ASCAC-BERAC Joint Subcommittee on Computational Biology and GTL

Rick Stevens

John Wooley

Co-chairs

The Subcommittee Charge

- Convene a joint panel with BERAC to examine the issue of computational models for GTL, including:
- How progress could be accelerated through targeted investments in applied mathematics, and
- How computer science can be incorporated to meet the needs of computational biology.
- The joint panel should consider whether the current ASCR long-term goal is too ambitious, given the status and level of buy-in from the community.
- It needs to consider what is happening in the computational-science and life-sciences communities. It should discuss possible intermediate goals that might be more relevant to the two programs.
- And it should identify the key computational obstacles to developing computer models of the major biological understandings necessary to characterize and engineer microbes for DOE missions, such as biofuels and bioremediation.

Joint Subcommittee Members

- Rick Stevens, Argonne-Uchicago (co-chair)
- John Wooley, UCSD (co-chair)
- Barbara Wold, Caltech
- David Galas, Battelle and ISB
- Thomas Zacharia, Oak Ridge-UT
- Michael Banda, Berkeley Lab
- Virginia Torczon, William and Mary
- David Kingsbury, Moore Foundation
- Chris Somerville, Carnegie Institution

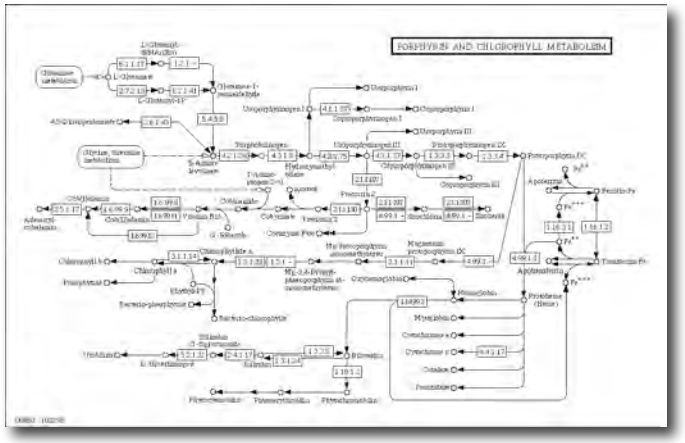
Overall Plan of Attack

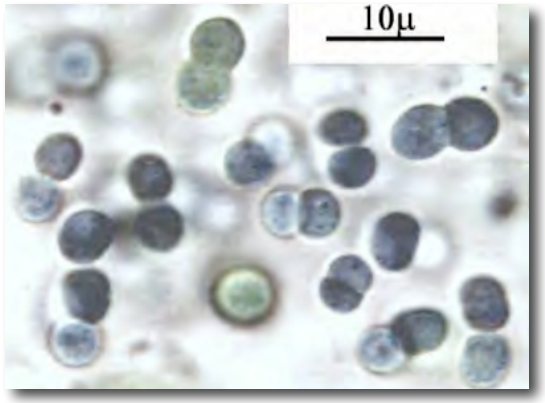
- Organizational and Framing Meeting (8/6/07)
- Teleconference (mid 9/07)
- Two Day Community Input Meeting at Planned for early October at the Moore Foundation in the Bay Area (~10/04/07)
- Writing Meeting/Teleconference (mid 10/07)
- Report Concurrence Teleconference (11/07)
- Reporting out at the November ASCAC and BERAC meetings.

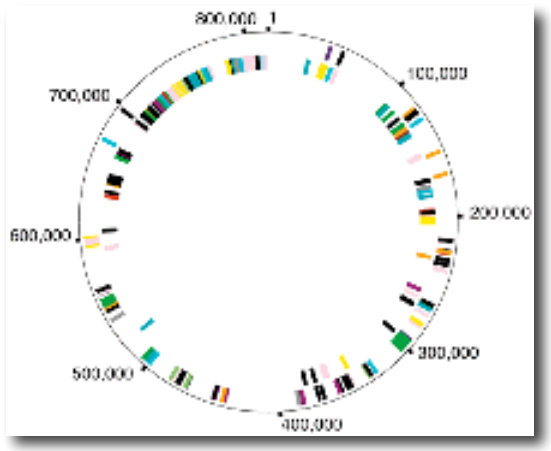
Observations from the First Meeting

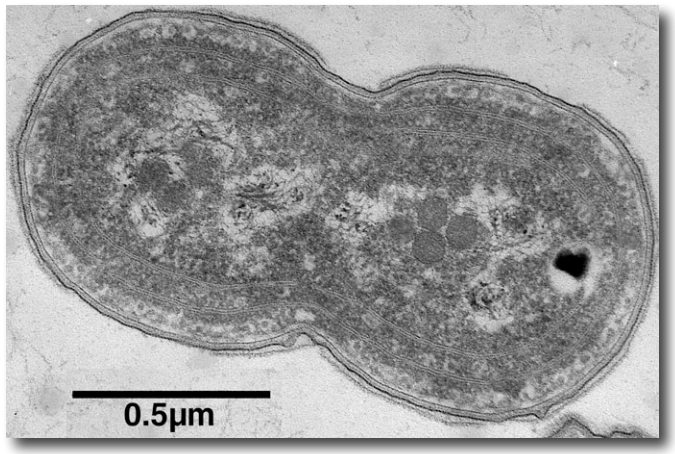
- Both BER and ASCR need a common set of explicit science goals to drive advances in bioinformatics, computational and theoretical biology and and the associated high-throughput experimental techniques in systems biology
- The existing PART Performance Target is not an ideal goal as currently stated since it is somewhat vague and not focused on a scientific outcome
- Also progress towards the existing PART Performance Target is also not easily measured
- The recent shifting of the BER agenda to nearer term bioenergy research may make it more difficult to focus on basic joint goals

Genes → Proteins → Cell Networks → Cells → Populations → Communities → Ecosystems



