

The role of the SciDAC II in Predictive DOE Mission-Driven Systems Biology

- **Julie Mitchell**, University of Wisconsin, Madison
- **Nagiza Samatova**, ORNL
- **Tjerk Straatsma**, PNNL

The success of the Genome Projects through the partnership between biological sciences & computer science, math. & IT

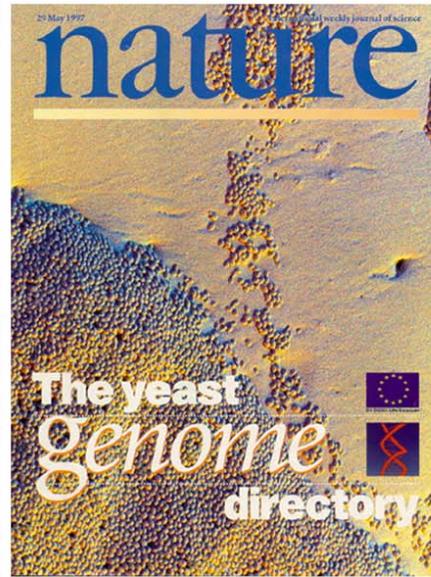
1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* 269: 496]



1997

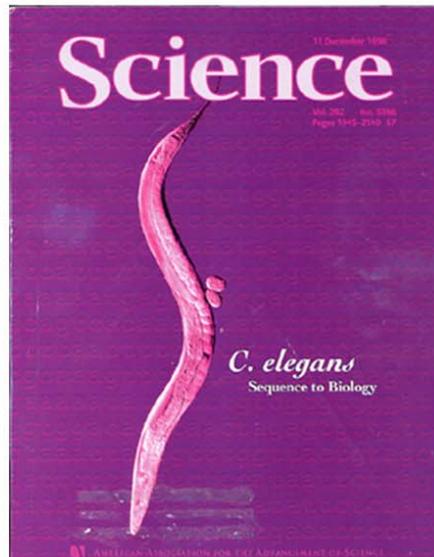
Eukaryote,
13 Mb,
~6K genes
[*Nature* 387: 1]



Highlights
of the
Genome
Projects

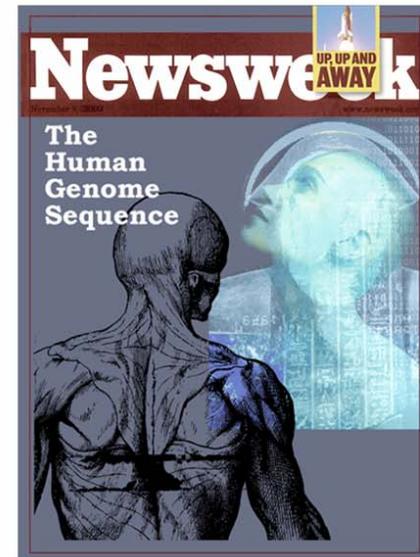
1998

Animal,
~100 Mb,
~20K genes
[*Science* 282:
1945]

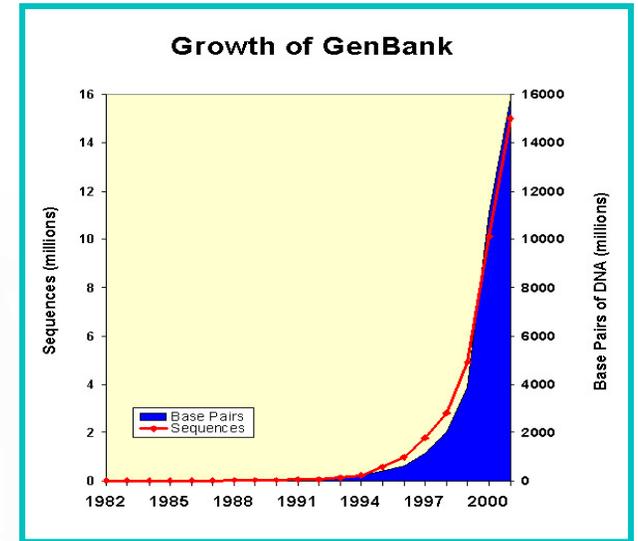
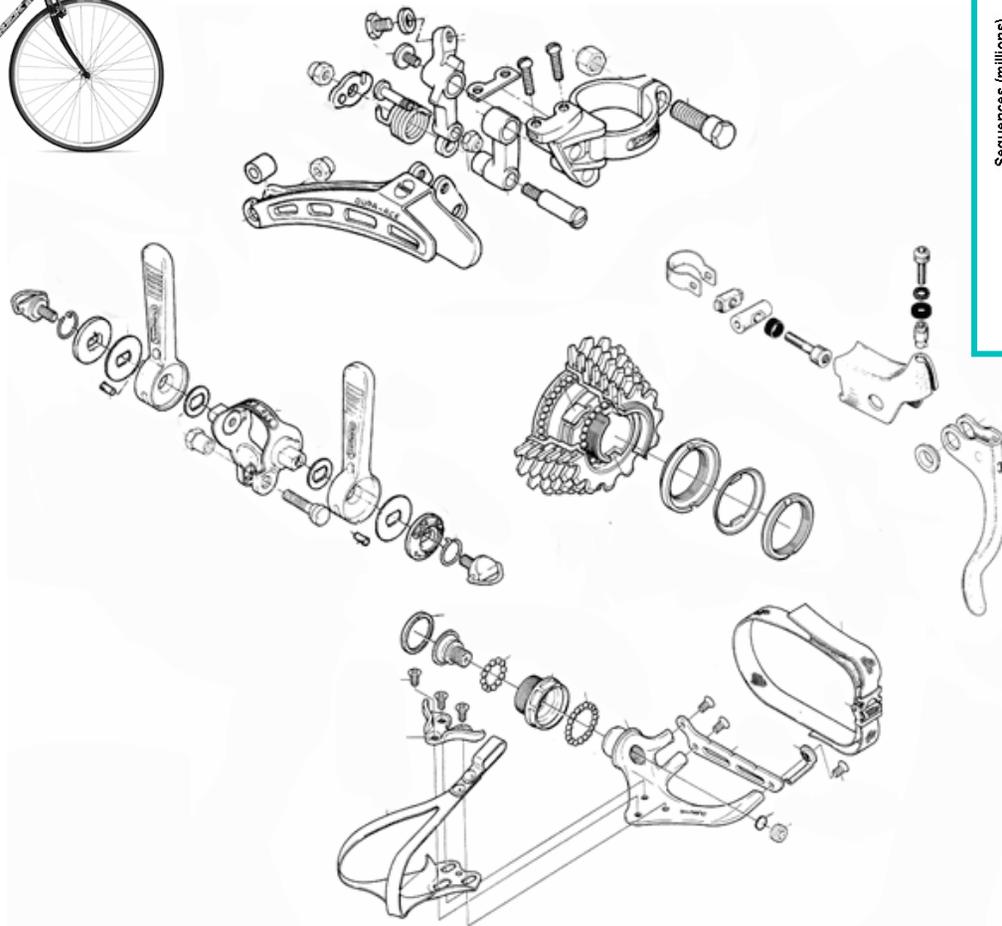


