The Department of Energy: Genomics:GTL

NABIR PI Meeting

Daniel Drell, Ph.D.

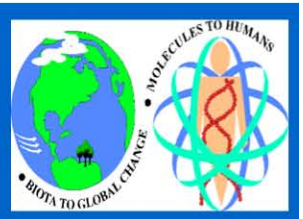
Life Sciences Division
Biological and Environmental Research Program
U.S. Department of Energy
Germantown, Maryland

April 19, 2005

Major DOE Missions

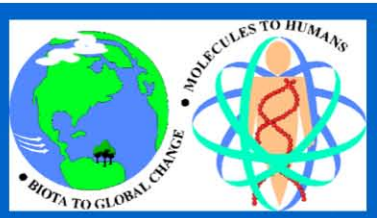
- Biological generation of energy compounds (biofuels) e.g. Hydrogen, Methane, Ethanol
- Carbon Dioxide Capture and Sequestration
- Bioremediation of legacy wastes (metals, radionuclides, toxic chemicals, etc.)

Microbial Genomes: Rationales



- **Microbes have evolved for over 3.5 billion years**
- **Est. 4 – 6 x 10^{30} on Earth (~ 5 x 10^{11} tons of C)**
- **They reside everywhere**
  T: -10°C to +121°C
  P: <1 atm to >230 atms
  Salinity: fresh water to Great Salt Lake/Dead Sea
  Acidity: pH < 0.5 to pH >10
  Radiation: up to 1.5 Mrads
  Location: >1 mile into subsurface, >1.5 miles in ocean
- **They can handle just about every element**
- **Most have never been characterized or identified**
- **Most do NOT cause diseases**

Microbial Genomes: Rationales (continued)



- **Importance to global processes, bioremediation, biogeochemistry, carbon sequestration, energy**
- **Modest sized genomes**
- **Lateral gene exchange**
- **25-45% unknown ORFs in newly sequenced genomes**
- **Vast repertoire of uncharacterized capabilities that could be useful in biotechnologies:**
  - **-- Medical**
  - **-- Pharmacological**
  - **-- Environmental**
  - **-- Industrial**
- **Scientific paradigm shift; holistic view of life**

It's Their World!

Humans: 6,400,000,000
[6.4×10^9]

Microbes:

6,000,000,000,000,000,000,000,000,000,000
(+/- 100x) Whitman, et. al, PNAS (1998)
[6×10^{30}]

Environmental Cleanup

Understanding How Bacteria Reduce Metals

Atomic Force Microscopy Image of *Shewanella* on An Iron Oxide Particle (Fredrickson, PNNL)

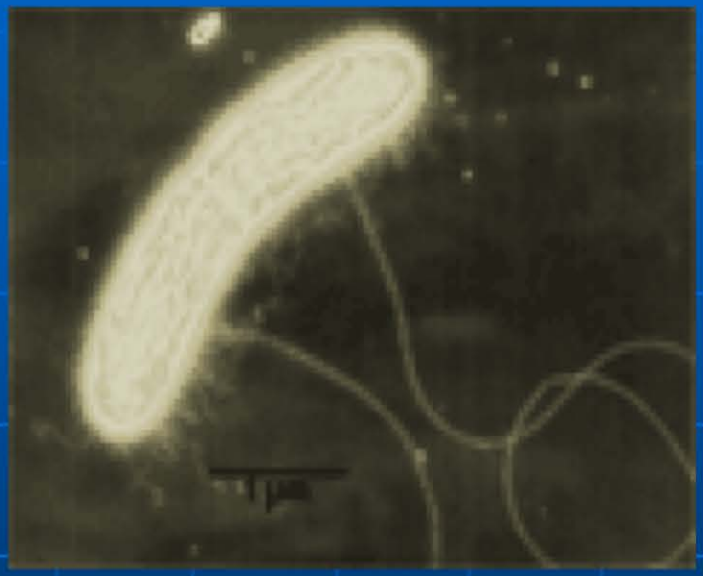


Image of *Geobacter metallireducens* (Lovley Lab, U Mass)

