# A STUDY ON ROLE OF BIG DATA IN BIOINFORMATICS

**K.Ganeshbabu**
Assistant Professor,
Department of Computer Application,
Sri Kaliswari College, Virudhunagar, India

**Abstract: –** In this paper the role of big data in bioinformatics is the 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 emerging big data, bioinformatics field is also evolving continuously. This paper presents the concept of biological big data and its associated challenges in the bioinformatics field which will provides a snapshot of the importance of biological big data in the bioinformatics future research.

***Keywords: Keywords: Big Data, Bioinformatics, privacy, security, genomics, and proteomics***

## I. INTRODUCTION

Next generation sequencing technologies contributed to Science New paradigm "Data Deluge". This deluge in data has fostered bioinformatics field to be more focused on "Computing Data" along with increasing demand of sequencing. Moreover, an interdisciplinary field 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].

### Bioinformatics

Bioinformatics is the Science of integrating, managing, mining and interpreting information from biological datasets at genomic, metabalomic, proteomics, phylogenetic and cellular or whole organism levels.

According to (National Institute of Health) NIH organization, the Bioinformatics and Computational Biology have been defined as "Bioinformatics is research and development or application of computational tools and approaches for expanding the use of biological, medical, health data including those to acquire store, organize, active, analyze or visualize such data".

### Genomics

DNA (Deoxyribonucleic Acid) is a molecule encoding the genetic instructions used in the development and functioning of all known living organisms many viruses. DNA is one of the three major macromolecules that are essential for all known forms of life.

Genetic information is encoded as a sequence of nucleotides (Guanine, Adenine, Thymine, and Cytosine) recorded using the letters G, A, T, and C. Most DNA molecules are double-stranded helices, consisting of two long polymers of simple units called nucleotides with the nucleobases (G, A, T, C) attached to the sugars. DNA is well-suited for biological information storage, since the DNA backbone is resistant to cleavage and the double-stranded structure provides the molecule with a built-in duplicate of the encoded information [3].

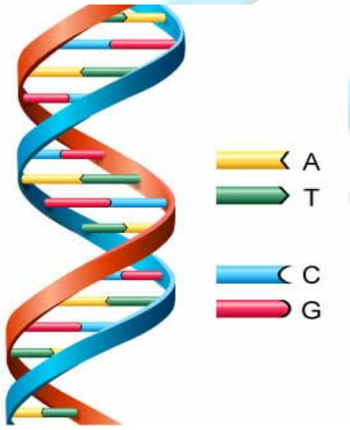


**Figure 1: Structure of DNA**

## II. 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 ($10^{12}$) and petabytes ($10^{15}$) of biological data are producing at an ever increasing rate.

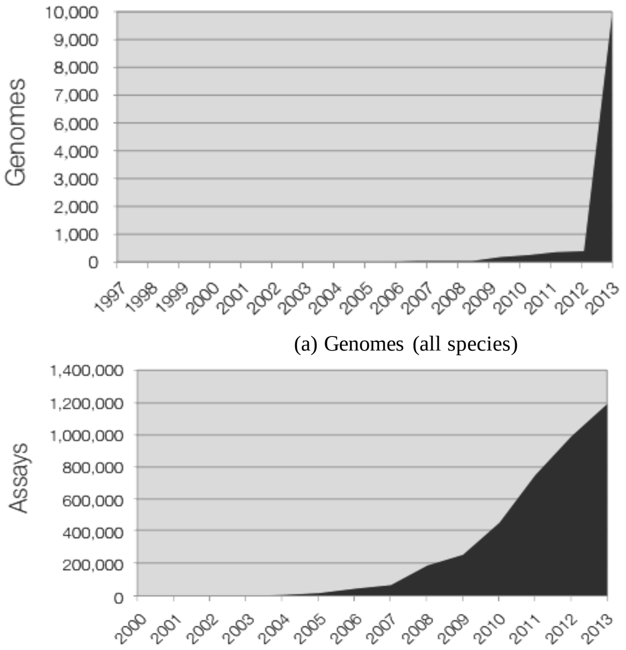
Thus, with technological advancements, the cost and the time to sequence a genome has been significantly dropped. For instance, the cost of sequencing has been reduced from the millions of dollars to some thousands of dollars [6] that made sequencing benchtop to be accessible to thousands of institutions, 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.

