Biology and Grids

Rick Stevens
Argonne National Laboratory
University of Chicago
stevens@mcs.anl.gov

Grids and Biology

- Biology perhaps more than any other discipline can benefit from Grids
  - Biology is data and computing intensive
  - Biology is highly distributed
  - Biology is moving quickly
  - Biology is transitioning from an experimental science to a theory and computing driven science
  - Biologists are already comfortably dependent on the web
- The productivity gains from BioGrids will directly impact drug development and delivery of medical care

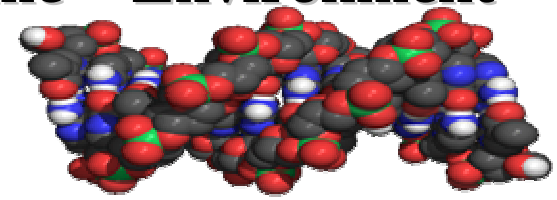
Future milestones

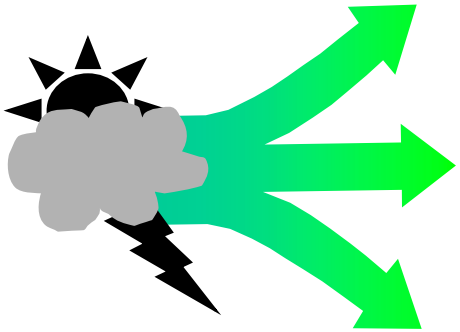
- First synthetic model prokaryotic organism
- Characterization of human microbial ecology
- Global index to life on earth
- Characterization of microbial life
- Theory of cell evolution and organization
- Theory of evolution of intelligence
- First synthetic eukaryotic organism
- Confirmation of extra-solar earthlike planets
- Synthetic self-reproducing biomemetic nanosystem

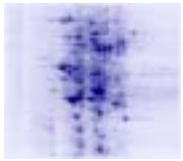
Biology is an Societally Critical Grid Domain

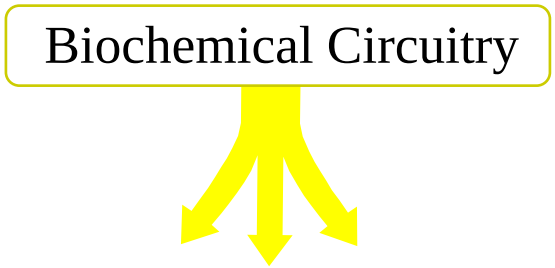
- Safe and abundant food supplies
- Sustainable and benign energy sources
- Effective management of disease and aging
- Novel materials and renewable industrial feedstocks
- Advanced computational devices beyond Moore's law
- Wide variety of molecular scale machinery
- Self-assembly and self-reproduction technologies

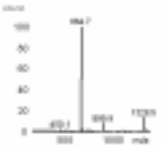
Reverse Engineering Living Systems:
Genome + Environment = Phenotype

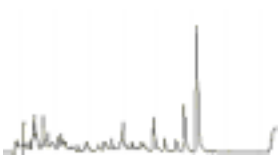


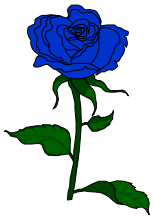


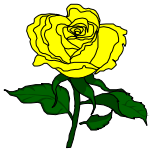


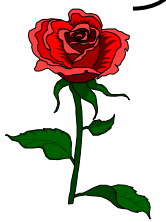






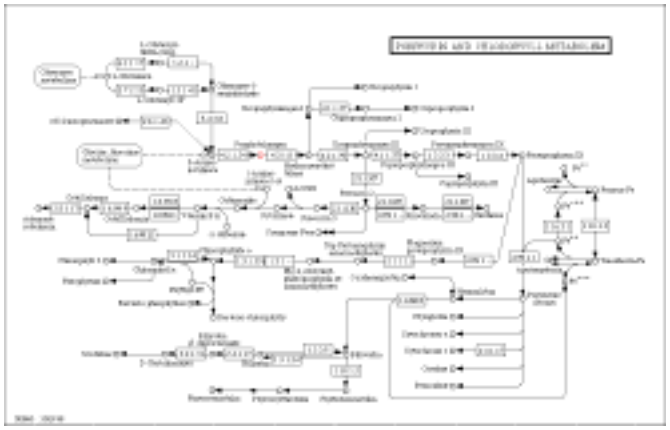


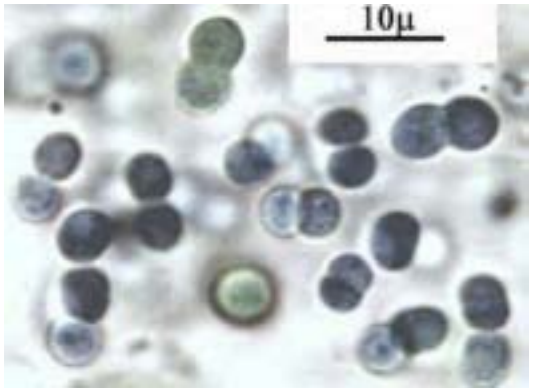


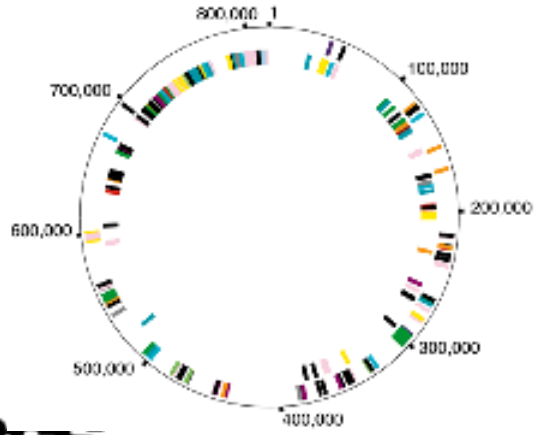


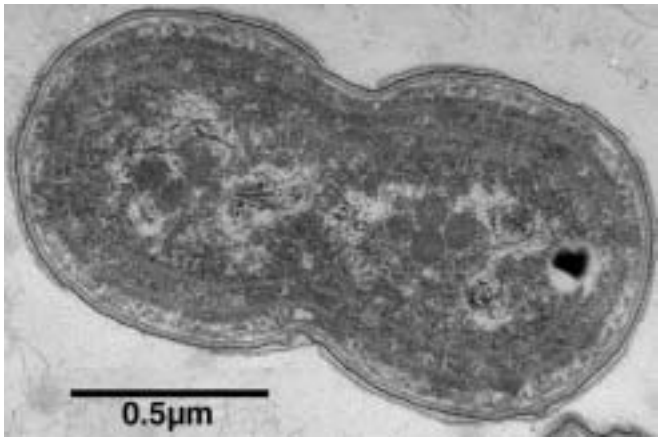
Adapted From Bruno Sobral VBI

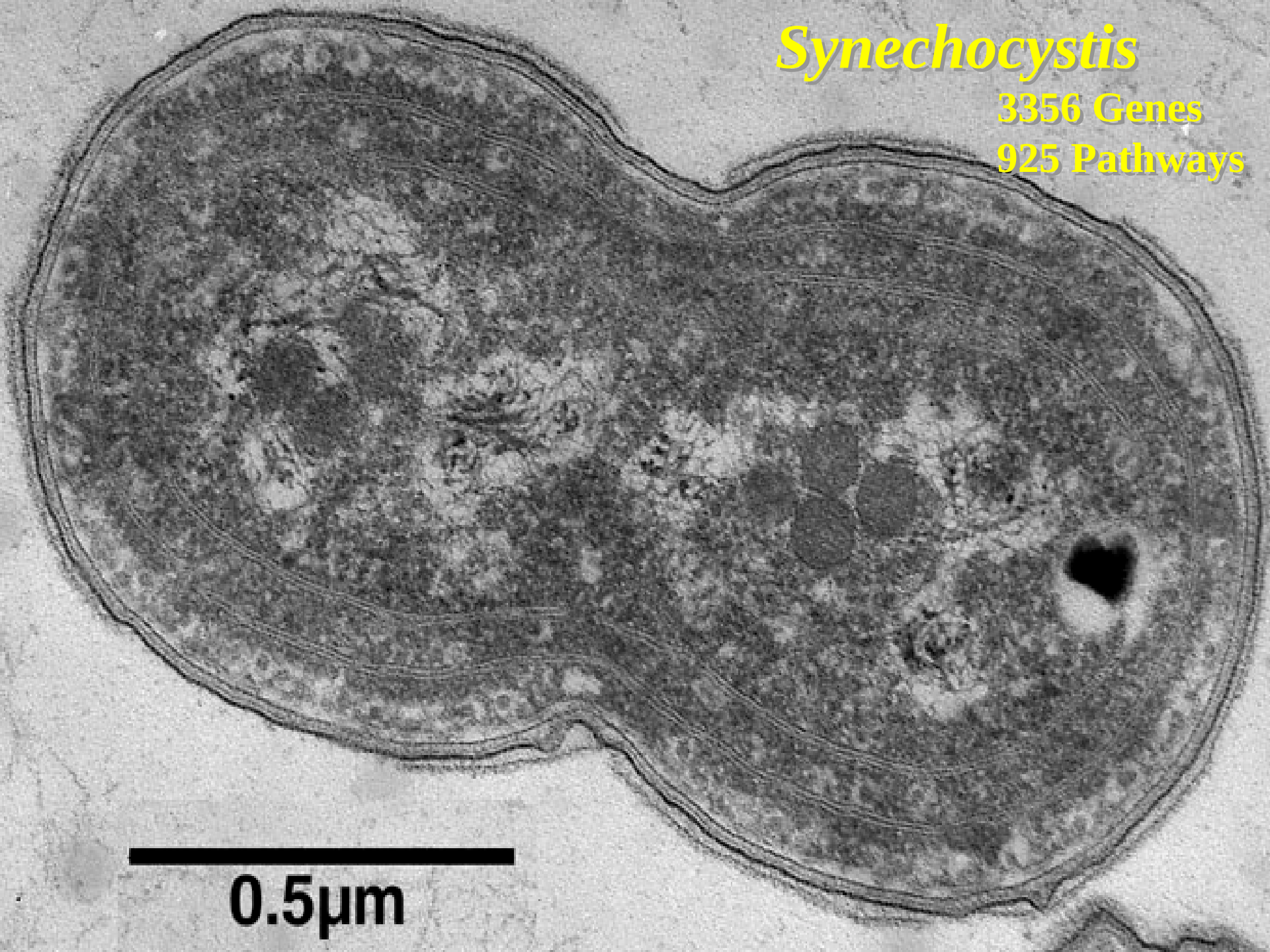
Systems Level Understanding
Genes → Cell Networks → Populations → Communities





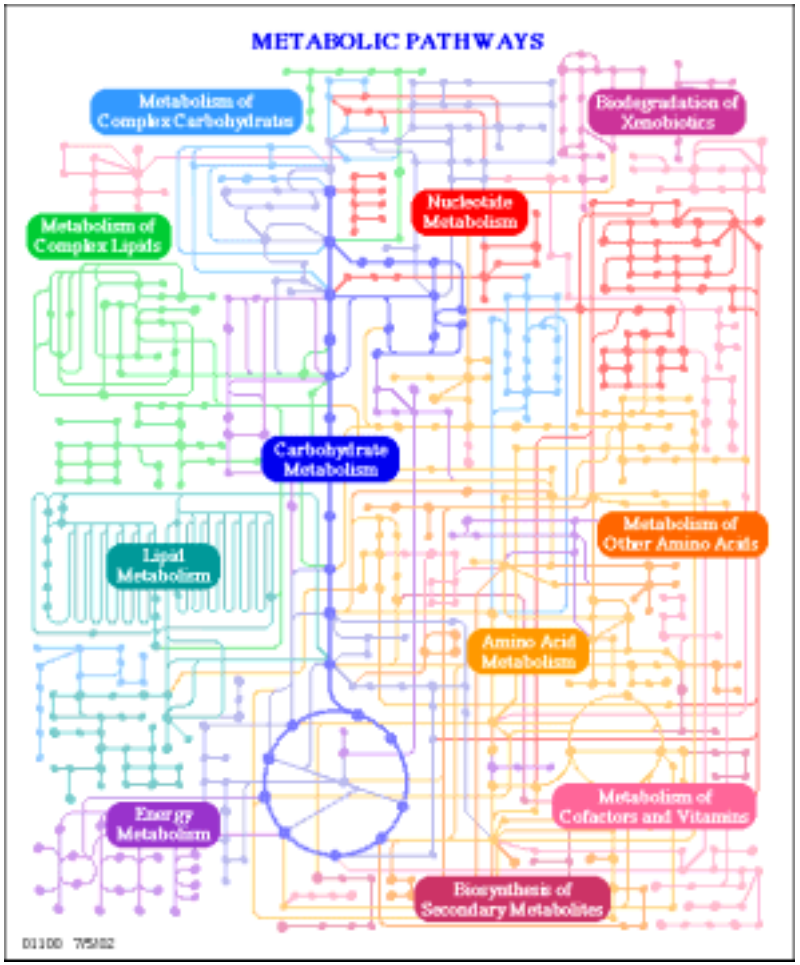




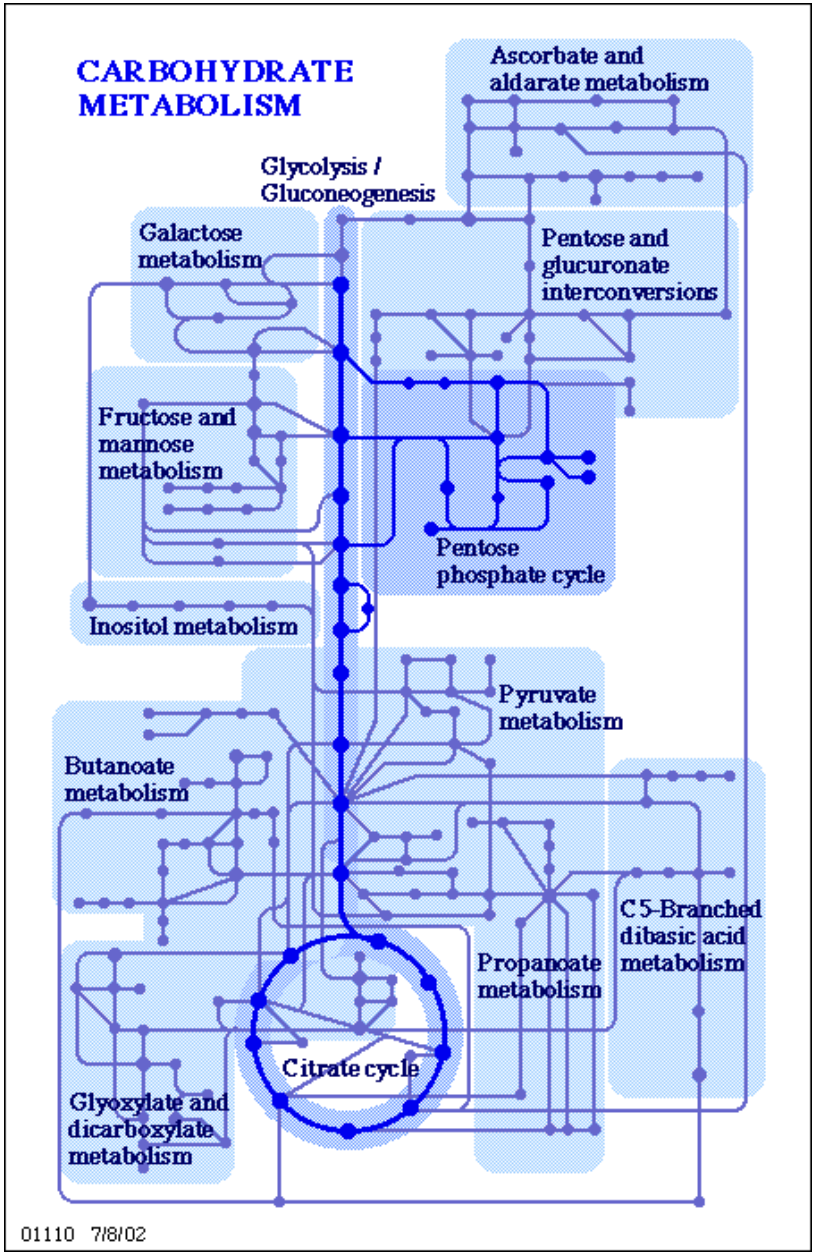

Synechocystis
3356 Genes
925 Pathways
0.5µm

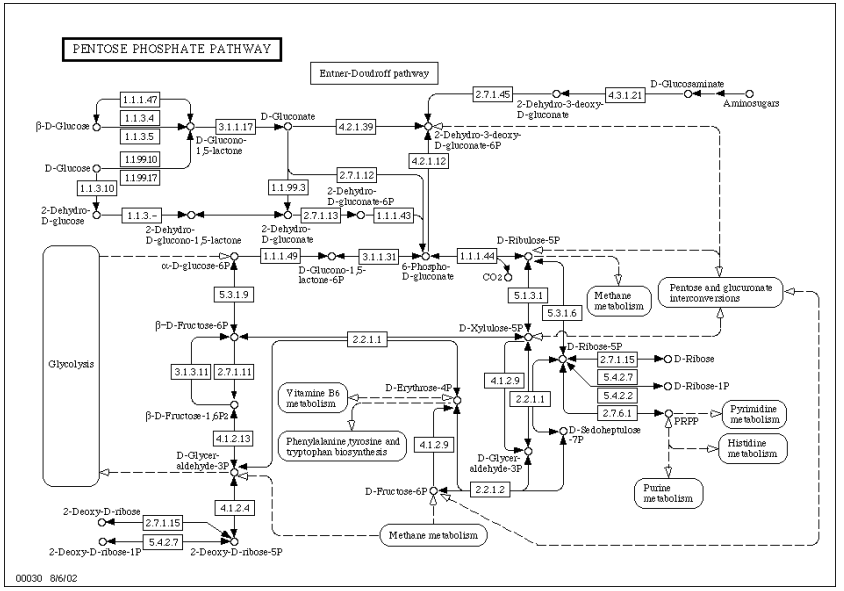
Systems Biology is Different Way of Thinking

