

Modeling the Earth's Microbiome A real world deliverable for microbial ecology

Jack A Gilbert^{1,2} and Folker Meyer^{1,3}

¹Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, U.S.A.

²Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A., ³Computation Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.,

Microbiology is awesome, awe inspiring and seemingly unfathomably complex. Globally, microbial cells are estimated to be a billion times more abundant than stars in the known universe. When one considers the complexity of the n-dimensional hyper volume of niche space in which microbial life finds itself, multiplied by the immensity of time (~3.8 billion years), it is hardly surprising. Yet, it is a daunting task to try and understand this complexity in a way that could be useful to humanity. Useful because, cataloguing this vastness, while essential, does not immediately provide useful products. These surveys are fundamental, but they are natural history, a cataloguing of events linked to a limited number of contextual variables, e.g. location and time. A useful product is one that we can use; the form of that interaction, or 'what you want the tool to do' should define how we design these surveys, so that specific questions can lead to specific products that can help to refine the specific questions. This is to paraphrase the scientific method.

It is axiomatic that microbes are important to ecosystem function; **if they weren't, they wouldn't be so ubiquitous**. Indeed, great strides have been made in the last hundred years describing how microbial species then consortia and finally communities interact with their biological, physical and chemical world. The way in which this has been achieved is continuous investigation at both ends of the perspective ladder. At the base of the ladder is intracellular metabolic dynamics; exploration of gene to transcript expression regulation, the folding of macromolecules, protein function and biochemistry at the level of individual cells, usually from individual taxa, has enabled us to create a window into metabolism. This window means that we can now visualize the metabolic pathways that allow a cell to interact with the environment to create more cells, thus increasing biomass of a taxon. At the other end of the ladder is the so-called '30,000 feet perspective'; this perspective is used to explore the sum of an ecosystem's taxonomic and functional capability.

The Earth Microbiome Project (www.earthmicrobiome.org) is a collaborative initiative to create multiple comparable data sets of taxonomic and functional sequence data from a broad range of ecosystems. The EMP is currently in its pilot phase, with several projects running in parallel. The main aim of one of these projects is to generate 16S rDNA amplicon and shotgun metagenomic sequence data from 10,000 environmental samples. Obviously this is a significant sized task, yet absolutely within our grasp; the generation of 15 trillion base pairs of sequence data (e.g. using Illumina HiSeq2000 and Pacific Biosciences RS1) from a diverse array of complex microbial ecosystems generates a

informatics challenge that requires a coordinated data management plan. Hence the EMP has been built by direct interaction between bioinformaticians and microbial ecologists; the communication between these groups has been constant, enabling multiple avenues of feedback, and constant revision of the frameworks for data handling and processing through the QIIME database and MGRAST processing capability.

Goals are always moving targets and in science can be considered floating points, defined by available evidence. The primary goal of the EMP is enable comparison of many environments across the planet leveraging comparable data sets. This means, the data must have been generated using a defined, standard protocol in much the same way as samples within individual projects (e.g. the Global Ocean Sampling expedition, TARA Oceans, etc). For nucleic acid analysis this means several things. *Firstly*, a standard DNA extraction protocol; the method used to lyse cells and release the DNA for downstream analysis determines the breadth of the microbial community that is accessed. Hence, comparing two or more samples with two different DNA extraction protocols will likely lead to a statistical result whereby the most significant discriminatory factor between the two samples is the extraction protocol, hence the data is of limited biological and ecologically relevance. *Secondly*, a standard amplicon protocol; it is well known that DNA amplification is biased. PCR has many steps that can introduce bias, including primer selection, the type of taq polymerase, the temperature or timing of amplification cycles, the amount of ionic strength in the mixture, etc. These biases, as with the DNA extraction procedure, will lead to incomparable datasets if different protocols are followed. *Thirdly*, standard sequencing protocols; this is especially important for shotgun metagenomic sequencing, whereby the absence of an amplification step means that sequencing (after DNA extraction) can be the primary source of bias. This is especially important in the current technological climate with a number of relevant sequencing platforms to choose from, each with different idiosyncratic biases. Essentially, by standardizing these elements of bias it is possible to feed comparable sequence data sets from many different ecosystems across time and space into comparative informatics and statistical analysis to enable comparisons of gene and taxa diversity and abundance in different environmental contexts.

