Computational Biology Discussion

Gary M. Johnson
Krell Institute

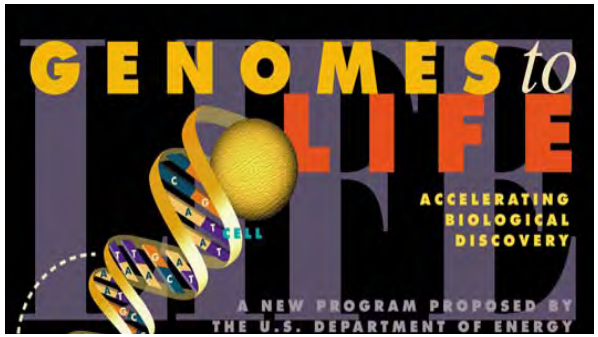
Prepared for:

Advanced Scientific Computing Advisory Committee Meeting
October 25 and 26, 2001
Crowne Plaza Hotel
14th and K Streets
Washington, DC

Outline of Discussion

1. Why are DOE, OBER and OASCR engaged in computational biology and systems biology research?
2. Specific research activities
3. Summary of GTL Program
4. Summary of FN 01-21 Awards
5. Agency funding levels
6. GTL program planning activities
7. Research opportunities in computational biology
8. Where should we go from here?

Systems Biology for Energy and Environment - Genomes to Life

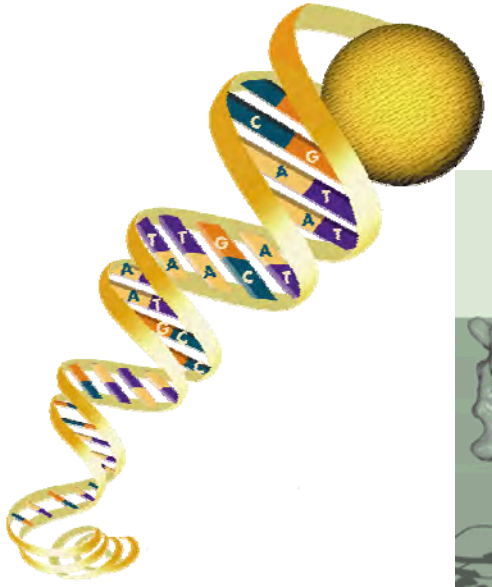


Systems biology is

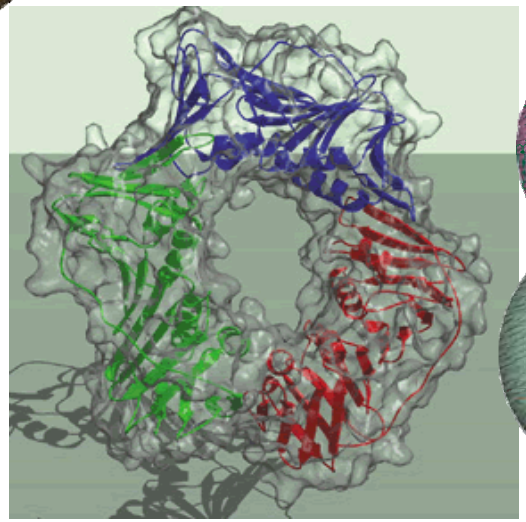
- A systems analysis and engineering approach to biology to understand the workings of entire biological systems
- It requires the integrated application of methods from modern biology, computational science, and information science and technology
- It requires advanced measurement and analytical technologies

Systems biology provides biological solutions to DOE problems through understanding biological systems…

from the genome



to the proteome



to the cell and organism and microbial communities

The bridge between physical, computational and life sciences
Enabling scientific breakthroughs impacting DOE missions

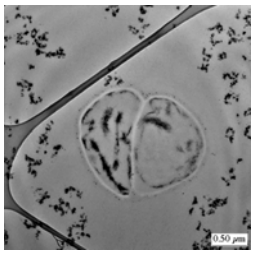
Why Systems Biology and DOE?

- Only a systems approach can lead to biological solutions for complex energy and environmental problems
- DOE is the only agency that can integrate the physical, computational and biological science expertise at a large scale and scope required for successful systems biology solutions to energy-related problems

Payoffs in the near term

Significant savings in toxic waste cleanup and disposal

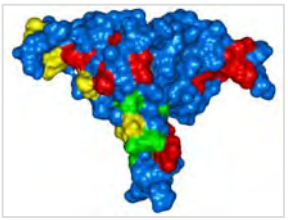
Bioremediation methods for accelerated and less costly cleanup strategies



Understanding metabolic pathways and mechanisms of native microbes

Improve the scientific basis for worker health and safety

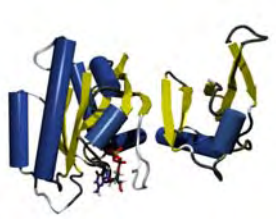
Improved diagnostics and standards for ecological and human health