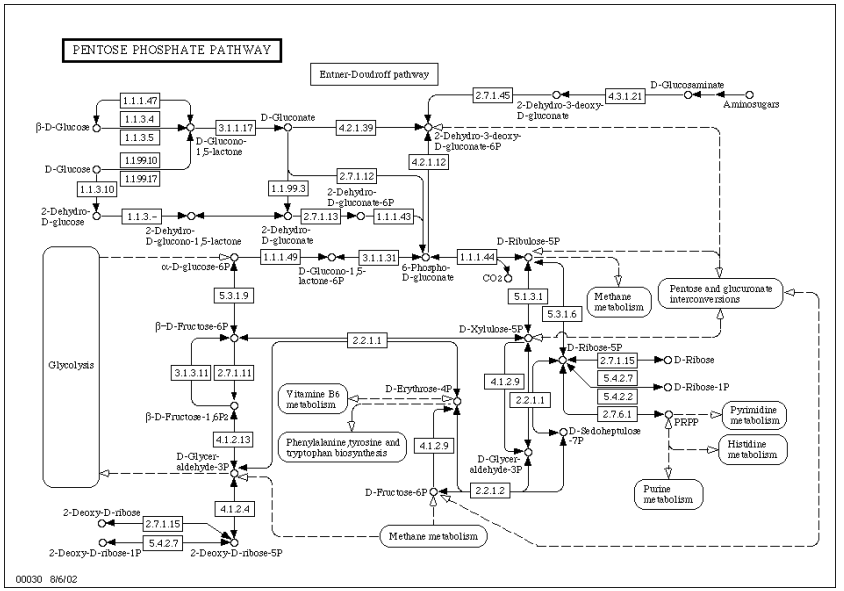
- Integrative understanding of a biological system
  - Cell, organism, community and ecosystem
- Counterpoint to reductionism
  - Requires synthesizing knowledge from multiple levels of the system
- Discovery oriented not necessarily hypothesis driven
  - Data mining vs theorem proving
  - Need both types of reasoning



KEGG pathways

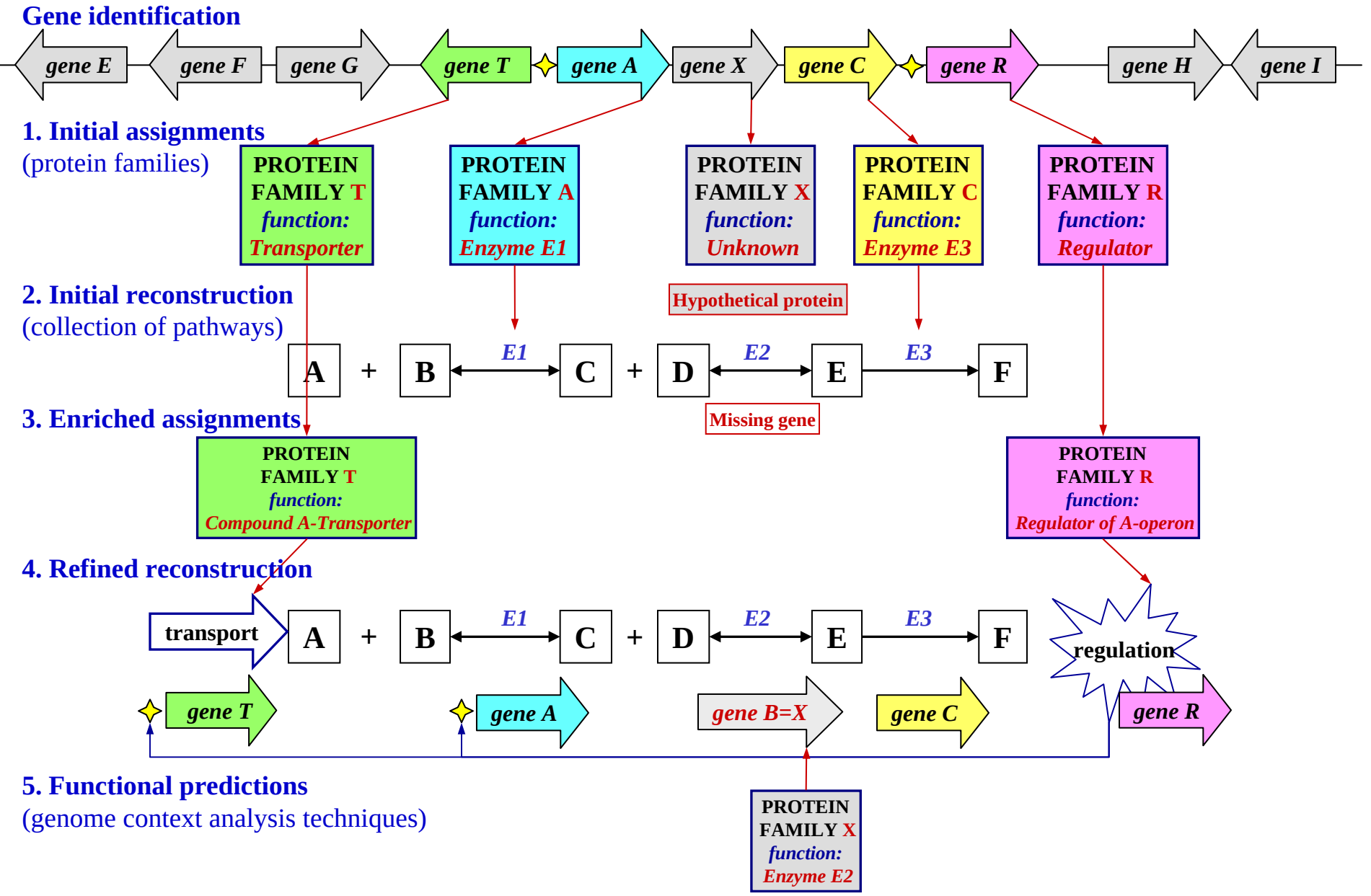

CARBOHYDRATE METABOLISM
Ascorbate and aldarate metabolism
Glycolysis / Gluconeogenesis
Galactose metabolism
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Pentose phosphate cycle
Inositol metabolism
Pyruvate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
Propanoate metabolism
Citrate cycle
Glyoxylate and dicarboxylate metabolism
01110 7/8/02


PENTOSE PHOSPHATE PATHWAY
Entner-Doudroff pathway
Glycolysis
Methane metabolism
Pentose and glucuronate interconversions
Vitamine B6 metabolism
Phenylalanine,tyrosine and tryptophan biosynthesis
Pyrimidine metabolism
Histidine metabolism
Purine metabolism
PRPP
00030 8/6/02


PENTOSE PHOSPHATE PATHWAY
Entner-Doudroff pathway
Glycolysis
Methane metabolism
Pentose and glucuronate interconversions
Vitamine B6 metabolism
Phenylalanine,tyrosine and tryptophan biosynthesis
Pyrimidine metabolism
Histidine metabolism
Purine metabolism
PRPP
00030 8/6/02

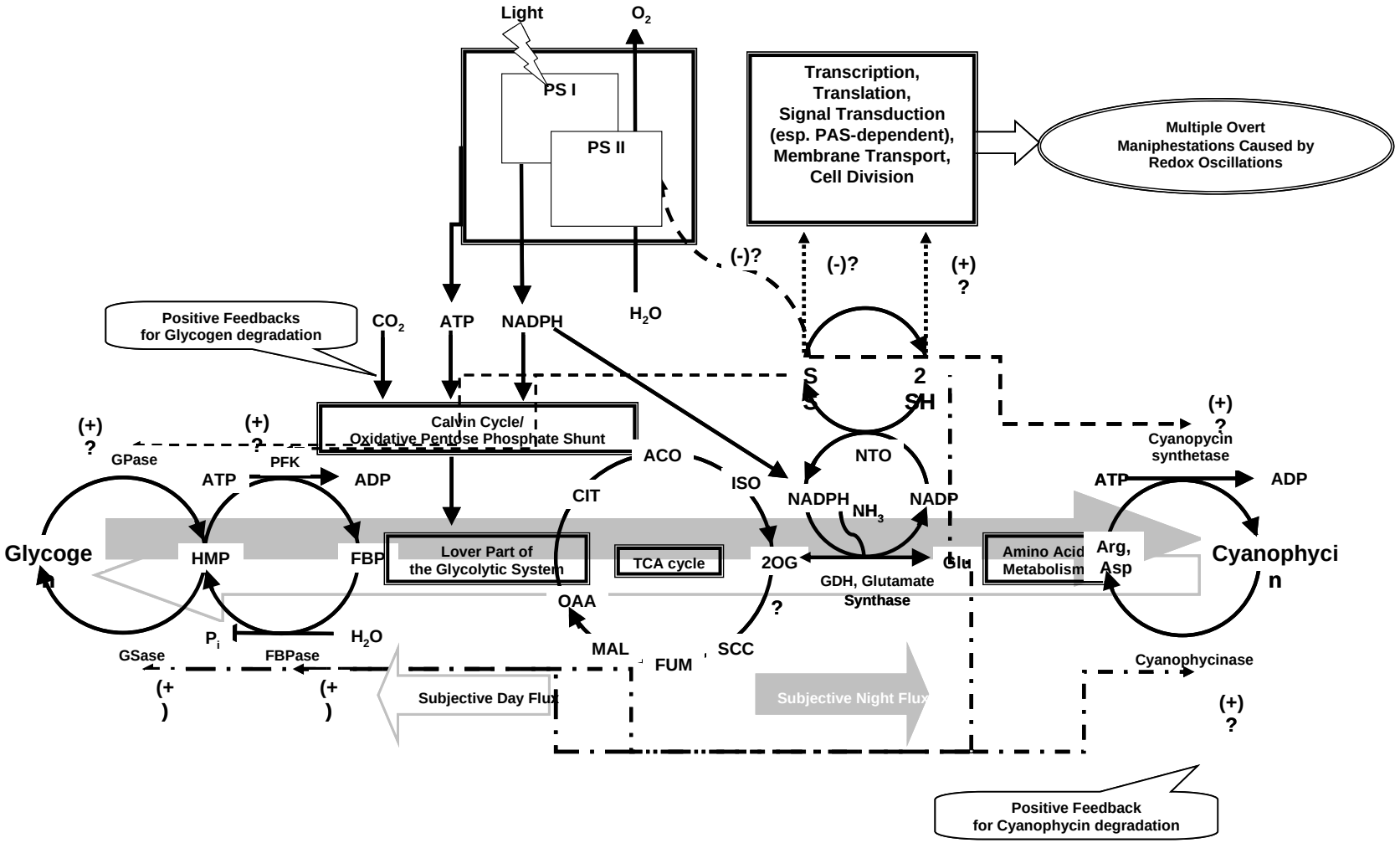
From KEGG

ERGO: Genome Annotation/Reconstruction Pipeline



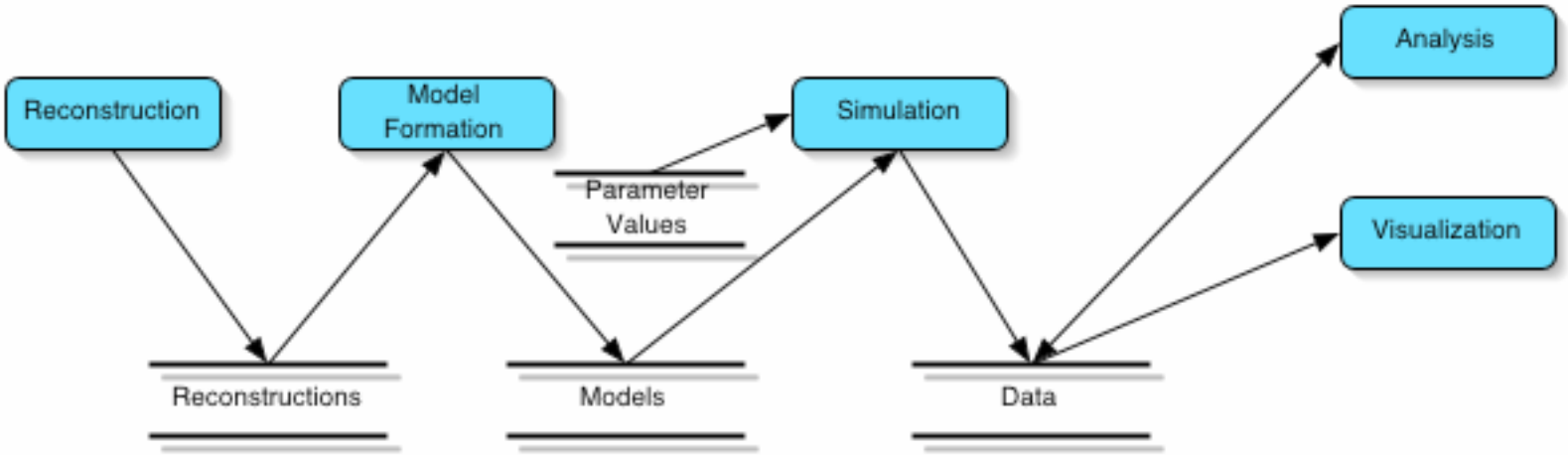
From Integrated Genomics

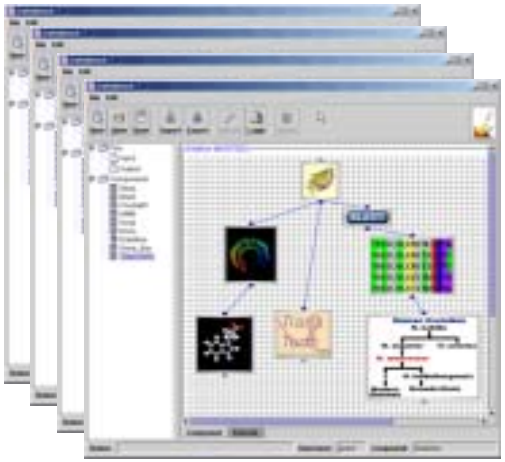
Simplified Model of Cell Clock



From ANL (Selkov et. al.)

