Advanced Computational and Modeling Needs in Biological Sciences

Michael Colvin
Biology & Biotechnology Research Program
Lawrence Livermore National Laboratory

On Detail to

DOE Office of Biological and Environmental Research

Outline

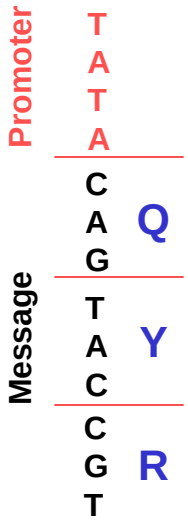
I. Computation and the biological sciences
II. Examples of challenges in computational biology
III. Computational Components of Genomes to Life Project
IV. Concluding comments

Computational analysis and simulation have important roles in the study of each step in the hierarchy of biological function

DNA sequence

Sequence Annotation

Protein sequence and regulation

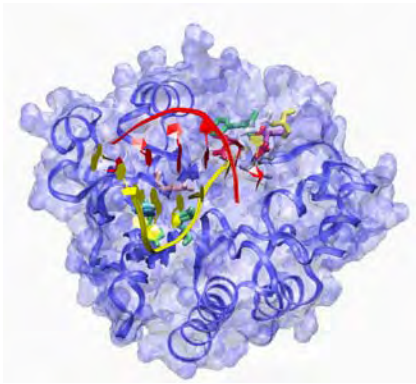


Homology based protein structure prediction

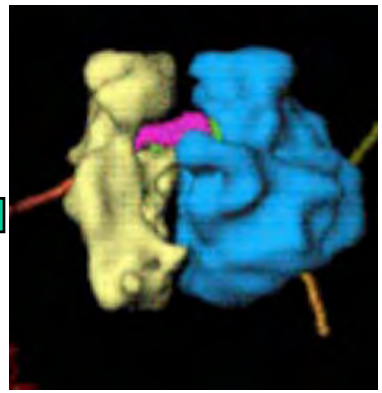
Protein structure

Molecular modeling

Protein/enzyme function

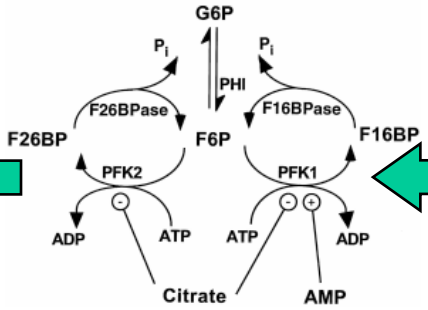


Network elucidation



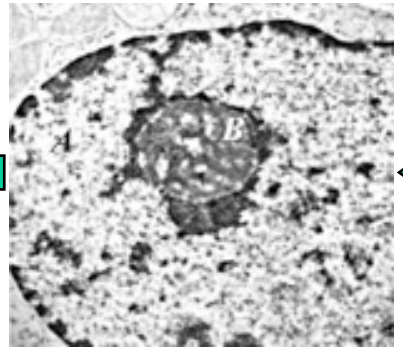
Multi-protein machines

Pathway simulations

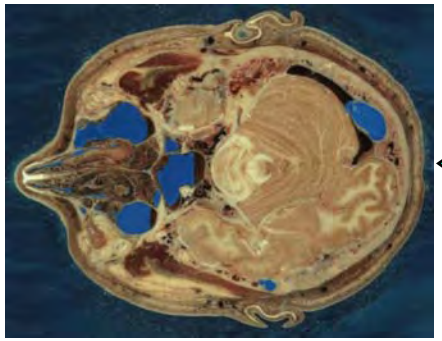


Metabolic pathways & regulatory networks

Organism simulations



Bacteria and cells



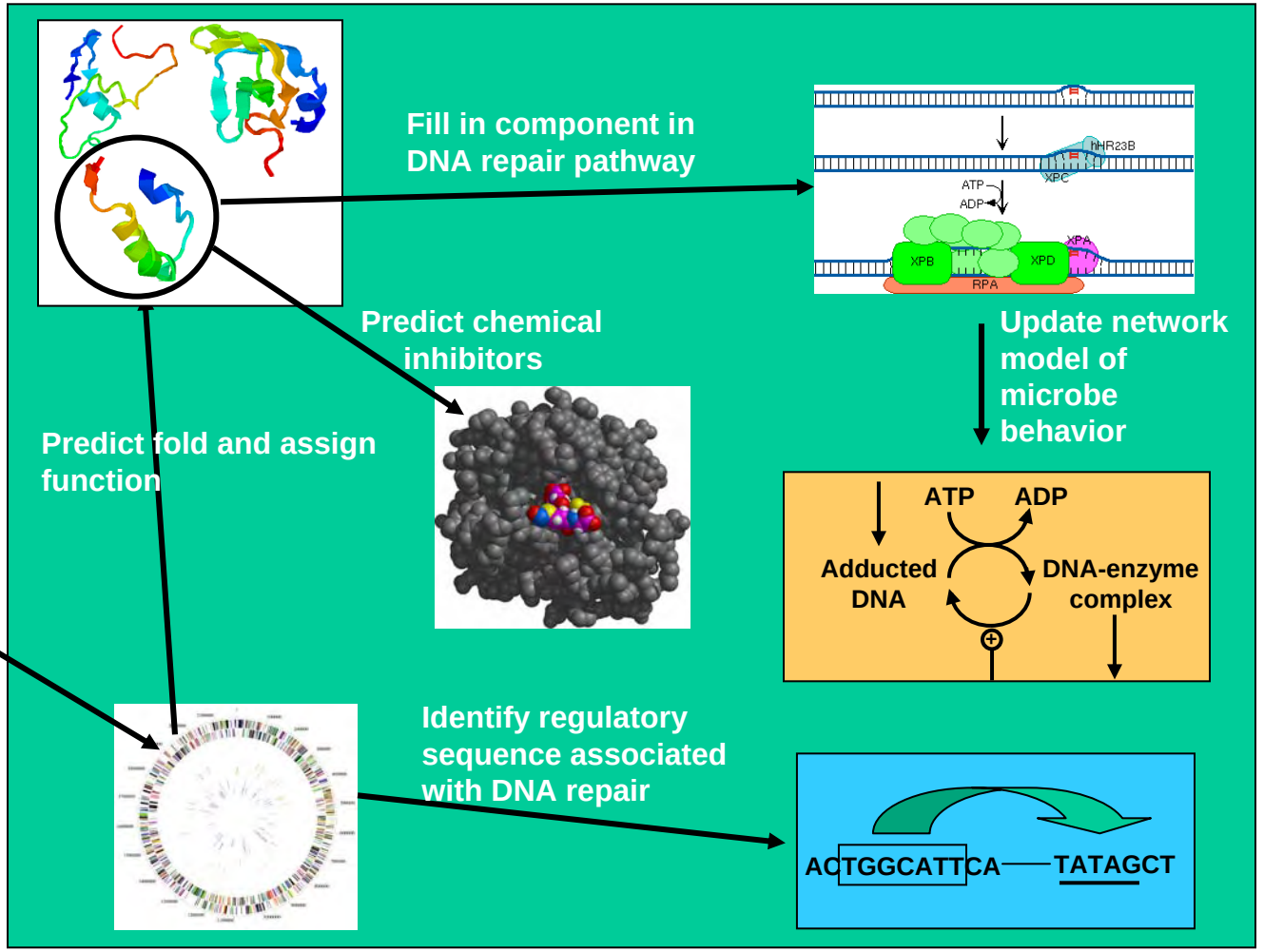
Bacterial communities & multicellular organisms

Overall vision for biology: Provide a framework for integrating new biological data to create new understanding

Hypothetical example:

New data showing that a gene forms complex with DNA repair enzyme

Add function to genome database

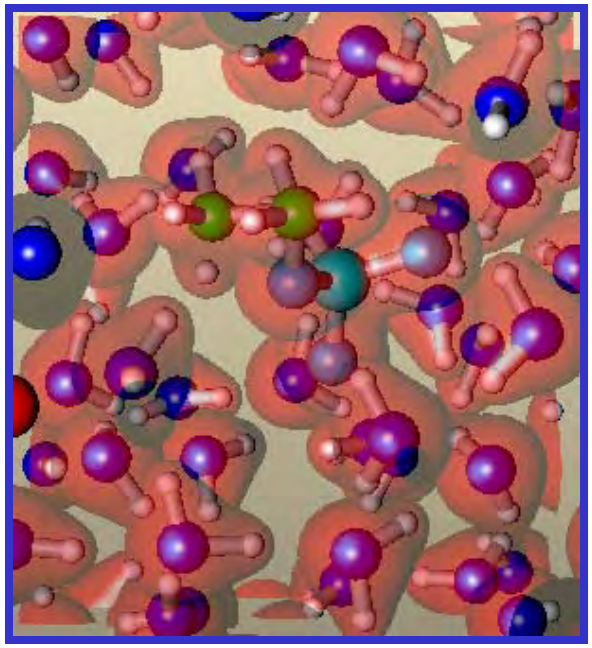


Example 1: Predictive molecular simulations of most biochemical processes require new algorithms and computers

E.g. First principles dynamical simulations of enzyme activity

State-of-the-art

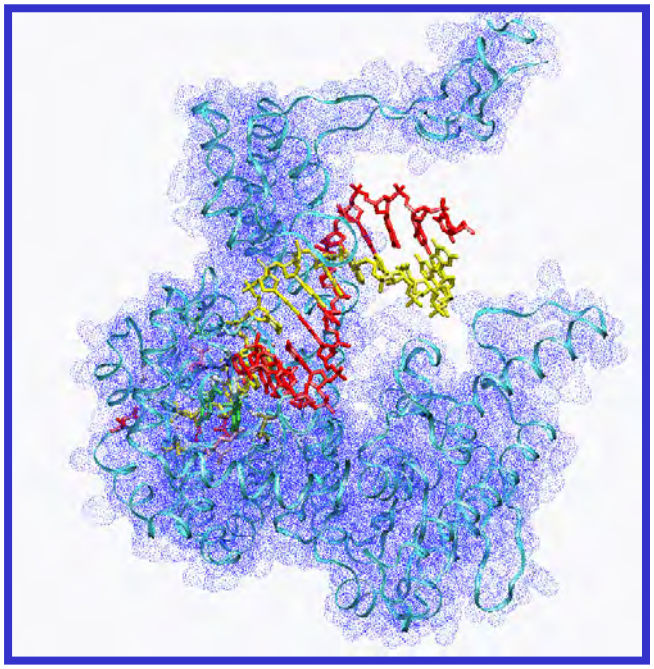
(DNA fragment in water)



~600 atoms for 10^{-12} seconds
(3840 processors for 12 days)

Long-term goal

(DNA replication machine)



~100,000 atoms for 10^{-3} seconds

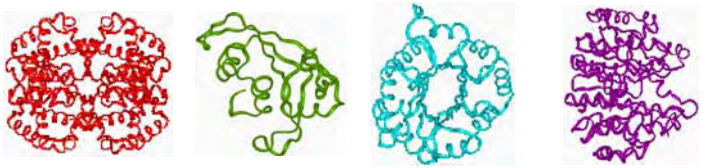
Need 11 orders of magnitude improvements in computers and algorithms

Example 2: Homology-based protein "threading" will allow structural characterization of many newly sequenced genes

Protein sequence:

CLVASLDNVRNLFTVDKAIH...

"thread" onto templates



Fold recognition methods

evaluate fitness

Best match

State-of-the-art

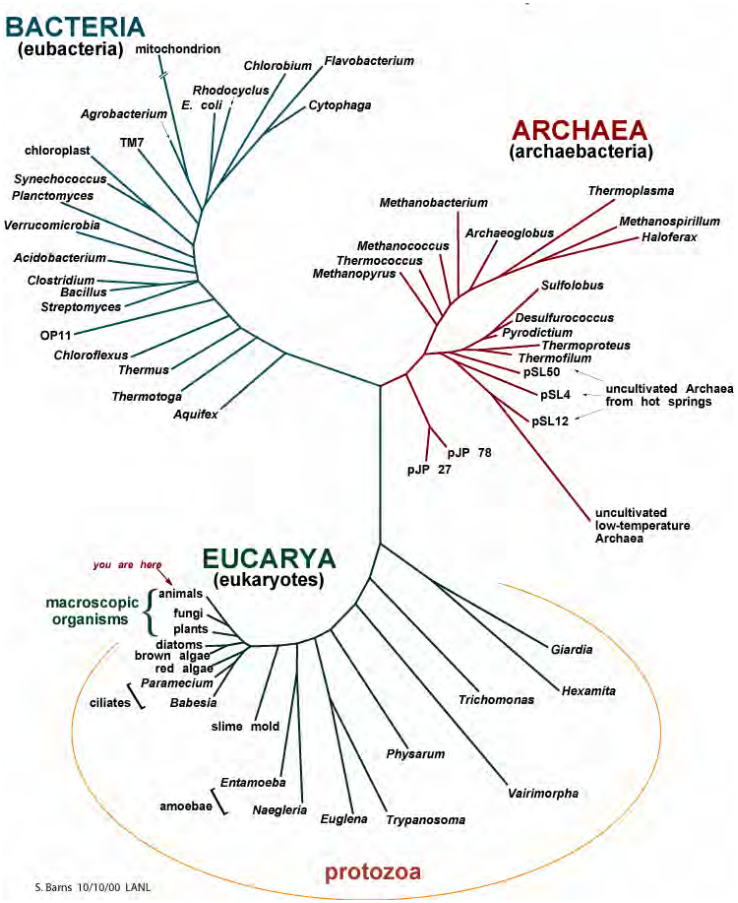
- Threading 5000 genes/day on teraFLOP computer
- Structural assignments possible for ~30% new genes

Research needs

- New algorithms for finding more distant homologies
- High-thoughput structural refinement methods
- Automated management of predicted structures databases

Example 3: New computational algorithms are needed to assemble rigorous phylogenetic trees from full genome data

Goal: Build tree structure of evolutionary relationships involving minimum number of DNA changes



State-of-the-art:

Using NP-hard methods:

~10 bacteria

~3000 genes

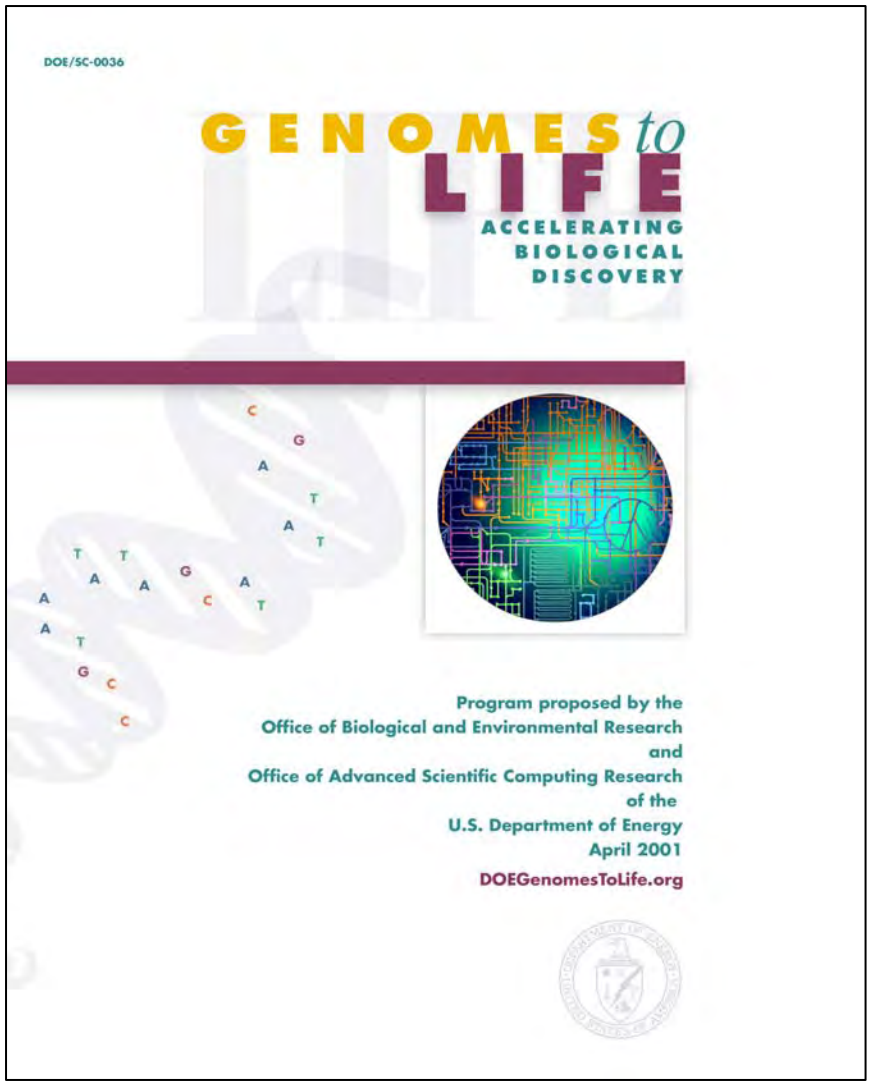
Research needs:

Rigorous trees for:

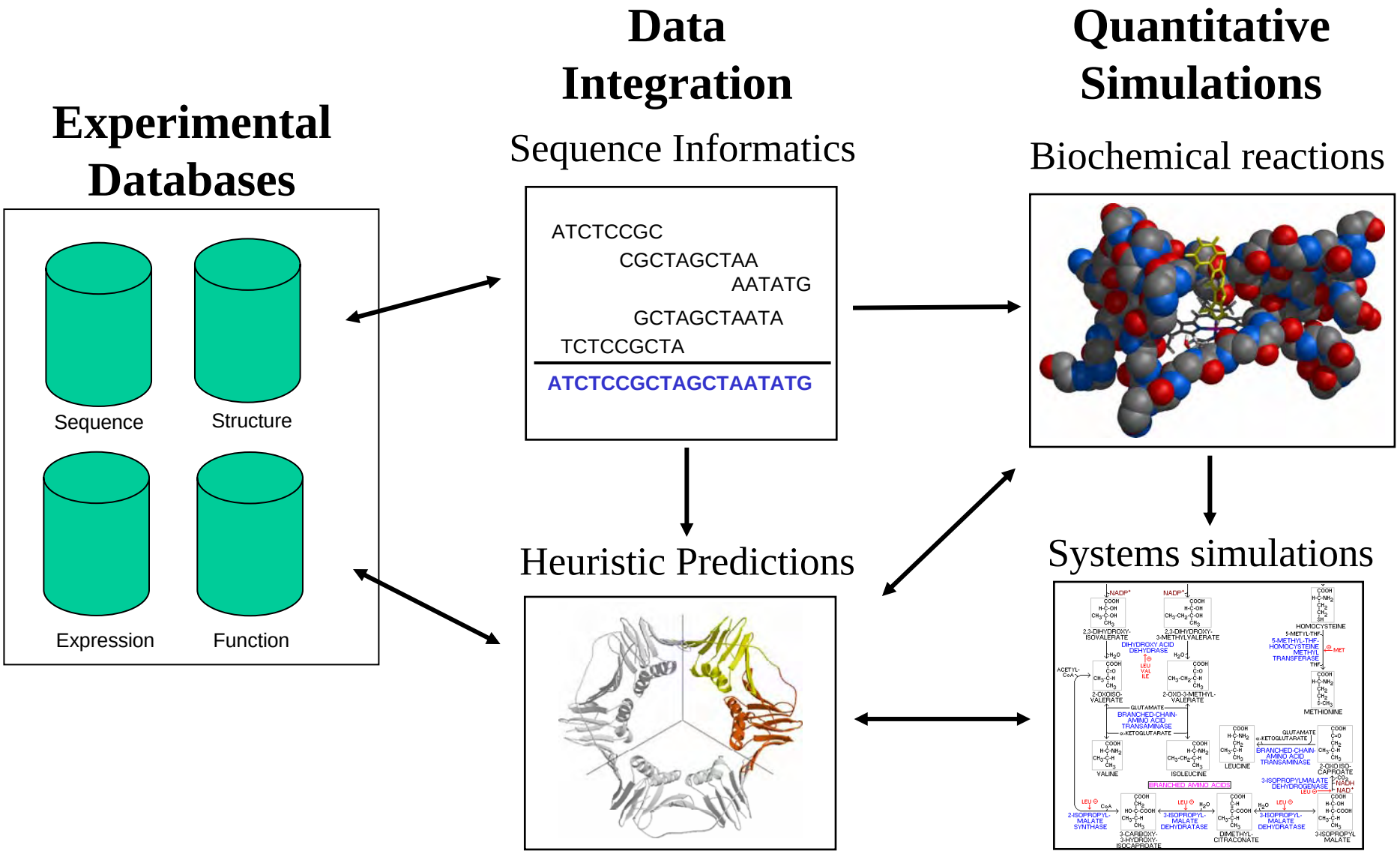
100's of organisms

Improved heuristic algorithms

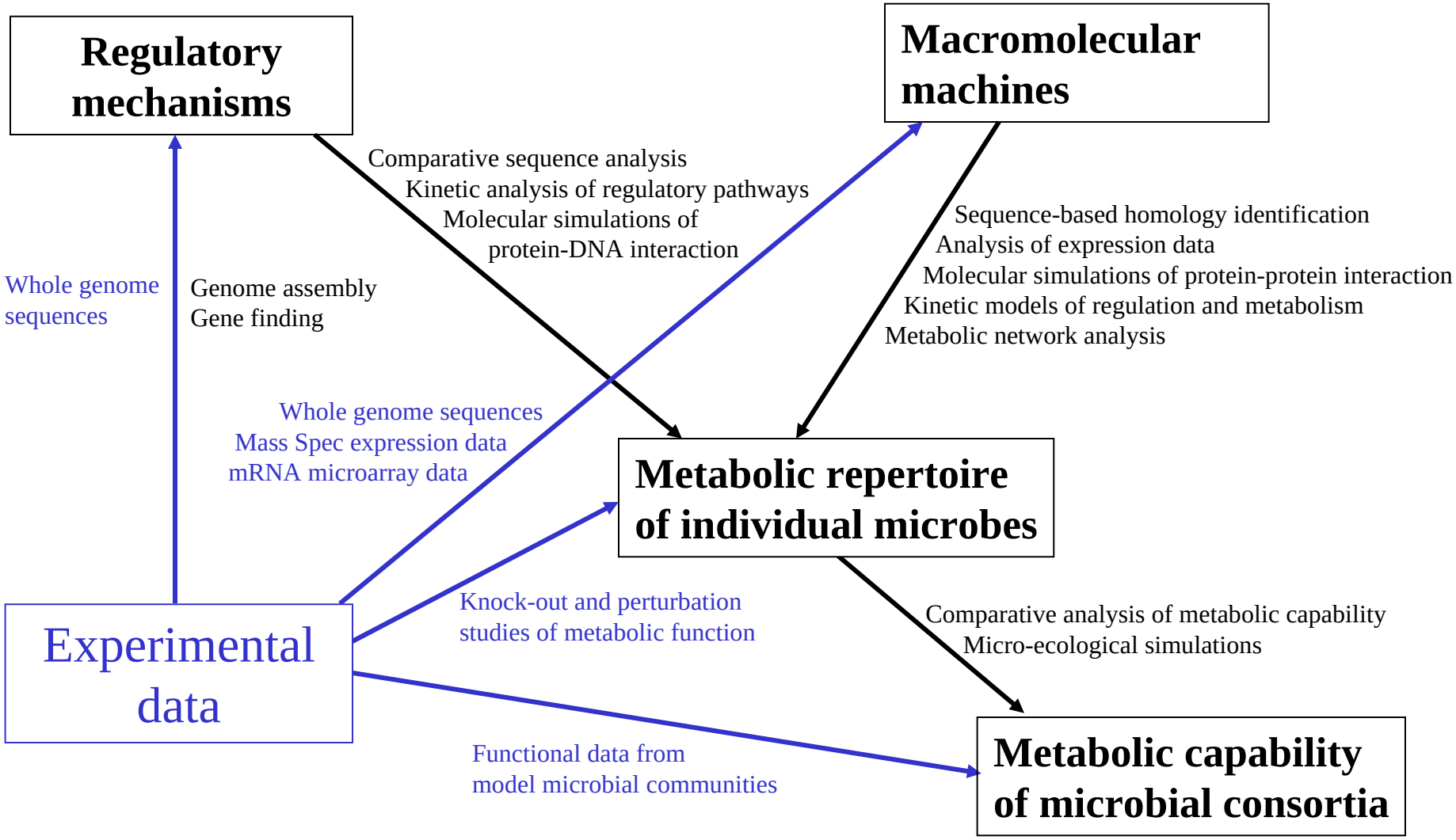
New developments in the computational sciences are central to the Genomes to Life Initiative