Understanding responses of metabolic and regulatory pathways of organisms to environmental conditions

Technologies and systems for detecting and responding to biological terrorism

Sensors for detecting pathogens and toxins; strategies to enable strain identification; and improved vaccines and therapeutics for combating infectious disease

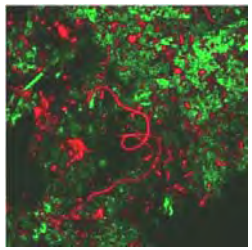


Investigating protein expression patterns, protein-protein interactions, and molecular machines

Payoffs in the mid to long term

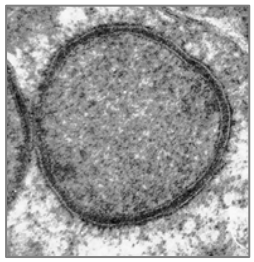
Enable independence of foreign oil

Clean, efficient biological alternative to fossil fuels



Harnessing metabolic pathways/mechanisms in H_2-producing microbes

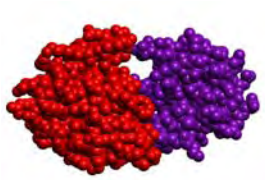
Designer plants for easily convertible biomass for fuels, chemical feed stocks, products



Understanding metabolic pathways and networks, and cell wall synthesis

Stabilize atmospheric carbon dioxide to counter global warming

Strategies and methods for storing and monitoring carbon



Investigating enzymes, regulation, environmental cues, and effects

Specific research activities

- Joint OBER-OASCR program on Genomes to Life
- Joint OASCR-OBER project on Advanced Modeling and Simulation of Biological Systems
  - Office of Science Notice 01-21
- OBER Microbial Cell Project
  - Office of Science Notice 01-20

GTL Scientific Plan —*To understand how genes, proteins, and cells work in intricate networks to form dynamic living systems exquisitely responsive to their environments.*

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

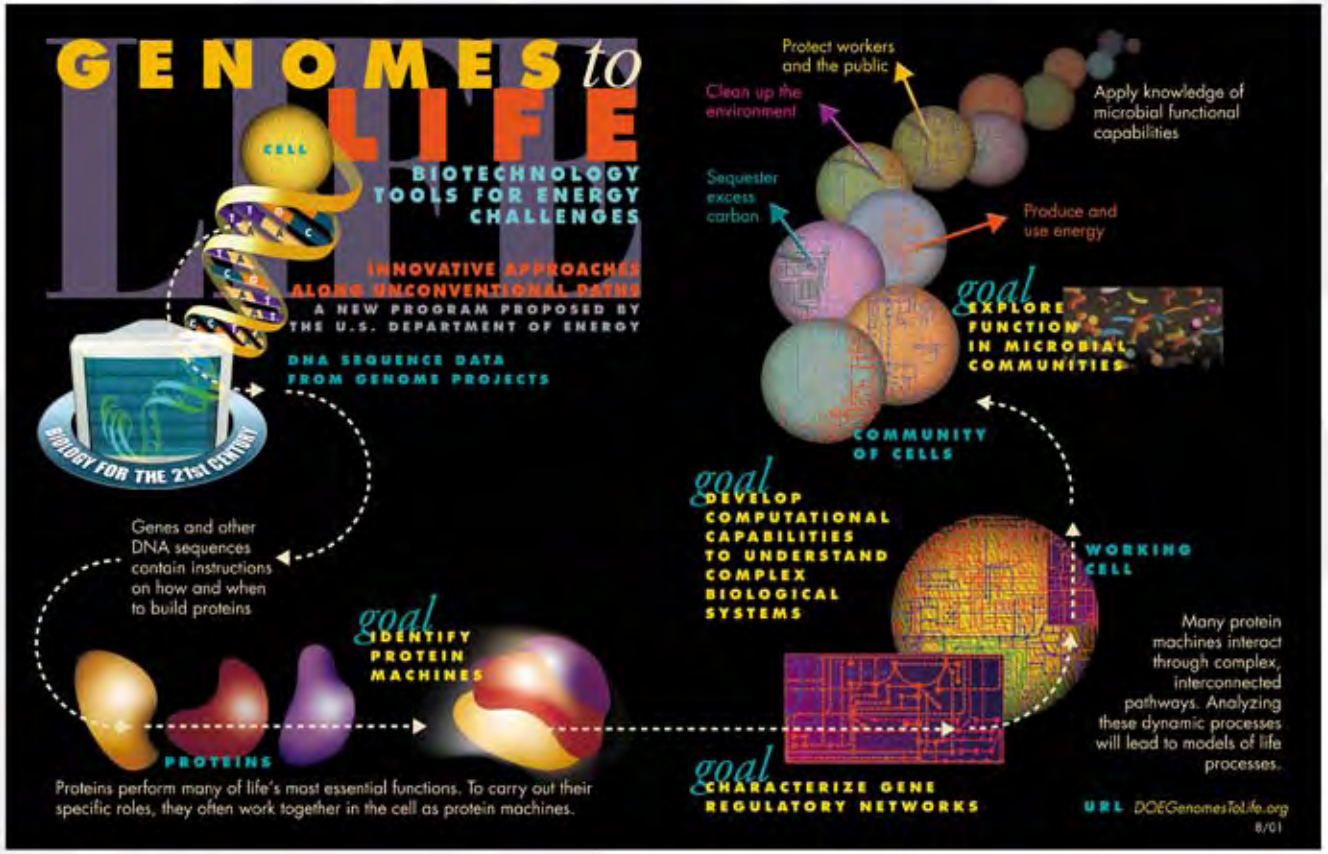
Cells contain DNA—the hereditary material of all living systems.

