Integrating Globus into a Science Gateway for Cryo-EM

Michael Cianfrocco
Life Sciences Institute
Department of Biological Chemistry
University of Michigan

Globus World 2018
Executive summary

Biologists need atomic-detailed structures of proteins in order to understand healthy and diseased states of organisms at the molecular level.

A **revolution** has occurred in structural biology - these questions can now be answered with **cryo-electron microscopy (cryo-EM)**

**But…**

Cryo-EM is a big data technology, generating **20+TB per person per project**

**So…**

We **incorporated Globus into a web-platform** to allow point-and-click data upload & analysis on HPC resources
COSMIC\(^2\): a platform for determine cryo-EM protein structures

Cryo-EM Open Source Multiplatform Infrastructure for Cloud Computing
Understanding biology requires an understanding of the ‘microcosmos’

Example human cell

100,000 times smaller
Understanding biology requires an understanding of the ‘microcosmos’

**Model** illustrations

Nucleus

Example human cell

Cytoplasm

Plasma membrane

Illustrations: David Goodsell, TSRI
Understanding biology requires an understanding of the ‘microcosmos'

Example subset of proteins (~100)
Understanding biology requires an understanding of the ‘microcosmos'

Diagram of ~100 proteins / molecules

Example human cell

Example subset of proteins (~100)

Amino acid

- Humans have 21,000 genes
- Predicted to form 100,000 different types of proteins

Proteins are linear chains of amino acids strung together (Usually > 300)

Illustrations: David Goodsell, TSRI
Cryo-EM utilizes transmission electron microscopes to take images of proteins.
Cryo-EM: 30 second overview

NOISE AVERAGING

EM IMAGING

Class X

Class Y

Class Z

3D RECONSTRUCTION

Class Average X

Class Average Y

Class Average Z

ALIGNMENT

CLASSIFICATION

20,000 - 100,000 core hours
Cryo-EM relies on movies instead of images of protein samples

1-5 GB / movie
1 movie / minute (24/7)

Bai et al. 2015
Cryo-EM data is low SNR...

Grouped into classes and averaged to create averaged images:

Averaged image

Individual particles in averaged image
...but can solve atomic protein structures

Combine data together in 3D dimensions

Atomic protein structure from ~400,000 images (12 TB)
Impact & growth of structural biology

2017 Nobel Prize in Chemistry

Jacques Dubochet
Joachim Frank
Richard Henderson
Cryo-EM is undergoing a rapid expansion

Instrument collects ~2 TB / day (24/7)
-> Each individual requires collecting ~10 - 20 TB

~60 instruments in the US, growing quickly
Building a science gateway for cryo-EM

**Short term:**
1. Remove command-line interface for cryo-EM job submission
2. Remove decisions regarding HPC job environment
3. Create centralized location for data analysis software

**Long term:**
1. Connect users to cloud storage
2. Become platform for training and implementing advanced algorithm development - e.g. Machine learning / neural networks
3. Integrate educational materials to train next generation of structural biologists
Comet supercomputer at San Diego Supercomputer Center

46,656 CPUs
288 GPUs

XSEDE
Extreme Science and Engineering Discovery Environment
Computing allocations

Current allocation on COSMIC\(^2\): \(\sim 25,000\) GPU-hours

All users: 500 GPU hours (free, no questions asked)

If you run out:

- Apply for supplement from COSMIC\(^2\)
- Receive guidance on how to submit XSEDE computing allocation from COSMIC\(^2\)
Welcome to COSMIC$^2$!

This is a freely available, science gateway for cryo-EM structure determination.

Please login below with your university credentials and then you are ready to go!

Login

Questions about this login process? Please read more here.
Welcome

Current Folder Details

<table>
<thead>
<tr>
<th>Label</th>
<th>Gateways 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Demo project for Gateways 2017</td>
</tr>
</tbody>
</table>

Create New Folder  Create Subfolder  Edit Folder  Delete Folder
Impact & growth of structural biology
Folders
Total Storage: 0 bytes

| Data (2) |
| Tasks (1) |

All Data
Upload Relion directories, particle stacks, and 3D volumes using the Globus data transfer service.
Upload 3D volume and other small files (<200 MB) using your browser.

<table>
<thead>
<tr>
<th>Data</th>
<th>Name</th>
<th>Bytes</th>
<th>Format</th>
<th>Date Created</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>belagatess/particles.star</td>
<td>345460</td>
<td>STAR</td>
<td>10/9/17, 06:11</td>
</tr>
<tr>
<td></td>
<td>Particles/shiny_2sets.star</td>
<td>28378</td>
<td>STAR</td>
<td>10/21/17, 18:22</td>
</tr>
</tbody>
</table>

20 records on each page

Move selected to First test Go
Delete Selected
Tool "RELION_2D_CLASS_COMET" successfully set to current task.

Create new task

Task Summary  Select Data  Select Tool  Set Parameters

You may edit your task using the tabs above. Current CPU Hr Usage: 0 

Description
2D classification - 150 classes

Input
1 Inputs Set

Tool
Relion 2D classification

Input Parameters
14 Parameters Set

Click here to provide job parameters

Click for more info

Save Task  Save and Run Task  Discard Task

Saved tasks can be run later from the task list.
XSEDE tasks are limited to 188 hours. Non-XSEDE tasks are limited to 72 hours.
Create new task

Relion 2D classification: Calculate 2D class averages using Relion (S. H. W. Scheres)

Simple Parameters

Number of classes: 00
Particle diameter (Angstroms): 1
Pixel size of data (Angstroms/pixel): 1
Number of iterations: 25
Output directory name: output_direct
In-plane angular sampling: 5
Pixel search range (pixels): 5
Pixel search range step size, in pixels (offset_step): 1

Advanced Parameters

Save Parameters  Reset  Cancel
Create new task

Task Summary

Select Data  Select Tool  Set Parameters

Description

2D classification - 150 classes

Input

1 inputs set

Tool

Relion 2D classification  Click for more info

Input Parameters

15 Parameters set

Run job!

Save Task  Save and Run Task  Discard Task

Saved tasks can be run later from the task list.
XSEDE tasks are limited to 168 hours. Non-XSEDE tasks are limited to 72 hours.
Task Details

Task: 2D classification - 150 classes
Owner: rmichael
Group: rmichael
Date Created: 10/22/17, 16:45
Tool: Felion 2D classification
Input: View (1)
Parameters: View (14)
Output: View (2)
Intermediate Results: None
Status: COMPLETED

Task Messages

Particles/shny_2sets.star
Lessons learned

• Globus integration was straightforward
• User account management & authentication
• Very easy data movement

Next steps:
Linking up to cloud storage
Moving data between NSF HPC resources using Globus
For more information

Website:
cosmic-cryoem.org

Github repo:
https://github.com/cianfrocco-lab/COSMIC-CryoEM-Gateway

PEARC17 paper:
https://goo.gl/gnQ79a