The volume of data is growing fast in bioinformatics research. Big data sources are no longer limited to particle physics experiments or search-engine logs and indexes. With digitization of all processes and availability of high throughput devices at lower costs, data volume is rising everywhere, including in bioinformatics research. For instance, the size of a single sequenced human genome is approximately 200 gigabytes [7]. This trend in rising data volume is also supported by decreasing computing cost and increasing analytics throughput with growing big data technologies. Biologists no longer use traditional laboratories to discover a novel biomarker for a disease, rather they rely on huge and continuously growing genomic data made available by various research groups. Technologies for capturing bio data are becoming cheaper and more effective, such as automated genome sequencers, giving rise to this new era of big data in bioinformatics. Fig. 2. illustrated that the quantity of data stored by EBI over the years.

EBI has installed a cluster, the Hinxton data centre clus- ter, with 17,000 cores and 74 terabytes of RAM, to process their data. Its computing power is increased in almost every month. More importantly, EBI is not the only organization involved in massive bio-data store. There are many other organizations, who are storing and processing huge collections of biological databases and distributing them around the world, such as National Center for Biotechnology Information (NCBI), USA and National Institute of Genetics, Japan.

Availability of high volume of data is helpful for more accurate analytics, particularly in a highly sensitive field of research like bioinformatics. However, the big data challenges here are much different from other well known big data problems, such as particle physics data captured at CERN or high resolution satellite data received at NRSC/ISRO open data archive.



(a) Genomes (all species)

(b) Gene expression data

**Figure 2: Quantity of data stored by EBI over the years**

The difference comes mainly in two aspects. First, bioinformatics data are highly heterogeneous in nature. Many analytics problems in bioin- formatics require multiple heterogeneous and independent databases for inference and validation. Moreover, bioinfor- matics data are generated by many uncontrolled organiza- tions and consequently, the same types of data are repre- sented in different forms by their sources. Second, bioinfor- matics data, massive and growing in terms of dimension and number of instances, is geographically distributed all over the world.

While part of these data may be transferred over the Internet, the remaining are not transferable due to their size (and hence inefficient), cost, privacy, and other ethical issues [9]. This sometimes forces to perform part of the analysis remotely and share the results. Therefore, big data problems in bioinformatics are not only characterized by volume, velocity, and variety, but also by geographically distributed data.

In order to tackle these challenges of big data in bioinformatics, cloud computing technologies have been used, with a lot of success. The best policy is to use cloud for both data store as well as for computation [9]. In fact, this policy helps to handle the big data challenges imposed by bioinformatics research over massive, growing and remotely distributed data. BGI, formerly known as Beijing Genomics Institute, one of the world's premiere genome sequencing centers, has installed a cloud-based analysis workflow called Gaea, using Hadoop framework.

Gaea can be used to perform large-scale genome analysis in parallel across hundreds of cloud-based computers. Another notable cloud-based genome analytics solution is provided by Bina Technologies, a Stanford University and UC Berkeley spin- off, in terms of a hardware component, called Bina box, to do the pre-processing on genome data and a cloud-based component to perform analytics on the pre-processed data. Bina box also reduces the size of genome data for their effi- cient transfer to the cloud component. This solution claims to improve the throughput of genome analytics by orders of magnitude higher than the traditional approachesBut, 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 $10^{15}$ 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.

## III. 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 of pathogens and characteristic genes responsible for antibiotic resistance and toxins.
4. Enhancement of testing process with high-performance databases and parallel computing software technologies.

