

## Appendix 4

# Opportunities and Requirements for Research in Carbon Cycling and Environmental Remediation

Underlying problems in understanding the global carbon cycle and processes important to environmental remediation are similar in that both topics involve the mechanistic understanding of the fate and transport of species as they flow through ecosystems. Both topics are challenged by the classic scaling problem—connecting spatial and temporal scales of molecular processes to the macro-scales of ecosystems and beyond.

## Carbon Cycle

A pressing national need is for comprehensive understanding of the planetary carbon cycle across terrestrial and ocean environments to provide scientific underpinnings for more robust modeling of climate change and to define carbon biosequestration options over the coming decades. To reach a consensus on projections for future climate, the scientific community needs a better understanding of fundamental mechanisms that control carbon sources and sinks. In the past two decades, much progress has been made in understanding historical trends in atmospheric CO<sub>2</sub>, and biogeochemical modeling of carbon in oceans and terrestrial systems continues to advance. The current state of carbon cycle science, however, still cannot quantitatively address several key questions:

- Where are major carbon sinks in ocean and terrestrial ecosystems?
- Which mechanisms control behaviors of carbon sinks and sources?
- How long will biologically sequestered carbon remain stored?
- Will current carbon sinks persist or become carbon sources in a warmer, higher-CO<sub>2</sub> world?
- How do human activities impact carbon storage and release in ecosystems?

Biological processes are fundamental to planet-wide carbon cycling. Thus a mechanistic, systems-level understanding of complex biogeochemical processes at multiple scales will be essential for predicting climate-ecosystem feedbacks. Key issues revolve around photosynthetic productivity; partitioning of photosynthate into energy or biomass pathways; respiration mechanisms; paths to recalcitrant compounds and structures with long environmental residence times; and the effect of environmental variables, nutrients, and water in the context of climate change. Research details from DOE's Carbon Cycling and Biosequestration Workshop can be found in the report, *Carbon Cycling and Biosequestration: Integrating Biology and Climate Through Systems Science* (U.S. DOE 2008, <http://genomicsgtl.energy.gov/carboncycle/>).

## Environmental Remediation

DOE is faced with a daunting legacy of the Cold War: environmental management of the nuclear weapons complex consisting of more than 5000 surplus facilities and associated land located at 144 sites in 31 states and a U.S. territory. More than 380,000 m<sup>3</sup>



of high-level radioactive waste are estimated to be stored at several sites. Furthermore, the total volume of soil, groundwater, and sediment present in highly diverse environments and contaminated with complex mixtures of radionuclides, metals, and organic contaminants may exceed 1800 million m<sup>3</sup>. Waste treatment, environmental remediation, and long-term stewardship of these sites are estimated to require hundreds of billions of dollars and major breakthroughs in science and technology.

A key to making informed decisions regarding DOE site remediation and stewardship is to understand the interdependent biological, chemical, and physical processes that interact at multiple scales to control contaminant form and transport in the environment. New knowledge and tools are evolving rapidly from DOE Office of Science programs and initiatives to address important gaps in our understanding of the molecular sciences and the complex machinery of life and to develop the computational power and infrastructure for simulating and scaling complex interdependent phenomena. Taking advantage of these emerging resources will be critical to providing the scientific foundation for environmental remediation and stewardship of DOE sites.

## Similar Requirements for Common Elements of Domains

Research problems in carbon cycling and environmental remediation have many common elements that render their knowledgebase requirements very similar, if not identical. Ultimately, a major goal is to understand biogeochemical processes at a level required for predictive modeling of both carbon cycling and contaminant fate and transport at the field scale. Examples follow.

- Scaling of processes occurring at the molecular and microscopic scales to the field scale is critical for understanding and predicting biogeochemical processes involving carbon cycling and contaminant transformations.
- Both domains deal with complex environmental systems in which organisms interact intimately with and are affected by their surrounding chemical and physical environment.
- Key processes and properties controlling carbon cycling or contaminant solubility may be occurring in only a small subset over possible domains in a given environment, yet these microscale processes and properties exert a disproportional influence on system behavior.

