Systems Biology and the Grid

Rick Stevens
Argonne National Laboratory
The University of Chicago
Stevens@cs.uchicago.edu

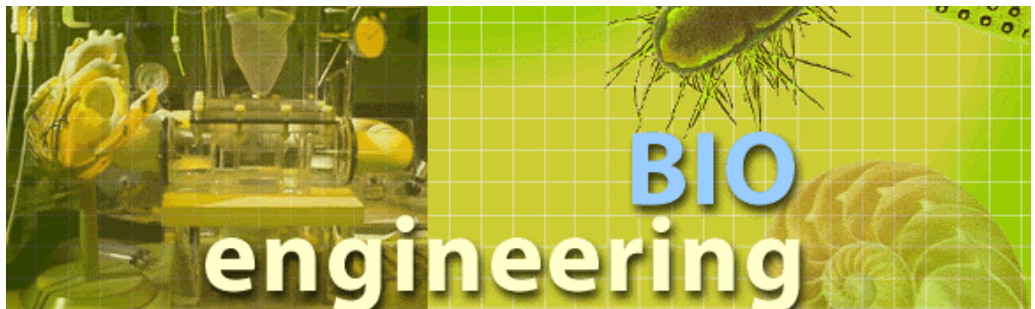
Acknowledgements

Many people have contributed to this work

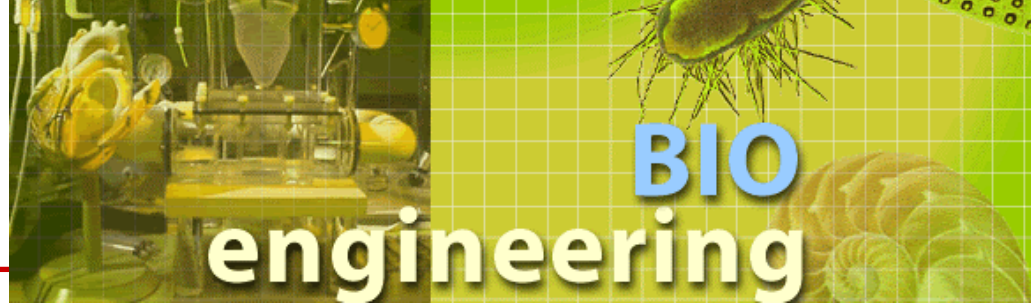
- Ross Overbeek
- Natalia Maltsev
- Ed Frank
- Alex Rodrigez
- Dina Sulakhe
- Veronkia Vonstein
- Terry Disz
- Bob Olson
- Mark Hereld
- Mike Papka
- Mike Wilde
- Ian Foster
- Rick Stevens
- Miron Livny