2000

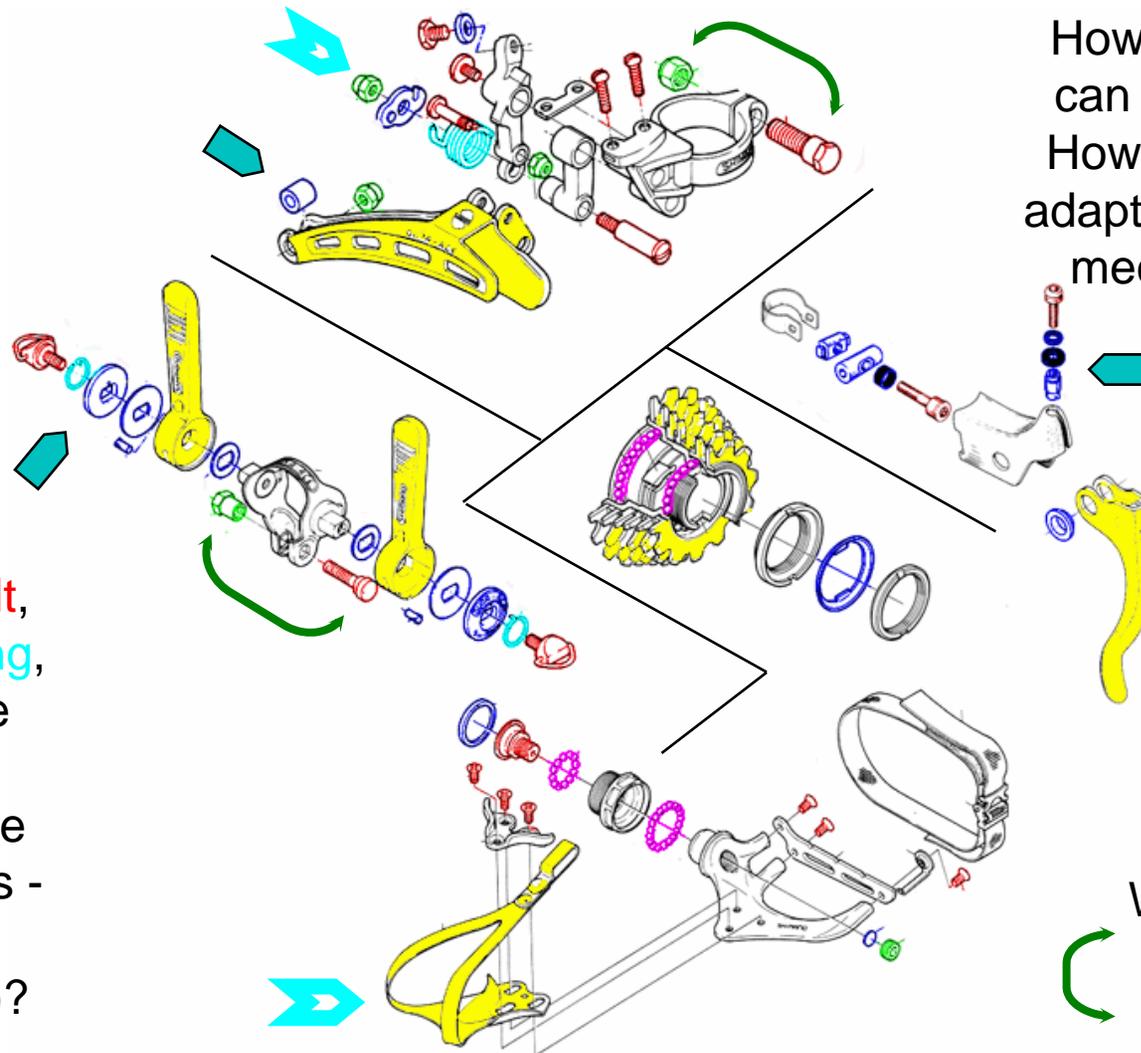
Human,
~3 Gb,
~30K genes



A Collection of Parts



Assembly & Maintenance



What are the shared parts (**bolt**, **nut**, **washer**, **spring**, **bearing**), unique parts (**cogs**, **levers**)? What are the common parts - types of parts (**nuts** & **washers**)?

How many roles can these play?
How flexible and adaptable are they mechanically?

Where are the parts located?
Which parts interact?

What is After the Genome?

Genome

Proteome

Transcriptome

Physiome

Metabolome

Phenome

Morphome

Interactome

Glycome

Secretome

Ribonome

Orfeome

Regulome

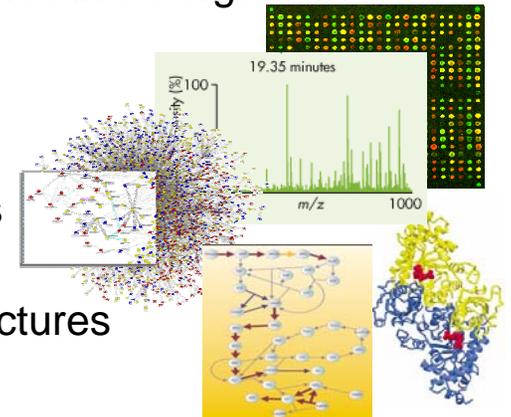
Cellome

Operome

Transportome

Functome

- **Genomics projects have led to a fundamental way in which science is done. Genomics has opened the door to “*discovery driven*” studies as contrasted with traditional “*hypothesis driven*” research.**
- **An approach that seeks a global and simultaneous view of:**
 - Transcriptome – what are all the transcribed genes in an organism
 - *Gene expression microarrays*
 - Proteome – what are all the expressed proteins
 - *Two dimensional gel / Mass spectrometry*
 - Interactome – all the protein-protein interactions
 - *Two hybrid or tagging experiments*
 - Structurome – all three dimensional protein structures
 - Metabolome – metabolites
- **Overall goal is to use the genome as a “parts list” and to build higher order models of function integrating this complete list and comprehensive high-throughput technologies**



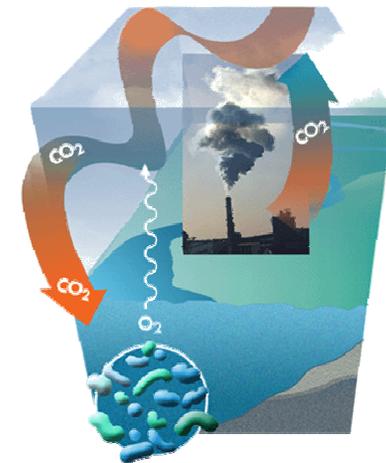
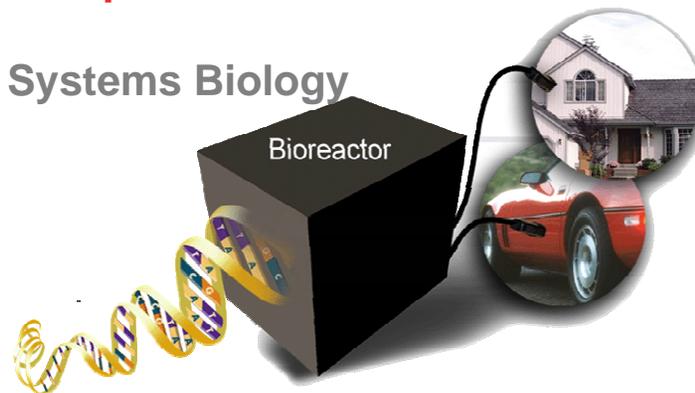
National Quest for Systems Biology