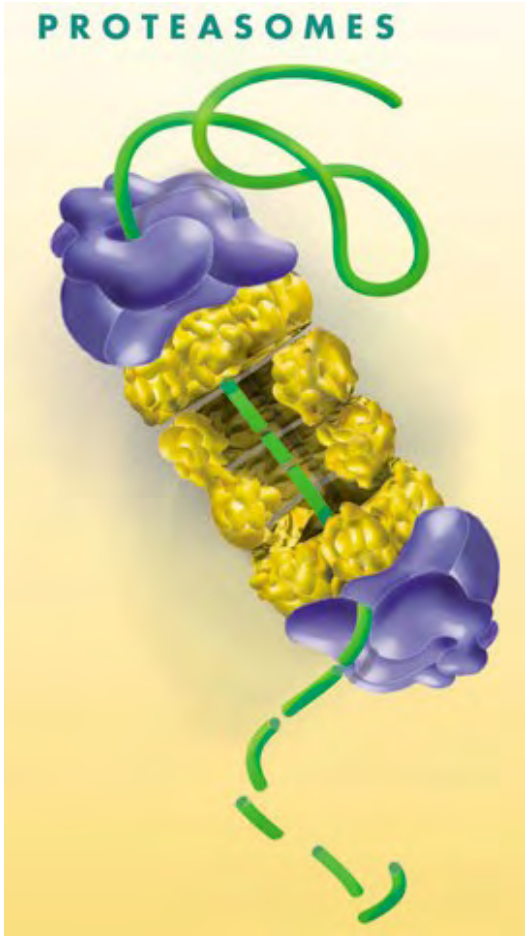
Genomes to Life describes a new form of biological science that is heavily dependent on computations



Computations provide the linkage between levels of biological description involved in the GTL initiative



GTL Goal 1: Identify the molecular machines of life

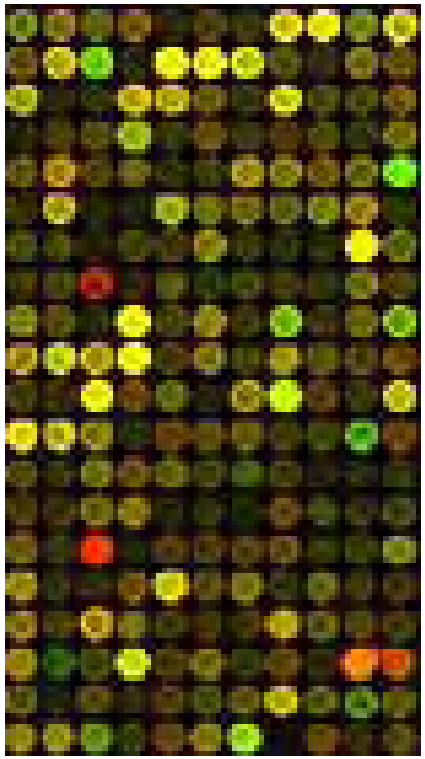


Computational Needs

- Improve bioinformatics methods needed to analyze experimental protein expression data
- Adapt and develop databases and analysis tools for integrating experimental data on protein complexes
- Develop algorithms for integration of diverse biological databases and provide functional and structural annotations of protein-sequence data.
- Develop modeling capabilities for simulating the function of multiprotein machines in cell networks and pathways.

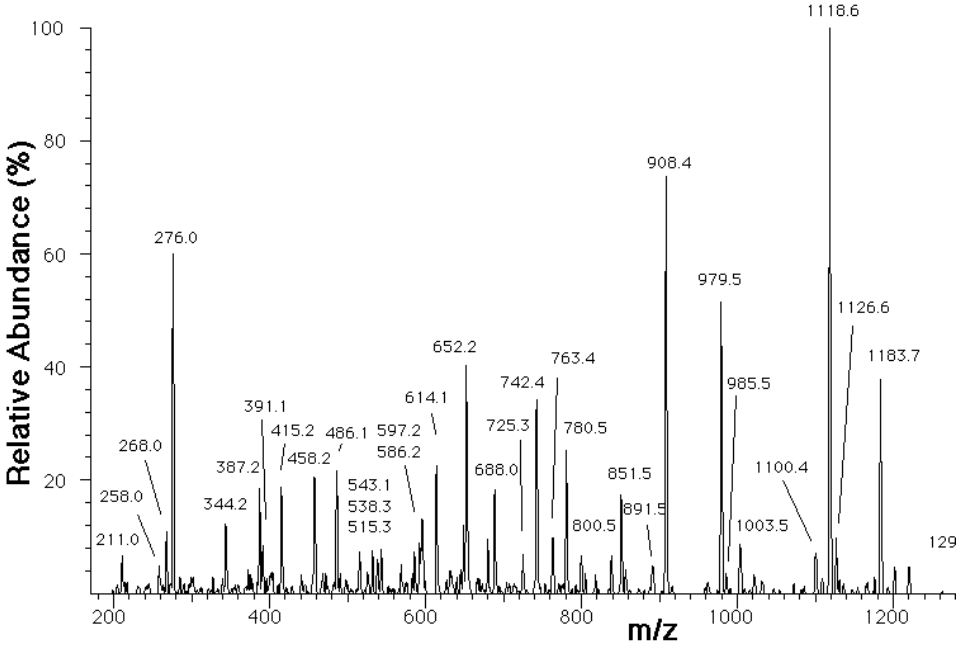
Advanced algorithms will be required to effectively analyze protein expression and interactions data

