Scope International Journal of Science, Humanities, Management and Technology. ISSN : 2455-068X
Vol.2 Issue 1 (2015) 1-7. Submitted 03/01/2016. Published 1/02/2016

OMICS TECHNOLOGY IN BIG DATA

[1]Geetha*, Assistant Professor, Mahendra Engineering College for Women, Tamilnadu, India

[2]Dr. M. Thangamani, Assistant Professor, Kongu Engineering College, Tamilnadu, India

ABSTRACT

"Big data" is a branch of health care informatics that pools large and disparate data sets and applies a suite of mathematical approaches that derives associations, facilitates comparisons and generates insights in medical aspects. The data sets can be comprised of Electronic Health Record data, insurance claims, pharmacy utilization, care management systems, consumer as well as government information, public health etc., Life Sciences have been highly affected by the generation of large data sets, specifically by overloads of omics information (genomes, transcriptomes, epigenomes and other omics data from cells, tissues and organisms). Big data in health is concerned with meaningful datasets that are too big, too fast, and too complex for healthcare providers to process and interpret with existing tools. This paper attempts to integrate omics data with other data sets such as clinical data from patients.

Keyword: Big data, Omics, Electronic Health Record, Health care.

I. INTRODUCTION

In this paper, several ways [1] of defining big data exist as a broad term to encapsulate the challenges related to the processing of a massive amount of structured and unstructured data. Clearly, the size (or volume) of data is an important factor of big data. Indeed, the US healthcare system alone already reached 150 exabytes (1018) five years ago. Several open-source frameworks such as Hadoop have been considered to store distributed databases in a scalable architecture, as a basis for tools (e.g., Cascading, Pig, and Hive) that allow developing applications to process vast amounts of data on commodity clusters. EHRs describing patient treatments and outcomes are rich but underused information. The availability of the genomic, proteomic, and metabolic databases allows a better understanding of the development of complex diseases such as cancer. They also allow the search of new biomarkers using different pattern mining and clustering techniques. The clusters can be either partitional or hierarchical (tree-like nested structure). These methods can be further accelerated by using multicore CPU, GPU, and field-programmable gate arrays with parallel processing techniques.

The multiscale data generated [2] from individuals is continuously increasing, particularly with the new high-throughput sequencing platforms, real-time imaging, and point of care devices, as well as wearable computing and mobile health technologies. They provide genomics, proteomics and metabolomics, as well as long-term. A single whole human genome obtained by the next-generation sequencing is typically 3 GB. Depending on the average depth of coverage, this can vary up to 200 GB, making it a clear source of big data for health.

1.1 Big data in Clinical trials

Clinical trials and studies [3] are all about conducting research into disease and conditions, and the various treatment methods that may ease symptoms or eradicate the illness altogether they explore which treatments work best for which illnesses and in which groups of patient. All around the world, electronic business machine is the established standard for the provision of healthcare. But, in the age of big data, that might be about to change. Clinical trials work by testing new treatments in small groups at first, looking at how well the treatment works and to identify any side effects. If a trial proves promising, it is expanded to include larger groups of people. Often the trial will include comparing the new treatment to other treatments by separating patients into different groups, each trialing a different treatment. This is usually done by a process called randomization, where patients are assigned to the various groups randomly. A staggering 80% of medical and clinical information about patients is formed of unstructured data, such as written physician notes, consultant notes, radiology notes, pathology results, discharge notes from a hospital, etc. Schulte, a physician who was Apixio's Chief Medical Officer before being appointed CEO, says, "If we want to learn how to better care for individuals and understand more about the health of the population as a whole, we need to be able to mine unstructured data for insights."

1.2. Genome with big data

Whole genome sequencing by Next Generation Sequencing is important to the study of complex diseases such as cancer. It has been a long-standing problem in cancer treatment that drugs often have heterogeneous treatment responses even for the same type of cancer, and some drugs only show profound sensitivity in a small number of patient [4,5] . A bottleneck in analyzing big data is to obtain fast inference in real time from large and high-dimensional observations. For instance, high-dimensional spaces may arise from an extensive set of biomarkers [6, 7], health attributes, and sensor fusion. From a software point of view, processing big data is usually linked to parallel programming paradigms such as MapReduce [8].

II. RELATED WORKS

Big data analytics has an indispensable role in fostering those enhanced relations because it vastly enriches the remarkable but isolated wonder of the genome-on-a-thumb drive. Healthcare providers and drug makers now have the ability to explore and analyze omics data not only for an individual, but also in an aggregate from an increasing number of patients in specific population studies [9]. Online tools, such as the General Practice Research Database, applied [10] to clinical studies in drug discovery and assessment exemplify how IT is impacting biomedicine.

