A Dynamic, User-centric Big Data Analytics Framework for Genome Data.

Shalini Ravishankar, Meeta Pradhan and Mathew Palakal

Department of Computer and Information Science, IUPUI School of Science; Department of BioHealth Informatics, IU School of Informatics and Computing

The cost to sequence DNA today has reduced from $100million to mere over $1000 and this has significantly increased the generation of genomic data multifold. However, analysis of such large data requires meeting user needs and computational challenges. There are different tools that exist to process the sequenced DNA information for alignment and research. These tools are made adaptive to work in a big data processing environment like Hadoop. However, the analysis of such sequence data is dependent on user specific needs, and hence, a unique data analysis pipeline is needed for each user. We propose a barcode driven technology to instruct a Hadoop-based big data analytics system that would allow the user to select the necessary tools to process the input genome data file. The proposed framework can dynamically generate customized barcodes for each user based on the user's data analysis need and a pipeline is created and driven by the barcode. This approach will revolutionize the way NGS data analytics pipelines are being setup by the user. This new method will provide the user with a seamless way to analyze the data. The time taken to process a genomic file was significantly reduced from 2 hours on a traditional Linux server to just 3.81 minutes on Hadoop. Our results indicate that a barcode-based approach will enable the user to customize NGS data analysis in a very efficient manner.