DNA microarray



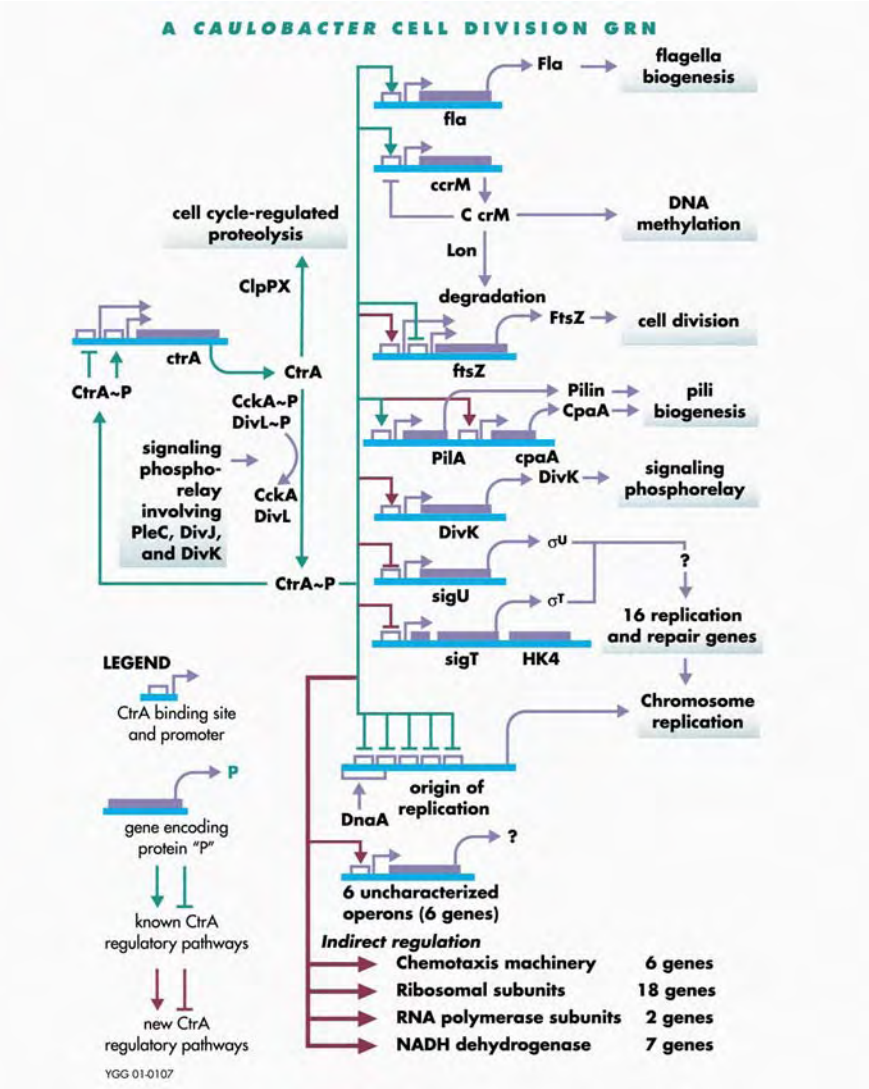
mRNA expression levels as a function of time and conditions

Mass Spectrometry



Identify proteins and expression levels based on mass

Goal 2: Characterize gene regulatory networks

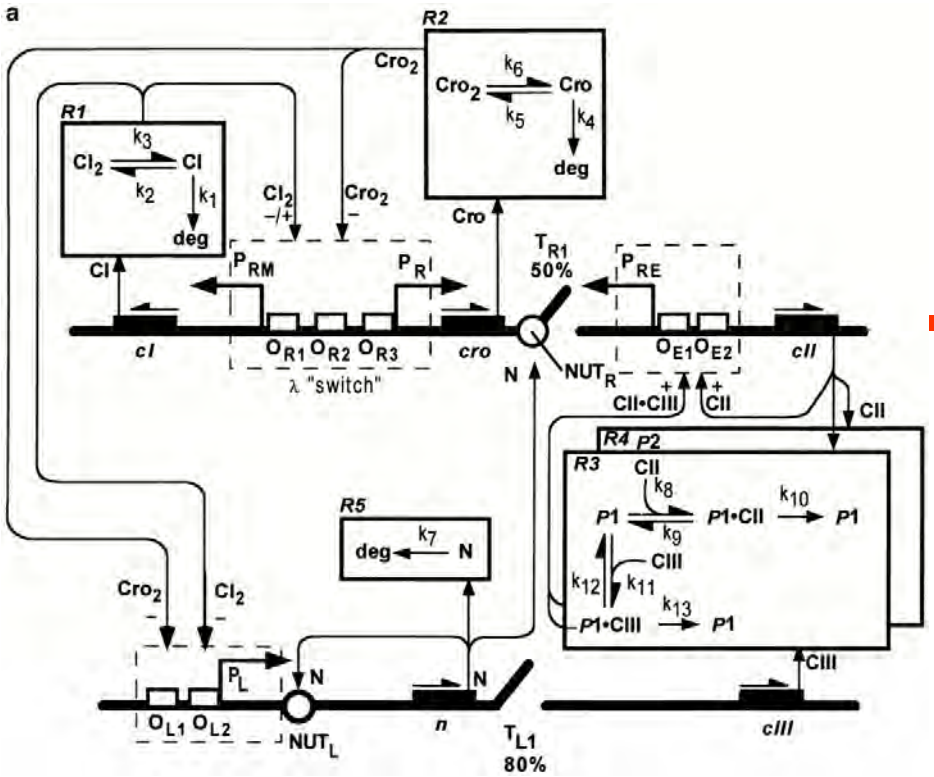


Computational Needs

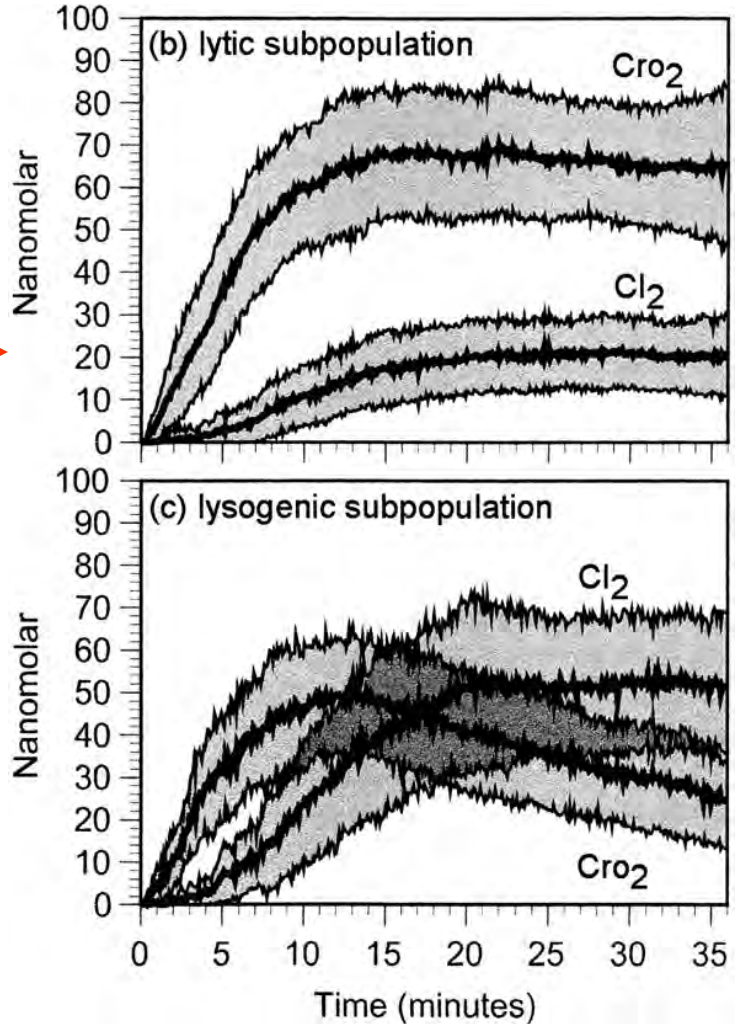
- Extract regulatory elements, including operon and regulon sequences, using sequence-level comparative genomics.
- Simulate regulatory networks using both nondynamical models of regulatory capabilities and dynamical models of regulatory kinetics.
- Predict the behavior of modified or redesigned gene regulatory networks.