Why Systems/Synthetic Biology is Important

- 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

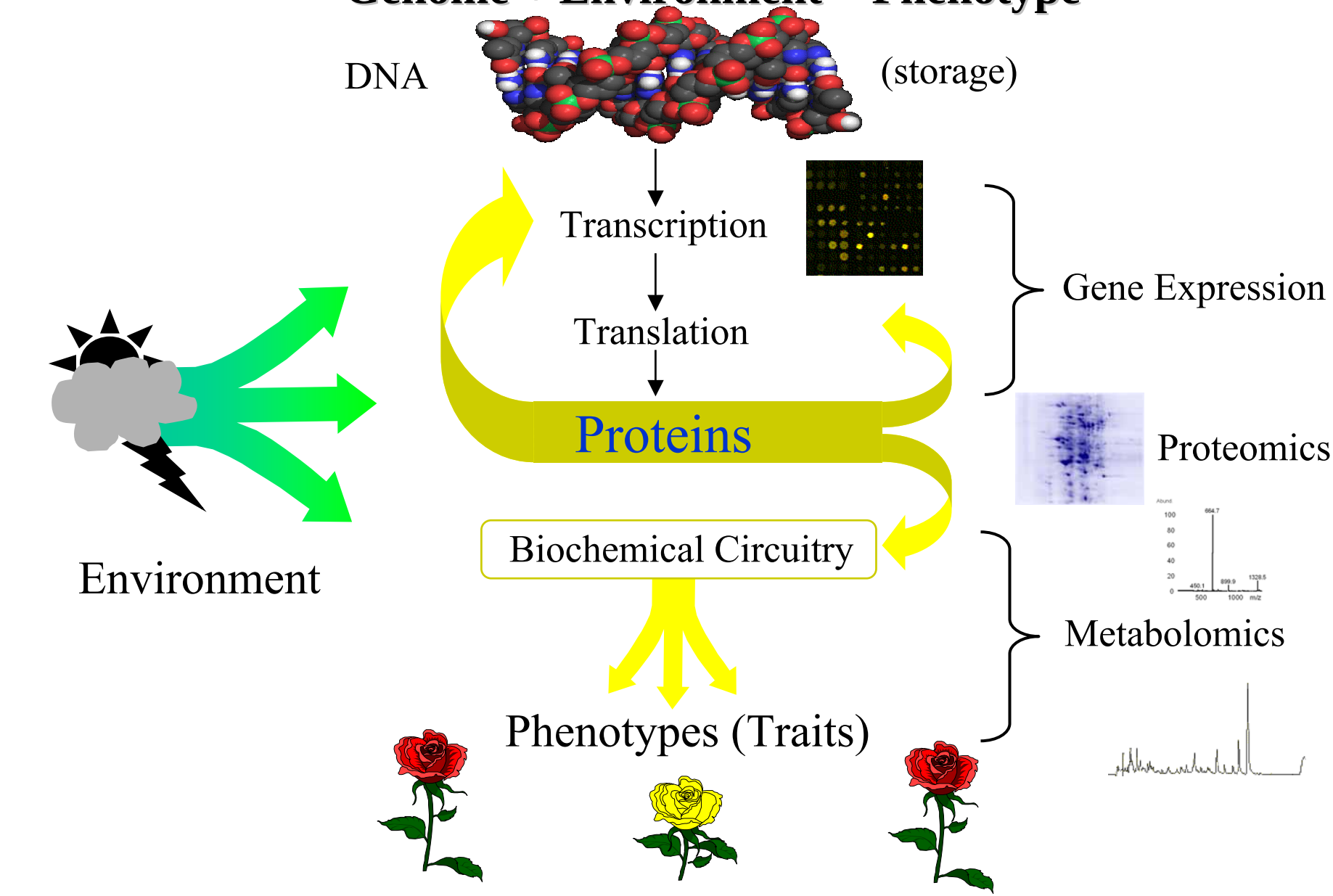


Future milestones



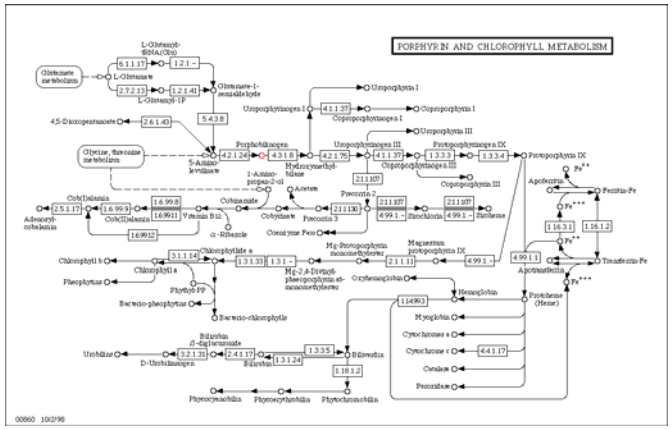
- First synthetic model prokaryotic organism
- Characterization of human microbial ecology
- Global index to life on earth
- Characterization of microbial life
- Theory of cell evolution and organization
- Theory of evolution of intelligence
- First synthetic eukaryotic organism
- Confirmation of extra-solar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

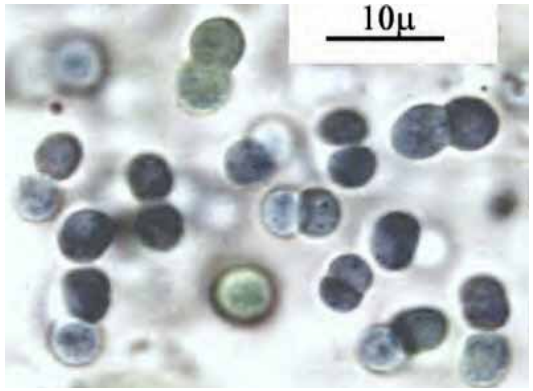
Reverse Engineering Living Systems: Genome + Environment = Phenotype

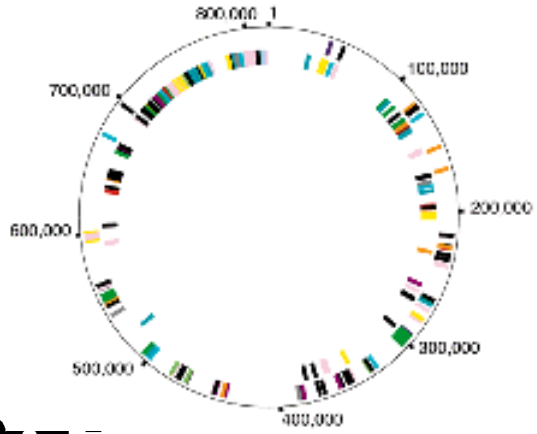


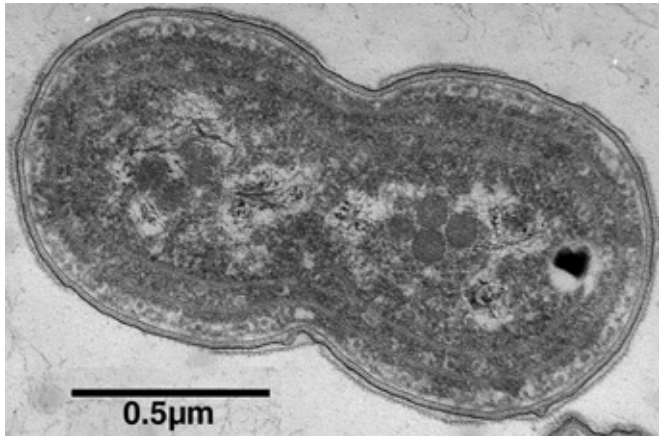
Adapted From Bruno Sobral VBI

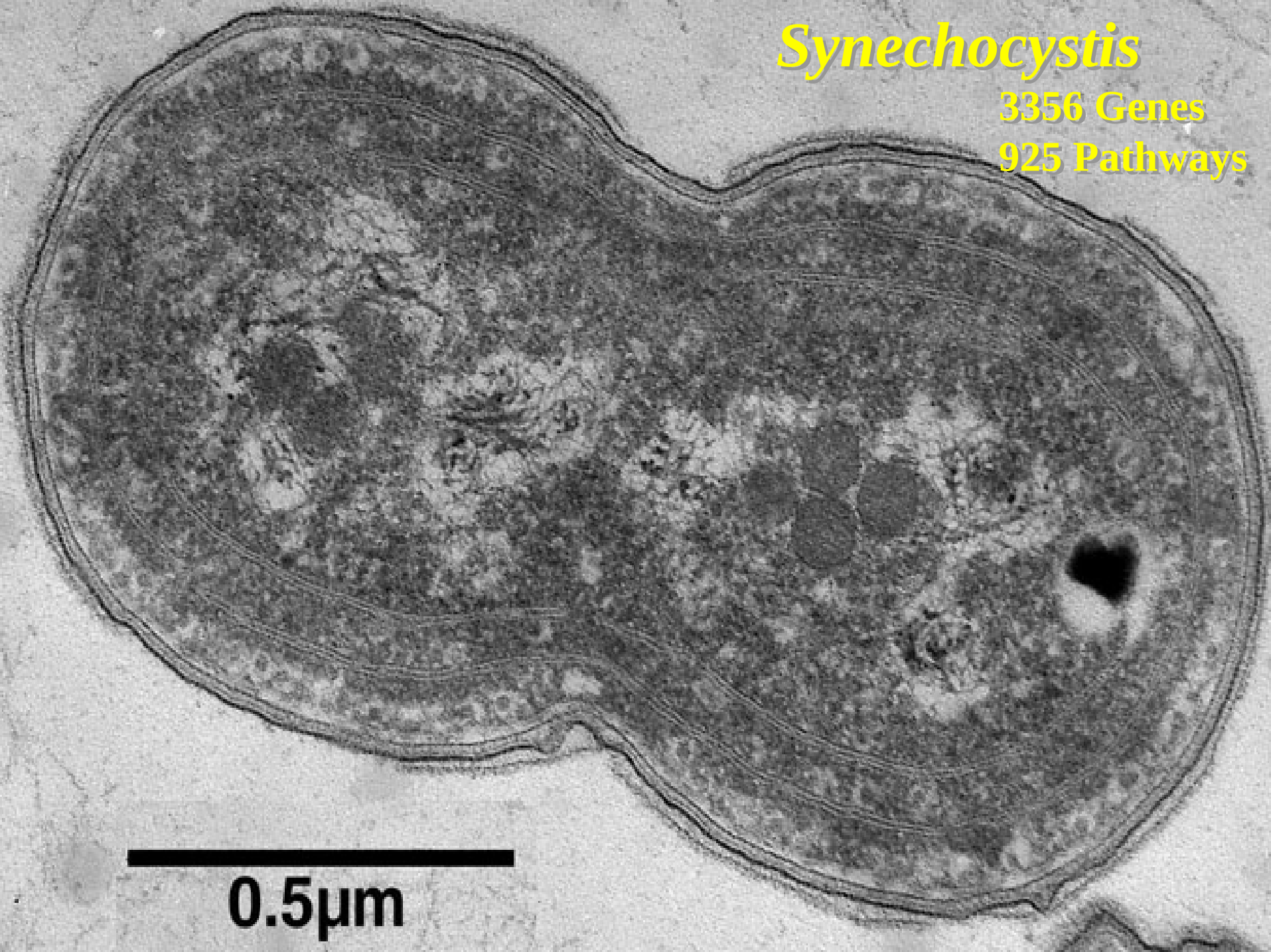
Systems Biology Systems Level Understanding
Genes → Cell Networks → Populations → Communities