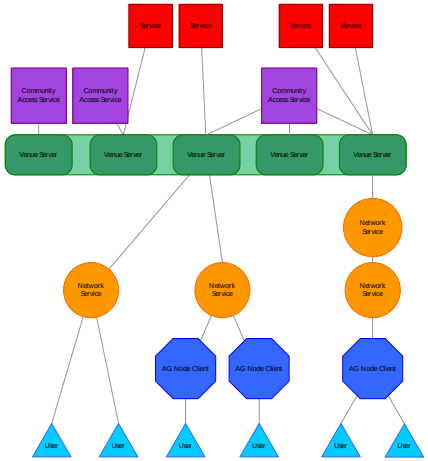
Workflow for Cell Modeling and Simulation





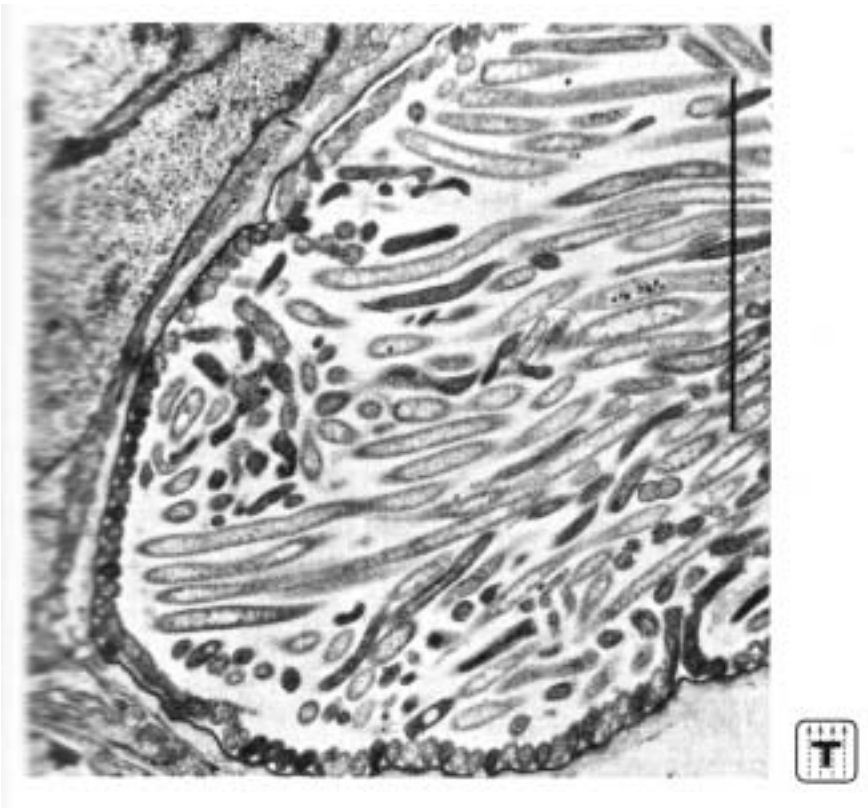
TurboWorx

⊕



Access Grid 2.0

A Diverse Bacterial Community

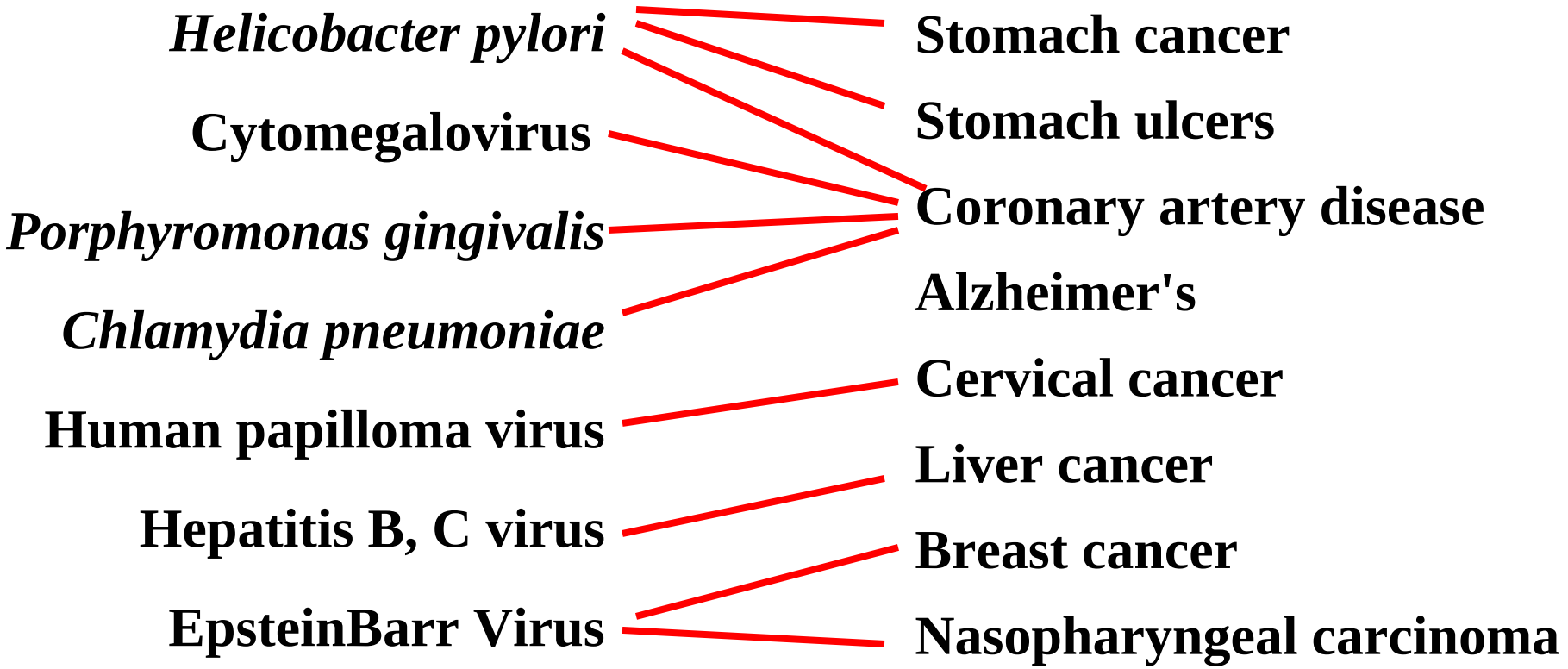


From Five Kingdoms

Lynn Margulis and Karlene Schwartz

- Pocket in the hindgut wall of the Sonoran desert termite *Pterotermes occidentis*
- 10 billion bacteria per milliliter
- Anoxic environment
- ~30 strains are facultative aerobes
- Many/most are unknown

Human Microbial Ecology: Sorting Out Cause, Effect + Cure



Intracellular Environment – Gel Like Media

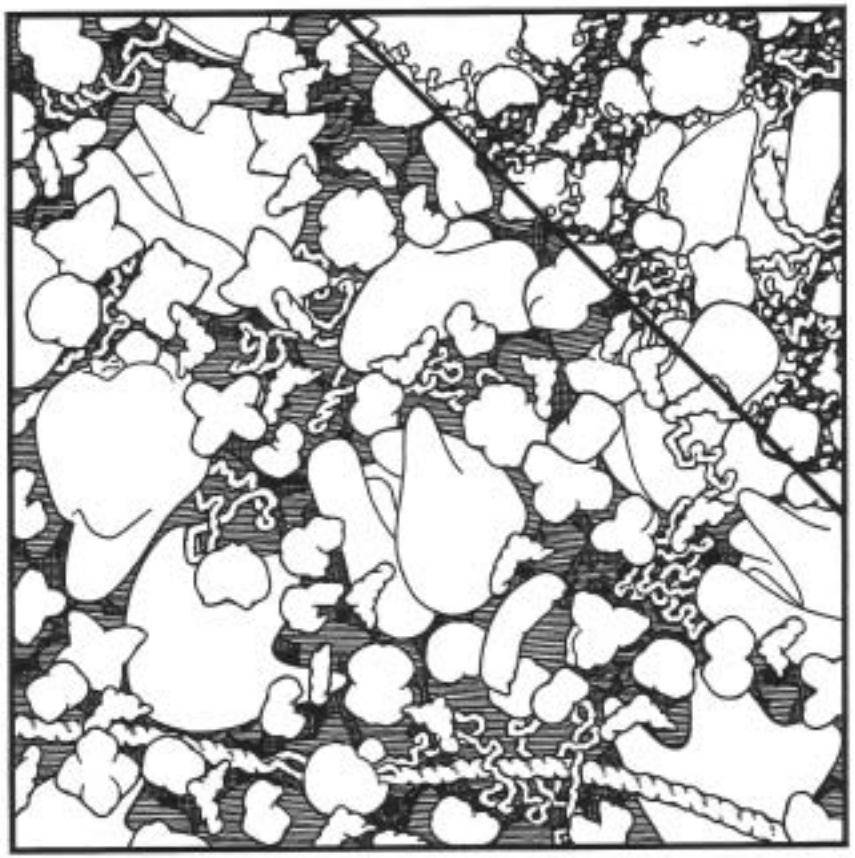


Figure 4.2 Cytoplasm

- 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%

From: David Goodsell, The Machinery of Life

Cell Membranes and Cell Wall

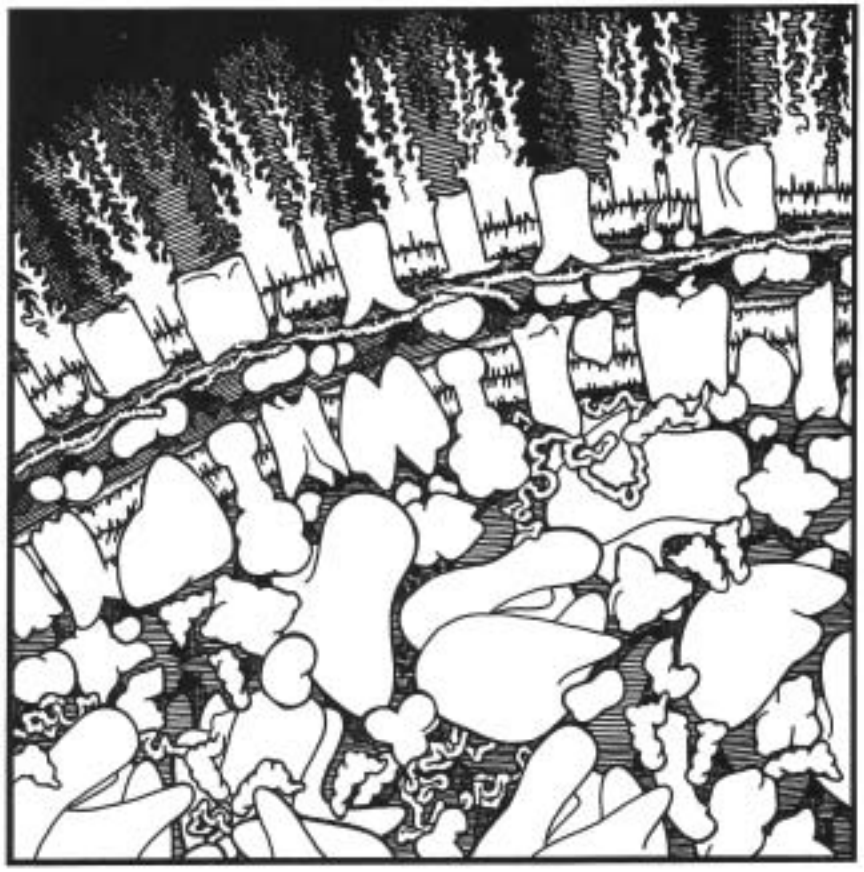


Figure 4.3 Cell Wall

From: David Goodsell, The Machinery of Life

- Cell wall
  - Polysaccharides
  - Porin pores
- Peptidoglycan
  - cross linked
- Periplasmic space
  - Small proteins
- Complex inner membrane
  - < 50% lipids

Flagellum and Flagellar Motor: Nanotechnology

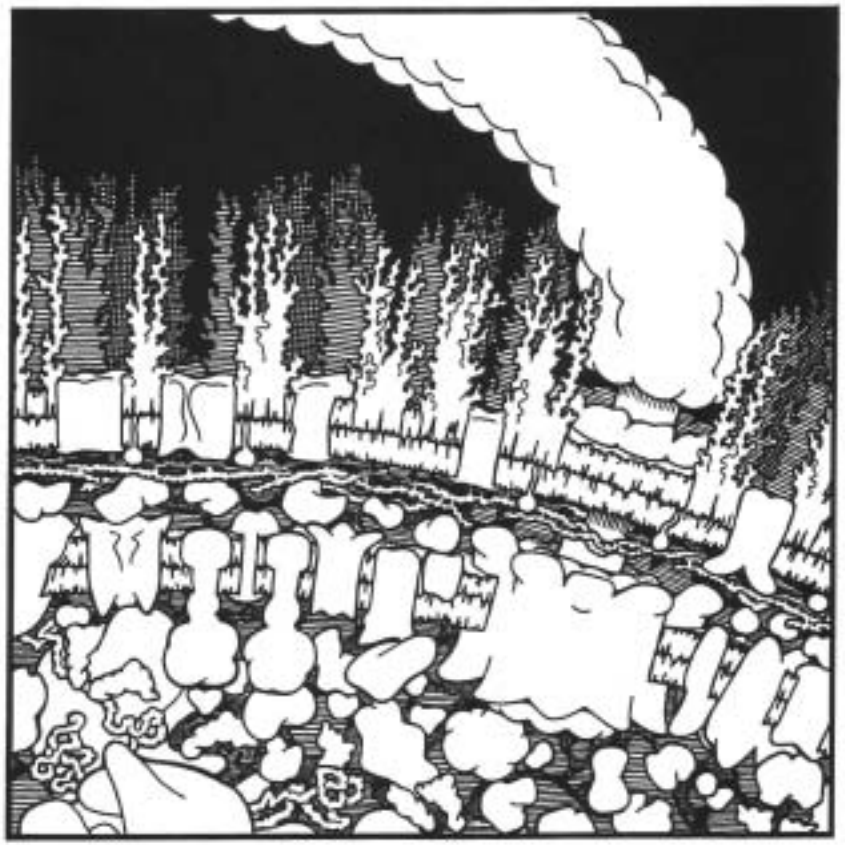


Figure 4.4 Flagellum and Flagellar Motor

- Transmembrane proton powered rotating motor
- About 10 Flagella per cell
- 5-10 um long
  - Built from the inside out
- Propels cell ~10-20 um/sec
  - Medium is extremely viscous
  - 10-20 body lengths/sec
  - ~100KM/hr scaled velocity

From: David Goodsell, The Machinery of Life

DNA Replication via DNA Polymerase

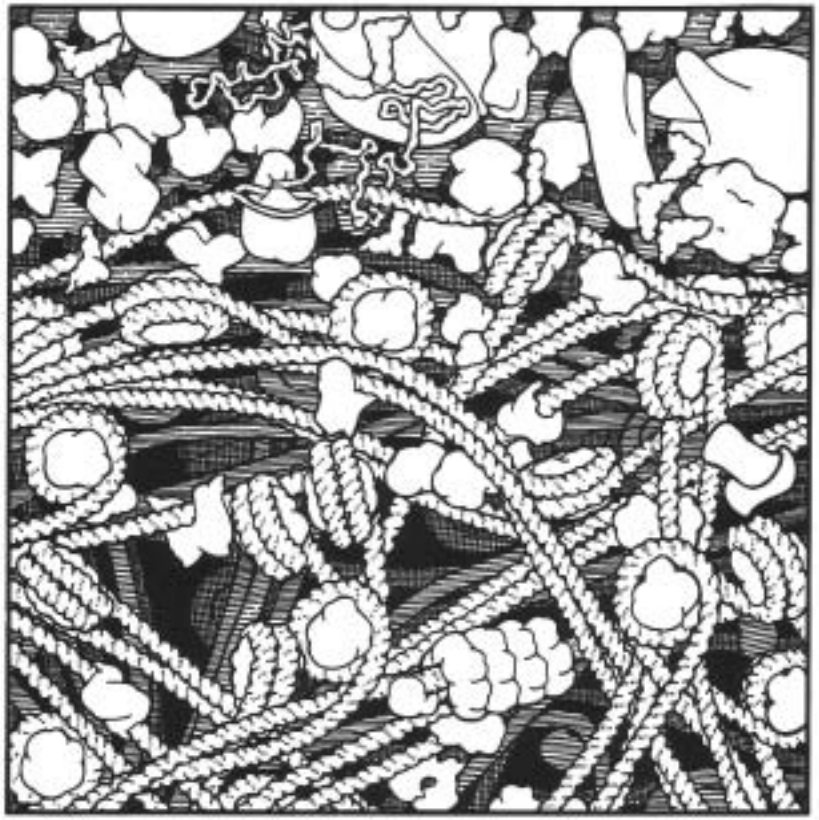


Figure 4.5 Nuclear Region

- DNA replication about 800 new nucleotides per second
- In circular DNA both directions at once
- 50 minute to duplicate entire circle of 4,700,000 nucleotides
- With cell replication ~30 minutes
- DNA replication is pipelined!!