**Microarray data analysis:**

The size and number of microarray datasets are growing rapidly, mainly due to decreasing cost and widespread use of microarray experiments. Moreover, microarray experiments are also been performed for gene-sample-time space, in order to capture the changes in expression values over time or over different stages of a disease. Big data technologies are required for fast construction of co-expression and regulatory networks using voluminous microarray.

**Gene-gene network analysis:**

Gene regulatory networks (GRN) alterations underlie many anomalous conditions, such as cancer. Inferring GRN and their alterations from high-throughput microarray data is a fundamental but challenging task. With the rapid growth of high throughput sequencing technologies, system biologist are now able to infer gigabytes of data. In many cases, movement of such large volume of data is not feasible. Integration of large multiple GRNs from different sources help in reconstruction of a unified GRN. Reconstruction of GRNs locally and then their integration through cloud infrastructure may help system biologists to better analyze a diseased network.

**PPI data analysis:**

PPI complexes and changes in them inhibit high information content about various diseases. PPI networks are being studied in various domains of life sciences with production of voluminous data. The volume, velocity, and variety of data make PPI complex analytics is a genuine big data problem. It demands for an efficient and scalable architecture to provide fast and accurate PPI complex generation, validation, and rank-aggregation.

**Sequence analysis:**

With the increasing volume (in order of petabytes) of DNA data deluge originated from thousands of sources, the present DNA sequencing tools have been found inadequate. So, development of a high throughput and compact architecture for DNA sequence analysis with renewed focus for big data management is a bioinformatics problem with high demand in the recent days.

The next generation genome sequencing provides information on the complete genome of an individual, in orders of magnitude bigger in size than microarray based methods for genetic assessment. Large scale methods are needed to study the specific changes in genome sequences due to a particular disease and to compare with the existing results of the same or different related diseases.

**Pathway analysis:**

Pathway analysis associates genetic products with phenotypes of interest, in order to predict gene function, identify biomarkers and traits, and classify patients and samples. The genetic, genomic, metabolomics, and proteomic data has increased rapidly and big data technologies are required to perform association analysis on huge volumes of these data.

**Disease network analysis:**

Large disease networks have been formulated for many species, including human. These networks are continuously growing and new networks are being added by different sources in their own format. The multi-objective associations among diseases in heterogeneous networks are useful for understanding the relations among diseases across networks. Traditional network analytics techniques would not perform well over unstructured and heterogeneous data without compromising information quality, and intelligent and efficient analytics are required. Big data technologies are required to effectively deep mine the associations among heterogeneous disease networks.

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.

### *3.1 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 decreased. For scientists having large storage and computational infrastructure can be difficult to maintain, also implementing cost of this infrastructure may be extremely high.

### *3.2 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.

### *3.3 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.

### *3.4 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.

### *3.5 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.

### *3.6 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 problem in this pathway of the better visualization.

### *3.7 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 is time-consuming, but also demands appropriate computer infrastructure. But with the present available infrastructure for scientists and researcher that seems to be difficult.

### *3.8 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 facilitates the storage of these data along with efficient retrieval and handling of data from these huge databases [10].

## IV. 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.

## V. 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 biological big 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.

## VI . REFERENCES

[1] Eisenstein, M. 2012. Oxford Nanopore announcement sets sequencing sector abuzz. Nature Biotechnology, 30, 295–296.

[2] Science.gov. 2014. Science.gov Trivia Answers

[3] EMBL–EBI. 2012. Annual Scientific Report 2012.

[4] Howe D., Costanzo M., Fey P., et al. 2008. Big data: the future of biocuration. Nature, 455, 47-50.

[5] Collins, F. S., Green, E.D., Guttmacher, A. E. and Guyer, M. S. 2003. A vision for the future of genomics research**.** Nature, 422, 835-847.

[6] Vance, A . 2014. Illumina's New Low-Cost Genome Machine Will Change Health Care Forever. Bloomberg Businessweek Technology.

[7] Gartner.inc. 2014. Gartner IT Glossary Big Data.

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

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

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

[11] Costa, F. F. 2013. "Big Data in Biomedicine," Drug Discovery Today, 2013, in Press.