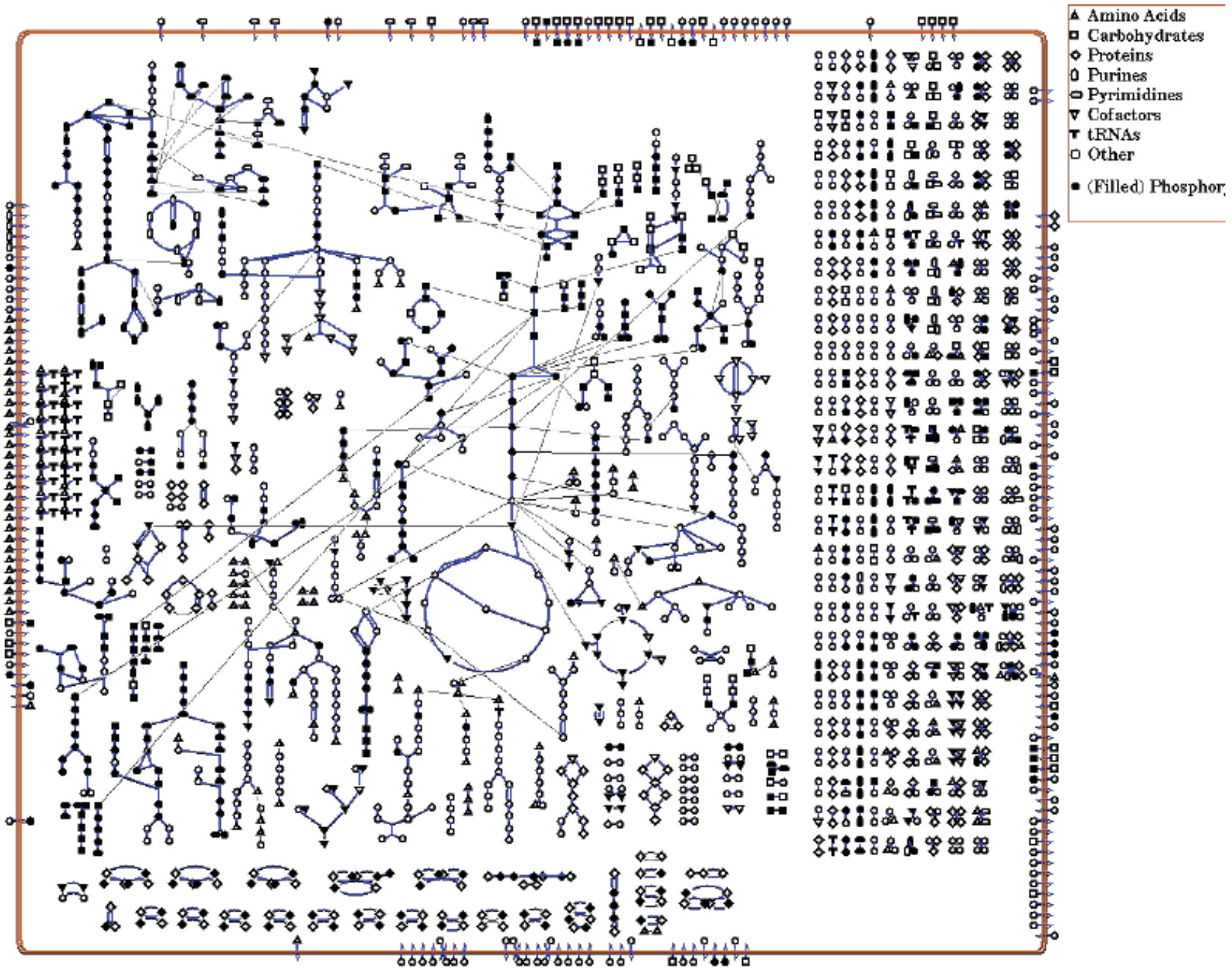






Microbial Cell Modeling an Example of one Approach

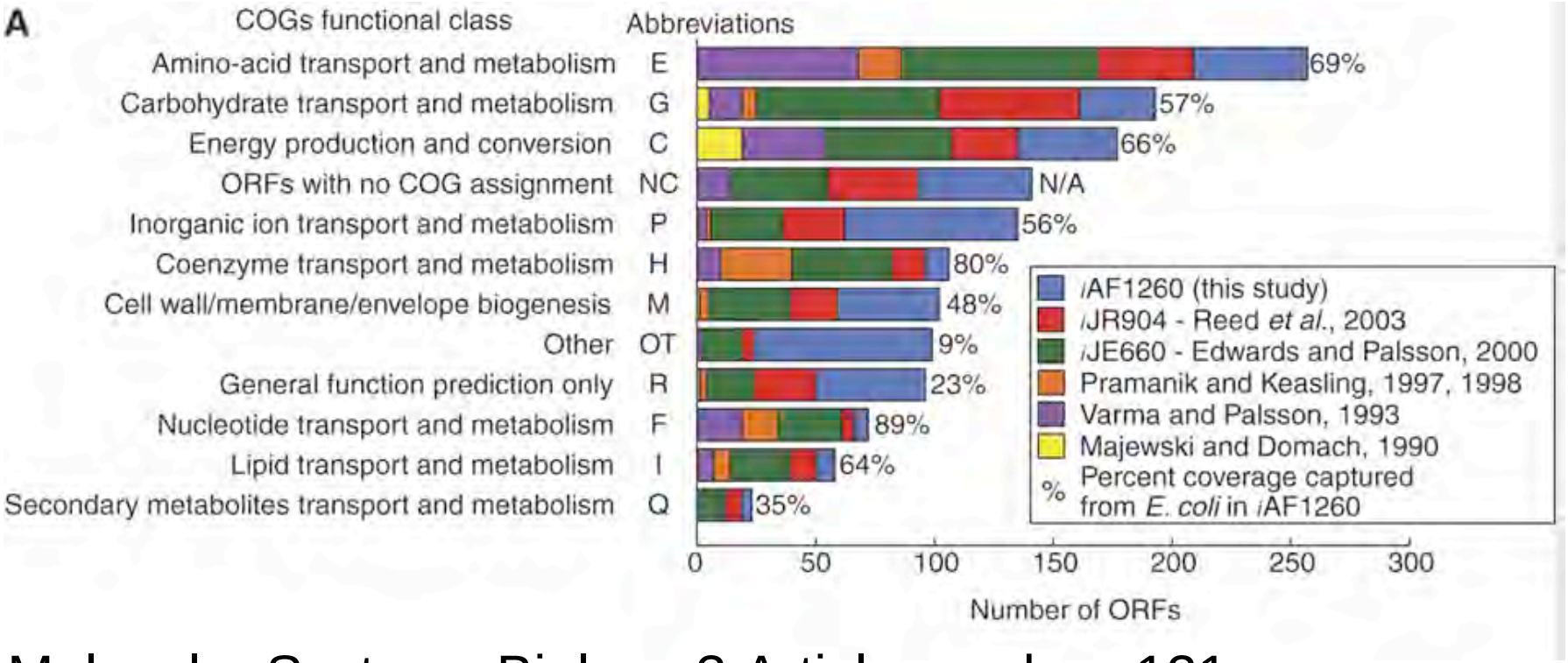
- There are many groups building models of cells at different levels of abstraction
- Significant progress has been made in the last 17 years on extending flux balance methods to whole genomes
- The state-of-the-art reconstruction now boasts over 1200 reactions incorporated into the model and is now covering nearly 70% of metabolic genes and increasing fraction of other cellular functions
- Soon it will be possible to semi-automatically produce 100’s of reconstructions from existing microbial genomes



