

Demo: COSMIC² – A science gateway for cryo-electron microscopy

Michael A. Cianfrocco,* Life Sciences Institute & Dept. of Biological Chemistry - University of Michigan; Mona Wong, San Diego Supercomputer Center – UCSD; and Choonhan Youn, San Diego Supercomputer Center – UCSD.

*Corresponding author address: Life Sciences Institute & Dept. of Biological Chemistry, University of Michigan, Ann Arbor, MI, 48109, USA; email: mcianfro@umich.edu

Abstract: Recent advances in cryo-electron microscopy (cryo-EM) have led to the wide-spread adoption of the technique worldwide to allow the determination of atomic protein structures. In order to use this powerful technique, however, scientists must use high performance computing resources in order to calculate protein structures from terabytes of image data. We have built the COSMIC² science gateway to provide researchers with access to the computing resources available on the Comet supercomputer located at the San Diego Supercomputer Center. In order to handle terabyte-sized data uploads, we have integrated the file transfer service Globus into the gateway. During this demonstration, we will highlight the user interface and data ingestion process used by the gateway, while also discussing considerations for user interface and design.

1. Introduction

Cryo-EM is one of the fastest-growing techniques used in structural biology. Due to recent advances in electron microscope optics and detectors, cryo-EM has been used to determine protein structures for samples that were previously unattainable with traditional structural biology techniques [1]. These structures are calculated from 10+ terabytes of data, where individual protein molecules are imaged in the electron microscope and then extracted for image analysis. The low signal-to-noise of these individual protein images requires the analysis of >500,000 images, which necessitates the use of high-performance computing (HPC) resources. Then, after utilizing >10,000 computing hours, users are able to determine atomic biological structures.

*Presented at Gateways 2017, University of Michigan, Ann Arbor, MI; October 23-25, 2017.
<https://gateways2017.figshare.com/>.*

1.1 COSMIC² science gateway

Given the large computing requirement for cryo-EM, access to HPC resources has become a limiting factor in the spread of the technique. In the current system, each cryo-EM researcher must 1) determine which HPC resources are available; 2) install relevant cryo-EM software; and 3) optimize HPC job submission and management on the HPC resource.

Given that this system requires that users have advanced knowledge of Linux and HPC job management tools, we built the COSMIC² science gateway (COSMIC²: Cryo-EM Open Source Multiplatform Infrastructure for Cloud Computing) [2, 3] to remove these barriers. This gateway has incorporated Globus services for the movement of terabyte-sized datasets into the Comet supercomputer at the San Diego Supercomputer Center through a web interface. Users are then able to use standardize software packages and commands to submit and monitor jobs on Comet through a web interface. Finally, all output data is downloaded back to the users local machine using Globus services.

2. Demonstration of COSMIC² functionalities

Below we will describe and show screenshots that will be included during the demonstration of the gateway. Please note that we have included a small selection of images due to space for this submission.

2.1 User authentication

We have integrated Globus services into the COSMIC² gateway, which provided us with the ability to use Globus for user authentication [2]. This system allows users to authenticate with the

COSMIC² gateway using their university credentials, Google, or personal Globus account, which removes the need for administration and security of handling usernames and passwords in the gateway.

2.2 Data upload with Globus

My Endpoints

Endpoint	Path	List Files
Laptop in Leschziner lab (Source)	/~/Desktop/	<button>List</button>
Leonora	/~/	<button>List</button>

XSEDE Endpoint

Endpoint	List Files
COSMIC2 Science Gateway storage on XSEDE Comet	<button>List</button>

Globus Transfer Switch Source and Destination

Destination Endpoint: COSMIC2 Science Gateway storage on XSEDE Comet
Source Endpoint: Laptop in Leschziner lab

File / Folder	Size	Select
AL_Lab	3808	<input type="checkbox"/>
Emplar_10061	102	<input type="checkbox"/>
RP Lab	4114	<input type="checkbox"/>
cryoem-cloud-tools	646	<input type="checkbox"/>

Fig. 1: Data transfer using Globus in the COSMIC² gateway. Endpoints available to the user, where ‘My Endpoints’ indicates Globus endpoints managed by the user and ‘XSEDE Endpoint’ is the Comet supercomputer endpoint at the San Diego Supercomputer Center. When a user selects ‘List’ for a location on a personal endpoint, the gateway will provide a list of the directory contents under the ‘Globus Transfer’ section of the webpage.