A genome is an organism's complete set of DNA.

DNA contains genes, whose sequence specifies how and when to build proteins.

Proteins perform most essential life functions, often working together in the cell as "protein machines."

Supercomputers will analyze how protein machines interact through complex, interconnected pathways. Computer models of these life processes will be applied to help solve energy challenges.

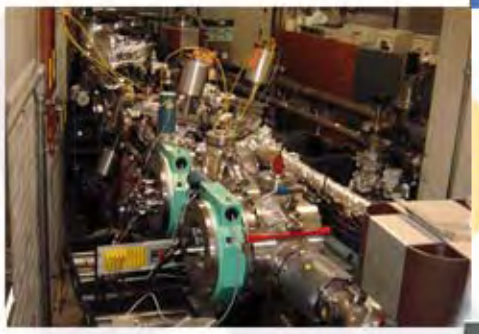


YGG-01-0397

DOE Cutting-Edge Facilities for Multidisciplinary Research

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Production Sequencing Facility at DOE's Joint Genome Institute



Beamlines at the National Synchrotron Light Source at Brookhaven National Laboratory and Stanford University's Linear Accelerator

Neutron sources at the High Flux Isotope Reactor at Oak Ridge National Laboratory and Los Alamos Science Center at Los Alamos National Laboratory. Under construction, the Spallation Neutron Source (site plan at left) at ORNL in collaboration with five other national laboratories.

Advanced Photon Source at Argonne National Laboratory

Environmental Molecular Sciences Laboratory's 800-MHz nuclear magnetic resonance spectrometer at Pacific Northwest National Laboratory

Advanced Light Source at Lawrence Berkeley National Laboratory

Supercomputers at six national laboratories

YGG-01-0417

Supercomputers Will Decipher How Genes Work—*This knowledge will aid development of new applications to solve energy and environmental challenges.*

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Living systems are complex and not well understood.

Computer simulations and models have been used to understand many complex systems, such as nuclear reactions and global climate. DOE has much experience in fielding problems of this computational magnitude.

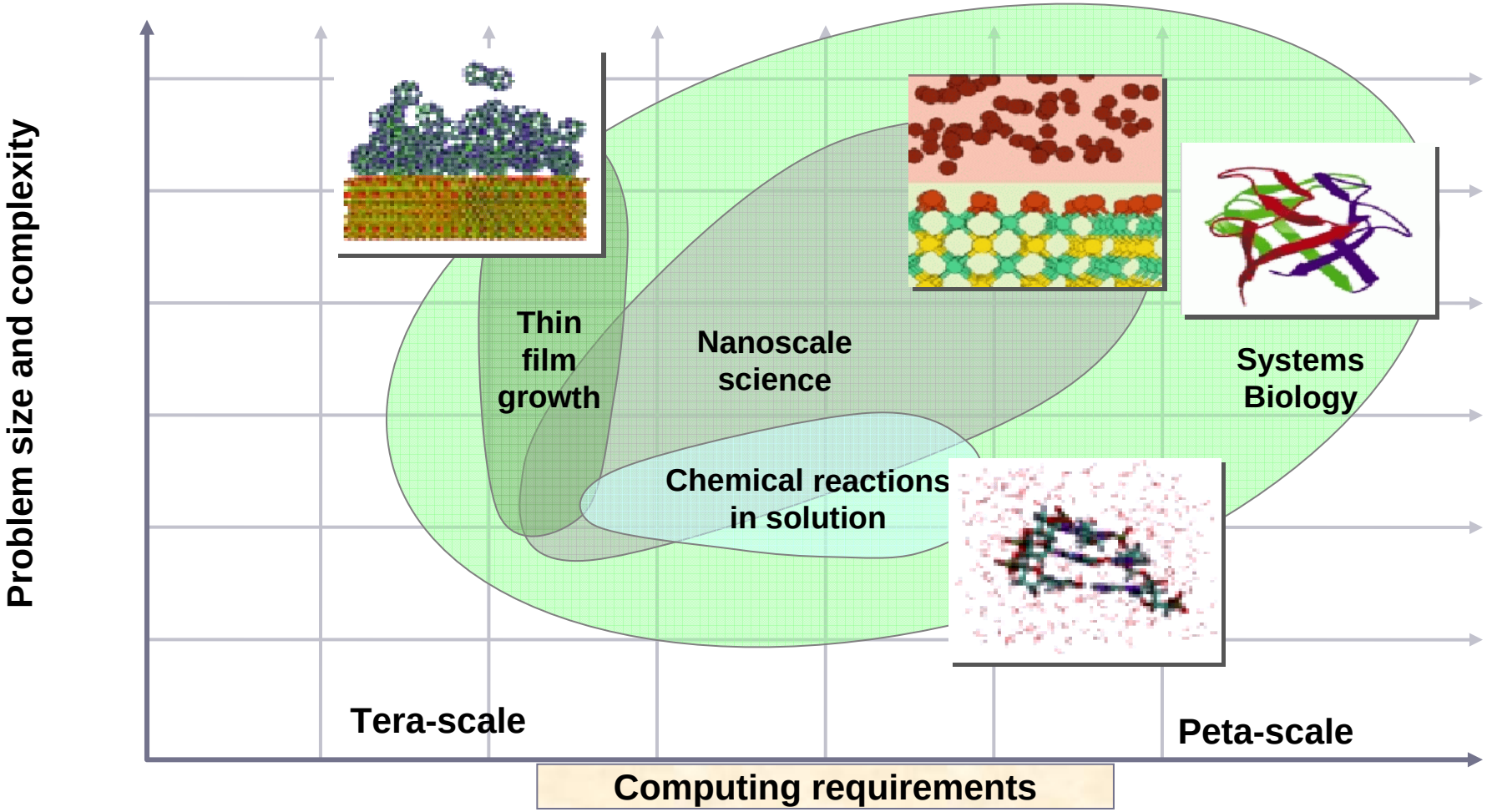
BER and ASCR of the Office of Science have formed a strategic alliance in GTL to develop the computational and mathematical capabilities to model living systems on the scale and complexity of living organisms.

