ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Cloud Computing and Big Data for Genomics: A Review

Sayantani Basu*
School of Computer Science and Engineering,
VIT University,
Vellore, India.

Sathyaraj R.
School of Computer Science and Engineering,
VIT University,
Vellore, India.

Abstract: The Human Genome Project has led to the massive growth and deluge of genomic sequencing data. Though new methods have decreased gene sequencing costs, this has increased the need of large-scale genomic data analysis. Such data sets are valuable and informative for scientists and medical researchers for extracting valuable associations about genes and diseases. As a result, proficient technologies capable of handling such biological big data are the need of the hour. Cloud computing and big data both serve to manage, store and analyze such data. This is because such technologies offer concurrent and distributed data processing for evaluating petabyte (PB) range genomic data sets. This paper provides a review and comparative study of cloud computing and big data approaches in Genomics that have been proposed in the last decade. This paper can be used by other researchers to develop more futuristic and robust data genomic storage systems in the future.

Keywords: cloud computing; big data; genomics; gene sequencing; bioinformatics

I. INTRODUCTION

The Human Genome Project was started as a pioneering effort by international researchers in 1993 in order to understand the genes constituted in humans. It involved mapping of genes from the human genome. New techniques of sequencing have resulted in unmatched levels of sequence data. A current global task is the challenge of handling and managing sequence data in Computational Biology. A single strand of human DNA consists of nearly 3 billion base pairs (bp), which occupies nearly hundreds of gigabytes (GB) of data. This deluge of data in Bioinformatics has led to problems in note only storage, but also in analyzing the stored data. It is because of this reason that data pertaining to biological domains will gather at an even accelerated pace [1]. Moreover, there is even more proliferating data in areas such as cancer genomics [2]. Hence, there arises a necessity to store this information so that it can be accessed and managed safely.

Cloud Computing and Big Data are required for handling the proliferating genomic sequence data. Such technologies offer parallelized and distributed data processing. However, there are certain drawbacks such as large data transfer time and less bandwidth of the network. Currently, a single DNA sequence of a human contains almost 3 billion base pairs, which is nearly 100 gigabytes (GB) in size. Hence, sequencing a mass of tumor cells may be in the range of petabytes (PB) [3]. However, if incorporated successfully, this will offer new insights on genes and diseases.

The most popular bioinformatics big data tools developed using Apache Hadoop are MapReduce, CrossBow and CloudBurst. CloudEra is also a pioneer in the domains of Cloud Computing and Big Data. Transfer of data is facilitated in cloud using Peer-to-Peer (P2P) model. Considerable programming knowledge is required for setting up a cloud environment using Hadoop. Hence, current analysis of large biological data has been made possible due to the existence of such technologies.

However, there are substantial shortcomings [4] of migrating genomics into the cloud. One noteworthy point is the cost of transferring the present systems into a completely different setup. Genome databases as well as integrators require making considerable changes in the procedures and models as capital expenses are moved to recurrent costs; all genomics users will also have to acclimatize to the new environment. One more problem that has to be addressed is management of discernible genetic data, which involve data from disease sequencing projects or whole genome association studies. Such data are presently stored in databases having limited access. Moving such data sets into public clouds like that of Amazon or that of other cloud service providers will require a considerable level of encryption and cloud security. Software also needs to be developed accordingly so that only authorized users are permitted access to such cloud resources. Implementing such a system would be governed by various privacy regulations and would eventually take time for development at both the legal as well as technological levels.

In this paper, a review has been provided on the efforts of genome sequencing that have taken place in the last decade. The analysis shows that with advances in Big Data and Cloud Computing, we will be capable of storing genomic data provided the correct methodology is used.

II. RELATED WORKS

Rosenbloom et al. [6] have discussed the 2015 updates and further developments of a Genome Browser initially developed at the University of California Santa Cruz (UCSC) [5]. The new model is provided more functionality to handle large-scale genomic datasets by using big data technologies. The authors have proposed the use of "data hubs" for storing collections of datasets. The authors have suggested improving the servers used in the previous model by expanding them to include more species. The missing DNA sequences which are required have been searched from the existing NCBI and GenBank databases and incorporated in the Genome Browser. The primary key used for the database tables are the GenBank accession number, NCBI identifiers and scientific names of organisms. The final steps involved in construction of the browser include completion of the web pages, reviewing and conducting automatic and manual testing. The authors have suggested expanding the browser for the human genome. This project will be developed later to make it compatible for cloud deployment. They have also stated that other components such as gene sets, expressions and pathways will be included in the future. Further work is also required to incorporate interaction and pathway based databases.

