Role of Big Data in Bioinformatics: Brief Study

OPEN ACCESS

Volume : 6

Special Issue : 1

Month : November

Year: 2018

ISSN: 2321-788X

Impact Factor: 3.025

Citation:
Ganeshbabu, K. "Role of Big Data in Bioinformatics: Brief Study." *Shanlax International Journal of Arts, Science and Humanities,* vol. 6, no. 1, 2018, pp. 21–25.

DOI:
https://doi.org/ 10.5281/zenodo.1614242

K.Ganeshbabu
Assistant Professor, Department of Computer Applications
Sri Kaliswari College, Sivakasi

Abstract

In an era of high-throughput sequencing, drilling of biological data to extract hidden valued information plays an important role in making critical decisions across every branch of science, whether it is genomics or proteomics or metabolomics or personal medicine. For example, the genome sequence of the patients contains much valuable information about the myriad of disease causes, easy extraction of which from the sequences will enable science to develop patient-specific medicine, thereby accelerating curing process and minimizing drug's side effects. Today, true solutions of the several problems in the biological field are hidden in the analysis of exponentially increasing data, so-called Big Data. Big data has become currently hot and open issue for the biological community to handle, collect, store, analyze and manage such vast amount of data. Due to this, the computing Big Data has become the new paradigm of the science and big data in bioinformatics. While, big data is playing central role in the continuity of the progress of research in the biological field, but it presents challenges in terms of scalability, complexity, privacy and security. Big Data in the biological field have revolutionizing power to bring dramatic changes in our current understanding about several solved and unanswered problems. With these big emerging data, bioinformatics field is also evolving continuously. This paper presents the concept of big biological data and its associated challenges in the bioinformatics field which will provide a snapshot of the importance of biological big data in the bioinformatics future research.

Keywords: Biological Big Data, Bioinformatics, privacy, security

Introduction

Bioinformatics imparts its function by utilizing mathematical and computational power to store, retrieve, analyze data and extract hidden information or knowledge from the biological data. Earlier, sequencing was the key factor in the research progress due to its long time completion requirement and extremely high cost. But now the sequencing is occurring much at the faster pace accommodating the genomic sequences of thousands of diverse organisms including animals, plants and microbes apart from the thousands of human genome sequences. For instance, GridION and MinION, two nanopore sequencing platforms, can produce ultra-long sequencing reads (~100kb) with higher throughput at much lower cost [1]. These huge amounts of genomic data are maintained at both public and private repositories that are continuously retrieved by the others for further research and analysis. For instance, National Center for

Biotechnology Information or NCBI is a public repository comprised of petabytes — thousands of terabytes— of data, and biologists worldwide are extracting information from 15 petabytes of sequences [2]. Another public repository, the European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory, one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 1015 bytes) of data and back-ups about genes, proteins and small molecules [3]. Thus, high-throughput next-generation technologies have contributed to the continuously increasing data in terms of volume, variety and velocity of data. Scientists and researchers are facing difficulty in capturing, storing, and analyzing this large amount of data so-called "Big Data." Therefore, on one side where more data, information and derived knowledge presents significant opportunities for looking the organism system as a whole in bigger picture, on other side it also puts considerable challenges including data- handling, -integration, -analysis, - modeling and -simulation, knowledge extraction and management [4]. Along with this, studies involving biological big data are at the beginning phase, issues related to it are still to be resolved and thus presents an open and hot area for bioinformatics research. This paper presents the concept of big data in bioinformatics, its associated challenges and its related future perspective. This paper has been organized into five sections: The second section presents the concept of biological big data in bioinformatics. The third section deals with the challenges associated with biological big data. The fourth section provides the discussion and final section presents conclusion.

