Journal of Computer Science

Review

Big Data and its Future in Computational Biology: A Literature Review

[1]Iman Ahmed ElSayed, [2]Kamal ElDahshan, [1]Hesham Hefny and [2]Eman Karam ElSayed

[1]Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt
[2]Department of Computer Science, Faculty of Science, Al-Azhar University Cairo, Egypt

Article history
Received: 11-12-2018
Revised: 06-03-2019
Accepted: 10-04-2019

Corresponding Author:
Iman Ahmed ElSayed
Department of Computer Science, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt
Email: imanelsayed2729@gmail.com

Abstract: Big data technologies are used majorly for computational biology research nowadays due to the emerge of massive amounts of biological data that have been generated and collected at an unprecedented speed and scale and with different structures. As an example, the processing of billions of DNA sequence data per day represents the new generation of sequencing technologies. It is expected that the cost of acquiring and analyzing biomedical data will decrease with the assistance of technology upgrades in the computational biology field. The application of big data in health care is a fast-growing field, with many new discoveries and methodologies. In this study, an overview of recent developments and technologies in big data in the computational biology field. Moreover, in this study, we summarize and highlight the challenges, gaps and opportunities to improve and advance big data applications in computational biology. Also, this review presents a useful starting point for the application of Big Data in future Bioinformatics research. Finally, some of the Big Data sources in computational biology that help in applying these new emerging technologies are mentioned.

Keywords: BigData, Computational Biology, Bioinformatics, Literature Review, Healthcare, Hadoop, MapReduce

Introduction

Most of the world is stepping now into the new vision of big data. Data is becoming bigger than we knew before, not just concerning the plenty of patterns existing (i.e.,: Instances), but also putting in mind the data dimension itself (i.e.,: Attributes) (Wang *et al*., 2016). Over the last decade, most industries have resorted to changing their long lasting records into a digital form in order to well preserve them. One of the major fields that caught the eyes of big data was especially the computational biology and health care field (Palanisamy and Thirunavukarasu, 2017) (Mehta and Pandit, 2018). The available healthcare data offers a future vision for delivering optimized health and medical care. Unfortunately, this healthcare data is sometimes recognized only as a by-product, rather than a central huge asset source for competitive advantages in the medical informatics field (Mehta and Pandit, 2018).

As the electronic health data remains largely underutilized and hence wasted, there is a need for converting the raw data into meaningful and actionable information (Mehta and Pandit, 2018). Consequently, a marvelous huge amount of biomedicine data that is daily produced, collected and stored has also passed over this digital transformation due to the increase in use of Electronic Medical Records (EMRs); Healthcare Information Systems (HIS); and handheld, wearable and smart devices in the bioinformatics and healthcare field. Consequently, with massive amount and variety of health-related data today, it has been stored in a digital form such as: Omics data, socio-demographics data and insurance claims data apart from clinical data (Mehta and Pandit, 2018).

In spite that there still doesn't exist an exact, clear and agreeable definition; “big data” as an expression mainly points to datasets that are often characterized by their enormous volume, variety and veracity. These large datasets are now available thanks to affordable and easy-to-use technologies. These large datasets have enabled the easy access for public health researchers to generate large amounts of data (Grant, 2012; Lohr, 2012; Marx, 2013; Murdoch and Detsky, 2013; Young, 2015). For example, the Human Genome Project (HGP), was an

Science Publications

© 2021 Iman Ahmed ElSayed, Kamal ElDahshan, Hesham Hefny and Eman Karam ElSayed. This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

international collaboration to sequence all the base pairs in the human genome. Individual labs were tasked to contribute data from certain areas of the human genome to the HGP database. The combination of these data and the additional combined data have made HGP a clear example of big data genome.gov, 2014 (Young, 2015).

Time being, big data technology has a favorable and promising value in bioinformatics and healthcare. Recently, big data has created an increasing interest in academic and industrial researches and investigations. Nevertheless, there have been only few literature reviews in this point and the literature remains broadly crumbled.

