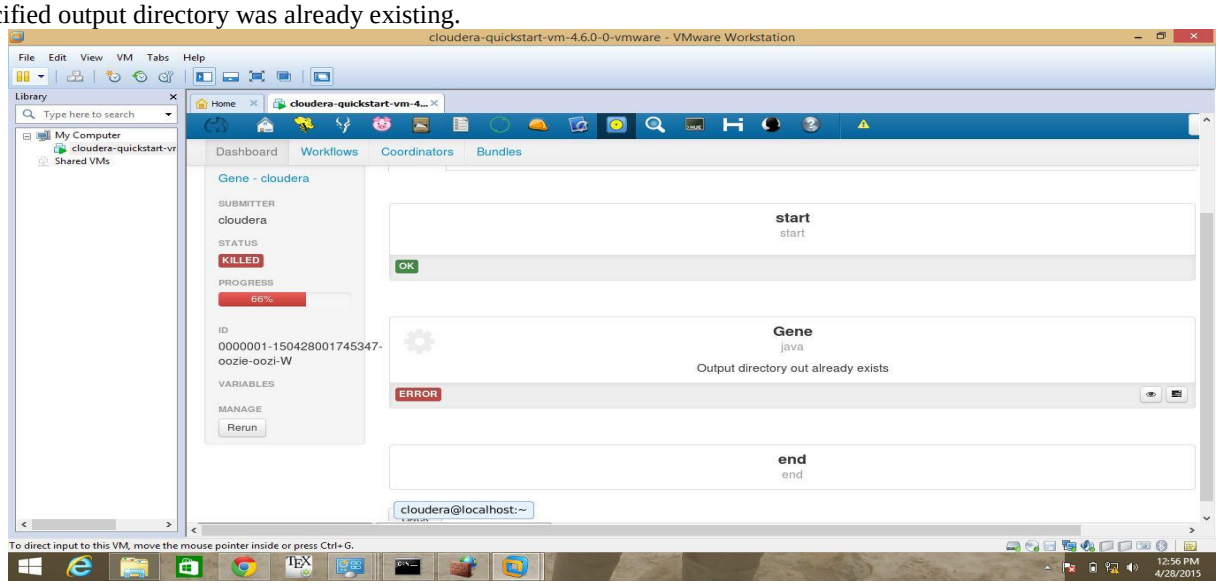
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

The two actors are User and the Admin. The user has functionalities such ass registration to create a new account, login/logout, selection of organism, providing input query, selecting type of mutation, and viewing of the final report. The admin also has the functionality of registration, login/logout. Apart from that the admin can upload the source file, provide the rule set and field separator, and also manage the user account and the final report.

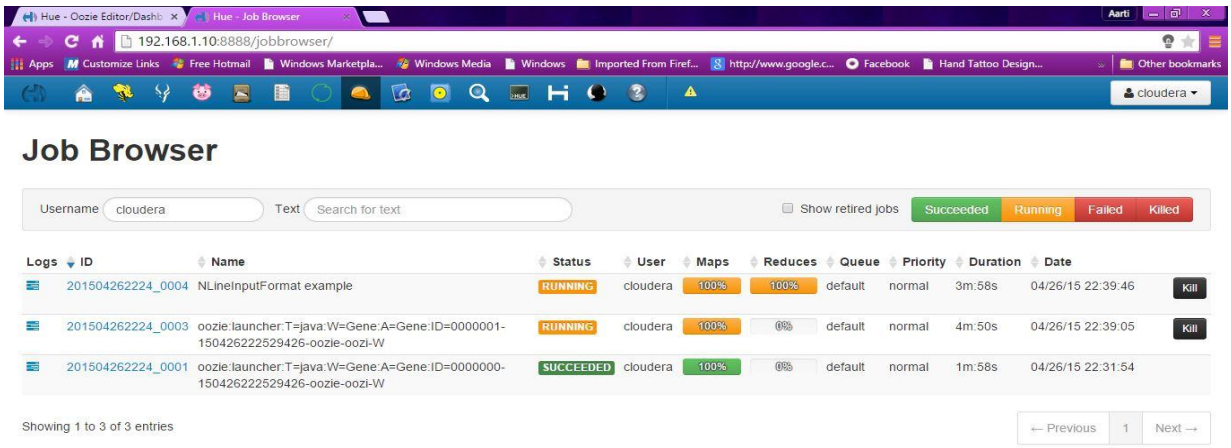
IV. RESULTS

The snapshots given below gives a glimpse to the working of the proposed system. Snapshot1 depicts the job dashboard where we can infer briefs details such as the progress of the job, the alert on whether the job is started or whether its ended. The Snapshot 1 shows that the process was started but terminated at 66% due to the error that the specified output directory was already existing.



Snapshot1: An erroneous output

Snapshot 2 shows the Job Browser wherein we can find the status of the job as of how much percentage it is running or succeeded or failed, its duration, priority etc.



