Advanced Science and Technology Letters
Vol.107 (Ubiquitous Science and Engineering 2015), pp.70-73
http://dx.doi.org/10.14257/astl.2015.107.17

A High Efficient Encoding Scheme of Big-Volume Bio-informatics Data using a Linear Block Buffering

Seokjin Im[1] and HeeJoung Hwang[2],

[1]Dept. Of Computer Engineering, SungKyul University, Anyang,
Gyeonggido, 430-742, Korea
[2]Dept. Of Computer Science, Gachon University, Seongnam,
Gyeonggido, 461-701, Korea
[1]imseokjin@sungkyul.edu, [2] hwanghj@gachon.ac.kr

Abstract. Cloud computing upon high speed communication technologies makes it possible to use high computing power and mass storage provided by a cloud service. Especially, in order to analyze big-volume bioinformatics data, the computing power by cloud computing can be an alternative to use. However, the transmission of the big-volume data to the cloud server in order to analyze the data can be a bottleneck because of lots of time consumed for transmitting. In this paper, we propose an efficient scheme transmitting while encoding big-volume bio-informatics data to a cloud server. Also, the proposed scheme uses a linear block buffering to improve the transmission throughput. We evaluate the proposed scheme with respect to the time to be consumed for encoding and transmitting to a cloud server, and show the effectiveness by comparing the scheme with other encoding algorithms.

Keywords: Bio-informatics data Encoding, Cloud computing, Linear block buffering.

1 Introduction

Advanced medical technologies produce high-quality and big-volume bio-informatics data through the analysis of data obtained various biomedical experiments [1]. For example, Solexa technology platform produces bio data of a volume of hundreds of gigabytes and Life/SoLiD platform produces bio data of about 30GB to 50GB. To obtain meaningful information by analyzing the big-volume data, we need huge computing power and mass storage.

The computing power and storage services from a cloud computing can be an alternative to be used for analyzing the bio data [2], [3]. When we use the power and storage for the analysis, we have to transmit the big-volume bio data to a cloud through a communication connection established between a local storage and the cloud. After the transmission of the data, we analyze the data in a cloud site and obtain meaningful information. In this environment, however, the limited bandwidth can be an obstacle to transmit the data to the cloud site because the bandwidth causes delay during the transmission [4], [5].

ISSN: 2287-1233 ASTL

To avoid the bottleneck, several encoding algorithms have been developed, aiming to reduce the volume of the bio data by compression. The reduced volume of the data leads to the shortened time to be consumed for the transmission. The developed algorithms use the strategy of the transmission after complete compression of whole data. This causes additional delay and deteriorates the performances.

In this paper, we propose a scheme of encoding algorithm of big-volume bio data with high compression efficiency and transmitting the compressed data with a linear block buffering. The proposed scheme is different from the existing encoding algorithms in the aspect of transmitting the bio data while compressing the data.

The rest of the paper organizes as follows: chapter 2 presents the proposed scheme. We evaluate the performance of the proposed scheme by comparing with existing schemes in chapter 3. We conclude this paper in chapter 4.

2 The Proposed Encoding Scheme

In order to transmit big-volume bio data, e.g., NGS data, to a cloud effectively, we present an encoding scheme with high compression efficiency and transmission scheme using block buffering of compressed bio data stream. Fig. 1 shows the entire procedure of encoding and transmitting to a cloud. The proposed scheme consists of an encoder and a transmitter.

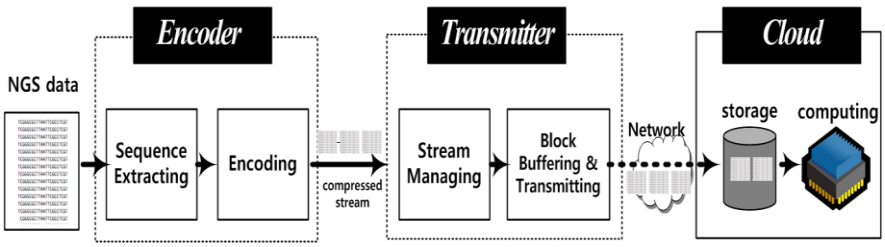


Fig. 1. The proposed scheme for encoding and transmitting bio data.