Thus, the purpose of this research is to earn and acquire a comprehensive and thorough understanding of current vision on this technology. Its aim is to give an answer and explore the conceptual aspect of applying Big Data to healthcare and bioinformatics major field.

Review Methods

We have applied and adapted a systematic review for handling relevant literature from different sources, focusing on the following research questions:

RQ.1: What are the main concepts in Big Data? And what are the different concepts of big data in computational biology?
RQ.2: What are the different sources of big data available in computational biology field?
RQ.3: What are the different tools of big data available in computational biology field?
RQ.4: What are the impacts of Big Data in computational biology? And what are the challenge that might face big data usage in bioinformatics?

Information Resources

A search has been taking place to find related research articles on following databases:

(1) El Sevier
(2) Oxford University Press
(3) IEEE
(4) Science Direct
(5) Springer
(6) Scopus
(7) PubMed

Selection Criteria

While performing the search in these databases, we used the following main keywords for the search criteria: {"big data in health care", "bioinformatics in big data", "big data" and “computational biomedical techniques.” As a result, papers were selected upon the following selection criteria:

SC.1: The paper has to be written and published in English
SC.2: The article has to be published within this period (2007-2018).
SC.3: The paper discusses and evaluates the usage of big data applications in the bioinformatics and health-care domains.
SC.4: The paper discusses the impacts and challenges facing big data in the computational biology field.
SC.5: The paper reports new methods for processing big data and discusses the performance of each method

Exclusion Criteria

The following exclusion criteria was used to filter out irrelevant papers:

EC.1: The paper which doesn't discuss any of our related focusing results from our research questions.
EC.2: The document is a tutorial or a course material not a published revised article.

Search Results

In the first search based on the keywords of our research questions, all related papers were collected by reviewing the title and abstract and checking them thoroughly. This initial search resulted in 342 papers from 2005 to 2017.

In the second search, all 342 chosen papers were investigated based on the above mentioned inclusion and exclusion criteria and as a result, 74 candidate papers were selected.

Finally, all these 74 papers were evaluated for the final selection by reading the content of the papers required and consequently, 21 papers were selected for this study.

Big Data Concepts

Big data is not like normal traditional term for any data. Big data has appeared and raised rapidly with new opportunities and challenges in many different research fields. Big Data has a complex nature that requires powerful technologies and advanced algorithms. It's clear obvious that the size of data is considered to be a very important factor for big data. Big data refers to large, massive (we will be dealing with exabytes (10^{18}), zettabytes (10^{21}) and yottabytes (10^{24})) complex and linkable amount of data sets that include heterogeneous formats: Structured, unstructured and semi-structured data sets (Oussous *et al*., 2018; Andreu-Perez *et al*., 2015).

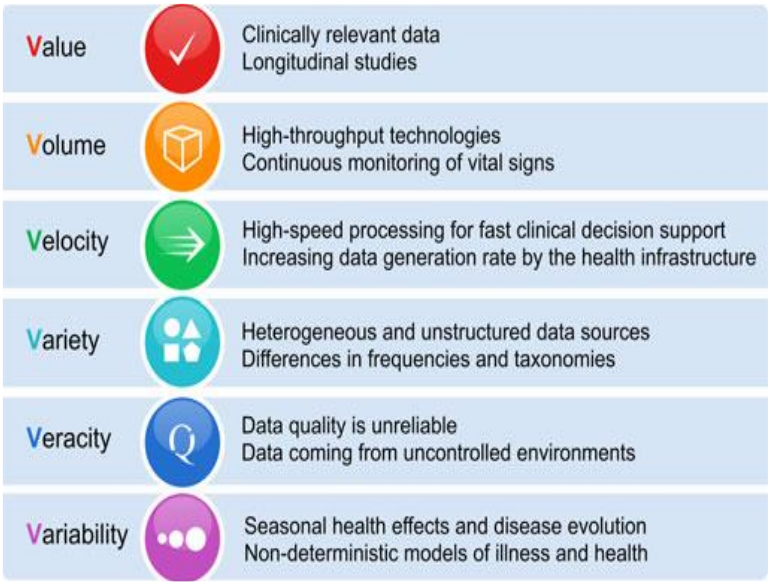

Fig. 1: Six V's of big data (value, volume, velocity, variety, veracity and variability) (Andreu-Perez *et al*., 2015)