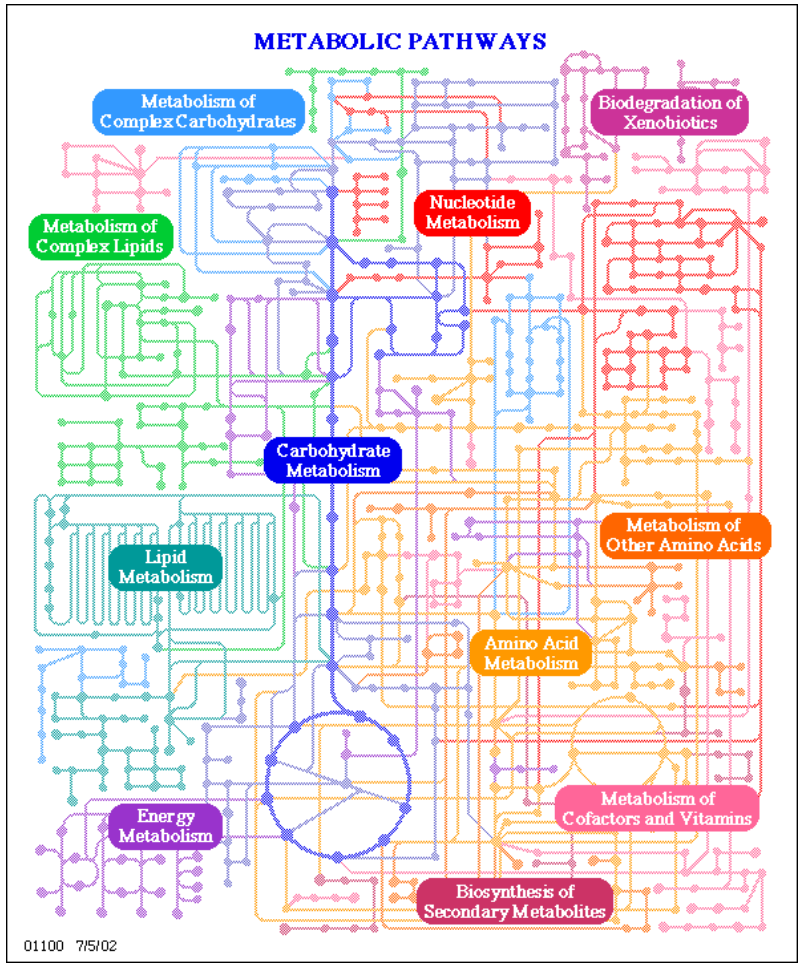




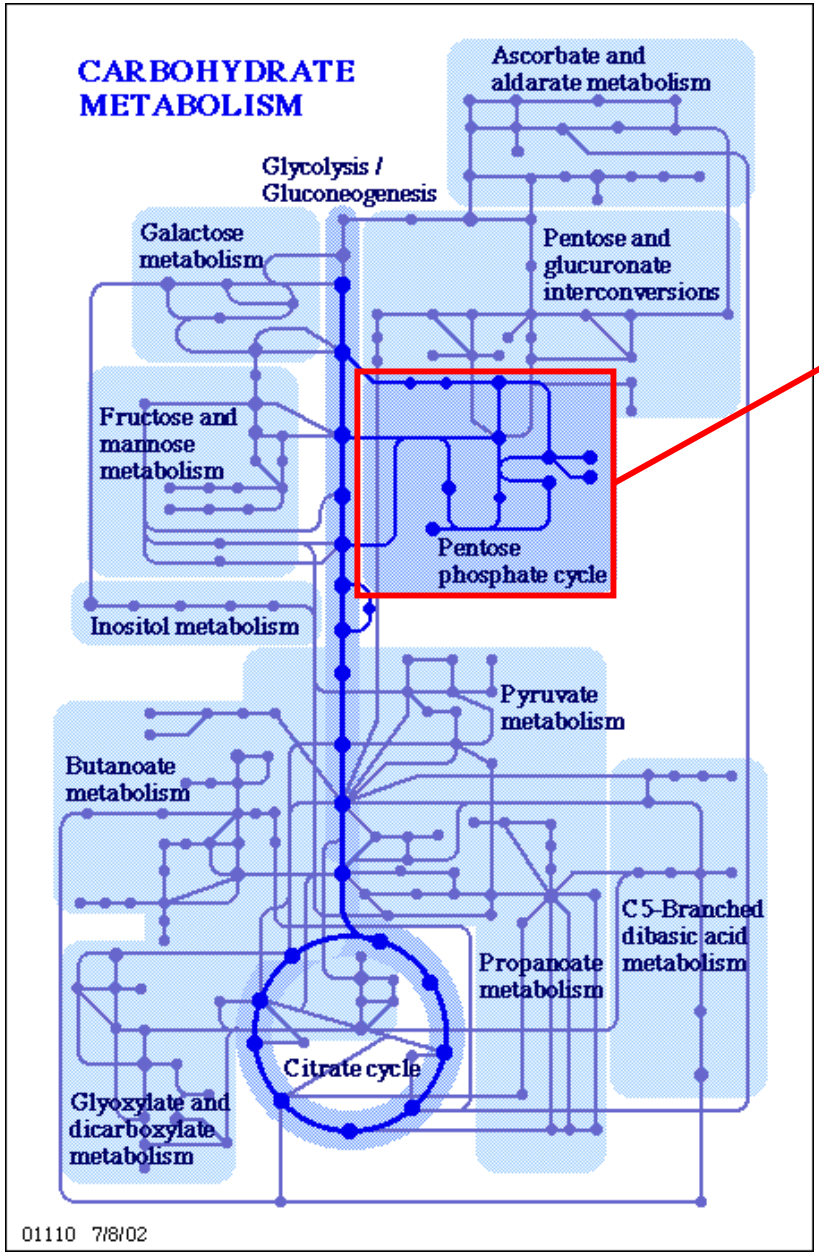

Synechocystis
3356 Genes
925 Pathways
0.5µm

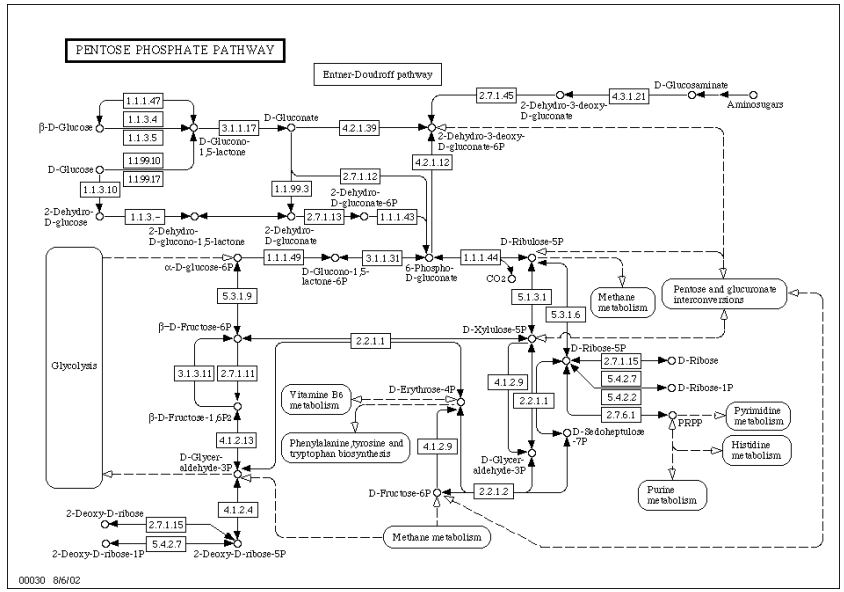
Systems Biology is Different Way of Thinking

- 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



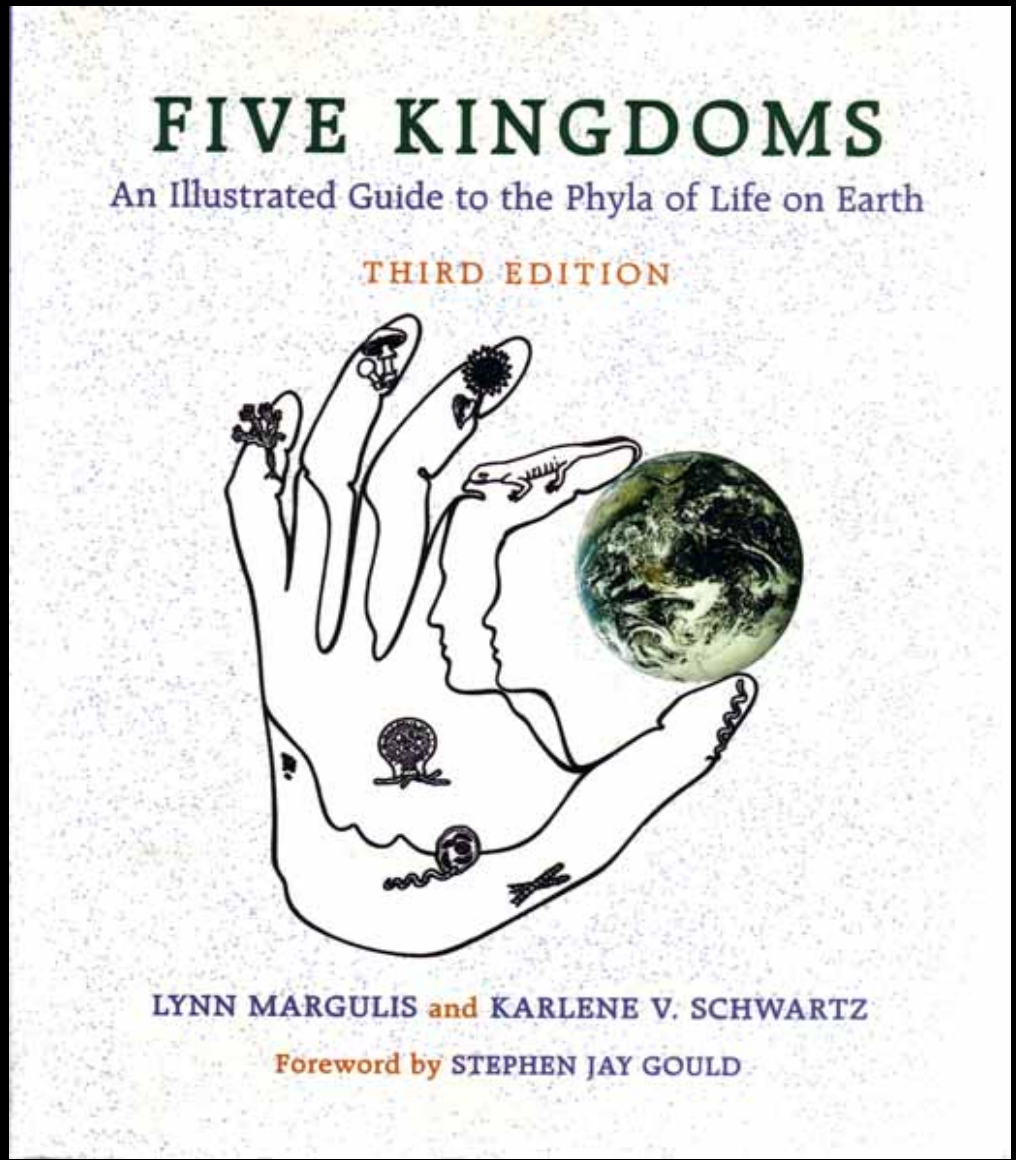
KEGG pathways


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
β-D-Glucose
D-Glucose
2-Dehydro-D-glucose
D-Gluconate
2-Dehydro-3-deoxy-D-gluconate-6P
D-Glucosaminate
Aminosugars
2-Dehydro-3-deoxy-D-gluconate
2-Dehydro-D-gluconate-6P
2-Dehydro-D-gluconate
2-Dehydro-D-glucono-1,5-lactone
α-D-glucose-6P
D-Glucono-1,5-lactone-6P
6-Phospho-D-gluconate
D-Ribulose-5P
CO2
Methane metabolism
Pentose and glucuronate interconversions
Glycolysis
β-D-Fructose-6P
D-Xylulose-5P
D-Ribose-5P
D-Ribose
D-Ribose-1P
FRPP
Pyrimidine metabolism
Histidine metabolism
Purine metabolism
Vitamine B6 metabolism
D-Erythrose-4P
Phenylalanine,tyrosine and tryptophan biosynthesis
D-Sedoheptulose-7P
β-D-Fructose-1,6P2
D-Glycer-aldehyde-3P
D-Fructose-6P
2-Deoxy-D-ribose
2-Deoxy-D-ribose-1P
2-Deoxy-D-ribose-5P
Methane metabolism
00030 8/6/02