***E. coli K-12* Metabolic Overview**

Source: EcoCyc

17 Years of Progress in FBA Model Development



Molecular Systems Biology 3 Article number: 121
doi:10.1038/msb4100155
Published online: 26 June 2007

Prokaryotic Intracellular Environment – Gel Like Media

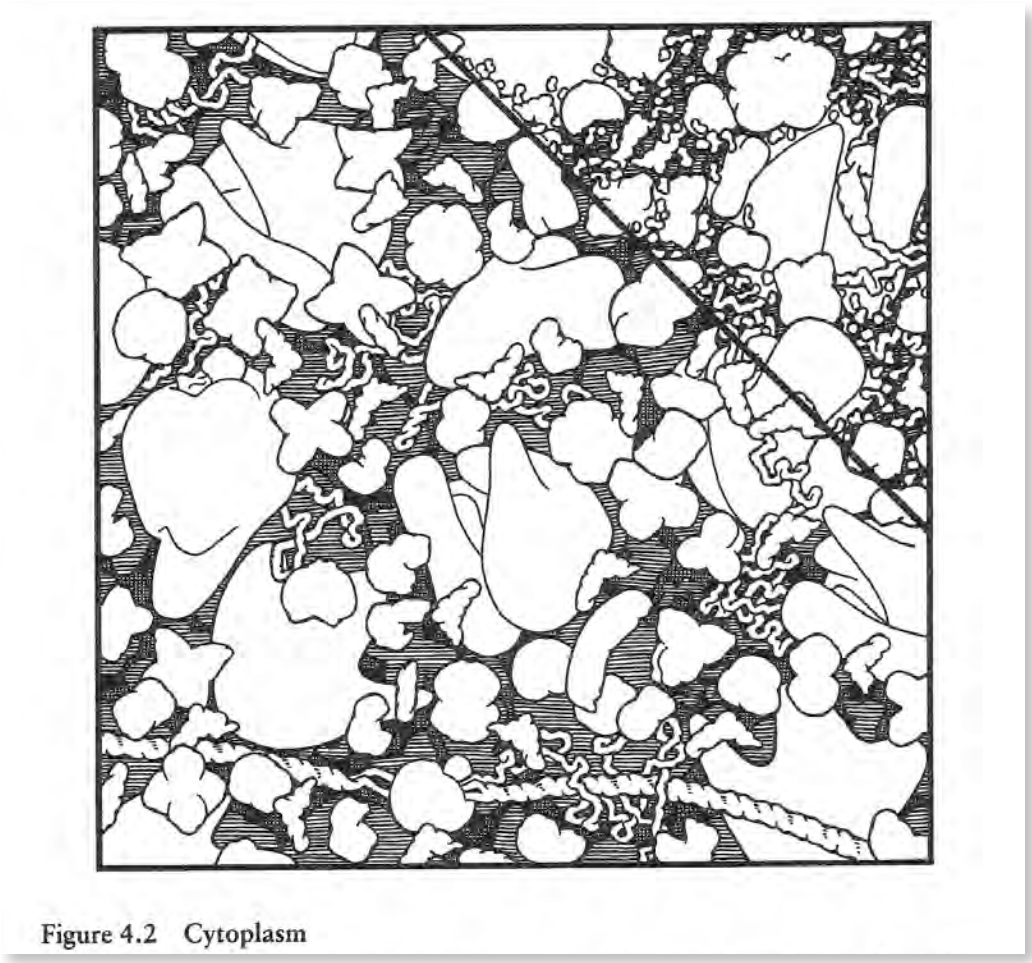


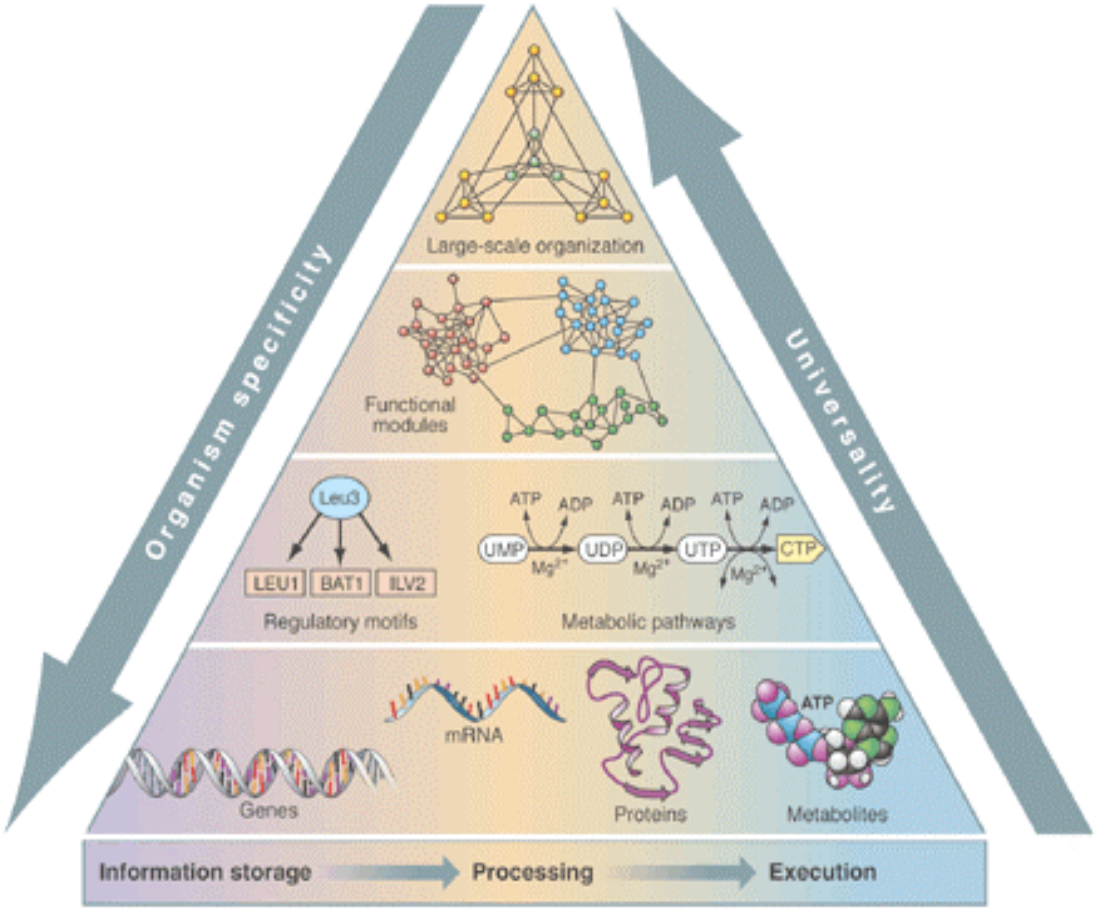
Figure 4.2 Cytoplasm

From: David Goodsell, The Machinery of Life

- 100 nm^3
- 450 proteins
- 30 ribosomes
- 340 tRNA molecules
- Several long mRNAs
- 30,000 small organic molecules
- 50,000 Ions
- Rest filled with water 70%

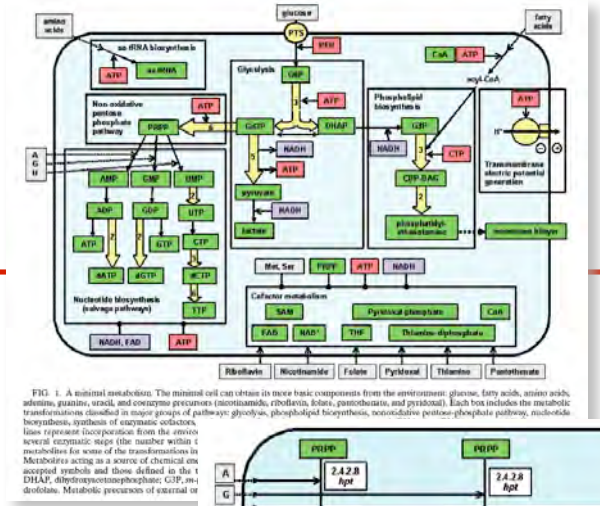
Hierarchical Modeling in Biological Systems

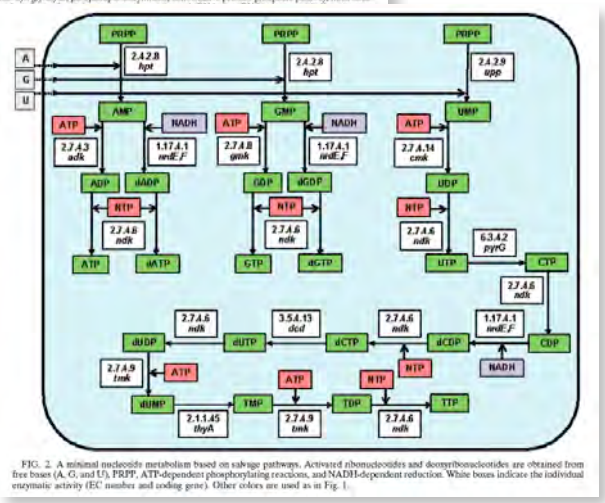
- Genetic Sequences
- Molecular Machines