With the unique challenges for the DOE

To gain functional understanding and control of **microbial organisms** to help solve challenges in global climate change, environmental cleanup, and energy security

“The limits of component-by-component analysis are becoming apparent. To be fully understood, complex biological process – such as signal transduction, malignant transformation, or gene regulation – need to be studied **using systems-wide approaches developed at the interface of multiple disciplines.**”

MIT Computational and Systems Biology Initiative (CSBi)



DOE OS's Systems Biology Programs

<http://doegenomestolife.org/research/currentprojects.shtml>



- *GTL Center for molecular and cellular systems* (ORNL, PNNL)
- *Molecular Machines to Hierarchical modeling* (ORNL and SNL)
- *Rapid Deduction of Stress Pathways in Bacteria* (LBL, ORNL, SNL)
- *Microbial Ecology, Proteogenomics and Computational Optima* (Harvard Medical School)
- *Analysis of the Genetic Potential and Gene Expression of Microbial Communities* (U. Mass.)
- *Institute for Biological Energy Activities* (C. Ventor's Inst.)
- *Shewanella Federation* (PNNL, ORNL, Biotech)

- *Institutes for the Advancement of Computational Biology R&D:*
 - a) Johns Hopkins University, the University of Delaware and LANL
 - b) University of California, Merced; Rensselaer Polytechnic Institute; and LLNL
 - c) University of Wisconsin, Madison

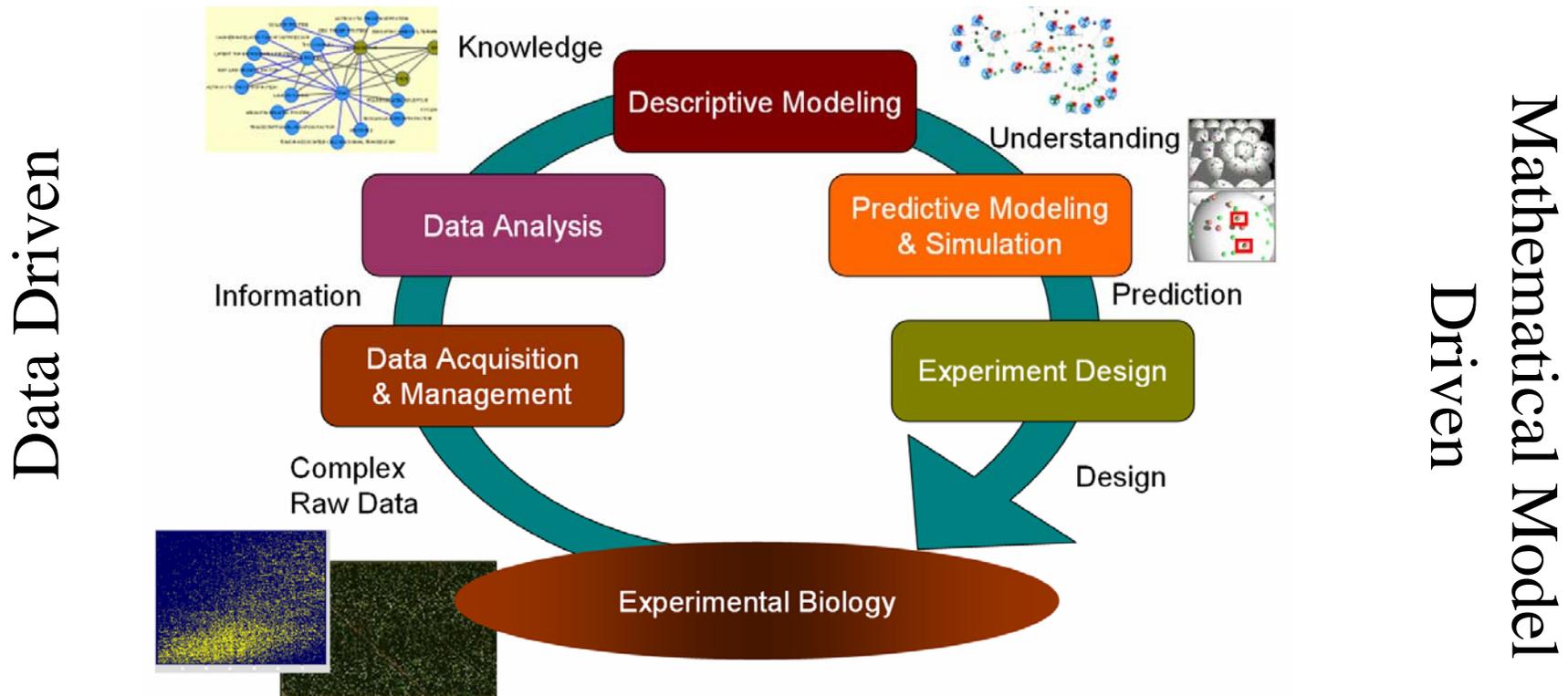
What is the Role of the SciDAC-II for DOE Systems Biology?

Reports: SCaLeS, HECRTF, Genomics:GTL

Neither current software nor current HEC hardware are adequate for modeling & simulation of biological systems

- Investments in scientific computing hardware and software, over the past decade, has overwhelmingly focused on scientific simulations from “**first principles**” including astrophysics, climate modeling, fusion, and materials.
- The result is that high-end computing (HEC) HW is relatively well configured to run scientific simulation software and to store the massive amounts of data that it produces.
- However...

Fundamental differences exist between the scientific discovery cycle of applications from “first principles” and the ones of systems biology



- Unlike simulations from “first principles”, systems biology starts from analysis and interpretation of the massive amounts of data.
- These data are needed for **building the model** and **constraining the space of feasible models** which make simulations computationally tractable.

Biology Presents Unique Challenges to Mathematics and Computer Science

- To offer the capability needed to obtain and understand the avalanche of complex biological data
- To provide the theory, principles, software and hardware infrastructure to construct and run predictive simulations of biological systems.

To successfully address these challenges

Requires *specialized* high-performance computing architectures & advanced *data management* technologies.

Optimized for both
data-intensive and *compute-intensive* science

Objectives of the White Paper

- Translate the DOE systems biology goals into **high-end computing requirements** and identify existing and emerging **resources** relevant to these goals;
- Describe the current **state-of-the-art capabilities** in relevant computational and biological research areas enabled by the DOE systems biology program and scientific community at large;
- Identify **technical challenges** and needs for further development of computational methods, data-analysis tools, and modeling and simulation of biological systems **under the SciDAC umbrella**;
- Identify **high-performance computing infrastructure requirements** to accomplish DOE predictive systems biology goals.