DOE will discover how microbial genes, proteins, and microbial communities work together and will apply that knowledge to develop tools to solve energy and environmental challenges.

Biological research problems will drive computer science for the coming decades.

YGG-01-0399

Systems Biology depends on high-performance computing



Office of Science
Notice 01-21
Advanced Modeling and Simulation of Biological Systems

The goal of this program is to enable the use of terascale computers to explore fundamental biological processes and predict the behavior of a broad range of protein interactions and molecular pathways in prokaryotic microbes of importance to DOE.

FN 01-21 Awards

- 19 proposals received
- Proposals in areas of protein folding/docking and cell modeling
- 9+1 awards made
- First year awards totaled about $3M

Computational Biology Portfolio

FN 01-21 Projects:

ID	Institutions	Title	Total Funds Requested	2001 Funding
83078	Scripps Research Institute	Biomolecular Simulation Using Amber and CHARMM	$673,464	$216,572
Project Description				
Build on the existing CHARMM and Amber simulation packages, adapting them in novel ways to massively parallel architectures and high-performance CPUs.				
83086	Indiana University	Cyber Cell: Automated Physico-Chemical Cell Model Development Through Information Technology	$830,102	$268,340
Project Description				
Integrate the comprehensive reaction-transport-genetic cell simulator, Cyber-Cell, with experimental data, resulting in an automated model development methodology. The model will be developed and tested using data on *E. coli*.				
83088	Columbia University	Computational Analysis and Simulation of Bacterial Molecular Networks	$1,461,968	$240,000
Project Description				
The goal of the proposed research is to enhance our ability to predict and control bacterial phenotypic behavior by proper manipulation of genomic information and environmental stimuli. Such bacterial phenotypic behavior can be steered towards DOE-releva				
83089	University of Notre Dame	Organization of Complex Metabolic Networks	$1,436,747	$331,509
	Northwestern			
Project Description				
The purpose of this project is to develop semi-quantitative models that capture the structure and function of the E. coli metabolism. The investigators plan to complement the purely topologic, pathway based methodologies with dynamical information quanti				
83090	University of California, San Diego	Parallel Protein Docking and Interaction Dynamics with Adaptive Meash Solutions to the Poisson-Boltzmann Equation	$1,900,200	$348,792
	Scripps Research Institute			
Project Description				
This project involves the improvement of tools for determination of the structures of protein complexes through docking with an energy function of high quality. Particular emphasis is given to electrostatic interactions, and much of the work involves need				
83125	LLNL	Advanced Molecular Simulations of E. Coli Polymerase III	$1,781,369	$446,612
Project Description				
The project involves the use of advanced molecular simulation methods on terascale computers to improve understanding of bacterial multicomponent protein machines. The research will involve performing dynamical simulations of *E. coli* DNA polymerase III b				