- Molecular Complexes and modules
- Networks + Pathways [metabolic, signaling, regulation]
- Structural components [ultrastructures]
- Cell Structure and Morphology
- Extracellular Environment
- Populations and Consortia etc.



In Quest of the Minimum Genome

- What are the Smallest number of genes needed to create a viable organism?
  - Free living on a rich, but defined culture medium
- Experimentally determined essential genes
  - Bacillus subtilis ~300 CDS
  - Escherichia coli ~400 CDS
- Reduced organisms in nature
  - Mycoplasma ~500
  - Nanoarchaea ~400
- Bioinformatics predicts a conserved core
  - ~200-400 CDS





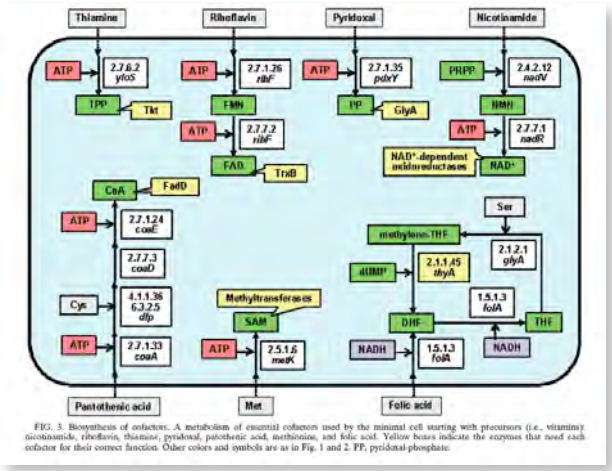


Table 1

Genome-scale gene essentiality studies in bacteria.

