OPEN ACCESS
Freely Available Online

JOURNAL OF BIG DATA RESEARCH

ISSN NO: 2768-0207

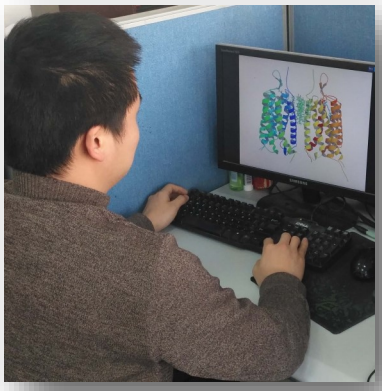
Editorial

DOI: 10.14302/issn.2768-0207.jbr-17-1925

Big Data Research: Database and Computing

Qifeng Bai[1,*]

[1]Key Laboratory of Preclinical Study for New Drugs of Gansu Province, School of Basic Medical Sciences, Lanzhou University, Lanzhou, Gansu 730000, P. R. China



Abstract

Big data research has become popular and exciting studies in almost all scientific fields such as biology, chemistry, epidemiology, medicine and drug discovery. The various systems and platforms produce large amounts of data every day. It will be very helpful for the researchers and workers to deal with big data if the practical database and useful software are introduced in time. The Journal of Big Data Research (JBR) supplies an efficient and open access publishing platform for big data research. The first issue of JBR aims to foster the dissemination of high-quality big data studies in the biological, medical and chemical database as well as the new algorithm and software for big data processing. The database and computing framework are selected to introduce the development of big data in the biological, medicine and drug discovery. The mature and functional database can be serviced in big data research of scientific fields. It promotes the scientists to extract the useful and essential dataset from the massive data. The grid computing and cloud computing supplies a new paradigm that offers an effective framework of computing and services. The research papers are welcomed from the scopes of the practical database, new algorithm and software for big data studies. All these kinds of papers not only provide the effective application methods and platforms, but also give a good promising future for big data research.

Corresponding Author: Qifeng Bai, Key Laboratory of Preclinical Study for New Drugs of Gansu Province, School of Basic Medical Sciences, Lanzhou University, Lanzhou, Gansu 730000, P. R. China. Email: baiqf@lzu.edu.cn

Received: Dec 28, 2017 **Accepted:** Apr 02, 2018 **Published:** Apr 06, 2018

Big Data and Database

With the development of network, storage and computing technology, the big data research has been permeated widely into biology, chemistry, medicine, physics and so on. For instance, the PubChem and ZINC databases (see Figure 1), which are on behalf of chemical molecules big data and information, have recruited 2D, 3D structures, chemical and physical properties, bioassay, vendors, literature, names and identifiers of compounds. ZINC database is a popular and free database of commercially available compounds for drug discovery. In the earlier version of ZINC, it only contains 727,842 purchasable ligands in 3D structure[1]. As the recently reported in official ZINC, the number of purchasable compounds in ready-to-dock and 3D formats has reached over 35 million[2]. PubChem database, which is maintained by the National Center for Biotechnology Information (NCBI), has contained about two hundred million chemical substances with the information of activities and biological assays[3,4]. The PubChem supplies the big data of chemical structures, bioactivity, spectra data, health & safety, patent identifier of small molecules and so on. In combination with the drug designing software, the small molecules big data accelerates greatly the speed of drug discovery. In the past two decades, more and more ligands are found for treating the disease by screening the chemical molecules database[5,6]. The big data is not only coming into the drug discovery field, but also promotes the development of other fields such as medicine, chemistry and finance. The Journal of Big Data Research (JBR) is born at the right moments. It gives the opportunities for the scientists who want to publish their original and high-quality research papers. The authors are welcome to contribute their popular studies into JBR which are involved in the databases with the capabilities of functional and service in scientific fields.

Big Data and Computing

If the big data is just the huge and mountains of data, it is worthless and wastes the time of researcher. The big data need be classified and mined for the important and useful information by the algorithm, program and software. In order to deal with the big data, the computer cluster can be used to run program and software for data mining. However, because the computer cluster need take enough infrastructure capital, it is not suitable for the small labs and research groups. Moreover, the file system of computer cluster has the limit number of files. For example, if the number of files exceeds 4,294,967,295 (2^{32}-1) on NTFS file system[7], the operating system will be halted. So the single file system does not seem to deal with big data which contains the huge amount of files. Grid computing is a way to use the internet idle computing resources from multiple locations for the big data processing and analysis. Because grid computing does not rely on the single file system, it can handle the huge amount of data on different computers with Linux, Windows, and MAC operating systems. Grid computing can be built on basis of a distributed computing system through many computers in different regions. For example, the JPPF (www.jppf.org), which is a good open source parallelize framework for grid computing, can be used to deal with the intensive tasks for big data by forming the grid computing network (see Figure 1). The JPPF have been reported to build grid computing network for drug discovery via drivers, nodes components and MolGridCal[8]. Grid computing is very helpful in the commercial enterprises and scientific research such as economic forecasting, marine climatic forecast, earthquake simulation, weather modeling, protein folding and drug discovery. Because grid computing does not use the universal file system, it must consider the format features of files in different operating systems such as Windows, Linux, and Mac. In addition, it need cost a lot of operating expenses and capital to maintain the large data backup, system images and associated hardware. Cloud computing supplies the higher-level services with minimal management effort from full computer infrastructure via the Internet[9] (see Figure 1). Apache Hadoop is a typical cloud computing open source framework which provides the Hadoop Distributed File System (HDFS) for distributed data backup and MapReduce framework for dealing with big data. Besides, Apache Spark and Apache Flink have the faster engine than Hadoop for big data processing. Cloud computing has been applied to commercial enterprises and scientific research such drug design[10] and molecular dynamics simulations[11]. In the first issue

