

# Computational Biology Discussion

Gary M. Johnson  
Krell Institute

*Prepared for:*

Advanced Scientific Computing Advisory Committee Meeting

October 25 and 26, 2001

Crowne Plaza Hotel

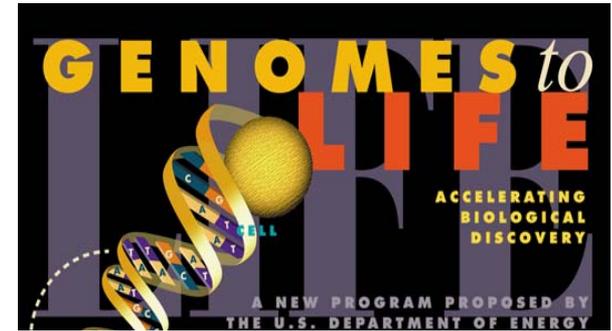
14th and K Streets

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

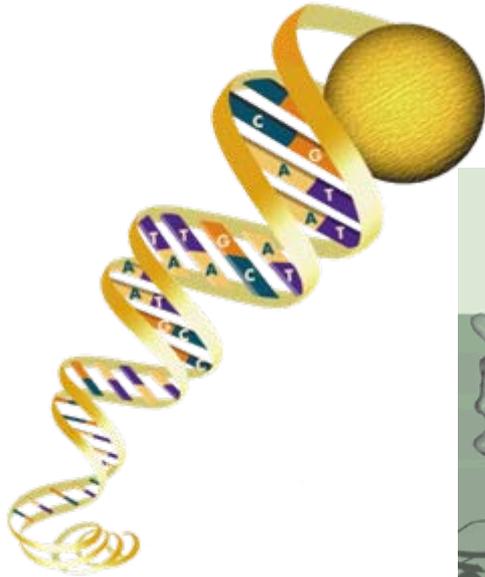


Systems biology is

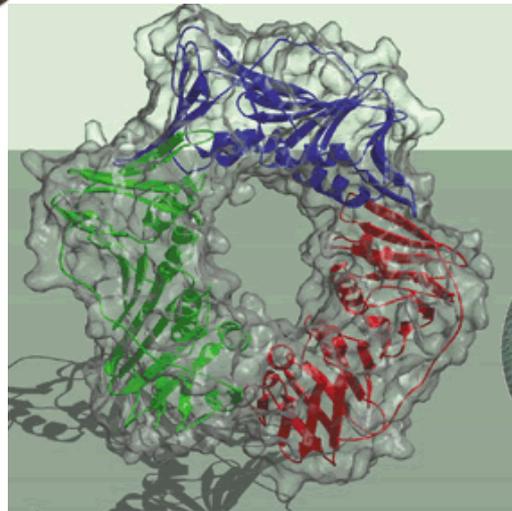
- A systems analysis and engineering approach to biology to understand the workings of entire biological systems
- It requires the integrated application of methods from modern biology, computational science, and information science and technology
- It requires advanced measurement and analytical technologies

# Systems biology provides biological solutions to DOE problems through understanding biological systems...

*from the genome*



*to the proteome*

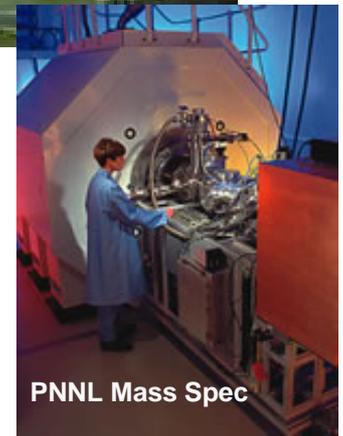


*to the  
cell and  
organism  
and  
microbial  
communities*

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

# Why Systems Biology and DOE?

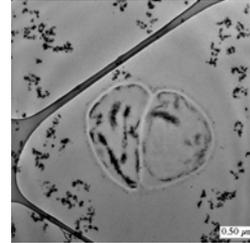
- Only a systems approach can lead to biological solutions for complex energy and environmental problems
- DOE is the only agency that can integrate the physical, computational and biological science expertise at a large scale and scope required for successful systems biology solutions to energy-related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



