Structural Biology Data Grid Dataverse

Piotr Sliz,
data.sbgrid.org
Harvard Medical School, Dept. of BCMP
Boston Children’s Hospital, Dept. of Pediatrics
SBGrid Consortium

July 11th, 2016
Dataverse Community Meeting 2016
Dataverse Repositories Around the World
Sonia Barbosa
2000-2016: Growth of the community

~300 structural biology groups
(DFCI, BCH, HMS, HU, Tufts, Genzyme)

22 countries

sbgrid.org
Additional Services Provided by SBGrid in 2016

- **Software Support**
  - (~300 OS, academic, commercial applications)

- **Local Research Computing Infrastructure**
  - Structural Biology (clusters, storage, workstations)
  - Lattice Light Sheet Microscopy

- **Access to HPC Resources**
  - Open Science Grid (Portal)
  - XSEDE

- **Teaching and Training**
  - YouTube Program
  - Structural Biology Training Curriculum (workshops)

---

Example I: Lin28 crystal structure provides a static interpretation of experimental data and demonstrates limitation of traditional research strategy.

- Different software packages (e.g. XDS vs HKL2000: 3D vs 2D profile fitting)
- Different assumptions (e.g. symmetry, mosaicity, radiation damage, number of frames)
- New software packages (e.g. DIALS)
- Improved criteria (e.g. resolution limits such CC1/2, Karplus and Diederichs, 2012, or anisotropic correction)
- New corrections (e.g. data anisotropicity)
- Additional features: (e.g. anisotropic diffuse scattering signals)

Motivation

Nam, et al., 2011 Cell
Storage Requirements for Primary Datasets
(based on 100,000 PDB files and average dataset size in SBDG)

Molecular models and reduced datasets are stored in Protein Data Bank

SBDG:
110 datasets
~0.5TB

PDB:
100,000 models
0.3 TB

Primary Diffraction Datasets proposed to be stored on SB Data Grid

Some of “All Experimental Data” preserved at national synchrotrons e.g. Tardis or Diamond

Data Access

Example II: EM structure - management of datasets from EM facilities

Local Processing
2x20 Core Xeon 2600 V4s (20TB scratch)

Drift Correction (GPU)

Fast Storage

“summed” images

Local Archives

Orchestra/XSEDE

Titan Krios 300kV TEM
Talos-Arctica 200 kV

Motivation

HARVARD MEDICAL SCHOOL

K2 Computer

Drift Correction (GPU)

UMASS

K2 Computer

uncorrected data (~10TB/dataset)

WORLDWIDE PROTEIN DATA BANK

PDB
data.sbgrid.org website

Live-analysis of X-ray diffraction datasets can inject new data analysis methods.

SBGrid Consortium

Phase 1 (2015)
Prototype

Phase 2 (2016-2018)
Dataverse Implementation

The National DATA SERVICE

Dataverse Project

Prototype Implementation
X-Ray Diffraction data from Structures of the E. faecalis LiaR DNA binding domain complexed to the putative consensus sequence, source of 4WU4 structure

Shamoo, Yousif, 2016, “X-Ray Diffraction data from Structures of the E. faecalis LiaR DNA binding domain complexed to the putative consensus sequence, source of 4WU4 structure”, doi:10.15785/SBGRID/71, Root Dataverse, V1

Learn about Data Citation Standards.

Related Publication

4WU4 Coordinates
PDB, MMDB

Biological Sample
Structures of the E. faecalis LiaR DNA binding domain complexed to the putative consensus sequence

Dataset Type
X-Ray Diffraction

Subject Composition
DNA

Data Creation Date
2014-04-19
Registration followed by Data Transfer Step

- Dataset Registration
- Dataset Upload
- DOI Assignment
- HOLD
- Dataset DAA Replication
- DOI Release
- Data Release

