BioHDF : Open binary file formats for large scale data management

Todd Smith(1), Christian Chilan (2), Rishi Sinha(3), Elena Pourmal(2), Mike Folk(2).

Geospiza, Inc. 100 West Harrison St. North Tower #330, Seattle WA 98119. 2. 1901 S. First St., Suite C-2 Champaign, IL 61820. 3. Microsoft Corporation, Redmond WA.

Presenting Author: Todd Smith todd@geospiza.com

Project URL: http://hdfgroup.com/projects/bioinformatics/index.html

Project Code: http://hdfgroup.org/projects/bioinformatics/bio_software.html

Source License: BSD

The first wave of Next Generation (“Next Gen”) sequencing technologies are providing large numbers of laboratories with “Genome Center” kinds of throughput to make discoveries and develop new assays never before imagined. However, widespread adoption of Next Gen will be hindered because current bioinformatics programs do not scale; they are inefficient in data storage, processing, and memory utilization. The most popular programs typically copy and recopy data to new files many times during processing, require that all data be maintained in random access memory (RAM) when running, and cannot incrementally process data. To overcome these issues, fundamental changes in data management and processing are needed.

Geospiza and The HDF Group are collaborating to develop portable, scalable, bioinformatics technologies based on HDF5 (Hierarchical Data Format - http://www.hdfgroup.org). We call these extensible domain-specific data technologies "BioHDF.” BioHDF will implement a data model that supports primary DNA sequence information (reads, quality values, meta data) and results from sequence assembly and variation detection algorithms. BioHDF will extend HDF5 data structures and library routines with new features (indexes, compression, graph layouts) to support the high performance data storage and computation requirements of Next Gen Sequencing. BioHDF will include APIs, software tools, and a viewer based on HDFView to enable its use in the bioinformatics and research communities. Using BioHDF, researchers will be able to perform *de novo* sequencing, do resequencing-based SNP discovery, analyze genotyping data, and export datasets in formats ready for submission to key databases. As a programming environment, BioHDF can be easily extended to accept data from new data collection platforms and format data for interchange with many databases. BioHDF will be delivered to the research community as an open source technology.

In preliminary studies, HDF5’s feasibility for managing large volumes and complex biological data was tested. The first test case looked at DNA sequencing-based SNP discovery. Through this study, HDF’s strengths and data organization features (groups, sets, multidimensional arrays, transformations, linking objects, and general data storage for other binary data types and images) were evaluated to determine how well these features would handle SNP data. Other test cases were added to test the ability of HDF to handle extremely large datasets associated with HapMap data and chromosomal scale LD (Linkage Disequilibrium) calculations. Data from preliminary studies and new work with Next Gen sequence data will be presented.