2.1. Big data in biomedical field

Big data in biomedicine is driven by the single premise of one day having personalized medicine programs that will significantly improve patient care. Constant advances in understanding of different omics information are providing the footholds into establishing, for the first time, the causal genetic factors that could help manage the golden triangle of treatment: the right target, the right chemistry and the right patient. Solutions to deal with this overload of information are becoming a reality. However, challenges ahead include funneling clinical data, omics data, administrative data and also financial information securely into an unified system [11] to achieve better patient outcomes, advance research and continually improve the quality of patient care while reducing costs.

2.2. Big data Analytics

Most of the big data surge is unstructured information and is not typically easy for traditional databases to analyze it. Therefore, the predictive power of big data has been explored recently in fields such as public health, science and medicine [12]. Big data describe a new generation of technologies and architectures, designed to extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery and analysis [13]. This world of big data requires a shift in computing architecture so that researchers can handle both the data storage requirements and the heavy server processing needed to analyze large volumes of data in a secure manner. Cloud computing is the only storage model that can provide the elastic scale needed for DNA sequencing, whose rate of technology advancement could now exceed Moore's Law [14] One of the greatest scientific discoveries of our generation is the mapping of the human genome. Our genome can be sequenced [15] for between $1000 and $4000, and scientists predict that $100 individual genome sequencing can be seen in the next few years.

2.3. Omics with big data

Pipelines to deal with increasing amounts of omics data will be needed to store, transfer, analyze, visualize and generate 'short' reports for researchers and clinicians (Fig. 1). In fact, an entirely new genomics industry could result from cloud computing, which will transform medicine and life sciences. Indeed, cloud computing opens a new world of possibilities for the genomics industry to transform the way that it approaches research and medicine. Success in biomedical research to deal with the increasing amounts of omics data combined with clinical information will depend on the ability to interpret large data sets that are generated by different emerging technologies. Big corporations, such as Microsoft, Apple, Oracle, Amazon, Google, Facebook and Twitter, are masters in dealing with big data sets [16]. Other solutions to deal with big data, especially when analyzing complex genomics information, include the use of graphics processing units (GPUs). GPUs have the potential to improve quickly and drastically computational power over conventional processors, even when compared with the cloud [17]. For example, GPUs can be used as a tool to detect gene–gene interactions in genome wide studies [18]. Compared with the currently used central processing units (CPUs), GPUs are highly parallel hardware providing massive computation resources. GPUs have been recently used for proteomic analysis [19] and metagenomic sequence classification [20], and could be applied to deal with heterogeneous sources of data, such as clinical and genomic information. NextBio uses data about the human genome to aid providers in making personalized medical decisions. Their big-data technology uses both public and

proprietary molecular and genomic data, as well as clinical information from individual patients which is uploaded by the provider into a HIPAA-compliant secure domain [21]

Using high-throughput technologies [22] ,an exhaustive number of measurements can be performed over a short period of time giving access to individuals' DNA, transcribed RNA from genes over time, DNA methylation and protein profiles of specific tissues and cells, metabolites,that relates to the types of omics such as genomics, transcriptomics, epigenomics and proteomics and metabolomics respectively.

Initially, there were great expectations for omics data to provide clues on the mechanisms underlying disease initiation and progression as well as new strategies for disease prediction, prevention and treatment [23]. The idea was to translate omics profiles into subject-specific care based on their disease networks shown in Fig.1. However, our ability to decipher molecular mechanisms that regulate complex relationships remain limited despite growing access to omics profiles. Biological processes are very complex, and this coupled with the noisy nature of experimental data (e.g. cellular heterogeneity) and the limitations of statistical analyses (e.g. false positive associations) poses many challenges to detecting interactions between "networks" and "networks of networks". Fig.1 shows the omics Profile of personalized medicine.

