Int J Adv Life Sci Res. Volume 1(4)35-38
doi: 10.31632/ijalsr.2018v01i04.005

International Journal of Advancement in Life Sciences Research

Online ISSN: 2581-4877

journal homepage http://ijalsr.org

Review Article

Progress Towards and Challenges in Biological Big Data

Manisha Sritharan[1], Farhat A. Avin[2]**

[1]Department of Science and Biotechnology, Faculty of Engineering and Life Sciences, University of Selangor, 45600 Bestari Jaya, Selangor, Malaysia
[2]Department of Biotechnology, Faculty of Science, Lincoln University College, 47301 Petaling Jaya, Selangor, Malaysia

**Correspondence E-mail : farhat.avin@gmail.com*

Abstract

Biological big data represents a vast amount of data in bioinformatics and this could lead to the transformation of the research pattern into large scale. In medical research, a large amount of data can be generated from tools including genomic sequencing machines. The availability of advanced tools and modern technology has become the main reason for the expansion of biological data in a huge amount. Such immense data should be utilized in an efficient manner in order to distribute this valuable information. Besides that, storing and dealing with those big data has become a great challenge as the data generation are tremendously increasing over years. As well, the blast of data in healthcare systems and biomedical research appeal for an immediate solution as health care requires a compact integration of biomedical data. Thus, researchers should make use of this available big data for analysis rather than keep creating new data as they could provide meaningful information with the use of current advanced bioinformatics tools.

Keywords: *Big data, Bioinformatics, Biomedical, Database, Sequence*

Introduction

Big data is a term that is used to elucidate the huge data with certain features such as data volume, data type (variety), data generation speed (velocity), data unpredictability (variability) and data quality (veracity) (Greene et al., 2014, Yang et al., 2017). The generation of big data which evaluate the complex biological system is feasible with modern technologies such as next-generation sequencing (Nekrutenko and Taylor, 2012, Schuster, 2008). The big data grows rapidly due to the generation of data from the diverse sources and quick conversion of digital technologies in this modern era (Kashyap *et al.*, 2015, Yang *et al.*, 2017, Stephens *et al.*, 2015). In addition, the terabytes of data are generated in the large depository databases every day through modern information systems (Marx, 2013). Hence, big data analysis has become the recent area of interest in research and development (Acharjya and Ahmed, 2016). Managing the big data and connecting them is critical particularly in the biological/ biomedical research area (Marx, 2013). Moreover, accessible data are not completely employed as it requires basic skills and upgraded solutions (Wang *et al.*, 2017). The aim of this short review is to discuss the progress towards and challenges in biomedical and biological Big Data.

The progress towards

The blast of data in healthcare systems, biotechnology and biomedical research appeal for an immediate solution as health care

requires a compact integration of biomedical data (Luo *et al.*, 2016, Liang and Kelemen, 2016). This is important for stimulating personalized medicine as well as better treatments (Alyass *et al.*, 2015). Nevertheless, the presence of the latest solution to engage with this large volume of data is significant as it is directly linked to such progress (Merelli *et al.*, 2014, Kozubek, 2018). Apart from that, sequencing price continuously declining and this leads to the rapid increase of sequence data (Wetterstrand, 2013). Hence, it requires a new method for storing these data and analyzing them (Levy and Myers, 2016). Therefore, arise of cloud computing assures the place for dealing the generation of a large amount of sequence data (Liu *et al.*, 2014, Zhao *et al.*, 2017). Cloud computing is the use of network server for storing, managing, and processing the big scale data (Hashem *et al.,* 2015). However, variability in price pattern of the latest computing trend will vitally affect the response of funding agencies. Besides that, it also affects the researchers to approach data analysis as accessing cloud computing is costly (Muir *et al.*, 2016, Wall *et al*., 2010).

The challenges

Genomic is looking forward to producing data within 2 to 40 exabytes per year (Check Hayden, 2015). This is caused by data volume being generated in genomics day by day that is doubling every seven months (Check Hayden, 2015). However, the huge size of the data is not the only issue the field needs to overcome. Collection of data from many locations and various format causes difficulty in utilizing the over datasets according to a study (Gebelhoff, 2015). On top of that, most of the data produced nowadays are not properly structured, standardized and organized (Kho *et al.,* 2013, Lathe *et al*., 2008). Thus, this leads to many technical issues and inappropriate interpretation (Simon, 2008). There is a perspective that Big Data might cause extra problems than it can be solved by now to 2020 (Anderson and Rainie, July 20, 2012). Besides that, processing big data in a timely mean and analyzing them eventually to provide meaningful inference are much challenging (McAfee *et al.,* 2012). In addition, Hadoop is open source software that provides a vital platform for analysis of big data. It allocates immense storage and facilitates of rapid data processing (Chen et al., 2014). Thus, steps are taken by Google via MapReduce in order to develop Hadoop (Ware *et al.*, 2017). Furthermore, a high level of Java expertise is needed for developing parallelized programs so that Hadoop can be programmed, yet it could be another hurdle (O'Driscoll *et al.*, 2013).