From: David Goodsell, The Machinery of Life

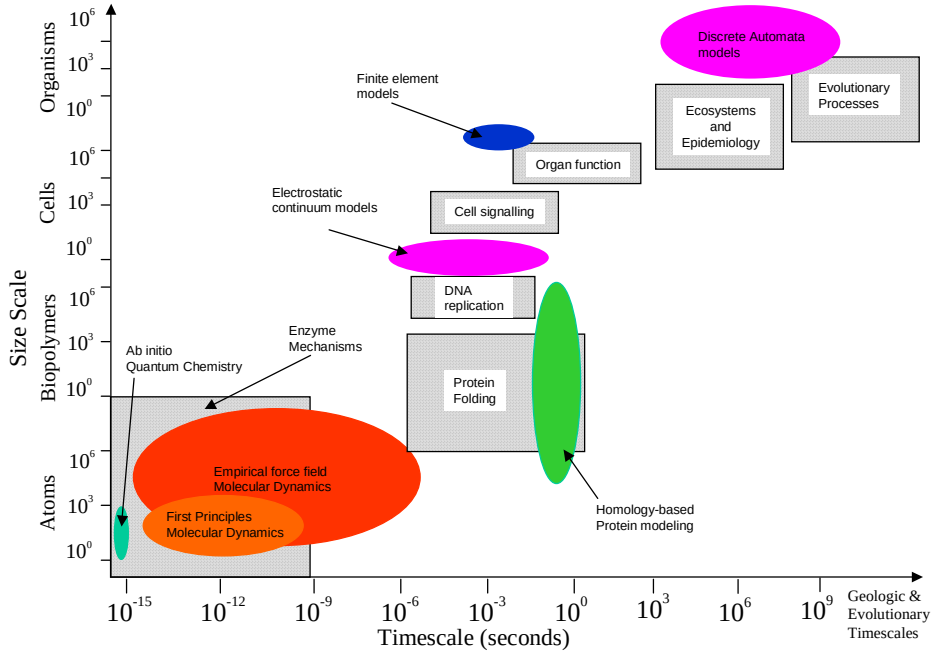
Systems Biology Model Development

Systems	Director	Institution	Features
ERATO/SBW,j	John Doyle	Caltech	planned workbench
Gepasi,w	Pedro Mendes	Santa Fe	MCA, systems kinetics
JarnacScamp,wx	Herbert Sauro	Caltech	MCA, Stochastic
StochSim,w+	Dennis Bray	Cambridge	Stochastic
BioSpice,u	Adam Arkin	LBL	Stochastic
DBSolve,w	Igor Goryanin	Glaxo	enzyme/receptor-ligand
E-Cell,u+	Masaru Tomita	Keio	metabolism. Net ODE
Vcell,j	Jim Schaff	U.CT	geometry
Xsim,u	J.Bassingthwaighte	Seattle	enzymes to body physiology
CellML,x+	Peter Hunter	U.Auckland	geometry, model sharing
GENESIS,u	James Bower	Caltech	neural networks
Simex,u+	Lael Gatewood	U.MN	Stochastic micro populations

J = java, w = windows, u=unix, x=XML, + = source/community input

Hierarchical Modeling in Biological Systems

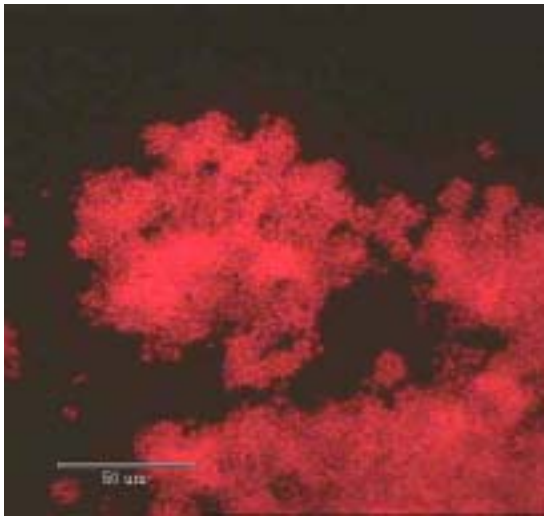
- Genetic Sequences
- Molecular Machines
- Molecular Complexes and modules
- Networks + Pathways [metabolic, signaling, regulation]
- Structural components [ultrastructures]
- Cell Structure and Morphology
- Extracellular Environment
- Populations and Consortia etc.



MONERA: A Mathematical Toolkit for Modeling Biological Systems

"A Mathematica for molecular, cellular and systems biology"

- Core data models and structures in the bio domain
  - From genes and molecules to populations and communities
- Optimized functions for high-performance computers
  - Parallel and hardware accelerated kernels
  - Simulation interfaces to available tools
- Grid enabled Scripting environment [e.g. Python, PERL, etc.]
  - Distributed data access
  - Distributed computation
- Database accessors and built-in schemas
- Visualization interfaces [info-vis and sci-vis]
- Collaborative workflow and group use interfaces
  - AG services model integration

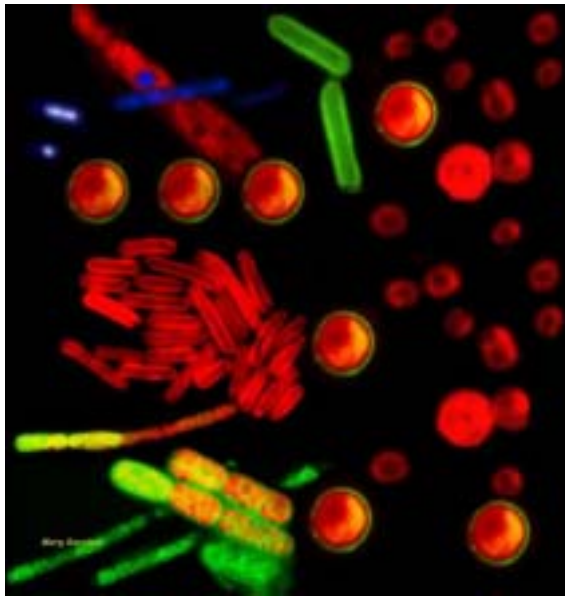


mcs

A = Algorithms
C = Compute
P = Parallelism
I = Integration

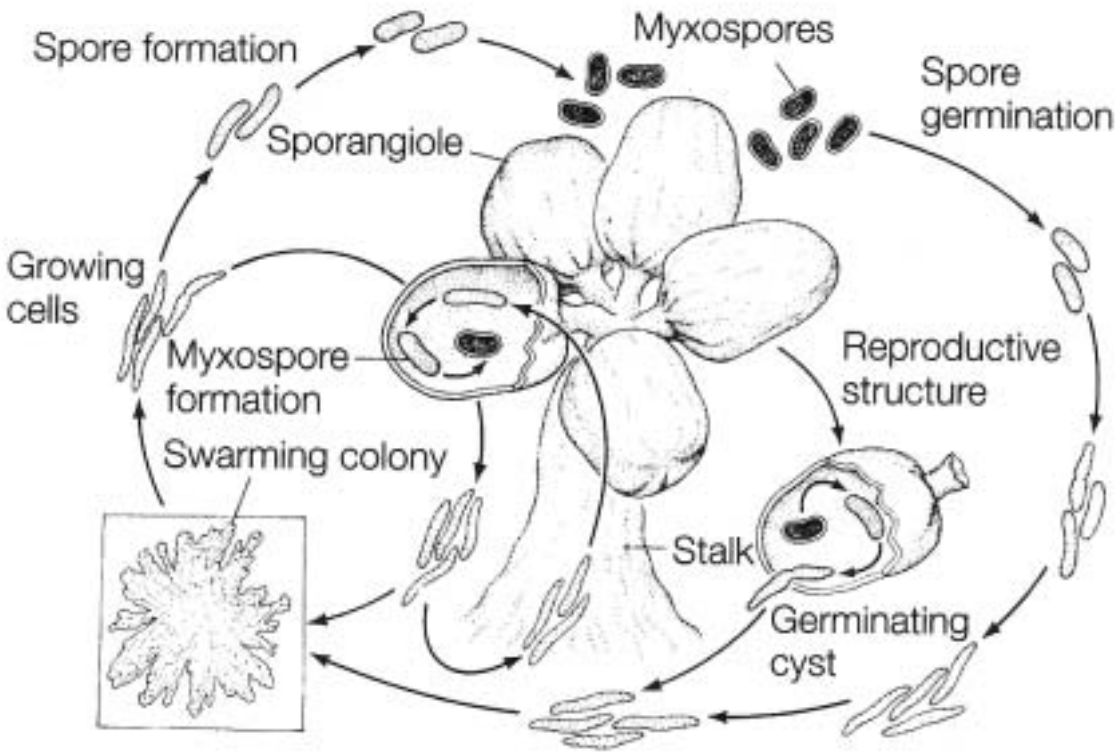
Paths to Whole Cell Simulations

- Unregulated metabolic model (flux analysis)
- Allosteric regulation (binding changes conformation) (A)
- Gene Regulated + Metabolic Model (A, C)
- Heterogeneous/Compartmentalized/Diffusion (A,C,P)
- Active Regulation + Transport (A,C,P,I)
- Complete Integrated Cell (geometry) (A,C, P,I)

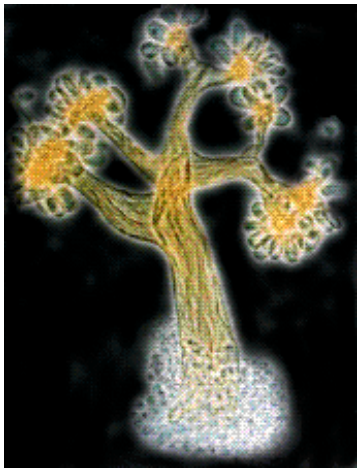


- Multicellular models (homogeneous) (P)
- Multicellular (homo) with complex communication (P)
- Multicellular (hetero) mixed population (P, I)
- Multicellular differentiation and motility (A, C, P, I)
- Multicellular structures with complex geometry $(A,C,P, I)^2$

Understanding Bacterial Life Cycles

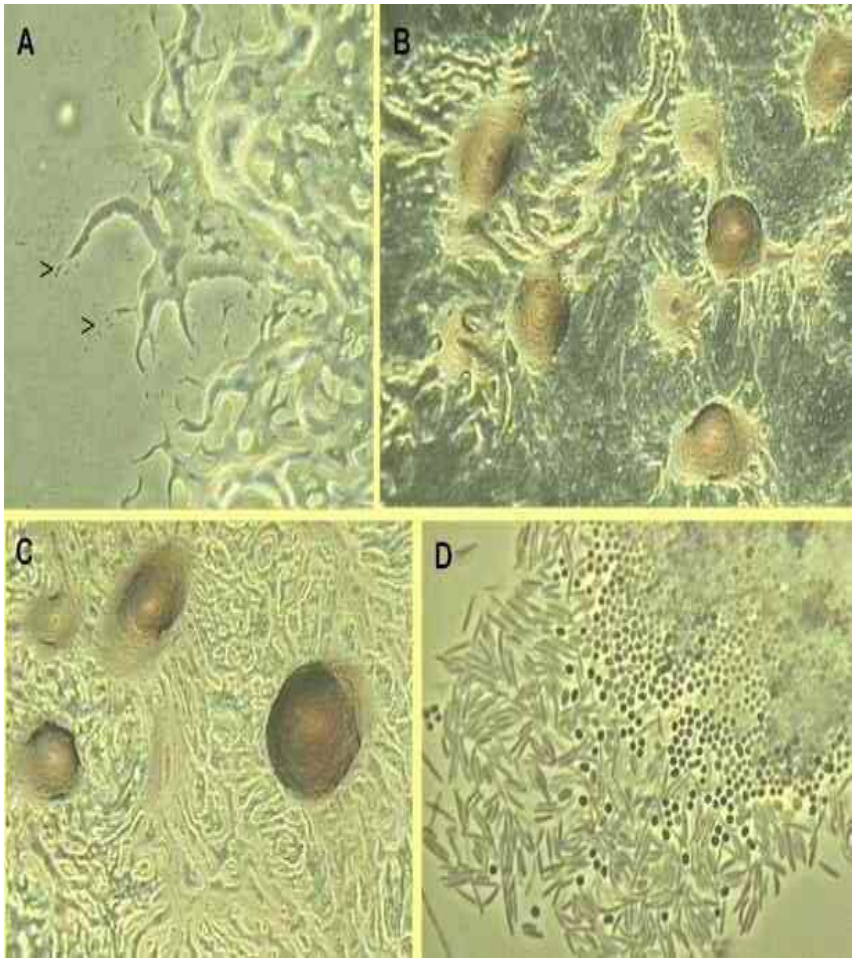


F Life cycle of *Stigmatella aurantiaca*. [Drawing by L. Meszoly; labeled by M. Dworkin.]



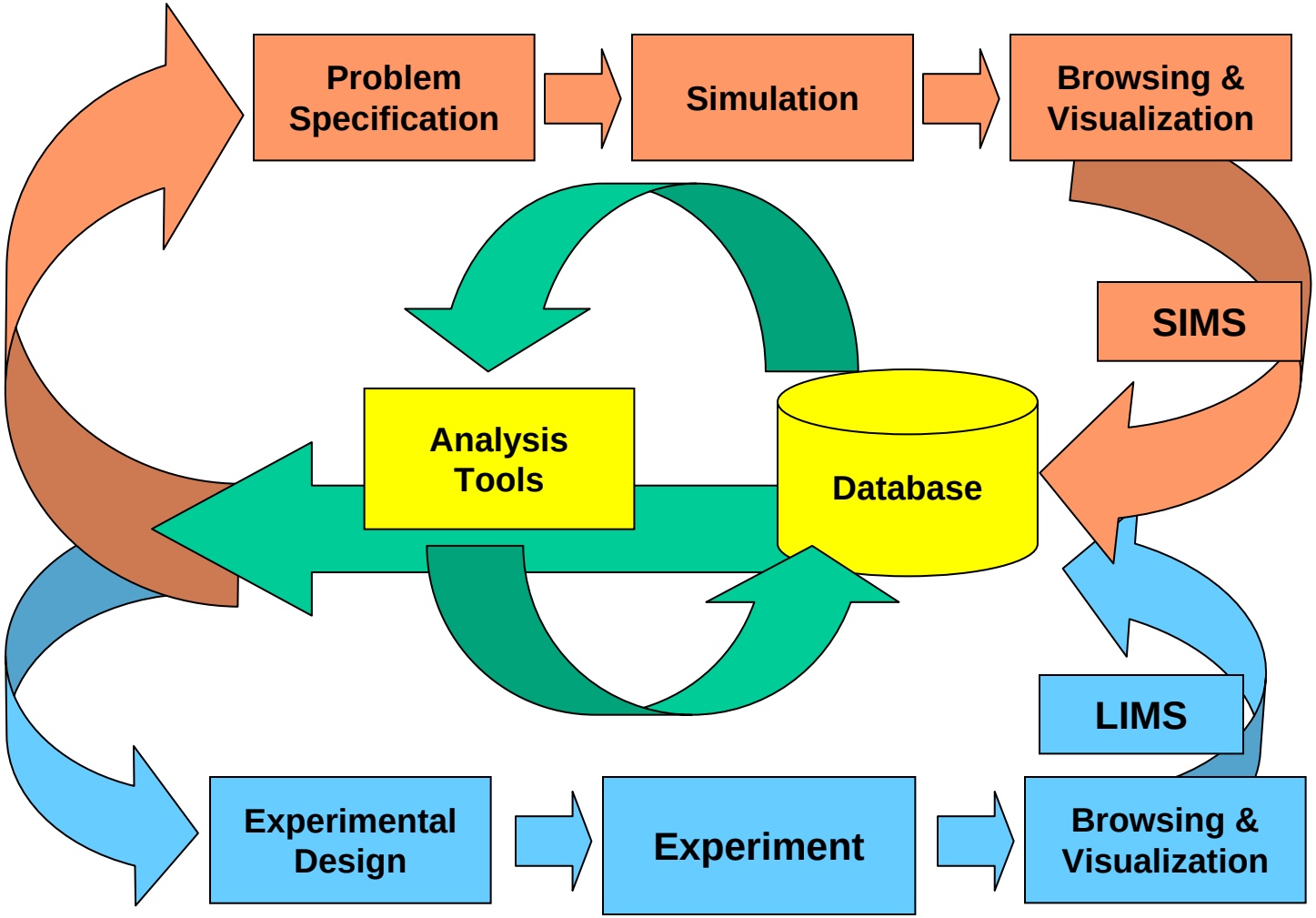
From Lynn Margulis and Karlene Schwartz

Modeling Swarming Behavior in *Myxobacteria*



- 100,000 cells swarm to form fruiting bodies
- 80% of the cells lyse
- 20% form spores
- Involves chemotaxis and quorum sensing
- Most complex bacterial genome currently known at > 9Mbp
- Very little is understood