2.1 The Encoding Procedure

The proposed scheme begins with encoding, i.e., compressing the NGS data. The encoder of the proposed scheme consists of two processes of sequence extracting and encoding. In the sequence extracting, meta data blocks of NGS are extracted and then the blocks are compressed using the compression algorithm, SolidZipper [6]. The extractor separates csfasta and fastaq characteristics of DNA and additional information, e.g., ID and base, from NGS data. The encoder compresses the extracted using SolidZipper that compresses csfasta and fastaq data in units of block, adopting multi-threaded parallel process in CUDA for improve the efficiency.

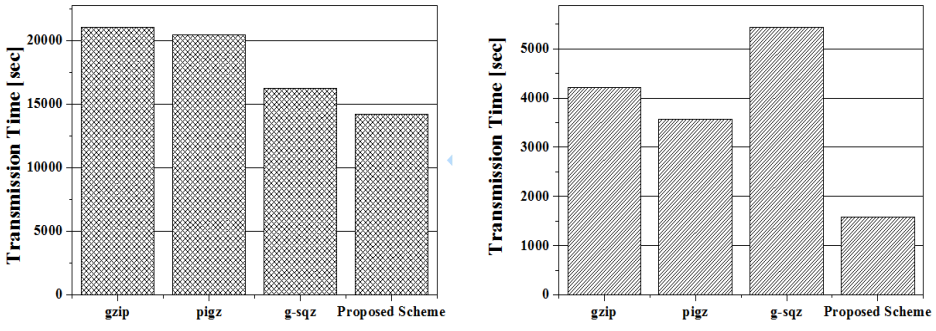
2.2 The Transmitting Procedure

The transmitting procedure uses a stream manager and transmitter to a cloud. The stream manager receives compressed blocks of bio data and pushes them into a linear buffer. Then, the manager establishes a communication connection to a cloud storage for the transmission. The transmitter pulls a compressed block from the linear buffer then sends it to the cloud storage. Using the linear buffer, the stream manager and transmitter can send each compressed block to the cloud as soon as the completion of the compression. The proposed scheme can reduce the transmission time by sending a bio data block right after compression, unlike the existing schemes that send the data after the compression of whole bio data.

3 Performance Evaluation

We evaluate the proposed scheme with respect to transmission time by comparing the existing compression schemes, i.e., gzip, pigz, and s-sqz. We implemented the proposed scheme in Java on a 64-bit Linux machine equipped with 64 bit Intel CPU and a main memory of 4 GB. We use big-volume NGS data of about 70 GB.

Fig.2 shows the effectiveness of the proposed scheme with respect to the transmission time. Fig. 2(a) reveals the transmission times of the existing compression algorithms and the proposed scheme at transfer bandwidth 10 Mbps. The transmission time of the proposed scheme is about 67% of that of gzip and 88% of that of g-sqz. This means that the strategy of the proposed scheme reduces the entire transmission time, that the compression of big-volume bio data in units of blocks and transmission the block as soon as the compression. Fig. 2(b) depicts the comparison of the transmission time at transfer bandwidth 100 Mbps. The figure shows that the proposed scheme outperforms the existing schemes.



(a) At transfer bandwidth 10Mbps (b) At transfer bandwidth 100Mbps

Fig. 2. Comparison of the transmission time for various transfer bandwidth.

4 Conclusion

In this paper, we have proposed a scheme for a quick transmission of big-volume bio-informatics data using a strategy, transmission of blocks right after compression of the data in units of block. For the strategy, the proposed scheme consists of an encoder extracting blocks from NGS data and compressing, and a transmitter pushing each compressed blocks into a linear block and sending each block to a cloud from the linear buffer. We show the effectiveness of the proposed scheme by experiments with 70 GB NGS data. The results of the experiments disclose that the proposed scheme outperforms the existing compression schemes.

References

1. Metzker, M.L.: Sequencing technologies the next generation. Nature Reviews Genetics 111, 31--46 (2010)
2. Stein, D.: The case for cloud computing in genome informatics. Genome Biology 115, 207--215 (2010)
3. Langmead B.: Searching for SNPs with cloud computing. Genome Biology, 10R134 (2009)
4. Kudtarkar P.: Cost-Effective Cloud Computing: A Case Study Using the Comparative Genomics Tool, Roundup. Evolutionary Bioinformatics 226, 197--203 (2010)
5. ZHENG W.M.: An introduction to Tsinghua Cloud. SCIENCE CHINA Information Sciences 537, 1481--1486 (2010)
6. Jeon Y.J., Hwang H.J.: SOLiDzipper: A High Speed Encoding Method for the Next-Generation Sequencing Data. Evolutionary Bioinformatics 107, 1--6 (2011)

