IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIG DATA IN BIOINFORMATICS ESPECIALLY IN BIOMEDICAL RESEARCH.

P Sai Vasantha Lakshmi
Student, Amity University Chhattisgarh, Raipur

Mr.Advin Manhar
Assistant Professor, Amity University Chhattisgarh, Raipur

ABSTRACT

In this paper , we review and discuss about the major biomedical subdiscipline "bioinformatics". Specifically, in bioinformatics, the thorough with better results which are being done facilitating the new-genome wide association studies of diseases and are beneficial in health care. The explosion of the data both in the biomedical research and in the health care systems demands urgent solutions. Effective examination and understanding of large information opens new roads to investigate atomic science.with the rapidly ever increasing amount of data being generated with the advanced tools and techniques, a number of suitable ways have been simultaneously developed to handle this vast measure of information to make it satisfactory, available and orchestrated in a legitimate request for expanding the usefulness with the information.

INTRODUCTION

Big data technologies are widely used for biomedical and health-care informatics research. The new sequencing technologies enables the processing of millions and billions of DNA sequence data per day. Big data applications provide the new and productive opportunities to discover new knowledge and create novel- efficient methods to improve the quality of health care. Medical care is continually requesting a more tight coordination with biomedical information to elevate customized medication and to flexibly better therapies.Due to the nature of the data being voluminous, methods of big data management have shown their capabilities to make the biological data effectively managed in terms of both accessibility as well as cost.

What is Big data ?

In biomedical informatics region, big data is an new ecosystem that transforms case-based studies to data-driven research. A simple definition of big data is based on the topic and concept of data sets whose size is beyond the management capabilities of typical relational database software.It is widely accepted, as the characteristics of big data that are defined by three major features, known as the 3V's: Volume, Variety and Velocity.

First and the foremost significant feature, the "volume of data" is exponentially increasing in the biomedical informatics areas. let us take an example of ProteomicsDB covers 92% of known human genes that are commented in the Swiss-Prot database. In the clinical domain, the advancement of the HITECH demonstration has almost significantly increased the appropriation pace of electronic wellbeing records (EHRs) in medical clinics to 44% from 2009 to 2012.. From millions of patients data have already been collected and stored in electronic format, thus this accumulated data could potientially enhance and nourish the health-care services and increase the research opportunities.

The second important feature of the big data is "variety of data types and structures". The paradigm of biomedical big data provides many unique levels of data sources to create a rich array for researchers. We can consider an example of sequencing technologies produce "omics" data very systematically at almost all levels of cellular

components, from proteomics, genomics and metabolomics to protein interaction and phenomics. Much of the unstructured data i.e., for example notes from EHRs, clinical trail results, medical images ans medical sensors, provide many opportunities and a different type of challenge to formulate and set up new investigations.

The third feature of big data," velocity", refers to producing and processing data. The new generation of sequencing technologies enables the production of billions of DNA sequence data every day at a comparatively low cost. Because faster speeds are required for gene sequencing, technologies of big data will be tailored to match the producing data speed, as it is required to process them.

Big Data Technologies

In the field of public health, Big data technologies will provide biomedical researches with time-saving tools for inventing new patterns among population groups using social media data. Biomedical scientists are constantly facing new challenges of storing, managing and analyzing massive quantities of datasets. The big data characteristics require novel and powerful technologies to extract the essential information and enable more broad-based health-care solutions.

Parallel computing is one of the most fundamental infrastructures for the efficient managing of big data tasks. It is capable of executing algorithm related tasks simultaneously on a cluster of machines or supercomputers. Usage of multiple technologies that are used together such as Artificial Intelligence (AI) along with Hadoop and data mining tools. In the previous years, novel parallel computing models have been proposed by Google such as MapReduce for a new bigdata infrastructure. And again recently, an open-source MapReduce package called Hadoop was released by Apache for distributed data management. The Hadoop Distributed File System (HDFS) enables and supports the concurrent data access to clustered machines.