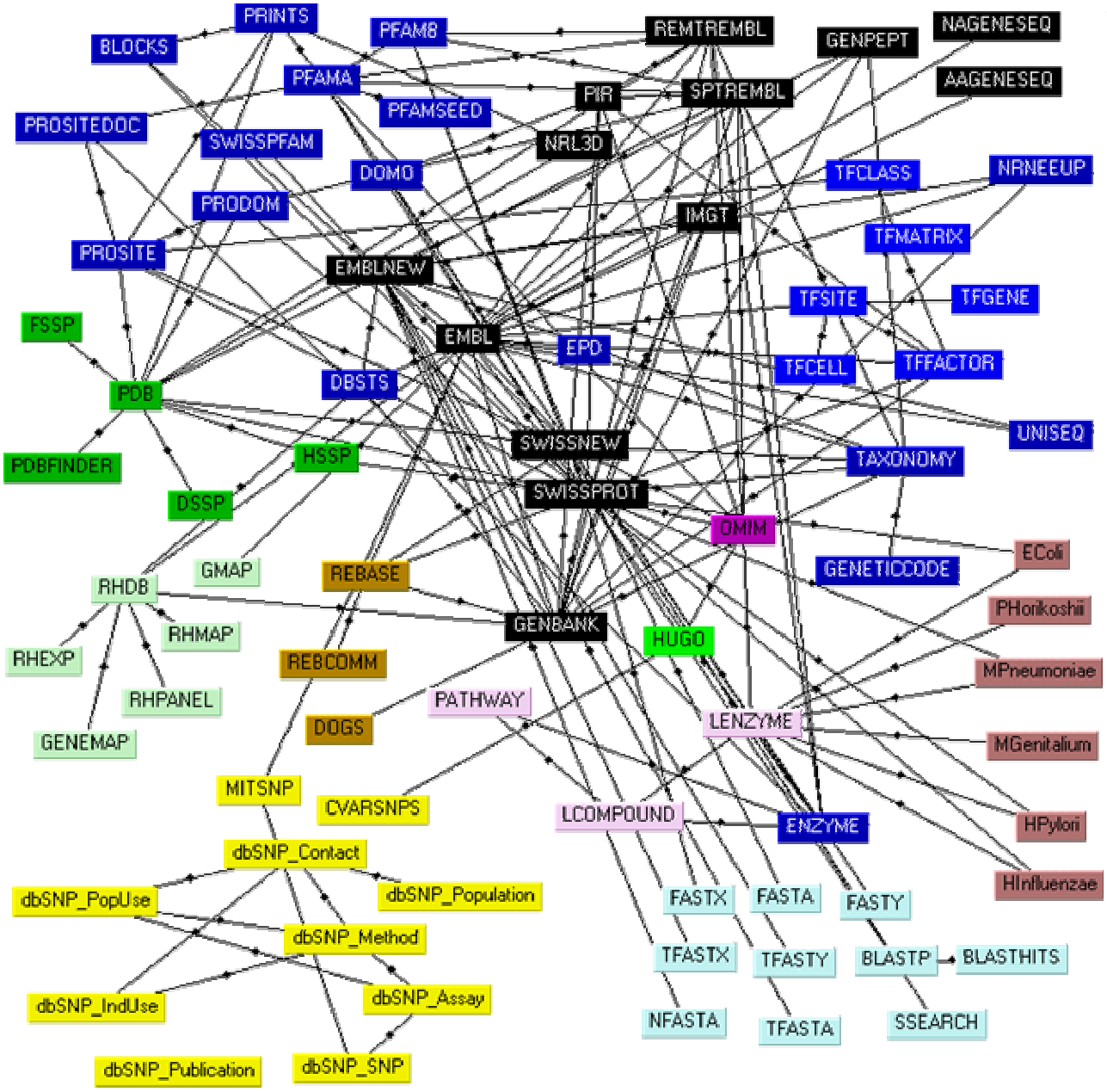
An Integrated View of Simulation, Experiment, and Bioinformatics



Biology Databases (335 in 2001)

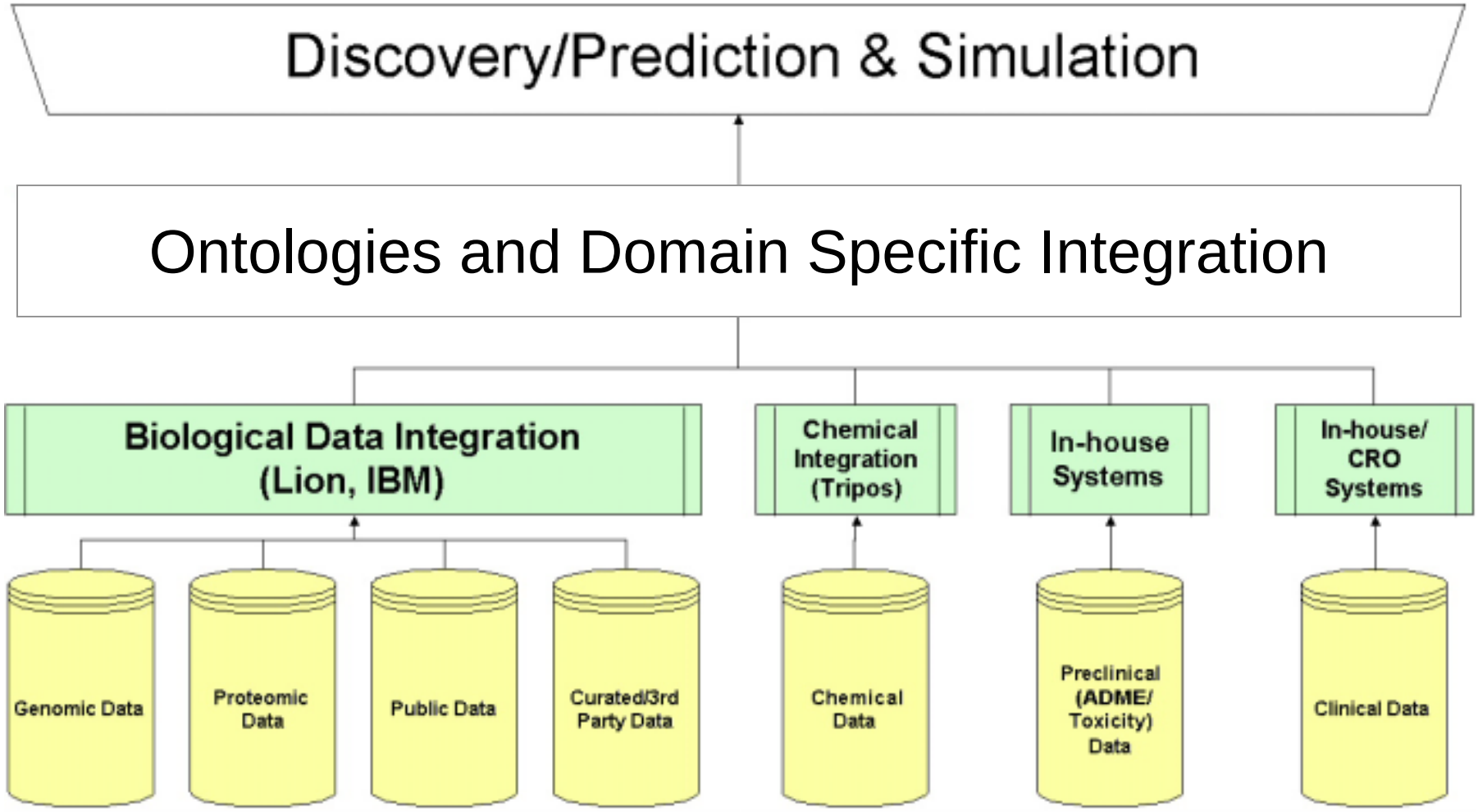
- Major Seq. Repositories (7)
- Comparative Genomics (7)
- Gene Expression (19)
- Gene ID & Structure (31)
- Genetic & Physical Maps (9)
- Genomic (49)
- Intermolecular Interactions (5)
- Metabolic Pathways & Cellular Regulation (12)
- Mutation (34)
- Pathology (8)
- Protein (51)
- Protein Sequence Motifs (18)
- Proteome Resources (8)
- Retrieval Systems & DB Structure (3)
- RNA Sequences (26)
- Structure (32)
- Transgenics (2)
- Varied Biomedical (18)

Baxevanis, A.D. 2002. *Nucleic Acids Research* 30: 1-12.


PRINTS
BLOCKS
PFAM8
REMTREMBL
GENPEPT
NAGENESEQ
PFAMA
PIR
SPTREMBL
AAGENESEQ
PFAMSEED
PROSITEDOC
SWISSPFAM
NRL3D
DOMO
TFCLASS
NRNEEUP
PRODOM
IMGT
TFMATRIX
PROSITE
EMBLNEW
TFSITE
TFGENE
FSSP
EMBL
EPD
TFCELL
TFFACTOR
PDB
DBSTS
UNISEQ
SWISSNEW
PDBFINDER
HSSP
TAXONOMY
SWISSPROT
DSSP
OMIM
EColi
GMAP
REBASE
GENETICCODE
RHDB
PHorikoshii
GENBANK
RHMAP
HUGO
RHEXP
REBCOMM
MPneumoniae
PATHWAY
RHPANEL
LENZYME
DOGS
GENEMAP
MGenitalium
MITSNP
CVARSNPS
LCOMPOUND
ENZYME
HPylori
dbSNP_Contact
HInfluenzae
dbSNP_PopUse
dbSNP_Population
FASTX
FASTA
FASTY
dbSNP_Method
TFASTX
TFASTY
BLASTP
BLASTHITS
dbSNP_IndUse
dbSNP_Assay
NFASTA
TFASTA
SSEARCH
dbSNP_Publication
dbSNP_SNP

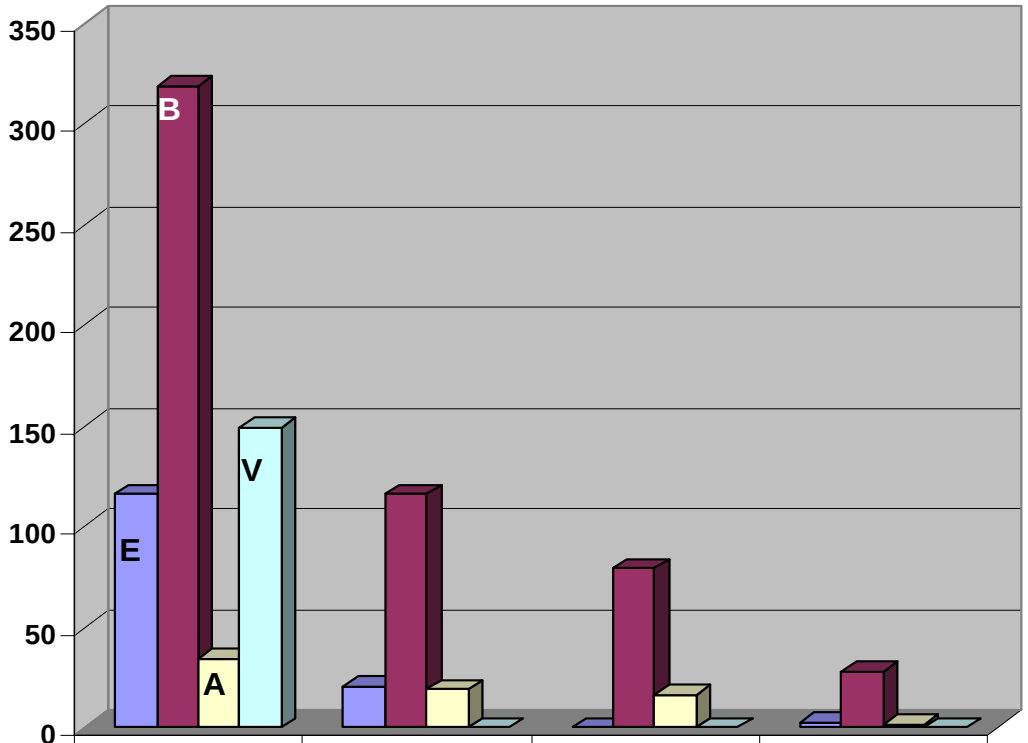
LION

Software Infrastructure in Drug Discovery



Adapted From Richard Gardner(InCellico)

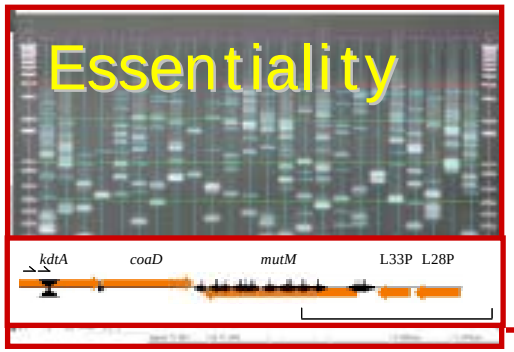
ERGO: Integration of Genomic Data

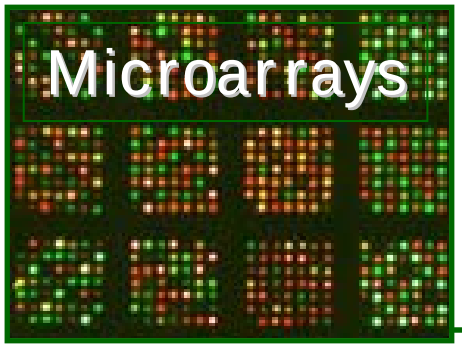


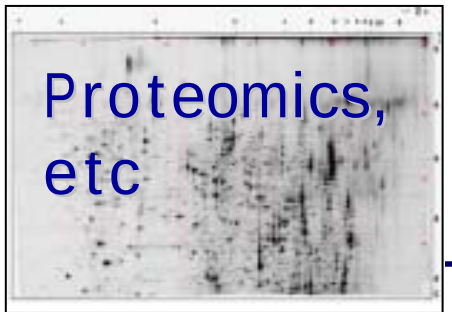
	ERGO	PEDANT	TIGR	GTC
Eukarya	116	21	0	3
Bacteria	319	116	79	28
Archaea	34	19	16	2
Viral	149	0	0	0