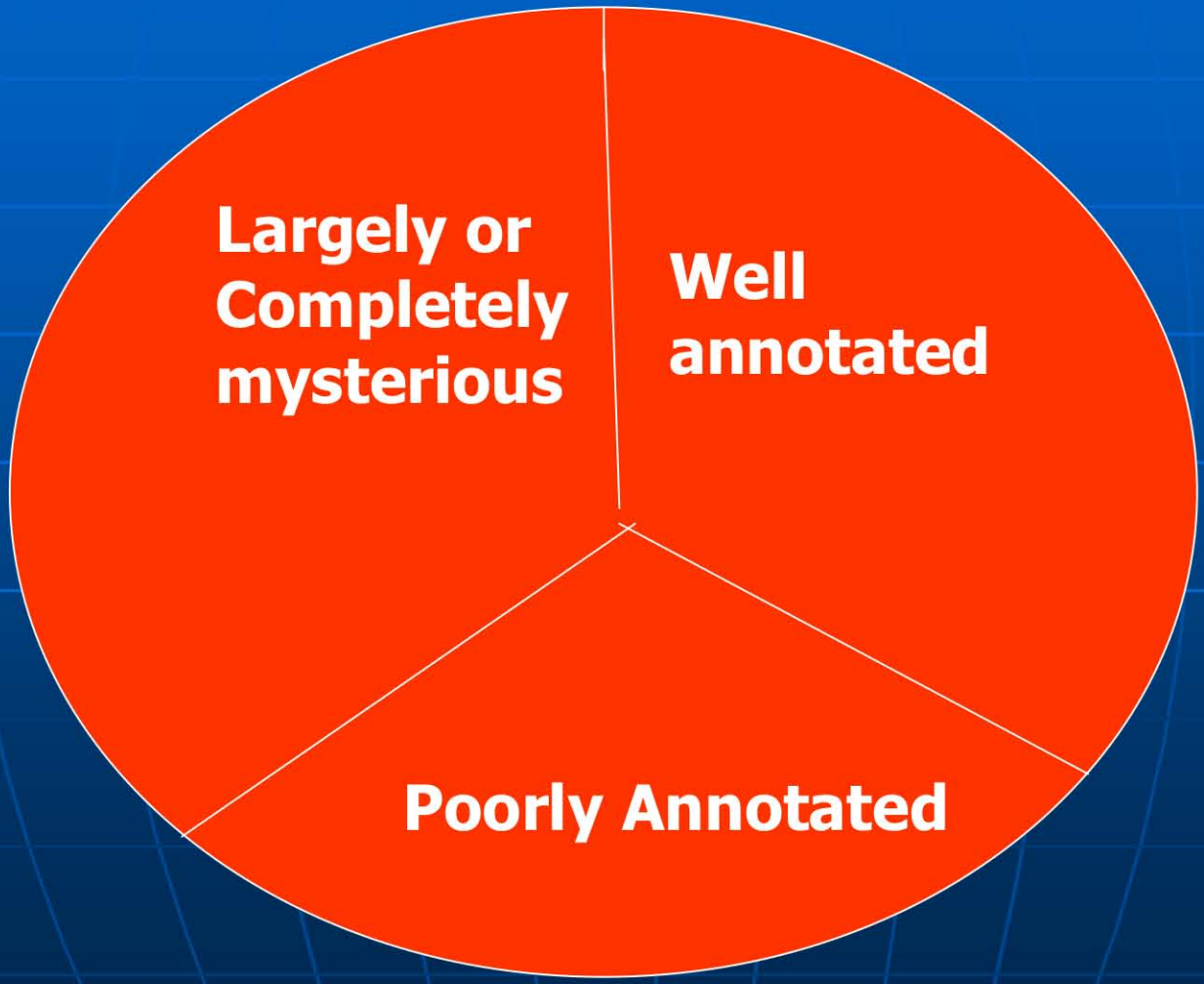
Sequencing to Date (4/12/05)

- http://www.genomesonline.org/

- Published Complete Genomes: **261**
- Prokaryotic Ongoing Genomes: **669**
- Eukaryotic Ongoing Genomes (including 7 chromosomes): **489**
- Total: **1421**

- **July 14, 2005: Nomination Deadline**

The Annotation Dilemma



Systems Biology: Sequencing is only the beginning

- "parts list" to functions
- Protein complexes and pathways
- Regulation (internal and external)
- Environmental diversity
- Computational modeling

Genomics:GTL Program Goals

Using DNA sequence and high-throughput technologies

goal 1

Identify and characterize the molecular machines of life

goal 2

Characterize gene regulatory networks

goal 3

Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level

goal 4

Develop the computational capabilities to advance understanding of complex biological systems and predict their behavior

Systems Biology

Gain a comprehensive and predictive understanding of the dynamic, interconnected processes underlying living systems

Why are Facilities Important to Complement the GTL Science?