Snapshot 2: Job Browser

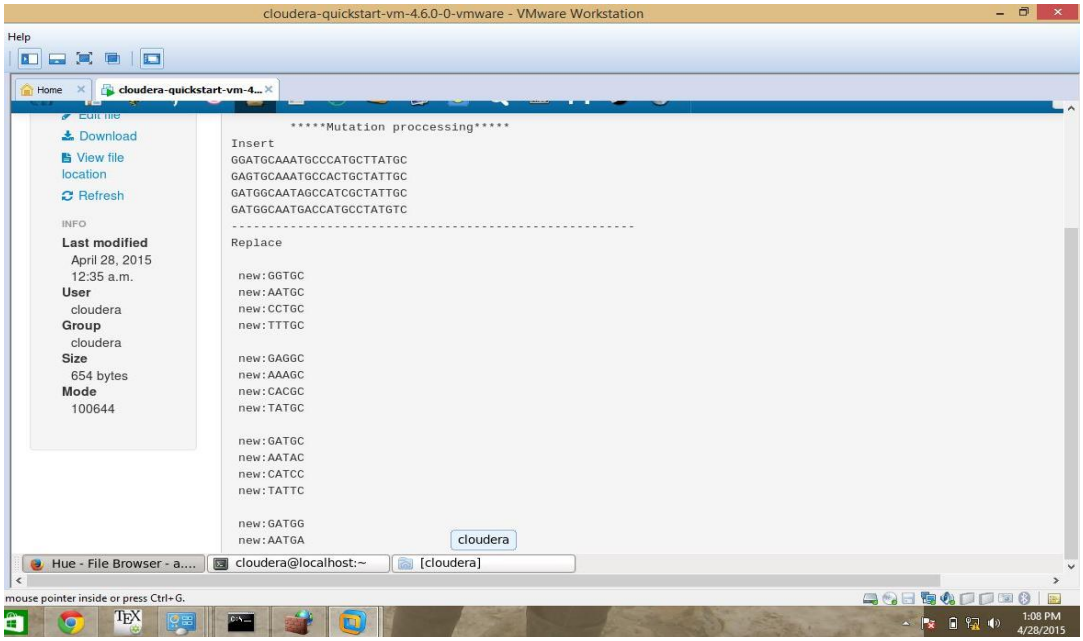
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

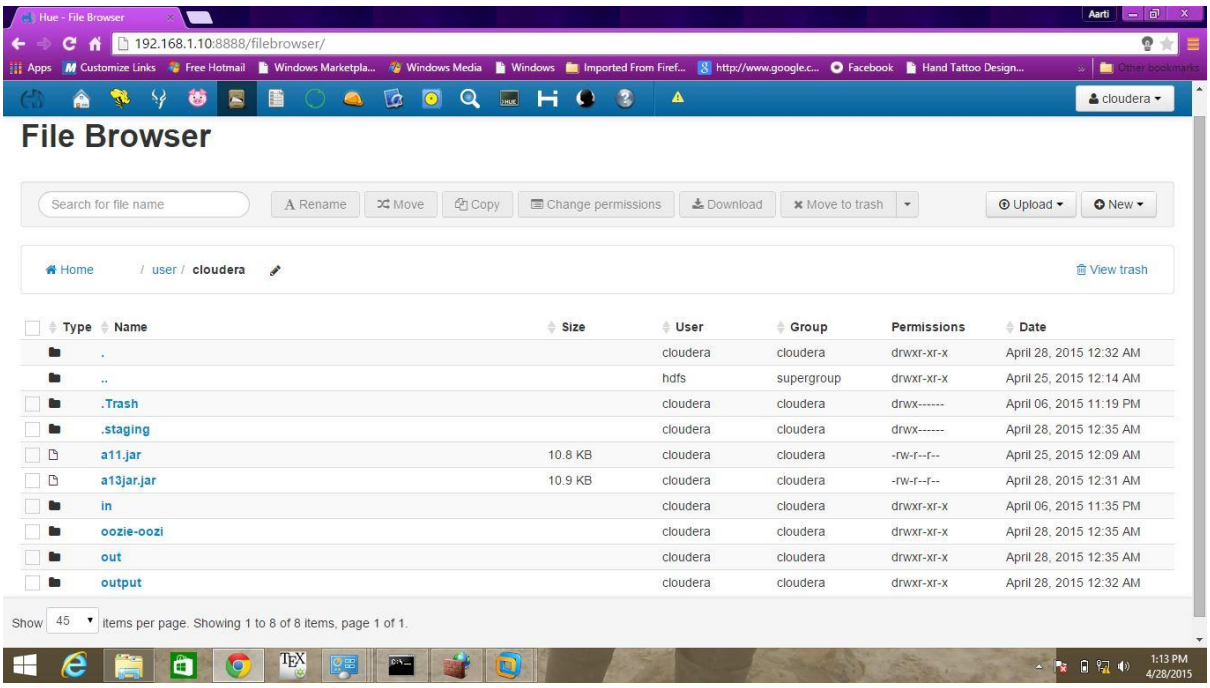
Vol. 3, Issue 5, May 2015

Snapshot 3 shows the mutation output file derived. The search string being the gene annotation "ATGC", all the possible mutations are shown.