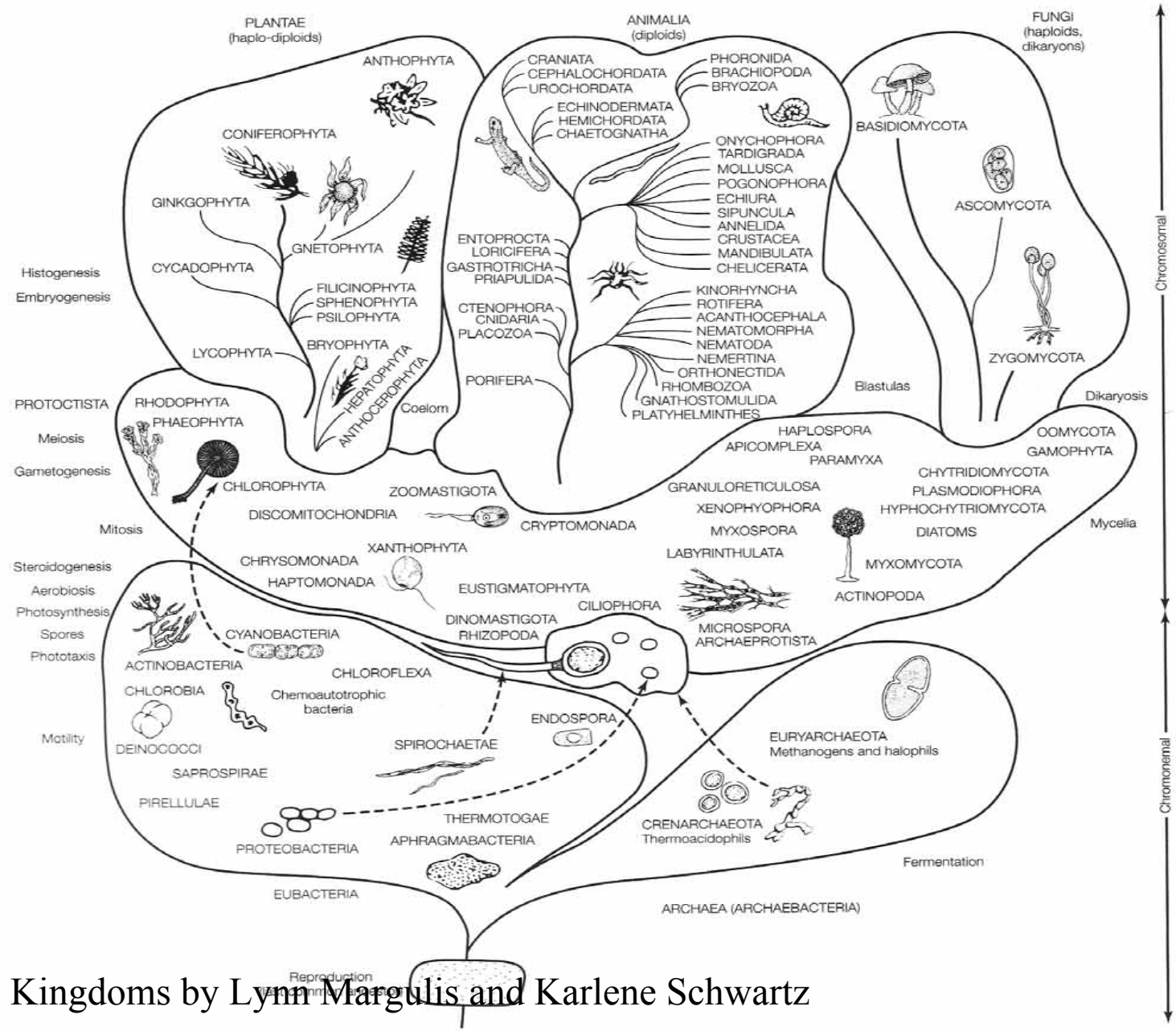
Biology is BIG!!

- 3500+ Millions years of evolution
  - ~10M extant species (distinct genomes)
    - 400 genomes sequenced so far many to go!!
    - >100M extinct species
  - 10^8 average genome size (coding region)
  - ~10^8 bp x ~10^7 sp = 10^{15} bp of genetic diversity
  - 10^{11}-10^{13} total genes and gene products
    - ~3,000 protein structures determined
    - typical bacteria has about 3,000 types of proteins



- Overview of Life on Earth
- Basic Microbiology Introduction
- Excellent Summary

The phyla of life on Earth based on our modification of the Whittaker five-kingdom system and the symbiotic theory of the origin of eukaryotic cells.

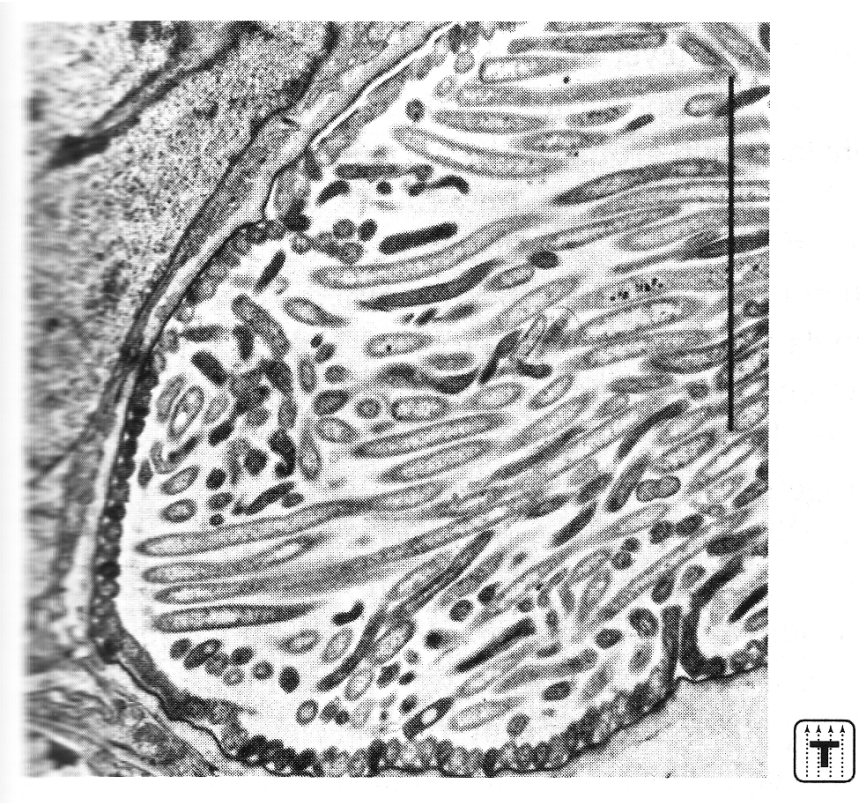


From Five Kingdoms by Lynn Margulis and Karlene Schwartz

Could Biology Dominate the Grid?