Organism	Mutagenesis	Mutant outgrowth	Essentiality assesment					Ref.
			Readout	ORFs total	N	E	%E	
M. genitalium, *M. pneumonia*	Random	Population	Sequencing	482	130	265–350	55–73%	[1]
M.genitalium	Random	Clones	Sequencing	482	100	382	79%	[14•]
S. aureus WCUH29	Random	Clones	Sequencing	2600	n/a[b]	168[c]	n/a	[2]
S. aureus RN4220	Random	Clones	Sequencing	2892	n/a[b]	658[c]	23%	[3]
H. influenzae Rd	Random	Population	Footprint-PCR	1657	602	670	40%	[5]
S. pneumoniae Rx-1	Targeted	Clones	Colony formation	2043	234	113[c]	n/a	[4]
S. pneumoniae D39	Targeted	Clones	Colony formation	2043	560	133[c]	n/a	[13]
M. tuberculosis H37Rv	Random	Population	Microarray	3989	2567	614	15%	[6•]
B. subtilis 168	Targeted	Clones[a]	Colony formation	4105	3830[d]	271[d]	7%	[7••]
E. coli K-12 MG1655	Random	Population	Footprint-PCR	4308	3126	620	14%	[8]
E. coli K-12 MG1655	Targeted	Clones[a]	Colony formation	4308	2001	n/a[e]	n/a	[12]
E. coli K-12 BW25113	Targeted	Clones[a]	Colony formation	4390	3985	303	7%	[15••]
P. aeruginosa PAO1	Random	Clones[a]	Sequencing	5570	4783	678	12%	[9]
P. aeruginosa PA14	Random	Clones[a]	Sequencing	5688	4469	335[f]	6%	[16•]
S. typhimurium	Random	Clones	Sequencing	4425	n/a[b]	257[c]	~11%	[10]
H. pylori G27	Random	Population	Microarray	1576	1178	344	22%	[11]

This table provides a short summary. We refer the reader to the Supplementary material (Table S1) for details. Genome-wide screens for genes essential for virulence are beyond the scope of this review and are not listed here. The complete gene essentiality datasets obtained in these studies have been incorporated in the SEED (http://theseed.uchicago.edu/FIG/index.cgi) and National Microbial Pathogen Data Resource (NMPDR; http://www.nmpdr.org/) genomic databases. ORFs total, an estimate of a total number of protein-encoding genes in a genome; N, genes detected as nonessential; E, deemed essential for survival (for datasets generated via clonal strategy) or essential for fitness (for datasets generated via populational screens); %E, fraction of essential genes in a genome.

[a] Mutant collection is available for public distribution.

[b] Direct essentiality screening methods (e.g. via antisense RNA) do not provide information about nonessential genes.

[c] Only partial dataset is available.

[d] This list also includes predicted gene essentiality and data compilation from published single-gene essentiality studies.

[e] Project in progress.

[f] Deduced by comparison of the two gene essentiality datasets obtained independently in the *P. aeruginosa* strains PA14 [16•] and PAO1 [9].

Current Opinion in Biotechnology 2006, 17:448–456

Modeling of Simple Microbial Communities

- Many environmental metagenomics projects have identified modest sized microbial communities in novel environments
  - Dozens of Studies have been done and hundreds are being planned
- Reconstruction of these communities (genomic, proteomic, metabolic) has begun
  - When the diversity is small, it is possible to essentially sequence the dominate organisms without culturing
- In addition there are well studied model communities that can serve as laboratory models for development (e.g. Winogradsky columns, biofilms, etc.)

World J Microbiol Biotechnol (2007) 23:15–21

Acid Mine Drainage Sediment Community

Table 2 BLAST analysis of 16S rDNA sequences of acidophiles in the sediment

Representative sequence	Length (bp)	Frequency	Microbial group affiliation	Closest relative (Genebank accession number)	Similarity
H5	1461	4.4%	*Acidobacteria*	Uncultured bacterium (AB179509)	1378/1462 (94%)
H6	1494	4.4%	*γ-Proteobacteria*	Uncultured bacterium clone 1013-28-CG34 (AY532575)	1423/1492 (95%)
H11	1517	51.1%	*δ-Proteobacteria*	Uncultured bacterium BA71 (AF225447)	1418/1451 (97%)
H12	1521	6.7%	*Nitrospira*	Uncultured bacterium clone ASL9 (AF544226)	1476/1515 (97%)
H24	1502	4.4%	*β-Proteobacteria*	Uncultured bacterium clone DSBACT9 (AY762628)	1030/1109 (92%)
H40	1432	2.2%	Candidate Division TM7	Uncultured soil bacterium clone C129 (AF507687)	1297/1411 (91%)
H50	1520	6.7%	*Nitrospira*	Uncultured bacterium (DQ223212)	1473/1517 (97%)
H65	1500	4.4%	*γ-Proteobacteria*	*Acidithiobacillus ferrooxidans* strain QXS-1 (DQ168465)	1491/1498 (99%)
H70	1491	2.2%	Low G + C Gram-positives	Uncultured Low	782/799 (97%)
H74	1484	13.3%	*δ-Proteobacteria*		967/992 (97%)