Faster Better Cheaper

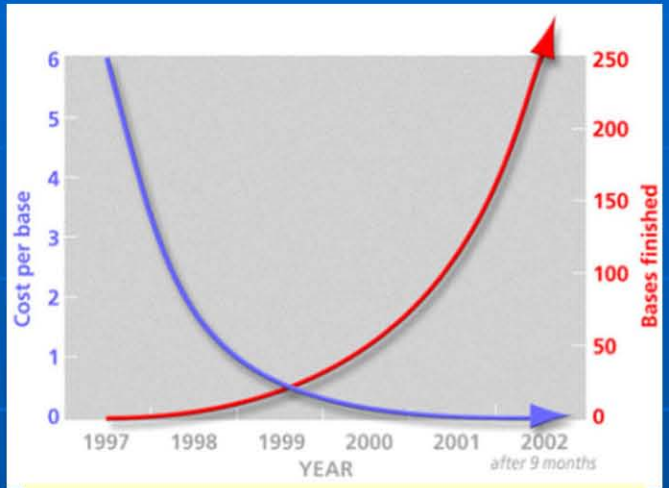
Efficiencies, Economies

Quality Control

Standardization

Democratiztion

Let Scientists do science.



Large-scale facilities spur cost and productivity improvements (data from the DOE Joint Genome Institute).

Instrumentation from a GTL protein production pilot project.

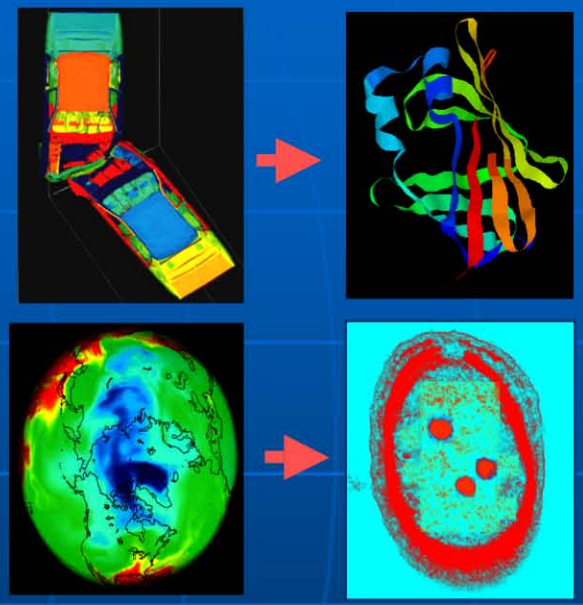
Scientists will be freed from production activities to do fundamental scientific research.

GTL User Facilities will Create

- **Research infrastructure for the new biology**
- **Comprehensive informatics and computing infrastructure**
- **Comprehensive microbial knowledgebase: From genomics to predictive biology**
- **A new multidisciplinary community of scientists trained in the new biology**
- **A new biotechnology base for DOE missions**
- **Training and education**

Ultimate Goal is to Provide Predictive Models of Microbes

This goal drives data collection and computing strategy.



Experimental:

- Complete datasets
- Quantitative measurements
- Comprehensive physical characterization:
  - **Protein expression and interactions**
  - **Spatial distributions**
  - **Process kinetics**

Computational:

- Automated data analysis and validation
- Automated integration of diverse data sets
- Human and computer-accessible databases
- Molecular, Pathway and cell-level simulations

The goals require a new synergy between computing and biology.

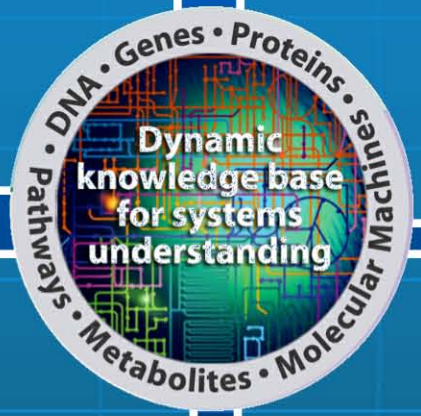
Key Points About GTL Facilities:

- **The purpose of the GTL Facilities is to enable access to data, greater analytic power, and resources – to democratize microbial research.**
- **Their outputs will COMPLEMENT individual investigator research (not supplant it – viz. synchrotrons).**
- **Biology is poised to enter a post-genomic "exponential" phase and these facilities will be vitally necessary.**
- **Community consultation is critical and we are committed to it; workshops, reactions solicited on web site, e-mail us.**
- **We have no choice if we are to exploit the genomics revolution.**