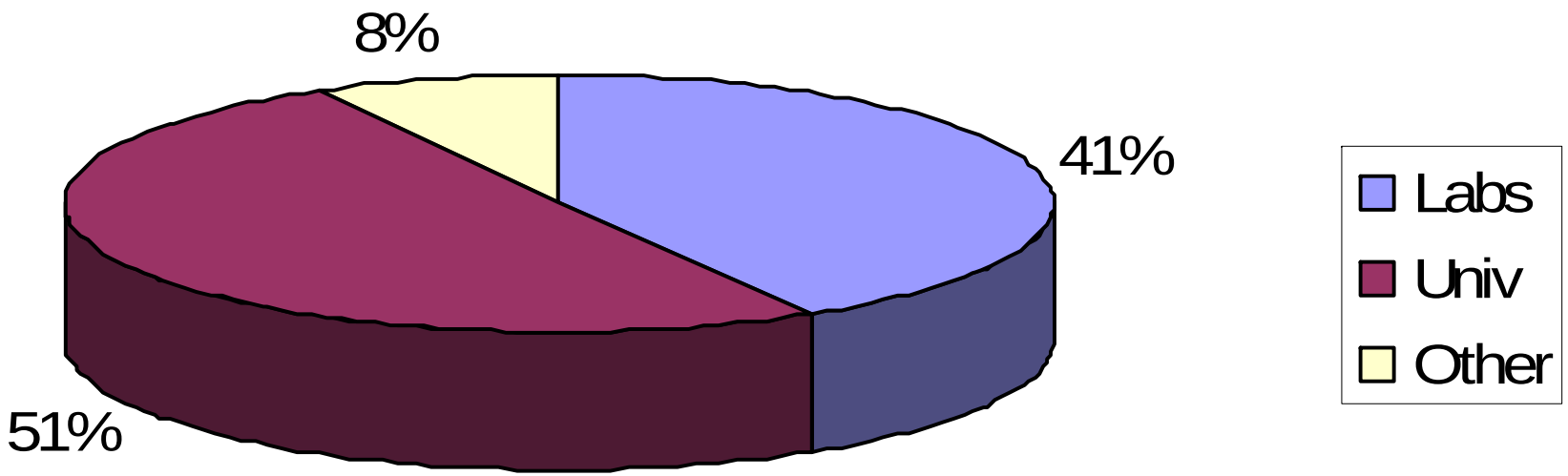
FN 01-21 Projects Continued:

ID	Institutions	Title	Total Funds Requested	2001 Funding
83136	PNNL	Computational Approaches and Framework for Microbial Cell Simulations	$1,470,000	$360,000
Project Description				
The investigators propose to develop a wide-ranging set of computational tools in support of intracellular model building. These tools will be applied to build computable representations of the core energy and material pathways in Rhodobacter sphaeroides				
83137	PNNL	Molecular Modeling of Complex Enzymatic reactions: The Respiratory Enzyme Flavocytochrome c_3 Fumarate Reductase of *Shewanella frigidimarina*	$952,000	$313,000
Project Description				
This project involves highly detailed simulations of complex reaction mechanisms in bacterial enzymes such as a flavocytochrome. It involves ongoing development of a program that permits such simulations, and specific application to the study of metal ion				
83095	Genomatica, Inc.	Development of the Next Generation of Genome-scale Constraints-Based Cellular Models	$2,204,360	$0
	Penn State University			$187,000
Project Description				
The proposed research will extend the PI's work on a top-down approach to metabolic modeling, which begins with a stoichiometric network model and successively constrains the set of admissible solutions with conditions that are derived, for instance, from				

FN 01-20 Projects:

ID	Institutions	Title	Total Funds Requested	2001 Funding
83108	Institute for Systems Biology	Interdisciplinary Study of Shewanella putrefaciens MR-1's Metabolism & Metal Reduction	$4,498,512	$100,000
Project Description				
The project is an integrated systems approach to study S. putrefaciens MR-1 in an attempt to delineate the organisms response to environmental perturbation. MR-1 is a suitable organism for this study because it can function in aerobic conditions, reduces				
		Grand Totals:	$17,208,722	$2,811,825

FN 01-21 Distribution of Funds in 2001



Office of Science
Notice 01-20
Microbial Cell Project

The MCP is focused on fundamental research to understand those reactions, pathways, and regulatory networks that are involved in environmental processes of relevance to the DOE, specifically the bioremediation of metals and radionuclides, cellulose degradation, carbon sequestration, and the production, conversion, or conservation of energy (e.g. fuels, chemicals, and chemical feedstocks).