Snapshot 3: Mutation Output

Snapshot 4 depicts the File Browser where the directories can be viewed. In this snapshot we can find the input, output directories, the java jar file a11.jar that compromises the code regarding the data analysis that's to be run.



Snapshot 4: File Browser

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

V. CONCLUSION

Hadoop handles big data very efficiently. The performance and speed of computation is quite appreciable. And so Hadoop forms a perfect tool to handle even complicated area like computational gene analysis that require analysis of large data samples for research purposes. Such researches open new path towards getting to know in depth the miracle of life and various hidden facts about living organisms. Hence the necessity of merging these researches with Hadoop for faster and reliable computations. And so a Hadoop based system is proposed that provides reliable and efficient gene analysis computationally.

ACKNOWLEDGMENT

I express my sincere gratitude to my guide, to the staff and management of my college SCEM, for their support and guidance to carry out the research.

REFERENCES

1. Simone Leo, Federico Santoni, Gianluigi Zanetti, "Biodoop: Bioinformatics on Hadoop", *International Conference on Parallel Processing Workshops,* August 2009.
2. Poonm Kumari, Shiv Kumar, "Analyze Human Genome Using Big Data", International Journal of Science and Research, May 2014.
3. David Chicco, Leo Andradae, and Marco Masseroli, "*Computational Prediction of Gene Functions Through Machine Learnig Methods", IEEE Transactions on Systems and Cybernetics, Vol.2, No.4,* July 2014.
4. Srabanti Maji and Deepak Garg, "*Progress in Gene Prediction: Principles and Challenges, IEEE Transactions on Computational Biotechnology and Bioinformatics", Vol: 13 No: 1*Year 2011.
5. Xiao Wang, Robert Clarke, and Jinghua Gu, *A Markov Random Field-Based BayesianModel to Identify Genes With Differential Methylation, IEEE Transactions on Information Technology, Vol.9, No.7,* January 2014.
6. Luis M. O. Matos, Diogo Pratas, and Armando. J, *A Compression Model for DNAMultiple Sequence Alignment Blocks, IEEE Transactions on Information Theory, Vol.6, No.1,* March 2010.
7. P. Hanus, J. Dingel, and J.Hagenauer, *Compression of Whole Genome Alignments, IEEE Trans. Inf. Theory, vol.56, no.2, pp.696-705,* Feb 2012.
8. J.Dean and S. Ghemawat, *"MapReduce: simplified data processing on large clusters,"* in *OSDI '04: 6th Symposium on Operating Systems Design and Implementation,* 2004.
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "*Basic local alignment search tool*," *J. Mol. Biol., vol. 215, no. 3, pp. 403–410,* 1990.
10. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "*Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,*" PNAS, vol. 102, no. 43, pp. 15545–15550, 2005.
11. Y. Aulchenko, D. de Koning, and C. Haley, "*Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis*," Genetics, vol. 177, no. 1, p. 577, 2007.
12. N. Amin, C. van Duijn, and Y. Aulchenko, "*A genomic background based method for association analysis in related individuals,*" PLoS ONE, vol. 2, no. 12, 2007.
13. D. Abrahams and R. Grosse Kunstleve, "*Building hybrid systems with Boost.Python,*" C/C++ Users Journal, vol. 21, no. 7, pp. 29–36, 2003.
14. Y. Aulchenko, S. Ripke, A. Isaacs, and C. van Duijn, "*GenABEL: an R library for genome-wide association analysis,*" Bioinformatics, vol. 23, no. 10, p. 1294, 2007.
15. B. Devlin and K. Roeder, "*Genomic control for association studies,*" Biometrics, vol. 55, no. 4, pp. 997–1004, 1999.
16. G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "*The international HapMap project web site,*" Genome Research, vol. 15, no. 11, pp. 1592–1593, 2005.
17. S. Asgharzadeh, R. Pique-Regi, R. Sposto, H. Wang, Y. Yang, H. Shimada, K. Matthay, J. Buckley, A. Ortega, and R. C. Seeger, "*Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking MYCN gene amplification,*" J. Natl. Cancer Inst., vol. 98, no. 17, pp. 1193–1203, 2006.
18. Y. Sun, S. Zhao, H. Yu, G. Gao, and J. Luo, "ABCGrid: application for bioinformatics computing grid," Bioinformatics, vol. 23, no. 9, pp. 1175–1177, 2007.
19. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215, no. 3, pp. 403–410, 1990.
20. Michael C. Schatz, Ben Langmead1, Jimmy Lin, Mihai Pop, Steven L. Salzberg "Whole Genome Resequencing Analysis in the Clouds", June 15, 2011

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

BIOGRAPHY

Arathiis a 4th Semester MTech student, in the Computer Science and Engineering Department, Sahyadri College of Engineering and Management, Visveswaraya Technological University, India. Her research interests are Programming, Web Designing, Bioinformatics etc.

Ankitha Kis a Assistant Professor, in the Computer Science and Engineering Department, Sahyadri College of Engineering and Management, Visveswaraya Technological University, Mangaluru, India She received her Master of Technology (MTech) degree in CSE on 2014 from SCEM, Mangaluru, India. Her research interests are Computer Networks, Programming, etc.