Genomics:GTL Facilities to Understand Cellular Systems

Facility for Production and Characterization of Proteins and Molecular Tags

Facility for Whole Proteome Analysis



Facility for Characterization and Imaging of Molecular Machines

Facility for Modeling and Analysis of Cellular Systems

A New Infrastructure for Biological Research

GTL Facilities: Resources for the Community

- JGI: 40% for BER, 50% for CSP, 10% for internal use.
- Facility 1: Initially most for GTL, evolving over time towards user facility for Science.
- Constant Principle: Rigorous Peer Review

Office of Science
U.S. Department of Energy

Functions of Facility I

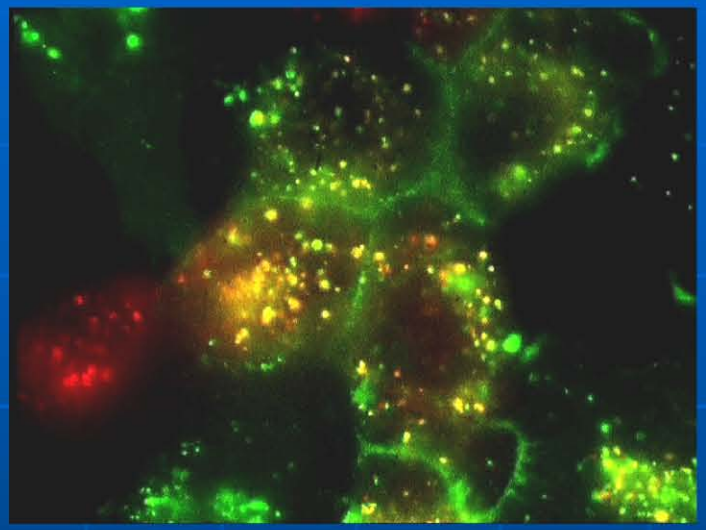
High-Throughput Production of:

- **Each gene in formats suitable for protein expression**
- **Active, full-length, purified proteins (~2 mg quantities each, 10-25,000 proteins/year, ~6 genomes/year)**
- **Protein variants or mutants**
- **Economical affinity reagents for each protein**
- **Biophysical characterizations; *e.g.* solubility, disorder & the structure-function paradigm**
- **Accessible databases & computational tools**

Facility for Production and Characterization of Proteins and Molecular Tags

What It Will Be

- **150,000-sq. ft. facility**
- **Advanced automated robots or production lines**
- **Advanced micro electro mechanical systems – lab on a chip, microfluidics**
- **Synchrotron and other characterizations**
- **Informatics and computing infrastructure**
- **Cryogenic storage, handling, and shipping**
- **Comprehensive R&D program**
- **Workshops to specify requirements**



Tracking proteins with tags in live cells.

Characterizing proteins with synchrotron X-rays.

Office of Science
U.S. Department of Energy

Exported Products from Facility I

- **What will Facility I deliver?**
  - **Validated clones and expression protocols for proteins & affinity reagents**
  - **Proteins and protein sets**
  - **Affinity reagents**
  - **Detailed, standardized biophysical characterizations**

Facility I: Production and Characterization of Proteins and Molecular Tags: First Uses (?)

- **Attack the problem of unannotated genes**
- **Validate novel approaches to annotation**
- **Generate and characterize entire classes of proteins (e.g. marine bacteriorhodopsins, hydrogenases, cytochromes, RubisCO, etc.)**
- **Supply reagents to enable studies of "molecular machines"**
- **"Build" sets of tagged proteins for intracellular biochemical investigations.**
- **Others……!**