Findings and Recommendations

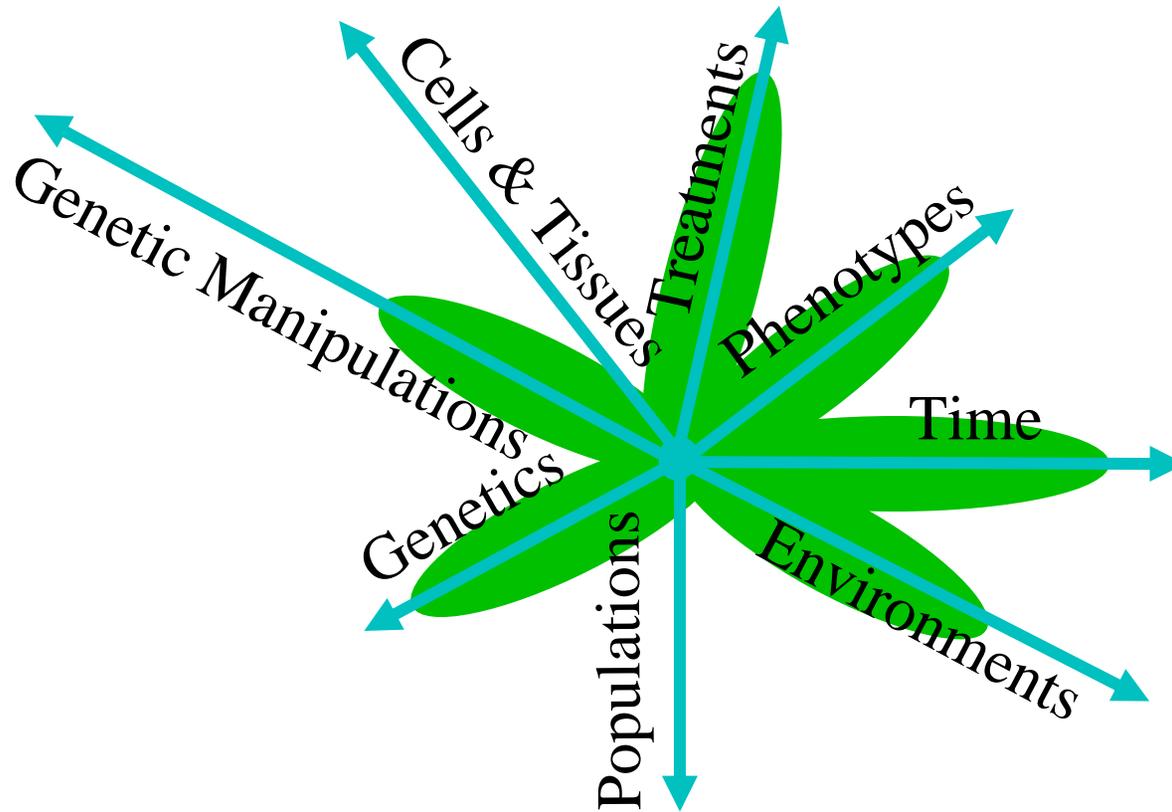
- Dynamic modeling and simulation of microbial cells and microbial communities
- High-throughput techniques for data collection, storage and analysis, and model building
- Structural modeling and simulation
- High-end biological computing and mathematics infrastructure

Findings and Recommendations

- High-throughput techniques for data collection, storage and analysis, and model building
- Structural modeling and simulation
- Dynamic modeling and simulation of microbial cells and microbial communities
- High-end biological computing and mathematics infrastructure

High Dimensional Spaces

Complexity Challenge



The experiment paradigm is changing to statistically capture the complexity. We will get **maximum value when we explore three or more dimensions in a single experiment.**

Mass Spectrometry Analysis

Scale Challenge

Now

Analyze proteome of *Rhodospseudomonas palustris*
35,000 MS/MS spectra against 4000 proteins- .01 TF for a day

In 2-5 years

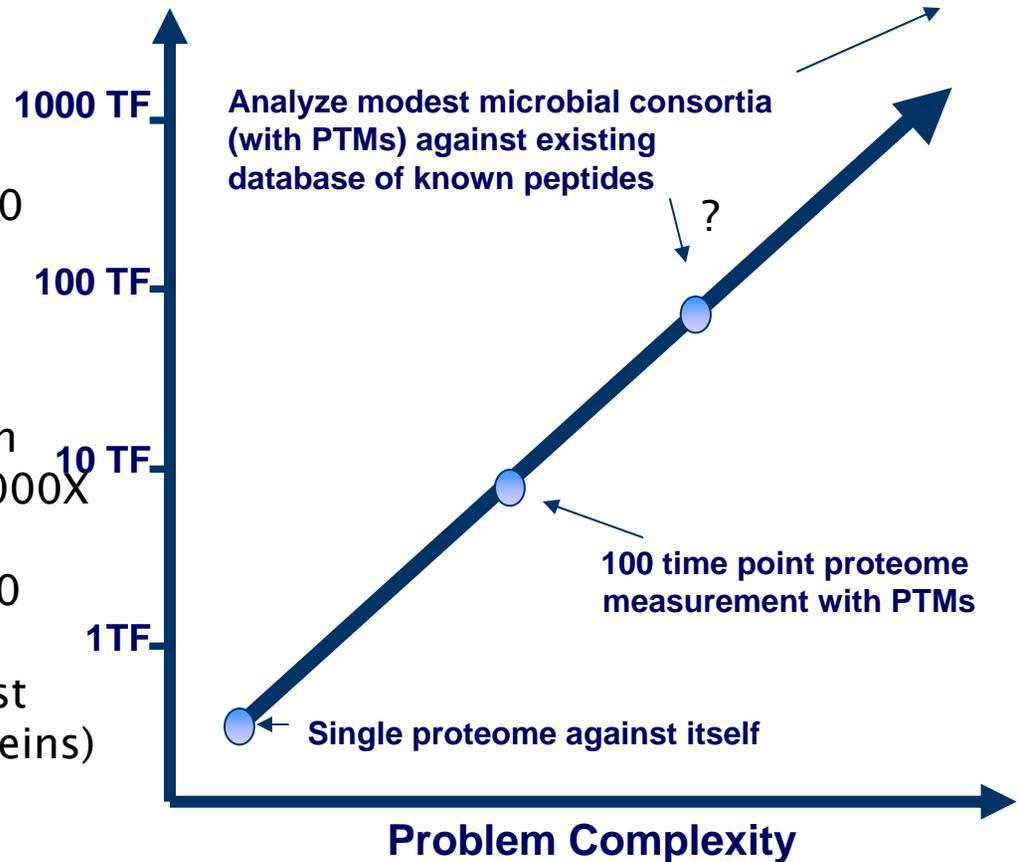
Measure 100 time points with protein post-translational modifications - 1000X

