発表概要

遺伝子情報圧縮伸張プログラムCRAMの高速化

村田 浩樹[1,a)]

2013年11月12日発表

近年，遺伝子解析技術の進歩により，ヒトの遺伝子情報を 1 人当たり 1,000 ドル程度の費用で取り出すことができるようになりつつあり，日々大量のデータが生成されている．現在，この取得した遺伝子情報からその機能を解析する多くのプログラムが開発され，アルゴリズムの改良が進められている．これらのプログラムの開発者の関心の中心はアルゴリズムの改良による精度の向上にあるため，著名なプログラムであっても性能や並列化を意識せずにアルゴリズムに忠実に実装されていることが多く，マルチコアのコンピュータで十分な性能を発揮するにはプログラムの並列化・最適化が必要となるが，その際，扱う遺伝子情報の大きさが問題となる．シーケンサで取り出された生のヒトの遺伝子情報は 100 GB 程度あり，多くのデータ圧縮方法が提案されている．我々は，このような遺伝子解析プログラムの代表例として，CRAM 圧縮アルゴリズムのリファレンス実装の並列化を試みた．これは現在遺伝子情報の保存に標準的に用いられている BAM 圧縮フォーマットを，より圧縮効率が高く次世代の標準として提案されている CRAM 圧縮フォーマットに変換するもので，並列化を意識せず，アルゴリズムに忠実に実装されている．本発表では，この CRAM プログラムをアルゴリズムへの忠実さを維持しつつ並列化するにあたって重要だった点を議論する．並列化バージョンでは，BAM から CRAM への変換速度が 10.6 倍，CRAM から BAM への変換速度が 2.1 倍という結果が得られた．

Optimization of Genome Information Compression/Decompression Program: CRAM

Hiroki Murata[1,a)]

Presented: November 12, 2013

Recently huge genomic data are generated daily, based on the progress of genome analysis technology and eliminating cost to pick up human genome information with about $1,000. As the focus of these program developer is to improve the accuracy, even popular applications are implemented according to the algorithm without considering performance or parallelization. To extract enough performance from multi-core computers, the huge size of genomic information becomes problems. As the size of raw human genome data picked up by sequencer is about 100 GB, many methods of data compression have been proposed. We attempted to parallelize the reference implementation of CRAM compression algorithm. It transforms the BAM, current most popular format using reference genome information, to CRAM compression format, has higher compression efficiency and proposed as the next standard, and is implemented without considering parallelization and regarding to the algorithm. This paper describes the points to keep CRAM program naïve with the algorithm and parallelize it. The parallelized version speeds up 10.6 times to transform from BAM to CRAM and 2.1 times to transform from CRAM to BAM.

[1] 日本 IBM 東京基礎研究所
IBM Japan, IBM Research-Tokyo, Koto, Tokyo 135–8511, Japan
[a)] mrthrk@jp.ibm.com

© 2014 Information Processing Society of Japan