Integrating Biology and Climate Through Systems Science

Genomics and Systems Biology

Genome Analyses Related to Climate Response of Ecosystems

The genomic and postgenomic eras have provided unprecedented potential to understand plants', microbes', and even entire communities' molecular and cellular responses to global change. Early efforts to apply genomic technologies and concepts to climate change research are already improving our capability to predict organism response to such change. For example, genome-wide analyses of soybean and poplar have revealed that elevated CO_2 down-regulates key genes in the octadecanoid pathway—a biosynthetic process producing an important defense hormone. Such findings may explain why ecosystems dominated by these plants show increased susceptibility to insect herbivory and delayed canopy senescence. Genomic technologies are now tractable in a wide range of organisms, including important agronomic (e.g., maize and rice) and forest (pine) species. Incorporating these high-throughput "omic" tools into current and emerging global change experiments will accelerate discovery and strengthen the predictive power of this research.

A main goal of such studies is determining whether individual genes or small groups of genes play keystone roles in controlling ecosystem capacity to store atmospheric CO_2. Related to such research is investigation into whether the similarity or dissimilarity in different ecosystems' reactions to global change is explained by coordinated and synergistic genetic responses across taxa.

Metaomics

Microbial communities inhabiting soils, oceans, and other types of terrestrial and aquatic environments play crucial roles in the global carbon cycle, yet these organisms and the processes they catalyze remain poorly understood. Although hundreds of terrestrial and aquatic microbial genomes have been completely sequenced (and even fewer have been studied in sufficient detail to develop robust models of metabolism and regulation), they represent only a small fraction of the total diversity of microorganisms, most of which defy laboratory cultivation. New and emerging technologies in metagenomics, metatranscriptomics, and metaproteomics, which can probe whole communities, offer insight into the metabolisms and lifestyles of diverse microbes, including those that remain uncultivated. The daunting complexity of most terrestrial and aquatic communities and the inability to directly translate gene sequence into potential biological function thus far have limited our ability to extract detailed insights into functionality.

Overcoming these obstacles will require, in part, developing and pursuing techniques that enable targeted metagenomic (or other omic) research. Using narrowed, highly specific approaches makes study of a microbial community manageable and progressively helps unravel the complexity of the overall system. Specifically, methods such as stable-isotope probing or metabolic labeling with bromodeoxyuridine will allow scientists to effectively target important segments

of a microbial community without cultivation and thus begin to understand these particular segments' functional roles. Metatranscriptomics and metaproteomics, which by nature primarily target the metabolically "active" microbial community and its expressed macromolecules, will provide real-time insight into actively occurring processes. Single-cell genomics, using cells obtained via flow sorting or micromanipulation, offers the potential for even more targeted analyses of microbial community members, further reducing the challenges arising from the incredible complexity within microbial communities.

These developing techniques and other similar approaches can begin to surmount some of the technical complications related to studying complex and heterogeneous soil and marine microbial communities. Equally important is an overall understanding of entire communities associated with key environments. Such an understanding would serve as an invaluable baseline from which to view future metaomic studies relevant to carbon cycling. As DNA sequencing becomes increasingly accessible and less expensive, a human genome–type project would target the microbiome in a spectrum of representative habitats, as suggested in a recent report by the National Academies Press (http://www.nap.edu/openbook.php?record_id=11902&page=R1). Serving as models for this type of large-scale endeavor are the National Institutes of Health human microbiome project (http://nihroadmap.nih.gov/hmp/) and the Global Ocean Sampling survey (http://collections.plos.org/plosbiology/gos-2007.php). The latter, which resulted in a massive metagenomic dataset, also is a useful resource for data-mining information relevant to carbon cycling and for applying complementary omic methods for testing hypotheses regarding community function. The Department of Energy's Joint Genome Institute, also a valuable resource in this regard, has begun sequencing numerous ecologically relevant organisms and communities, including those inhabiting soils, plant biomes, and oceanic environments (http://www.jgi.doe.gov/).

Systems Biology

Achieving a predictive, systems-level understanding of carbon processing by plants, microbes, and biological communities will require integration of fundamental science and technology. A key emphasis should be developing and employing genomic and systems biology approaches to model, for example, the regulatory networks that control carbon flow and fate, from assimilation by phototrophs to processing of organic matter by heterotrophs. Succeeding in this endeavor will allow scientists to predict molecular-network states under untested conditions, such as in anticipated climate change scenarios or gene modifications. The power of systems biology to accomplish such goals is illustrated by the recent success in predicting gene regulatory networks that control the physiology of a free-living bacterial cell in response to genetic and environmental perturbations (Bonneau et al. 2007). A central goal of systems biology approaches in this context is to develop models of metabolic and regulatory networks in keystone soil and marine microbes, plants, and biological communities that ultimately will inform climate and biogeochemical models (see sidebar, Systems Biology, p. 97).

Systems Biology

Systems biology can be defined as "the exercise of integrating the existing knowledge about biological components, building a model of the system as a whole, and extracting the unifying organizational principles that explain the form and function of living organisms" (von Bertalanffy 1968). Genome-scale analyses in microbes and plants have the potential to provide the necessary data to understand on a systems level how an entire organism works. In a practical sense, a systems approach to understanding biology can be described as an iterative process including (1) collection and integration of all available data (ideally for all of an organism's components and their relationships), (2) system modeling, (3) experimentation at a global level, and (4) generation and testing of new hypotheses (Ideker et al. 2001). The ultimate goal of a systems approach is not to describe and model what is known, but to predict how a system will react under untested conditions or in response to perturbation. Only then can researchers use systems-based models in a predictive fashion to manipulate biological systems for optimizing a specific process or function.

Ecologists and physiologists for many years have used systems biology to study organisms, yet applying this approach to examine molecules is only now feasible with the advent of genomics-inspired technologies able to supply a sufficient volume of information at many levels of organization. Thus, the postgenomic era offers the prospect of integrating knowledge across different levels of biological organization and anchoring this insight at the molecular level.

