Computational Biology

Reinhold Mann
Associate Laboratory Director, Biological and Environmental Sciences

Oak Ridge National Laboratory

April 20, 2004
2004 High Performance Computing Conference
Salishan Lodge, Oregon

The Computing Barometer

Source: Ed Uberbacher

Simulation of modified microbes in real world community situation	**>2015**
Robust models of machines, networks, cells, community interactions	**2005-2015**
GTL Facilities Planning Initiated	**2003**
Initiated GTL Program	**2002**
Joint Genome Institute formed	**1999**
Started Microbial Genome Program	**1994**
First complete microbe (*M. genitalium*)	**1994**
Initiated the Human Genome Project	**1986**
Created first DNA sequence database	**1979**
Computing used for early crystallography	**<1979**
Classical biology	**<1960**

Age of *In Silico* Biology

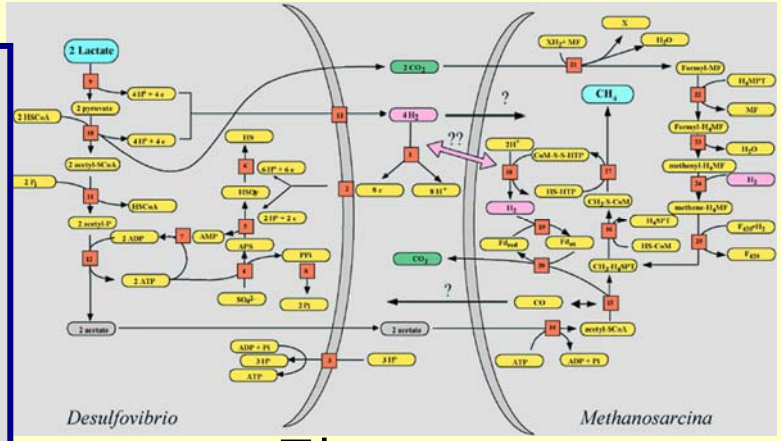
Models/Simulation Drive Experiment

Biology becomes an 'Information Science'

Niche Computing

No computing

To understand organisms from a "systems perspective," we must couple theory and experiments and analysis in a recursive manner.