Biological Big Data Concept

The completion of goal of Human genome project (HGP) revealed billions of not only the bases in the human genome sequence but also identified and mapped the total number of genes in the human genome [5]. HGP has fostered the development of high-throughput measurement tools and strategies as well as stimulated the development of new computational tools and software for acquiring, storing and analyzing sequencing data [5]. Moreover, earlier science was known for its experiments, theoretical explanations and computational techniques but now in today's world, high- throughput next-generation sequencing technology marked its role in providing ultra high speed and lower cost to sequencing where terabytes (1012) and petabytes (1015) of biological data are producing at an ever-increasing rate. Thus, with technological advancements, the cost and the time to sequence a genome has been significantly dropped. For instance, the cost of sequencing has been reduced from the millions of dollars to some thousands of dollars [6] that made sequencing benchtop to be accessible to thousands of institutions, laboratories and hospitals. Earlier, sequencing single human genome need several years, but now half dozen can be sequenced in around ten days. This indicates revolutionary momentum in the sequencing data generation, marking its name as "Big Data." Biological Big Data refers to extremely large and complex datasets that is far exceeding the capacity of computer technology (traditional databases, tools and techniques) to collect, store, process, manage, organize and analyze these data. But, this definition describes superficial features of the big data and in order to understand its deep meaning, the word "Big" should be considered as what required amount hard disk capacity (terabytes or petabytes or exabytes) is attributing data to be Big. Moreover, defined big term, in terms of a certain amount of memory space, is also the time- and technology- dependent as the required capacity that seems to be big to date will be changed to even larger capacity requirement as the time progresses and the technology will advances. Gartner defined "Big data" to be high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [7]. Some of the facts regarding biological big data are as follows [8]:

- Even a single sequenced human genome is around 140 gigabytes in size.
- European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 1015 bytes) of data.
- The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.
- Each day in 2012, the EBI received about nine million online requests to query its data, 60% increase over 2011.

Biological Big Data Challenges-Bioinformatics

Since a long time biologists are struggling with the 'big data' and the condition are getting more and more severe. Moreover, it is realized that while according to Moore's law, the computing power is doubling roughly for every two years, but this rate is not matching the speed with which sequencing data is accumulating. The tremendous growth of Biological Data is currently driving the development of technology to extract "Knowledge" from the "Big Data" due to which bioinformatics has become an active area of research that is capable of extracting the absolute benefits from the Big Data.

Bioinformatics requires variety of graph analysis, string matching and database technologies to meet the computational challenges of big data. Thus, big data handling requires improvement in existing algorithm along with the development of new ones. But limited computational power for capturing, storing and analyzing data are the most important constraint in front of bioinformatics to effective and efficient use of the algorithm for

1. Sorting: to distinguish between the clinical and real-world environmental sample.
2. Reassembly of constituent genome
3. To Identify pathogens and characteristic genes responsible for antibiotic resistance and toxins.
4. Enhancement of testing process with high-performance databases and parallel computing software technologies.

That's the reason the major focus of bioinformatics has been moved to the computation of biological Big Data. In last three years, semiconductors or nanotechnology [9]; the two Next Generation Sequencing (NGS) platforms have exponentially increased the rate of generation of biological data. Big Data generation and acquisition gives birth to profound challenges for storage, transfer and security of the information. Even if companies were forced to limit their data collection and the storage space, still the big data analytics would be needed. However, in the coming years, the doctors will use the individual DNA for providing personal medicine.

Thus, now the focus has been shifting from sequencing to computation of biological big data. But big data has brought considerable challenges to bioinformatics, which are as follows:

Storage demand of big data

Big data not only demand large storage space but also need increasing space to accommodate rapidly increasing data. Processing of big data needs greater computational time, so, for fast processing computational time needs to decrease. For scientists having large storage and computational infrastructure can be difficult to maintain, also implementing cost of this infrastructure may be extremely high.

Data transfer

Data transfer from one location to another is also a major that is carried out mainly by the use of external hard disks or by mail. Thus, due to the large size, big data transfer and access need large amount of time that lead to reduced processing time. Big data handling system will become efficient if data simultaneously processed and computed that result in faster outputs generation.

Security and the privacy

Bioinformatics Big data handling including data storage or transfer through external hard drive or servers raises the issue of security. Therefore, authenticity and confidentiality of the data are the two important challenges associated with big data that has to be considered.

Deriving value

Volume, variety and velocity are the three main challenges in deriving the meaningful values from the big biological data as it represents challenges such as data processing, data management and data infrastructure. Researchers need high- performance computers in order to extract the hidden information from the big data.

Heterogeneity

Biological data are much more heterogeneous than physics. The data generated from an array of different experiments that produce different types of information such as nucleotide sequences, protein interactions, and findings of medical records of patients.