Connecting Omics to Biochemical Function

A major DOE objective is developing methods that use knowledge of genome-based microbial ecophysiology (i.e., functionality) to ultimately assess global carbon biosequestration strategies and climate impacts and feedbacks. The challenge of this objective can be stated simply as the need to advance scientific understanding from "sequence to physiology to activities." Accomplishing these goals requires a clear strategy for selecting which processes and systems are most important for developing a predictive understanding of carbon cycling and biosequestration (see Chapter 2, Technical Strategy, including Fig. 2.1. Scales and Processes of the Global Carbon Cycle, p. 16).

Identification of these elements could be aided greatly by an approach centering on the concept of intensive characterization of keystone genes and organisms. This method, for example, first could involve genomic and systems biology laboratory studies of relevant, experimentally tractable organisms or communities. Research then would progress to field experiments to answer fundamental questions such as which genes are functioning under various environmental conditions. The latter studies must include sensitive high-throughput methods not requiring large concentrations of biomass. Genomic and functional genomic approaches also can be used to reveal organismal processes and characteristics important in the environment and thus necessary for incorporation into models.

Critical for global-scale climate and biogeochemical models are accurate estimates of process rate constants, which influence biochemical functionality in organisms (see Fig. 2.2b. Knowledge Integration and Synthesis, p. 20, in Chapter 2, Technical Strategy). This functionality (V_{max}, rate per unit biomass) generally is defined as catalytic property plus rate constant, which can be incorporated into system models operating at larger scales. A crucial enabling research need is the ability to use omic information to provide estimates of catalytic rates and identify the types of processes and mechanisms occurring in organisms. Currently, genomic, transcriptomic, and proteomic measurements can give, at best,

relative abundances of functional molecules whose activity is inferred largely from sequence homology. Thus, more precise assessments of biochemical function are needed and will require concerted, extensive research as well as new and innovative approaches and technologies. Achieving this level of functionality understanding has the potential to tremendously advance not only carbon cycling objectives, but all DOE science missions, including those related to environmental remediation, bioenergy, and beyond.

Another difficulty in progressing to a genome-to-activity understanding involves challenges associated with annotation—predicting protein function from DNA sequence and homology. In some cases, defining a specific protein's general functional class, such as an amino acid transporter, is relatively easy. However, identifying its substrate range (i.e., which amino acids it transports) can be extremely difficult, yet doing so can help answer important ecophysiology questions and determine the function of these molecules within metabolic networks. A potentially powerful approach for determining gene function and ultimately improving predictive capabilities combines comparative genomics with experimental techniques such as those used by Yang et al. (2006) to characterize the *N*-acetylglucosamine utilization pathway in *Shewanella*. The study identified genes involved in this particular metabolic pathway. A complementary research method would target specific enzyme systems that process important extracellular compounds key to carbon cycling in terrestrial and marine systems.

Connections to Phylogeny

An important functionality question relating to variability in V_{max} is the extent of sequence divergence in orthologs (i.e., similar genes or gene segments appearing in different species and arising from a common ancestor). Studying this divergence, in extracellular hydrolases for example, can provide useful insight into how phylogenetic information [structure of bacterial small-subunit ribosomal RNA (abbreviated as ssu rRNA) and multilocus sequence typing (MLST)] relates to functionality, both substrate catalysis and environmental-stress responses.

Stable-isotope probing offers one approach to further advance these studies to determine biochemical function. For example, labeling key organic substrates with ^{13}C could help identify important taxa (phylogenetic designations) that function in the carbon metabolic process. This labeling could be conducted in representative habitats worldwide. Resulting phylogenetic information would then be used to isolate representative microbes from taxa carrying out the functional processes important to carbon cycling. (Amann, Ludwig, and Schleifer 1995; Madsen 2005). Next, a set of microbes covering the phylogenetic breadth of a key taxon would be identified, and functional process rates (V_{max}) under optimal conditions or functional response to environmental stress would be measured. Variance in these properties then could be determined across the taxon's phylogenetic breadth, and if little or none is observed, phylogeny can directly inform functionality. In summary, these three steps connect omic approaches to biochemical function:

- Obtain relevant sequence information (using MLST for phylogenetic placement and DNA sequences for specific functional genes).
- Apply analysis of variance techniques to determine if functional rate is predictable from gene-sequence information.

- Conduct comparative genomic investigations of strains within a species to provide an estimate of core capabilities of a specific taxon.

Tracking carbon via ^{13}C labeling and determining rate constants for carbon processing through different ecosystems could become a very important tool with direct linkages to larger, perhaps even global, scales of the carbon cycle. Moreover, this approach offers the additional virtue of obtaining phylogenetically informative tagged macromolecules.

Value and Challenges of Visualization Tools and Modeling

Linking genomics-based information to function requires both genome-scale data generation and systems biology tool development. Generation and collection of transcriptomic, proteomic, and metabolomic data require critical parameters that must be assayed and quantified. Computational requirements include development of visualization and other types of tools to integrate genome-scale data over various time scales of experimentation. Also critically needed are predictive modeling tools.

The value of visualization tools in showing genomic relationships is illustrated by the use of multinetworks to graphically display information about the manifold connections among genes, proteins, and molecules—all generically referred to as "nodes" in a network (see Fig. 7.1. Multinetwork Analysis of *Arabidopsis* Genome, p. 100). Node-linking "edges" are drawn based on experimental evidence or predictive algorithms. For example, protein:protein and protein:DNA edges could be determined experimentally but also might be predicted based on, for the former, two homologous proteins interacting in a different species or, for the latter, the presence of a transcription-factor binding site in the promoter of a gene. Another experimentally derived edge, for instance, could originate by determining that a gene encoding a certain enzyme uses a particular metabolite in a nonreversible catalytic reaction. Thus, a gene-encoding enzyme:metabolite edge would represent this interaction. An edge connection between genes also could be drawn based on transcriptional activation of a target gene by a transcription factor, depicted as a protein:DNA interaction edge. The latter two examples include nodes connected by "directed edges" (e.g., the transcription factor regulates the target gene, not vice versa, and thus is represented by a directional arrow). Alternatively, an edge might be nondirected, as is the case for those representing a protein:protein interaction.

A great obstacle to connecting genomic data to biological function centers on incorporating this information into models that can be tested dynamically. For example, researchers are faced with mathematical and computational challenges—automating and integrating into models the massive volumes of high-throughput data from experimental systems biology as well as that from ecological observations. Generation of these omic data should be motivated by the specific need to build larger-scale models rather than indiscriminate collection of information. In turn, the larger scale will drive data development to populate these models, thus enhancing their predictive capabilities.