A second, but no less important, goal of the EMP is to determine the actual impact the biases associated with DNA extraction, PCR amplification and sequencing technology on interpretation of ecologically relevant patterns and community reconstruction in sequencing data. This is vital, as it is not sensible to assume that prescriptive approaches will be adopted by the global community, or will be future-proof. Technological advances, in recent years specifically sequencing platforms, move at an astonishing rate, hence it is necessary to understand how data generated by updated standard protocols can be back compared to existing data. One alternative, to re-sequence everything, does not address changes in PCR primer sequences to improve taxonomic coverage of metagenetic (amplicon metagenomics) screens, or the need for different DNA extraction approaches which deal with different sample substrates. Currently, most researchers have their own ideal approach for extraction good quality DNA from their ecosystem of interest. These questions are not being ignored.

Among the additional goals (including a global database for exploring environmental samples by their niche space characteristics, e.g. location, temperature, nutrients, moisture, pH, light availability, etc.; a global atlas of protein functions orientated along the same niche partitioning parameters; a catalogue of re-assembled genomes and their taxonomic distribution), there is also a fundamental aim to create predictive mathematical models of ecosystem services. Ecosystem models are usually defined as abstract representations of an ecosystem, and are created to represent scales ranging from the intracellular dynamics to the regional and global scale predictions. Invariably these are networks of interactions between the biological, chemical and physical variables in a system overlaid with algorithms, which describe the relationships, so that an alteration to one variable will generate predictions about the response of each other variable. Models are essential for ecology and humanity for two basic reasons; firstly, they enable us to make predictions about a system *in silico* without the need for inappropriate experimental alterations of whole ecosystems, e.g. heating up or acidifying the ocean. Secondly, they help us to predict the impact of change on the ability of an ecosystem to deliver the vital services upon which we rely.

There are a large number of ecosystem models, spanning a wide range of ecosystems. A large proportion of these exist for marine systems, probably because the fluid dynamics in a body of mixed water provide a homogeneity capable of conforming to one of the fundamental characteristics of many predictive ecosystem/bioclimate models; the assumption that there is no barrier to dispersal of organisms within the predicted spatial or temporal range. Indeed the ubiquity and age of microbial life means that in a dynamic ocean environment there should be very few limitations to absolute dispersal. In the One Ocean Model of Biodiversity, O'Dor and colleagues, demonstrate that a microorganism could reach any other location in a global ocean in approximately 10,000 years, with a dispersal rate of approximately 1 mile per year. Given oceanic currents, which are themselves agents of and potential barriers to dispersal, it is likely that this rate is both underestimated and overestimated for particular groups on a global scale. Yet this perceived dispersal could lead to the development of a predictive model of global microbial community composition and community structure. In a recent review by Follows and Dutkiewicz they successfully demonstrate this by modeling the distribution of specific Eukaryotic and bacterial taxa throughout the global ocean over an ocean-current dynamics model. In this model the predicted distribution of each taxon is defined by its enforced distribution through global currents and its observed capability to survive and thrive within different environmental constraints, e.g. temperature. Modeling the spatial characteristics of microbial community structure in terrestrial ecosystems is conversely extremely difficult, primarily because of the perceived heterogeneity in, for example, soil-derived communities. The static nature of terrestrial systems results in patchiness of predictive capability, so that predictions for a pasture will be very different to a forest even if separated by a few meters. However, this perceived heterogeneity might not be as 'patchy' as once thought. Just as marine systems have the potential to allow universal microbial distribution every 10,000 years, the potential for terrestrial distribution must have a temporal limit. Soil is not static; animals, hydrology and wind, not to mention longer-term geological erosion, continental drift and hydrothermal activity, enable local, regional, continental and global distribution. While determining the