Agency Funding Levels

Agency	Funding Level	Focus Areas
NIH	$50M to $100M	Human Health
NSF	$48M	Human, Animal, & Plant Science
DARPA	$15M to $18M	Applications of Biotechnology to Defense
DOE		
OBER	$9M	Systems Biology, GTL
OASCR	$3M	Systems Biology, GTL
USDA	$3M	Food Crops

GTL Program Planning Activities

August 2001 Workshop

Computational Biology Workshop for the Genomes to Life Program

Organizers Mike Colvin, LLNL & Reinhold Mann, ORNL

Report: http//www.doegenomestolife.org/compbio/draft/index.html

username gtl

password workshop

September 2001 Workshop

Computational and Systems Biology: Visions for the Future

Organizer Eric Lander, MIT

Report pending

GTL Program Planning Activities

Future Workshops:

January 2002

- Computational Infrastructure for the Genomes to Life Program

February 2002

- Computer Science for the Genomes to Life Program

March 2002

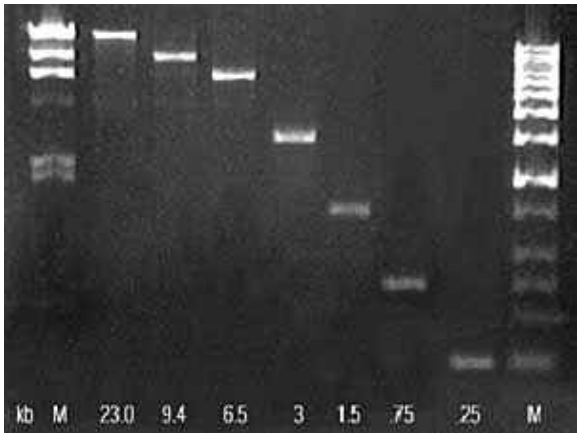
- Mathematics for the Genomes to Life Program

Research Opportunities in Computational Biology

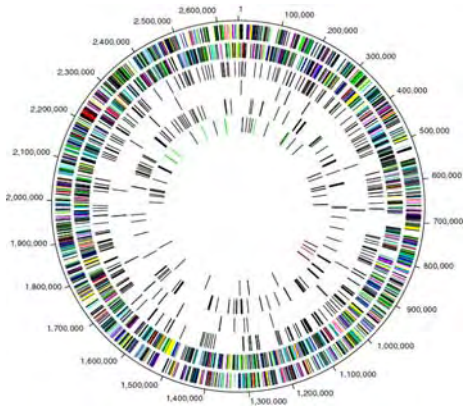
- Methods to model and simulate biological networks and pathways
- Methods to support the study of proteins, protein complexes, protein-protein interactions
- Methods to link models of biological processes and systems at various temporal and spatial levels of resolution
- Data management, access and analysis specifically focused on diverse data sets generated by modern biology experiments
- Tera-, peta-scale tool kits to support computational biology, e.g., pattern recognition algorithms, data mining, optimization, discrete math, multi-spectral image analysis, etc.

Biology is undergoing a major transformation that will be enabled and ultimately driven by computations