Tyner et al. [7] have discussed the 2017 updates of the open source, scalable display, web-based Genome Browser developed at the University of California Santa Cruz (UCSC). The 2015 update included graphical and command-line utilities for manipulating and interacting with the application. The authors have included gateways and new design methodologies to track the genetic components. They have also introduced three additional species into the existing browser configuration. New data types are now supported for genome database entry. Long-range chromatin interaction and sequence expression data can be handled in the newly updated browser. The future improvements by the authors include development of Genome Browser in the Cloud (GBiC) which will be capable of handling even larger data. Efforts are also underway so that data can automatically be synced from the existing databases. Other suggestions include improving the database searching techniques, enhancing the outlook and display of the application. There is also a necessity of tracking the human genome assembly, which will be very useful in the domain of medical research.

Dewey et al. [8] have discussed Sequence to Medical Phenotypes (STMP), an integrated pipeline system for deciphering human DNA sequence data. STMP performs targeted genotyping of variants with known clinical associations, rich functional annotation of discovered variants, and classifies genetic variants according to potential impact, mode of inheritance, and phenotypic presentation. It also offers individual predictions about disease traits (inherited) and the subsequent therapeutic response. It uses input interval call files representing previously reported Mendelian disease associated loci and loci associated with drug response to provide targeted genotype calls. STMP also provides metrics for coverage of loci with known importance to human health and disease. The STMP approach to genotype interrogation allows downstream variant annotation while reducing storage requirements for genotype data. The default STMP module comes pre-loaded with a gene co-expression network topology representing gene expression microarray data from 75 normal unused human donor hearts, tissue from 49 human hearts with right- or left-ventricular hypertrophy, and 436 explanted human hearts with dilated cardiomyopathy. When STMP is applied to WGS (Whole Genome Sequence) data from single probands, STMP provides rich functional annotation and prioritization of potential Mendelian disease risk alleles, including novel variants, structural variants, and important regulatory variants. Comprehensive annotation of standard.vcf and .gff format variant files in a median of 96 (range 90–102) minutes per genome was performed on a six-core Intel Xeon X5670 processor running 64-bit linux with 128 GB of RAM, utilizing five concurrent threads. It is highly agreeable with parallel processing architecture, produces parsimonious variant sets for manual review, and interrogates both Mendelian disease risk and genetic drug response.

Kim and Chu [9] have proposed an Internet-based data repository called the Korean Cancer Genome Database (KCGD) for finding associated gene signatures. The database is capable of handling RNA-sequence data, continuous numeric data or any other gene expression profiling method. In this repository, 1,403 cancer genomics data from Korean hospitals have been gathered, processed and stored. It uses common statistical methods such as Kaplan-Meier plot, log-rank test and the Cox proportional hazard model. These provide instant significance estimation for searched molecules. Three cancer types (i.e., liver, breast, and stomach) were known to be most frequently occurring in Koreans. The database also has non-Korean datasets for comparison or validation purposes. All datasets stored in the database were normalized using quantile normalization. Most data in the current database was created by gene expression profiling and the remaining were by methylation profiling methods. The system architecture consists of various software frameworks mainly implemented with a JAVA-based environment. The ICEfaces framework was used to lend user-friendliness to the system. The MySQL database management system was used to store and handle the datasets. Data queries on MySQL from JAVA are controlled by MyBatis, an XML-based SQL mapping framework, the statistical analysis methods were implemented using R with Bioconductor plugins, calling R modules from JAVA is managed by the RCaller framework while all these services were hosted on an Apache Tomcat web server.

Huang et al. [10] have introduced Starfish, a self-tuning system that uses big data analytics. It builds on Hadoop while adapting to user needs and system workloads to provide good performance automatically; hence, users don’t need to understand and manipulate the many tuning knobs in Hadoop. Hadoop is a MAD system (Magnetism, Agility and Depth) that is gaining popularity for big data analytics and consists of two components: a distributed file system and a MapReduce execution engine. The functionality of the components in the Starfish architecture can be categorized into job-level tuning, workflow-level tuning, and workload level. These components interact to provide Starfish’s self-tuning capabilities. Starfish’s Just-in-Time Optimizer addresses unique optimization problems and automatically selects efficient execution techniques for MapReduce jobs. The optimizer takes the help of the Profiler, which uses dynamic instrumentation to learn job profiles for unmodified MapReduce programs, and the Sampler, which collects statistics efficiently about the input, intermediate, and output key-value spaces of a Map Reduce job. The Sampler enables the Profiler to collect approximate job profiles at a fraction of the full job execution cost. Starfish’s language, Lastword, accepts and reasons about analytics workloads. Language translators automatically convert workloads from higher-level languages to Lastword. It offers Starfish to be additionally used as a recommendation engine; in this mode, Starfish uses its tuning features to recommend suitable configurations. The Jumbo operator can process any number of logical SPA (Select-Project-Aggregate) workflow nodes over the same table in a single MapReduce job. The Workflow-aware Scheduler makes decisions by considering producer-consumer relationships among jobs in the workflow. It works in conjunction with the What-if Engine and the Just-in-Time Optimizer. It performs a cost-based search for a good layout for the output data of each job in a given workflow. Starfish focuses simultaneously on different workload granularities—overall workload, workflows, and jobs (procedural and declarative)—as well as across various decision points—provisioning, optimization, scheduling, and data layout. This makes it different from other approaches and enables it to handle the significant interactions arising among choices made at different levels.