Yellowstone Microenviroments

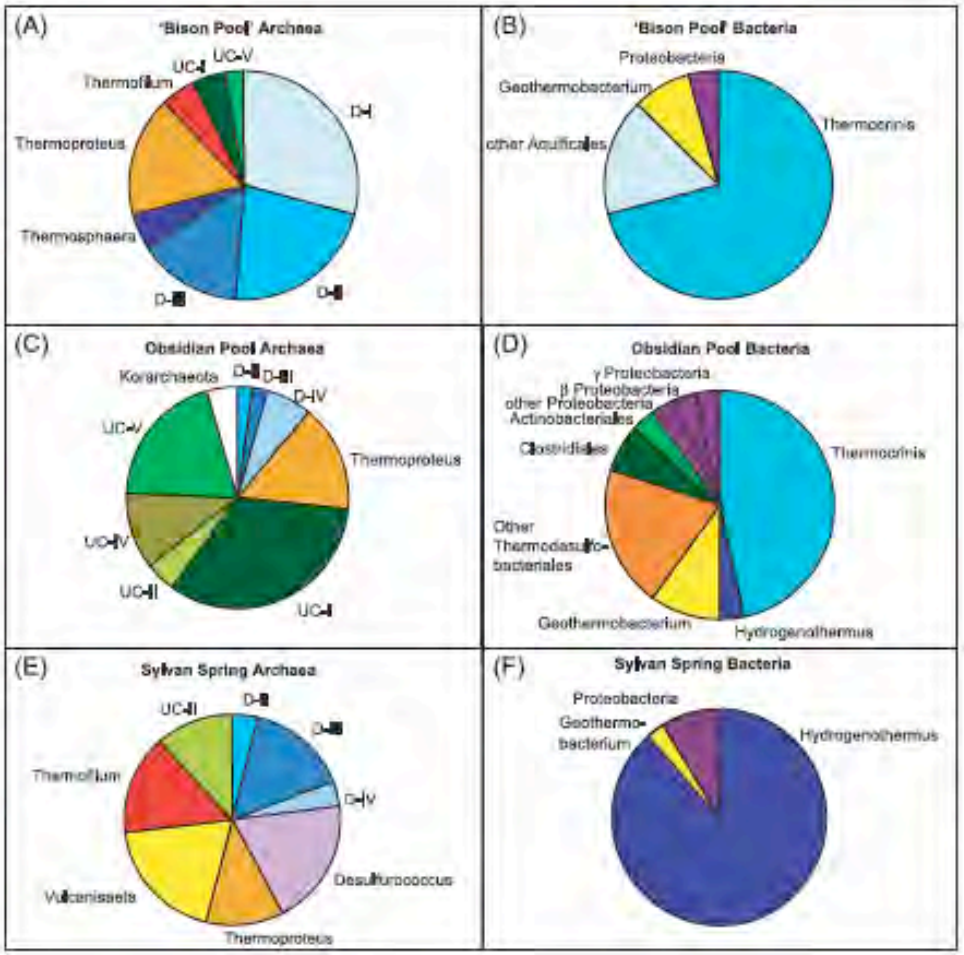
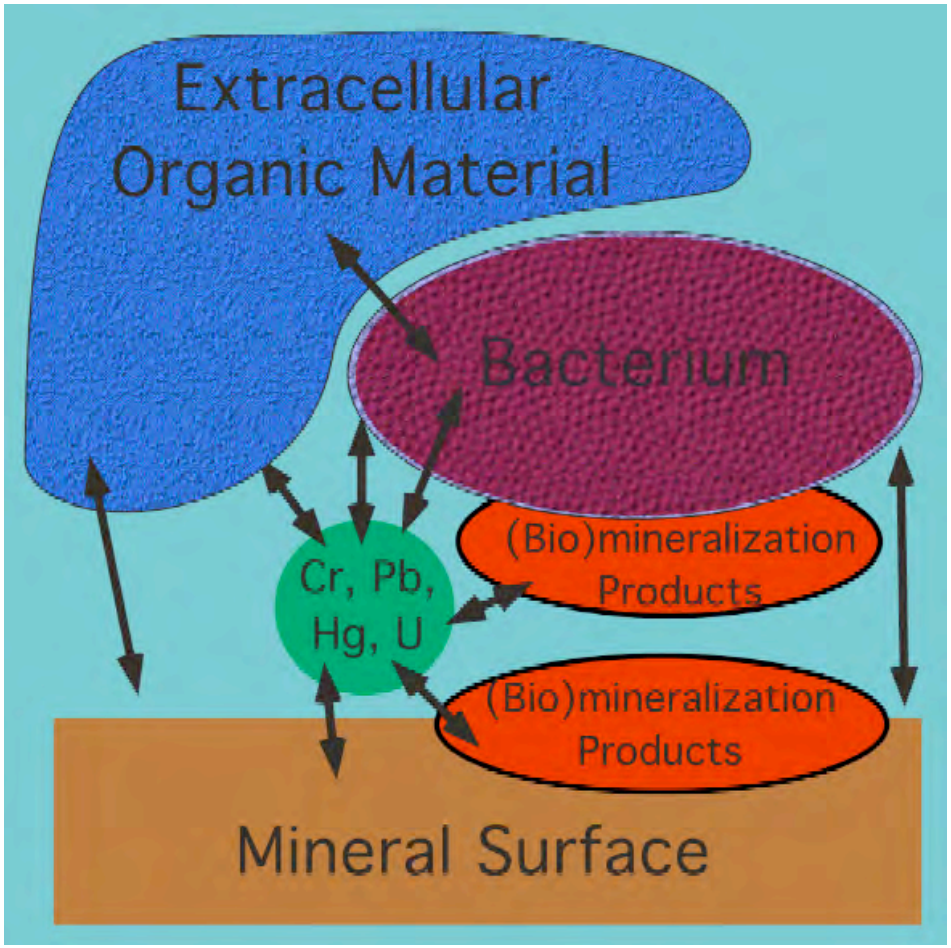


Fig. 3 Diversity and distribution of plasmids within archaeal and bacterial clone libraries for 'Bison Pool' (A, B), Obsidian Pool (C, D), and Sylvan Spring (E, F). Pie wedges indicate the percentages of each taxon in the total library, as determined by screens of partial (~700 bp) 16S rRNA sequences (see Table 1). Colours are coordinated for all archaeal libraries (A, C, E) and all bacterial libraries (B, D, F), but not across domain boundaries.