# Payoffs in the near term

**Significant savings in toxic waste cleanup and disposal**

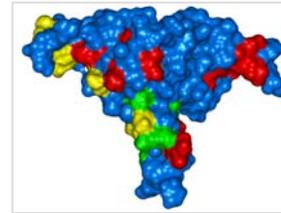
Bioremediation methods for accelerated and less costly cleanup strategies



Understanding metabolic pathways and mechanisms of native microbes

**Improve the scientific basis for worker health and safety**

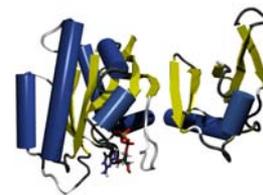
Improved diagnostics and standards for ecological and human health



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



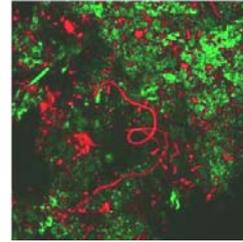
Investigating protein expression patterns, protein-protein interactions, and molecular machines



# Payoffs in the mid to long term

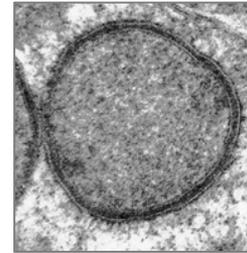
**Enable  
independence  
of foreign oil**

Clean, efficient  
biological  
alternative to  
fossil fuels



Harnessing metabolic  
pathways/mechanism  
s in H<sub>2</sub>-producing  
microbes

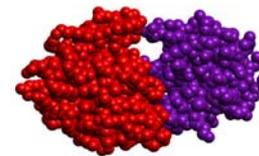
Designer plants for easily  
convertible biomass for  
fuels, chemical feed  
stocks, products



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



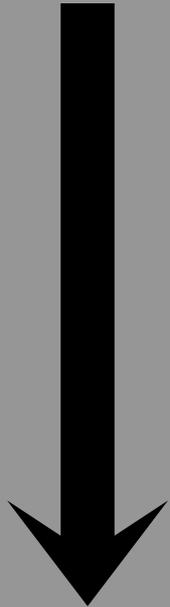
Investigating  
enzymes, regulation,  
environmental cues,  
and effects

# Specific research activities

- Joint OBER-OASCR program on **Genomes to Life**
  - Now known as: "Microbes for Energy and the Environment"
- Joint OASCR-OBER project on **Advanced Modeling and Simulation of Biological Systems**
  - Office of Science Notice 01-21
- OBER-OBES-OASCR **Microbial Cell Project**
  - Office of Science Notice 01-20