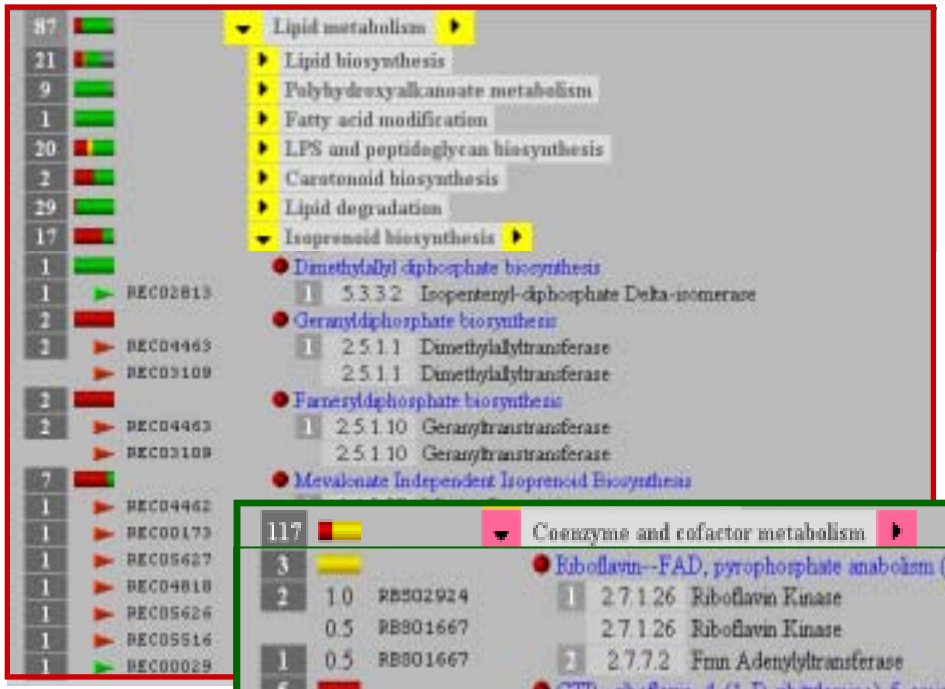
A framework for integrating functional data with genomic sequence data

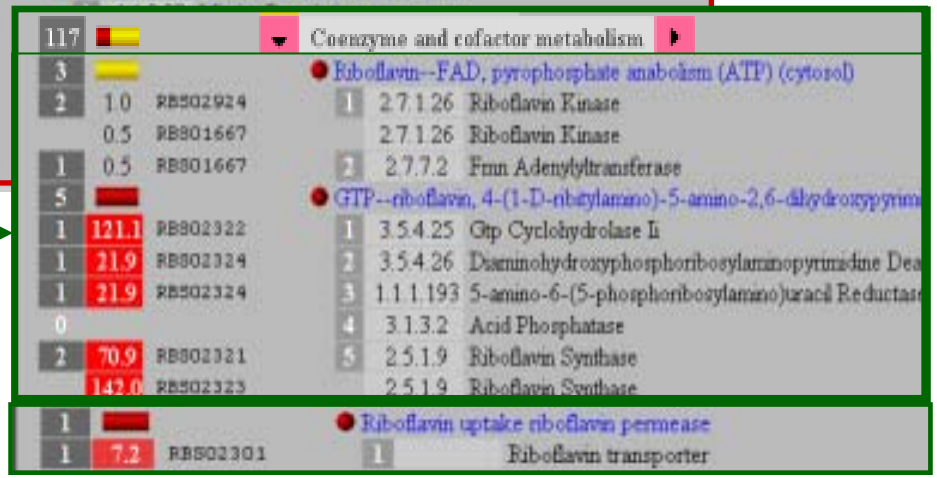
projection of post-genomic data onto functional reconstruction