Examine microbial consortia with 100 proteomes (P) with Post-translation modifications (PTM). Compare against existing proteomes DB (4×10^6 proteins)

$$100 * 10 * 1 * 1000 = 1,000,000X$$

$$C \sim P * PTM * t * DB$$

Also need capability to move Petabyte datasets to computing site



- Need significant algorithm improvements
Plus >100 TB shared memory capability

Molecular Interaction Image Analysis

Scale Challenge

Now

Single event detection
(find face in picture)
Recognize FRET event
in a microbial cell- single processor
problem

In 5 years

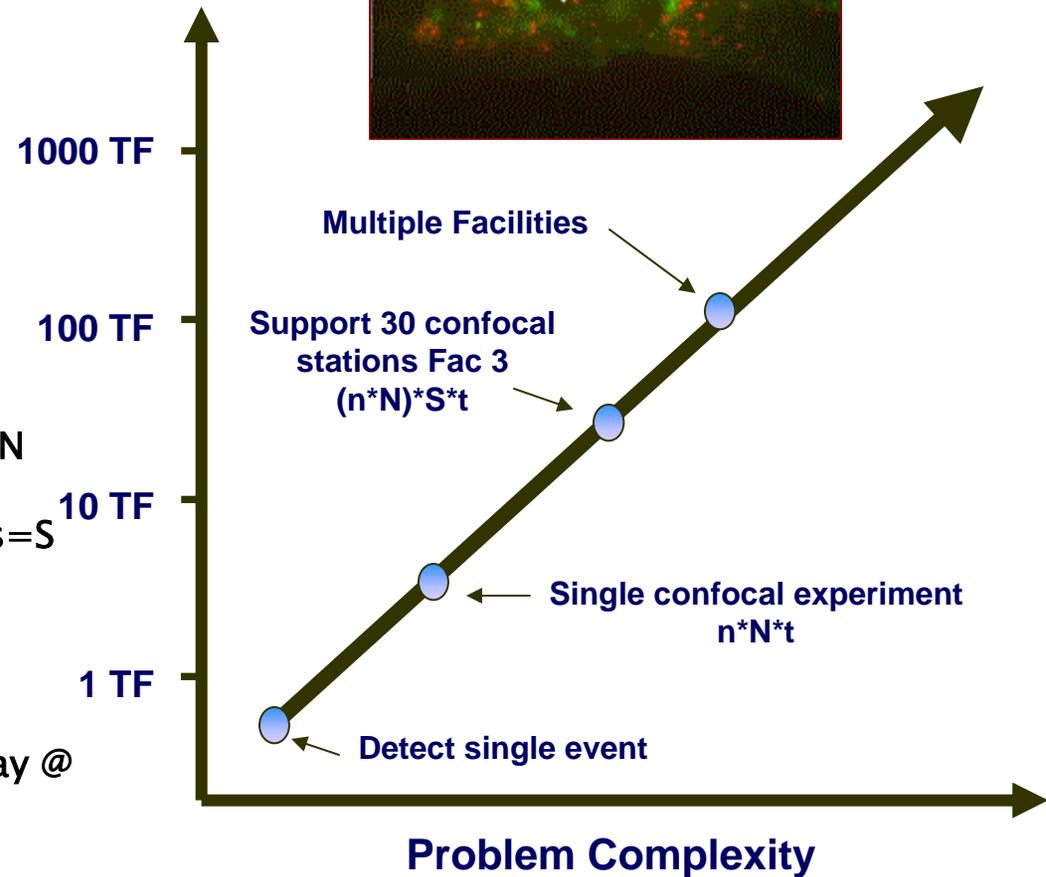
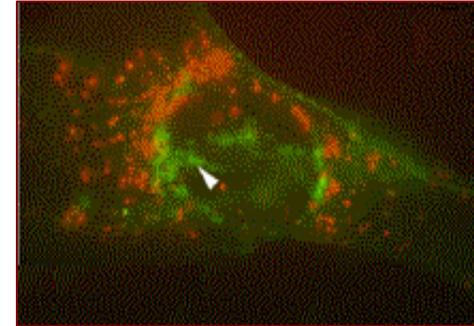
Track 100 proteins in microbe = n
Dynamic segment cell ultrastructure = N
Track for 100 time points = T
Interpret output of 30 confocal stations = S

$$C \sim (N \cdot n) \cdot S \cdot T$$

$$(100 \times 100) \times 100 \times 30 = 3 \cdot 10^7$$

Need to generate 10,000 images per day @
4 megabyte each

Query and retrieval issues
Infrastructure for moving Image Datasets



- Need sustained 100 TF capacity
- Need significant algorithm development

Comparative Genome Analysis

Scale Challenge

Computing the optimal phylogenetic tree based on the entire genome of 10 species will remain intractable even with peta-scale computers

Complexity

Maximum likelihood

$C \sim n^2 * m$ where

n is number of contemporary species

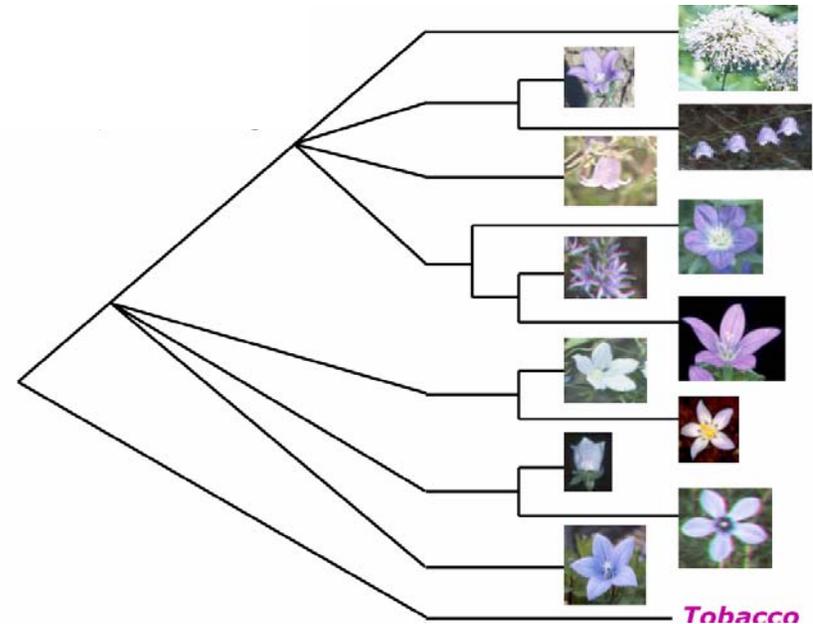
m is size of the genome in question

Requirements

For 10^6 species on a gene sequence of about 1000 pairs or 1000 species on the entire genome:

Runtime: 1-3h at 1P flop

Memory: ~3 TB



Multiple Sequence Alignment

Scale Challenge

Complexity

Run-time of dynamic programming solution
= $O(2^k * n^k)$

where n = length of each sequence
 k = number of sequences

Space, $O(n^k)$, is prohibitively large!

Example:

6 sequences of length 100 \Rightarrow
 6.4×10^{13} calculations!