# Genome Development

**Information**



**Knowledge**

**Genome  
Sequence**



**Understand  
Genes**



**Understand  
Proteins**



**Understand  
Basis of Life**

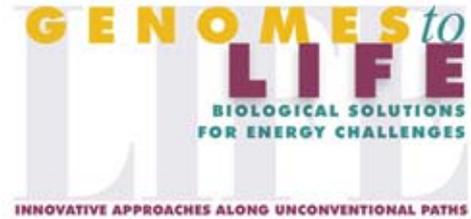
**ahmtlnikhteerorelh**

**thilnk ehtrre foral ma**

**lthi nkthe refore I am**

**I think therefore I am**

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



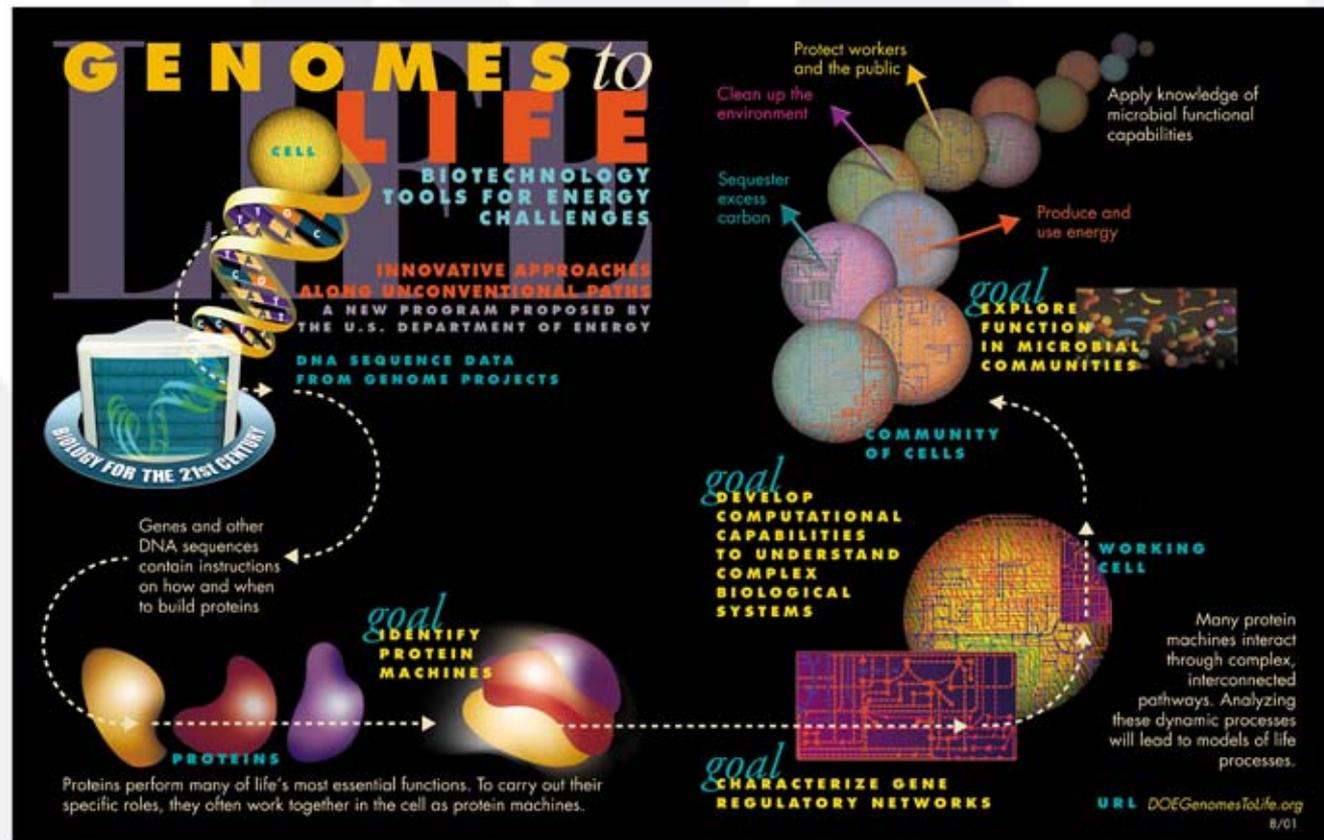
*Cells* contain DNA—the hereditary material of all living systems.

*A genome* is an organism’s complete set of DNA.

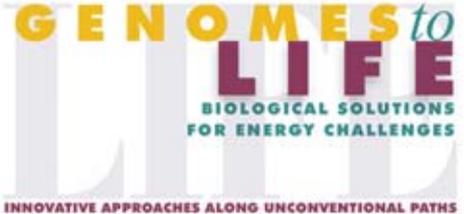
*DNA* contains genes, whose sequence specifies how and when to build proteins.

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

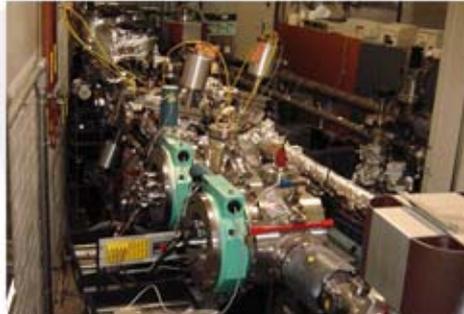
*Supercomputers* will analyze how protein machines interact through complex, interconnected pathways. Computer models of these life processes will be applied to help solve energy challenges.



# DOE Cutting-Edge Facilities for Multidisciplinary Research



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 Light Source at Lawrence Berkeley National Laboratory



Advanced Photon Source at Argonne National Laboratory

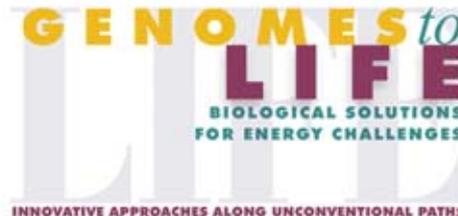


Environmental Molecular Sciences Laboratory's 800-MHz nuclear magnetic resonance spectrometer at Pacific Northwest National Laboratory

Supercomputers at six national laboratories



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



*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.

