{tag} {/tag}
International Journal of Computer Applications

Volume 86 - Number 8

Year of Publication: 2014

Authors:
Mohammad Nassef

Amr Badr

Ibrahim Farag

10.5120/15002-3134
{bibtex}pxc3893134.bib{/bibtex}

Abstract

Genome resequencing produces enormous amount of data daily. Biologists need to frequently mine this data with the provided processing and storage resources. Therefore, it becomes very critical to professionally store this data in order to efficiently browse it in a frequent manner. Reference-based Compression algorithms (RbCs) showed significant genome compression results compared to the traditional text compression algorithms. By avoiding the complete decompression of the compressed genomes, they can be browsed by performing partial decompressions at specific regions, taking lower runtime and storage resources. This paper introduces the inCompressi algorithm that is designed and implemented to efficiently pick sequences from genomes, that are compressed by an existing Reference-based Compression algorithm (RbC), through partial decompressions. Moreover, inCompressi performs a more efficient complete genome decompression compared to the original decompression algorithm. The experimental results showed a significant reduction in both runtime and memory consumption compared to the original algorithm.

Refer

ences

- R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, et al. The international hapmap project. Nature, 426(6968):789–796, 2003.
- N. Siva. 1000 genomes project. Nature biotechnology, 26(3):256–256, 2008.
- G. M. Church. The personal genome project. Molecular Systems Biology, 1(1), 2005.
- D. R. Bentley. Whole-genome re-sequencing. Current opinion in genetics & development, 16(6):545–552, 2006.
- J. Shendure and H. Ji. Next-generation dna sequencing. Nature biotechnology, 26(10):1135–1145, 2008.
- R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, et al. A map of human genome sequence variation containing 1. 42 million single nucleotide polymorphisms. Nature, 409(6822):928–933, 2001.
- L. S. Heath, A. Hou, H. Xia, and L. Zhang. A genome compression algorithm supporting manipulation. In Proc LSS Comput Syst Bioinform Conf, volume 9, pages 38–49, 2010.
- S. Kuruppu, S. J. Puglisi, and J. Zobel. Reference sequence construction for relative compression of genomes. In String Processing and Information Retrieval, pages 420–425. Springer, 2011.
- S. Deorowicz and S. Grabowski. Robust relative compression of genomes with random access. Bioinformatics, 27(21):2979–2986, 2011.
- C. Wang and D. Zhang. A novel compression tool for efficient storage of genome resequencing data. Nucleic Acids Research, 39(7):e45–e45, 2011.
- A. J. Pinho, D. Pratas, and S. P. Garcia. Green: a tool for efficient compression of genome resequencing data. Nucleic Acids Research, 40(4):e27–e27, 2012.
- B. G. Chern, I. Ochoa, A. Manolakos, A. No, K. Venkat, and T. Weissman. Reference based genome compression. In Information Theory Workshop (ITW), 2012 IEEE, pages 427–431. IEEE, 2012.
- A. D. Wyner and J. Ziv. The sliding-window lempel-ziv algorithm is asymptotically optimal. Proceedings of the IEEE, 82(6):872–877, 1994.
- M. C. Brandon, D. C. Wallace, and P. Baldi. Data structures and compression algorithms for genomic sequence data. Bioinformatics, 25(14):1731–1738, 2009.

Index Terms

Computer Science

Keywords

Reference-based Genome Compression; Partial Genome Decompression; Browsing Genome Sequences; inCompressi.