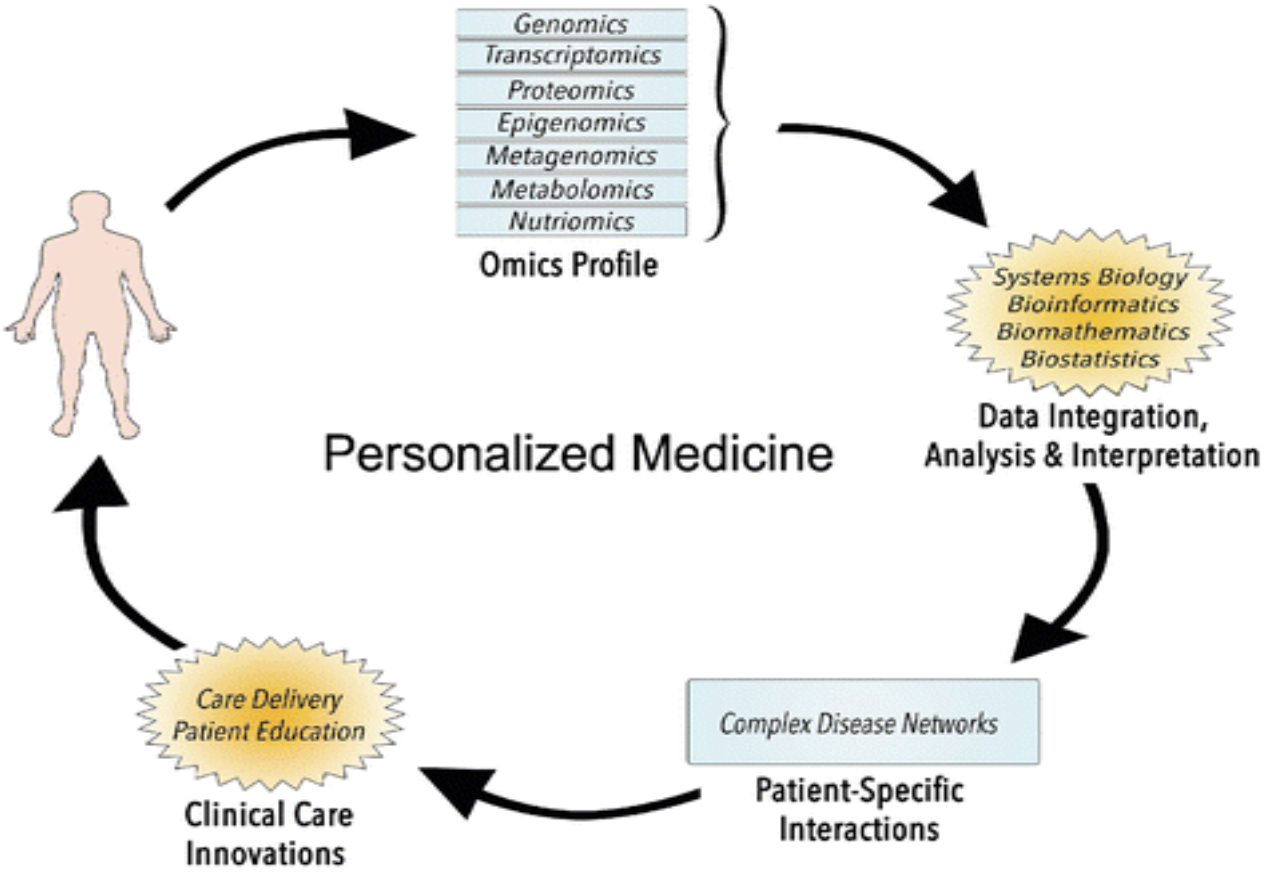

Fig.1. A Basic framework of personalized medicine

2.4. Challenges in Omics data

Major investments need to be made in bioinformatics. Classic research laboratories do not possess sufficient storage and computational resources for processing omics data. Laboratory-hosted servers require investments in informatics support for configuring and using software. Such servers are not only expensive to setup and maintain, but do not meet the dynamic requirements of different workflows for processing omics data, leading to either extravagant or sub-optimal servers. One promising technology to close the gap between generation and handling of omics data is cloud computing [24, 25] . The integration of omics data is both a challenge and an opportunity in biostatistics and biomathematics that is an increasing reality with the decreasing costs of omics profiles. Omics data embody a large mixture of signals and errors, where our current ability to identify novel associations comes at the cost of tolerating larger error thresholds in the context of big data. Major investments need to be made in the fields of bioinformatics, biomathematics, and biostatistics to develop translational analyses of omics data and make the best use of high-throughput technologies [26]

III. CONCLUSION

Despite the remarkable progress in these technologies, the analysis and mining of existing genomics data are still challenging due to the complexity of genetic inheritance, metabolic partitioning, and developmental regulations. Integration of knowledge from omics-based research is an emerging issue because the risk factor is to identify significance, gain biological insights and promote translational research. From these perspectives, omics" technologies provide exciting new opportunities and imminent breakthrough in human DNA vaccines incorporating latest technologies.

