Commentary

Big Data Analytics Enhancing: Cloud Computing

Jiang Lei*

Department of Digital Technology, Raytheon Technologies Research Center, Waltham, Massachusetts, USA

ABOUT THE STUDY

Researchers can use high-throughput studies to investigate complicated multifactorial disorders using large-scale omics data analysis. Storage, analysis, and sharing of such high-dimensional data collections are all challenges. In the bioinformatics sector, recent advances in computational technologies and methodologies, particularly in cloud computing, provide a promising, low-cost, and highly adaptable alternative. Cloud computing is quickly becoming more useful in molecular modeling, omics data analytics (e.g., RNA sequencing, metabolomics, or proteomics data sets), and phenotypic data integration, analysis, and interpretation. We look at how modern cloud-based and big data technologies are being used to handle and analyze omics data, as well as providing insights into cutting-edge cloud bioinformatics applications.

Several unique big data innovations and scalable cloud-based solutions have been suggested and developed in recent years to address the storage and analytical issues brought by the advent of omics technology. Advanced big data analytics frameworks provide the deployment of scalable analytic infrastructures, such as the Hadoop Distributed File System (HDFS) for storage and the Spark Machine Learning libraries (MLlib) for analysis, to speed up the storage and processing of massive omics data. Big data and cloud computing technologies must be closely integrated and utilized in a consistent manner to support enhanced bio-data analytics. For the study of genetic data, cloud computing has been shown to be scalable across single workstations, clusters, and public cloud infrastructures.

Security, confidentiality, and restricted cloud management technologies are among the limits of existing data processes targeted at high-throughput experimental analytics (also known as multi-omics data). Users may utilise complex pipelines or processes to convert and analyse multi-omics data in the cloud, reducing upload and download time while taking use of cost-effective computing resources. The Cancer Genome Atlas (TCGA) 2 project, for example, is now making its data available via an Application Programming Interface (API) on a variety of public and private cloud repositories, making it one of the largest and most extensive cancer genomics data sets available. These initiatives enable secure, effective, and repeatable analysis of shared data sets and provide feasible solutions for redundant and expensive local infrastructure setups.

Scalable, cloud-based systems, such as the NCI Cloud Pilots Programme Fire Cloud, may therefore be constructed, reducing the requirement for ad hoc, in-house high-performance computing architectures and costly data transport. Currently, cloud infrastructures confront considerable issues in delivering the resources required to process the quickly growing, diverse, and large-scale omics data. These obstacles have a direct impact on our capacity to better understand illness pathobiology and pathophysiology, leading to the identification of multifactorial genetic disease–related biomarkers for enhanced customised and targeted health care solutions.

As a result, there is an undeniable need to develop novel standardized approaches that will cater efficient multimodal multi-omics integrative analytics that are leveraging cloud computing infrastructures, which are edging us, closer to the tantalizing potential of a sustainable, secure, scalable, and cost-effective technology that can address this challenge.

Correspondence to: Jiang Lei, Department of Digital Technology, Raytheon Technologies Research Center, Waltham, Massachusetts, USA, E-mail: jianglei34@jl.edu