Fostering communication between modelers and metaomic researchers is a first step in identifying the data most important for improved models. One way to do so is mutual list building and intercomparison of such lists between the two groups. For example, biologists would itemize the level of metabolic and biogeochemical

information they currently (or in the near future) can provide to large-scale modelers. Meanwhile, computational scientists studying global change would identify their metabiogeochemical data needs. Comparisons between the groups would identify key areas of overlap and facilitate concept development and expansion of intersecting research.

Simple list comparisons also can be valuable in helping modelers and metaomic researchers readily identify—and thus ultimately connect—experimentally observed enzymatic or protein functions and associated gene sequences. Making these connections involves leveraging gene expression to determine the sequence underlying a metabolic pathway of particular importance to modeling. This process can be considered classical annotation run both forward (using DNA sequence to determine protein function) and backward (using observed protein function to sequence DNA). For example, within sequences derived from metagenomic surveys, the mapping of genes to enzymes remains largely incomplete.

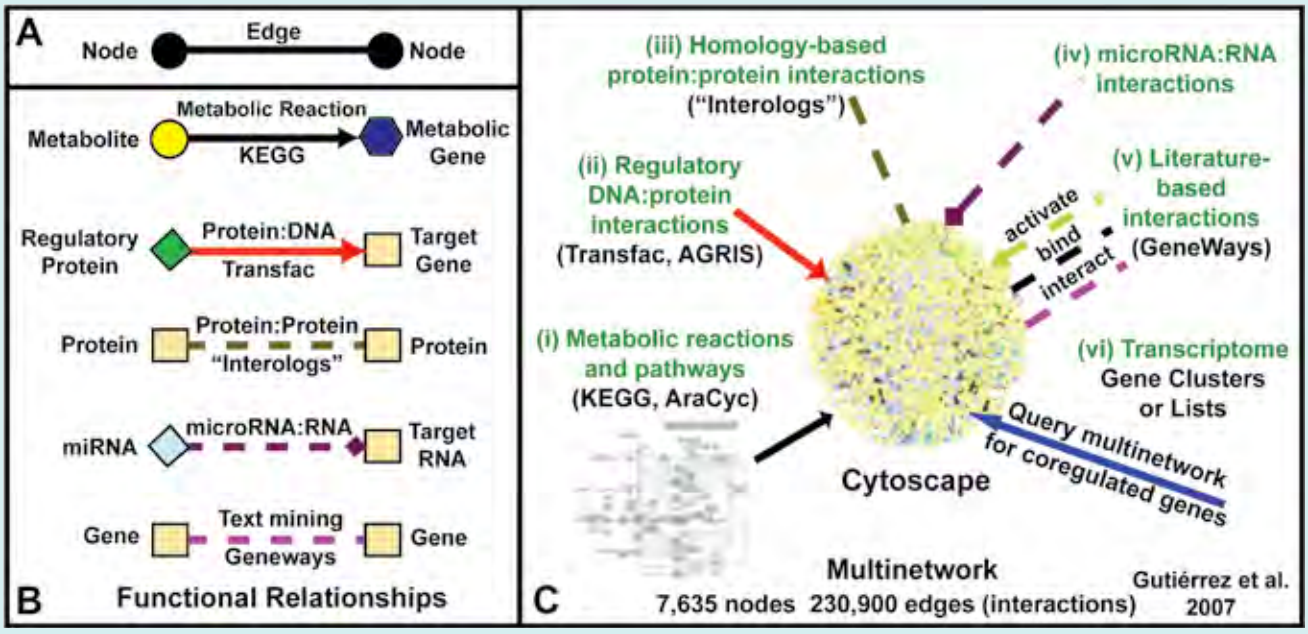

Fig. 7.1. Multinetwork Analysis of *Arabidopsis* Genome. (A) Networks in their simplest forms are made up of "nodes" and the "edges" that connect them. (B) Various types of node connections can be displayed using colors and shapes that indicate different types of molecules and the relationships between them. (C) Multinetwork representation of *Arabidopsis* metabolic, regulatory, and predicted regulatory connections between genes, proteins, and metabolites (Gutiérrez et al. 2007; Gifford et al. 2006). Specifically, the data used to draw edges comes from (i) information about metabolic reactions and pathways from the KEGG and AraCyc databases; (ii) known DNA:protein regulatory interactions from the Transfac and AGRIS databases; (iii) predicted protein:protein interactions based on homology to experimentally verified protein:protein interactions in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans* using the "Interolog" approach (Gutiérrez et al. 2007); (iv) predicted relationships between microRNAs and their targets (provided by Pam Green and Blake Meyers, Delaware Biotechnology Institute, University of Delaware, USA); and (v) known interactions between genes or proteins gleaned from published literature using the text-mining tool GeneWays (A. Rhetzsky, University of Chicago). The Cytoscape software is used to visualize the multinetwork in an interactive way (Shannon et al. 2003); (vi) the multinetwork can then be queried to find regulatory subnetworks and interactions between a subset of genes.

[Source: Gifford, M. L., R. A. Gutiérrez, and G. M. Coruzzi. 2006. "Modeling the Virtual Plant: A Systems Approach to Nitrogen–Regulatory Gene Networks," Essay 12.2, Chapter 12: Assimilation of mineral nutrients (http://4e.plantphys.net/article.php?ch=12&id=352). In *A Companion to Plant Physiology, Fourth Edition* by L. Taiz and E. Zeiger. Sinauer Associates, Inc., Publishers, Sunderland, Mass. (see also http://www.virtualplant.org and Gutiérrez et al. 2007).]

However, laboratory experiments with relatively simple, defined model systems can demonstrate at the metabolic level the activity of certain key enzymatic processes whose genetic controls may be unknown. (Such experiments have been used to study marine organisms, including cyanobacteria, diatoms, and other eukaryotes along with certain classes of heterotrophs.) These observed metabolic pathways, if not apparent from initial genomic analyses, can be mapped in reverse to determine the gene sequence directing them. Reverse mapping thus identifies a subgenome containing a piece of critical biogeochemical information. List comparisons between modelers and experimentalists can accelerate this process by pinpointing important pathways whose genetic bases can be determined by quick laboratory and field studies.