In order to allow large dataset uploads (> 1 terabyte), Globus services provides a third-party mechanism for moving data. Briefly, when the user uploads data to the COSMIC² gateway, Globus coordinates the file transfers from the local Globus endpoint on the local machine of the user with the endpoint located on the Comet supercomputer at the San Diego Supercomputer Center [2]. Globus will then monitor job status, restarting failed attempts, and then notify the user when the upload completes.

Once a user authenticates with the COSMIC² gateway using Globus, they are able to select their local endpoint and the files for upload. On the data

transfer page of the gateway, the user will see local endpoints and the name of the endpoint on Comet (Figure 1). By default, the gateway does not allow users to view files on the gateway using Globus; instead, users are only able to see data on their local Globus endpoints (Figure 1). Once the user selects their data, they select ‘Transfer’ and then Globus will move data into the COSMIC² gateway from their personal endpoint.

2.3 Job submission

Once the data are ingested into the gateway, users are able to perform a series of different analysis routines for cryo-EM structure determination. We are initially offering three separate analysis routines through the gateway to use the commonly used program RELION [4] for image analysis in two and three-dimensions. We can easily expand the number of available software programs depending on user need.

For a given ‘task’, the user will provide input parameters for the job submission. Using our gateway, we are able to populate almost all fields of the input parameter list with default values. Furthermore, we provide detailed help information for each input parameter, guiding the user through the list of input parameters (Figure 2).

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

Simple Parameters

Number of classes * 100

Particle diameter (Angstroms) * -1

Pixel size of data (Angstroms/pixel) * -1

Number of iterations * 25

Output directory name * output_dir

In-plane angular sampling * 5

Pixel search range (pixels) * 5

Pixel search range step size, in pixels (offset_step) * 1

Fig. 2: Example input parameter page for a RELION job.

Once a job is submitted, the user can monitor job status through the gateway by monitoring the standard output file for the RELION job running on the Comet supercomputer.

2.4 Data retrieval with Globus

After the user finishes cryo-EM data analysis on the COSMIC² gateway, the user will be able to retrieve data using the same Globus transfer page that was used for data upload (Figure 1). For

downloading the output data, the user will ‘switch’ source and destination, and will be able to download output files from the gateway.

3. Conclusion & future directions

The COSMIC² gateway is the first science gateway capable of handling terabyte-sized datasets by including the Globus file transferring service [2]. The demonstration shown here highlights the minimal set of features we chose to include for the launch of the gateway. Further work is necessary to streamline the pipeline for image analysis, such as displaying input and output image files within the webpage.

In the coming years, we will build upon these core features to include a full processing pipeline that will handle all pre-processing steps that were not included in this initial design. We will continue to leverage Globus as a system for user authentication and file transfers, continuing to offer users with HPC resources that run on the Comet supercomputer.

4. Acknowledgments

This work has been supported through grants from XSEDE-ECSS and Science Gateways Community Institute.

5. References

- [1] Eva Nogales. The development of cryo-EM into a mainstream structural biology technique. *Nature Methods* 13, 1 (2016), 24–27.
- [2] M.A. Cianfrocco, M. Wong, C. Youn, R. Wagner, A.E. Leschziner. COSMIC²: A Science Gateway for Cryo-Electron Microscopy Structure Determination. *PEARC 2017, New Orleans, LA. Accepted.*
- [3] www.cosmic-cryoem.org
- [4] Sjors H.W. Scheres. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* (2012) 180(3): 519-530.