Elazhary et al. [11] have reviewed the scope of Cloud Computing for Big Data. Examples include computational biology applications such as genome or DNA sequencing. Amazon, Walmart, Facebook and E-Bay are examples of business applications of Big Data. Business applications of Big Data: Facebook, Amazon, Twitter, Flipkart. According to NIST, cloud computing is a model comprising five essential characteristics (Rapid elasticity, Measured Service, On-demand self-service, Broad network access, and Resource pooling), three service models (Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS) and Cloud Infrastructure as a Service (IaaS)), and four deployment models ((Private cloud, Public cloud, Community cloud and Hybrid cloud). Cloud computing has various other advantages such as sharing facilities, Big Data paradigms, data reliability as well as easy upgrade and maintenance. Cloud computing offers unlimited on-demand storage and computation capacity at affordable cost. However, while there are several promising advantages, the challenges in Big Data are resolving security issues, handling the internet connection, portability of applications and data, complicated pricing models and offering Quality of Service (QoS) assurance. Some examples of service tools for Big Data are IBM Softlayer, ProfitBricks and Amazon EC2. Software as a Service is a set of Software as a Service tools are developed and being developed to aid in the processing of Big Data. These include Plex and Opani Pathway Tools version 19.0 update for pathway/genome informatics and systems biology.

Karp et al. [12] have used Pathway Tools, a bioinformatics software environment possessing a variety of functionalities. It offers tools such as sequence alignments, genome browser, comparative-genomics operations and a genome-variant analyzer. It also provides metabolic-informatics tools, such as metabolic reconstruction, quantitative metabolic modeling, pre- diction of reaction atom mappings and metabolic route search. Additionally, regulatory-informatics tools are also present, such as the ability to visualize and represent a wide range of regulatory interactions. Pathway tools support development of organism-specific databases that integrate many bioinformatics data types. Some of the new pathway tools capacities added since 2010 are as follows: MetaFlux metabolic modeling component, metabolic route search and computation of reaction atom mappings and data storage of organism phenotype. A new signaling pathway editor allows user-interactive constructing and editing of signaling pathway diagrams using a set of icons and operations. Pathway tools have designed and implemented drawing code for a special ETR diagram, which shows the enzyme complex embedded in a membrane, and which schematically depicts the flow of electrons from one redox half reaction to another. Programmatic access is through APIs. Programmers can access and update PGDB data directly by writing programs in the Python, R, Java, Perl and Common Lisp languages. Its metabolic modeling capabilities include flux-balance analysis modeling for individual organisms and organism communities, with model gap filling and the ability to model gene knockouts. Omics data analysis tools paint genome-scale data sets onto a complete genome diagram, complete metabolic network diagram and complete regulatory net- work diagram.

Merelli et al. [13] have given a review of open problems in Big Data pertaining to Medical Informatics. The paradigms on which Big Data is based include volume, velocity and variety. Access and management of Big Data is the primary concern faced by researchers in bioinformatics. File system is the first level of the architecture. The second level or architecture is the framework that supports development of user specific solutions. Disco is a distributed computing framework aimed at providing a MapReduce platform for Big Data processing using Python application. Computational facilities for analyzing Big Data comprise cloud computing, GPU computing and cluster computing. Semantic web is used to promote standard for the annotation and integration of data by encouraging the inclusion of semantic content in data accessible through the Internet. Languages specially designed for handling Big Data are Resource Description Framework (RDF), Web Ontology Language (OWL), SPARQL (which is a protocol and query language for semantic web data sources), and Extensible Markup Language (XML). Ontology is a set of terms pertaining to a specific domain organized in a hierarchical form that enables searching at various levels of specificity. The following ontologies are commonly used for annotation and integration of data in biomedical and bioinformatics: (i) Gene ontology (GO) (ii) KEGG ontology (KOnt) (iii) Brenda Tissue Ontology (BTO) (iv) Cell Ontology (CL) (v) Disease ontology (DOID) (vi) Protein Ontology (PRO) (vii) Medical Subject Headings thesaurus (MESH).