The entire progression of data processing—from genome sequences to biogeochemical function—may be viewed as a unified (or potentially unifiable) information-sciences challenge. Many of the individual steps spanning this progression already are automated. For example, genome sequencing (molecular-level data) has driven development of databases that now feature modular ecosystem (global-scale) information. In the near future, research must attempt to automate intermediate data collection, including information on a system's full complement of RNA transcripts (transcriptome), expressed proteins (proteome), and metabolites (metabolome). Useful to automation efforts is viewing the genome and transcriptome as vectors of the most fundamental biogeochemical data, the proteome as an amino acid matrix, and the metabolome as a multidimensional space containing stoichiometries and process rates. Integrating model assembly to higher scales then becomes a matter of mathematically manipulating the resulting datasets from each of these stages of biogeochemical function. Data 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 possible theoretical representation of global biotic systems.

Knowledge Integration and Synthesis with Biogeochemical Models

Accurately quantifying contemporary terrestrial carbon sinks and projecting their future stability require continuous improvement of models via integration and synthesis of various datasets and greater understanding of key mechanisms.

Scientists have developed various terrestrial biogeochemical models that simulate ecosystem carbon processes (e.g., Parton et al. 1987; Luo and Reynolds 1999; Cramer et al. 2001; McGuire et al. 2001). These models generally incorporate current understanding of ecosystem activity and use carbon-process data for parameterization and validation. In fact, qualitative knowledge of major carbon-transfer processes within ecosystems is fairly well developed. For example, as discussed in Chapter 3, Carbon Flows in Ecosystems—Ecosystem Processes, p. 27, scientists have established that (1) a portion of photosynthetically fixed carbon is used for plant growth, and some is released via plant respiration; (2) plants store carbon in live structures for periods ranging from several months to hundreds of years; and (3) dead plant materials (i.e., litter) are partially incorporated into soil organic matter (SOM), which can sequester carbon in soil for centuries and longer before it is broken down into CO_2. Knowledge of carbon-transfer processes has been incorporated into a common structure shared by most biogeochemical

models. This modeling structure partitions photosynthetically fixed carbon into several pools (Rastetter et al. 1997; Luo et al. 2001), with transfers among pools controlled by the carbon-donor pool (Luo and Reynolds 1999).

Critical for effective climate–carbon cycle models is robust representation of nitrogen, whose availability strongly regulates carbon biosequestration amid rising atmospheric CO_2 concentrations**.** More and more, carbon cycling models are incorporating nitrogen processes according to stoichiometric relationships observed between carbon and nitrogen in all plant and soil pools. However, little is known about shifts in ecosystem nitrogen availability in response to global change. Alterations in the nutrient's total amount in an ecosystem are related to microbially mediated nitrogen fixation and processes resulting in nitrogen loss (see Fig. 3.5. Nitrogen Cycle, p. 41, in Chapter 3, Carbon Flows in Ecosystems—Ecosystem Processes). Understanding these shifts thus requires more research on nitrogen fixation in natural ecosystems under steady state and in response to elevated CO_2 and other climatic changes and disturbances. Also needed is greater insight into how denitrification, leachage, volatilization, and other nitrogen-loss processes respond to increased atmospheric CO_2 and global change.

Climate warming affects almost all physical, chemical, and biological processes. Experimental studies have identified several key regulatory mechanisms underlying ecosystem responses to warming. Such responses include acclimation of photosynthesis and respiration, shifts in phenology and nutrient dynamics, and ecohydrological regulation (Luo 2007). Most models, however, still are incapable of quantitatively representing how climate change alters basic ecosystem processes.

Carbon allocation and partitioning among plant parts and autotrophic respiration and among different soil pools are not well understood or represented in models. Rising atmospheric CO_2 concentration, climate warming, altered precipitation, and nitrogen deposition likely change trophic cascades from plant to litter to soil organic matter, resulting in shifts in concomitant community structures of plants and microbes. Critical to predicting the implications of such shifts are improved models, particularly Dynamic Global Vegetation Models (DGVM) used to study how plant functional types respond to disturbances and other factors. Improving DGVMs requires enhancing model-response functions that link alterations in community structure to global change factors at different time scales.

Several **requirements** are necessary to enhance carbon cycle modeling capability, including the following.

1. **Model Structure.** Terrestrial carbon cycling models require multiple carbon pools with different accumulation and residence times.
2. **Initial Value Problems.** Models should accurately quantify contemporary carbon sinks; attribute them to different historical causes, such as disturbances and climate change; and relate sink state to age.
3. **Response Functions.** Models should represent ecosystem-response functions—as they relate to major carbon processes—to environmental variables of global change. Key areas include:
    a. Nitrogen fixation, nitrogen loss, and nutrient limitations for plant and heterotrophic processes in response to rising atmospheric CO_2 concentration and climate change.

b. Climate change effects on basic biological, chemical, and physical processes represented in models (e.g., acclimation of photosynthesis and respiration at the enzyme level and shifts in phenology).

c. Alterations in carbon allocation and partitioning, including autotrophic and heterotrophic respiration response to elevated CO_2 and other global changes.

d. Trophic-cascade (plant → litter → SOM) sensitivity to environmental factors.

e. Carbon-nitrogen-phosphorus-water interactions coupled to nitrogen fixation.

f. Climate-induced shifts in plant and microbial community structure. (DGVMs must link these alterations to global change at various time scales).

g. Hydrological controls.

Parameter Values and Their Variability. Better integration of field data into models is needed to improve predictions of terrestrial carbon biosequestration and feedbacks to climate. Field data are used to constrain model parameters, characterize dynamic disequilibrium of carbon cycling, and quantify carbon biosequestration over space and time (from years to centuries). Acquiring these valuable data requires careful experimental design to optimize sampling.

Spatial Patterns of Carbon Sinks. Scientific knowledge and data on carbon cycling and biosequestration largely are derived from research in temperate climates. Broadening our understanding of the global carbon cycle thus requires more information from several understudied areas such as tropical and high-latitude zones.