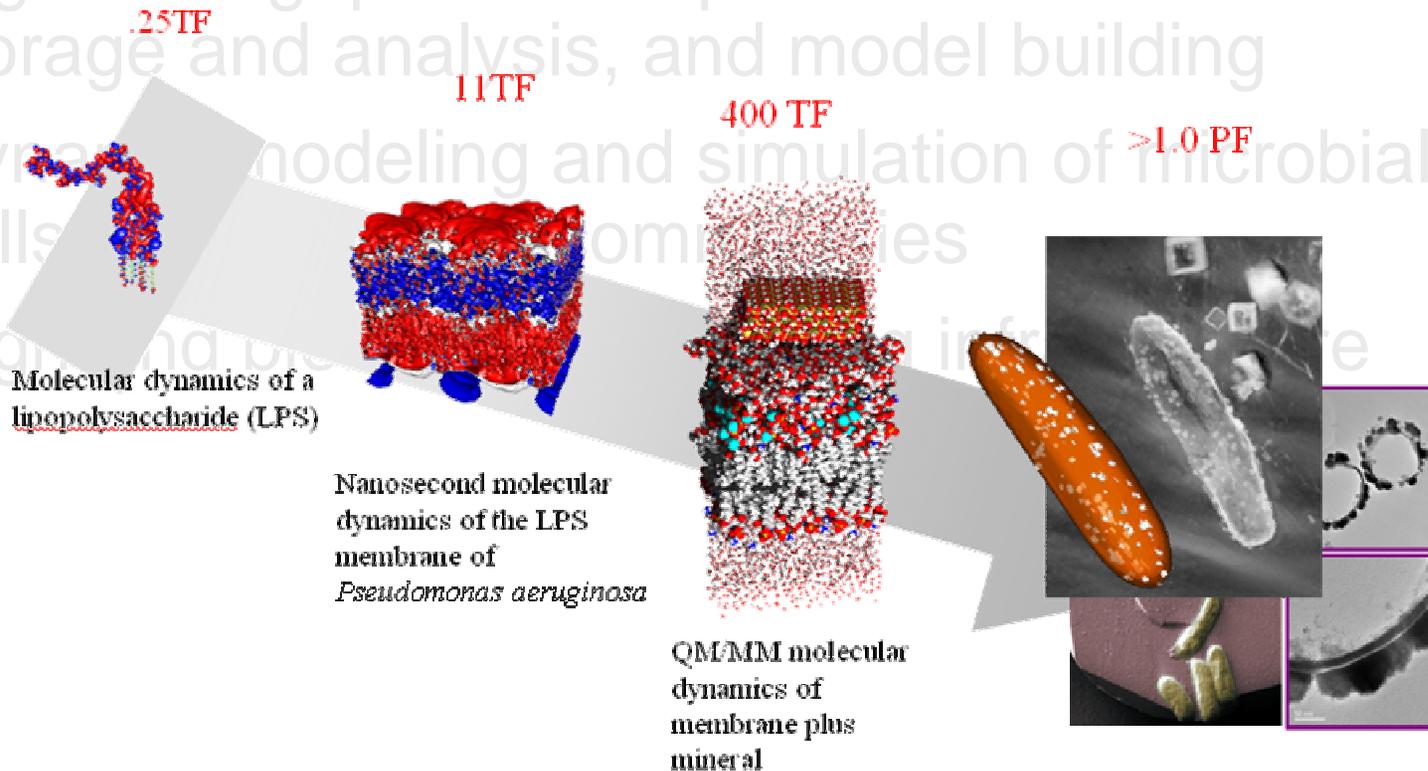
```
ARAGEEGRGFSVIADDEVRS LAAQSAEATAAMEALIVTI 120
ARAGEQGKGF SVVAEEVRKLAQDSQAATQQVNAILGDI 373
ARAGEHKGKGF AVVADEVKRLAEQSRQSSSEVSNIVKNI 260
ARAGESGKGF SVVANEIRKLATNSKENVSQINDITNTI 235
SRAGEKKGKGF AVVANEVRKLADQTKASTNTVSQ LIEKT 351
:**** *:***:***:***:***:*.** ::      :. :
```

Recommendations

- Development of scalable bioinformatics technologies.
 - Must be able to handle large amounts of (possibly heterogeneous) data derived from high-throughput genomics, proteomics and structure determination efforts.
- Development of methods for automated data analysis and model generation.
 - Should enhance our knowledge of genomics (sequence), active site motifs (structure) and metabolic/regulatory pathways (network).
- Development of shared, common testbeds for regulatory-network prediction
 - Along the lines of the highly successful CASP experiments for protein-structure prediction.
 - To accelerate progress in bioinformatics research.

Findings and Recommendations

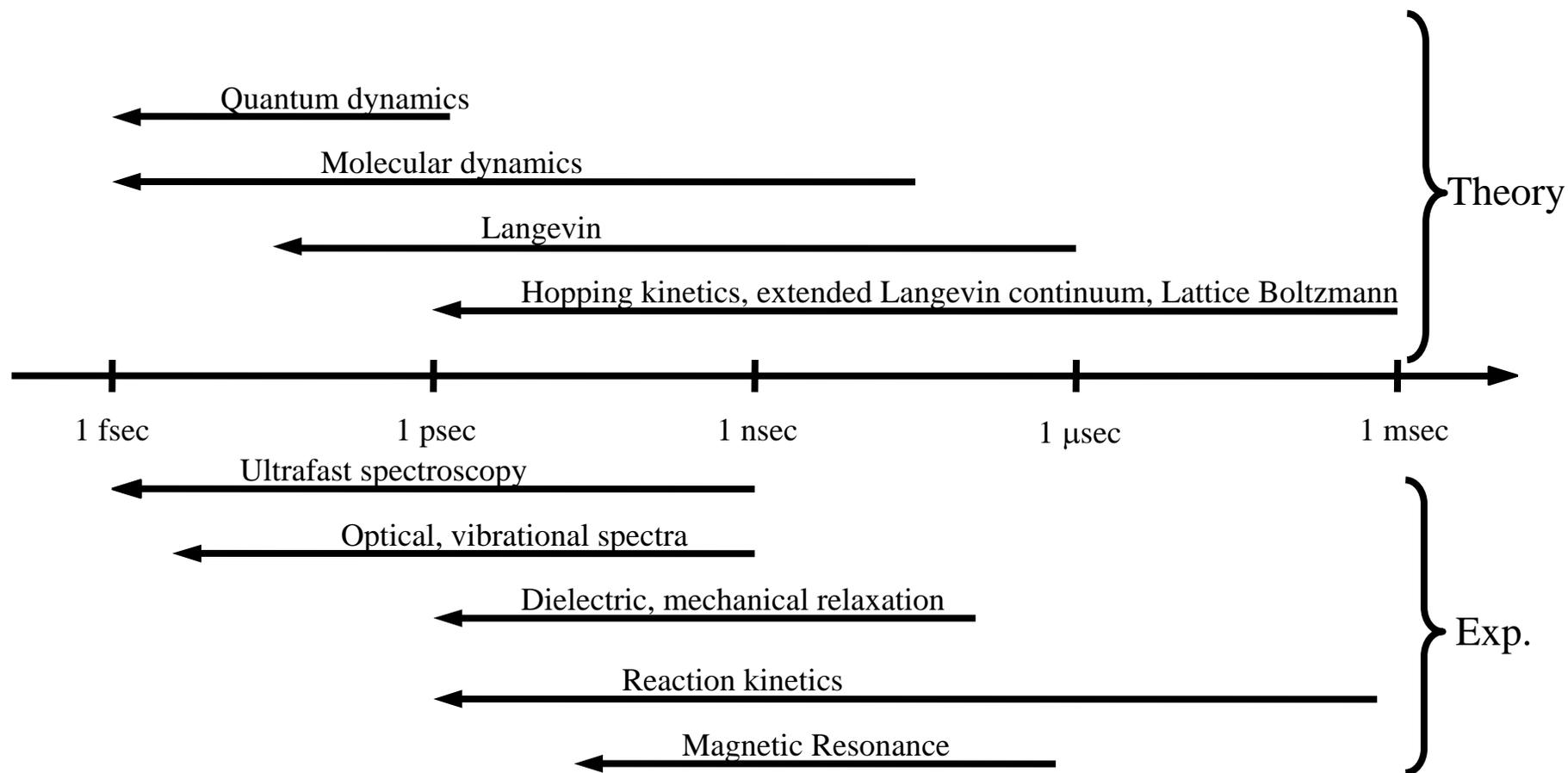
- Structural modeling and simulation
- High-throughput techniques for data collection, storage and analysis, and model building
- Dynamic modeling and simulation of microbial cells
- High-resolution microscopy