DataCite Citation Profiles/Dataverse Extension

Upload scripts/Verification

California Digital Library
EZID System

Persistent Dataset URL
Data Access Alliance:
Globus Infrastructure

Endpoint
Globus
Connect Server

Globus Network

Yale Med. Endpoint
Globus Connect
Personal

Transfers coordinated by Globus

REST API calls

Control System
(activates endpoints)
- reporting new datasets to other systems for replication
- user initiated replication
  - retention policies
- admin initiated replication
Dataset Access Alliance: Local access through SBGrid Satellites

SBx: Desktop Tool for Structural Biologists

Jason Key

<table>
<thead>
<tr>
<th>Package</th>
<th>Installed Version</th>
<th>Current Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2ps</td>
<td>4.14</td>
<td></td>
</tr>
<tr>
<td>Aline</td>
<td>1.0.025</td>
<td></td>
</tr>
<tr>
<td>AMPS</td>
<td>2.3a</td>
<td></td>
</tr>
<tr>
<td>ARIA</td>
<td>2.3.1</td>
<td></td>
</tr>
<tr>
<td>AutoDock</td>
<td>4.2.5.1</td>
<td></td>
</tr>
<tr>
<td>BLAST</td>
<td>2.2.26</td>
<td></td>
</tr>
<tr>
<td>BLAST+</td>
<td>2.2.31</td>
<td></td>
</tr>
<tr>
<td>Bowtie</td>
<td>1.0.0</td>
<td></td>
</tr>
<tr>
<td>Bowtie 2</td>
<td>2.2.1</td>
<td></td>
</tr>
<tr>
<td>bresseq</td>
<td>1.00rc8</td>
<td></td>
</tr>
<tr>
<td>BWA</td>
<td>0.7.7-r441</td>
<td></td>
</tr>
<tr>
<td>CCP4</td>
<td>6.5</td>
<td></td>
</tr>
<tr>
<td>Chimera</td>
<td>1.10.2</td>
<td></td>
</tr>
<tr>
<td>CNS</td>
<td>1.3r8</td>
<td></td>
</tr>
<tr>
<td>CTF</td>
<td>20140609</td>
<td></td>
</tr>
<tr>
<td>Cufflinks</td>
<td>2.1.1</td>
<td></td>
</tr>
<tr>
<td>cython</td>
<td>0.21.1</td>
<td></td>
</tr>
<tr>
<td>DireX</td>
<td>0.7.0</td>
<td></td>
</tr>
<tr>
<td>EMAN</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>EMAN2</td>
<td>2.12</td>
<td></td>
</tr>
<tr>
<td>EPMR</td>
<td>15.04</td>
<td></td>
</tr>
<tr>
<td>FASTA</td>
<td>36.3.8a</td>
<td></td>
</tr>
<tr>
<td>FREALIGN</td>
<td>9.11_151013</td>
<td></td>
</tr>
<tr>
<td>GROMACS</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>HADDOCK</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>IGV</td>
<td>2.3.34</td>
<td></td>
</tr>
<tr>
<td>MAFFT</td>
<td>7.245</td>
<td></td>
</tr>
<tr>
<td>MGLTools</td>
<td>1.6.7rc1</td>
<td></td>
</tr>
</tbody>
</table>

**Cufflinks**

**Description**

A reference-guided assembler that assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.

**Links**

- [Website](#)
- [Manual](#)
- [Forum help](#)
Choose Style:
- Preset
- Cartoon
- Tube
- Lines
- Line Trace
- Smooth Line Trace
- Trace
Dataverse Team: Gustavo Durand, Leonid Andreev, Stephen Kraffmiller, Phil Durbin, Raman Prasad, Eleni Castro, Danny Brooke


- The Leona M. and Harry B. Helmsley Charitable Trust 2016PG-BRI002 to PS and MC.
- NSF SI2 1448069 (to P.S.)
- NCRR 1S10RR028832 (HMS)
- NIH, NIH Intramural Program, HHMI, EU Infrastructure Grant, The Swiss National Science Foundation, National Science and Engineering Research Council of Canada, McKnight Scholar Award, Welcome Trust, Canadian Institutes of Health, ANRS/Fondation de France, Fundação para a Ciência e a Tecnologia, Portugal, Welch Foundation, Edward Mallinckrodt, Jr. Foundation, CPRIT.