Data-Assimilation Techniques. Further development is required to improve integration of information with models. Such techniques are new to ecology but are well established within the climate research community. Early data-assimilation papers (e.g., Williams et al. 2004; Braswell et al. 2005) state that for model-data integration to advance, consistent information is vitally needed on long- and short-term processes across biomes, climate zones, and disturbance classes. Furthermore, measurements of long-term ecosystem fluxes of carbon, nutrients, water, and energy are essential to develop, test, and apply carbon cycling models.

Development and Refinement of Spatial-Temporal Carbon Cycling Models across Scales

Critical to climate-mitigation and carbon biosequestration strategies is the ability to conduct predictive modeling. Needed are models that anticipate how a system will react under specific conditions rather than those that simply reproduce results already established through experimentation or observations. Achieving this predictive capability requires equipping models with increasing levels of detail over different space and time scales (see Fig. 2.1. Scales and Processes of the Global Carbon Cycle, p. 16, in Chapter 2, Technical Strategy). However, representing key processes at the necessary scales is a central challenge of global carbon cycle research. Part of the complication arises from the disconnect in information from scientists working at different spatial and temporal scales. For example, environmental scientists can measure ecosystem functions and phenomena but have difficulty relating results to higher and lower scales and in extrapolating behavior outside the range of observations. Understanding and effectively modeling carbon processes thus require data from scientific investigation across all scales. For instance, researchers examining system attributes at lower scales can capture

important details, while those working at higher scales can provide data needed for model parameterization. Moreover, since most climate effects on carbon cycling are manifest at the macroscale, efforts should be made to generate data relatable to higher scales.

Current climate change models rely on geophysical data obtained at widely varying scales. Much of this data is from well-characterized sources with longstanding methods of incorporating such information into climate change models. However, researchers envision a continuous progression of modeling science characterized by increasingly accurate predictions and critical new capabilities to ask and answer "what if" questions concerning climate change. Attaining the desired level of predictability will require models dramatically more detailed and mechanistically based. Advanced climate change models also must span ever-increasing lengths and time scales and draw upon more precise and quantitative data on all ecosystem processes relevant to carbon cycling.

The lengths and time scales of carbon cycling processes represented in future climate change models likely will range from microscopic to aggregate (mm to cm) to field and beyond. In particular, model development that includes carbon processing across scales will generate data yielding fundamental understanding of complex biological systems—from single cells to microbial communities to organisms with multiple cells and tissues to diverse ecosystems with many species. Furthermore, these data also will aid development of parameterized dynamic models capable of quantitatively predicting ecosystem response to climate change and disruptions. Such model development in some cases will require measuring and quantitatively characterizing carbon cycling processes specifically for model parameterization and validation as opposed to meeting needs of general scientific interests. Other development requirements include new methods for coupling parameterized models of system response at various levels of complexity and scales to informatics data derived from ecological observations or high-throughput systems biology studies of cellular processes. In particular, improved model scalability and coupling of mathematically heterogeneous representations are necessary for developing increasingly sophisticated and detailed models that include complex processes contributing to and ultimately governing carbon cycling. Moreover, current climate change models have "hooks" to incorporate parameterized versions using more detailed carbon cycling data, but next-generation models probably will require new methods for submodel parameterization and coupling.

Integration of Different Types of Data into Models

The science of carbon cycling and biosequestration across hierarchical levels from genomics to ecosystems requires integration of measurements and models using systems approaches. First, conceptual frameworks must be developed to guide the integration of theoretical understanding, knowledge of carbon processes, data, and quantitative relationships. Such frameworks also would enable scientists to connect nodes—relationships among genes, proteins, and molecules—at different hierarchical levels and evaluate scalable variables. Thus, development of quantitative models should be based on these conceptual frameworks.

Although existing carbon cycle models can connect information from leaf-level photosynthesis to global flows of carbon, advanced models are needed to link

knowledge, data, theory, and quantitative relationships from genomic studies to subcellular and cellular processes and eventually to those occurring at organismal and ecosystem scales. Such model enhancement can be aided by recently developed data-assimilation techniques integrating observational data into ecological models with rigorous statistical and mathematical approaches. Data assimilation is a valuable tool to improve model parameterization, choose between alternative model structures, design better sensor networks and experiments for data collection, and analyze uncertainty of model predictions. The ecology research community recently explored, examined, and developed various data-assimilation techniques (e.g., inverse analysis, hierarchical Bayesian analysis, model-selection approaches, and state-space modeling) to analyze multiscale ecological data in space and time.

Uncertainty in Model Projections

Although carbon cycling models have been used extensively to predict carbon biosequestration in terrestrial ecosystems, uncertainty associated with model parameters and predictions has not been analyzed carefully. If such uncertainty is inadequately assessed, carbon sink potentials cannot be understood fully. In fact, some carbon sinks may be underestimated, while others overestimated, even to the extent resulting in contradictory source-sink designations. In such situations, policies to stabilize CO_2 concentrations based on current understanding will fall short in meeting environmental-mitigation targets.

Considering the importance of uncertainty analysis to policymaking, the research community investigating global climate change recently directed considerable attention to studying the stochasticity and uncertainty in ecosystem processes and how various sources of randomness affect prediction of ecosystem changes (Murphy et al. 2004; Dose and Menzel 2004; Forest et al. 2002; Wang et al. 2001). Expert-specified probability density function [(PDF) e.g., Murphy et al. 2004] has been used to quantify key uncertain properties of climate change simulations. Researchers have introduced the Bayesian paradigm to incorporate *a priori* PDFs with measurements to generate *a posteriori* PDFs for parameters of ecosystem models (Braswell et al. 2005; Knorr and Kattge 2005). With a probabilistic approach, Mastrandrea and Schneider (2004) presented a cumulative probability function (CDF) to assess dangerous anthropogenic interference and showed CDF utility by applying it to analysis of uncertainty in model predictions of future changes. On a global scale, the Bayesian approach has been applied to constrain parameters in biosphere models against atmospheric CO_2 concentration data and to assess biosphere carbon fluxes and uncertainties (Kaminski et al. 2002; Rayner et al. 2005). A probabilistic inversion within a Bayesian framework was conducted by Xu et al. (2006), who used six datasets and a terrestrial ecosystem model to evaluate uncertainty in parameter estimation and projected carbon sinks. In this analysis, measurements were treated as random variables with certain probability distributions. A joint PDF was constructed for model parameters to analyze information within observed datasets. Samples were taken from the joint PDF using a Markov chain Monte Carlo technique appropriate for sampling high-dimensional PDFs of model parameters and widely used in inverse problems in engineering and geosciences (e.g., Dosso and Wilmut 2002; Oh and Kwon 2001; Geman and Geman 1984). The samples were used to construct marginal distributions for model parameters, calculate parameter correlations, and make CDFs for simulated pool sizes in forward modeling.

