DATA SHARING PLAN
Roman Stocker

MMI Investigator Award

The data and knowledge that will be generated in this project will be primarily of three types:

1) new microfluidic approaches to marine microbial ecology, consisting of fabrication methods, specific devices, and insights into their operation in the lab or in the field
2) data on the behavior and physiology of marine bacteria, in particular their growth efficiency and nutrient uptake, under different conditions of nutrient availability
3) a mathematical model of bacterial foraging in the ocean, consisting of a framework to integrate information on the spatial structure of the nutrient landscape, the behavior and physiology of the organisms, and (in output) the relative fitness of different foraging strategies.

None of these three data types lends itself to a standard approach for data deposition. However, it is our vested interest to actively share this information, because

1) we hope that microfluidic methods will be broadly adopted in microbial ecology, and in particular in marine microbial ecology; detailed methods papers, sharing of protocols through our wiki page https://wikis.mit.edu/confluence/display/romanstocker/Home, and openness to collaborations with (or shipping of microdevices to) interested users, will be some of the approaches we will use to ensure that this occurs. Furthermore, we have recently added a site on our webpage that describes the tools we are using and will update this to include microfluidic methods developed as part of this project. Selected Illustrator files for microfluidic design will be made available through the wiki (publishing them *in toto* would make the wiki site unwieldy), and both the wiki and the 'tools' webpage will make it clear that all Illustrator files will be available upon request to any investigator.
2) we will generate both behavioral data and physiological data. The behavioral data relate to the spatial and temporal distribution of bacteria around microscale hotspots (e.g., a lysing phytoplankton). The raw data are in the form of (rather large, due to spatiotemporal resolution) movies, the processed data can be spatiotemporal distributions of bacteria around hotspots (e.g. accumulation of bacteria around the phytoplankton, as a function of space and time): samples will be posted on our wiki and additional sets shared upon request (unless we intend to publish them). We will also generate physiological data, primarily in the form of (long) time series of, e.g., cell size and nutrient uptake as a function of different temporal nutrient supply regimes. Again, samples of these data will be posted on our wiki and additional sets shared upon request (unless we intend to publish them).
3) we envisage that our mathematical model, while microscale in nature, will have the possibility to link microscale processes with larger-scale consequences, and see it as a first step for integrating microscale understanding and knowledge in larger scale models, making it highly appealing to share the model, any information acquired in developing it, and its results. The code will be made available through out wiki site (see above for link), along with the types of model outputs that we will generate.

Data-release check-ins will be handled with an eye to make public all information that will either published very soon, or where there is no potential for competition. Data that will not be published will be shared, though this will occur seldom, because if we do not use certain data it is typically because we have found a way to generate better quality data. The group's approach is to include extensive supplementary material in publications, precisely with the purpose of making those data available for everyone.

The timescale for making data available will be highly dependent on progress on each individual topic. On microfluidic approach, we foresee substantial progress on the ‘bacterial growth efficiency’ device within the first year, and this will represent a tool with broad applicability to a number of different research questions where the temporal nature of a chemical signal could influence microbial fitness. This device, subject to appropriate training and resources for collaborations (it is not immediately portable in its current configuration), will thus be broadly available to colleagues for collaborative or independent experiments, likely half-way through year 2. The mathematical model will take approximately 1.5 years to build with a reasonably complete time frame, and after initial publication the framework and the input parameters will be made broadly available. Finally, there is a possibility that some of the microfluidic techniques will be patented: this will be evaluated case by case, based on the breadth of applicability of each device.

DATA SHARING PHILOSOPHY

DATA SHARING PHILOSOPHY

The Gordon and Betty Moore Foundation's goals of scientific advancement, environmental conservation and health care improvement will best be served through a culture of open access to data. It is our philosophy that:

- All data used in or developed in whole or in part by foundation-funded projects (and that can be shared in a manner consistent with applicable laws) will be made widely available and freely shared as soon as possible[1]. If data used in foundation-funded projects are owned by an additional party other than the grantee, we do not require it to be released, but the grantee will use its best efforts to encourage the data owners to make it openly and freely available.
- Data are shared with full and proper attribution to the data provider.
- Data developed in whole or in part by foundation grant funding are the property of the grantee unless otherwise specified. The grantee may protect its property through patent, copyright and/or other intellectual property protection instruments, except that it may not impede the effective access and use of the data by the public.
- The foundation is not responsible for any liabilities associated with errors in the data or misrepresentations or misinterpretations of publicly available data.
- The foundation supports grant funding for costs associated with data sharing and open access publication of scientific findings, where appropriate.
- The foundation and prospective grantees will jointly develop a Data Management and Sharing Plan prior to the finalization of a grant agreement.