Metabolic and regulatory models

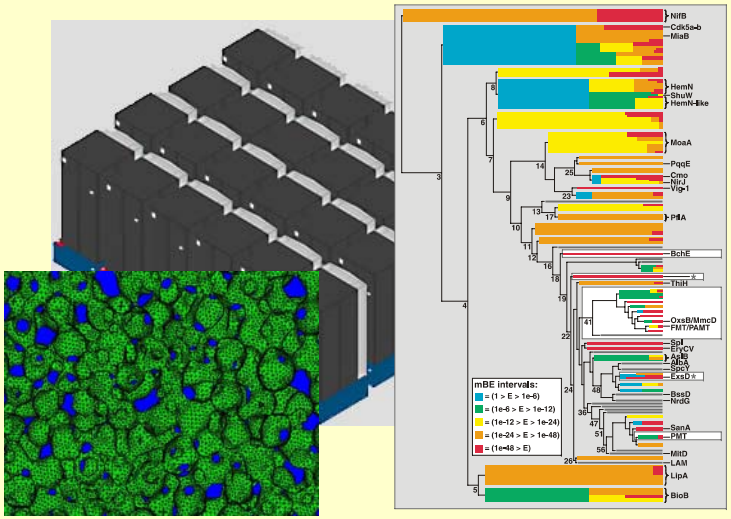
Theory

Analysis

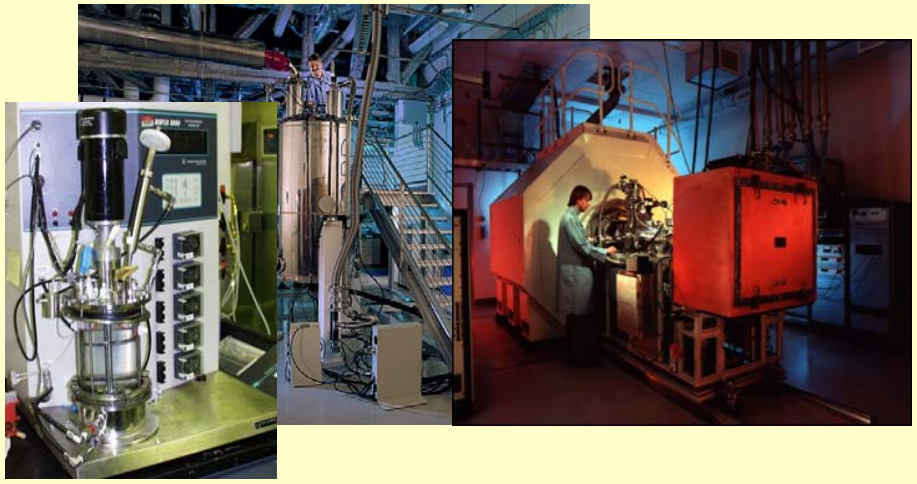
Experiment

Source: Steve Wiley, PNNL

Advanced Computation and Information Management

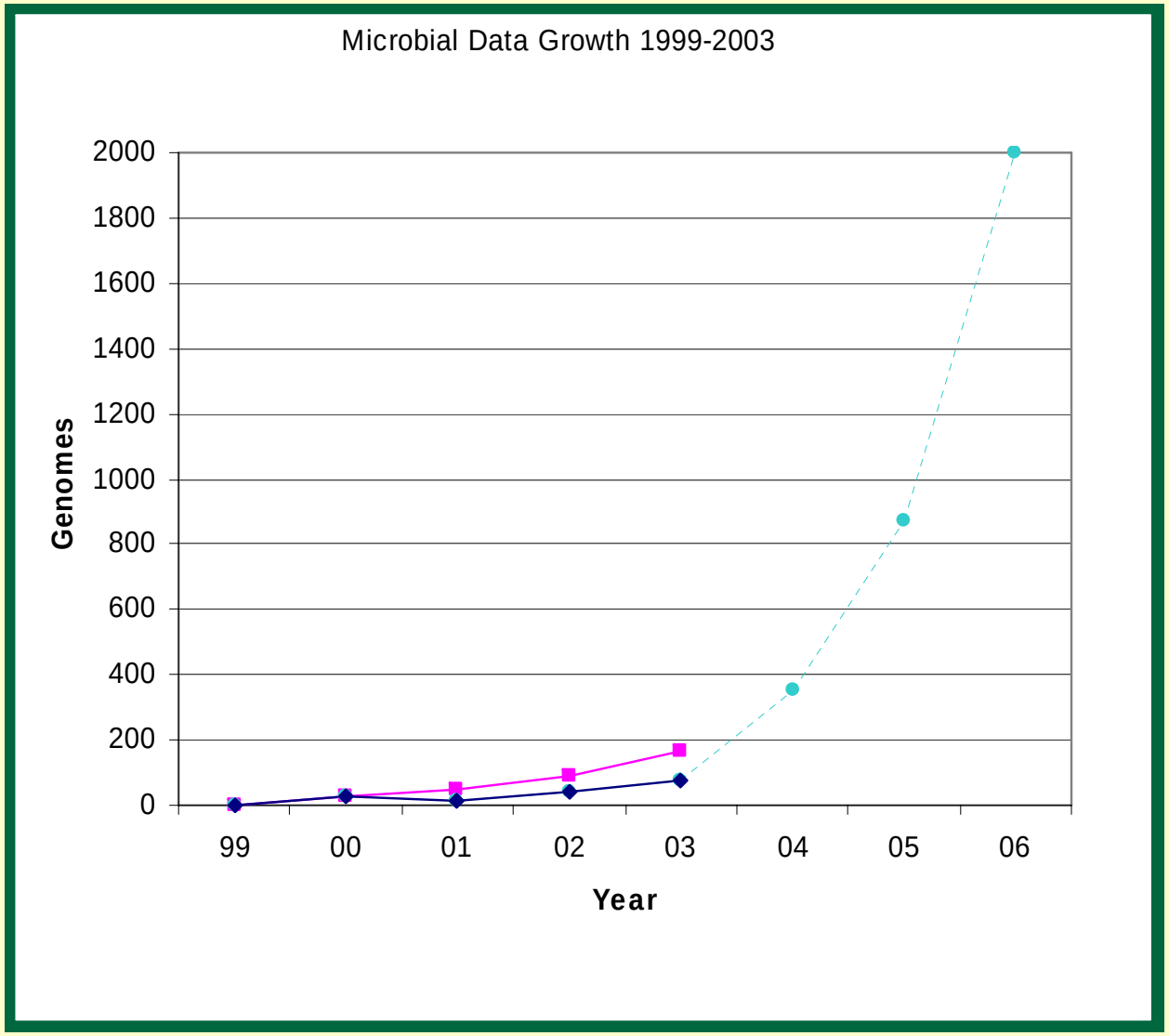


Advanced High-Throughput Technologies



Human Genome Project

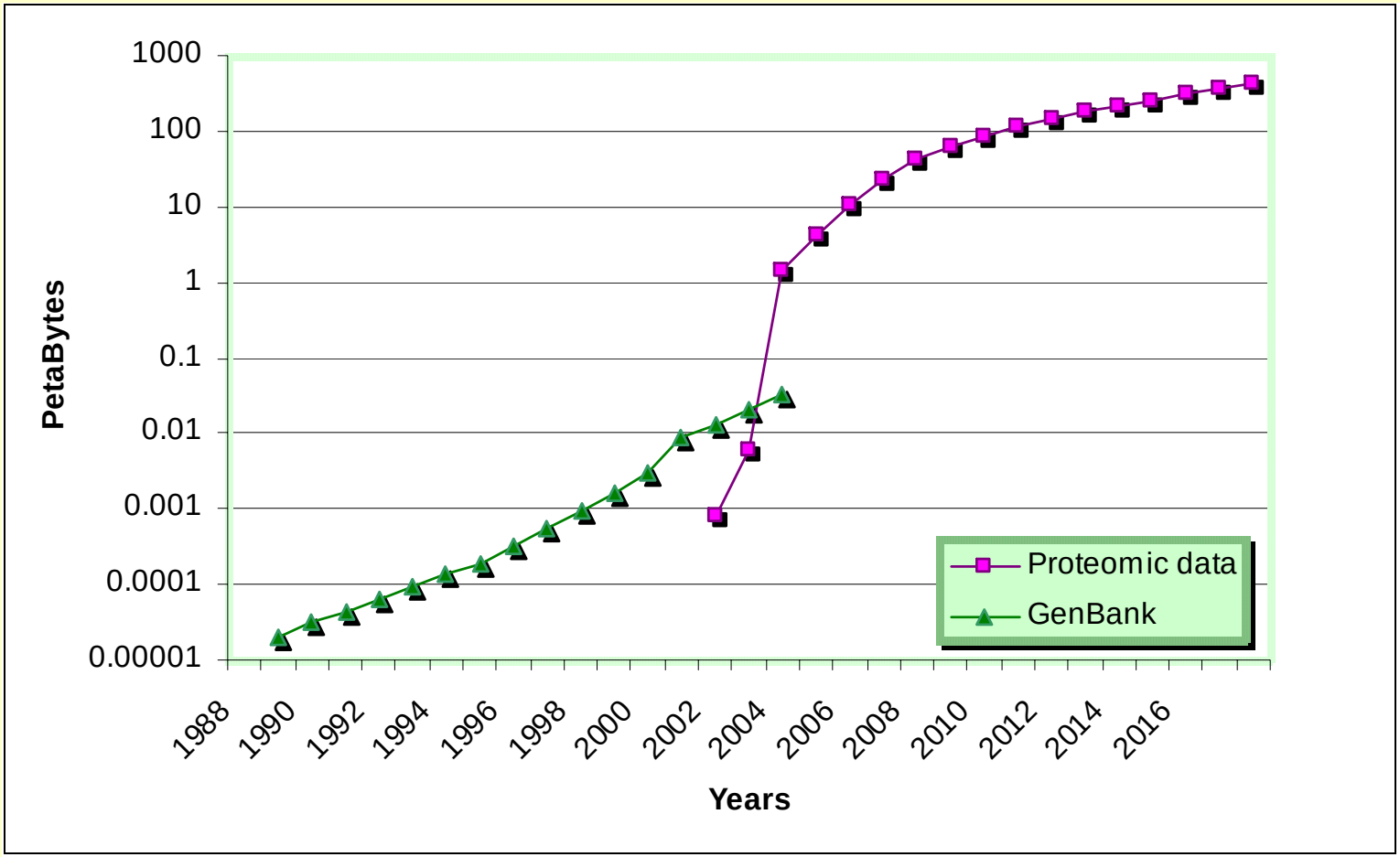
. . . Making the Impossible Routine



Providing the Paradigm for High-Throughput Biology Facilities

Source: Ed Uberbacher

Growth of Proteomic Data vs. Sequence



Each Data Dimension Multiplies the Scale of Data and Analysis

DOE Genomics:GTL Example Experiment Template

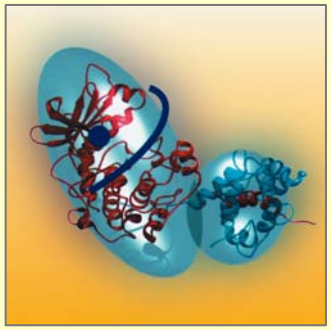
Experiment templates for a single microbe									
class of experiment	time points	treatments	conditions	genetic variants	biological replication	total biological samples	Proteomics data volume in TB	Metabolite data in TB	Transcription data in TB
simple (scratching the surface)	10	1	3	1	3	90	18.0	13.5	0.018
moderate	25	3	5	1	3	1125	225.0	168.8	0.225
upper mid	50	3	5	5	3	11250	2250.0	1687.5	2.25
complex	20	5	5	20	3	30000	6000.0	4500.0	6
real interesting	20	5	5	50	3	75000	15000.0	11250.0	15
Profiling method									
Proteomics	Looking at a possible 6000 proteins per microbe assuming ~200 GB per sample								
Metabolites	Looking a panel of 500-1000 different molecules assuming ~150GB per sample								
Transcription	6000 genes & 2 arrays per sample ~100 MB								
	Typically a single significant scientific question takes the multidimensional analysis of at least 1000 biological samples								