Regional and Geographic Issues

Both on land and in the ocean, certain regional-scale ecosystems require special treatment in Earth System Models because they are either sensitive to climate change or poorly understood. Terrestrial examples include tundra—which already may be transitioning to shrub lands—and tropical and boreal forests. In the ocean, distinctive areas include (1) coastlines subject to rapid nutrient recycling from river inputs and proximate, underlying sediment; (2) the continental shelf, which is considerably more complex than the pelagic zone and can be defined in such a way that it influences a large fraction of planetary geocycling surface area; and (3) the poles, further distinguishable by Arctic and Antarctic regions. The former consists of an enclosed peripheral sea surrounded by landmasses while the latter consists of the opposite. Although climate change will affect polar biota hardest and fastest, Arctic and Antarctic organisms may react very differently to induced stresses.

In some cases, critical geochemical processes occur at the intersection of special terrestrial and oceanic ecosystems. Methane clathrates, for example, form preferentially on continental shelves below the coldest and most productive waters. These clathrates harbor carbon stocks comparable to those of known global coal reserves. If even a tiny fraction of the clathrates is destabilized, the implications for further climate change would be huge and carbon biosequestration efforts overwhelmed.

Visualization Tools

Various visualization tools are excellent catalysts for discovery and communication across disciplines and at multiple scales. Moreover, applying systems biology approaches to carbon cycling and biosequestration research will require development of these and other tools and methods for integrating disparate data types across different scales. In general, new tools are needed for informatics, imaging, math, and statistics to enable dynamic modeling and visualization of processes ranging from molecular networks in cells to populations in ecosystems. Integrating these tools in a common platform is a key goal that will facilitate a better understanding of how internal and external perturbations affect processes, pathways, and networks controlling organism growth and development and how these disruptions impact ecosystem "nodes."

The visualization aspect of research often is underappreciated but can catalyze cutting-edge research. Furthermore, visualization tools and approaches can greatly enhance communication and information sharing across scientific disciplines. In particular, the genomics–to–cell–to–global ecology concepts of interest to the carbon cycling research community hold rich potential for detailed, color-enhanced visual representations. Multilevel biological networks, three-dimensional biochemistry, and large-scale biogeography can be displayed and animated simultaneously. For ocean studies, the interaction of turbulent fluid flow with ecodynamics can be simulated and graphically displayed along time coordinates. Advances in visualization technologies have made associated tools extremely useful and, in some cases, critical to research. As scientists reach milestones in network and geochemistry mapping, workshops should be held for collective examination of these advances. Detailed, mobile data fields associated with such examinations tend to stimulate new directions of analysis and cross-disciplinary discussion.

Crosscutting Issues, Measurement Methods, and Strategies

State-of-the-Art Instrumentation and Methods

Imaging and Microspectroscopy

A major crosscutting technology need identified at the DOE Carbon Cycling and Biosequestration Workshop is development of advanced imaging and microspectroscopy tools. In particular, new imaging technologies are required to analyze at appropriate scales key ecosystem components, processes, and properties, including macromolecular complexes, microbes, plant-root cells, and soil microaggregates. For ocean systems, there are parallel needs to characterize phytoplankton and marine snow as it forms and decomposes. These analyses demand new and sensitive approaches (e.g., sensors, probes, and stable isotopes) for measuring and monitoring biological activities and physical and chemical processes in situ. Such approaches could give valuable insight into extracellular enzymes and their activities, which are particularly important in organic carbon processing in both soil and marine environments. For example, many of the polymers produced by phototrophs are too large to be transported into microbial cells and require depolymerization by extracellular enzymes. Relatively little is known, however, about the nature of these proteins and their in situ catalytic activities, requiring further study aided by new imaging tools. Technologies also are needed for measuring at high resolution macro- and micronutrient concentrations and chemical forms in situ. Such information is critical for providing the context for biological properties and processes and assessing how they influence carbon cycling. This type of environmental characterization data is critical for determining how organisms respond to changes in their surroundings and, when coupled with genomic and other omic data, is particularly effective for understanding these responses. Similarly, more insight also is needed into how organism responses alter the environment. Numerous in situ observing systems are beginning to provide long-term data on how organisms' environments respond to such shifts. Fully exploiting this detailed information requires improvements in measurement technologies, such as automated soil moisture profiling, precipitation sensors, and tools to assess soil enzyme activities.

Understanding the fundamental mechanisms that control the biogeochemical cycling of carbon requires analyzing the physical and chemical micro- and macroenvironments at soil-water-microbe-plant-fungi interfaces. The physical and chemical microenviroments at these interfaces potentially are some of the more critical controls on biogeochemical cycling of elements, yet characterizing them is difficult. Overcoming this challenge will require high-throughput imaging and chemical and structural analysis of bacteria, roots, and soil aggregates coupled with investigations of microbial communities and metabolic expression and activity.

To accomplish the desired level of analysis, new facilities and high-throughput instrumentation—used in parallel with omic approaches—should be developed for physical and chemical characterization of environmental systems at many scales. These facilities and techniques must be available to the entire scientific community. Some will require additional technical experts who must be knowledgeable about carbon cycling and carbon biosequestration. Furthermore, an integrated use of standard and exotic techniques should be employed often to enable new insights.

Key Research Question

1. **How can we assess uncertainty in model estimates of ecosystem carbon sinks and improve confidence in these projections?**

In addition to studies of natural materials, integrated approaches must be used to investigate defined yet representative systems in the laboratory. Some technologies important or potentially important to understanding carbon cycling and biosequestration include (1) microelectrodes; (2) focused ion beam; (3) secondary ion mass spectrometry (nano-SIMS); (4) time-of-flight (TOF) SIMS; (5) nuclear magnetic resonance (NMR); (6) synchrotron-based approaches; (7) electron microscopies (e.g., transmission electron microscopes and conventional and environmental scanning electron microscopes); (8) atomic force microscopies; and (9) gas chromatography, liquid chromatography, and mass-spectrometry approaches.

