Genome Sequence Compression with Distributed Source Coding

Shuang Wang*, Xiaoqian Jiang*, Lijuan Cui[$], Wenrui Dai[+], Nikos Deligiannis[!], Pinghao Li[+], Hongkai Xiong[+], Samuel Cheng[$], and Lucila Ohno-Machado*

*Division of Biomedical Informatics, University of California, San Diego, San Diego, CA, 92122, USA {shw070, x1jiang, lohnomachado}@ucsd.edu

[$]School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK, 74135, USA {lj.cui, samuel.cheng}@ou.edu

[+]Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, 200240, China {fdaiwenrui, lipinghao, xionghongkai}@sjtu.edu.cn

[!]Department of Electronics and Informatics, Vrije Universiteit Brussel - iMinds, 1050 Brussels, Belgium ndeligia@etro.vub.ac.be

Introduction

Genome data are playing a significantly important role in modern medicine, e.g., personalized medicine and earlier detection of diseases. With the increasing demand in genome data, advanced sequencing techniques have been developed, among which the flexible miniaturized sequencing devices [1] are very promising, especially due to their portability and efficiency. These portable devices commonly have low processing capabilities, constrained power budget, small storage or memory, and limited communication bandwidth. In contrast, genome data are usually large in size and are highly redundant within or across sequences. There is an obvious need to compress genome data for the sake of computation and storage. Unfortunately, traditional genome compression techniques [2] fail to satisfy the aforementioned miniaturized devices, due to the facts of high computational complexity, large space and memory requirements for storing references at the encoder side.

Methods

In this paper, we develop a novel genome compression framework based on distributed source coding (DSC)[3], which is specially tailored to the need of miniaturized devices. At the encoder side, subsequences with adaptive code length can be compressed flexibly through either low complexity DSC based syndrome coding or hash coding with the decision determined by the existence of variations between source and reference known from the decoder feedback. Moreover, to tackle the variations between source and reference at the decoder, we carefully designed a factor graph based low-density parity-check (LDPC) decoder, which automatically detects insertion, deletion and substitution.

Results

The genome sequences used in our experiments are the Arabidopsis thaliana, TAIR8 dataset with TAIR9 dataset as reference. Table 1 shows a side-by-side comparison among the raw file size, the compressed file sizes, encoding complexities of GRS algorithm [2] and the proposed method, respectively. We can see that both the proposed method and GRS algorithm achieve significant file size reductions. Although, there is still existing performance gaps when comparing with GRS's, the proposed GeDSC encoder shows a significantly lower encoding complexity. To the best of our knowledge, this is the first study of DCS based genome compression. There is no doubt that it opens many possibilities for the portable miniaturized applications in which energy consumption and bandwidth usage are of paramount importance.

Table 1: Comparison of compression performances and encoding complexities of the proposed GeDSC codec and the GRS codec [2] with the Raw file sizes as reference for all 5 chromosomes. ET: encoding time; CR: Compression rate.

#	Raw file	GRS		Proposed Method		
	size (MB)	Size(KB)	ET(s)	Size(KB)	CR	ET(s)
1	29.4	0.6982	7	8.0264	3702.1	1.8292
2	19.0	0.3760	4	3.9775	4836.4	0.8918
3	22.7	2.9189	6	6.0225	3804.1	1.3947
4	17.9	1.9053	5	19.9268	910.8	1.3702
5	26.1	0.5858	6	3.1924	8251.9	1.0864

Acknowledgment

SW, XJ and LOM were funded in part by AHRQ(R01HS19913), NLM (K99LM011392) CTSA (UL1TR000100) and iDASH (NIH grant U54HL108460). L. Cui and S. Cheng were funded by NSF (CCF 1117886). N. Deligiannis is funded by the FWO Flanders projects G.0391.07 and G.0047.12.

References

[1] MinION: A miniaturised sensing instrument. Oxford Nanopore Tech. [Online]. Available: http://www.nanoporetech.com/technology/minion-aminiaturised-sensing-instrument

[2] C. Wang and D. Zhang, "A novel compression tool for efficient storage of genome resequencing data," Nucleic Acids Res., vol. 39, no. 7, p. e45, Apr. 2011.

[3] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): Design and construction," IEEE Trans. Inform. Theory, vol. 49, no. 3, pp. 626–643, Mar. 2003.

1068-0314/13 $26.00 © 2013 IEEE
DOI 10.1109/DCC.2013.104