Presentation and visualization of big data

Often data produced by one researcher is shared and used by other researchers. However, better presentation of big data is a major challenge in the present era as rapid extraction of valuable information based on it. Thus, data presentation should be in such a format that is easy to be finding out and analyzed. Better visualization of big data can make it easy to understand and analysis faster and help the scientist in making quicker solution of a given problem. But the large volume, complexity of big data and who is the concerned user are the some of the problems in this pathway of the better visualization.

Time and space-constraint

If health practitioners or scientist wants to compare their research results or patient's disease condition with all thousands of other previous results or patients records having similar conditions, they have to download all related huge amount of data. Retrieving such large amount of data not only s time-consuming but also demands appropriate computer infrastructure. But with the present available infrastructure for scientists and researcher that seems to be difficult.

Availability of software tools

Big data analysis also demands up-to-date software and tools for their analysis that begins with the specific development of sequence algorithm. Sequence algorithm development accounts several sequence formats that needs the construction of multiple copies of data in parallel. Also, each copy of data need to be pass through a computation of the number of statistics and the selection of a specific algorithm based on these formats and statistics.

Further analysis proceeds and repeated with the testing of selected algorithm over a range of parameters across all formats of data until an optimal range met satisfying the given aim. Thus, sequence algorithm development and testing needs an extensive data storage and retrieval system. Sequence algorithm consists of three phases- collections, storing and comparing with query.

In first step, data is collected from different sources and is parsed into suitable format for further analysis. In second stage, the collected data are stored in the database so as to allow the retrieval of data through their queries. In the third stage, the queried data in different combination are returned required by the users. Therefore, big data produced by these high- throughput techniques in the genomic and proteomic field need databases that are built to facilitate the storage of these data along with efficient retrieval and handling of data from these huge databases [10].

Discussion

Like many other fields, big data has brought dynamic changes in the science as well as in medicine too. DNA sequencing application has become a remarkable tool in the area of medicine. But the increasing capacity of computers and speed of the internet at present does not seem to meet the need to produce, transfer and analyze biological big data securely such that omics data can be successfully integrated with other data sets, such as patient's clinical datasets. In addition to this, the tremendous growth in biological (such as sequencing and biomedical) data, have empowered us to see the different aspects of biological science with entirely new prospective (such as in the areas of cancer drugs and personal medicines). But these continuously accumulating huge data in computers and large servers across the world [11] have also raised concerns over security, privacy and ethical issues.

Conclusion

Big data is one of the general attribute of biological studies, and today, researchers are capable of generating terabytes of data in hours. Over the last decade, biological datasets have been grown massively in size, mostly because of advances in technologies for collection and recording of data. Therefore, big data posses a great impact on the bioinformatics field and a researcher in this field faces many difficulties in using big biological data. Thus, it is essential that bioinformatics develop tools and techniques for big data analysis so as to keep pace with our ability to extract valuable information from the data easily thereby enhancing further advancement in the decision-making process related to diverse biological process, diseases and disorders.

References

Baxevanis. A. D. & Francis Quellette. B.F. B., F., F. (2005) Bioinformatics: A practical guide to the analysis of genes and proteins (2005).

Clarke. J. et al. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 4 (2009), pp. 265–70.

Collins. F. S., Green. E.D., Guttmacher. A. E. & Guyer. M. S. (2003) A vision for the future of genomics research. Nature, 422, pp. 835-847.

Costa. F. F. (2017) "Big Data in Biomedicine," Drug Discovery Today, 2017, in Press.

Eisenstein. M. (2014) Oxford Nanopore announcement sets sequencing sector abuzz. Nature Biotechnology, 30, pp. 295–296.

EMBL–EBI. 2012. Annual Scientific Report 2012.

Gartner.inc. 2016. Gartner IT Glossary Big Data.

Howe. D., Costanzo. M. & Fey. P., et al. (2008) Big data: the future of biocuration. Nature, 455, pp. 47-50.

Marx. V. (2013) The Big Challenges of Big Data. Nature, 498 (2013), pp. 255-260.

Science.gov. 2015. Science.gov Trivia Answers

Vance. A. (2014) Illumina's New Low-Cost Genome Machine Will Change Health Care Forever. Bloomberg Businessweek Technology.

Websites

https://research.ijcaonline.org/volume101/number11/pxc3898841.pdf

http://www.ijcrcst.com/papers/IJCRCST-OCTOBER15-03.pdf