Enabling the connection of various omic measurements to biochemical and biogeochemical processes is a universal need in the environmental sciences. To this end, the Genomics:GTL program (GTL) has identified as one of its major mission-related goals the development of methods to relate genomics-based microbial ecophysiology (functionality) to the assessment of global carbon biosequestration strategies and climate impacts. The problem can be stated simply as the need to move from sequence to physiology to activities.

## Example Problems and Key Scientific Issues

### Carbon Cycling in Ocean and Terrestrial Ecosystems

Genomics governs the molecular processes that control cellular, organism, ecosystem, and, ultimately, global phenomena. Connecting mechanistic understanding at the molecular level to the physiological changes observed in organisms, ecosystems, and the global

climate represents a major challenge (see Box 1.1, Global Carbon Cycling Research, beginning on p. 10). Historically, the climate, ecosystem, and molecular biology research domains have had limited overlap because they differ widely in experimental and modeling approaches used, and results, in many cases, do not translate well across scales.

## Marine Environments

The following are key questions surrounding the carbon cycle:

- How do metabolic processes of microbial communities in marine habitats link to the global carbon cycle, with special attention to integration of processes across genetic, organismal, community, and ecosystem scales?
- What are the links among the composition of dissolved organic matter, nutrient limitation, and the structure of heterotrophic microbial communities in marine systems?
- How do environmental, ecological, and physiological factors interact to set the pathways and regulate the flows of carbon and other elements through upper-ocean ecosystems?

Phytoplankton (microscopic marine plants) and photosynthetic bacteria convert dissolved CO<sub>2</sub> into organic compounds in surface waters. By reducing the partial pressure of CO<sub>2</sub> in the upper ocean, photosynthetic marine microbes enhance oceans' physical absorption of CO<sub>2</sub> from the atmosphere. Without phytoplankton photosynthesis, atmospheric CO<sub>2</sub> concentration would be 150 to 200 ppmv higher (Laws et al. 2000). Large oscillations in phytoplankton abundance, therefore, significantly impact the oceans' ability to take up atmospheric CO<sub>2</sub>. Several challenging issues, such as the following, need study:

- Understanding the composition of microbial communities that dominate primary production in oceans (in a beginning phase).
- Determining differences in functional potential and metabolic processes of various types of photosynthetic microbes (poorly understood).
- Predicting how communities might be affected by climate change and its impact on the marine carbon cycle (difficult).

Metagenomics and related omic measurements have the potential to provide detailed insights into the structure and function of marine phytoplankton communities.

## Terrestrial Environments

In terrestrial ecosystems, plants also use photosynthesis to convert atmospheric CO<sub>2</sub> into organic compounds for building plant biomass and driving metabolic processes. Key issues include the following:

- How can we better distinguish between regulatory systems and molecular controls for partitioning carbon among plant structures versus cellular respiration among different soil pools, and how can we represent this new knowledge in models?
- How will climate change influence enzymes and biochemical reactions underlying water use efficiency, nutrient uptake, and many other processes? These processes control photosynthetic productivity as plants are subjected to levels of atmospheric CO<sub>2</sub> and other conditions that have not existed for the past 650,000 years and possibly millions of years.



- How can we better quantify photosynthesis, respiration, and other biological processes that influence carbon cycling? The problem is that the metabolic flux of material and energy through cells, organisms, and ecosystems is tightly linked to the abiotic environmental factors (e.g., temperature, precipitation amounts and timing, geographical features, nutrient availability, length of days and seasons, and sunlight exposure) that define a particular region. Climate is both a product and a catalyst of interactions between a region's physical environment and the biosphere, all of which are driven by the sun and affected by human activities.
- How do soil microbes and their activities respond to climate change, and what are the consequences for carbon cycling processes?

Critical for global-scale climate and biogeochemical models are accurate estimates of process rate constants, which influence biochemical functionality in organisms. Biochemical functionality generally is defined in terms of a catalytic property plus its rate constant ( $V_{\max}$ ) that can be plugged into models operating at larger-system scales.