The Data Sharing Philosophy applies to all activities that are financially supported in whole or in part by the foundation that include, but are not limited to:

> Data collection and analyses, data, meta-analyses and information derived from pre-existing datasets, and database development

Data sharing includes, but is not limited to, data contained within the following:

> Publications, databases, derived data products, mathematical models and model code, metadata (defined as appropriate documentation describing the data, relevant specifics of their collection and the data format) and statistical and other forms of data reduction and analysis

[1] Examples of when data should be released: For data created for scientific and environmental conservation purposes, public release should occur not more than six months from the "date of collection" (defined as the date when data enters an electronic database), unless otherwise specified in the grant agreement between the grantee and the foundation; for DNA sequence data, "public release" (i.e. submission to an appropriate public database), should occur not more than six months after "completion" of the DNA sequence determination (as defined in the grant agreement between the grantee and the foundation).

DATA SHARING AND MANAGEMENT PLAN

As part of the foundation grant development process, potential grantees are required to develop a Data Management and Sharing Plan with their foundation grant team. In these cases, before submitting a final grant proposal to the foundation for approval, both the potential grantee and the foundation grant team must approve a final version of the plan that is consistent with this Data Sharing Philosophy. Any exceptions to the Data Sharing Philosophy must be clearly articulated in the plan and approved by the grant team. Funds needed for data sharing and management may be requested as part of the proposal. Once finalized, the plan will be referenced in the grant agreement for the approved grant.

The plan should address the following three topics and any other topics identified by the foundation and/or grantee:

1. **Data description**. Questions to consider as appropriate:

   - What data will be collected during this project?
   - How many different data formats are anticipated? Please list formats.
   - When will the data be collected, when will they be entered into electronic databases and what databases will harbor the data?
   - Does this project involve organization or analysis of pre-existing data, and what are the data sharing arrangements for these data?
   - What are the anticipated data products (e.g., databases, analyses, tools)?
   - What kinds of metadata will be associated with the data?
   - Who is the owner of the data?

2. **Data management**. Questions to consider as appropriate:

   - Where (physically) will the data be stored?
   - What type of data access or data distribution mechanism and software will be used?
   - Will the location or software for initial data entry differ from the data archive?
   - How will metadata be stored, and what provisions will be made to enable metadata searching capability?
   - Who will be responsible for entering and maintaining data archives, and over what period of time will archives be maintained?
   - What data quality controls and assurances will be provided?
   - Who will contribute to the database?
   - Will proprietary data be used? If so, describe the permissions obtained to use the data.

3. **Data Sharing**. Questions to consider as appropriate:

   - Who are the potential data users?
   - What is the appropriate timing for release of data to the public or relevant users, and why?
   - When will archived data be openly available to other users?
   - If data from non-foundation-supported or previous projects are integral to the successful completion of the Grant Purposes, will the non-foundation-supported and/or pre-existing data also be made freely available?
   - How will other users (i.e., beyond the grantee and the foundation) access data and metadata?
   - Are the publicly available data in raw form? If not, what treatments have been

applied to the data prior to their being released to the public?
- How long beyond the grant term will the data be maintained and by whom?
- Does the proposed grant include provisions for future hardware upgrades in the event that data is to be stored and maintained well beyond the project period of the grant?
- If data analysis tools are to be created as a consequence of the grant, will a tutorial be available for training of future users of the data, and if so, how can it be accessed?
- Will a data sharing agreement be required between outside vendors? If so, a brief description of the agreement needs to be provided in the grant proposal.
- Is a Creative Commons type-license appropriate for sharing the data? Why or why not?
- How will appropriate attribution to the data provider be provided?
- Do you anticipate publishing a “Data Release Paper” for referencing and sharing the data?