Wiewiórka et al. [14] have proposed the SparkSeq software which utilizes a MapReduce framework, Apache Spark, for sequencing of data. Many time-consuming data analyses procedures can be enhanced with the help of cloud computing. Apache Hadoop-based solutions have gained popularity in genomics owing to their scalability in a cloud infrastructure. The SparkSeq software has been created to take advantage of a new MapReduce framework, Apache Spark, for next-generation data used in sequencing. SparkSeq is a general-purpose, flexible and easily extend- able library for genomic cloud computing. It can be used to build genomic analysis pipelines in Scala and run them in an interactive manner.

The comparative study of all the approaches discussed above has been tabulated as shown in Table 1.

Table 1: Comparative Study of Approaches used in storing Genome Data

Sr. No.	***Reference***	***Year***	***Name of Proposed Tool***	***Database Used***	***Cloud Technology Used***
1	[5-7]	2002-2017	UCSC Genome Browser	MySQL server	Genome Browser in the Cloud (GBiC)
2	[8]	2015	Sequence to Medical Phenotypes (STMP)	Human Gene Mutation Database (HGMD)	None
3	[9]	2015	Korean Cancer Genome Database (KCGD)	MySQL, Java, XML	None
4	[12]	2015	Pathway Tools	Lisp, JavaScript,	None

			version 19.0	Pathway/Genome Database (PGDB)	
5	[14]	2014	SparkSeq	Apache Spark, Scala, Hadoop	Data Serializer (KryoSerializer)

III. CONCLUSION

Cloud and big data based genomics is capable of revolutionizing medical science. Cloud-based resources are promising in handling big data, integrating software to analyze and manage data and as a fast transfer method to help manage large-scale genome data sets. In this paper, some such technological advances in the last decade have been highlighted which can help researchers and scientists in designing and implementing even more superior and futuristic genome databases in the cloud. The appropriate incorporation and integration of such far-reaching technologies can transform genomic research.

IV. REFERENCES

[1] Sawicki, M.P., Samara, G., Hurwitz, M. and Passaro, E., 1993. Human genome project. The American journal of surgery, 165(2), pp.258-264.

[2] Noor, A.M., Holmberg, L., Gillett, C. and Grigoriadis, A., 2015. Big Data: the challenge for small research groups in the era of cancer genomics. British journal of cancer.

[3] Singh, P., 2016. Big Genomic Data in Bioinformatics Cloud. Appli Microbio Open Access, 2(1000113), p.2

[4] Stein, L.D., 2010. The case for cloud computing in genome informatics. Genome biology, 11(5), p.207.

[5] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D., 2002. The human genome browser at UCSC. Genome research, 12(6), pp.996-1006.

[6] Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. and Harte, R.A., 2015. The UCSC genome browser database: 2015 update. Nucleic acids research, 43(D1), pp.D670-D681.

[7] Tyner, C., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C.M., Gibson, D., Gonzalez, J.N., Guruvadoo, L. and Haeussler, M., 2016. The UCSC Genome Browser database: 2017 update. Nucleic Acids Research, p.gkw1134.

[8] Dewey, F.E., Grove, M.E., Priest, J.R., Waggott, D., Batra, P., Miller, C.L., Wheeler, M., Zia, A., Pan, C., Karzcewski, K.J. and Miyake, C., 2015. Sequence to medical phenotypes: a framework for interpretation of human whole genome DNA sequence data. PLoS Genet, 11(10), p.e1005496.

[9] Kim, S.K. and Chu, I.S., 2015. A Database of Gene Expression Profiles of Korean Cancer Genome. Genomics & informatics, 13(3), pp.86-89.

[10] Huang, T., Lan, L., Fang, X., An, P., Min, J. and Wang, F., 2015. Promises and challenges of big data computing in health sciences. Big Data Research, 2(1), pp.2-11.

[11] Elazhary, H., 2014. Cloud Computing for Big Data (Vol. 2, No. 4, pp. 135-144). MAGNT Research Report.

[12] Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P. and Spaulding, A., 2015. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Briefings in bioinformatics, p.bbv079.

[13] Merelli, I., Pérez-Sánchez, H., Gesing, S. and D'Agostino, D., 2014. Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. BioMed research international, 2014.

[14] Wiewiórka, M.S., Messina, A., Pacholewska, A., Maffioletti, S., Gawrysiak, P. and Okoniewski, M.J., 2014. SparkSeq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics, p.btu343.