Source: George Michaels

GTL Data will be 1000's of times larger than Genomes

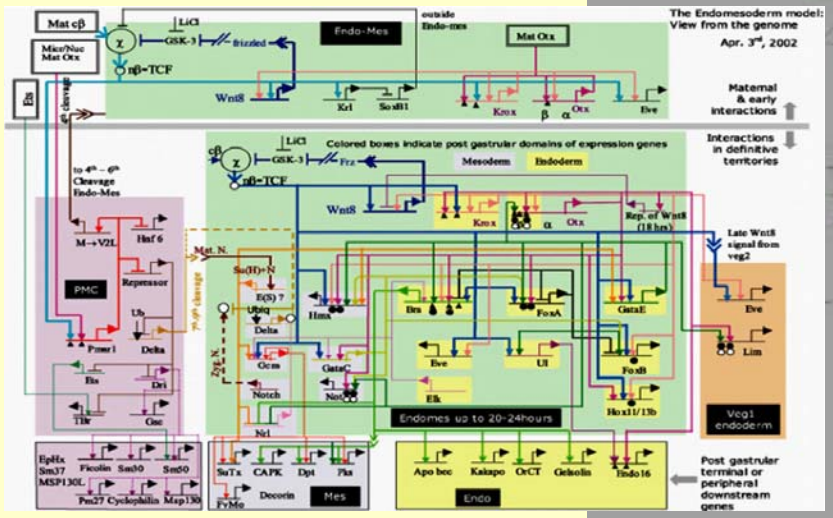
Example of overarching scientific objective from biology that drives HPC

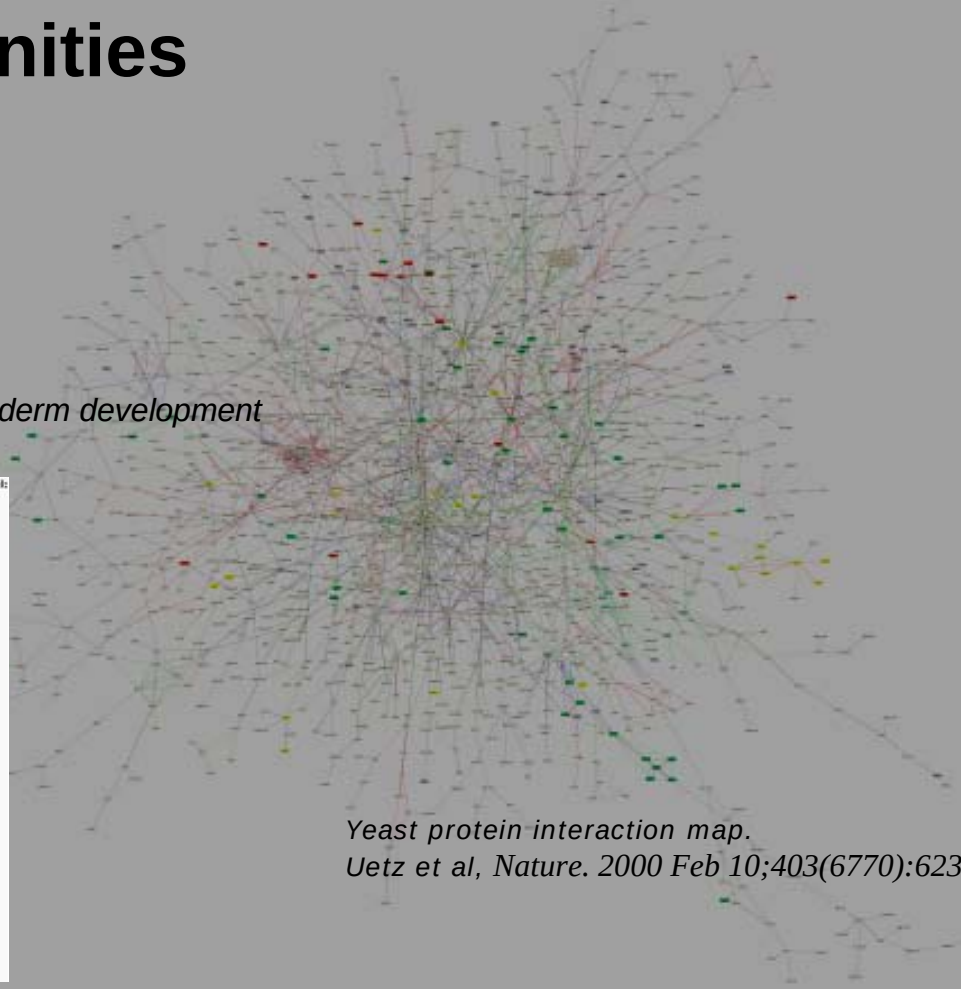
- **Understand molecular interactions in cells and cellular communities**



The enzyme CAM kinase II and its activator protein calmodulin
Jill Trewhella

Gene network for endomesoderm development
Eric Davidson





Yeast protein interaction map.
Uetz et al, Nature. 2000 Feb 10;403(6770):623-7.

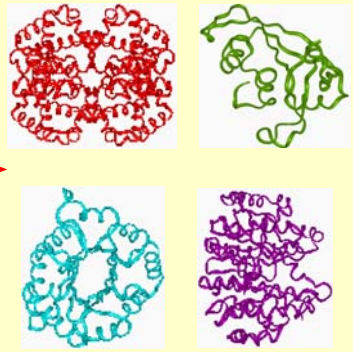
Taxonomy of frequently used algorithms in computational biology

- **Genome annotation and comparative genome analyses**
  - **Pattern recognition, dynamic programming, optimization, …**
- **Network and pathway analyses and simulations**
  - **Graphs, ODEs, stochastic and ODE networks, …**
- **Structure prediction**
  - **MD, dynamic programming, docking, ...**

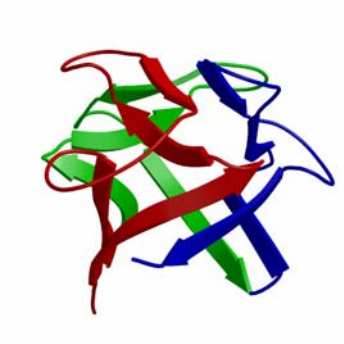
New biology facilities will require HPC for both C*apacity* and *Capability*

ATCGTAGCAATCGACCGT...
CGGCTATAGCCGTTACCG...
TTATGCTATCCATAATCGA...
GGCTTAATCGCATACGAC...

