PERSPECTIVES IN GENOMICS

Perspectives of Bioinformatics in Big Data Era

Maozu Guo[1] and Quan Zou[2,3,*]

[1]*School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, P.R. China;* [2]*Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, P.R. China;* [3]*Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, P.R. China*

With the development of sequencing techniques, a growing number of data has been generated, especially the sequencing data. "Big data era" is a coined term from the internet and computer science. However, it is also popular in the bioinformatics researches due to the big scholar fundings, such as Human Genome Program, Human Microbiome Project, 1000 Genomes, *etc*. We would like to discuss the bioinformatics big data problems, and share some of our comments and perspectives.

First of all, we should discuss the next generation sequencing data. Massive short reads are generated from the latest sequencers, including Illumina, Hiseq, Roche 454, *etc*. They should be mapped into the reference genomes, or assembled with ab initio techniques. Although related software tools have been developed for decades, single thread or stand-alone computer program could not satisfy the users. Parallel platforms are being employed for the massive data, such as Hadoop [1, 2], Spark [3, 4], Cuda [5], MIC [6], *etc*. However, most of these paralleled works focused on the alignment problems, including multiple sequence alignment and sequencing reads mapping. *Ab initio* assembling problem is neglected because big graphs are difficult to handle in parallel. Big graphs always require huge memory and it is an uneasy process to divide and conquer. Therefore, it is suggested that traditional progressive assembling with parallel mechanism would be an interesting direction, where the big graph problem can be resolved.

Big data are difficult to handle for limited memory and low level configuration computers. Therefore, it is suggested that the bioinformatics researchers should develop more fast algorithms and parallel programs. However, sometimes they are gospel for coders. Some of the benefits of big data is deep learning. As the genomic sequences have grown, researchers could employ deep learning and Convolutional Neural Network (CNN) to solve the DNA/protein sequence feature extraction problem [7]. The deep learning techniques are used in protein subcellular localization [8], DNA/RNA/protein modification prediction [9, 10], drug and target prediction [11], *etc*. A number of surveys [12] have been conducted on the deep learning in bioinformatics.

Besides deep learning and parallel computation, traditional bioinformatics researches also generate interesting points, as shown in our special issue before, including genome reconstruction [13], protein submitochondrial locations [14], and noncoding variants functional prioritization [15]. Traditional techniques, such as support vector machine, are not weeded out for big data. It is not easy to assess how much data belong to big data. Therefore, it is assumed that there are some overfitting works that claimed to be big data and employed deep learning related techniques arbitrarily. In the future, researchers should pay more attention on the overfitting problem in the big data era. Moreover, visualization is also an essential topic for big data research, including sequence alignment, network illustration, sample distribution, molecular space simulation, *etc*. The future of big data seems to be promising for the bioinformatics researches.

*Address correspondence to this author at the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China; E-mail: zouquan@nclab.net

1389-2029/19 $58.00+.00

REFERENCES

[1] Zou, Q.; Li, X.B.; Jiang, W.R.; Lin, Z.Y.; Li, G.L.; Chen, K. Survey of MapReduce frame operation in bioinformatics. *Brief. Bioinform.*, **2014**, *15*(4), 637-647.

[2] Zou, Q.; Wan, S; Zeng, X.; Ma, Z.S. Reconstructing evolutionary trees in parallel for massive sequences. *BMC Sys. Biol.*, **2017**, *11*(6), 100.

[3] Guo, R.; Guo, R.; Zhou, Y.; Fang, X.; Peng, S. Bioinformatics applications on apache spark. *Gigascience*, **2018**, *7*(8), 98.

[4] Wan, S.; Zou, Q. HAlign-II: Efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing. *Algorithms Mol. Biol.,* **2017**, *12*(1), 25.

[5] Chen, X.; Wang, C.; Tang, S.; Ye, C.; Zou, Q. CMSA: A heterogeneous CPU/GPU computing system for multiple similar RNA/DNA sequence alignment. *BMC Bioinform.*, **2017**, *18*(1), 315.

[6] Peng, S.L.; Cheng, M.; Huang, K.; Cui, Y.; Zhang, Z.; Guo, R.; Zhang, X.; Yang, S.; Liao, X.; Lu, Y.; Zou, Q.; Shi, B. Efficient computation of motif discovery on Intel Many Integrated Core (MIC) Architecture. *BMC Bioinformatics*, **2018**, *19*(1), 10.

[7] Peng, L.; Peng, M.; Liao, B.; Huang, G.; Li, W.; Xie, D. The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform*., **2018**, *13*(4), 352-359.

[8] Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distribut. Comput.,* **2018**, *117*(1), 212-217.

[9] Wei, L.; Su, R.; Wang, B.; Li, X.; Zou, Q. Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. *Neurocomputing*, **2019**, *324*(1), 3-9.

[10] Zou, Q.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene subsequence embedding for prediction of mammalian n6‐methyladenosine sites from mrna. *RNA,* **2018**, *25*(2), 205-218.

[11] Yu, L.; Sun, X.; Tian, S.; Shi, X.; Yan, Y. Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr. Bioinform.*, **2018**, *13*(3), 253-259.

[12] Zhang, Z., Zhao, Y.; Liao, X.; Shi, W.; Li, K.; Zou, Q.; Peng, S. Deep learning in omics: A survey and guideline. *Brief Funct. Genom.*, **2018,** doi: 10.1093/bfgp/ely030.

[13] Bing, F.; Lingxi, and T. Jijun, Ancestral genome reconstruction on whole genome level. *Curr. Genome.,* **2017**, *18*(4), 306-315.

[14] Pu-Feng, D. Predicting protein submitochondrial locations: The 10th Anniversary. *Curr. Genome.*, **2017**, *18*(4), 316-321.

[15] Haoyue, F.; Lianping, Y.; Xiangde, Z. Noncoding variants functional prioritization methods based on predicted regulatory factor binding sites. *Curr. Genome*., **2017**, *18*(4), 322-331.