Energy transduction across membranes is still out of reach at 1.0 PF

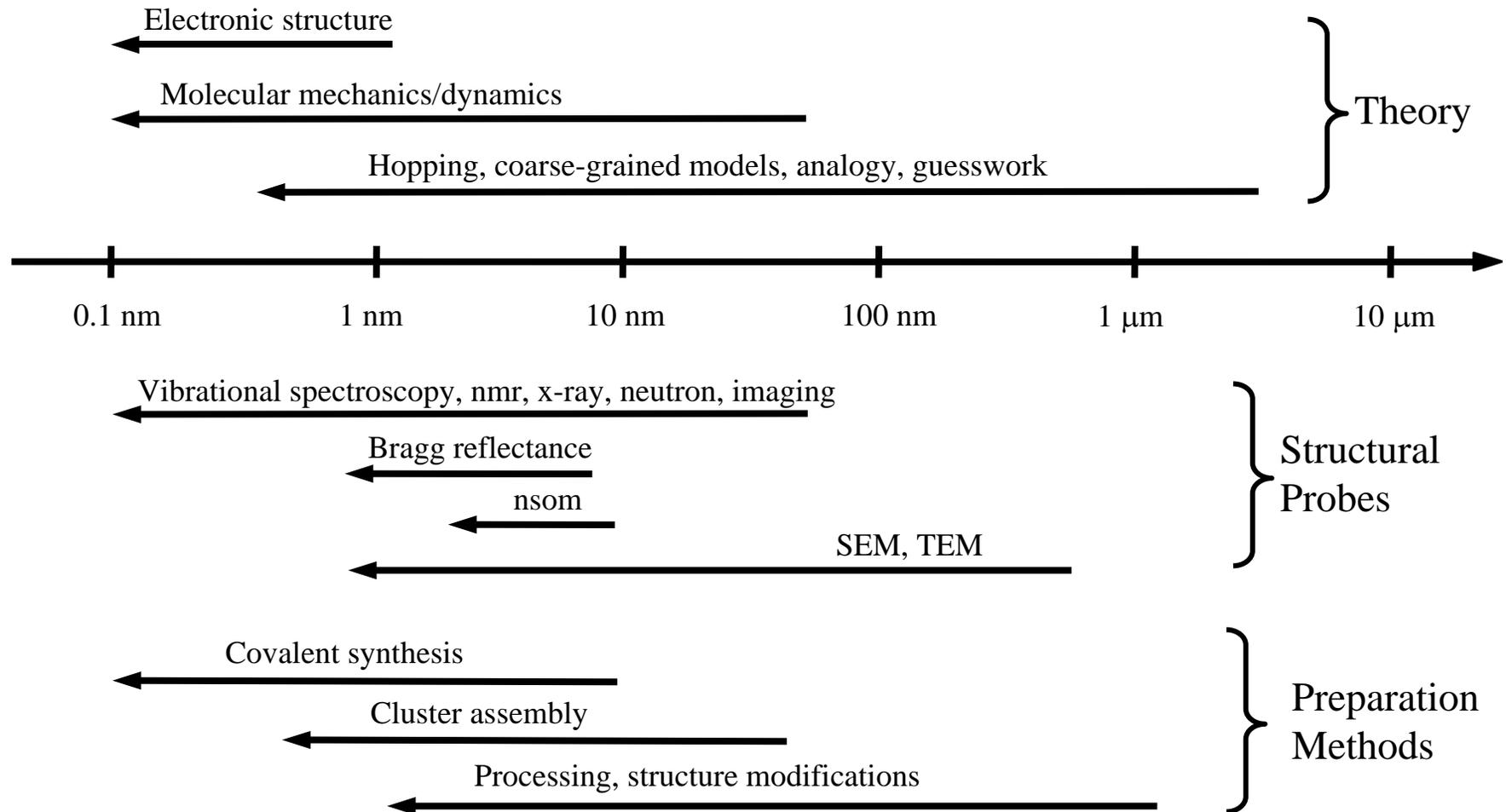
Temporal Scales

Complexity Challenge



Spatial Scales

Complexity Challenge



Molecular Docking and Dynamics

Scale Challenge for Docking

Now

Rigid docking of Protein complex with ligand

$C \sim n * (t + F)$ where

n is number of possible orientation

t is time to find a possible orientation

F is time to fit ligand

1 TF day

In 5 years

Multi-component docking

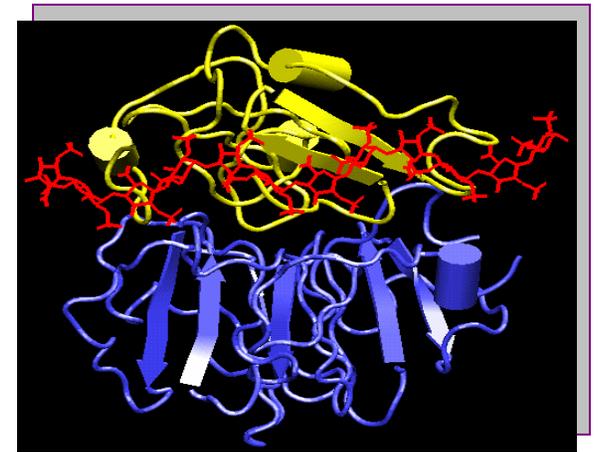
$C \sim (N-1)! * n * (t + F)$ for **N=6**