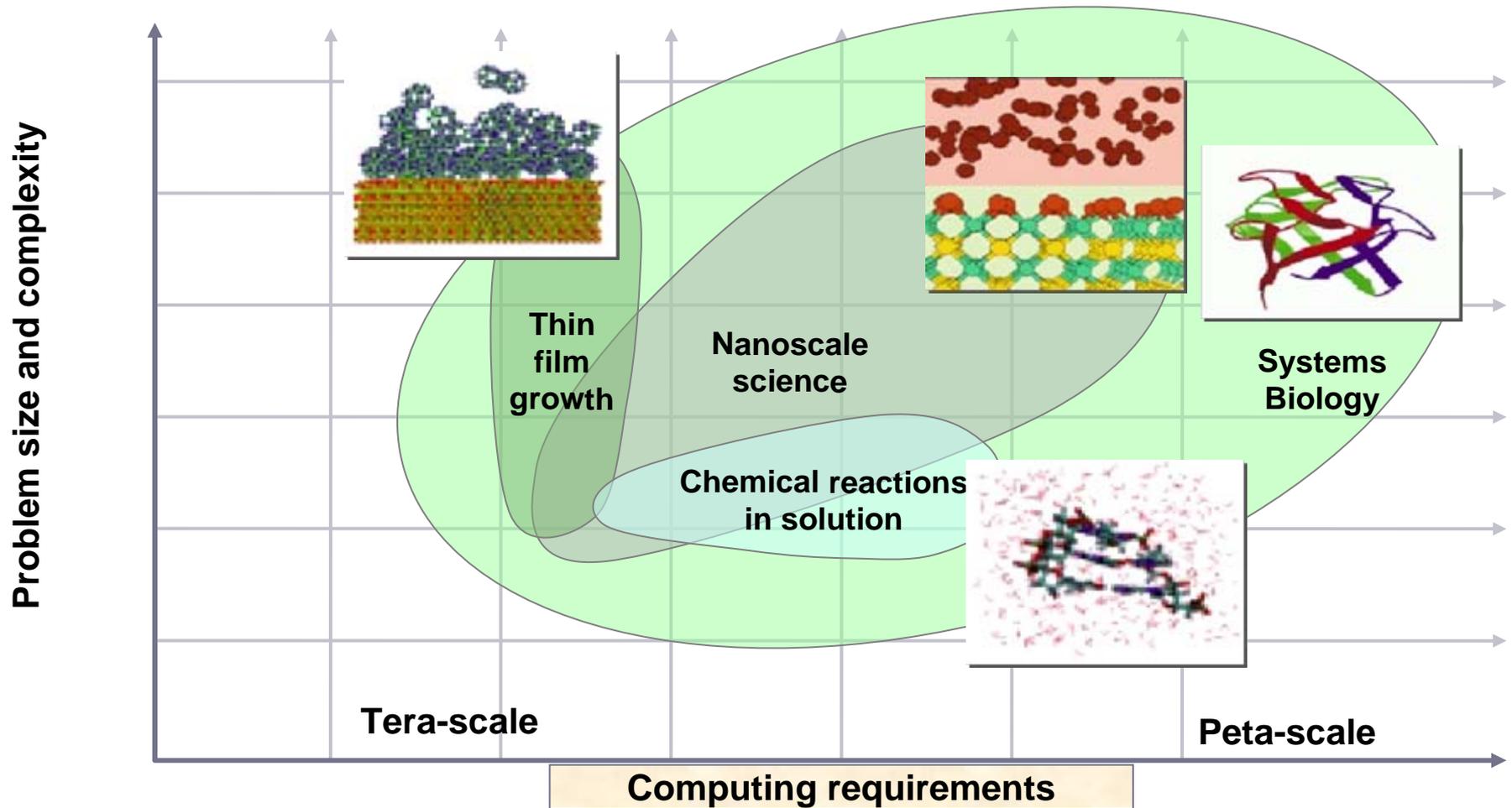
*BER and ASCR* of the Office of Science have formed a strategic alliance in GTL to develop the computational and mathematical capabilities to model living systems on the scale and complexity of living organisms.