Kinetic modeling can predict quantitative behavior of well-characterized regulatory networks

Circuit model of *e. coli* phage-λ lysis vs. lysogeny decision circuit

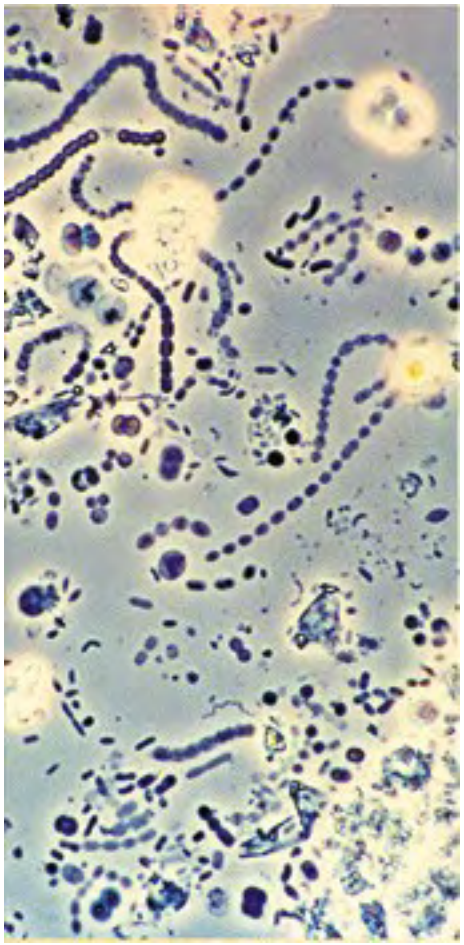


Signal concentrations in lytic and lysogenic sub-populations



Arkin, Ross, and McAdams, *Genetics* 149, 1633 (1998)

Goal 3: Characterize functional repertoire of microbial communities



Computational Needs

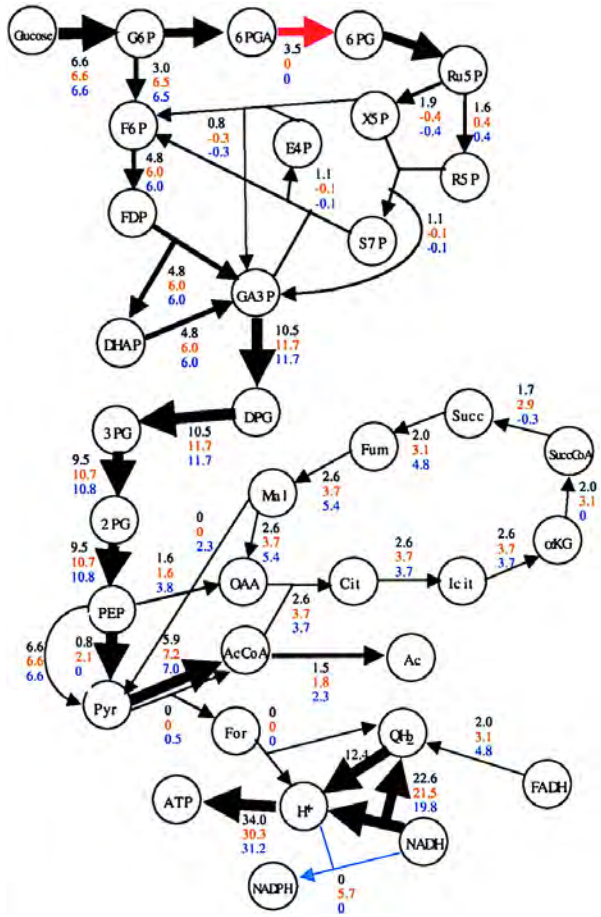
- Deconvolute mixtures of genomes sampled in the environment and identify individual organisms.
- Facilitate multiple-organism shotgun-sequence assembly.
- Improve comparative approaches to microbial sequence annotation and gene finding and use them to assign functions to genes.
- Reconstruct pathways from sequenced or partially sequenced genomes and evaluate the combined metabolic capabilities of heterogeneous microbial populations.
- Integrate regulatory network, pathway, and expression data into integrated models of microbial community function.

Network modeling and flux-based analysis provides semi-quantitative information about a microbe's metabolic capability

Stoichiometric matrix involving
426 metabolites x 720 genes

	Glu	G6P	6PGA	F6P	FDP	PEP	Fum	αKG	Vit	Pyr	•	•	•
aceA	5	-	-	-(6)-		-	-	-	-				
aceB	-	-	-	-	-	1	-	-	-				
ackA	2	-	-	-	-	-	4	-	-				
acnA	-	-	2	-	-	-	-	1	-				
atpH	-	1	-	-	-	-	-	-(2)-					
cydA	-	-	-	-	-	1	-	-	-	-	•	•	•
cydB	-	-	-	1	-	-	-	1	-	-			
cyoA	-	-	2	-	-	-	-	1	-				
cyoB	2	-	-	-	-	-	-	-	-				
fdnG	-	-	-	-	-	1	-	-	-				
ilvE	-	-	-	-	-	-	4	-	-				
leuA	1	-	2	-	-	-	-	1	-				
leuB			•										
•			•										
•			•										
•													

Predicted metabolic fluxes
in 3 *e. coli* genotypes

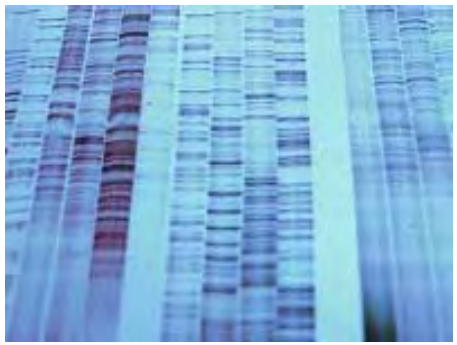


Edwards and Palsson, *PNAS* 97, 5528 (2000)

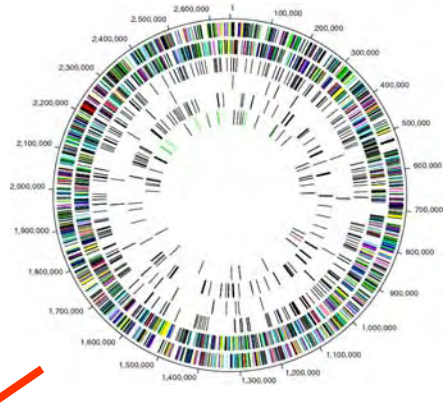
All models of microbial behavior will depend on information about the macromolecular machines that mediate function

Computation has roles at each step in deriving this data

Raw DNA sequencing reads

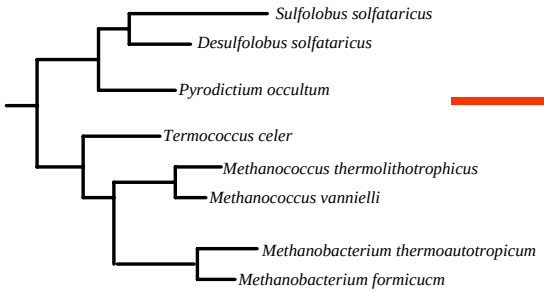


Assembled genome

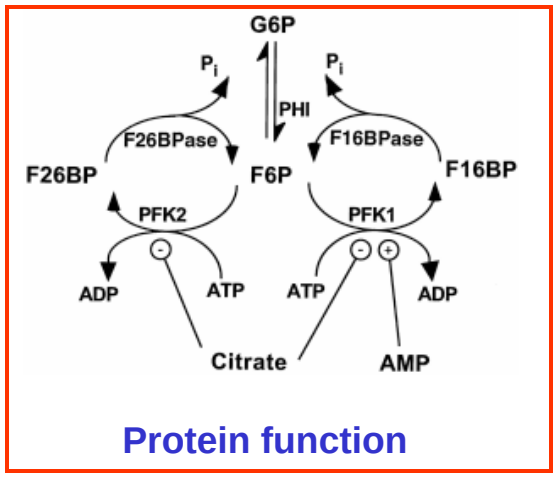


Sequence homologues

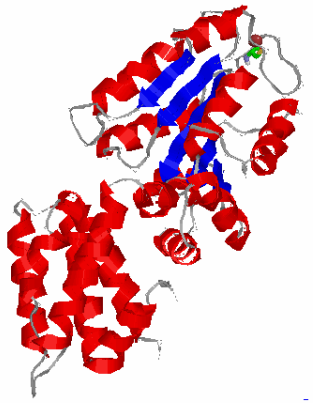
```
cagggccgcgtgcatcccgccactgtgg
||||||||  |||  |||||||  ||
cagggccgattgcgcgccgccaccctga

gcccaggcctggggcgagcggctgcgcg
||  | |||||||||| ||  ||||||||
gcggaagcctggggccagaagctgcgcg
```



