Data Management Plan

Data and Materials Produced

We will collect several types of data, including chemical and hydrographic data, DNA and RNA sequence data, cell physiology data, processing code data, and results from numerical model runs. Chemical, hydrographic, and DNA sequence data will be obtained from the Puget Sound in the North Pacific Ocean. RNA sequence data and information about cell physiology will be obtained from bacterial cultures grown under a broad range of growth conditions. These data will inform models to better predict latent marine phage activation. We will also build density-dependent and physiology-dependent models. From these models we will generate predictions with outputs including viral and host densities over time, as well as cell lysis rates, cell growth rates, and viral production rates. All process data and annotated data types will be stored in notebooks and on local servers, will be backed up on project specific laboratory hard drives, and will be made publicly available through the Biological and Chemical Oceanography Data Management Office website (BCO-DMO), NCBI, and Github. We will also construct a searchable and relational database in collaboration with Big Rose LLC, a minority and woman owned small business that has more than a decade of experience providing custom laboratory information management services to academic research laboratories throughout the United States, Europe, and Australia.

Standards, Formats and Metadata

We will record all metadata by taking careful notes in laboratory notebooks that are checked monthly by PI R. Morris and co-PI Knowles to ensure that they refer to associated data files (e.g. .txt, .csv, .tiff, .fasta) and that they describe all experimental conditions, units, abbreviations, and missing values. Physical notebooks will also be photographed, and digital images will be archived on a weekly basis. Notes and data recorded in laboratory notebooks will be transferred to electronic documents in standard formats and the PIs will ensure that they contain all associated information with clear identifiers that link experiments to data files. All data files generated at the University of Washington will be backed up on laboratory computers, a local server in the Center for Environmental Genomics, and on a designated project-specific backup hard drive that will be maintained by PI Morris. All models and associated data generated at the University of California in Los Angeles will be organized and archived weekly in the Knowles laboratory information management system (LIMS), along with README files of that week's progress and implementation (a digital companion to the physical notebooks) for easy retrieval by all lab members. Computational workflow/processing code will be stored as raw text files, and tabular data as comma separated tables (.csv). Model predictions will be stored as csv files whenever models are run, and uploaded along with accompanying Matplotlib graphs. All file formats will be compatible between the Morris and Knowles laboratories with open-source standards, such as mySQL, the most widely used open-source relational database management system. To the extent possible, proprietary file formats will not be used to store data. This will ensure that data products can be publicly disseminated in accessible formats.

Roles and Responsibilities

PI R. Morris is ultimately responsible for all data management associated with this research project, during and after data collection. Individuals (PI, graduate, and undergraduate researchers) will generate the data and are responsible for rapidly and regularly implementing quality control (PI cross checking) and archiving measures. All data, including notebooks, digital copies, backups, and associated records will be accessible to both PIs through the CEG sever at the University of Washington and through LIMS in the Knowles laboroatory at the University of California in Los Angeles. PIs Morris and Knowles will ensure that all chemical, hydrographic, and biological data, model process code, and model predictions are posted to online databases, such as BCO-DMO, NCBI, and Github, along with README documents. To ensure continuity, PI Knowles will subcontract with Big Rose LLC, a minority and woman owned small business that has more than a decade of experience providing custom laboratory information management services to academic research laboratories throughout the United States, Europe, and Australia. Projects that have utilized their expertise are supported by public and private funding agencies including NASA, USDA, NSF, NIH, Moore Foundation and the Pew Charitable Trust.

Dissemination Methods

Data will be disseminated through peer reviewed publications and through public repositories such as BCO-DMO, NCBI, and Github. In most cases, this includes re-distribution and derivation, following the policies of each online database. DNA sequences from seawater and RNA sequences from bacterial cultures will be deposited in the National Center for Biotechnology Information (NCBI) under accession numbers assigned by NCBI. Processed sequence data will also be available through MG-RAST, the iMicrobe, and the associated CyVerse data stores. Links to sequence data will be made available with associated metadata such as chemistry data and hydrographic data via PI Morris' profile on the BCO-DMO website, which uses community metadata standards and is routinely updated to provide links to associated raw data files (.csv, txt, tiff, .fasta). All raw and associated metadata generated in the Knowles laboratory will be transferred to the appropriate public database (e.g., Github, BCO-DMO, Figshare or Dryad).

Policies for Public Access

We are required to make all data publicly available. For public access, we will ensure that this takes place no more than two years after the data are collected. To that end, data will be made publicly available either at the time of the first publication that uses the data or within two years of its collection, whichever is first. Information about the location of published data files and a description of how to obtain the data will be provided through peer reviewed publications and public repositories with README files. We will release our data and analysis code under broadly accessible licenses (e.g. BSD or CC-BY).

Policy for Data Sharing and Data Security

For internal sharing and security of the data, a web application interface (WAI) used by the CEG and LIMS will provide users with a secure and a standards-compliant way to work with the databases. Several layers of security will be implemented to protect the integrity of the database by Big Rose LLC. Transactions between users and the WAI will be encrypted using a secure SSL connection. Authentication services at the server and database layers are used to control access to the WAI. Static VPN and institutional NetID-based authentication services will be used to manage access to the web, file, and database servers. User privileges in the WAI are assigned by the administrator group. In the event of an emergency, a user’s privileges can be rescinded, locking them out of the system. As an additional security measure, an audit log automatically captures when, by whom, and what changes are made to records. Depending on their access privileges, users can view, search, add, edit, delete, and export data.

Archiving, Storage, and Preservation

Data file structure and organization will routinely be checked by PIs R. Morris and Knowles to ensure that the most accurate and up to date files are available in public repositories such as BCO-DMO, NCBI, and Github. This will be done to ensure accurate long-term public archiving of all project products. Until public dissemination, as outlined above, all chemical and biological data, process code, models, and model predictions will be stored in laboratory databases (CEG sever and LIMS archives) and backed up regularly (i.e., daily) on project specific hard drives to avoid loss in the event of a power outage or computer crash. Hard drives will be maintained by the PIs for up to 10 years after the project is completed. In addition, weekly and monthly snapshots of the databases and data files will be mirrored to an off-site location as an additional security measure. Monthly snapshots of the database and data files will be preserved for the duration of the grant and then archived in perpetuity in deep storage.