-> 120 TF day

Goal - do a microbe (>200 complexes) in year 1

Need to use experimental constraints to limit **(N-1)!** and **n** terms

Constraints come from mass spectrometry and imaging



Atomistic Modeling of Proteins using MD

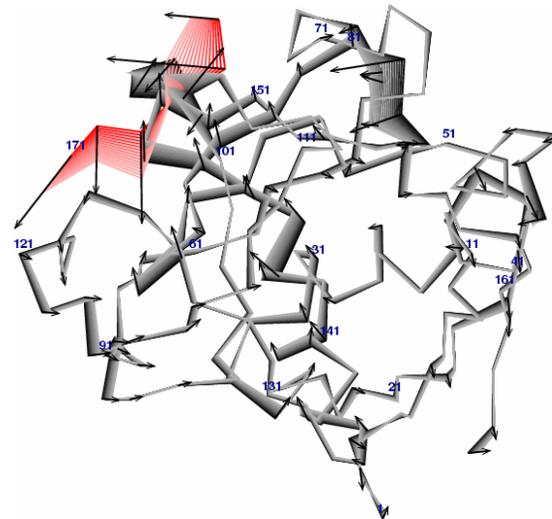
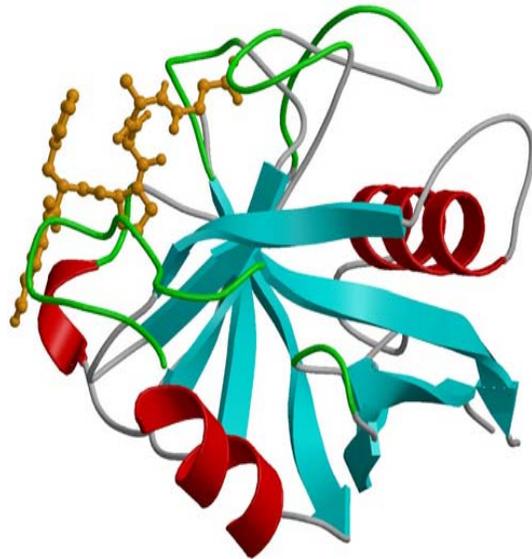
Scale Challenge

Protein/DNA complexes: structure, folding, dynamics
biochemical function & biomolecular recognition

Current: 100,000 atoms **Real systems:** >1,000,000 atoms

Current: $10^{-9} \sim 10^{-6}$ seconds **Real activity:** $10^{-3} \sim 1$ seconds

> 10^6 difference in computing power available & required



Protein Folding

Scale Challenge

Ab initio protein folding for ~32,000 atoms will require 1000 teraFLOP processors, with a bandwidth of 10^9 and a latency of $1/10^9$

Description	Count*	Comment
Atoms	~32,000	300 amino acid protein + water
Force evaluations / time step	10^9	Pairwise atom - atom interactions
FLOPs / force evaluation	150	Typical molecular dynamics
FLOPs / time step	1.5×10^{11}	
Each time step	$\sim 10^{-15}$ s	1 - 5 femto second
Total simulation time	10^{-3} s	Protein folds in ~1 milli second
Total time steps	2×10^{11}	
FLOPs / simulation	3×10^{22}	Total FLOP/s to fold a protein
Execution time	3×10^7 s	1 year
Required FLOPS	$\sim 1 \times 10^{15}$	1 Petaflop/s

Estimate is conservatively based on quadratic algorithm.

Better algorithms will reduce (somewhat) running time, but usual surprises will increase it!

And good science will require multiple simulations

Source: David Klepacki, IBM

Recommendations

- Develop the methodology for connecting low-resolution simulations of small protein backbones (3Å rmsd) and detailed full-atom simulations with the advanced force fields.
- Development of advanced methods for predicting the geometry of bound protein complexes, particularly multi-protein molecular machines.
- Development of multi-scale approaches to molecular modeling, such as hybrid molecular mechanics, quantum mechanics (QM/MM), to investigate biomolecular structure, dynamics and function.
- Development of advanced analysis infrastructures for globally distributed molecular trajectories.

Findings and Recommendations

- High-throughput techniques for data collection, storage and analysis, and model building
- Structural modeling and simulation
- **Dynamic modeling and simulation of microbial cells and microbial communities**
- High-end biological computing infrastructure

Whole-cell Simulation

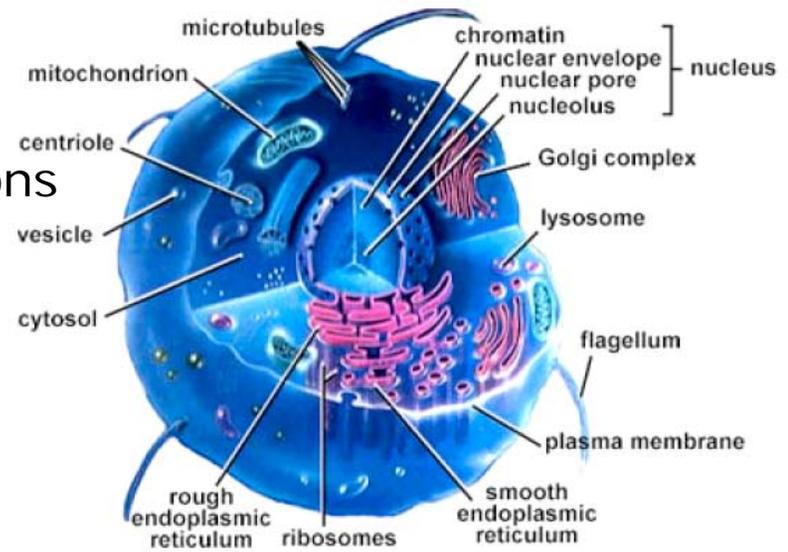
Scale Challenge

Varying timescale

- 10^{-15} s: 3D continuum transport
- 10^{-6} s: signal transduction
- 10^{-3} s: metabolic pathways
- 10^1 - 10^4 s: effects of gene-expressions

Stochastic Simulation

Model 10^{14} biochemical reactions in *E.coli* will take 4^{1024} s = 12 years on a single processor



Recommendations

- A program of research aimed at accelerating the development of high-fidelity models and simulations of metabolic pathways, regulatory networks, and whole-cell functions.
- Development of resources for large-scale analysis and comparison of biological pathways that provides advanced computational capabilities
- Development of simulation platforms and frameworks for studying microbial communities using detailed multiscale network models.

Findings and Recommendations

- High-throughput techniques for data collection, storage and analysis, and model building
- Structural modeling and simulation
- Dynamic modeling and simulation of microbial cells and microbial communities
- High-end biological computing and mathematics infrastructure

Recommendations

- Development of middleware solutions that allow simultaneous access to a large number of massive, distributed data repositories
- Software development efforts focused on emerging computer architectures
- Efficient systems for storage, retrieval and organization of massive biological datasets
- A multiscale modeling framework for structural and systems biology
- Advanced multiscale methods for solving differential equations and modeling complex systems
- Efficient methods for non-convex energy optimization in systems with many degrees of freedom
- Scalable methods for the integration of stochastic differential equations
- Development of open access and community wide repository (similar to Protein Data Bank) for sharing and archiving computational software and data

See also “Data Management” Report

It is all about the **scale** and the **complexity**