A key question is the extent of sequence divergence in orthologs (e.g., in extracellular hydrolases) related to variability in  $V_{\max}$  (rate per unit biomass). In particular, to what extent is phylogenetic information (SSU rRNA and MLST) related to functionality (both substrate catalysis and environmental stress responses)? A key need is to be able to use omic information to provide estimates of catalytic rates and to inform process type and mechanism (see Fig. A2.1. Systems Approach for Predictive Modeling of Cellular Responses, p. 70).

Current measurements of genomes, transcriptomes, and proteomes can give, at best, relative abundances of molecules whose function is largely inferred from sequence homology. This is a major problem that will require a concerted, extensive effort as well as new and innovative approaches, technologies, systems for data and information sharing, and models. The potential payoff for success is tremendous, not just for carbon cycling but for all DOE science missions, environmental remediation, bioenergy, and beyond.

### Optimizing Productivity and Carbon Biosequestration in Managed Ecosystems

Managed lands account for about 30% of current global terrestrial net primary production (NPP), and ongoing land-use changes will cause the proportion of global NPP from managed lands to continue to increase. By the end of the 21<sup>st</sup> Century, managed ecosystems will dominate the planet. To establish a basis for optimizing carbon fixation and biosequestration in this context, a fundamental approach is needed to provide a molecular mechanistic set of options. Along with natural ecosystem behaviors, these options will form a substantial component of Earth System Models. Many of these studies have a common basis and data management needs similar to those of bioenergy studies. Objectives include the following:

- Identify basic processes that underlie gross and net primary production (GPP and NPP, respectively) of terrestrial plants; examine molecular controls on above- and belowground components of NPP; and assess areas in which knowledge gained through mechanistic studies could lead to enhanced carbon biosequestration in plant biomass or soils.
- Consider how efficient acquisition and utilization of resources (e.g., nutrients and water) contribute to maximizing rates of GPP and NPP in terrestrial plants;

identify the molecular basis of efficient resource utilization; and assess interactions between carbon and other resources that might be important in determining the rate, magnitude, or sustainability of biosequestration.

- Evaluate how GPP and NPP could be maximized in plant populations and communities and consider the role of genetic diversity and resource utilization in carbon biosequestration. The objective is to maximize NPP and litter input to soils, for example, over a growing season.
- Generate dynamic models (in silico leaf and in silico plant) that predict how changes in genetic regulatory networks can be used to enhance GPP or NPP by altering metabolic and developmental pathways in response to external perturbations or genetic manipulation.

### Leaf-Level Strategies

Emergent mechanistic and systems-based models of GPP provide potential opportunities to substantially increase carbon fixation in managed ecosystems, with impact on both DOE carbon biosequestration and biofuel strategies. The following are examples:

- **Modifying the diffusion resistance to CO<sub>2</sub> transport in leaves.** Mesophyll resistance is a significant limitation on carbon acquisition (24% reduction) and on water and nutrient use efficiencies.
  - **Taking steps to suppress or bypass photorespiration.** RuBisCo (Ribulose-1,5-bisphosphate carboxylase/oxygenase) evolved without the ability to discriminate between its primary substrate, CO<sub>2</sub>, and the wasteful reaction with oxygen (a 35% reduction in carbon capture).
  - **Engineering maladapted RuBisCo in plants.** RuBisCo in current C<sub>3</sub> plants is optimized for historic concentrations of CO<sub>2</sub>, 200 ppmv. Introducing RuBisCo into C<sub>3</sub> plants from other species that have a higher catalytic activity (and are better suited for higher CO<sub>2</sub>) would dramatically increase carbon gain despite less ability to discriminate for CO<sub>2</sub> over O<sub>2</sub>.
  - **Optimizing the distribution of nitrogen within the photosynthetic apparatus.** Nearly half of nitrogen invested in soluble protein in leaves is in RuBisCo. Manipulating the partitioning of nitrogen resources (e.g., in the regenerative phase of the Calvin cycle) could greatly increase the potential for carbon acquisition without any increase in the total nitrogen requirement.