- The largest science discipline:
  - The most scientists (Globally ~500,000-1,000,000)
  - The most research funding (Globally >$100 Billion/year)
  - The most graduate students (>50,000 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

A Diverse Bacterial Community

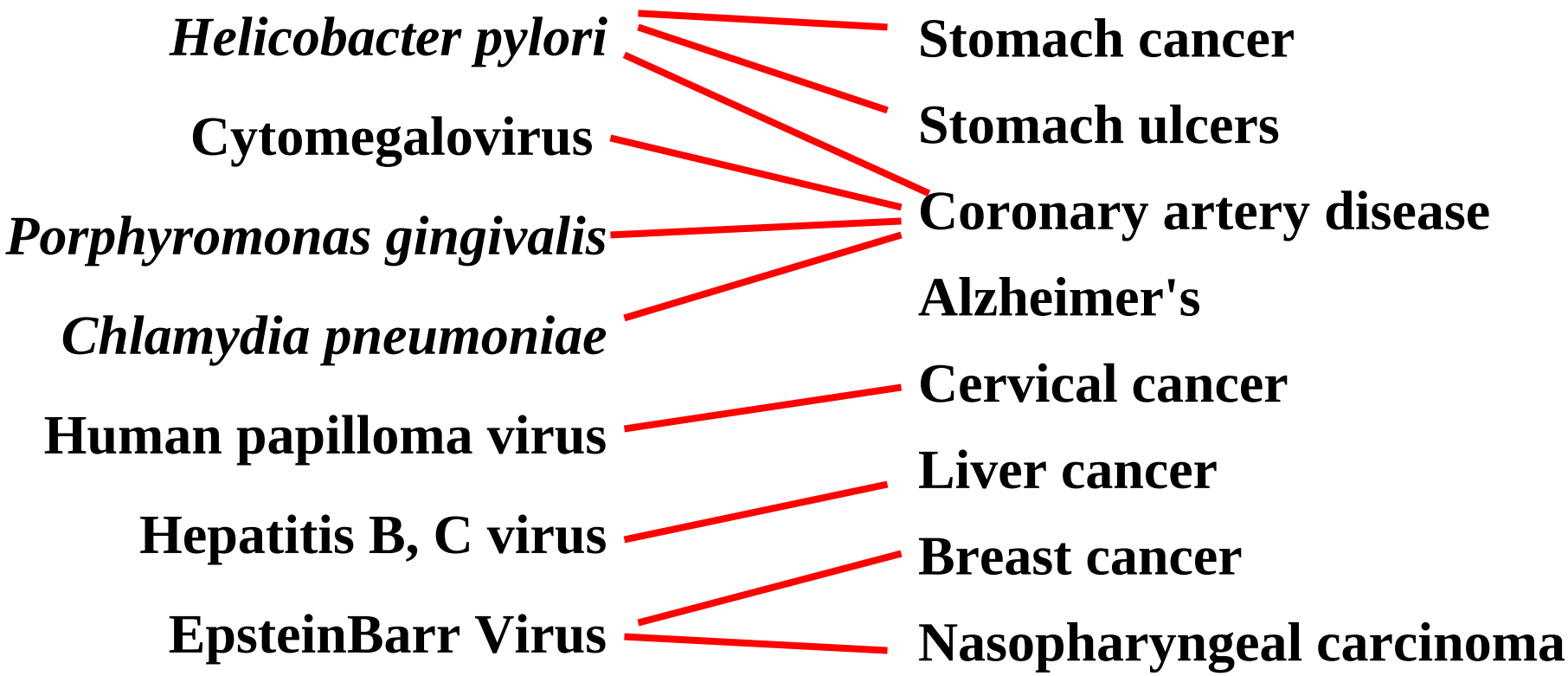


From Five Kingdoms

Lynn Margulis and Karlene Schwartz

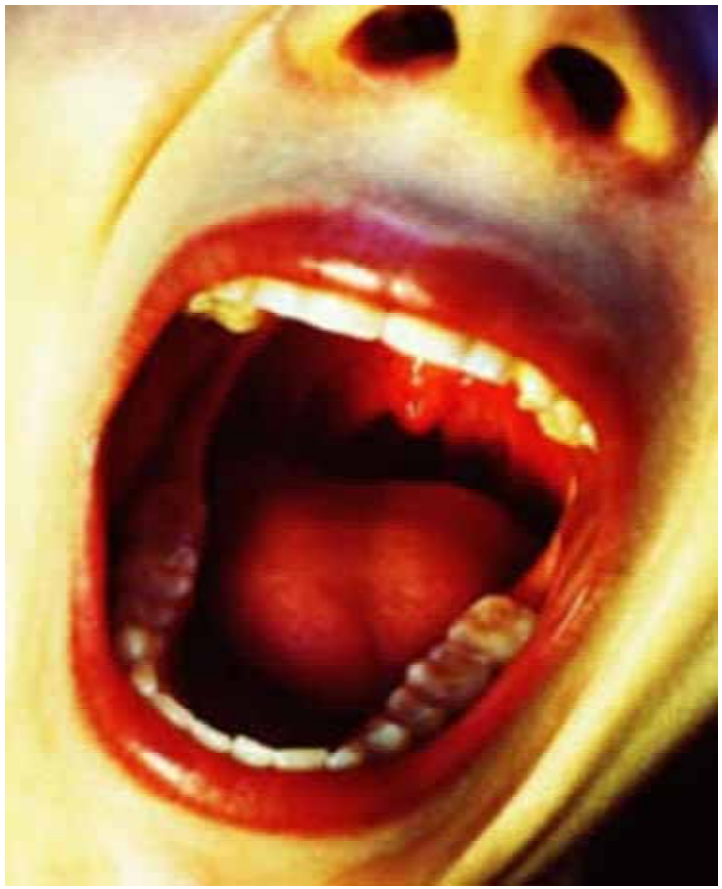
- Pocket in the hindgut wall of the Sonoran desert termite *Pterotermes occidentis*
- 10 billion bacteria per milliliter
- Anoxic environment
- ~30 strains are facultative aerobes
- Many/most are unknown

Human Microbial Ecology: Sorting Out Cause, Effect + Cure



= suspected linkages

Body Cavity Sequencing Project



- TIGR + Stanford
- Sequencing samples from mouth, gut and vagina
- 40% of mouth sequences in recent pilot have never been seen before
- Goal is development of diagnostic chips
- Ultimately, the new analyses should reveal which combination of bugs contributes to a healthy body and which to diseases. The team is already comparing the microbes on healthy teeth with those in a person suffering severe gum disease.