Most data scientists and experts define Big Data by the following three main major characteristics (called the 3Vs): volume, variety and velocity. These 3 V's were the original definition of the characteristics of big data. Lately other characteristics have been added and considered including Variability (i.e.,: consistency of data over time), Veracity (trustworthiness of the data obtained) and Value (Andreu-Perez *et al*., 2015) (Fig. 1).

With regards to Big Data in bioinformatics and health care field, initially and most important characteristic is "volume of data sets". Volume of data is increasing exponentially in the bioinformatics fields. For example, the ProteomicsDB8 covers 92% of known human genes. ProteomicsDB has a data volume of 5.17 TB. Data from millions of patients have already been collected and stored in an electronic format and these accumulated data could potentially enhance health-care services and increase research opportunities (Luo *et al*., 2016).

The characteristic of Variety in data types and its structures comes in the second place for big data characteristics. For example, sequencing technologies produce "omics" data systematically at almost all levels of cellular components, from genomics, proteomics and metabolomics to protein interaction and phenomics which provides a variety of unstructured data that helps in providing many and unique oppurtunities and challenges for new researching (Andreu-Perez *et al*., 2015; Luo *et al*., 2016).

The third characteristic for big data is velocity. Velocity for data here refers to data production and then data processing. As an example, in order to produce cheaply billions of DNA sequence data daily, a new generation of big data sequencing technologies are used to match the speed of data production in processing (Andreu-Perez *et al*., 2015) (Luo *et al*., 2016).

Finally, other challenges rather than the above mentioned three characteristics, there is also heterogeneity and data variety in health care fields. Due to different and variant data capturing devices, biomedical data is always increasingly generated at a very high speed and accordingly decision making and support has to be given at near real time.

Big Data Technologies in Computational Biology and Bioinformatics

Biomedical scientists are facing major burdened challenges for storing, managing and performing various analysis on large amounts of data sets. New technologies are needed to be powerful and able to extract useful information and as a result enables very well managed variant health care solutions. Moreover, research in the bioinformatics field has exceeded sequencing the genome of an individual moving into measuring the epigenomic data ((i.e., above the genome). This includes alterations in a gene expression rather than only primary DNA sequences changes (Oussous *et al*., 2018).

Lately, it has been realized that most bioinformatics big data problems are solved with multiple technologies along together such as; artificial intelligence, data mining tool, aside with Hadoop. Bioinformatics research analyzes biological system variations at the molecular level. With new and upcoming ways and methods in personalized medicine, there has been an increase in the production, storage and analysis of huge datasets in a quantifiable logical time interval. Big data techniques interfere at this point to help in bioinformatics applications in order to provide data repositories, computing infrastructure and efficient data manipulation tools for researchers to gather and analyze biological

information. Currently, Hadoop and Mapreduce are the most used tools within the computational biology and biomedical fields (Luo *et al.*, 2016; Oussous *et al*., 2018; Andreu-Perez *et al*., 2015; Mehta and Pandit, 2018).

Big Data and MapReduce in Bioinformatics

MapReduce is a very well-known supported programming model that helps in processing of a huge datasets. MapReduce users have to define the following items for the flow of data in the model and the items are: (1) An Input Reader, (2) A map function, (3) A partition function, (4) A compare function, (5) A Reduce function, (6) An Output Writer (Luo *et al*., 2016; Mehta and Pandit, 2018). MapReduce has been used in many points in the field of bioinformatics such as giving the ability to improve the performance of common signal detection algorithms for pharma-covigilance at approximately linear speedup rates. Moreover, MapReduce based algorithms can improve the performance of neural signal processing (Usha and Jenil, 2014). MapReduce framework has been also used for finding optimal parameters for lung texture classification and to increase the speed of medical image processing (Peek *et al*., 2014).

