New Algorithms for RNA-seq and ChIP-seq Data Compression

Olgica Milenkovic, University of Illinois at Urbana-Champaign; Vida Ravanmihr, University of Illinois; Zhiying Wang, University of California; Minji Kim, University of Illinois

We report on smallWig and ChiPWig, two new software solutions for RNA-Seq and ChiP-Seq data compression. Our new low-rate compression methods are especially designed for RNA- and ChIP-seq data, which contains real valued entries with patterns that cannot be efficiently compressed by general purpose compression software.

Our approaches are based on statistical modeling of position and expression values, and source coding techniques, which include transform coding, differential coding and arithmetic coding for integers and correlated real numbers. An additional compression technique, known as context-tree weighting, was used to achieve ultrahigh compression rates needed for archival storage applications

We tested our compression methods on different RNA-Seq and ChIP-seq data generated by the ENCODE project. The results reveal that the new methods offer, on average, a 10/20-fold decrease in file size compared to bigWig, while providing the same summary statistics and random access features. Compression and decompression times are comparable to those of wig formats.