Source: Nature 23 June 2003

Sequencing the Sargasso Sea

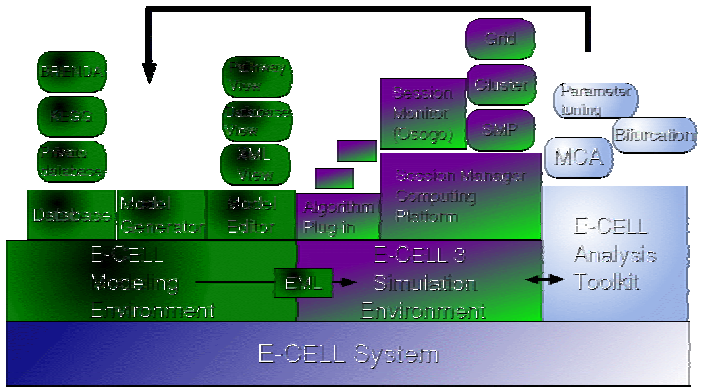
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- There's as much unique genetic information in a litre of seawater as in the human genome.
- 1.2 Million Genes
- 70,000 novel genes
- 1,800 species
- 148 new types of bacteria
- Craig Venter's IBEA

Challenges for Cell 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 may be a "killer app" for Petascale computing and beyond

E-CELL Development Overview



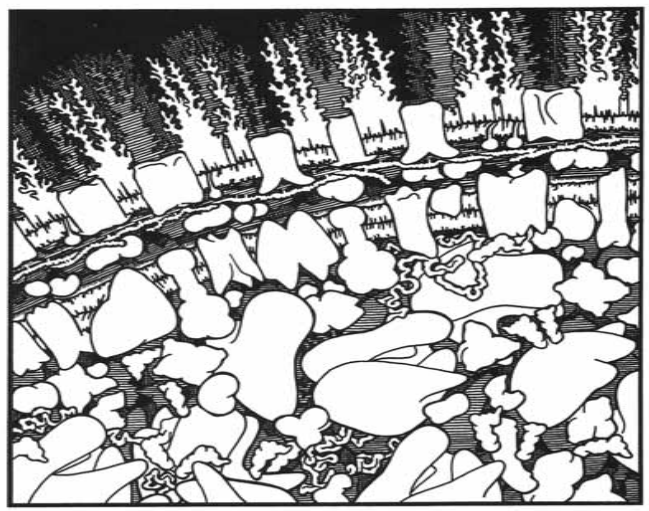
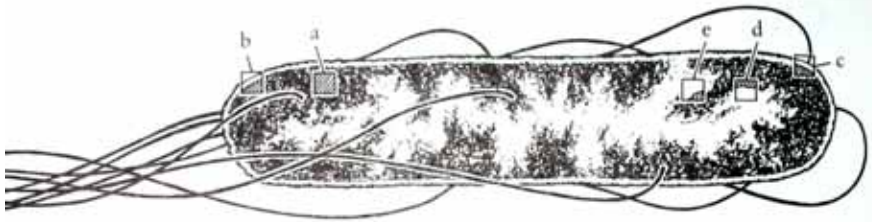
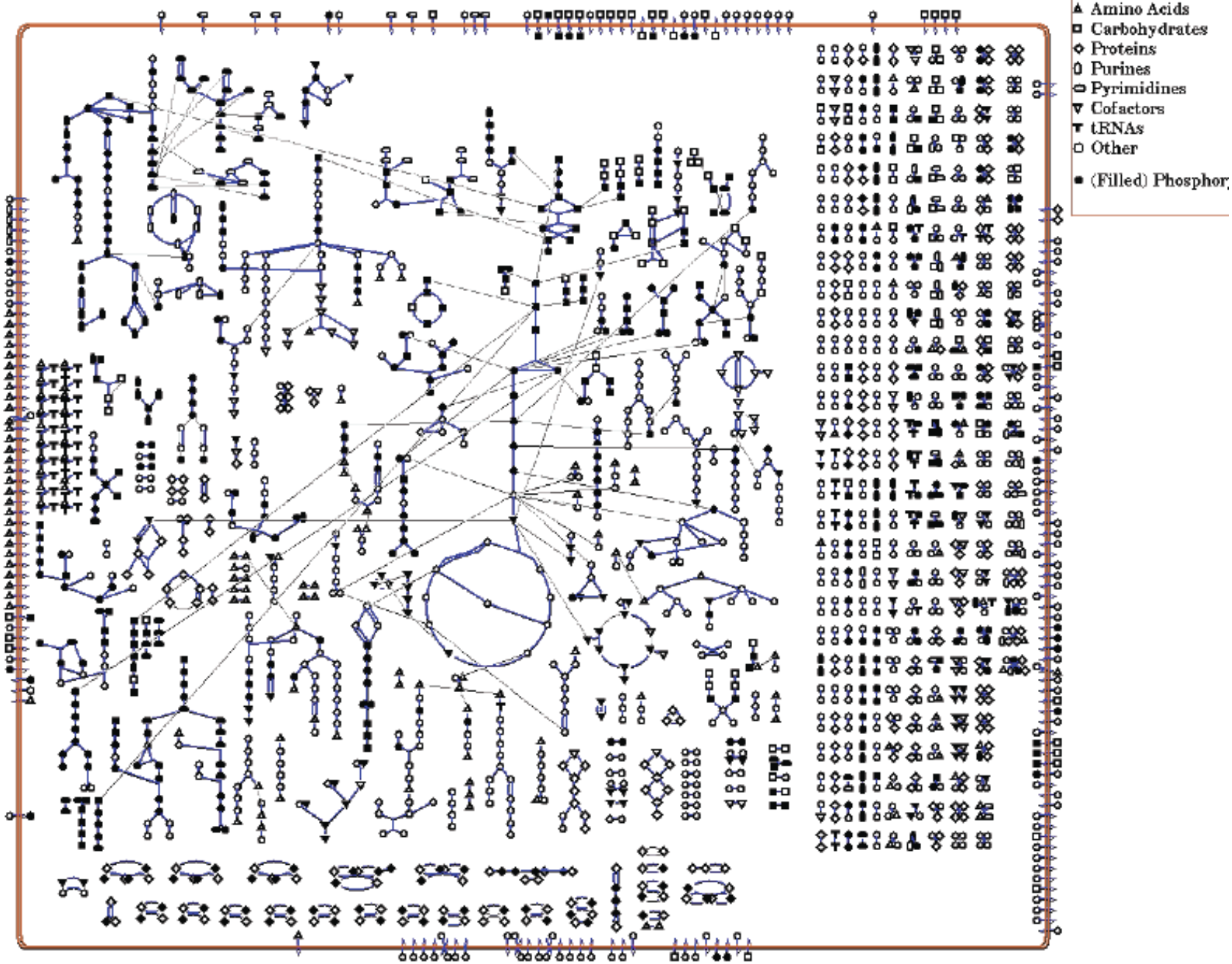


Figure 4.3 Cell Wall

Reconstructing and Modeling a Prokaryotic Cell?