References:

1. M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. W. Treister, Transforming Health Care Through Big Data, Institute for Health Technology Transformation, Washington DC, USA, 2013.
2. Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Big data for Health ,Ieee Journal Of Biomedical And Health Informatics, Vol. 19, No. 4, Pp. 1193-1208, 2015.
3. Bernard Marr, How Big Data Is Transforming into medicine,Forbes, Google Alerts, 2016
4. A. R. Zlotta, Words of wisdom: Re: Genome sequencing identifies a basis for everolimus sensitivity, vol. 64, p. 516, 2013.
5. G. Iyer, A. J. Hanrahan, M. I. Milowsky, H. Al-Ahmadie, S. N. Scott, M. Janakiraman, M. Pirun, C. Sander, N. D. Socci, I. Ostrovnaya, A. Viale, A. Heguy, L. Peng, T. A. Chan, B. Bochner, D. F. Bajorin, M. F. Berger, B. S. Taylor, and D. B. Solit, Genome sequencing identifies a basis for everolimus sensitivity, Science, vol. 338, pp. 221–223, 2012.
6. M. Hilario and A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, Briefings Bioinformat., vol. 9, pp. 102– 118, 2008.
7. G.-Z. Yang, J. Andreu-Perez, X. Hu, and S. Thiemjarus, Multi-sensor fusion, in Body Sensor Networks. New York, NY, USA: Springer, , pp. 301–354, 2014

8. E. A. Mohammed, B. H. Far, and C. Naugler, Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends, BioData Mining, vol. 7, no. 22, pp. 1–23, 2014.
9. Knoppers, B.M. et al. ,Sampling populations of humans across the world: ELSI issues. Annu. Rev. Genomics Hum. Genet. 13, 395–413, 2012
10. Skow, A. et al. ,The association between Parkinson's disease and anti-epilepsy drug carbamazepine: a case-control study using the UK General Practice Research Database. Br. J. Clin. Pharmacol, 2013.
11. Costa, F.F. Basic research, applied medicine and EHRs: are we on the right track? J. Cancer Sci. Ther. 3, 1948–5956, 2011.
12. Anon., and Illumina Forge Consumer Genomics Goliath. Bio-IT World Magazine, 2012
13. Villars, R.L. et al. Big Data: What It is and Why You Should Care. IDC, 2011
14. Scarpati, J. Big Data Analysis in the Cloud: Storage, Network and Server Challenges. CloudProvider,2012.
15. Vanacek, J. ,How cloud and big data are impacting the human genome: touching 7 billion lives. Forbes 16 ,2012.
16. Drug Discovery Today Volume 00, Number 00 November 2013 R, Big data in biomedicine, Fabricio F. Costa
17. Greene, C.S. et al., Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. Bioinformatics 26, 694–695 , 2010.
18. Yung, L.S. et al. ,GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 27, 1309–1310 , 2011.
19. Hussong, R. et al., Highly accelerated feature detection in proteomics data sets using modern graphics processing units. Bioinformatics 25, 1937–1943, 2009
20. Jia, P. et al. ,MetaBinG: using GPUs to accelerate metagenomic sequence classification. PLoS ONE 6, e25353, 2011.
21. Melissa McCormack, What Is Big Data in Healthcare, and Who's Already Doing It, Google Blogs, 2013.
22. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Hafen R, Ramey J, et al. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin Med Diagn.Vol.7, No.1, Pp.37–51, 2013.
23. Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. New Biotechnol. ,Vol.29, No.6,Pp.613–24, 2012.
24. Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotech,Vol. 28, No.7,Pp.691–3, 2010.
25. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet,Vol.12,No.3, Pp.224, 2011
26. Akram Alyass, Michelle Turcotte and David Meyre, From big data analysis to personalized medicine for all: challenges and opportunities, BMC Medical Genomics, vol.8, No.10, 2015

Authors Biography`

Ms. S. Geeitha has completed Master of Engineering in Computer Science and Engineering in Anna University Application. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing and ontology. She has presented 10 papers in national and international conferences in the above fields. She is currently working as Assistant Professor in Mahendra Engineering College for Women.

Dr. M. Thangamani possesses nearly 20 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 70 articles in refereed and indexed journals, books and book chapters and presented over 67 papers in national and international conferences in above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has organized many self-supporting and government sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She continues to actively serve the academic and research communities. She is on the editorial board and reviewing committee of leading research journals, which includes her nomination as the Associate Editor to International Journal of Entrepreneurship and Small & Medium Enterprises at Nepal and on the program committee of top international data mining and soft computing conferences in various countries. She is also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and countries like Malaysia, Thailand and China. She has Life Membership in ISTE, Member in CSI, International Association of Engineers and Computer Scientists in China, IAENG, IRES, Athens Institute for Education and Research and Life member in Analytical Society of India. She is currently working as Assistant Professor at Kongu Engineering College at Perundurai, Erode District.