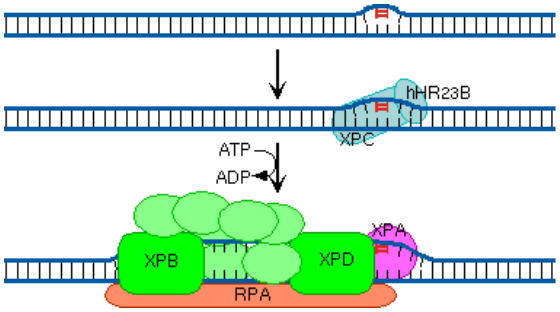
Data poor



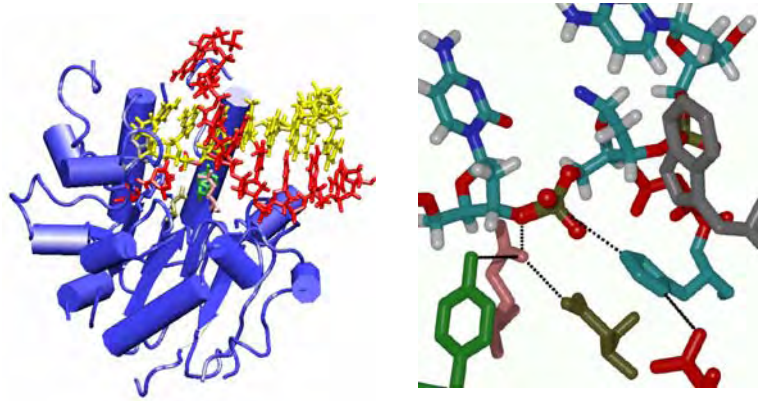
Data rich



Qualitative



Quantitative & predictive

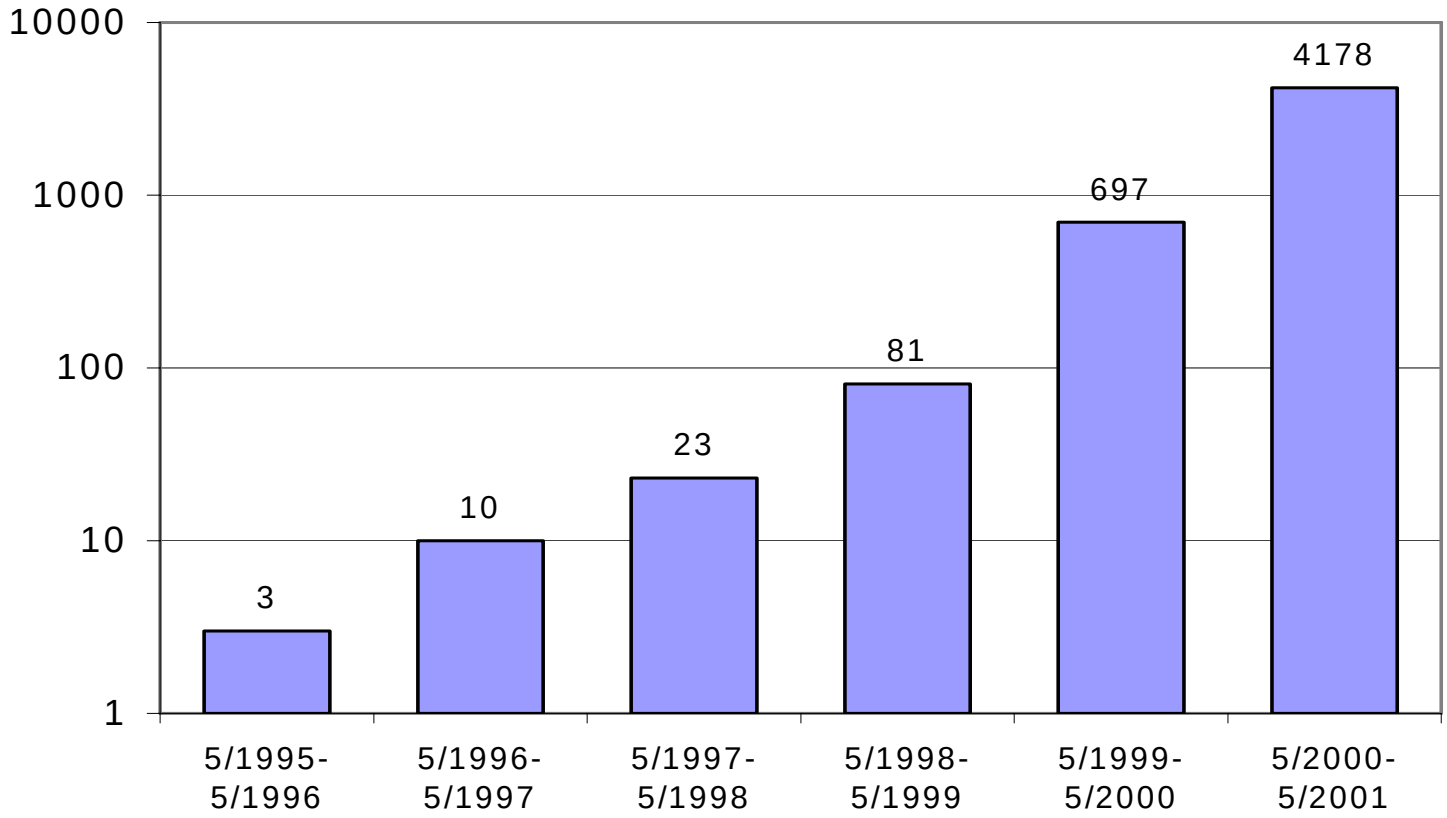


"It's time for biologists to graduate from cartoons to a real understanding of each protein machine ."

– Bruce Alberts, 9/6/01 (paraphrased)

Simulation and modeling are rapidly emerging as ways to explain biological data and phenomena

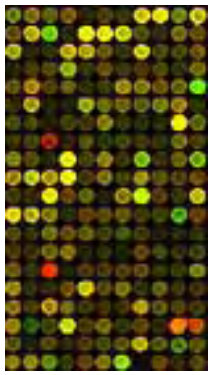
PubMed citations including “simulation” or “modeling” in title or abstract:

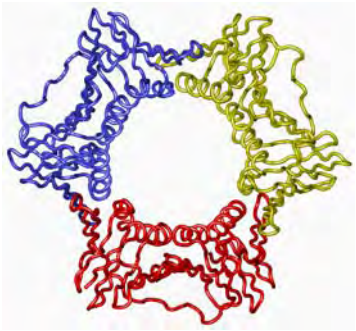


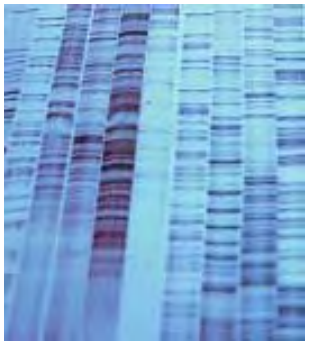
However, the field is still awaiting a major biological breakthrough achieved by supercomputer simulations

What capabilities are needed to be a leader in the emerging field of systems biology?