Thread onto templates

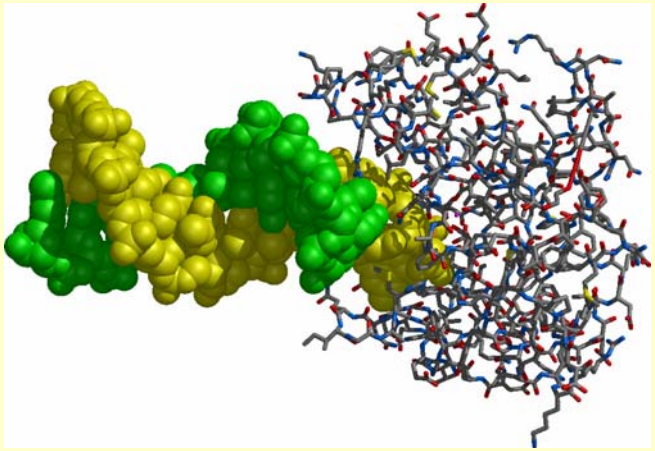


Best match

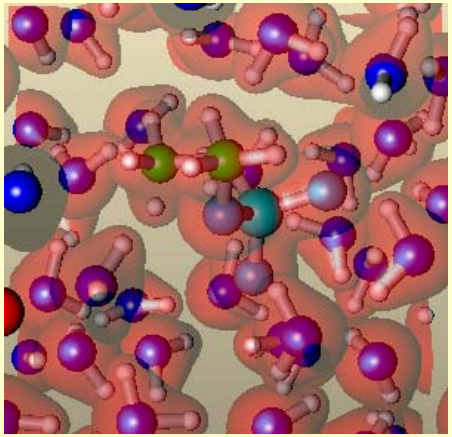


***Capacity*: e.g., High-throughput protein structure predictions, data analysis, sequence comparison**

***Capability*: e.g., Large scale biophysical simulations, stochastic regulatory simulations:**



Large size and timescale classical simulations



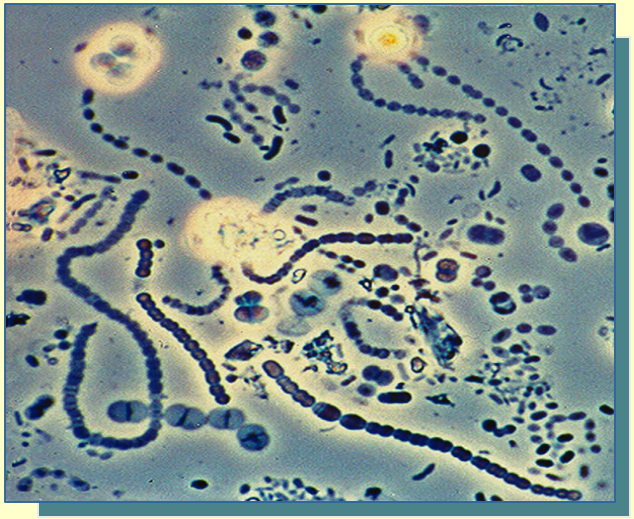
Highly accurate quantum mechanical simulations

Petascale *Capacity* Problems in Biology

Microbial and Community Genome Annotation

Now

Analyze and annotate 20 microbial genomes - (720,000 processor hours)



In 5 years

Assemble, Analyze and annotate community of 200 Microbes and phage (10,000,000 Processor hours)

Compare genome sequences (200 megabases) To previous genomes (4 gigabases) (5,000,000 processor hours)

Petascale *Capacity* Problems in Biology

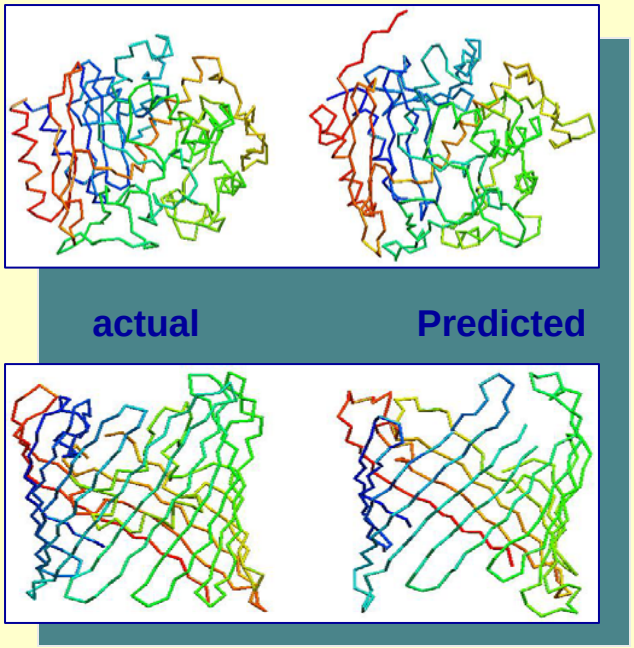
Protein Fold Prediction using Knowledge based potentials

Now

Protein fold prediction of 2000 proteins in a microbial genome using knowledge-based potentials - (100,000 processor hours)

In 5 years

Protein fold prediction for 200 microbes (400,000 proteins) in a microbial community (20,000,000 processor hours)



Petascale Capability Problems in Biology

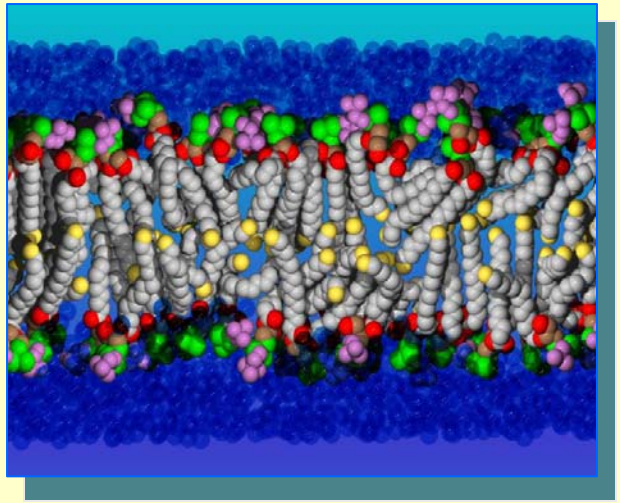
Membrane simulation using classical potentials

Now

Observe heterogeneous lipid segregation (patching) - (600,000 processor hours for 200 Nanosecond simulation)

In 5 years

Simulate membrane protein association and lipid interactions (7,000,000 processor hours for 1 millisecond simulation)



Petascale Capability Problems in Biology

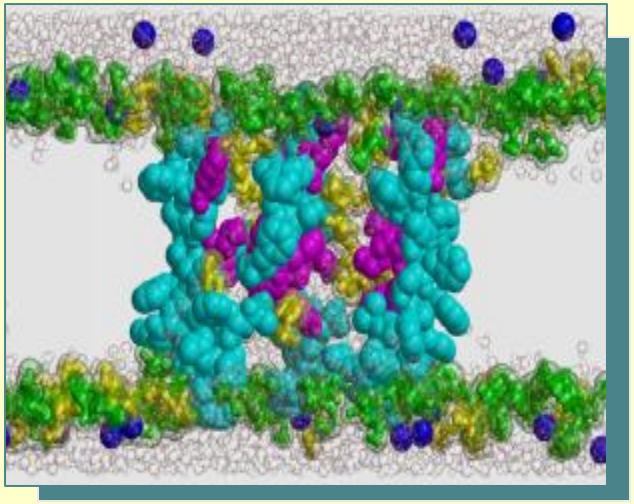
Membrane channel simulation

Now

Simulate non-flexible protein ion channel
K+ flow using quantum methods
(2,200,000) processor hours for
4 second simulation

In 5 years

Simulate flexible protein ion pump
for producing ATP from K+ gradient
(15,000,000 processor hours for 200
nanosecond simulation



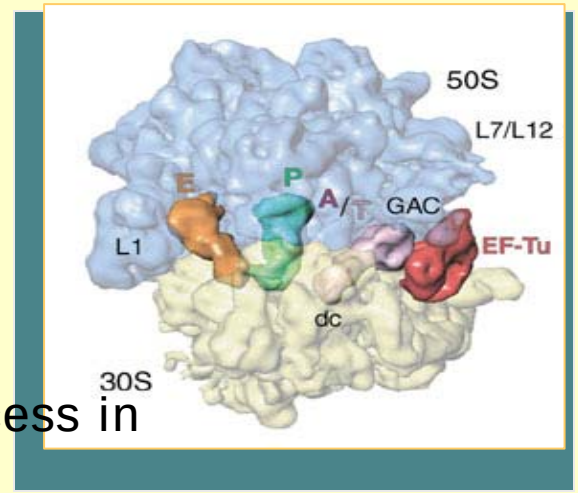
Petascale Capability Problems in Biology

Ribosomal Interactions and Dynamics

Now

Simulate Ribosome EF-tu interaction
Using classical molecular dynamics
(400,000) processor hours for 20
Nanosecond simulation