*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.*

# 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

# 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).

# Biosciences Funding levels

Agency	Funding Level	Focus Areas
NIH	\$20B	Human health and supporting basic biosciences
NSF	\$500M	Biological infrastructure Environmental microbiology Integrative bio & neuroscience Molecular and cell biology Plant Genomics
DARPA	\$15M to \$18M	BioSPICE modeling project Bioinspired/biohybrid devices BW Detection and defence Military medical and trauma care
DOE	(Computational Biology)	
OBER	\$17M	Genome informatics Protein structure analysis Cell and pathway modeling Sloan Fellowships in Comp. Bio.
OASCR	\$3M	HPC-based biological applications Algorithms for biological modeling
USDA	\$3M	Plant/animal genome informatics

# Computational Biology 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	\$17M	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

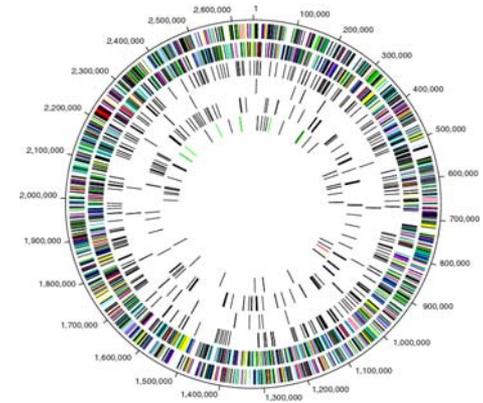
- Methods to model and simulate biological networks and pathways
- Methods to support the study of proteins, protein complexes, protein-protein interactions
- Methods to link models of biological processes and systems at various temporal and spatial levels of resolution
- Data management, access and 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

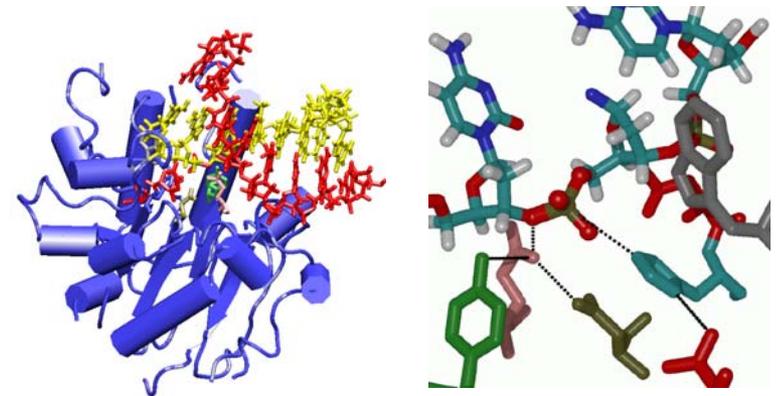
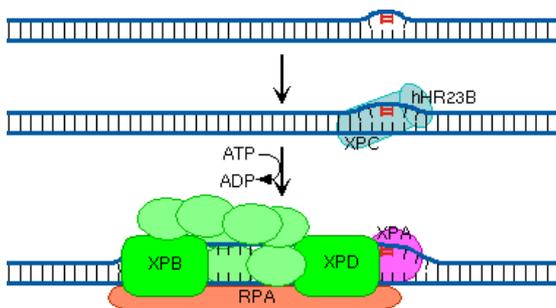