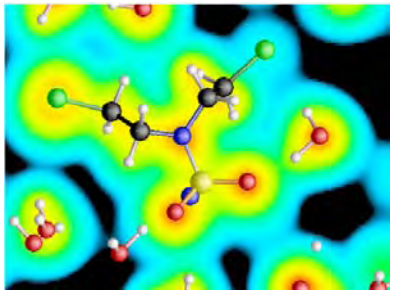
Strong experimental biology program







Theory and simulation



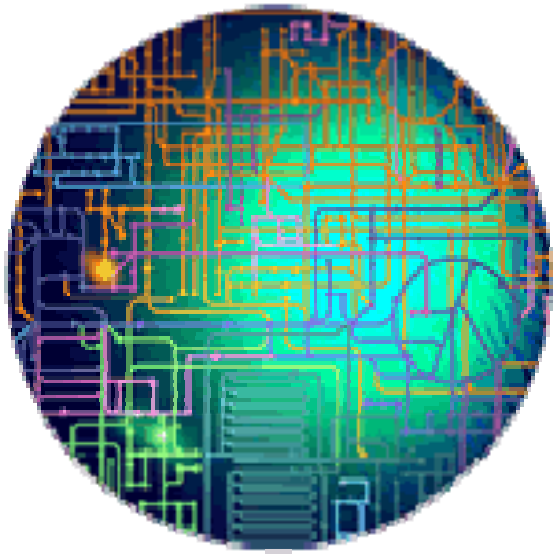
$$\left[\sum_i \left(\frac{-m_i}{2}\nabla_i^2 + \sum_{j\neq i}\frac{q_i q_j}{r_{ij}}\right)\right]\Psi = E\Psi$$

$$k\left[\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right] = \rho c_p \frac{\partial T}{\partial t}$$

High performance computing

Where should we go from here?

- Plan R&D agenda with components in:
  - Mathematics and statistics
  - Computer science
  - Informatics
  - Hardware and networking infrastructure
- Focus it on DOE mission opportunities to:
  - Use biological data to enable scientific discovery
  - Determine the structural details of biological “parts”
  - Model whole cells and microbial communities



Report on the Computational Biology Workshop for the Genomes to Life Program Summary of Recommendations

Modeling of Cells and Microbial Communities

- DOE should support a program of research aimed at accelerating the development of high-fidelity models and simulations of metabolic pathways, regulatory networks, and whole-cell functions.

Biomolecular Simulations

- DOE should ensure that advanced simulation methodologies and petaflop computing capabilities be available when needed to support full-scale modeling and simulations of pathways, networks, cells, and microbial communities.
- DOE should provide a software environment and infrastructure that allow for integration of models at several spatial and temporal scales.

Report on the Computational Biology Workshop for the Genomes to Life Program Summary of Recommendations

Functional Annotation of Genomes:

- DOE should support the continued development of automated methods for the structural and functional annotations of whole genomes, including research into such new approaches as evolutionary methods to analyze structure/function relationships.

Experimental Data Analysis and Model Validation:

- DOE should develop the methodology necessary for seamless integration of distributed computational and data resources, linking both experiment and simulation.
- DOE should take steps to ensure that high-quality, complete data sets are available to validate models of metabolic pathways, regulatory networks, and whole-cell functions.

Report on the Computational Biology Workshop for the Genomes to Life Program Summary of Recommendations

Biological Data Management:

- DOE should support the development of software technologies to manage heterogeneous and distributed biological data sets, and the associated data-mining and -visualization methods.
- DOE should provide the biological data storage infrastructure and the multiteraflop-scale computing to ensure timely data updates and interactive problem-solving.
- DOE should set a standard for open data in its GTL program and demonstrate its value through required universal use.

Report on the
Computational Biology Workshop
for the
Genomes to Life Program
Summary of Recommendations

General Recommendations:

- Continue the development of the GTL computational biology plan through a series of workshops focused on informatics, mathematics, and computer science challenges posed by the GTL systems biology goals;
- Ensure that the computing, networking, and data storage environment necessary to support the accomplishment of GTL goals will be available when needed. This environment should include computing capabilities scaling up through the multiteraflop and into the petaflop range; as well as a storage infrastructure at the multipetabyte level; and a networking infrastructure that will facilitate access to heterogeneous distributed biological data sets by a geographically dispersed collection of investigators. Further definition of this environment should be pursued through a dedicated workshop;

Report on the Computational Biology Workshop for the Genomes to Life Program Summary of Recommendations

General Recommendations:

- Establish policies for distribution and ownership of any data generated under the GTL program, prior to commencing peer review of GTL proposals or making any awards that would lead to the creation of such data; and
- Support sufficient scope of research to assemble the cross-disciplinary teams of biologists, computational biologists, mathematicians, and computational scientists that will be necessary for the success of GTL.