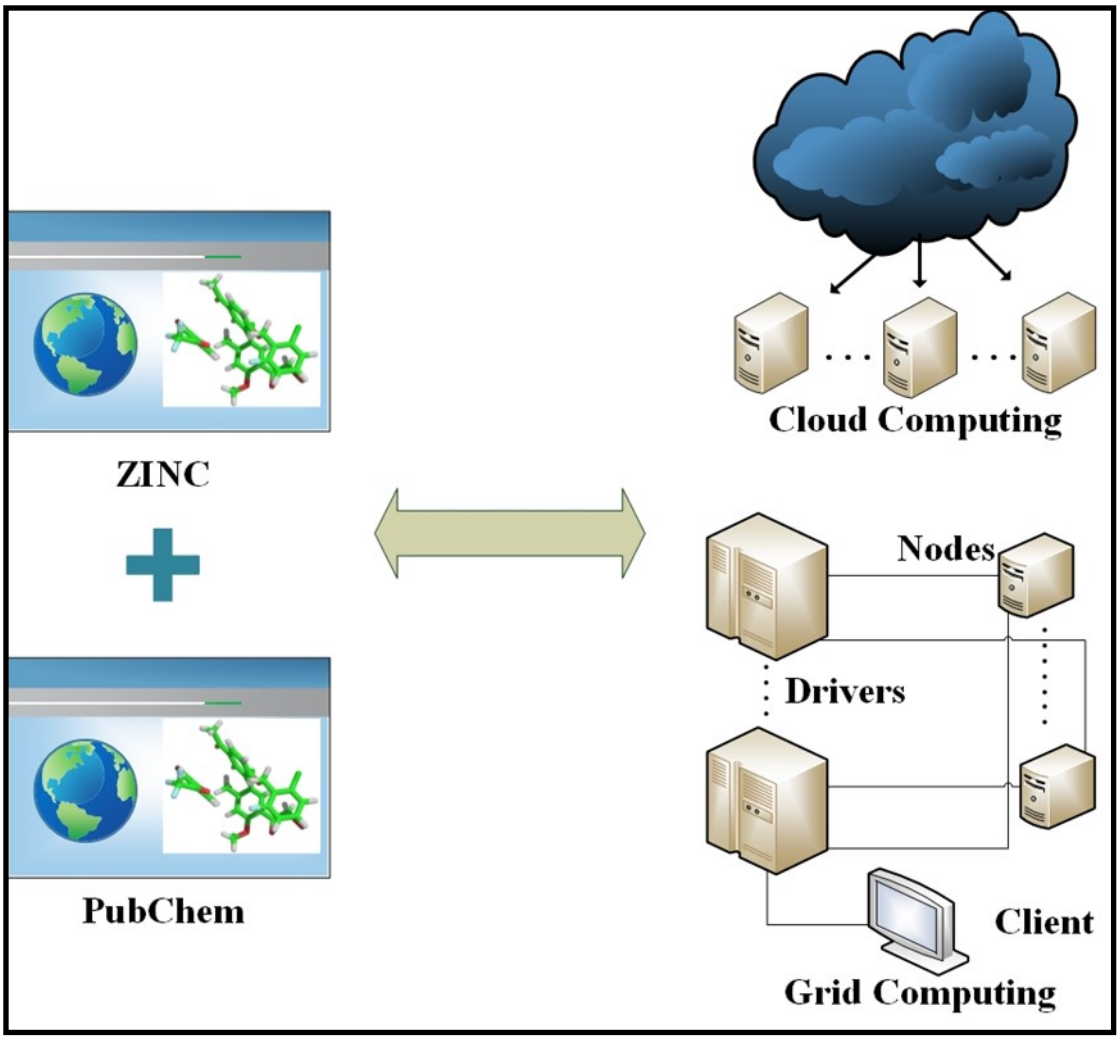


Figure 1. Databases, grid and cloud computing of big data.

of JBR, it aims to the new ideas, algorithms, software and applications which contribute to big data processing.

Conclusions

The practical database, new algorithm and software are important for big data processing and give a good perspective future for big data research. Generally, the first issue of JBR are honestly to receive the innovative and popular studies from professors, researchers, and students etc. who are interesting in publishing and updating their knowledge for big data research.

Acknowledgements

Gratefully acknowledge support from the National Natural Science Foundation of China (Grant No. 21605066)

References

1. Irwin, J. J. & Shoichet, B. K. (2005) J. Chem. Inf. Model. 45, 177-182.
2. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. (2012) J. Chem. Inf. Model. 52, 1757-1768.
3. Wang, Y., Bryant, S. H., Cheng, T., Wang, J., Gindulyte, A. *et al.* (2017) Nucleic Acids Res. 45, D955-D963.
4. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. (2008) Annu. Rep. Comput. Chem. 4, 217-241.
5. Shoichet, B. K. (2004) Nature 432, 862-865.
6. Fradera, X. & Babaoglu, K. (2017) Curr. Protoc. Chem. Biol. 9, 196-212.
7. Hayashida, S. in *SoMeT.* 203-211.
8. Bai, Q., Shao, Y., Pan, D., Zhang, Y., Liu, H. *et al.* (2014) PLoS One 9, e107837.
9. Hashemi, S. M. & Bardsiri, A. K. (2012) ARPN journal of systems and software 2, 188-194.
10. Hsu, C.-H., Lin, C.-Y., Ouyang, M. & Guo, Y. K. (2013) BioMed Research International 2013, 3.
11. Harvey, M. J., Giupponi, G. & Fabritiis, G. D. (2009) J. Chem. Theory Comput. 5, 1632-1639.