Big data analytics in bioinformatics

Zahra rajaei*
Department of IT in Water and Wastewater Zone 4 of Isfahan iran
*rajaei_z@yahoo.com

Abstract: Bioinformatics research is characterized by voluminous and incremental datasets and complex data analytics methods. The machine learning methods used in bioinformatics are iterative and parallel. These methods can be scaled to handle big data using the distributed and parallel computing technologies. Usually big data tools perform computation in batch mode and are not optimized for iterative processing and high data dependency among operations. In the recent years, parallel, incremental, and multi-view machine learning algorithms have been proposed. Similarly, graph-based architectures and in-memory big data tools have been developed to minimize I/O cost and optimize iterative processing. However, standard big data architectures are still lacking. Also appropriate tools are not available for many important bioinformatics problems, such as fast construction of co-expression and regulatory networks and salient module identification, detection of complexes over growing protein-protein interaction data, fast analysis of massive DNA, RNA, and protein sequence data, and fast querying on incremental and heterogeneous disease networks. This paper addresses the issues and challenges posed by several big data problems in bioinformatics, and gives an overview of the state of the art and the future research opportunities

Keywords: Big data; Bioinformatics; Machine learning; MapReduce; Clustering; Gene regulatory network

References

[2] Aggarwal CC, Reddy CK (eds)(2013) Data clustering: algorithms and applications. CRC Press

[3] Agrawal R, Imielin´ski T, Swami A (1993) Mining association rules between sets of items in large databases. In: ACM SIGMOD Record, vol 22. ACM, pp 207–216