Data rich



Quantitative & predictive

Qualitative

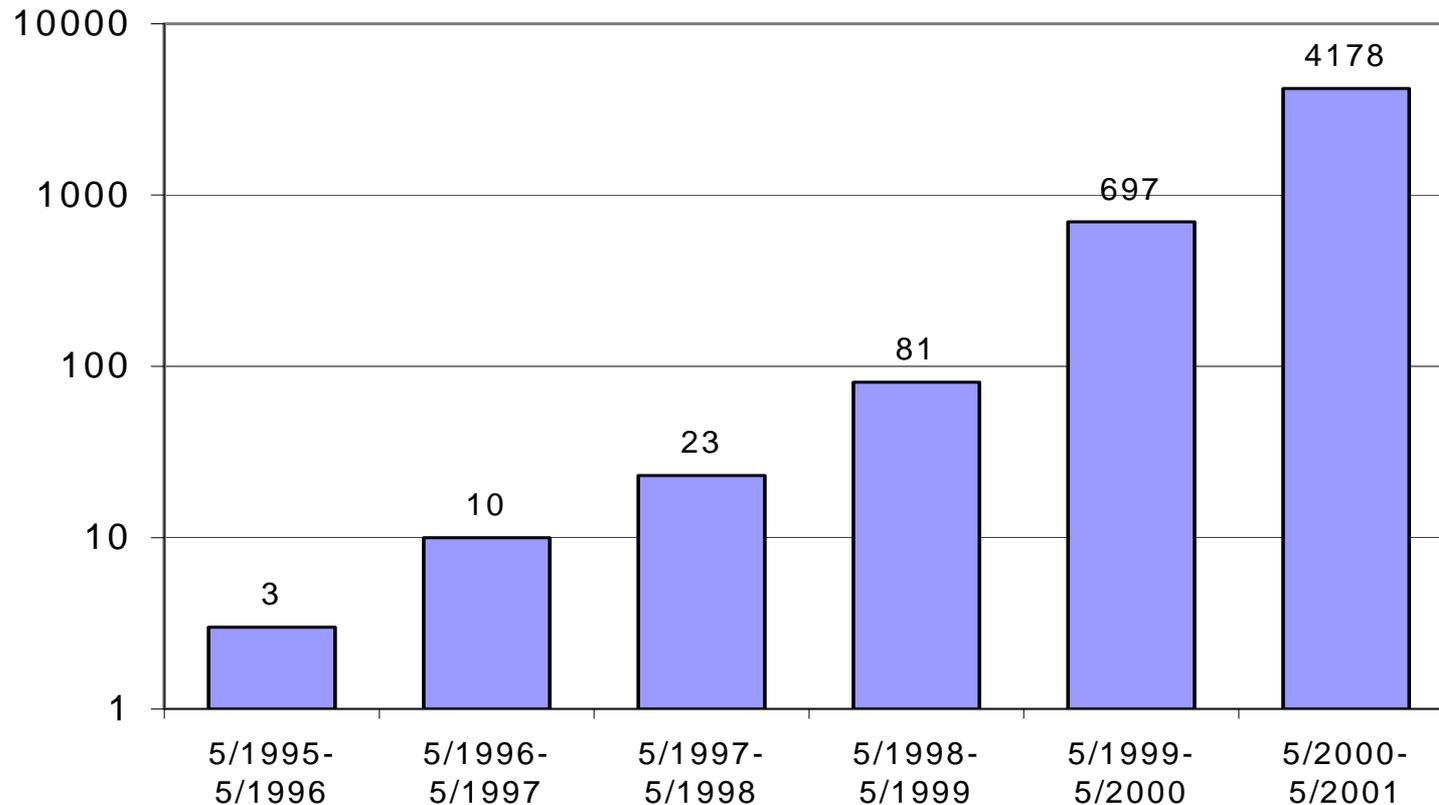


“It’s time for biologists to graduate from cartoons to a real understanding of each protein machine .”

– Bruce Alberts, President, National Academy of Sciences, 9/6/01

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

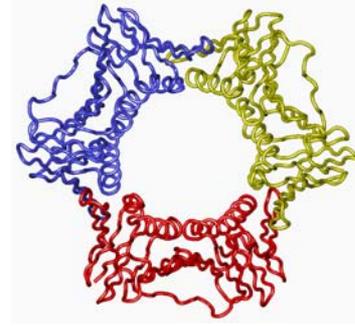
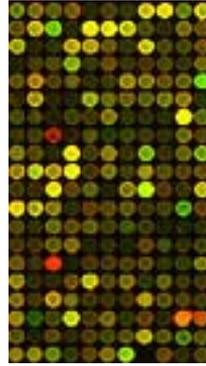
PubMed citations including “simulation” or “modeling” in title or abstract:



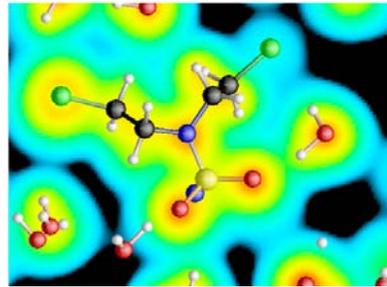
However, the field is still awaiting a major biological breakthrough achieved by supercomputer simulations

# 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*.