time scale for this distribution, i.e. the time taken for a microbe to travel 10,000 miles, may be extremely difficult, and defined by more potential barriers than in marine systems, it is not impossible. By creating a comparable global inventory of microbial taxa from thousands of disparate ecosystems, the EMP can hope to start to elucidate the level of overlap in taxonomic composition between different terrestrial systems across different spatial scales. A hypothesis to test using this dataset is that there exists a universal microbial community for most moderate ecosystems. To test this it will be necessary to sequence extremely deep in taxonomic space in many hundreds of different soil systems across time and space. In a comparable study in the marine systems, Caporaso et al generated 10,000 16s rDNA sequencing reads from 72 consecutive time points in the English Channel surface water research station, L4 (720,000 reads). They then picked one of these samples and sequence 10 million 16S rDNA reads. Strikingly they found 99.96% of all the taxa from the initial survey in the deep-sequenced data set from one time point, yet this species compliment only comprised 5% of the total taxonomic diversity identified in the 10 million reads. This suggested that despite, or perhaps because of, the dynamic flow of the English Channel, the same community was always present, which validated the assumption that there was no essential barrier to dispersal of this community. The EMP is already generating more studies of this kind to aid in validation of these assumptions regarding distribution. While taxonomic evolution of regionally isolated populations may lead to a perceived barrier to dispersal of absolute diversity, the concept of turnover of the system through geological time could render such differentiations redundant. And while, short-time scale predictions will have to consider such variation, it remains to be seen how this will impact, if at all, predictions of functional capability, which after all are the absolute goal of ecosystem modeling for the benefit of humanity.

Validation is essential for any model, yet it is often extremely difficult to obtain. However, models can be both generative of and validated by sampling strategies to fill this gap. Essentially, a model that predicts how an ecosystem responds to change, needs either a fundamental understanding of the biochemical mechanisms by which all the species in a system respond to changing biological, chemical and physical variables, or a set of observed correlations of changes in biological variables as a result of physical or chemical change. Both should be able to (a) predict changes in the system from conditions not used to train the model, and (b) feedback characterization of the environment through prediction of the impact of biological changes on physical and chemical variables. The method used depends on the type of data available; for example the first strategy requires a comprehensive understanding of the reaction limits and resulting interactions within the variables of a particular biological unit to changing parameters, such as that used by Follows and Dutkiewicz. The second strategy requires a comprehensive survey of the community through time and potentially space to define the range of the community and correlations of changes in structure (relative abundance of units in the system) to observed environmental parameters. Both methods rely on in-situ and/or experimental observations of how the biological unit or community structure will respond to change. There are however, very few long-term research stations which can be used to define how a community in a given location responds to the full suite of environmental variables that it is exposed to, for example, in a full seasonal cycle.

However, models that can utilize these limited data resources and extrapolate predictions through time and geographic space have immense power to inform future sampling strategies. In the short term they can identify anomalies in these predictions that can be explored using small-scale sampling trips to support or reject the prediction, and hence refine the model. In the long term, these validated observations of anomalous structure can be used to identify locations for future long-term ecosystem observatories.

