

**Name and brief description of initiative:**

**Department of Energy Genomics: GTL -- Systems Biology for Energy and Environment**

**Brief description of goals of initiative:**

GTL's goal is to reveal how the static information in genome sequences drives the intricate and dynamic processes of life. Through predictive models of these life processes and supporting research infrastructure, we seek to harness the capabilities of plants, microbes, and complex microbial communities, which are the foundation of the biosphere and sustain all life on earth. Gaining reliable use of plant and microbial processes requires understanding the whole living system, not just genomic DNA sequences or collections of proteins or cell by-products. GTL will study critical biological properties and processes on three systems levels— molecular, cellular, and community—each requiring advances in fundamental capabilities and concepts. GTL has the mission goal of tapping the powerful and diverse capabilities of microbes, microbial communities, and plants to provide breakthrough biotechnologies for renewable energy production, carbon sequestration, and environmental remediation.

**Principal investigator:** various, see

<http://doegenomestolife.org/pubs/2006abstracts/index.shtml>

**Program contact information:** John C. Houghton, Ph.D. 301-903-8288;

[John.Houghton@science.doe.gov](mailto:John.Houghton@science.doe.gov)

**Website address of initiative:** <http://doegenomestolife.org/>

**Brief description of informatics and computational biology components and their goals:**

Computation is essential to the GTL program goal of achieving a predictive understanding of plant and microbial cellular and community systems. The integrated GTL computational environment will link data with theory, modeling, simulation, and experimentation to derive principles and develop and test theory. GTL computation will employ data-intensive bioinformatics, compute-intensive molecular modeling, and complexity-dominated cellular systems modeling.

A comprehensive knowledgebase will be at the heart of GTL systems biology. The knowledgebase foundation is the DNA sequence code that will relate the many data sets emanating from plant and microbial systems biology research and discovery. Building over time to a detailed and annotated description of cellular functions, the GTL knowledgebase will assimilate a vast range of biological data as it is produced. It will grow to encompass program and facility data and information, metadata, experimental simulation results, and links to relevant external data and tools. Underlying the knowledgebase will be an array of databases, bioinformatics and analysis tools, modeling programs, and other transparent resources.

***Examples of core capabilities required by the overall GTL program***

**Bioinformatics: Collecting and Analyzing Data on Cellular Components.**

Sequence analysis, largely the prediction of genes and gene function by homology, has been a core task. GTL will generate many such data types as measurements of protein complexes, protein expression, and plant and microbial metabolic capabilities. Many new data sets must be correlated or annotated to genome data and archived to provide foundational data for computer models of biochemical pathways, entire cells, and, ultimately, plant and microbial communities and ecosystems.

**Molecular Measurements and Modeling: Revealing Processes Carried Out by Cellular Components.**

GTL seeks to understand fully the cell's biological machinery and its relationships with other cells and the environment. To reach this goal, investigators must know and be able to computationally model and test concepts in which cellular components interact directly with each other and with other molecules in a cell. They also must know how proteins dock structurally to form a complex, how the proteins of a complex interact dynamically to accomplish a biological function, as well as how these components are maintained and refreshed in response to changing conditions. For example, detailed characterization of protein complexes is the prerequisite for understanding the functions of molecules, cells, regulatory complexes, and networks as well as the interactions of cell surface proteins and complexes with the environment.

**Cell and Community Modeling: Coalescing the Cell's Components into a Whole-Systems Predictive Understanding.**

Biosystem models encapsulate our understanding of biology, and simulation is becoming a key tool to further understanding at the systems level. Through computational analysis of predictive mathematical models, we will understand how biological systems may be manipulated to solve problems, how plants and microbes regulate the expression of genes involved in environmental interactions, and how protein complexes are assembled to carry out important processes. Predictive models also will prove most useful in integrating and summarizing the vast amounts of data to be generated by the GTL program.

