MEDWIN PUBLISHERS
Committed to Create Value for researchers

Open Access Journal of Pharmaceutical Research
ISSN: 2574-7797

Cloud Computing and Biological Data

Dubey A*

Independent researcher and analyst, India

***Corresponding author:** Dr. Anubha Dubey, Independent researcher and analyst, Gayatri Nagar, Katni, M.P, India, Email: anubhadubey@rediffmail.com

Editorial

Volume 4 Issue 2

Received Date: April 08, 2020

Published Date: April 27, 2020

DOI: 10.23880/oajpr-16000201

Editorial

The diversity of data leads to volume, velocity, variety, variability and value, the "V"s of big data worldwide known. And these are the challenges for biomedical and data scientists to bring proper information out of them. Day by day a protein is isolated and biological databases are enhanced. The biological databases like Genbank, PDB etc are growing in a fast way. Not only protein, DNA, RNA database are enriched but also disease databases like eBioportal for cancer genomics [1] is widely used resource that integrates and visualizes cancer genomic data, including mutations, copy number variation, gene expression and clinical trial information. These are the first generation databases having complete information of particular nucleic acids, gene, and disease. Now as researches in bioinformatics are speeds up, our understanding of life and diseases are also achieving heights. With this we have second generation system cloud computing with biomedical data which enables researches of the world to compute over the data. BLAST Basic local Alignment search tool of NCBI is its very good example.

As the demand of information technology is increasing in biological field. The researchers are looking towards cloud computing. It provides on demand, large scale integrated computing infrastructure for storage and maintain security of biomedical data/biological data which is called data cloud [2]. One of the good example is cancer data cloud are developed by seven bridges Genomics [3] etc.

Importance of using Cloud Computing in Genomics

a. National cancer institute (NCI) Genomics data commons[4] is launched to analyze and harmonize genomic and associated all kind of data, i.e. mutation, clinical etc, including TGCA. And data harmonization is applied for cleaning, applying quality control criteria, processing and post processing of submitted data.
b. Development of three NCI cloud pilots, fire cloud etc which provide cloud based computing infrastructure to analyze TGCA data. They use GCP and AWS simultaneously.
c. Analysis of 280 whole genomes using multiple distributed public and private clouds for their information storage [5].

With the increasing demand of cloud computing in scientific or biological area. A "data commons" is borned which provides the co-location of data with cloud computing and commonly used software's, services, tools and applications for managing,integrating,analyzing and sharing data which require APIS for creating an interoperable resources. These data cloud, data lakes, data commons are all important to bring hidden information in data by the help of data analytics methods like machine learning or Artificial intelligence.

ML is an application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [6]. In genomics, bioinformatics, these techniques are very suitable for analytics: (a) researchers are using ML techniques to identify patterns within high volume genetic datasets; these patterns are then translated to computer models which will help in predicting individuals' probability of certain diseases or medical practitioners to design the potential therapy. (b) Genome sequencing and gene editing is the most highly researched field in medical science for future treatments of diseases.

Hence cloud computing methods provide data access, storage, reusability and interoperability. Most important path breaking is big data to Knowledge (BD2K) initiative [7]. Since data in a data common are like clinical, molecular, imaging and other data, which may be structured or unstructured. The structured data includes clinical data, demographic data, bio specimen data, variant data which form data schema. The unstructured data includes text, notes, articles, and other data that are not associated with data schema. If there is any query whether structured or unstructured data, there are several tools to make it easier [8,9].

By ML, classification models are developed to fulfils one own demand of data acquisition. As per need of data, data lakes, data catalog is formed which provide data access, analyses, in form of data model or schemas. Since importing data is labor intensive, we make our cloud as pay of compute models [10].

As bioinformatics has specialized workflow management system which is data intensive and complex. So according to our requirements the model is made (Figure 1).

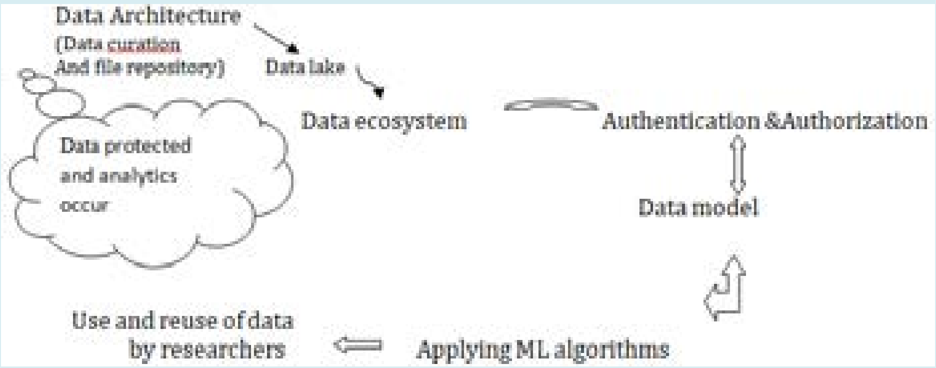

Figure 1: Cloud based data analytic model for data.

Conclusion

With the rapid growth of biological data with the high throughput techniques and medical system. There is a need to record data for further investigation, analysis and scientific study, leads to the development of cloud computing techniques. For analysis data in cloud, we have machine learning algorithms, deploying which will help medical practitioners to improve the treatment processes. The data is in the form of data commons, Data Lake, data model to support large scale computing and it is interoperable, sharing, highly secure and greater compliance. In near future data commons will provide patients to submit their own data to data commons and gain some understanding of it. This will develop new insights to patient's researcher partnership.

References

1. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 26(269): 1.

2. Heath AP, Greenway M, Powell R, Spring J, Suarez R, et al. (2014) Bionimbus: a cloud for managing, ana-lyzing and sharing large genomics datasets. J Am Med Inform Assoc 21(6): 969-975.

3. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, et al. (2017) The Cancer Genomics Cloud: collabora-tive, reproducible, and democratized-a new paradigm in large-scale computational research. Cancer Res 77(21): e3-e6 .

4. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, et al. (2016) Toward a shared vision for cancer genomic data. N Engl J Med 375(12): 1109-1112.

5. Yung CK, O'Connor BD, Yakneen S, Zhang J, Ellrott K, et al. (2017) Large-scale uniform analysis of cancer whole genomes in multiple computing environments. bioRxiv, pp: 1-45.

6. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, et al. (2015) The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc 22(6): 1114.

7. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

8. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, et al. (2016) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 45(D1): D712-D722.

9. Haendel M, McMurry JA, Relevo R, Mungall CJ, Robinson PN, et al. (2018) A census of disease ontologies. Annu Rev Biomed Data Sci 1: 305-331.

10. Grossman RL, Heath A, Murphy M, Patterson M, Wells W (2016) A case for data commons: toward data science as a service. Comput Sci Eng 18(5): 10-20.