### Plant-Level Strategies

- **Minimizing carbon-sink limitations and negative feedback on photosynthesis.** Source-sink interactions have a significant impact on photosynthesis and plant growth. Limited sink capacity results in decreased photosynthetic rates in leaf tissue. Photosynthetic activity is tightly regulated by sink demand. Therefore, increased productivity may be achieved by reducing sink limitations on photosynthetic rates. Opportunities to achieve these reductions come from recent experiments suggesting that sink regulation of photosynthesis is mediated by alterations in phloem loading (Chiou and Bush 1998; Vaughn, Harrington, and Bush 2002).
- **Optimizing carbon-nitrogen metabolism for increased plant productivity.** The interaction between CO<sub>2</sub> and nitrogen assimilation is of key importance to productivity. Assimilation of inorganic nitrogen into organic nitrogen requires photosynthetically derived carbon skeletons to serve as backbones for assimilation



of nitrogen into amino acids. Attempts to improve productivity or alter partitioning will be informed by improved understanding of central metabolism.

- **Partitioning carbon into organs and soil organic matter.** Modification of plant morphology and phenology can have substantial impacts on productivity by, for example, enhancing gas exchange in the shoots and increasing nutrient and water acquisition in the roots.
- **Identifying genes controlling biomass by using genetic screening approaches (forward and reverse genetics and natural variation).** The unifying question for this approach is whether previously unidentified genetic loci control plant productivity. Possible targets might include novel aspects of photosynthesis (from light reactions to RuBisCo and carbon metabolism) and water and nutrient use efficiencies. Completely novel pathways and master regulatory genes also may emerge from such genetic screens.
- **Optimizing biomass production versus respiration.** Partitioning of carbon between respiration and dry matter production is variable at the ecosystem, population, organismal, and tissue levels. Thus, establishing a mechanistic picture will enable optimization at all of them.

## Environmental Remediation

DOE contaminant fate and transport and biogeochemistry research focuses on natural microbial communities that significantly enhance or decrease the environmental mobility and toxicity of contaminants. The interdependent metabolic survival strategies used by microbial communities can directly or indirectly alter contaminant behavior in the subsurface or transform toxic contaminants into relatively benign forms. For example, *Shewanella* and *Geobacter*, two model dissimilatory metal-reducing microbes, currently are the subjects of systems-level research in GTL and OBER's Environmental Remediation Sciences Program. Via direct enzymatic reactions or indirectly via the generation of Fe(II), these microbes can reduce a range of radionuclides and toxic metals that are constituents of DOE wastes. Under anoxic conditions, for example, they can transform U(VI), which is relatively soluble and therefore can migrate in groundwater, to U(IV), which is insoluble and precipitates as a poorly soluble nanocrystalline solid (UO<sub>2</sub>). Studying these model microbes at a systems level is the first step in expanding our understanding of the structure, function, metabolic activity, and dynamic nature of microbial communities that have an important role in contaminant biogeochemistry. This knowledge is needed to predict microbe-mediated contaminant fate and transport and to understand how microorganisms may affect key biogeochemical processes in the environment. Of importance, however, is recognizing that these organisms do not function in isolation but rather are components of heterogeneous populations or microbial communities that may interact to various degrees and, in some cases, even exhibit interdependence.

A critical problem in biogeochemistry and environmental remediation science is the lack of understanding about how microbial processes are coupled to geochemical and hydrological processes influential in contaminant behavior and how these processes are scaled in heterogeneous environments. In addition, new tools are needed for measuring key microbial, geochemical, hydrological, and geological properties and processes in these systems. Less than 1% of all microorganisms collected at only a few DOE sites have been cultured and characterized in any great detail, and only a small

fraction of those have been sequenced. Even less is known regarding the interactions of microorganisms in communities. Metabolic processes observed in the subsurface often are the result of unique interactions between the microbial community and subsurface geochemistry. We have only begun to appreciate the existence of such systems, let alone understand them sufficiently to take advantage of their diverse capabilities and predict how they may influence contaminant behavior. Efforts are under way at several field sites (e.g., the Oak Ridge National Laboratory Field Research Center) to define the genomic potential of microbial communities using metagenomic and other molecular approaches to provide initial culture-independent insights into potential microbial functions such as the following:

- Linking genome sequence to functional potential, as described below, remains a key issue that must be resolved if the promise of genomics for understanding the function of microbial communities is to be realized.
- Collecting robust environmental data that can be linked directly to metagenomic sequence also is a critical issue. Environmental context is very important for interpreting such data.