Synchrotron-Based Approaches

Some synchrotron-based instrumentation already is available and being used to measure the chemical and physical characteristics of biological and environmental samples relevant to carbon cycling and biosequestration. These techniques include (1) protein crystallography (for determining protein structure); (2) small-angle X-ray scattering (for measuring the size distribution of solution-phase submicron particles); (3) X-ray tomography (for three-dimensional characterization of soil porosity and connectivity and for measuring organic carbon, water, and mineral distributions within soil microaggregates, aggregates, and microcosms); (4) hard and soft X-ray fluorescence and transmission microscopies [for providing suboptical spatial resolution information, such as size and chemical speciation of organic matter, micronutrients, and macronutrients (see, for example, Fig. 3.7a–b. Distribution of Micronutrients in Plant Roots and Associated Fungal Hyphae, p. 45, in Chapter 3, Carbon Flows in Ecosystems—Ecosystem Processes)]; (5) soft X-ray spectroscopy (for chemical speciation analysis of inorganic and organic carbon); and (6) hard X-ray spectroscopy (for chemical speciation analysis of macro- and micronutrients and identifying the valence state of redox-active elements). Although powerful, many of these techniques need to be made more readily available to novice users of synchrotron radiation. Several of the most useful and effective techniques still are "tour de force" measurements, underscoring the need for increased availability of these tools and approaches. Similarly, because researchers must characterize numerous biological and environmental samples to obtain statistically significant results, further development of present synchrotron-based techniques is greatly needed to enable standardized, user-friendly, and high-throughput measurements.

Isotope Techniques

Isotope-based technologies are particularly promising for measuring carbon cycling processes and linking such processes to the organisms and metabolic pathways that catalyze them. Isotope ratios of organic matter and CO_2 also provide powerful tools for understanding and tracing carbon flux and storage, from cellular to global scales. Stable isotopes of a given element differ in the number of neutrons they contain. For example, about 99% of carbon on Earth is ^{12}C, which has 6 protons and 6 neutrons ($6 + 6 = 12$). However, the other 1% is ^{13}C, which contains 7 neutrons. Both are stable isotopes, meaning they do not decay. Stable carbon and oxygen isotope ratios (defined as $^{13}C{:}^{12}C$ and $^{18}O{:}^{16}O$, respectively, and represented as $\delta^{13}C$ and $\delta^{18}O$) have been used successfully for cellular forensics and tracing carbon flow at cellular, tissue, organismal, ecosystem, regional, and global scales.

Also valuable for carbon studies is the radioisotope ^{14}C, which can be used to quantify the residence time and age of carbon in organic matter. Known as radiocarbon, ^{14}C—with its radiodecay and use as an isotope tracer—also can provide information on the time scales of carbon exchange with the atmosphere. Relatively recent technological developments have dramatically improved the usefulness and cost-efficiency of isotopic analyses, making previously inconceivable experiments now possible. As such, a significant opportunity exists to use isotopic techniques to understand the carbon cycle at cellular to global scales and in frameworks relevant to various environments and biological processes.

Isotopic analyses are valuable to multiscale studies of carbon biosequestration and the carbon cycle because stable isotopes integrate—over a temporal period—significant processes leading to biomass or CO_2 formation. Isotopes indicate key mechanisms resulting in carbon storage and fluxes because these processes often fractionate, or change the isotopic ratio, between source and product. Furthermore, isotopes record these same processes in organic, inorganic, or gaseous forms, such as in cell walls and tree rings. Stable isotopes also can be used to trace origins of carbon fluxes and pools. These powerful, multifaceted uses of isotopes make them a critical tool for future carbon cycling and biosequestration research (West et al. 2006; Dawson and Siegwolf 2007).

The $\delta^{13}C$ in organic matter is driven in part by isotopic fractionation that occurs during photosynthesis, which strongly depends on water stress, biochemical CO_2-uptake capacity, and type of carbon-fixation pathway used [e.g., C_3 or C_4 (Ehleringer et al. 1993)]. The isotopic signature of carbohydrate resulting from photosynthesis is used for organic-matter production and metabolic respiration for all autotrophic and heterotophic organisms within an ecosystem, thus providing both an organic record and a CO_2 tracer of photosynthetic processes at large scales.

The $\delta^{18}O$ of organic matter and ecosystem-respired CO_2 is driven largely by the $\delta^{18}O$ of water in the major water pools of ecosystems, the canopy, and soil. The $\delta^{18}O$ in this ecosystem water is in turn controlled by the regionally unique signature of incoming precipitation and is modified primarily by evaporative enrichment during dry periods. As such, evolved $\delta^{18}O$ in CO_2 carries a tracer of the ecosystem of origin and of drought. Furthermore, oceanic exchange of CO_2 with the atmosphere has only very small isotopic fractionations, thus atmospheric CO_2 contains a signature of terrestrial carbon cycle processes. Models of terrestrial carbon flows therefore can be uniquely constrained by measurements either of atmospheric CO_2 or of that respired from ecosystems, as has been done over the past few decades using flask-sampling techniques.

Radiocarbon produced cosmogenically is a valuable time-dependent tracer of carbon cycle processes and carbon biosequestration on time scales of centuries and beyond because of its half-life of 5730 years. Since 1964, however, radiocarbon injected into the atmosphere by aboveground nuclear testing has provided a global isotope tracer for the carbon cycle. In the past two decades, development of accelerator mass spectrometry, which measures ^{14}C atoms individually rather than waiting for them to decay, has increased sample throughput and decreased sample size dramatically.

Isotopes provide the means—unavailable via other methods—to understand the carbon cycle at scales ranging from the genome to globe. Observational, manipu-

lative, and pulse-label experimental approaches using isotopic techniques will facilitate achieving the following advances.