Cloud computing is also a novel model for sharing configurable computational resources over the network and can also serve as an infrastructure, platform an \ or software for giving an integrated solution. Hadoop-based services can also be considered as cloud-computing platforms, which allows for the storage of centralized data as well as remote access across the internet. Cloud-computing can improve speed of the system, agility and flexibility because it reduces the need to maintain software or hardware capacities and aslo requires only fewer resources for system maintenance, such as configuration, installation and testing. Numerous new huge information applications depend on cloud advances.

Big Data Applications

Bioinformatics applications: Bioinformatics research analyzes the variations of biological system at the molecular level. With the current trends in personalized medicine, there is an increase in the requirement to produce, store and analyze these massive datasets in a manageable frame of time. The big data techniques role in bioinformatics applications is to provide computing infrastructure, data repositories and efficient data manipulation tools for investigators to gather and analyze biological information.

This particular section mainly describes big data technologies or tools into four categories: (1) data storage and retrieval, (2) error identification, (3) data analysis and (4) platform integration deployment. These categories may overlap and are correlated.At present, our discussion and classification in the present study is based only on the main functions of each technology.

Data storage and retrieval. Nowadays, a sequencing machine can produce millions of short DNA sequencing data in one run. Thus the sequencing data needed to be mapped to specific reference genomes in order to be used for additional analysis or for any purpose, such as genotype and expression variation analysis.

Cloud Burst is a parallel computing model that comforts the genome mapping process. It parallelizes the short-read planning cycle to improve the adaptability of perusing colossal sequencing information.The cloud burst model was evaluated using a 25-core cluster and the results shows that the speed to process 7 million short-reads was almost 24 times faster than a single-core machine.

DistMap is a toolbox for conveyed short-read planning on a Hadoop bunch. The nine upheld mapper types incorporate BWA, Bowtie, Bowtie2, GSNAP ,STAR ,SOAP, Bismark, BSMAP and TopHat.. As for example, once an evaluation test was done using a 13-node cluster, making it an effective application for mapping short-read data. The BWA mapper can perform 500 million read sets (247 GB) in around six hours by utilizing DistMap, which is multiple times quicker than the single-hub mapper.

SeqWare is an inquiry motor based on the Apache HBase information base to help bioinformatics analysts access entire genome informational indexes in huge scope.In a prototyping analysis conducted, the U87MG and 1102GBM tumor data bases were loaded, and the team used this engine in comparing the Berkeley DB and HBase back end for loading and exporting the variant data capabilities. In the end, the results showed that Berkeley DB solution is much faster when reading 6M variants, while the HBase solution is faster when reading is more than 6M variants.

Error Identification. some number of tools have been developed to identify errors in data sequencing. SAMQA recognizes mistakes and guarantees to meet the huge scope genomic information satisfy the base quality guidelines.Actually built for the National Institutes of Health Cancer Genome Atlas to automatically identify and report errors. For organic tests, scientists can set a limit to channel the peruses that could be vacant peruses and report them to specialists for manual assessment.

ART provides stimulation of data sequencing analysis for mainly three major sequencing platforms: 454 sequencing, Illumina, SOLiD. It has built-in profiles of read error and read length and can also identify 3 types of sequencing errors : base substitutions, insertions and deletions.

Data analysis. The Genome Analysis ToolKit (GATK), is aMapReduce-based programming framework that is designed to support DNA sequence analysis in large scale. GATK upholds numerous information designs including SAM documents, paired arrangement/map (BAM), HapMap, and dbSNP.

SparkSeq is one of the fast, scalable, cloud-ready software package for interactive genomic analysis with nucleotide precision. It provides interactive queries for DND /RNA studies and the project is implemented on Apache Spark using the Hadoop-BAM library for the processing of bioinformatics files.

Platform integration deployment. The use of big data platforms generally requires a strong grasp of distributed computing and networking knowledge. To enable biomedical analysts to grasp large information innovation, novel strategies are needed to incorporate existing huge information advancements with use-accommodating activities.

SeqPig reduces the requirement for bioinformatics to obtain the technological skills needed to use MapReduce. With the assistance of Hadoop-BAM, seqpig tackles the issue of perusing BAM documents of enormous size to take care of examination appilications. It supports the commonly used formats of sequencing, such as SAM, BAM,FASTQ and QSeq.