Currently, most models of any ecosystem deal with microbes as a black box, which has known inputs and outputs of carbon, energy, nitrogen, etc. Even the predictive bioclimatic ecosystem models are used to predict the presence or absence of microbial taxa. What is lacking is a model that uses environmental parameters to predict microbial taxonomic community structure, and then uses these predictions to define the metabolic capability of that community. This provides excellent opportunities for defined feedback to the physical and chemical parameters in that ecosystem. Larsen and colleagues recently demonstrated two separate techniques, which when combined could provide a unique capability. Firstly, Predictive Relative Metabolic Turnover (PRMT) uses the relative abundances of enzyme activities annotated from comparative metagenomic studies to calculate the relative consumption or production of over 900 metabolites that could be generated by a marine microbial community. They validate this technique by comparing the predicted turnover of carbon and phosphorus to the *in situ* measurements of the turnover of carbon and phosphorus. This provides a cyclical approach whereby the environmental conditions that could define a community are predicted by the functional capability of that community. Secondly, Microbial Assemblage Prediction (MAP) uses Bayesian network construction to define relationships between physical, chemical and biological units as a direct acyclical graph, and then overlays an artificial neural network of non-linear mathematical descriptions of these relationships to enable predictions of the relative abundance of given taxonomic units from environmental parameters. Linking MAP to PRMT enables the extrapolative prediction of the relative consumption or production of metabolites as a function of the community structure predicted from environmental parameters. This enables a truly cyclical feedback loop from environment to taxon abundance to metabolite turnover to environmental parameters again.

A fundamental necessity for the future of microbial ecology and ecological modeling is the appropriate design of environmental sampling, and the coordination of sampling effort to minimize redundancy and improve statistical analytical comparability. Modeling has a role to play in this feedback loop, in that good experimental design can lead to informative models, which can be used to direct future experimental design, and identify appropriate geographic and temporal placement of sampling strategies. This is a call to the community to consider the full gambit of research when designing an experiment. It is no longer acceptable to only define how different a microbial community is between sample A and B and C, such natural history catalogues while useful have limited value unless a defined data management plan is available. To make them useful it will be necessary to explore collaborative studies to enable wider comparison, and to design experiments with ecological models in mind that enable predictive capability so as to refine future sampling effort.

Acknowledgements

This work was supported by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357.

Suggested Reading

Little, A. E., Robinson, C. J., Peterson, S. B., Raffa, K. F. & Handelsman, J. Rules of engagement: interspecies interactions that regulate microbial communities. *Annu Rev Microbiol* **62**, 375-401, doi:10.1146/annurev.micro.030608.101423 (2008).

Larsen PE *et al.* Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microbial Informatics and Experimentation* **1:4** (2011).

Follows, M. J. & Dutkiewicz, S. Modeling diverse communities of marine microbes. *Ann Rev Mar Sci* **3**, 427-451 (2011).

Caporaso JG, F. D., Paszkiewicz K, Knight R, Gilbert JA. Evidence for a persistent microbial community in the Western English Channel. *ISMEJ* (2011).

Jutla, A. S., Akanda, A. S., Griffiths, J. K., Colwell, R. & Islam, S. Warming Oceans, Phytoplankton, and River Discharge: Implications for Cholera Outbreaks. *Am J Trop Med Hyg* **85**, 303-308, doi:10.4269/ajtmh.2011.11-0181 (2011).

Jeschke, J. M. & Strayer, D. L. Usefulness of bioclimatic models for studying climate change and invasive species. *Ann N Y Acad Sci* **1134**, 1-24, doi:10.1196/annals.1439.002 (2008).

Gilbert JA, B. M., Field D, Fierer N, Fuhrman JA, Hu B, Jansson J, Knight R, Kowalchuk GA, Kyrpides NC, Meyer F, Stevens R. The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China. *Stand Genomic Sci* (in submission).

Gilbert JA, M. F., Antonoploulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Feng W, Huson D, Jansson J, Knight R, Knight J, Kolker E, Kostantindis K, Kostka J, Kyrpides, N, Mackelprang R, McHardy A, Quince C, Raes J, Sczyrba A, Shade A, Stevens R. Meeting Report. The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project. *Stand Genomic Sci* **3** (2010).

Gilbert JA, M. F., Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, Glockner F-O, Klenk H-P, Wommack KE, Glass E, Docherty K, Gallery R, Stevens R, Knight R. . The Earth Microbiome Project: Meeting report of the “1st EMP meeting on sample selection and acquisition”. *Stand Genomic Sci* **3** (2010)

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.