Many bioinformatics BigData toolkits use MapReduce framework and is based on it such as: The *Genome Analysis Toolkit* (GATK). GATK is designed and used for huge data sets of DNA sequence analysis. GATK supports many data formats, including SAM files, Binary Alignment/Map (BAM), HapMap and dbSNP. GATK has been used in the Cancer Genome Atlas and 1000 Genomes Projects (Van der Auwera *et al*., 2013; Vazirabad, 2016).

Big Data and Hadoop in Bioinformatics

Apache Hadoop is a very well recognized Big Data technology which was founded by Apache as a JAVA open source project and it is based on MapReduce model. It is a software framework for distributed processing of large dataset across large clusters of commodity servers which is used to optimize huge volume of data (i.e., ideal for BigData). Hadoop is complement to OLTP & OLAP and it provides reliability through replication.

Hadoop helped a lot in avoiding the bad performance that resulted from processing and analyzing Big Data using traditional technologies and increasing rapidly the processing of it. This occurred due to: (1) Parallel clustering, (2) distributed file system, (3) executing tasks where data are stored not in the memory and this helps the network to take a breath from the communication load and (4) running programs while ensuring fault-tolerance, usually encountered in distributed environment (Oussous *et al*., 2018; Usha and Jenil, 2014).

Table 1: Platforms supporting both hadoop and mapreduce in bioinformatics field

Name	Year	Description
Halvade	2015	A Hadoop-based tool for performing read alignment and variant calling for genomic data
Spark Seq	2014	Fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. It is implemented on Apache Spark using the Hadoop-BAM library for processing bioinformatics files
Seq Pig	2014	Simple and scalable scripting for large sequencing data sets in Hadoop
Cloud DOE	2014	A software package that provides a simple interface for deploying the Hadoop cloud because the Hadoop platform is often too complex for scientists without computer science expertise. It analyses high-throughput sequencing data with MapReduce, encapsulating the complicated procedures for configuring the Hadoop cloud for bioinformatics researchers
Dist Map	2013	Aims to increase the support of different types of mappers to cover a wider range of sequencing applications
Bio Pig	2013	A Hadoop-based analytic toolkit for large-scale sequence data
Hydra	2012	Is a scalable proteomic search engine that uses the Hadoop-distributed computing framework. It helps in processing large peptide and spectra databases, implementing a distributed computing environment that supports the scalable searching of massive amounts of spectrometry data.
SAMQA	2012	Identifies such errors and ensures that large-scale genomic data meet the minimum quality standards. It also includes a set of technical tests to find data abnormalities
The Eoulsan package	2012	Is a MapReduce-based framework which is used for analyzing the differential transcript expressions, including data imports from sequencer reads, data mapping to reference genomes, alignment filters, transcription expression calculations, expression normalizations and detection of differential expressions.
FX	2012	An RNA-Seq analysis tool on the cloud using Hadoop
Cloud Aligner	2011	Is MapReduce based framework for short DNA read mapping
Seq Ware	2010	Is a query engine built on the Apache HBase35 database to help bioinformatics researchers access large-scale whole-genome datasets.
Genome Analysis Toolkit GATK	2010	Is a MapReduce-based programing framework designed to support large-scale DNA sequence analysis.
Myrna	2010	Cloud-scale RNA-sequencing differential expression analysis using Hadoop
Cloud Burst	2009	Highly sensitive read mapping with MapReduce
Cloud BLAST	2008	Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications

Hadoop based algorithms are used in many areas in bioinformatics such as: Refining protein structure alignments more accurately than current used algorithms (Hung and Lin, 2013), local alignment search for similar regions between sequences, etc.