In 5 years



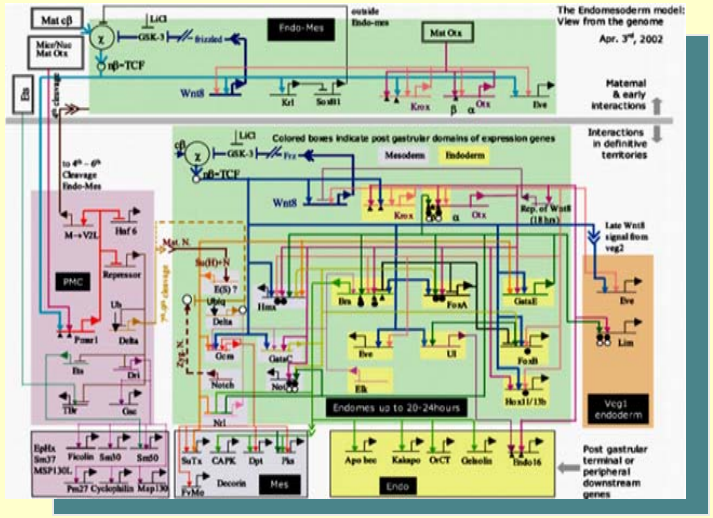
Simulate key steps in amino acid translation process in the Ribosome (25,000,000 processor hours for 1 millisecond simulation)

Petascale Capability Problems in Biology

Regulatory and Protein Interaction Networks

Now

Simulate moderately complex network using ordinary differential equations (10,000) processor hours for 1 millisecond simulation



In 5 years

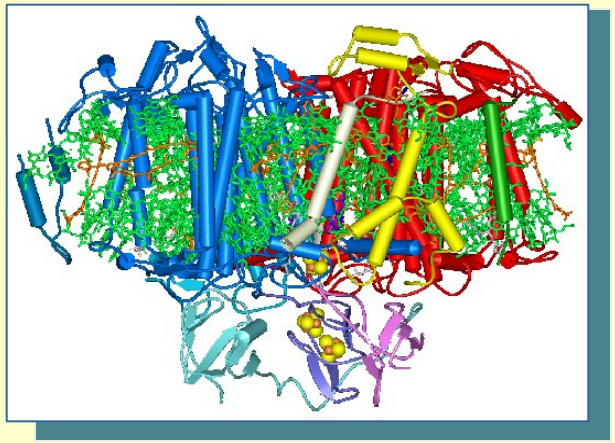
Multiscale stochastic and stochastic differential equation simulation of complex regulatory network (1,000,000 processor hours for 1 millisecond simulation)

Petascale Capability Problems in Biology

Molecular Machine Interactions

Now

Simulation of microbial gene regulatory factor binding (300,000 processor hours for 50 nanosecond simulation)



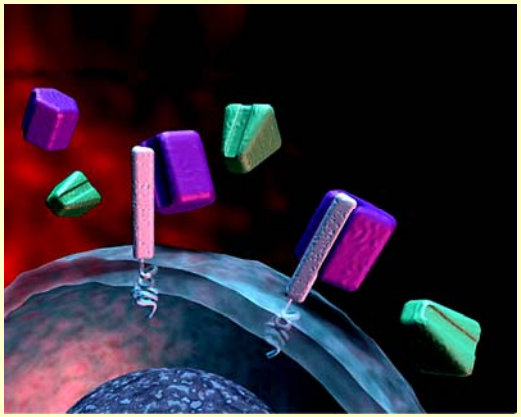
Rigid docking of multiple components of Protein complex and ligand (500,000 Processor hours for orientation search)

Petascale Capability Problems in Biology

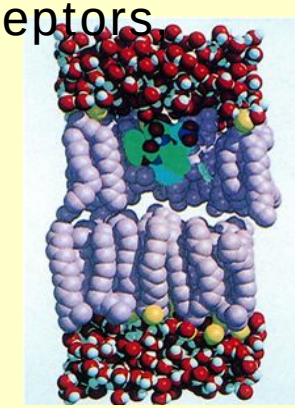
Molecular Machine Interactions and Dynamics

In 5-10 years

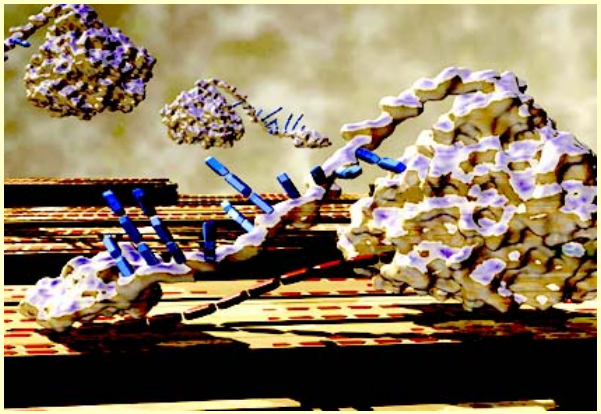
Simulate dynamics and chemistry of Cellulase, cell surface receptors, bionano structures