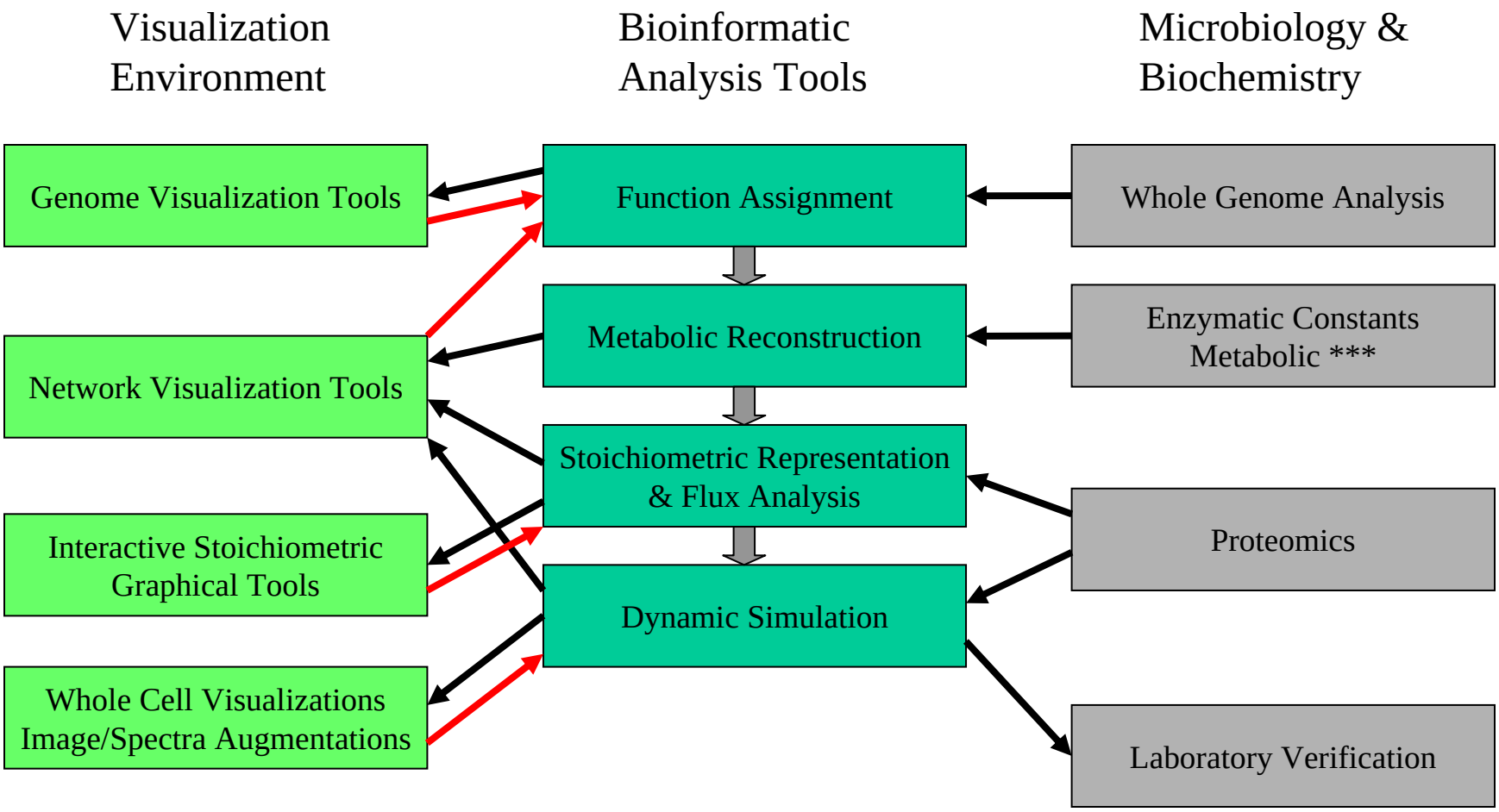


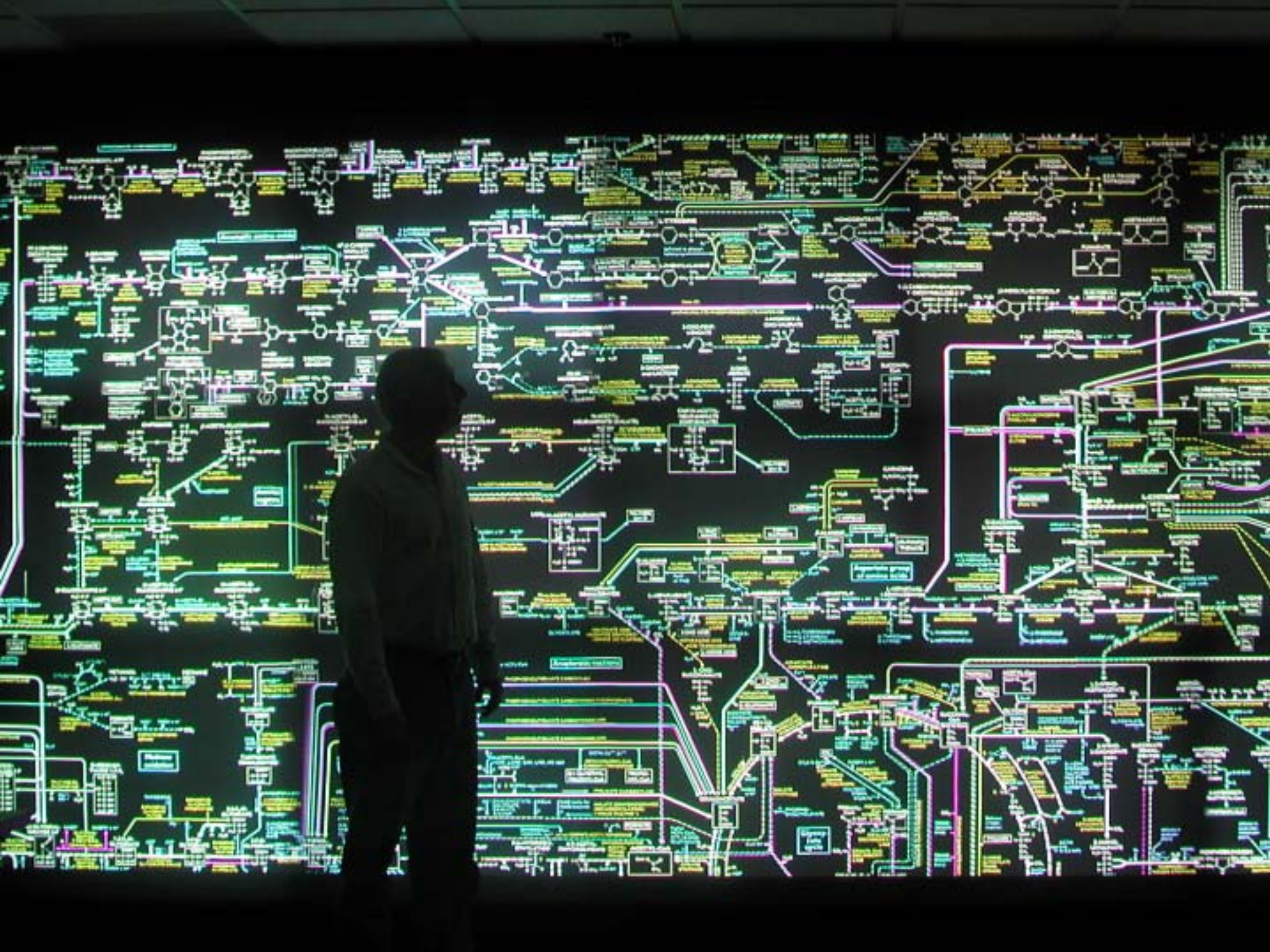




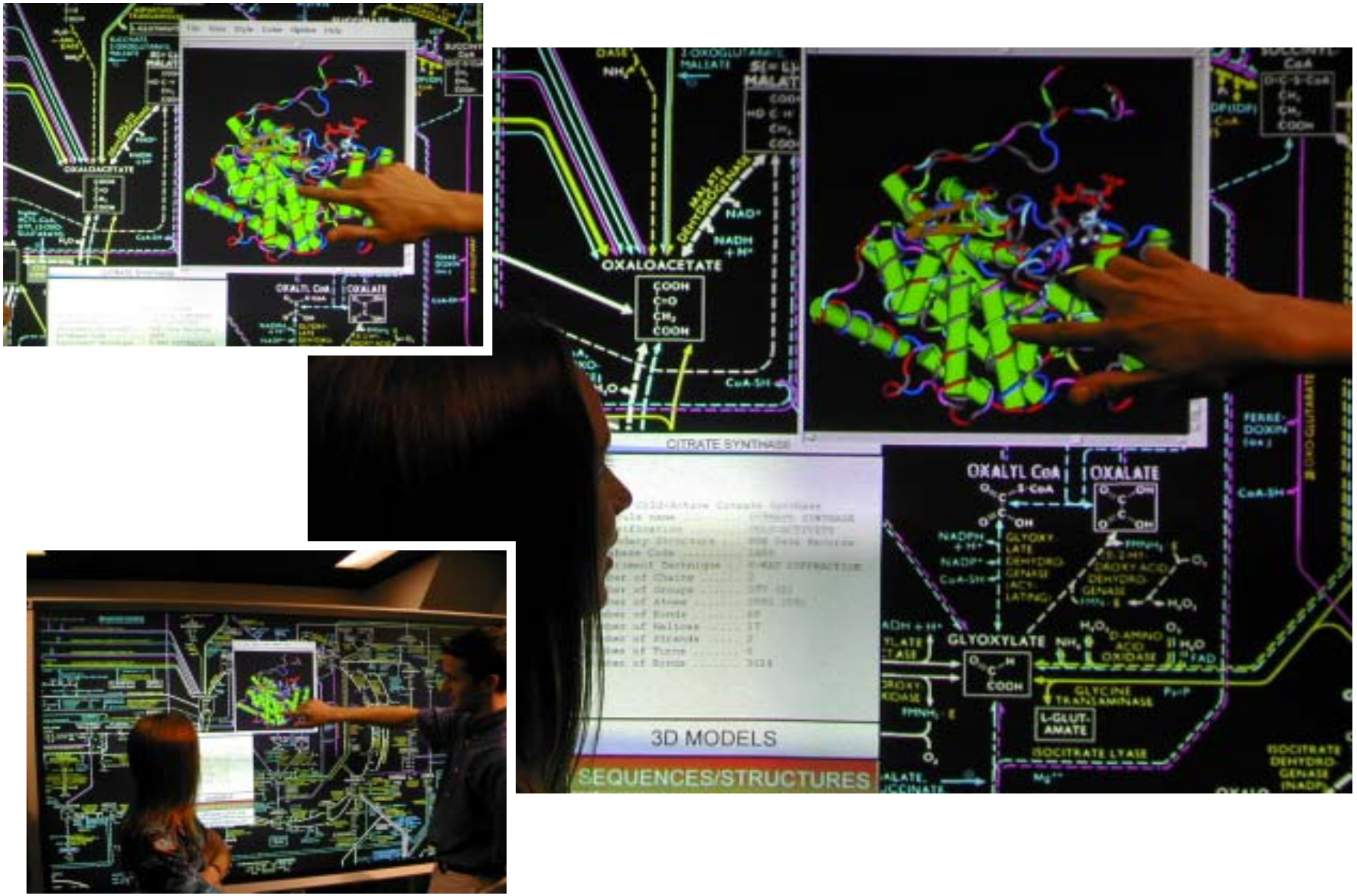


Visualization + Bioinformatics

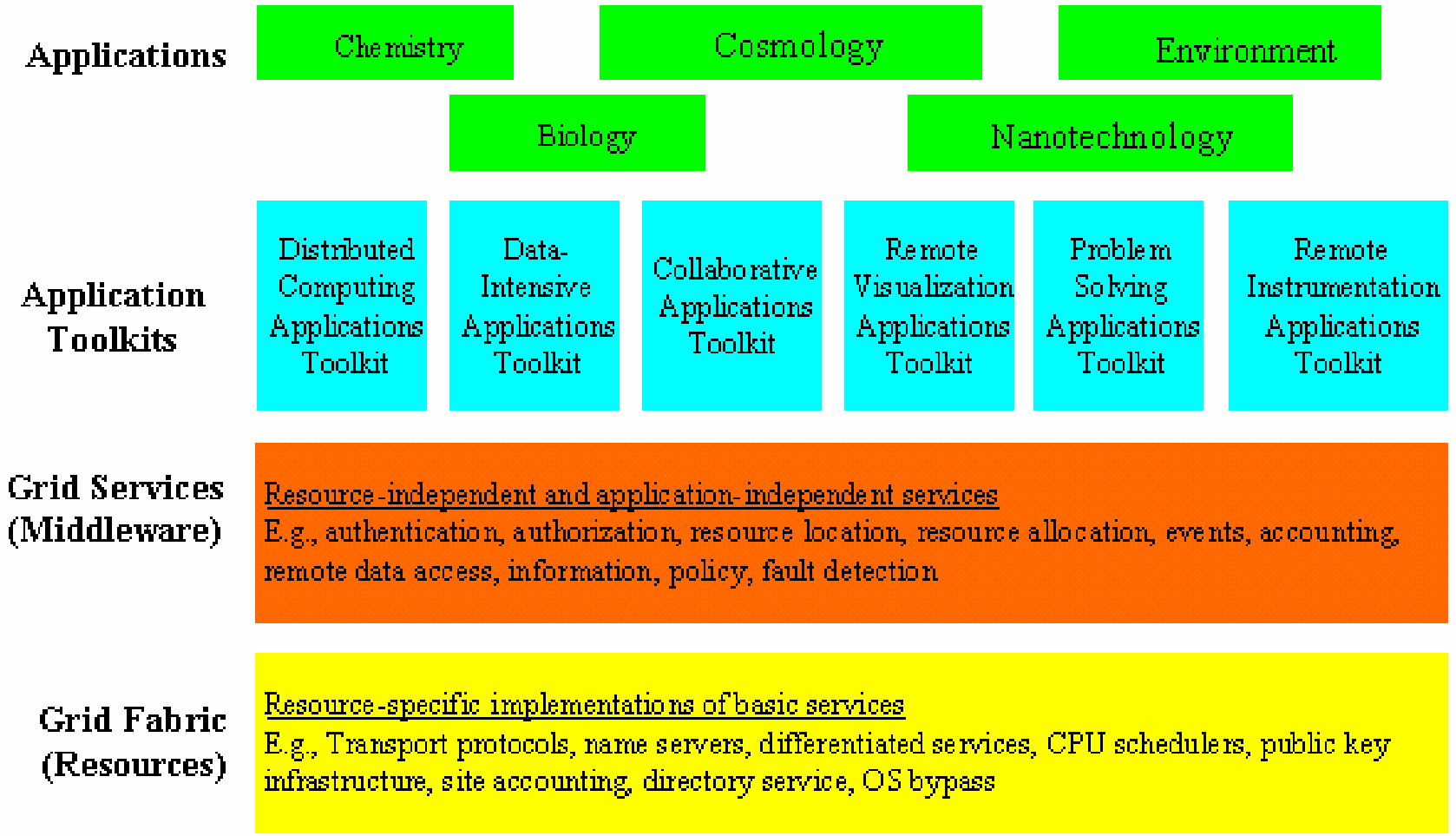




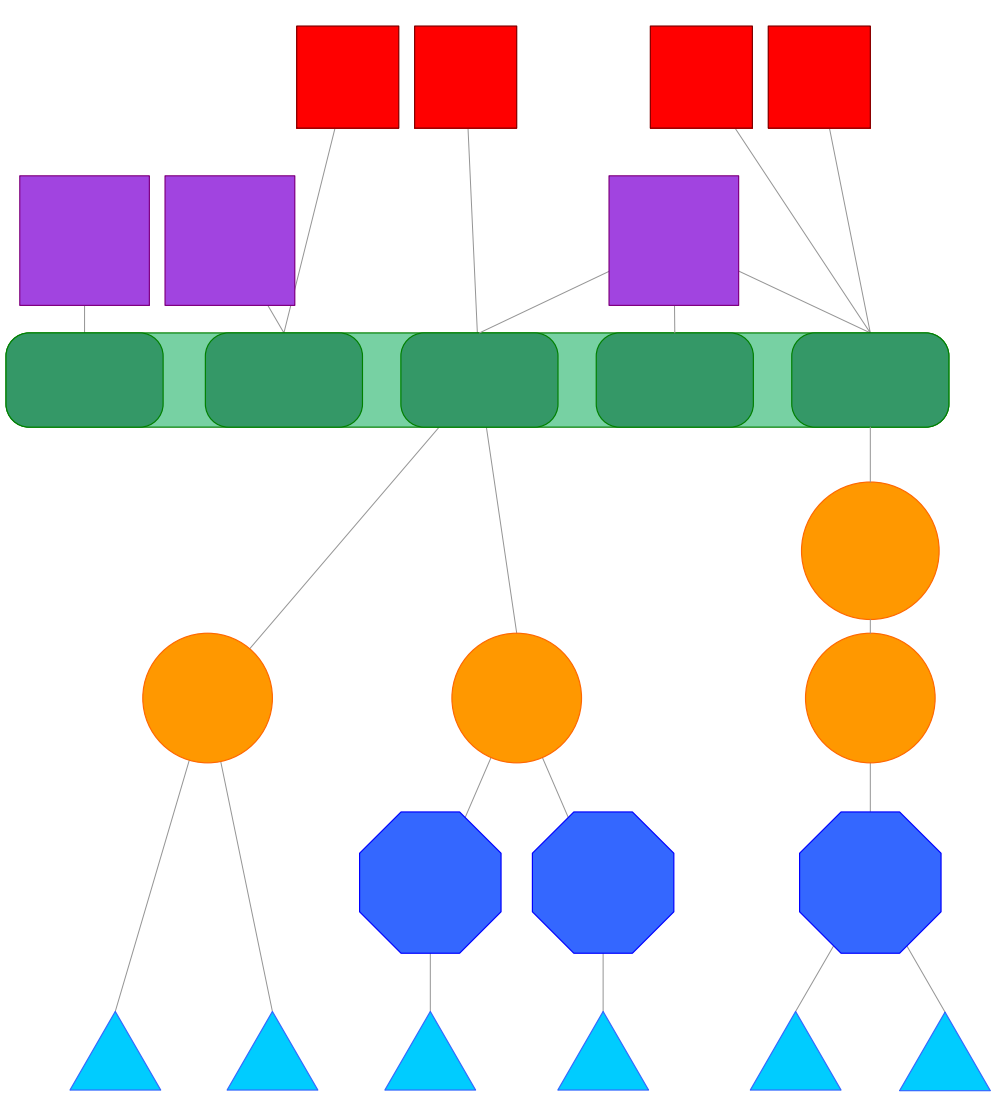
Pathway Explorer on µMural Tiled Display



The Grid Software Stack



Access Grid 2.0 Architecture



Application Services

Virtual Venues Services

Access Grid
Network Services

Service Servic

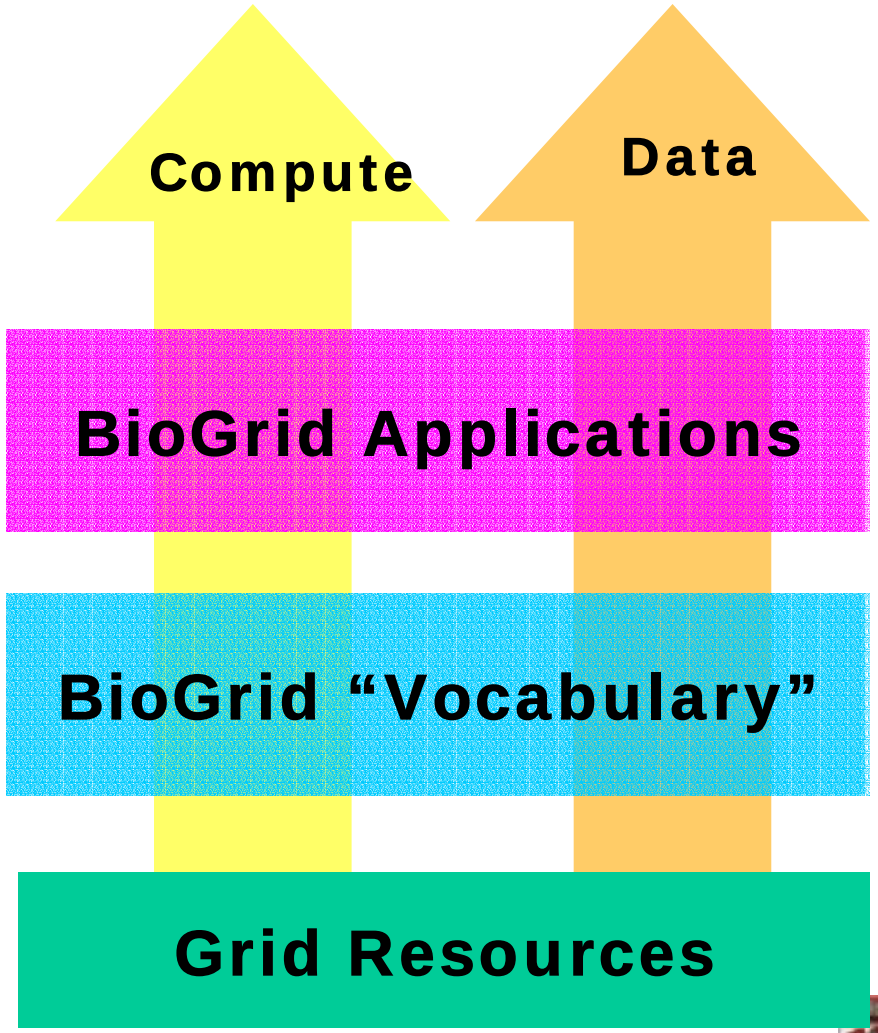
Access Grid
Nodes

Users

Community Community

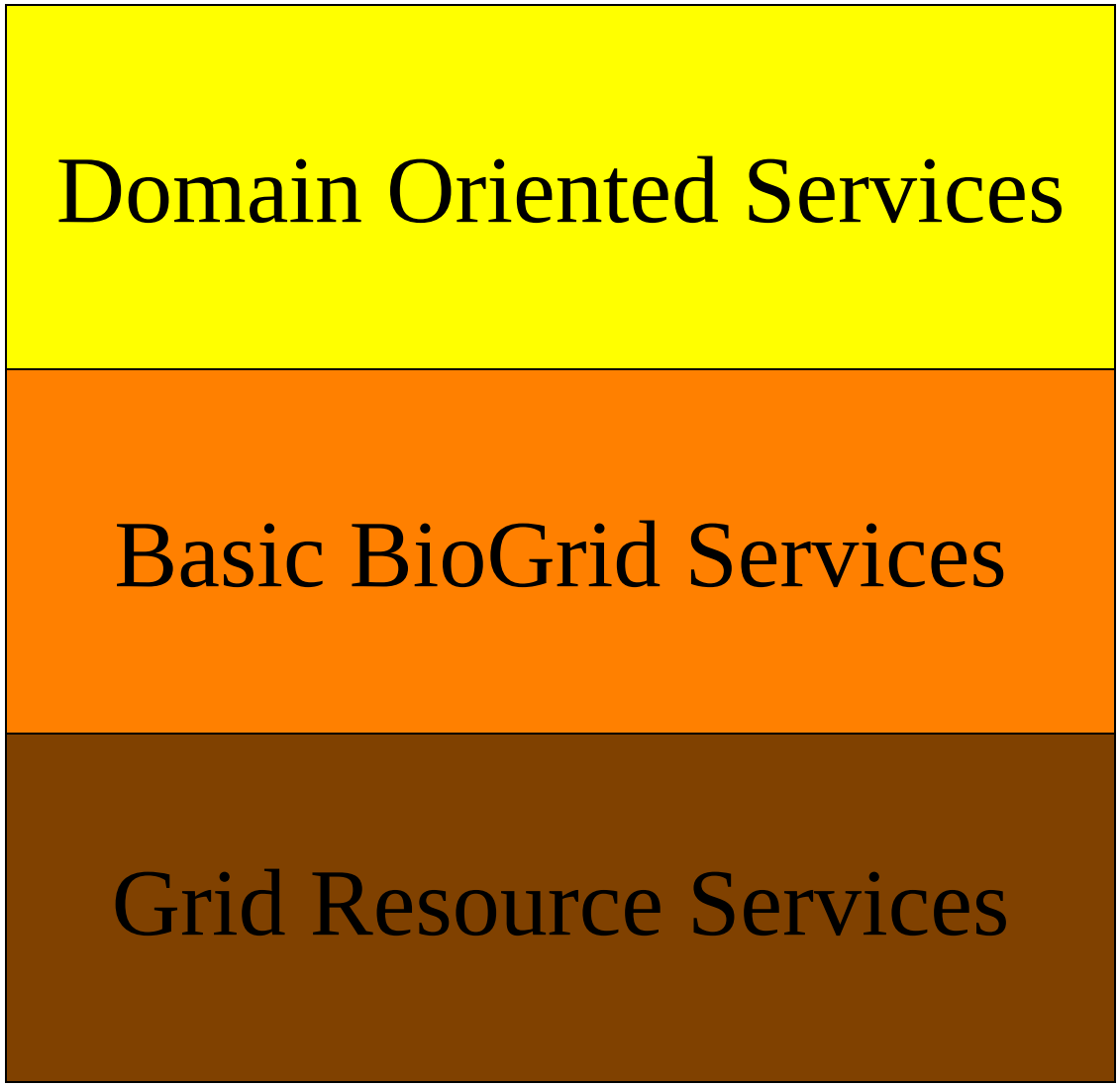
What We Need to Create

- Grid Bio applications enablement software layer
  - Provide application's access to Grid services
  - Provides OS independent services
- Grid enabled version of bioinformatics data management tools (e.g. DL, SRS, etc.)
  - Need to support virtual databases via Grid services
  - Grid support for commercial databases
- Bioinformatics applications "plug-in" modules
  - End user tools for a variety of domains
  - Support major existing Bio IT platforms



From Mark Miller, SDSC

An Example BioGrid Services Model

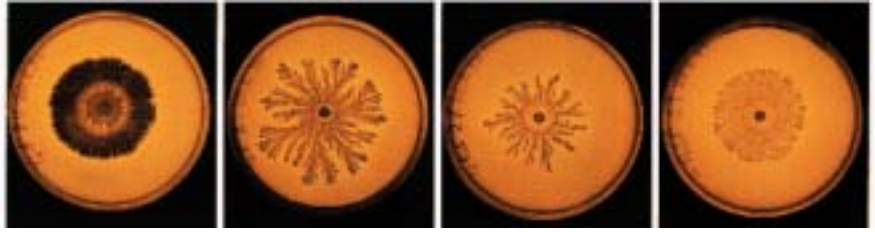


- Drug Discovery
- Microbial Engineering
- Molecular Ecology
- Oncology Research

- Integrated Databases
- Sequence Analysis
- Protein Interactions
- Cell Simulation

- Compute Services
- Workflow Services
- Data Service
- Collaboration Services

Conclusions



- Biology is well positioned to co-dominate Grid applications for the next several decades
- Biological and Biomedical applications of Grids will require dramatic increases in both capability computing and capacity computing
- Data intensive computing is an important aspect of biological applications and will help drive high performance and high-function databases
- Biology and Grids are well suited for each other

Acknowledgements

- DOE, NSF, ANL, UC, Microsoft and IBM for support
- Ian Foster (ANL/UC), Dan Reed (NCSA), John Wooley (UCSD), Mike Colvin(LLNL/DOE), Richard Gardner (InCellico), and others contributed to this talk

THE UNIVERSITY OF CHICAGO

ARGONNE NATIONAL LABORATORY
OPERATED BY THE UNIVERSITY OF CHICAGO

DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

NSF

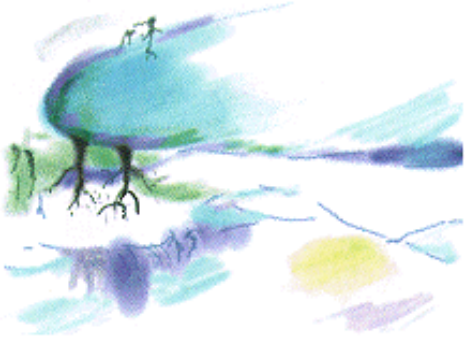
mcs

Extra Slides

Will Biology Dominate the Grid?