Below is a table for platforms supporting both Hadoop and Mapreduce in the bioinformatics field. [Table 1: Platforms supporting both Hadoop and Mapreduce in Bioinformatics field] (Hung and Lin, 2013; Lee, 2015; Xubin *et al*., 2012; Yang *et al*., 2017; Nguyen *et al*., 2011; Zicari, 2014).

Sources of Big Data in Bioinformatics

Data in healthcare are disorganized and distributed, having different structures and forms either structured or unstructured data. As Big Data technology in bioinformatics increased, it was able to deal with large volumes of both structured and unstructured data emerging in bioinformatics.

Bioinformatics field data contains many types and is collected from many variety of fields such as: Physiological, behavioural, molecular, clinical, environmental exposure, medical imaging, disease management, medication prescription history, nutrition, or exercise parameters. Here are examples of some of the primary sources of big data in bioinformatics:

1- Administrative databases (insurance claims and pharmaceuticals)
2- Clinical databases, electronic health record data and laboratory information system data
3- Biometric data
4- Patient-reported data (standardized health surveys)
5- Data from social media
6- Medical imaging data
7- Biomarker data, including all the spectrum of ‘omics’ data (that is, genomic, proteomic and metabolomic data)

All of these sources of data (Fig. 2) were the major input for applying the big data technologies that resulted later in many helpful researches and outcomes for evolving the field of bioinformatics (Zicari, 2014) (Markonis *et al*., 2012).

Results and Limitations

Results

This literature review enriched the role of BigData in Bioinformatics. In particular, it did focus about the concept of Big Data, defining Big Data in Bioinformatics, the most widely used frameworks in Big Data and how they were also used in bioinformatics and applied in it, some of the most important techniques in Bioinformatics applying Big Data frameworks in order to deal with the large healthcare datasets and finally some of the BigData sources which are used in these techniques in the bioinformatics field to find a proper outcome and help in assessing this prosperous evolving field.

As a result of this literature review, it was found that most of the studies published, reviewed and were concerned only with the application of big data with certain technology to find something specific or a certain application or a personalized medicine. Unfortunately, there doesn't exist no certain model that combined many of the technologies to perform a certain functionality and preserve a structured outcome that can doctors or clinics rely on directly without the need of another help and also help improve the time wasted for dealing with Big Data healthcare data itself.

Type	Description		Source
Clinical	Electronic Medical Records (EMRs)	Detailed patient-related information (physician prescriptions, medications, medical history)	Hospitals & Clinics
	Diagnostic	Diagnostic Results (imaging results, laboratory reports)	Laboratories Radiology Departments
	Biomarkers	Molecular data (genomic, proteomic, transcriptomic, metabolomic)	Diagnostic Companies
	Ancillary	Administrative data (admission, discharge, transfer) & financial data (claims)	Hospitals & Clinics Data Aggregators
Claims	Medical Claims	Medical reimbursement data (procedures, hospital stay, insurance policy details)	Payers Data Aggregators
	Prescription Claims	Prescription reimbursement data (drugs, dose, duration)	Payers Data Aggregators
Clinical Research	Clinical Trails	Design parameters (compound, size, end points)	Pharma Companies Medical Journals
Patient-generated Data	Social Media	Community discussions	Web Health Portals Social Media Websites
	Wearable & Sensors	Wellness & lifestyle data (smartphones, fitness monitors)	Device Data Systems

Fig. 2: Sources of Big data in Bioinformatics (Andreu-Perez *et al*., 2015)

Moreover in consequence, the act of integrating Big Data technology in Bioinformatics didn't help only in the improvement of biological findings, but also helps in identifying and realizing the highly risked patients by making the best usage of up to date and real-time analysis and eliminating any inefficient results concerning these patients.