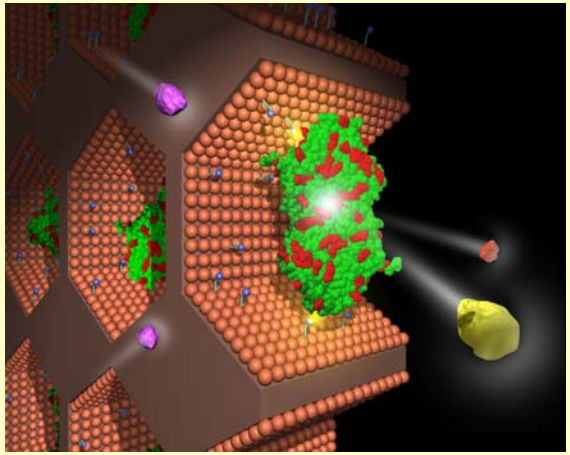
Cell-environment sensors



Membrane complexes



Cellulase molecular machine converting cellulose to glucose



Bionano chemistries

Virtual Microbial Cell Simulation

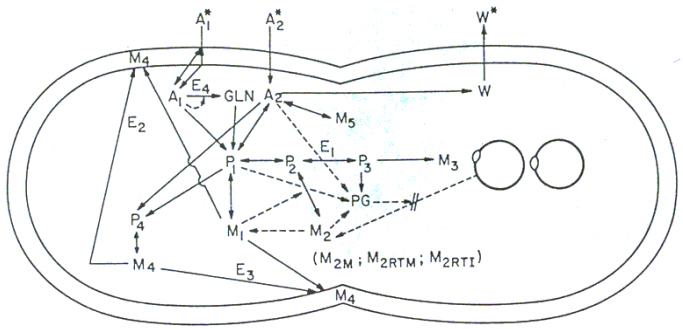
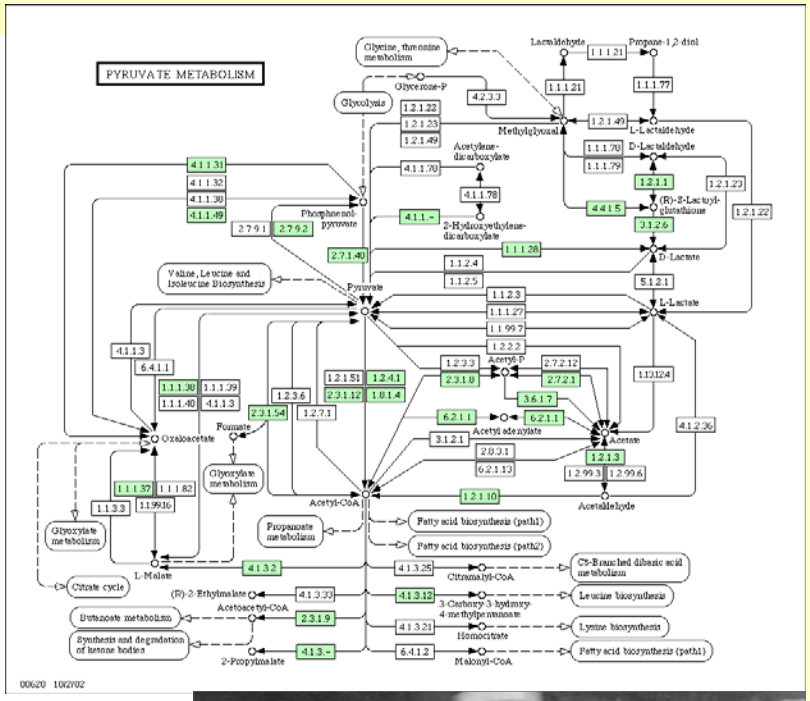
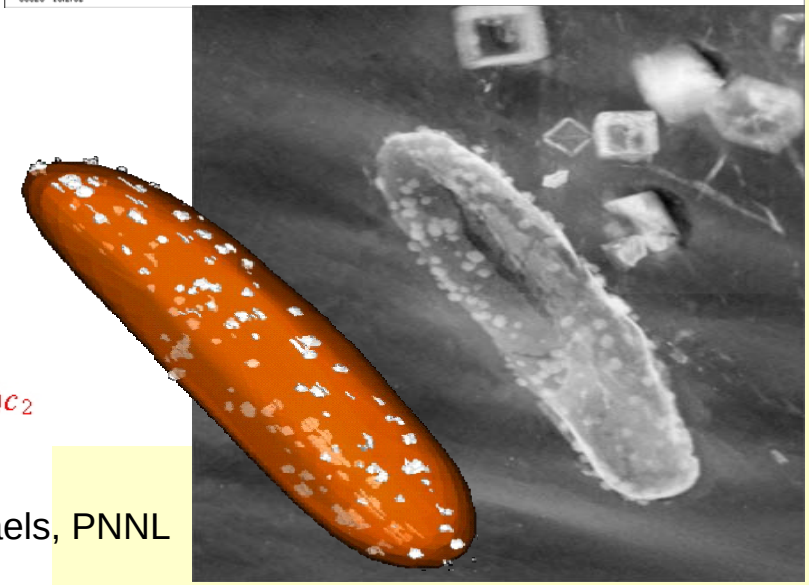


Figure 1. An idealized sketch of the model for E. coli *B/r A growing in a glucose–ammonium salts medium with glucose or ammonia as the limiting nutrient. At the time shown the cell has just completed a round of DNA replication and initiated cross-wall formation and a new round of DNA replication. Solid lines indicate the flow of material, while dashed lines indicate flow of information.*

A_1 = *ammonium ion*
A_2 = *glucose (and associated compounds in the cell)*
W = *waste products (CO_2, H_2O, and acetate) formed from energy metabolism during aerobic growth*
P_1 = *amino acids*
P_2 = *ribonucleotides*
P_3 = *deoxyribonucleotides*
P_4 = *cell envelope precursors*
M_1 = *protein (both cytoplasmic and envelope)*
M_{2RTI} = *immature "stable" RNA*
M_{2RTM} = *mature "stable" RNA (r-RNA and t-RNA—assume 85% r-RNA throughout)*
M_{2M} = *messenger RNA*
M_3 = *DNA*
M_4 = *non-protein part of cell envelope (assume 16.7% peptidoglycan, 47.6% lipid, and 35.7% polysaccharide)*
M_5 = *glycogen*
PG = *ppGpp*
E_1 = *enzymes in the conversion of P_2 to P_3*
E_2, E_3 = *molecules involved in directing cross-wall formation and cell envelope synthesis—the approach used in the prototype model was used here but more recent experimental support is available*
GLN = *glutamine*
E_4 = *glutamine synthetase*

*—the material is present in the external environment.





$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C_1 \rightarrow^{k_2} E + P$$

$$S + C_1 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} C_2 \rightarrow^{k_4} C_1 + P$$

$$\frac{ds}{dt} + \vec{u} \cdot \vec{\nabla} s = -\vec{\nabla} \cdot D_s \vec{\nabla} s - k_1 s e + k_{-1} c_1 - k_3 s c_1 + k_{-3} c_2$$

$$\frac{dc_1}{dt} + \vec{u} \cdot \vec{\nabla} c_1 = -\vec{\nabla} \cdot D_{c_1} \vec{\nabla} c_1 + k_1 s e - (k_{-1} + k_2) c_1 - k_3 s c_1 + (k_4 + k_{-3}) c_2$$

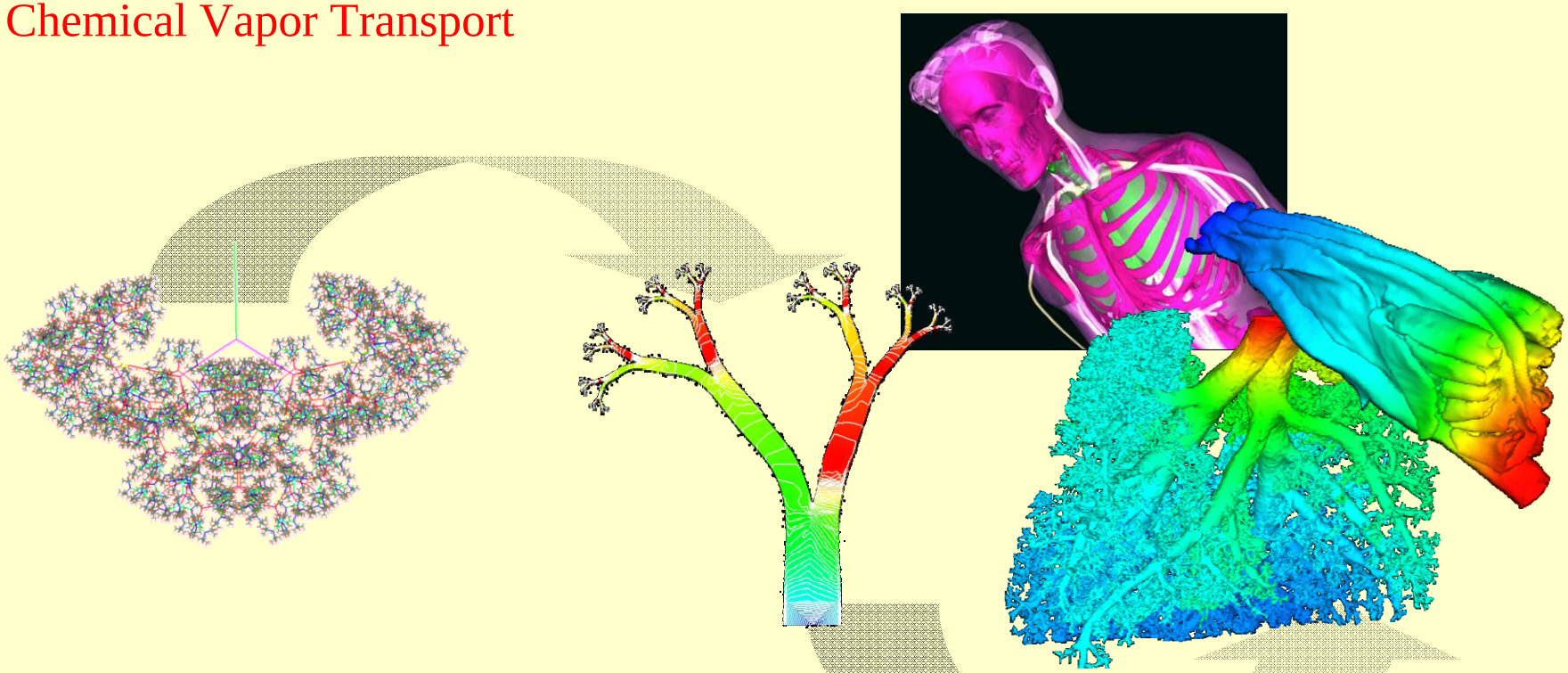
$$\frac{dc_2}{dt} + \vec{u} \cdot \vec{\nabla} c_2 = -\vec{\nabla} \cdot D_{c_2} \vec{\nabla} c_2 + k_3 s c_1 - (k_4 + k_{-3}) c_2$$