Scalability and validation of data is the general concern for analysis of bioinformatics genomic big data. The former can be handled by conceptually associated methods such as divide-and-conquer. It increases scalability and multiple executions for better validation. Besides that, an effective bioinformatics approach with a quality assurance could be implemented using metamorphic testing (Yang *et al.*, 2017). Metamorphic testing is a technique for identifying oracle problem and validating the scientific computing dependent machine learning program (Xie *et al*., 2009). The way of conducting research in medical science should be changed in this big data era. It should be changed from individual academic investigation to more collaborative research using the well-organized techniques. Researchers should focus on crucial life/ health networks including dynamics rather than static and statistic features of the big data (Li and Chen, 2014).

Conclusion

In a nutshell, it is very obvious that researchers are capable of producing huge amounts of data rapidly and cost-efficiently. This is due to the presence of advanced technology that leads to the current big data. However, there is a distinctive challenge caused by this big data such as storing, managing, transferring, and analyzing them. The available tools and skills can assist researchers to utilize the technology for interpreting the data efficiently. Hence, they should make use of the big data to create a network instead of focusing in a particular area. As there is no point of keep multiplying the data without analyzing the existing data.

References

Acharjya, D. P. & Ahmed, K. (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications,* 7(2)**,** 511-518.

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics,* 8**,** 33. https://doi.org/10.1186/s12920-015-0108-y

Anderson, J. & Rainie, L. (2012). The future of big data. *Pew Research Center Internet & Technology*.

Check Hayden, E. (2015). Genome researchers raise alarm over big data. *Nature News*.

Chen, M., Mao, S. & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications,* 19(2)**,** 171-209.

Gebelhoff, R. (2015). Sequencing the genome creates so much data we don't know what to do with it. *The Washington Post* 1-3.

Greene, C. S., Tan, J., Ung, M., Moore, J. H. & Cheng, C. (2014). Big data bioinformatics. *Journal of Cellular Physiology,* 229(12)**,** 1896-1900.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems,* 47**,** 98-115.

Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S. & Bhattacharyya, D. K. (2015). Big data analytics in bioinformatics: A machine learning perspective. *Journal of Latex Class Files,* 13**,** 20.

Kho, A. N., Rasmussen, L. V., Connolly, J. J., Peissig, P. L., Starren, J., Hakonarson, H. & Hayes, M. G. (2013). Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine,* 15**,** 772-778.

Kozubek, J. (2018). *Modern prometheus: Editing the human genome with crispr-cas9*, Cambridge University Press.

Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008). Genomic Data Resources: Challenges and Promises. *Nature Education,* 1**,** 1.

Levy, S. E. & Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics,* 17**,** 95-115.

Li, Y. & Chen, L. (2014). Big Biological Data: Challenges and Opportunities. *Genomics Proteomics Bioinformatics,* 12**,** 187-189.

Liang, Y. & Kelemen, A. (2016). Big Data science and its applications in health and medical research: Challenges and opportunities. *Austin Journal of Biometrics & Biostatistics,* 7.

Liu, B., Madduri, R. K., Sotomayor, B., Chard, K., Lacinski, L., Dave, U. J., Li, J., Liu, C. & Foster, I. T. (2014). Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *Journal of Biomedical Informatics,* 49**,** 119-133.

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**,** 1-10.

Marx, V. (2013). Biology: The big challenges of big data. Nature Publishing Group.

Mcafee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. & Barton, D. (2012). Big Data: The Management Revolution. *Harvard Business Review,* 90**,** 1-9.

Merelli, I., Perez-Sanchez-Sanchez, H., Gesing, S. & D'Agostino, D. (2014). Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed research international,* 2014**,** 13.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F. & Rozowsky, J. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology,* 17**,** 53.

Nekrutenko, A. & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews Genetics,* 13**,** 667-672.

O'Driscoll, A., Daugelaite, J. & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. *Journal of biomedical informatics,* 46**,** 774-781.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature methods,* 5**,** 16-18.

Simon, R. (2008). Interpretation of Genomic Data: Questions and Answers. Seminars in hematology. *Elsevier*, 196-204.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. & Robinson, G. E. (2015). Big data: astronomical or genomical? *PLOS BIOLOGY,* 13.

Wall, D. P., Kudtarkar, P., Fusaro, V. A., Pivovarov, R., Patil, P. & Tonellato, P. J. (2010). Cloud computing for comparative genomics. *BMC Bioinformatics,* 11**,** 259.

Wang, X., Williams, C., Liu, Z. H. & Croghan, J. (2017). Big data management challenges in health research—a literature review. *Briefings in bioinformatics***,** 1-12.

Ware, A., Janvale, G., Shaikh, F. & Harke, S. (2017). HADOOP: Solution for Big Data Challenges in Bioinformatics and its Prospective in India. *Journal of Computer Engineering***,** 51-54.

Wetterstrand, K. A. (2013). DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP).

Xie, X., Ho, J., Murphy, C., Kaiser, G., Xu, B. & Chen, T. Y. (2009). Application of Metamorphic Testing to Supervised Classifiers. Ninth International Conference on Quality Software. IEEE Computer Society, 135-144.

Yang, A., Troup, M. & Ho, J. W. (2017). Scalability and Validation of Big Data Boinformatics Software. *Computational and Structural Biotechnology Journal,* 15**,** 379-386.

Zhao, S., Watrous, K., Zhang, C. & Zhang, B. (2017). Cloud Computing for Next-Generation Sequencing Data Analysis. *Cloud Computing-Architecture and Applications, InTech, Rijeka***,** 29-51.