Limitations

Current researches have few limitations that need to be considered such as:

- Most studies reviewed, have nearly a narrow point of view with limited practical application as a whole model for big data application in Bioinformatics
- Most of these published research work discussed the application but there was no real-world cases as an evidence and example applied for this application

Conclusion

For the time being, we are in the major sight zone of BigData where BigData is rapidly and enormously applied to the computational biology and healthcare field. Big Data in these fields can help boosting the applicability of research studies and frame them into real-world scenarios concurrently and effectively.

In this literature review, it has been shown that Bioinformatics has emerged to be one of the most wanted and attractive fields in which BigData now plays a very important role due to the massive volume and veracity of bioinformatics data. Big data applications in computational biology is relatively mature, with sophisticated platforms and tools already in use to help analyse biological data.

Moreover, this review presents a useful starting point for the application of Big Data in future Bioinformatics research. As a result, advances in big data processing for computational biology will have a great impact on future clinical research.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G. Z. (2015). Big data for health. IEEE journal of biomedical and health informatics, 19(4), 1193-1208. doi.org/10.1109/JBHI.2015.2450362

Grant, E. (2012). The promise of big data. HSPH News, Spring/Summer.

Hung, C. L., & Lin, Y. L. (2013). Implementation of a parallel protein structure alignment service on cloud. International journal of genomics, 2013. doi.org/10.1155/2013/439681

Lee, H. (2015). Using bioinformatics applications on the cloud. Indiana University.

Lohr, S. (2012). The age of big data. New York Times, 11.

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. Biomedical informatics insights, 8, BII-S31559. doi.org/10.4137/BII.S31559

Markonis, D., Schaer, R., Eggel, I., Müller, H., & Depeursinge, A. (2012, September). Using MapReduce for large-scale medical image analysis. In 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (pp. 1-1). IEEE. doi.org/10.1109/HISB.2012.8

Marx, V. (2013). The big challenges of big data. Nature, 498(7453), 255-260. doi.org/10.1038/498255a

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, 57-65. doi.org/10.1016/j.ijmedinf.2018.03.013

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352. doi.org/10.1001/jama.2013.393

Nguyen, T., Shi, W., & Ruden, D. (2011). CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. BMC research notes, 4(1), 1-7. doi.org/10.1186/1756-0500-4-171

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-448. doi.org/10.1016/j.jksuci.2017.06.001

Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks-A review. Journal of King Saud University-Computer and Information Sciences, 31(4), 415-425. doi.org/10.1016/j.jksuci.2017.12.007

Peek, N., Holmes, J. H., & Sun, J. (2014). Technical challenges for big data in biomedicine and health: Data sources, infrastructure and analytics. Yearbook of medical informatics, 23(01), 42-47. doi.org/10.15265/IY-2014-0018

Usha, D., & Jenil, A. (2014). A survey of Big Data processing in perspective of Hadoop and MapReduce. International Journal of Current Engineering and Technology, 4(2), 602-606. http://inpressco.com/wp-content/uploads/2014/03/Paper29602-606.pdf

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), 11-10. doi.org/10.1002/0471250953.bi1110s43

Vazirabad, I. (2016). Hadoop and MapReduce in bioinformatics and healthcare. Department of Mathematics, Statistics, Computer Science, Milwaukee, WI.

Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods, 111, 21-31. doi.org/10.1016/j.ymeth.2016.08.014

Xubin, L., Wenrui, J., Yi, J., & Quan, Z. (2012). Hadoop applications in bioinformatics. Proceedings of the 7th Open Cirrus Summit, IEEE Xplore Press, Beijing, China, pp, 48-52. doi.org/10.1109/OCS.2012.40

Yang, A., Troup, M., & Ho, J. W. (2017). Scalability and validation of big data bioinformatics software. Computational and structural biotechnology journal, 15, 379-386. doi.org/10.1016/j.csbj.2017.07.002

Young, S. D. (2015). A “big data” approach to HIV epidemiology and prevention. Preventive medicine, 70, 17-18. doi.org/10.1016/j.ypmed.2014.11.002

Zicari, R. V. (2014). Big data: Challenges and opportunities. Big data computing, 564, 103.