Other processes important in modifying contaminant form and transport and in developing environmental remediation strategies include microbe-mineral interactions and resulting molecular structural and charge-transfer responses; microbial community responses (e.g., signaling, motility, biofilm formation, and other structural responses); and ensuing community functionality. The mechanistic linking of metabolism to contaminant transformation will represent an important advance from previous contaminant-fate models.

## Examining the Common Challenges of Carbon Cycling and Environmental Remediation

Microbial communities inhabiting terrestrial and aquatic environments are major players in the global carbon cycle and environmental remediation, but the organisms and biogeochemical processes they catalyze remain poorly understood.

While the genomes of hundreds of microorganisms from a range of terrestrial and aquatic habitats have been fully sequenced, they represent only a small fraction of total microbial diversity. New technologies such as metagenomics, metatranscriptomics, and metaproteomics offer a window into the metabolisms and lifestyles of the vast diversity of microbes, including uncultivated organisms. However, most successful applications have been applied to relatively “simple” microbial populations; the daunting complexity of most terrestrial and aquatic communities thus far has not yielded data that are easily translated into functionality.

### Annotation

A large generic annotation problem remains in genomics: predicting protein function from sequence and homology. In some cases, defining a general functional class of a specific protein, such as amino acid transporter, is relatively easy, but identifying substrate range (i.e., which amino acids it transports) can be extremely difficult. These issues can be important for answering ecophysiology questions and for determining function within metabolic networks. Even more challenging are situations in which homologies are poor or nonexistent. A potentially powerful approach for determining gene function and ultimately improving prediction is the combined



increasingly sophisticated and detailed models of complex processes contributing to and ultimately governing carbon cycling to produce increasingly quantitatively predictive models will require addressing model scalability and the coupling of mathematically heterogeneous representations. Furthermore, while current climate change models have “hooks” to incorporate parameterized models employing increasingly detailed carbon cycling data, next-generation models are likely to require new methods for sub-model parameterization and coupling that would rely on GKB data resources.

## Connecting Data to Function—Dealing with Complexity

To overcome the obstacles of translating omic data into function, researchers will need to develop techniques to enable targeted metagenomic (or other omic) research. Methods such as stable isotope probing or metabolic labeling with bromodeoxyuridine will allow us to effectively isolate important segments of the total microbial community without cultivation and thus begin to understand the functional roles of different community segments. Metatranscriptomics and metaproteomics, which target primarily the “active” microbial community and their expressed macromolecules, will result in unraveling complexity and provide insight into actively occurring processes. Single-cell genomics, using cells obtained via flow sorting or micromanipulation, has potential for even more targeted analyses of community members and for further reducing the impact of complexity on metaomic approaches.

While the native communities in soils and oceans are complex, techniques and approaches under development, such as those described above, can begin to overcome some of the technical issues associated with complexity. Additionally, understanding entire communities associated with key environments would be invaluable as a baseline.

As DNA sequencing becomes ever more accessible and less expensive, we can envision a human genome–type project such as that suggested in a recent National Academy of Sciences report, *The New Science of Metagenomics* ([http://www.nap.edu/openbook.php?record\\_id=11902&page=R1](http://www.nap.edu/openbook.php?record_id=11902&page=R1)), to target the microbiome in a spectrum of representative habitats. The National Institutes of Health’s Human Microbiome Project (<http://nihroadmap.nih.gov/hmp/>) and the Global Ocean Sampling (GOS; <http://collections.plos.org/plosbiology/gos-2007.php>) serve as models for this type of large-scale project. GOS is a useful starting point for mining these data for information relevant to carbon cycling research. The Department of Energy’s Joint Genome Institute is a valuable resource in this regard and already has embarked on the sequencing of numerous ecologically relevant organisms and communities, including those inhabiting soils and plant biomes (see sidebar, Analysis and Annotation at DOE’s Joint Genome Institute, p. 23).

