6-18-2018

Collaborative Research: Computational Modeling of How Living Cells Utilize Liquid-Liquid Phase Separation to Organize Chemical Compartments

Jia Zhao
Utah State University, jia.zhao@usu.edu

Data Management Plan

Policy for Data Storage

Extensive numerical data files will be produced in this project. For example, the preliminary in vitro simulations of liquid-liquid phase separation prior to coupling of viscoelasticity and hydrodynamics in cytoplasm generates ~1 GB data files per run, while the movies generated from the stored data are ~10 MB per movie. The incorporation of viscoelasticity, hydrodynamics, and confined cellular domain geometry are anticipated to generate 2 orders of magnitude more data and storage requirements. Algorithm and software development for these models are performed at USC (Wang group) and USU (Zhao group) using their computational facilities and likewise stored locally and shared as indicated below. Production runs of the various levels of in vitro and in vivo models will be performed at USC, USU and UNC (Forest, Gasior, Newby group). Additionally, experimental movie files of patterned cellular compartments (on the order of 1 TB per experiment) of the Gladfelter lab have been and will continue to be converted by Newby to particle tracking data, followed by data analytics and visualization of all tracked species time series (see Figure 1 in the proposal). All of the neural net conversion data, data analytics, and post-processed figures and movies are stored in Cloud Storage and on hard drives in the Gladfelter and Forest storage space; see below. The Neural network tracker is designed from the start to use distributed computing models (Apache Beam, Spark, Hadoop) to work with large video files. Over the course of the entire grant period, all data files essential for publication and dissemination, including the post-processed analytics, graphics and movies will be stored using services (Dropbox, Google Cloud Storage, Github) with secure backups, and will be available to all team members at each institution. All co-PIs and their respective undergraduate and graduate students will engage in secure data storage practices. Access to all resources will be managed through Google Cloud, Dropbox (with unlimited storage restriction business plan), and Github. Google Cloud access can be given to anyone with access to the Google Chrome browser and a Gmail account, which is protected by two-factor authentication. Likewise Github provides code repositories that are secured by two-factor authentication.

All numerical data, computer programs and other electronic files will be backed up on multiple machines and in cloud storage. Each PI will possess one copy and held for three years after the end of the award period, as required by NSF guidelines. Experimental data, including large TB data sets, are stored in Google Cloud Storage, which enables full lifetime management of all data. Image data begins in hot storage during pipeline processing and for initial stages of analysis (providing a time window to verify results, if required). After the initial stage post analysis, image data will automatically be loaded into long-term cold storage. Data conversion from the Newby et al. neural net pipeline analysis is maintained by Google BigQuery (Google's petabyte scale data warehouse cloud service) in hot storage indefinitely. All data associated with an image set will be linked and searchable using BigQuery. The pipeline is written in Python, and position-time series data is propagated through the pipeline using Pandas DataFrame objects, which can be serialized into a variety of formats and automatically interfaces via SQL with Google BigQuery. The data products of a given stage of the pipeline are generally orders of magnitude smaller than the preceding stages. These generally include CSV files containing all particle position time series, as well as other quantities such as PSF radius and localization uncertainty.

Policy for Data Sharing

Data will be published as soon as it is reasonable to do so (i.e., sufficient data generated to warrant a suitable paper). Supplemental data for each published paper will be made available as part of the

publication, and/or on the Publications page in the website of the respective PI under supplemental materials for each publication listing. Furthermore, all PIs and Senior Personnel will share all project results with other researchers in the community via each PI's research website and academic Github accounts for code hosting. Numerical simulation results and open access publications or links to journals will be uploaded regularly. The PIs will also post movies to YouTube and other social media to heighten awareness to the research. Research from the project will be disseminated in leading journals in mathematical biology, applied and computational mathematics, biology, biological physics, and rheology of biological systems, as well as through presentations of the entire research team, invited and submitted minisymposia at meetings of these communities. The data repositories will be advertised to draw attention to products of the project. The software developed over the entire funding period will be made available to the mathematical, biological, and computational science and engineering communities via Github accounts.

All co-PIs, Senior Investigator, and their groups will cooperate to ensure the external community is able to gain access to the data and software and to learn how to use these tools. We anticipate these tools to have broad national and international applicability to other cellular systems, and thereby the source of future collaborations.

We will conform to NSF policies on the dissemination and sharing of the research data and materials created or gathered in the course of the proposed project. The data acquired and preserved during the project will be further governed by the PIs' institute policies pertaining to intellectual property, record retention, and data management.

Training and Best Practices

At the start of the funding period, the PIs and all associated personnel will meet to outline data collection and storage protocols, including file nomenclature, documentation, and dissemination policies. A rigorous, documented folder system will be used to insure that results are stored in a manner easily understood by all researchers. Throughout the funding cycle, an annual meeting will be attended by all project participants to insure best practices are being maintained.

Other Resource Sharing

Resources will be made public through a variety of means, including publication in scientific journals, written or oral presentations at scientific conferences, and personal conversations with colleagues. We will make every reasonable effort to share materials and resources developed with NSF funding or any other source of funding. All academic researchers will have access to the necessary methods and protocols as described above. Wherever possible, we will transfer resources and materials to other academic researchers using simple letter agreements that conform to the intent of the Uniform Biological Material Transfer Agreement (UBMTA).