Geobiology (2005), 3, 211–227

Extending the Models to Include the Environment is Key to Progress



Emergent Biogeography of Microbial Communities in a Model Ocean

Michael J. Follows,1* Stephanie Dutkiewicz,1 Scott Grant,1,2 Sallie W. Chisholm3

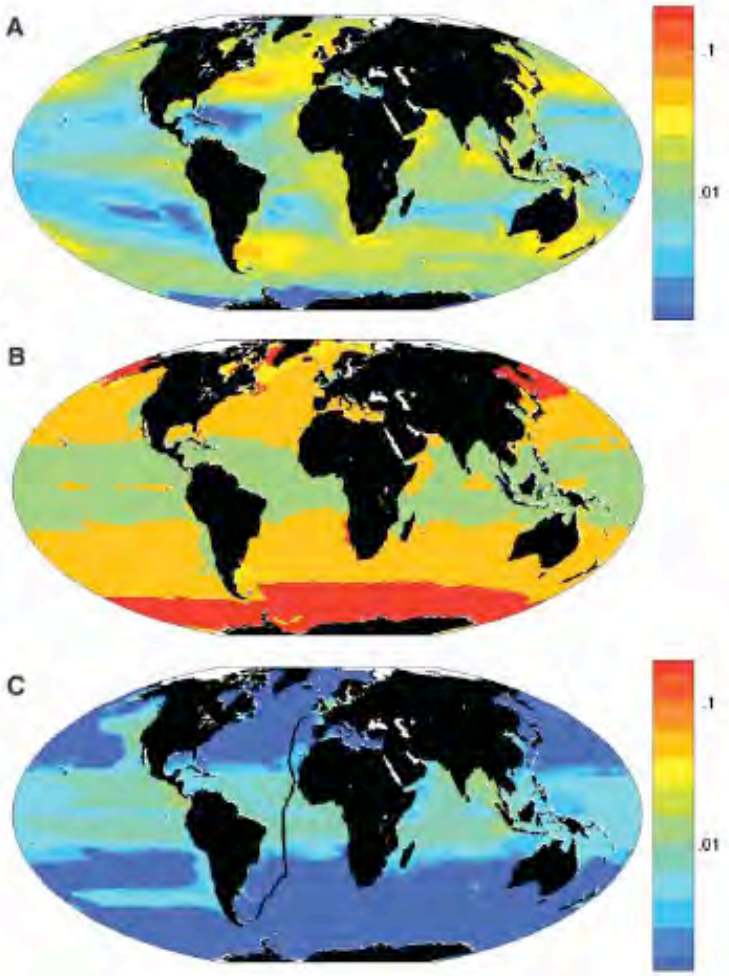


Fig. 1. Annual mean biomass and biogeography from single integration. (**A**) Total phytoplankton biomass (μM P, 0 to 50 m average). (**B**) Emergent biogeography: Modeled photo-autotrophs were categorized into four functional groups; color coding is according to group locally dominating annual mean biomass. Green, analogs of *Prochlorococcus*; orange, other small photo-autotrophs; red, diatoms; and yellow, other large phytoplankton. (**C**) Total biomass of *Prochlorococcus* analogs (μM P, 0 to 50 m average). Black line indicates the track of AMT13.

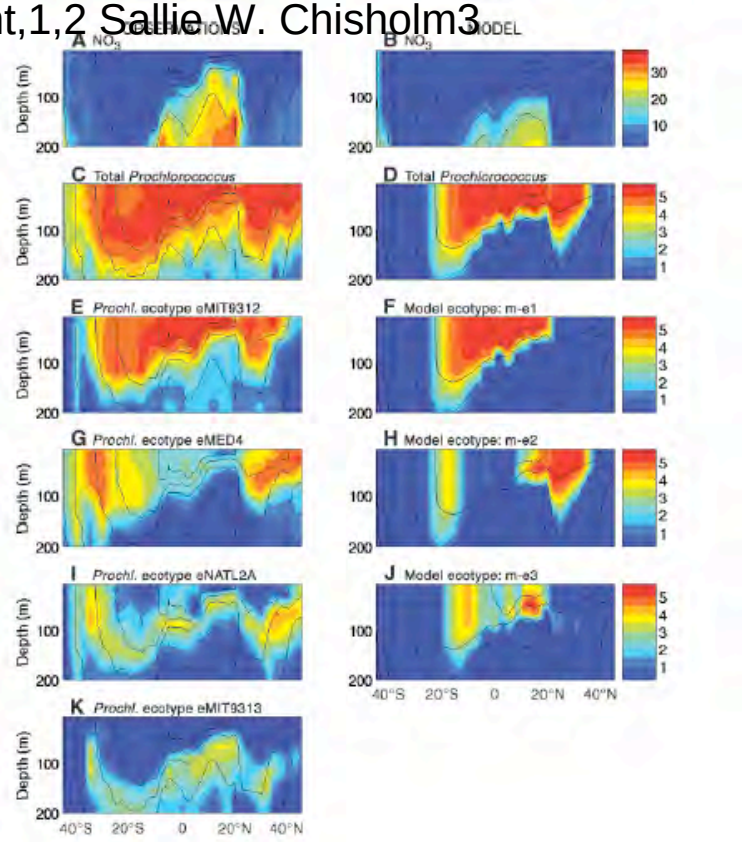


Fig. 2. Observed and modeled properties along the AMT13 cruise track. Left column shows observations (*17*), right column shows results from a single model integration. (**A** and **B**) Nitrate (μmol kg^{-1}); (**C** and **D**) total *Prochlorococcus* abundance [log (cells ml^{-1})]. (**E**, **G**, **I**, and **K**) Distributions of the four most abundant *Prochlorococcus* ecotypes [log (cells ml^{-1})] ranked vertically. (**F**, **H**, and **J**) The three emergent model ecotypes ranked vertically by abundance. Model *Prochlorococcus* biomass was converted to cell density assuming a quota of 1 fg P $cell^{-1}$ (*27*). Black lines indicate isotherms.

SCIENCE VOL 315 30 MARCH 2007

The Subcommittee is Focusing on a Short List of Exemplar Goals for Consideration

- Possibilities so far considered include:
  - Microbial Communities Associated with Carbon Sequestration
  - Microbial Communities Associated with Bioremediation
  - Communities Associated with Cellulose Degradation
  - A Synthetic Model Cell that Can be used for analysis of what needs to be measured for systems identification.