Public Health information

Public health has some core functions: such as (1)assessment, (2) assurance, (3) policy development etc. Among these assessment is the most prerequisite and fundamental function. Appraisal at first includes gathering and breaking down information to track and screen general wellbeing status, subsequently giving proof to dynamic and strategy improvement.Assurance is used to validate whether the services offered by health institutions have achieved their primary goals for increasing public health outcomes; as such, many large public health organizations , as the centres for Disease control and Prevention and the Administration of community Living, have collected and analyzed very large amount of population health data.

Here in this section no new approaches are introduced. Instead, we follow an integrated view of big data and health from a population perspective. This section mainly focuses on four areas such as infectious disease surveillance, population health management and chronic disease management.

Conclusion

We are currently living in the era of "big data", in which the big data is being rapidly used or applied to biomedical and health-care fields. In this review paper, we learned that bioinformatics is the primary field in which the big data analytics are currently being used, largely due to the massive and great volume, complexity of bioinformatics data. Application of big data in bioinformatics is quite relatively mature, with the help of clear and sophisticated platforms and tools that are already in use to help and analyze organic information, for example, quality sequencing planning instruments..

In this review paper, we demonstrated various examples in which big data technology has played an major role in modern-day health care activity. In the primary sections of this review paper we discussed that the big data applications facilitate three important clinical activities. We summed up ongoing advancement in the most significant territories in each field, including enormous information stockpiling and recovery, mistake distinguishing proof, information security, information sharing and information examination for electronic patient records, online media information and incorporated wellbeing information bases.However, in other biomedical research fields, such as medical imaging informatics, clinical informatics and public health informatics there is enormous and huge untapped potential for big data appilications.

REFERENCES

1. Hindawi publishing corporation, BioMed Research International, article ID 134023,2014.

2. Luo et al.Big data Application in Biomedical Research and Health care. A Literature review. Biomedical Insights 2016:8 1-10 doi:10.,2015.

3. Kumar, Dr & Simaiya, Sarita & Maheshwari, Shikha & Manhar, Advin & Kumar, Santosh & Chitkara,. (2020). Cloud Performance Evaluation: Hybrid Load Balancing Model Based on Modified Particle Swarm Optimization and Improved Metaheuristic Firefly Algorithms. Engineering Science and Technology an International Journal. 12315-12331.

4. Bioinformatics Big Data Problems using Natural language Processing: Help Advancing Scientific Discovery and Biomedical Research, Imam University.

5. Big data Analytics in Bioinformatics and Healthcare; Heuristic _Principal _Component _Analysis.

6. Applying Big data Analytics in Bioinformatics and Medicine; Proteomics _in _Personalized _Medicine,2018.

7. Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications, International Journal of Molecular Sciences.

8. Deep Learning in Bioinformatics: Introduction, Application and Perspective in the Big data Era, 2019.

9. Emerging trend of big data analytics in bioinformatics: a literature review, 2018.

10. Genomics in Big Data Bioinformatics, 2020.

11. An optimal big data workflow for biomedical image analysis, ELSEVIER, 2018.

12. Feature selection methods and genomic big data: a systematic review,2019.

13.Tavpritesh sethi, Big data to Big knowledge for Next Generation Medicine: A Data Science Roadmap.

14. Effy vayena ; “Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine”.

15. Aaron N. Richter & Taghi M. Khoshgoftaar : Sample size determination for biomedical big data with limited labels, Network Modelling Analysis in Health Informatics and Bioinformatics, 2020.

16. Building a Deep Learning Classifier for Enhancing a Biomedical Big Data Service.

17. Application of Big Data in Bioinformatics – A Survey, International Journal of Latest Trends in Engineering and Technology, 2016.

18. Big Data in Bioinformatics – Mathematical Biology and Bioinformatics, 2017.

19. Big data handling mechanisms in the healthcare applications: a comprehensive and systematic literature review, ELSEVIER, 2018.

20. A Review of the Literature on Big Data Analytics in Health Care,2019.

21. DNA and Bioinformatics : A Review of Different Aspects and Applications,2014

22. Integrating Molecular Biology and Bioinformatics Education, 2019.

23. Bioinformatics – Introduction and Applications ; Microbe Notes, 2018.