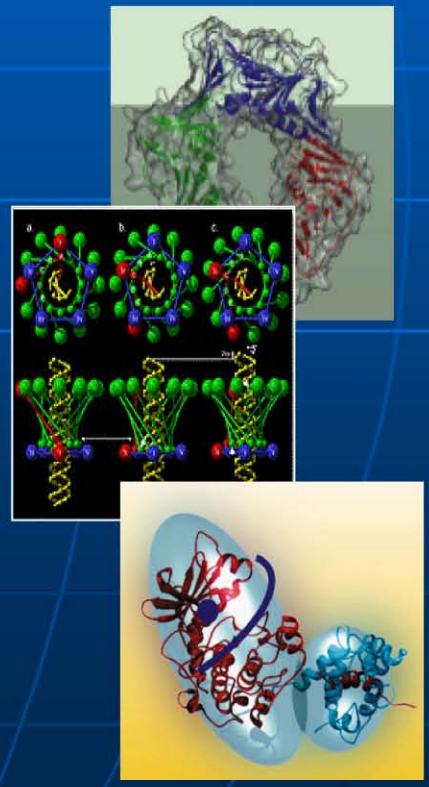
Facility II

Characterization and Imaging of Molecular Machines

Office of Science
U.S. Department of Energy

Exploring Molecular Machine Geometry and Dynamics

- **Computational Analysis, Modeling and Simulation**
  - Image analysis/cryoelectron microscopy
  - Protein interaction analysis/mass spec
  - Machine geometry and docking modeling
  - Machine biophysical dynamic simulation
- **Knowledge Captured**
  - Machine composition, organization, geometry, assembly and disassembly
  - Component docking and dynamic simulations of machines



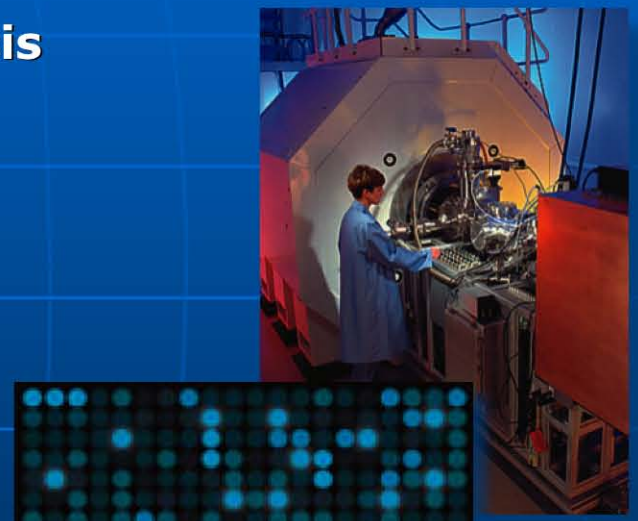
Facility III

Whole Proteome Analysis

Office of Science
U.S. Department of Energy

Modeling Proteome Expression, Regulation, and Pathways

- **Analysis and Modeling**
  - Mass spectrometry expression analysis
  - Metabolic and regulatory pathway/ network analysis and modeling
- **Knowledge Captured**
  - Expression data and conditions
  - Novel pathways and processes
  - Functional inferences about novel proteins/machines
  - Genome super annotation: regulation, function, and processes (deep knowledge about cellular subsystems)