$$e + c_1 + c_2 = e_0$$

Source: George Michaels, PNNL

Petascale Capability Problems in Biology

Virtual Respiratory Tract Modeling and Simulation with Particle Deposition and Chemical Vapor Transport



Simple 2-D geometries with fluid dynamics possible.

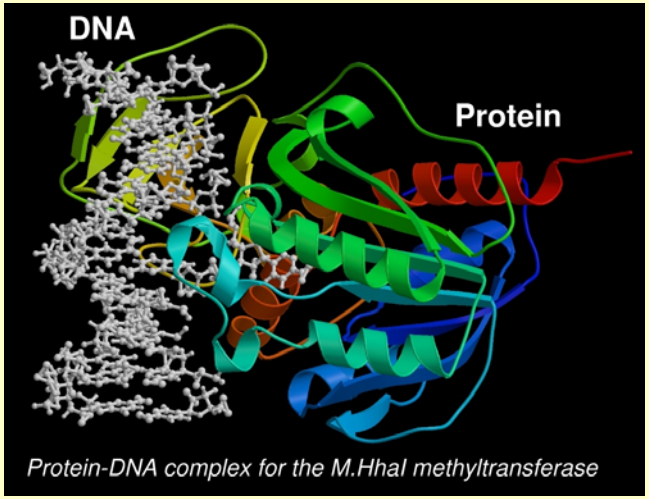
More complex 3-D geometries and fluid dynamics and transport.

Coupled upper and lower respiratory tract, with: 3-D geometries, fluid dynamics, tissue mechanics, transport and cellular response

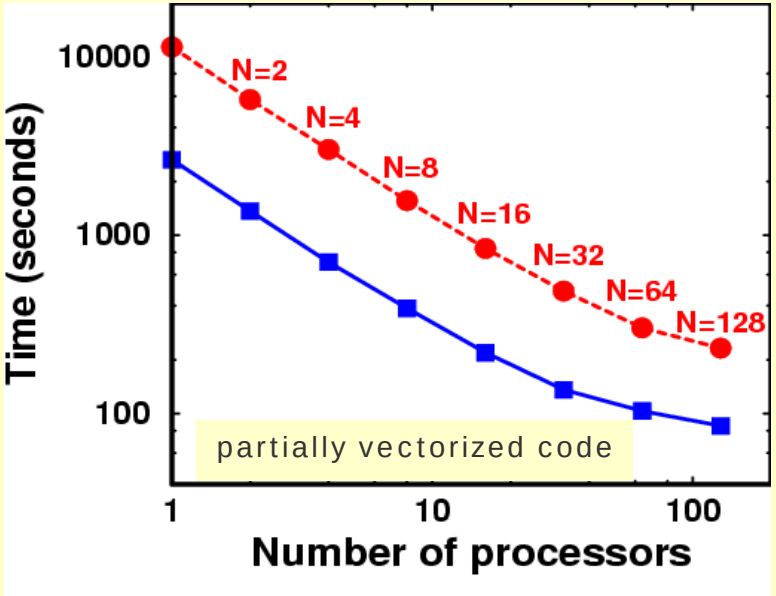
Source: George Michaels, PNNL

Large Scale Parallelization of Molecular Machines Dynamics Simulations

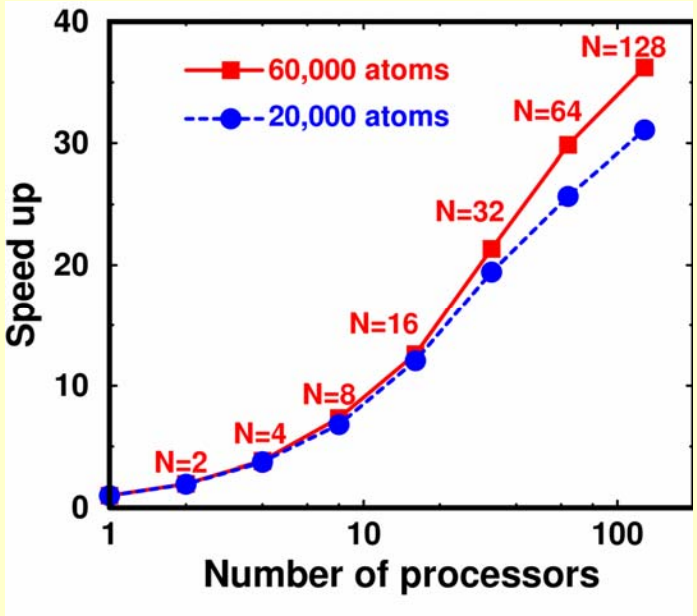
- Cray X-1 vs. Cheetah 60,000 atom DNA-protein benchmark
- 1 nanosecond per 2 days with preliminary code on X-1
- Collaboration with Jim Maltby (Cray) to fully vectorize Amber
- Extension of linear range to 1000(s) of processors possible
- Good machine for biology applications
- Will provide microsecond simulations in days
- This is needed for key steps in molecular machines dynamics



