P34

Aligned Genomic Data Compression via Improved Modeling

Idoia Ochoa, Mikel Hernaez, Tsachy Weissman

Department of Electrical Engineering, Stanford University, USA

Abstract

With the release of the latest Next-Generation Sequencing (NGS) machine, the HiSeq X by Illumina, the cost of sequencing the whole genome of a human is expected to drop to a mere $1000. This milestone in sequencing history marks the era of affordable sequencing of individuals and opens the doors to personalized medicine. In accord, unprecedented volumes of genomic data will require storage for processing. There will be dire need not only of compressing aligned data, but also of generating compressed files that can be fed directly to downstream applications to facilitate the analysis of and inference on the data. Several approaches to this challenge have been proposed in the literature; however, focus thus far has been on the low coverage regime and most of the suggested compressors are not based on effective modeling of the data.

We demonstrate the benefit of data modeling for compressing aligned reads. Specifically, we show that, by working with data models designed for the aligned data, we can improve considerably over the best compression ratio achieved by previously proposed algorithms. Our results indicate that the pareto-optimal barrier for compression rate and speed claimed by *Bonfield et al.* (2013) does not apply for high coverage aligned data. Furthermore, our improved compression ratio is achieved by splitting the data in a manner conducive to operations in the compressed domain by downstream applications.

Availability

The software is written in C and can be downloaded from:
http://www.stanford.edu/~iochoa/cbc.html.