Phylogenetic relationships

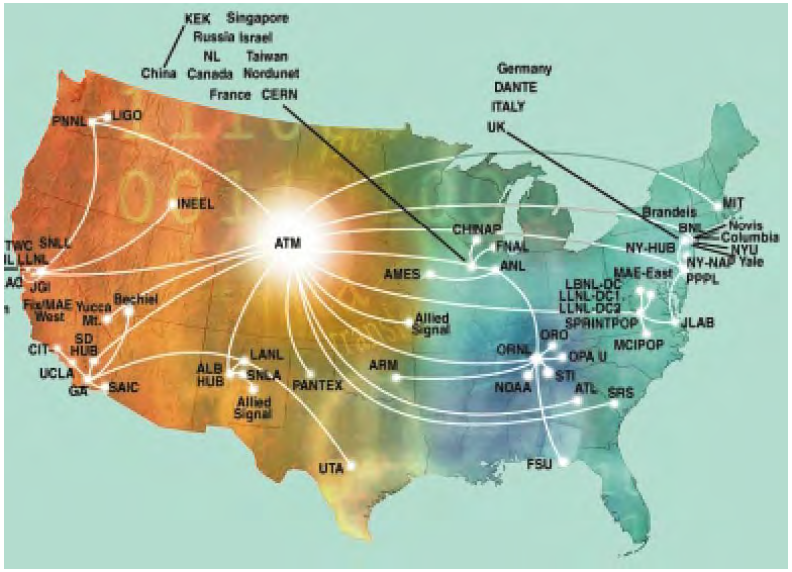


Protein function



Protein structure

Goal 4: Develop computational methods and capabilities to develop a predictive understanding of biological systems



- **Sequencing Informatics**
- **Sequence Annotation**
- **Structural Annotation**
- **Functional Annotation**
- **New Databases**
- **Data Integration**
- **Microbial Ecology**
- **Modeling & Simulations**
- **Visualization**

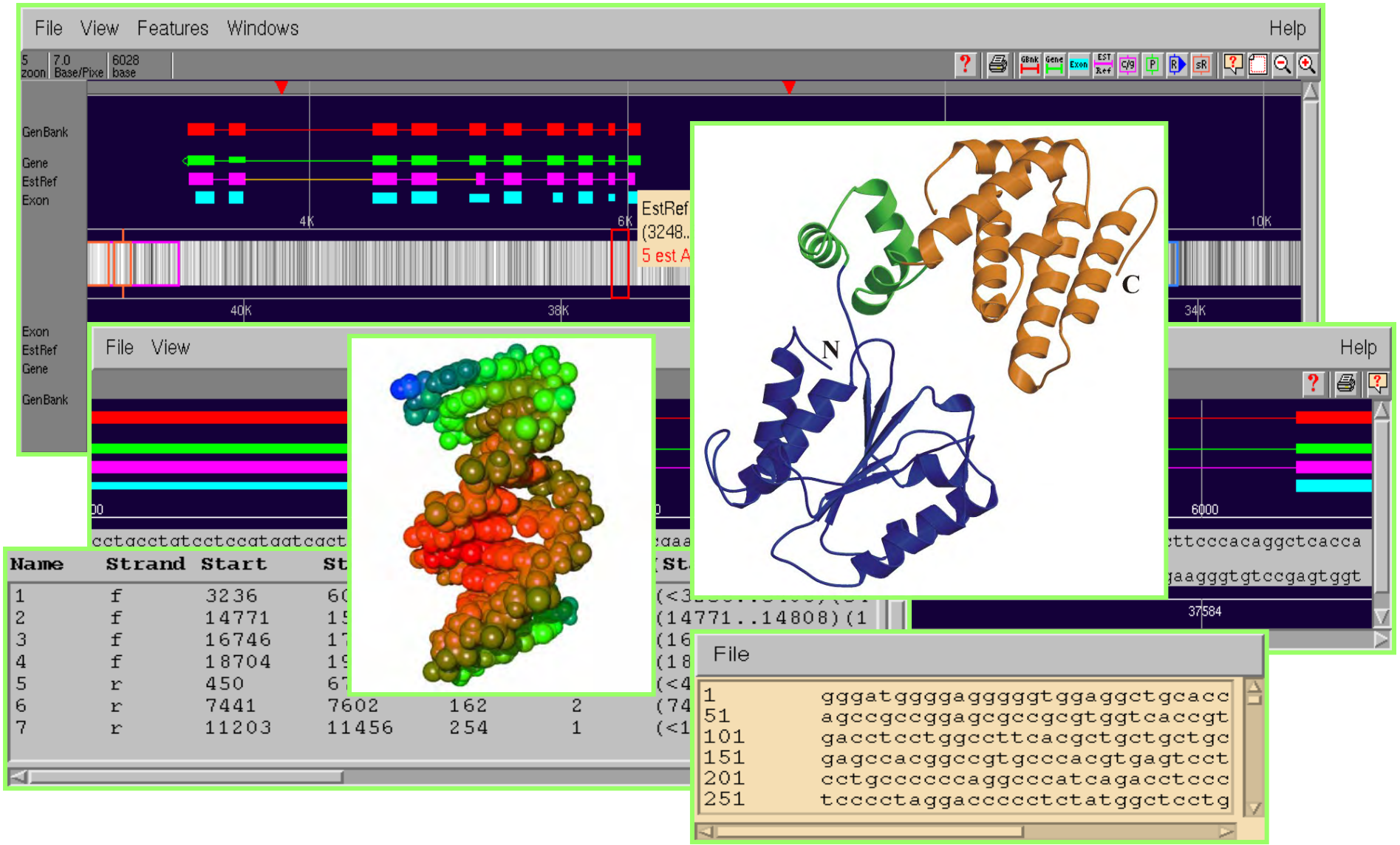
The development of the computational infrastructure for computational biology presents many challenges

- Need to integrate many different types of biological data
- Data derived from many different methods and laboratories
- Very rapid evolution in data collection methods
- Explosive growth in size of biological datasets
- Few standards and many incompatible databases