*Examples of Capabilities for an Integrated Computational Environment*

**Theory, Modeling, and Simulation Coupled to Experimentation of Complex Biological Systems:**

Build concepts and models of plant and microbial cells and communities that capture and extend our knowledge, based on a combination of experimental data types. Test and validate component models and use integrated models to understand mechanisms and explore new hypotheses or conditions to design new experimental campaigns.

**Sample and Experimental Tracking and Documentation – Laboratory**

**Information Systems (LIMS) and Workflow Management:** Provide systems for experiment design, sample specification, sample tracking and metadata recording,

workflow management, process optimization and documentation, QA, and sharing of such data across research centers and projects.

**Data Capture and Archiving:** Capture bulk data from many different measurements and instruments in large-scale data archives.

**Data Analysis and Reduction:** Provide analysis capabilities for systems biology data to enable insights, input, and parameters for systems models and simulations.

**Computing and Information Infrastructure:** Furnish hardware and software environments to support analysis, data storage, and modeling and simulation at the scales required in GTL.

**Community Access to Data and Resources:** Provide community access to data, models, simulations, and protocols for GTL. Allow users to query and visualize data, use models, run simulations, update and annotate community data, and combine community data and models with their local databases and models.

### ***Recent GTL Research Highlights***

- Progress is being made in data reduction and analysis for MS experiments, integration of databases containing heterogeneous data sets, and use of multiple approaches to metabolic and regulatory network modeling.
- The computational framework for comparative analysis of functional genomic data and computational models is being developed for data on the behavior of plant and microbial gene regulatory networks in response to environmental conditions.
- Whole-cell flux-balance models are being used to understand aspects of natural behavior and for comparative analysis of different plant species or microbial strains.
- Computational methods are being developed to predict the wiring diagrams of various response networks, which consist of signaling, regulatory, and metabolic components. These include carbon fixation, phosphorus- or nitrogen-assimilation, and electron-transfer networks. These methods use predictions of operons and regulons and interaction relationships among candidate genes whose proteins appear to be expressed together or coordinately.
- Computational models are being built to predict the activity of natural microbial communities for application of robust bioremediation technologies. Teams also are learning how to simulate growth and activity of metal-reducing organisms in their natural environments.
- Three institutes have been created to support the advancement of computational-biology research as an intellectual pursuit and provide innovative approaches to educating biologists as computational scientists. Using interdisciplinary teams of researchers drawn from the physical and life sciences, computational mathematics, and computer science, the institutes sponsor multidisciplinary scientific projects in which biological understanding is guided by computational modeling. They are training students to uncover biological mechanisms and pathways within microbial organisms through the use of computational biology and synergistic collaborations with experimental groups and will engage students in project-oriented research.

**Brief description of resources and tools available for sharing:**

The GTL program is committed to making available computational biological tools and data to the broader research community as they are developed.

**Brief description of selected integrative efforts:**

**Adoption of data standards:** The GTL program supports the efforts of the BioPAX: Biological Pathways Exchange effort to create a data exchange format for biological pathway data. <http://www.biopax.org/>

**Adoption of common software or workflows:** The GTL program helps support the Systems Biology Workbench, which is a software framework that allows heterogeneous application components-written in diverse programming languages and running on different platforms-to communicate and use each others' capabilities via a fast binary encoded-message system. <http://sbw.kgi.edu/research/sbwIntro.htm>

**Available inventories of resources and tools:** The DOE Joint Genome Institute provides integrated high-throughput sequencing and computational analysis to enable genomic-scale/systems-based scientific approaches to DOE-relevant challenges in energy and the environment. The sequencing and other data is readily available. Other resources, such as assembly and annotation software, are also available. <http://spider.jgi-psf.org/index.html>

**Opportunities for collaboration or synergy with the NCBCs:** Potential collaborations between the DOE GTL program and NCBCs include common interests in providing data standards, ontologies, and software tools for systems biology.

Prepared by John Houghton 06/30/2006