## Data Integration and Linking Analysis and Experimentation

Once data are generated, researchers face the challenging task of integrating metagenomic, metatranscriptomic, and metaproteomic data with physical and biogeochemical data and ultimately relating them to carbon cycling or subsurface biogeochemical processes. Tools must be developed that can correlate biogeochemical parameters with genomic information and generate metabolic predictions based on incomplete genomic, transcriptomic, and proteomic data. Databases such as IMG (<http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>) and CAMERA (<http://camera.calit2.net>) exist for comparative analyses of metagenomic data, and initial efforts



should be actively encouraged to integrate metadata into these frameworks. Metatranscriptomic and metaproteomic projects, on the other hand, are still very much individual laboratory efforts in need of more advanced tools to facilitate comparisons with metagenomic data and with data from other environments and to make the data and information available to the broader research community.

A potentially successful approach for connecting gene sequence to function centers on the concept of intensively characterizing keystone genes and organisms. This approach, for example, could involve bringing relevant, experimentally tractable organisms into the laboratory for genomics and systems biology–type investigations. Eventually, research would move into the field to address fundamental questions such as, “Which genes function in the environment?” Needed are high-throughput methods that are sensitive but do not require high concentrations of biomass. Genomic and functional genomic approaches also can be used to gather information about which organismal processes are important in the environment and which data should be incorporated into models. Arguably, the number of keystone genes and organisms involved in carbon cycling and environmental remediation may be immense. For this approach to be successful, high-throughput analyses will need to be coupled with robust systems for data capture and data analysis that can be used to develop models of metabolism and regulation. How can massive volumes of high-throughput experimental systems biology data from ecological observations be automated to convert the data into a model that can be tested dynamically? This is a mathematics and computing problem—not about getting omic data simply because it can be obtained, but about using larger-scale models to drive development of data needed to populate models and increase their ability to predict.

Another approach for linking genomes to function would be to foster communication and data and information sharing among researchers in the metagenomics and general metaomics realms. A specific initial concept for advancing the dialog is mutual list building with intercomparison. Cultural exercises could be supported in which biologists itemize the metabolic- and biogeochemical-level information they can provide currently or will be able to provide in the near term to large-scale modelers. Scientists on the computational side of the carbon cycling or environmental remediation issues would construct lists of their own, reflecting their metabiogeochemical information needs. Overlap would be identified and concepts developed for iteration plus expansion of the intersection zone. This process can be viewed as a simple Venn diagram with growing disciplinary area coverage and increasing conjunction.

A next level of interaction could then be attained by leveraging gene expression. This process can be considered as classical annotation run both forward and backward for modeling purposes. Within the available environmental sequences, mapping of genes to enzymes remains largely incomplete. Laboratory experiments, however, with relatively simple, defined model systems can demonstrate at the metabolic level that certain key processes are active. Marine organisms that have been studied in this manner include cyanobacteria, diatoms, and other eukaryotes along with certain classes of heterotrophs. Metabolic pathways can be mapped in reverse to the active genes if they are not apparent from analysis of the sequences themselves. A subgenome is thus identified as containing an initial kernel of critical biogeochemical information. The means for accelerating this process are in fact related to the above discussion; simple list comparisons will pinpoint processes in which the required laboratory and field work can be performed quickly.

At a more challenging level, the entire sequence of data processing from genome to biogeochemical function may be viewed as a unified or potentially unifiable information sciences problem. Many individual steps already have been automated. Examples include genome reads leading to library development on the biological end of the spectrum and modular additions at the field, ecosystem, and global scales. In the near term, only automating the gaps in between will remain. The genome can be viewed as a vector of the most fundamental biogeochemical data, the transcriptome likewise, the proteome as an amino acid matrix, and the metabolome as a multidimensional space containing stoichiometries and rates. Integrating model assembly upward then becomes a matter of mathematically manipulating the resulting datasets from each stage. They may be configured in a relational manner. Standard matrix algebra is then applied to yield biogeochemical source-sink relationships. In fact, data arrays and their mathematical relationships constitute the most concise theoretical representation possible for global biotic systems.