BioHDF

The BioHDF project is a collaborative effort to address the bioinformatics data deluge problem. Based on the established open-source HDF5 binary data storage technology, BioHDF strives to help biologists come to terms with the flood of data that the latest instrumentation can produce. The current focus of BioHDF is on next-generation sequencing (NGS) data storage.

As we envision it, there are three key parts to BioHDF:

- The data model and file organization.

This determines which data will be stored, how it will be arranged in the data file and how it will be queried. Data will be stored as fundamental building blocks such as "sequences", "alignments" and "MS/MS spectra". Unlike most file formats, which are set in stone, BioHDF files will are self-describing, flexible and extensible as they are based on HDF5.

- The C application programming interface (API) and library.

This is the library which will provide the basic means for manipulating the data stored in a BioHDF file. C is a useful language for the basic BioHDF API since it allows for easy interfacing with the HDF5 API, can be ported easily to many operating systems and can interoperate with most higher-level languages. Much bioinformatics work is done in higher-level languages, however, and we intend to make the BioHDF API easily wrappable for these languages using packages like SWIG and XS.

- Command-line tools

Command-line tools are provided for data I/O and manipulation. Interoperability with existing bioinformatics tools will be provided by functions which allow for import and export of the data from/to existing bioinformatics file formats.

We believe that a key factor to the success of BioHDF is the participation of interested parties in the development of the data model and API. If you are being drowned in data and would like to be a participant in the development of BioHDF we encourage you to follow our progress on this website and to subscribe to our mailing list (contact link on the left). **We welcome your input!**

Documentation

Getting started manual (pdf - 144 kB)

Users guide and tools manual (pdf - 541 kB)

API documentation

BioHDF is open-source software distributed under the same BSD license as HDF5. A copy of this license can be found here.

Publications

Increasing the Scale of Deep Sequencing Data Analysis with BioHDF

> Video from a talk by Todd Smith of Geospiza in which Todd discusses how BioHDF systems can be used with next generation DNA sequencing technologies. Delivered June 3, 2010 at the "Sequencing, Finishing, Analysis in the Future" meeting at the DOE Joint Genome Institute in Santa Fe, NM.

Standardizing the Next Generation of Bioinformatics Software Development with BioHDF

> A book chapter from Advances in Computational Biology (vol. 680) (requires subscription). The chapter refers to the BioHDF prototype and not the new API and tools.

Slides (pdf) from the SC09 conference BioHDF birds-of-a-feather meeting.

A recent article about storing biological image data in HDF5.

Geospiza's CTO Todd Smith has several in-depth blog posts about using BioHDF.

The case for HDF

Scalable Bioinformatics Infrastructures with BioHDF (A five-part bloginar series)
Part 1
Part 2

Part 3
Part 4
Part 5

Downloads

The current version is 0.4 alpha (December 2011)

This release focuses on improving the build system, testing and error checking.

Linux/Unix, Mac OS X and Windows are supported.

BioHDF source code (tar.gz format)

BioHDF source code (zip format)

BioHDF Subversion repository (read-only)

BioHDF is open-source software distributed under the same BSD license as HDF5. A copy of this license can be found here.

Contact Information

We welcome your feedback!

BioHDF is for the NGS community and, in order for us to serve you well, we need your feedback. If you have any comments, criticisms or feature requests, please contact us either via the mailing list or directly. We really do appreciate your feedback and would love to hear from you.

Primary Contact

Please contact Dana Robinson (derobins at hdfgroup dot org) with any questions or concerns about BioHDF.

Advisory Committees

We are in the process of setting up "steering committees" of interested people to help guide the development of BioHDF. Membership is informal and consists of people who are willing to answer our email questions and occasionally take part in a teleconference. If you would like to participate in a steering committee, please contact Dana Robinson at the email address given above.

Team Members

Principal Investigators
Dana Robinson (and primary contact: derobins at hdfgroup dot org)
Mike Folk

Weill Cornell Medical College
Chris Mason

University of Illinois
Jian Ma

Geospiza, Inc.
Todd Smith

Funding

Current Funding

BioHDF is currently funded as an internal project of The HDF Group.

Past Funding

National Institutes of Health
NIH SBIR Phase II Grant HG003792
BioHDF- Open Binary File Formats for Bioinformatics
Todd Smith (PI) Geospiza, Inc.
Mike Folk (PI for subcontract with The HDF Group)
March 2009 - February 2011

National Institutes of Health
NIH SBIR Phase I Grant 1R41HG3792-1
BioHDF- A Phase I Small Business Technology Transfer Grant
Todd Smith (PI) Geospiza, Inc.
Mike Folk (PI for subcontract with NCSA)
September 2005 - July 2006