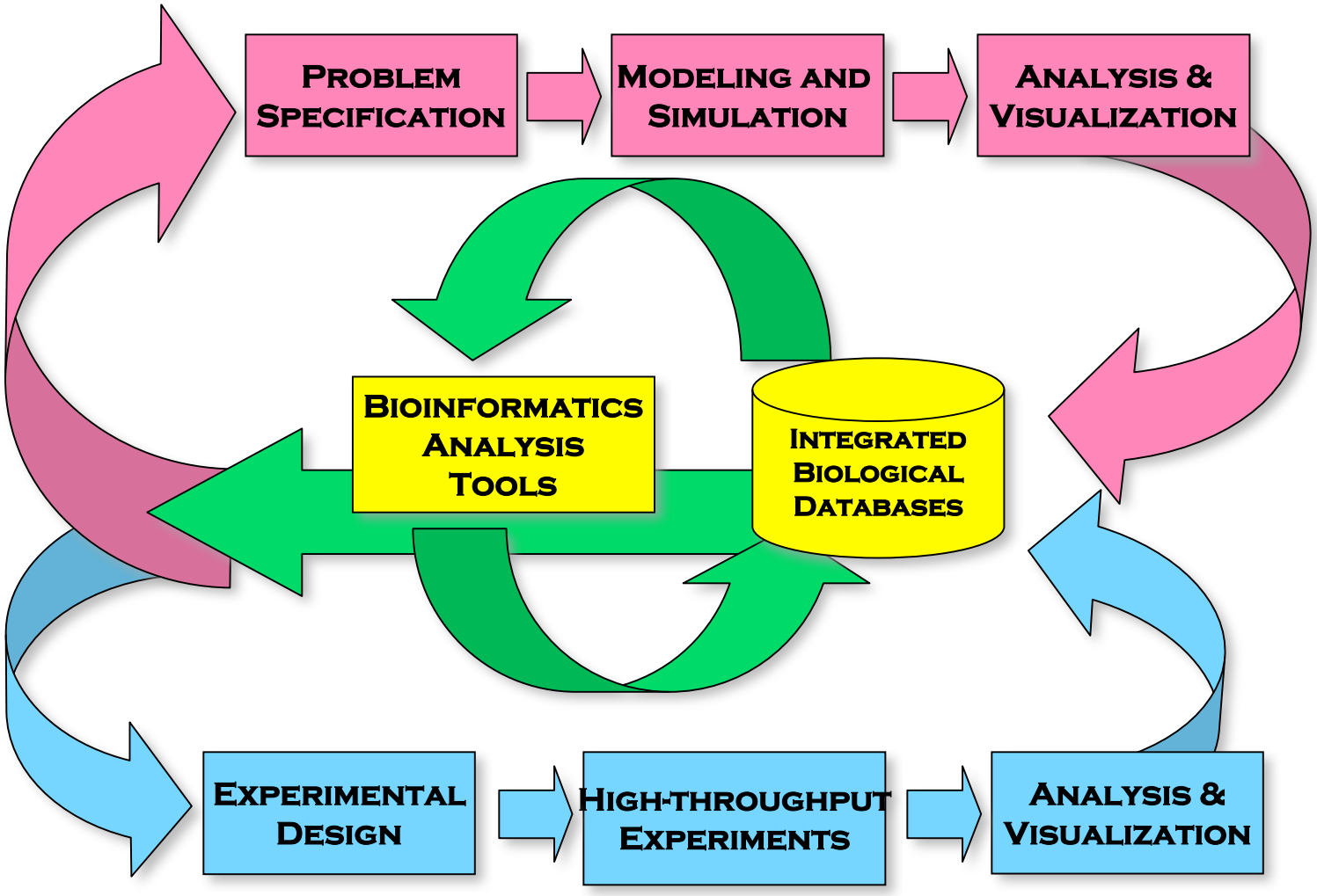
Predictive Modeling and Simulation

- Goal: Develop Integrated predictive models relating cell processes, phenotypes and response to environment

- Predictive Models
  - Metabolism
  - Transport
  - Regulation
  - Signaling
  - Replication
  - Development
  - Motility

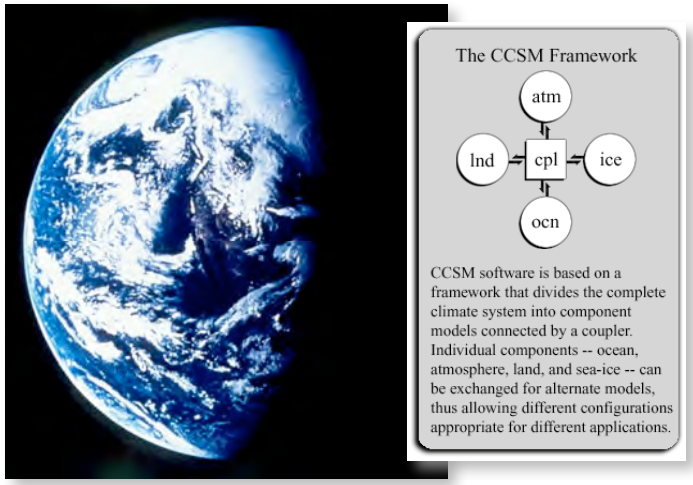
- Databases
  - Sequences and Expression
  - Phenotypes and Imaging
- Bioinformatics
  - Annotation and Informatics
  - Network and Model Reconstruction
- Modeling Targets
  - Model Organisms
  - Diverse Organisms
  - Limited Communities
  - Natural Communities

An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics



Challenges for Cell and Ecosystem Simulation

- Modeling cells rivals the complexity of climate and earth systems models
  - Multiple space and time scales
  - Millions of interacting parts
  - Populations of cells to understand emergent behavior
  - Integrated modeling necessary to advance theory in systems biology
- Cell modeling and systems biology could be a driver for Petascale computing and beyond



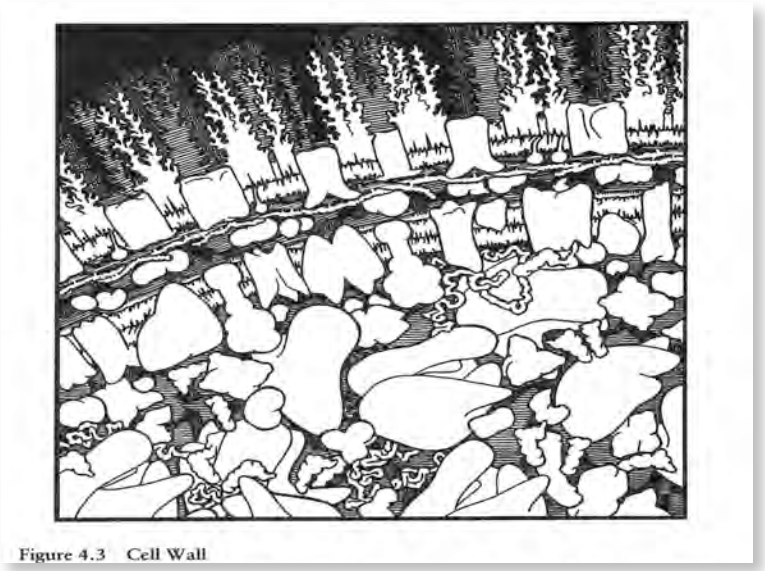


Figure 4.3 Cell Wall