- Determining how different carbon substrates contribute to biomolecule synthesis versus metabolic respiration (e.g., microbial efficiency).
- Tracing plant allocation and, in particular, relating changes in it to shifts in gene expression [e.g., using ^{11}C, ^{13}C, or ^{14}C to do so (Schwachtje et al. 2006; Carbone et al. 2007)].
- Using new technology to differentiate between autotrophic and heterotrophic respiration in soils (Trumbore 2006).
- Tracing carbon flow from plants to soils for a range of time scales [e.g., using ^{13}C pulse label for days to weeks (Bowling et al. 2002) or ^{14}C pulse label for years to decades]. For example, researchers could measure tracer concentrations in biomarkers to ascertain the timing of photosynthetic-product transfer to a microbial community.
- Using new tools such as nano-SIMS to allow isotope-tracer mapping at subcellular levels.
- Using natural-abundance ^{14}C to identify components of soil carbon that represent long-term stores and determining how this residence time responds to management strategies or climate change.
- Testing and constraining models of carbon cycling within microbial communities, soils, foliage, ecosystems, and ultimately the globe (Barbour et al. 2007; McDowell et al. 2008).
- Improving the deconvolution of global records of CO_2, $\delta^{13}C$, $\delta^{18}O$, and ^{14}C to assess the importance of particular regions as net carbon sources and sinks.

Although the value of isotopic measurements is widely accepted, such techniques historically have been time consuming and expensive, greatly limiting researchers' ability to capitalize on these powerful tools. Recent and continuing advances, however, are yielding breakthroughs in measurement frequency and cost, allowing previously unfeasible integration of isotopic measurements into field and laboratory experiments. For example, laser-based measurements of $\delta^{13}C$ and $\delta^{18}O$ in CO_2 have increased sampling frequency significantly—from once a week using field sampling and tedious laboratory analyses to once every couple of minutes using measurements taken directly in the field (Bowling et al. 2005; Griffis 2005; Barbour et al. 2007; McDowell et al. 2008). Furthermore, sampling improvements now allow measurements at the sub-Hertz time scale for ecosystem foliage and atmospheric fluxes. Proven applications of laser-based isotopic measurements include (1) assessing mesophyll conductance (Flexas et al. 2006), (2) testing photosynthetic models (Bickford et al., in review), (3) observing for the first time the $\delta^{13}C$ and $\delta^{18}O$ signature of light-enhanced dark respiration (Barbour et al. 2007), (4) examining the transient response of soil-respired $\delta^{13}C$ to precipitation (Powers et al., in review), (5) measuring ecosystem-scale partitioning of photosynthesis and respiration, (6) investigating ecosystem carbon cycle dependency on climate (Bowling et al. 2005), and (7) testing ecosystem carbon cycle models (McDowell et al. 2008).

Potential future applications of stable-isotope techniques are vast, offering tremendous opportunities for scientific creativity. Such laser-based systems are appropriate for cellular- to regional-scale monitoring, pulse-labeling experiments, and long-term observations. Moreover, within only the past year, new isotopic systems have emerged that are more portable and consume less energy and other resources, making these techniques further amenable to long-term field monitoring at remote locations. Additionally, accelerator mass spectrometry for ^{14}C analyses is becoming more economical, dramatically increasing the number of analyses that can be conducted to trace carbon flow and residence time in ecological systems. Smaller and easier-to-maintain accelerator mass spectrometers are being produced, including, for example, the new system at the University of California-Irvine, which is the first ^{14}C accelerator available exclusively for studying the carbon cycle. The significant increase in analytical capacity has accelerated carbon cycle research science and demonstrated that such facilities can be operated at lower per-sample costs. However, many facilities already have reached their analytical capacity because of existing needs (Trumbore 2006). Thus, extending the availability of such instrumentation is a critical requirement for better understanding the carbon cycle and devising strategies to enhance carbon biosequestration.

Organic-Matter Biogeochemistry and Analytics

A critical crosscutting need in carbon cycling and biosequestration science is characterization of natural organic matter, including its biochemical processing, changes in structure and physical biochemistry, and interactions with nonbiotic environmental factors. New, robust analytical and characterization methods are needed to measure compositional changes in organic material during its transport and degradation in both terrestrial and ocean environments. Challenges associated with such measurements in many ways are parallel to problems faced by scientists investigating biomass synthesis and deconstruction.

Previous characterization of organic matter in soils and oceans has been based largely on operational definitions, certain size fractions, and extraction with specific chemicals. Despite significant improvements in analytical methods for studying organic molecules (e.g., mass spectrometry, NMR, and synchrotron-based X-ray spectroscopy), few advanced technologies have been directed toward characterization of natural organic matter. Consequently, relatively little is known about the chemical composition and biotic and abiotic reactions involved in the biochemical degradation or alteration of organic matter. With their extensive infrastructure and capabilities in these innovative technologies (e.g., light sources and the Environmental Molecular Sciences Laboratory), DOE national laboratories and associated user facilities have the potential to fill in significant gaps in this knowledge.

A predictive understanding of the global carbon cycle requires linking theory, observations, experiments, and models because no single approach is sufficient. As noted, observations and experiments are profoundly valuable for informing researchers on how the carbon cycle works. However, implementation and interaction of empirical datasets from such studies must be done in conjunction with models so results can be integrated into a framework capable of forecasting future climate impacts on the carbon cycle.

Transport Between Reservoirs and Phases

Global climate change research communities have tended to focus their thinking and efforts within conceptual or geographical reservoirs distributed through the geochemosphere. Examples include land plants as a geochemical pool and processing unit; the soil ecosystem supporting them; the atmosphere as a medium of material transport from land-based continental systems to the ocean; and within the sea, the euphotic and twilight zones as well as central oceanic layers. The means of carbon transport among these subunits within the global system may determine rates of flow and introduce critical roadblocks. For example, the transition from litter to soil organic material is mediated in part by microfauna and links additionally to the hydrological cycle. Long-range transport of dust from terrestrial sources such as deserts carries iron to the remote ocean, but this nutrient becomes bioavailable only under appropriate hydrometeoric pH and photochemical conditions. Aerosol and cloud acidity in turn are determined by the flux of reduced sulfur and nitrogen from the ocean surface. Air-pollution sources also constitute critical modulators of such acidity in and of themselves. Furthermore, Asian economic growth is expected to trigger increases in acidity over the North Pacific even as North America and Europe cease pollution of the Atlantic atmosphere. The consequences of such increases remain underexplored but could result in significant feedbacks to carbon biosequestration strategies and climate change.