Multi-Scale Biological and Environmental Analysis: 21st Century Multi-Scale Environmental Challenges require a New Paradigm for Knowledge Discovery

Susan Hubbard, LBNL

The 20th Century was a period in which science was focused on the characterization of individual components of Earth systems. It is now clear that science in the 21st Century must consider the complexity of the natural environment as an integrated system, which necessitates quantification of biological, geochemical, hydrological and atmospheric process interactions that occur across great time and space scales. However, knowledge-based computational frameworks and tools that allow exploration of a continuum of interactions occurring across the molecular to the regional and global scales are nonexistent. New computational and collaboration paradigms are needed to facilitate the transition of 21st Century scientific research from one associated with distributed and disparate datasets, field observations, specific process knowledge, individual models and single investigators to a predictive understanding of couplings and feedbacks between natural system processes across scales. This transition from studying 'parts' to quantifying 'integrated system behavior' often requires multi-disciplinary scientific team expertise, whose members are typically geographically distributed.

As a prime example, microbial and plant communities and their interactions in natural environments regulate the fluxes of most life-critical elements, control the production of food and biofuel feedstock, control water quality, and regulate the flux of major greenhouse gases to the atmosphere. However, understanding how these processes influence geochemical cycles that are critical to both bioenergy and climate is complicated by our inability to predict how microbial and plant communities function in heterogeneous geochemical, physical and hydrologic environments that often vary over great spatial and temporal scales. Although biological processes occurring at the molecular and cellular scale can impact landscape to global scale processes, the lack of computational frameworks and tools limits our ability to develop a predictive understanding of multi-scale environmental phenomena. Partially due to this limitation, biogeochemical flows in terrestrial ecosystem models historically have been represented as the outcome of environmental influences on "black boxes" that represent large taxonomic groups—most commonly plants and microbes.

However, a more detailed understanding of microbial function in natural communities is evolving through pairing microbial community sequencing with larger environmental system observations and experiments (e.g., Wrighton et al., 2012; Castelle et al., 2013). An example includes the BER-supported LBNL SFA field studies conducted in a floodplain near Rifle, CO, which is exploring how the microbial metagenome influences biogeochemical functioning of the larger system relevant to contaminant transport and carbon cycling and how the metabolic activity is expected to change with climate and land use. Addressing such cell-to-watershed questions requires an ability to curate, integrate, simulate and analyze many different types of datasets, including those associated with molecules, microbes, microbial communities, aqueous geochemistry, groundwater fluxes, sediment physical and mineralogical properties, vadose zone moisture and fluxes, soil properties, organic matter, surface water distribution and fluxes, and vegetation.

Pairing plant functional information with other environmental and climate data in models is also challenging yet required to predict 21st century ecosystem questions, such as effects of water stress on primary production and the ability of a terrestrial ecosystem to serve as a carbon sink. BER's Next Generation Ecosystem Experiment in the Arctic is performing experiments to quantify the response of microbial and plant communities to climate-induced permafrost degradation and associated changes in hydrological conditions, which in turn is expected to alter terrestrial carbon storage and the ecosystem feedback to climate. Addressing such plant-through-climate questions requires an ability to handle many different types of data, including soil biogeochemistry, land surface deformation and erosion, topography, vegetation, surface water distribution and fluxes, surface energy balance, and meteorological parameters.

A specific bottleneck for advancing the BER studies described above is the lack of tools to facilitate team-based knowledge discovery and prediction using heterogeneous, multi-scale and distributed environmental datasets. To attain new knowledge and 'whole system' understanding using a variety of streaming data, a significant investment in software, hardware, and technology development, closely linked with process science and natural system expertise, is required to advance new ontologies, curation tools, visulization tools, data mining and analysis tools, multi-scale models, and workflows. Developing advanced environmental knowledgebases and associated tools is a prerequisite for developing 21st Century sustainable solutions to environmental and energy problems.

References

Castelle, C.J., L. A. Hug, K. C. Wrighton, B. C. Thomas, K. H. Williams, D. Wu, S. G. Tringe, S. W. Singer, J. A. Eisen, J. F. Banfield, Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Communications*, in press, 2013

Wrighton, K.C., B.C. Thomas, I. Sharon, C.S. Miller, C.J. Castelle, N.C. VerBerkmoes, M.J. Wilkins, R.L. Hettich, M.S. Lipton, K.W. Williams, P.E. Long and J.E. Banfield, Fermentation, hydrogen and sulfur metabolism in multiple uncultivated bacterial phyla, Science 337(6102), p. 1661-1665, 2012.