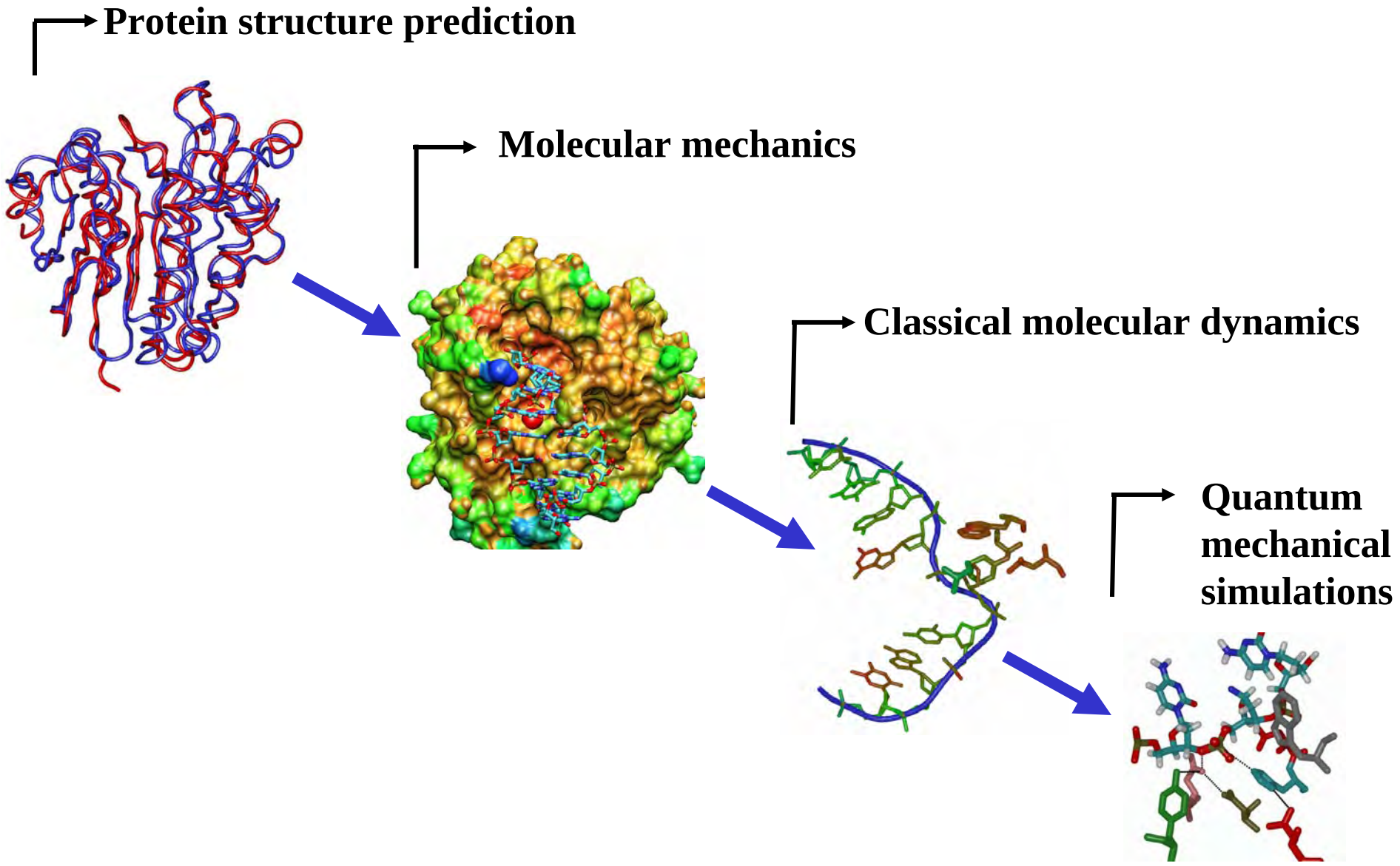
Near-term basic bioinformatics and annotation require ~10 teraFLOP computers

High Performance Computing Requirements for Genome Analysis 1999-2003 using GIST (Genomic Integrated Supercomputing Toolkit)								
	A) Computational Power	B) Query Time	C) Query Size	D)	E) Total Time	Total Time	Needed Frequency	Total Power needed (Top/s)
	(Top/secs)	(secs)			(secs)	(hours)	(days)	
1) Sequence Assembly								
adding new sequences	5	0.0002	1.00E+08	new basepairs	2.00E+04	5.56	1	1.15
rebuilding	5	0.0002	3.00E+09	basepairs	6.00E+05	166.67	60	0.57
2) Gene Modeling								
GRAIL-EXP	5	0.288	1.00E+06	exons	2.88E+05	80.00	7	2.38
3) Homology and Function								
Pairwise Sequence Comparision	5	0.0006	3.00E+07	megabases	1.80E+04	5.00	7	0.14
Multiple Sequence Alignment	5	0.1	1.00E+06	1 Kb sequences	1.00E+05	27.78	7	0.82
Protein Classification								
Database Maintenance	5	200	1.00E+00	database rebuild	2.00E+02	0.06	7	0.0016
Whole-Genome Searching	5	192	9.00E+02	models	1.73E+05	48.00	7	1.42
Phylogeny	5	120	1.00E+03		1.20E+05	33.33	7	0.99
4) Structure Modeling								
Protein Threading	5	5.528	2.00E+02	proteins	1.11E+03	0.31	1	0.063
Reanalysis	5	5.528	2.00E+05	proteins	1.11E+06	307.11	60	1.06
5) Systems and Pathways						research	research	research
6) Data Access and Storage								
User Query-by-sequence	5	0.0006	3.00E+07	megabases (1)	1.80E+04	5.00	1	1.04
TOTAL (Top/s):								**9.69**

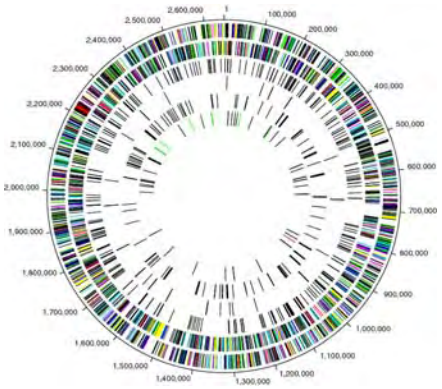
Major computational challenges remain in organizing, integrating, and visualizing biological data



Integrated multi-scale molecular models can provide information on macromolecular interactions and function

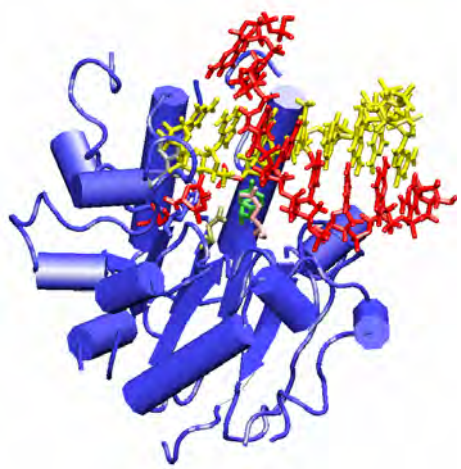


The necessary elements for leadership in computational biology are already core strengths of the DOE

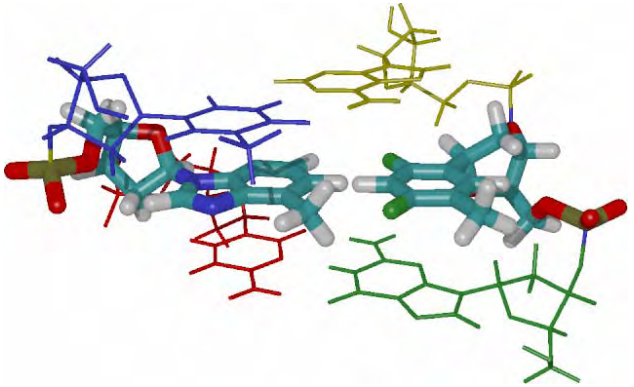


Bioinformatics

Advanced computing and algorithms



Protein structure prediction and macromolecular simulation



Molecular modeling and computational chemistry

Concluding comments

- Biology has many needs for advanced computer science and large scale computing
- GTL provides a framework for creating a new kind of biological research that is integrated with the computational sciences
- Key goal at this time is to develop effective partnerships between computational and biological scientists

"It is important to point out that there are plenty of problems that justify electronic [computing] speeds, and furthermore, the chances are that if these speeds become available, we will come to discover more and more how numerous these problems are."

--Von Neumann, 1946, on the need for electronic computers