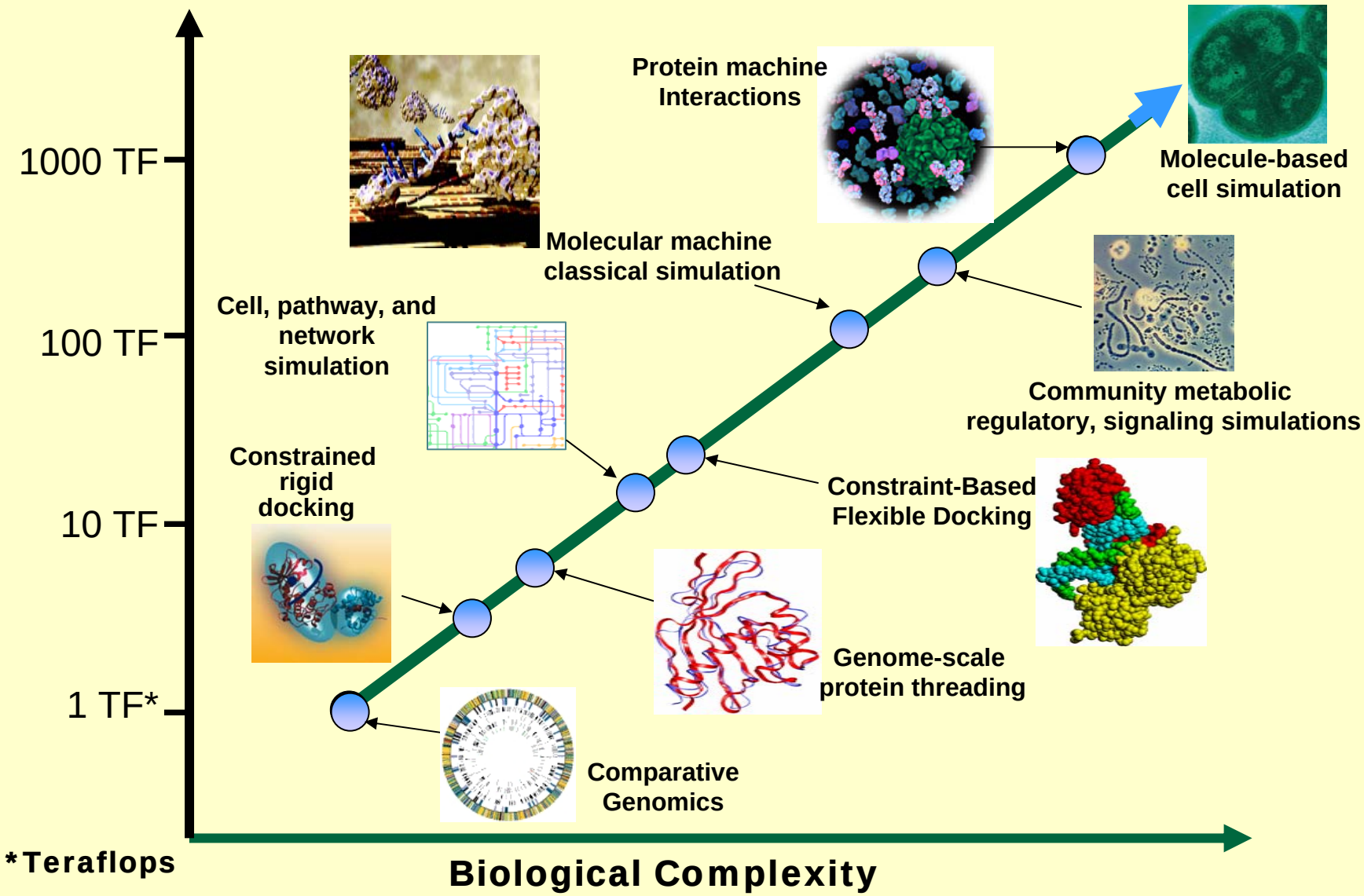
Much deeper vectorization is needed to
Extend linear scaling range to 1000 processors



Processor speed and communication speed are
optimal for large molecular problems (~60,000 atoms)



DOE Genomics:GTL High-Performance Computing Roadmap



New paradigms in Environmental Sciences require HPC

Experiments

Models

Theory

Observations

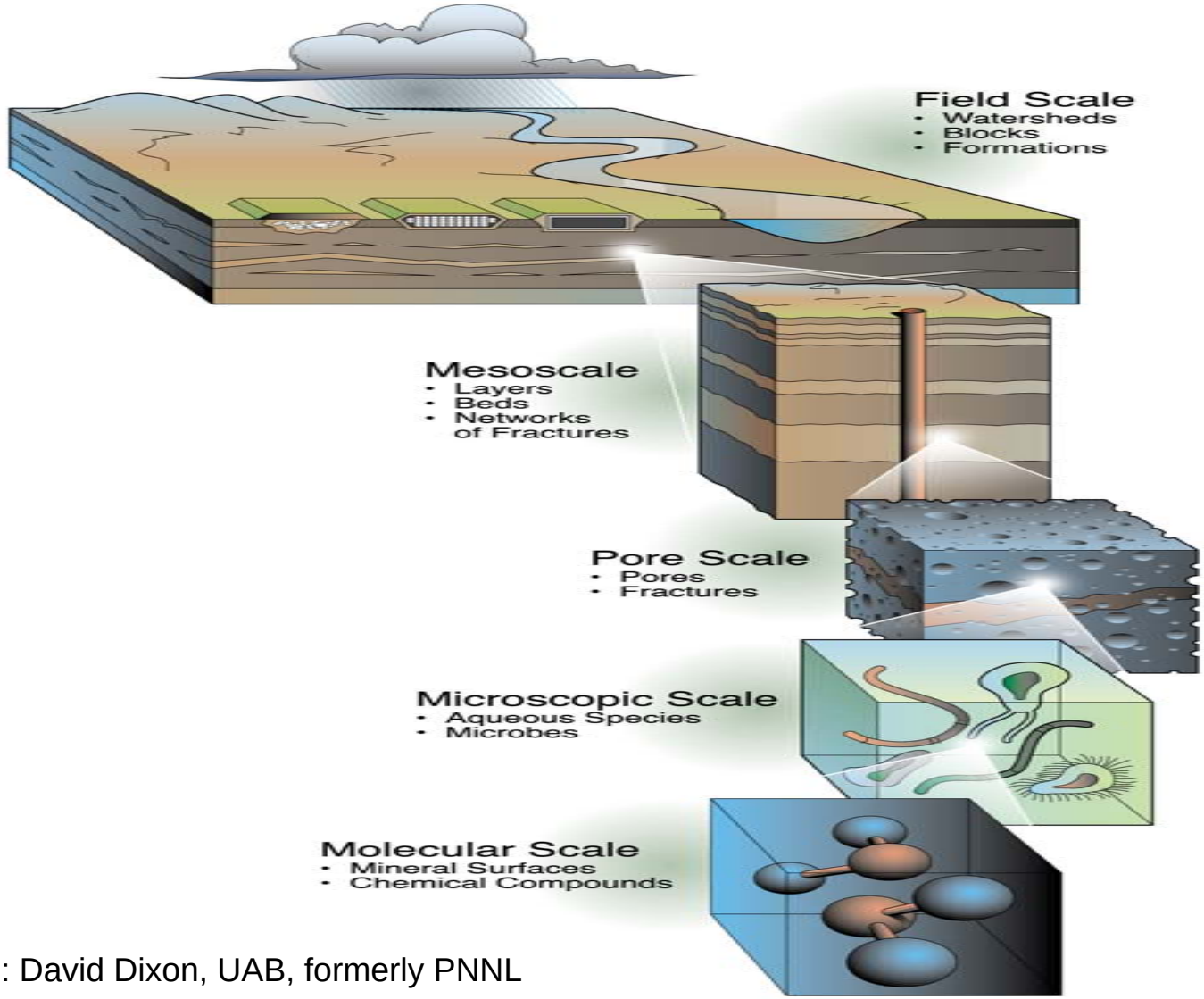
Real-Time Ecology
- Sensors
- Simulation
- Genomics
- Assessment

Ecosystem manipulation
- Global change hypothesis testing

Watershed-scale land-atmosphere interactions
- Biogeochemical cycling in aquatic and terrestrial ecosystems

Radioecology and systems ecology
- Radiation efforts research led to discoveries in energy and nutrient flow in terrestrial ecosystems

Subsurface science applications



Source: David Dixon, UAB, formerly PNNL

Summary

- **Biology and Biotechnology in the 21st Century require both high end capacity and capability computing**
- **Many bioinformatics applications benefit from large shared-memory architectures**
- **Applications in the area of structure determination; network and pathway analysis, modeling and simulation perform well on clusters or vector machines**
- **Community code development is maturing, e.g., AMBER, NWChem**