- The largest science discipline:
  - The most scientists (Globally ~500,000-1,000,000)
  - The most research funding (Globally ~$50 Billion/year)
  - The most graduate students (>20,000 year)
- Strong couplings to:
  - Medicine and human health
  - Agriculture and food supplies
  - Energy and ecology
  - Future industrial processes (bio-nano)
  - Consumer of other scientific technologies

Computer Science Barriers

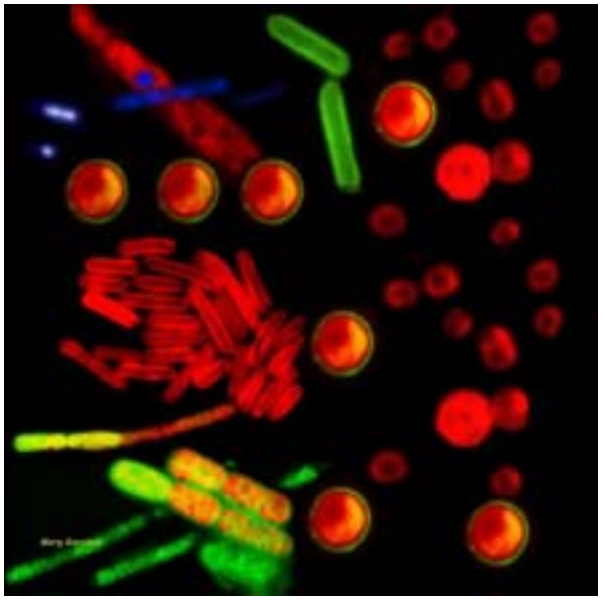


- Framework for Functional Composability
  - Multiple modules
  - Multiple time scales and space scales
  - Empirical, semi-empirical, phenomological, data driven
- Interpretation of output of complex models
  - Visualization and automated interpretation
- Algorithms
  - Parameter estimation, graph theory, combinatorics
- Architectures and Software
  - Issues with scaling models and performance
  - Control and synchronization of multi component models

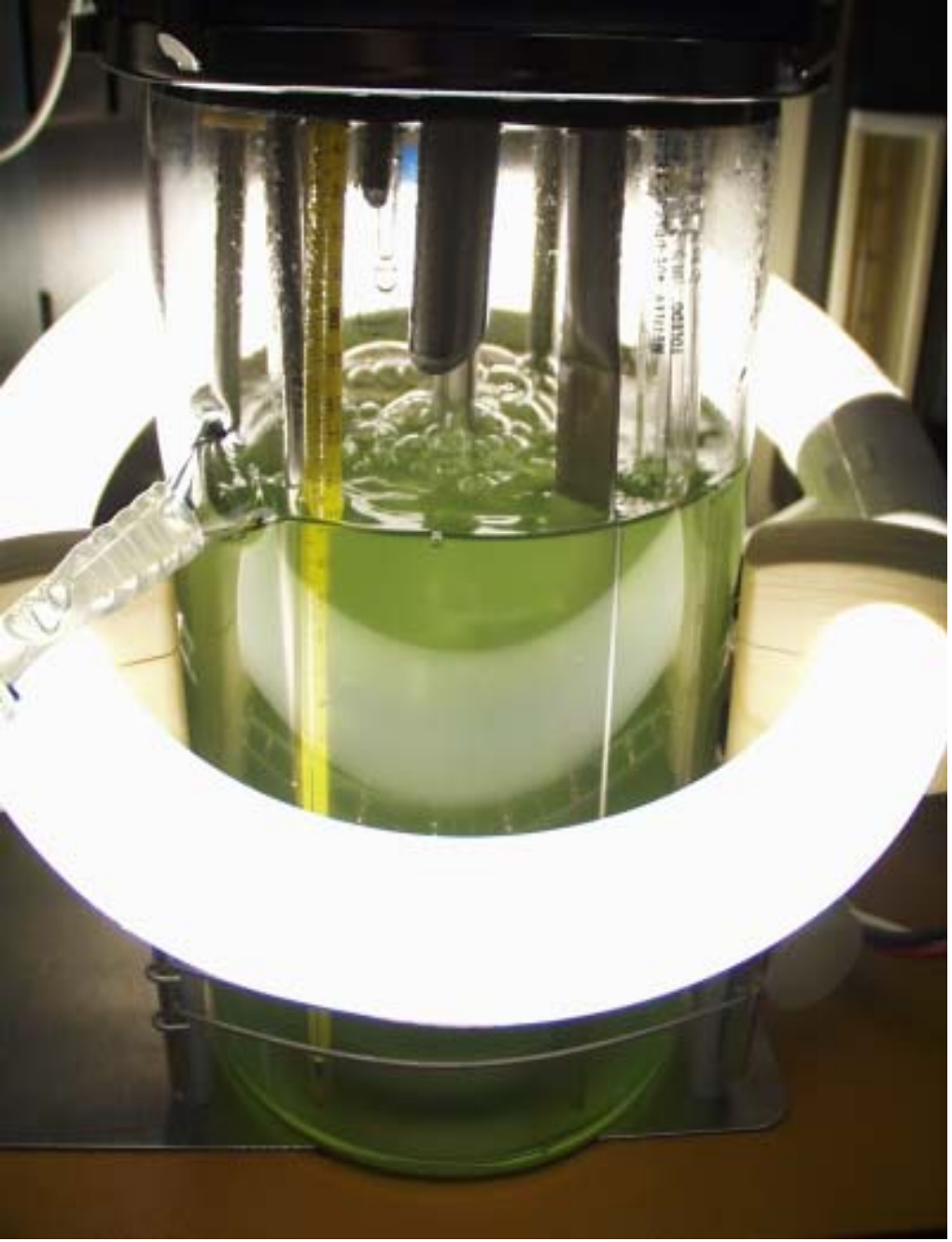
A = Algorithms
C = Compute
P = Parallelism
I = Integration

Paths to Whole Cell Simulations

- Unregulated metabolic model (flux analysis)
- Allosteric regulation (binding changes conformation) (A)
- Gene Regulated + Metabolic Model (A, C)
- Heterogeneous/Compartmentalized/Diffusion (A,C,P)
- Active Regulation + Transport (A,C,P,I)
- Complete Integrated Cell (geometry) (A,C, P,I)



- Multicellular models (homogeneous) (P)
- Multicellular (homo) with complex communication (P)
- Multicellular (hetero) mixed population (P, I)
- Multicellular differentiation and motility (A, C, P, I)
- Multicellular structures with complex geometry $(A,C,P,I)^2$



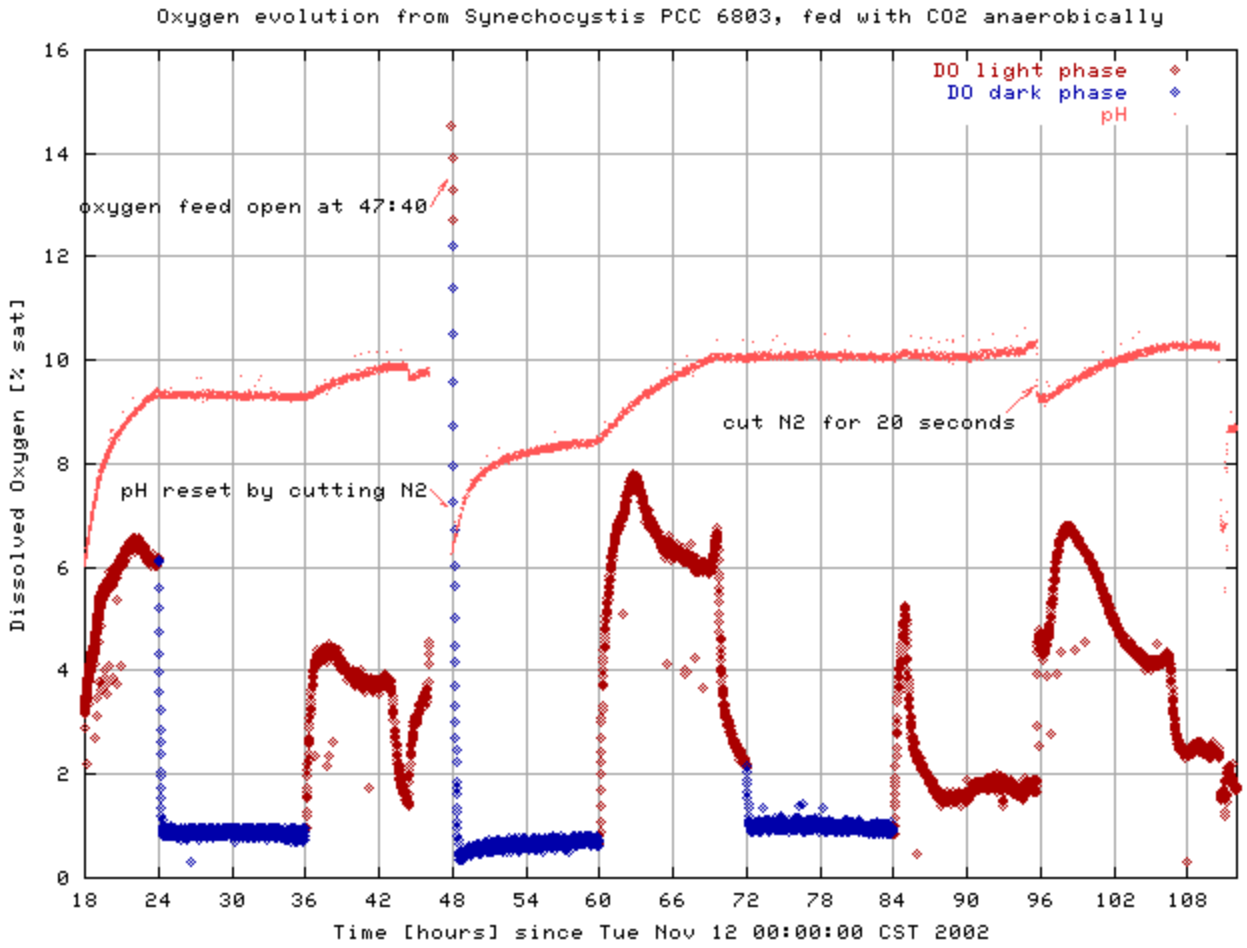
Brunswick fermenter during a light phase of cyanobacterium Synechocystis sp. PCC 6803 cultivation.

Here is the front view of the W. B. Moore's fermenter. It has two round glass windows for illumination of phototrophic cultures. One of them is seen at the bottom of the reactor steel vessel.

The back view of this reactor, tightly packed with controlling devices and microprocessors, shown in the next picture.

This fermenter is connected to the network. It will be used for long synchronous cultivation of cyanobacteria with automatic culture sampling for metabolic, proteomics and gene expression analysis.

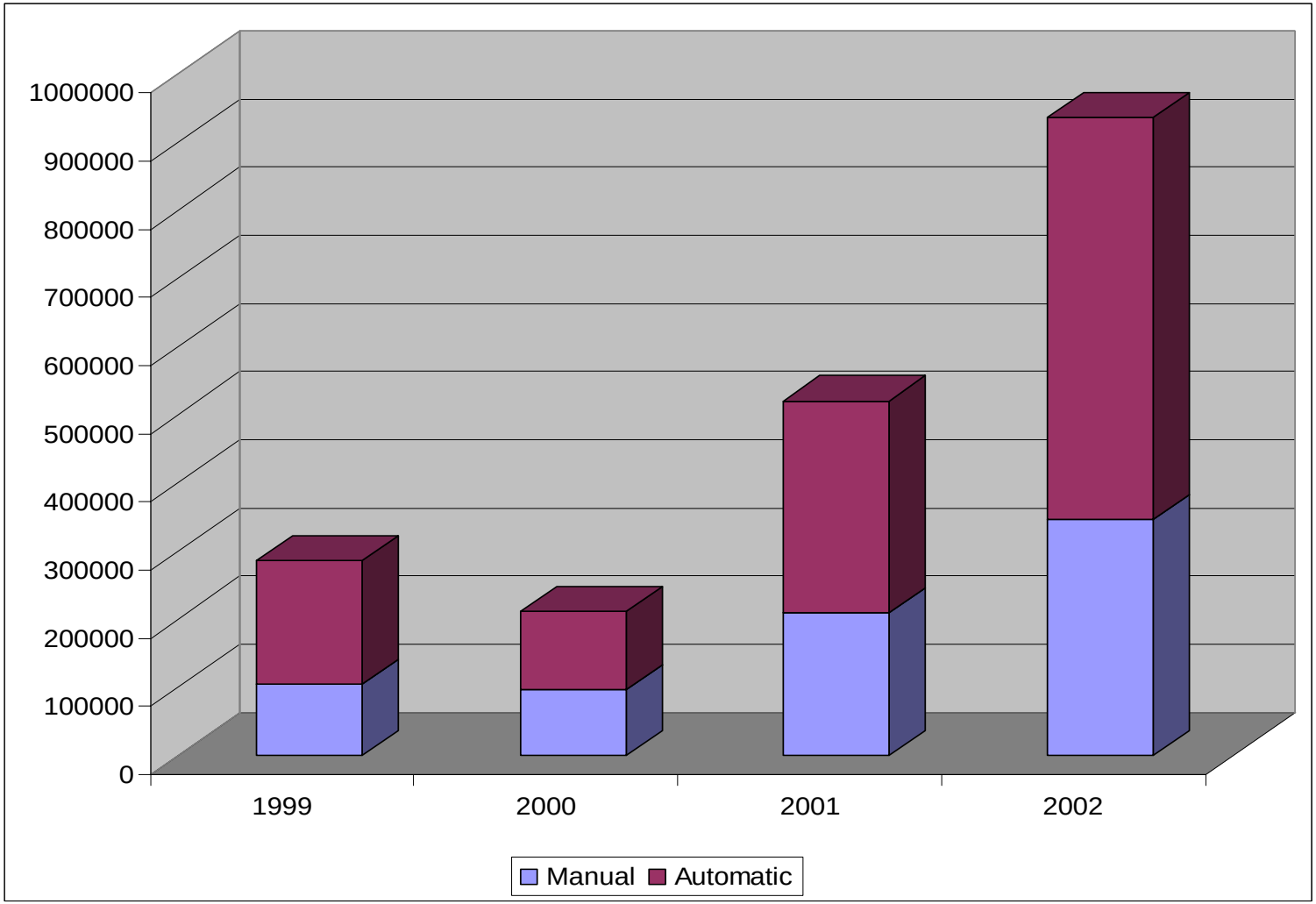
Oxygen Evolution from Synechocystis PCC 6803



Preliminary Science Results

- The *Synechocystis* PCC 6803 is easy to synchronize by dark periods. Note the photosynthesis phases at 36-43 hrs, 60-69 hrs, and 96-105 hrs. The photosynthesis is endogenously switched off by the cell clock before the light is off. Thus there is no need to use multiple entraining cycles in the illumination.
- The synchronous transitions between photosynthetic and fermentative (no or greatly reduced O2 evolution) phases are very fast. Such kinetics can be generated only by a positive feedback mechanism of the cell clock. A very short entraining period confirms this conclusion. Another cyanobacterium, oxygenic diazotrophic bacterium, *Cyanothece* sp. ATCC 51142 is being currently under cultivation experiment and gives the similar results (this experiment is not finished yet. We expect DO to be spontaneously down before or around the coming midnight. The light will stay 'on'.)
- The metabolic cell clock controls the photosynthetic activity of photosystem II (H2O splitting complex) so stiffly that it can override the illumination control. Such overriding is clearly seen in the previous record in the interval of 84-90 hrs where the clock switched off the O2 evolution in spite of the bright illumination is on after a 12hr dark period.
- The cell clock does not seem to depend on cell divisions. Although, this conclusion must be checked with cell counting yet, in the previous experiment there were practically no cell divisions after 72hr and yet the cyclic endogenous changing in DO continued.
- The fast modulation of photosystem II tells in favor of a purely metabolic regulation of its activity thus excluding involvement a much slower gene expression mechanism. We will check this statement with inhibitors of protein synthesis added to the cultures before the expected 'on' and 'off' transitions.
- The batch cultivation used so far and our simulation studies suggest that the cell synchrony is maintained via some short-living messenger molecules (most probably quorum-sensing type) secreted by the cells into the medium.

ERGO Annotations: Automatic vs Manual



A framework to support Cellular Modeling

EXAMPLE: purine biosynthesis in *B.subtilis*

Encoding, connection and curation:

- **pathways**
- **functional roles (enzymes)**
- **reactions**
- **compounds**