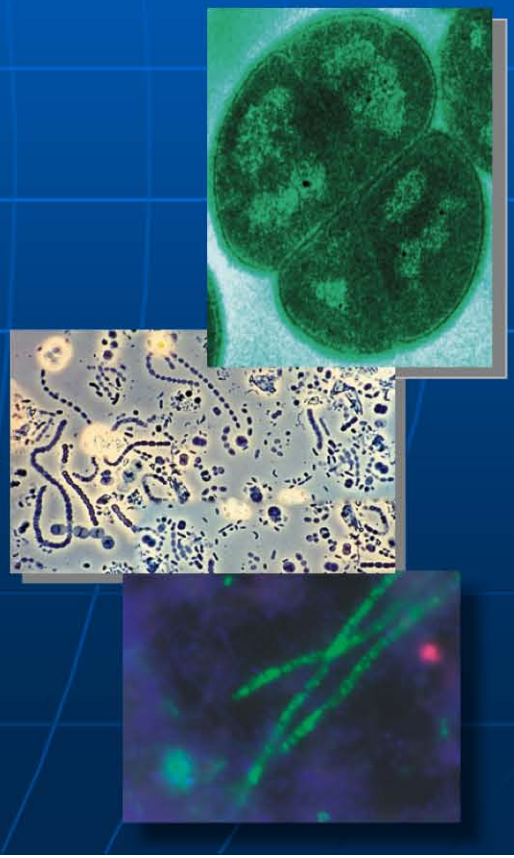
Facility IV

Office of Science
U.S. Department of Energy

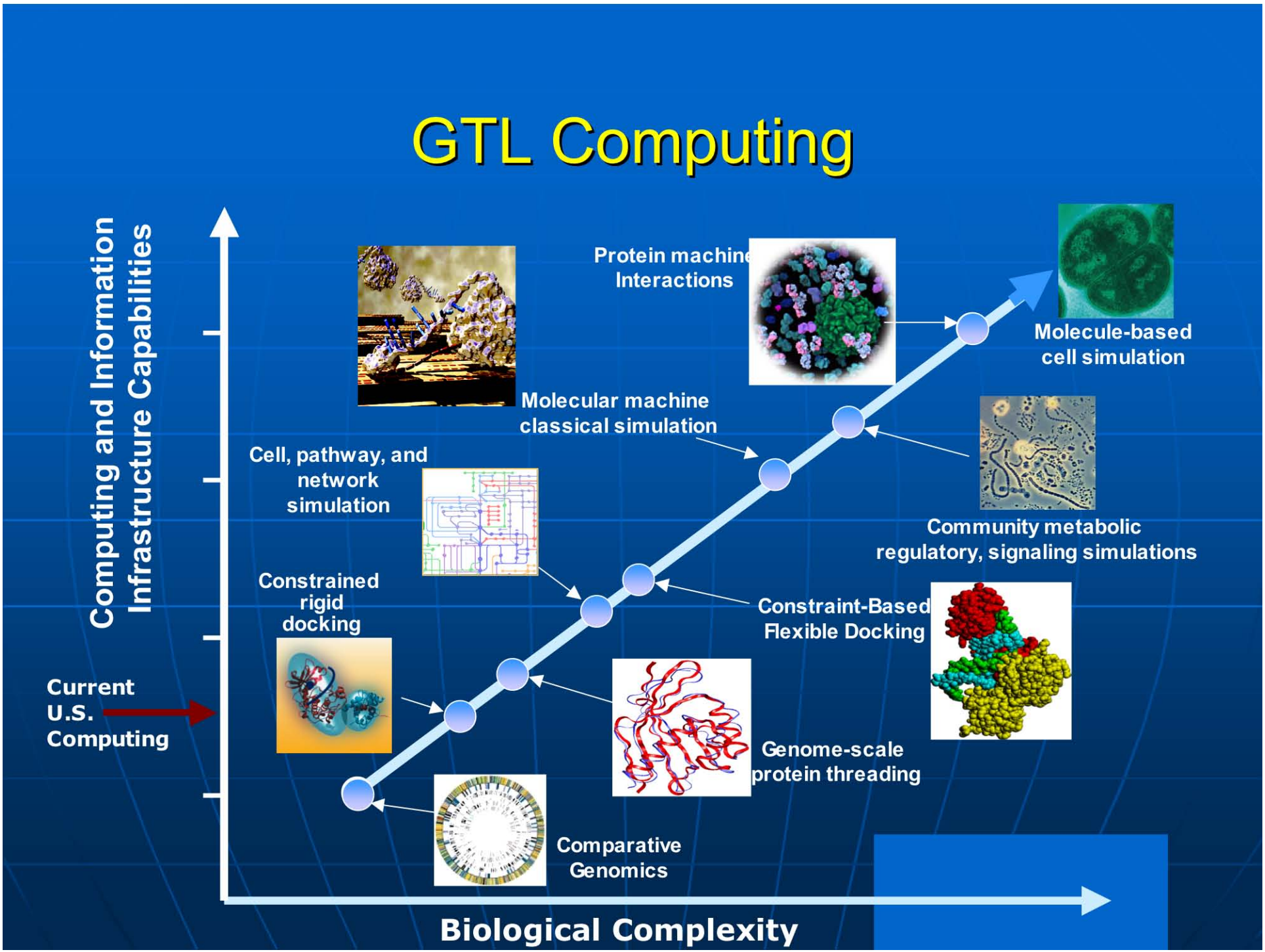
Analysis and Modeling of Cellular Systems

Simulating Cell and Community Dynamics

- **Analysis, Modeling and Simulation**
  - Couple knowledge of pathways, networks, and machines to generate an understanding of cellular and multi-cellular systems
  - Metabolism, regulation, and machine simulation
  - Cell and multicell modeling and flux visualization
- **Knowledge Captured**
  - Cell and community measurement data sets
  - Protein machine assembly time-course data sets
  - Dynamic models and simulations of cell processes



GTL Computing



?

GTL: The "Parts" → [Directions] → The Whole

Genomics: GTL

$$\text{The Whole} > \sum_{i=1}^{N} \text{part}_i$$

http://DOEGenomesToLife.org