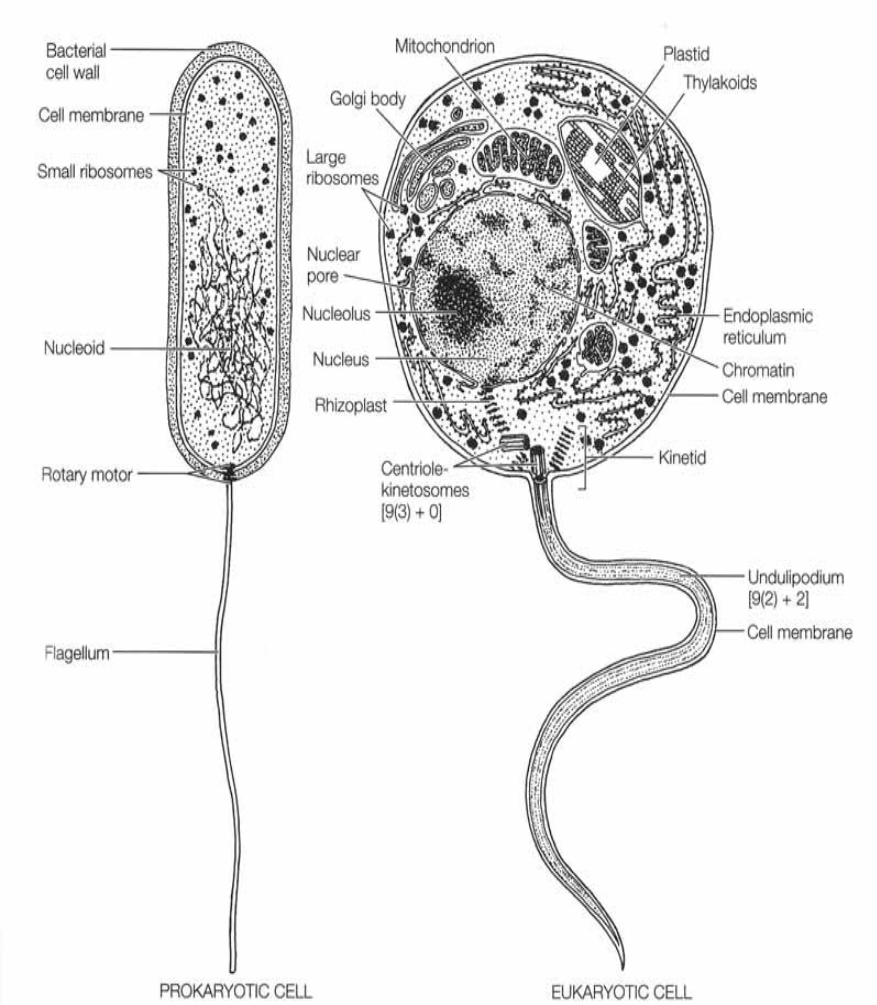
- Typical bacteria [*E. coli*]
  - 1000nm x 300nm x 300nm volume
  - ~4000 genes and gene products
    - 1/4 genes ⇒ protein translation, synthesis and transport
    - 1/4 genes ⇒ core metabolism
    - 1/4 genes ⇒ life cycle and environmental response
    - rest genes involved in regulation, synthesis and degrading tasks
    - Relatively few genes related to sensing and motility
  - 1,000's of small molecular species, not tracked individually
  - ~3 million total large molecules to track



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

Source: EcoCyc

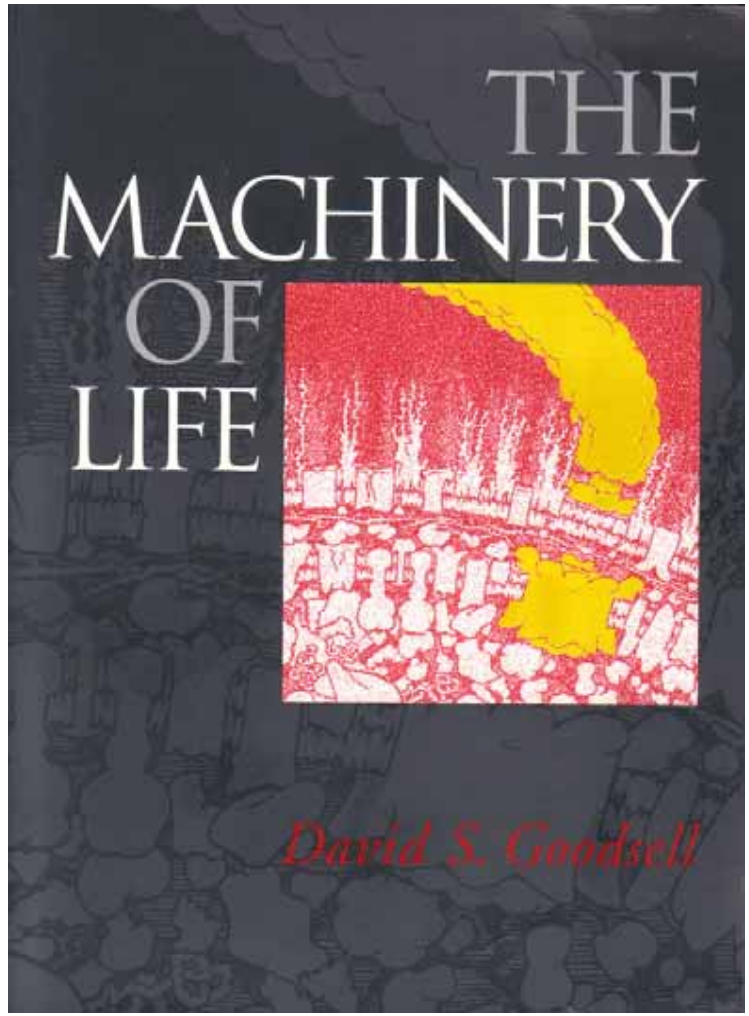
Prokaryotic v Eukaryotic Cell



- Membrane separated nucleus
- Multiple cellular compartments
- Introns
- More complex cell membrane
- Complex motility mechanisms
- Etc.

From Lynn Margulis and Karlene Schwartz

- David Goodsell
- Illustrations at 1Mx scale of many cellular domains
- Powerful use of scientific illustration
- Introduction to cellular biology and basic molecular biology



Intracellular Environment – Gel Like Media

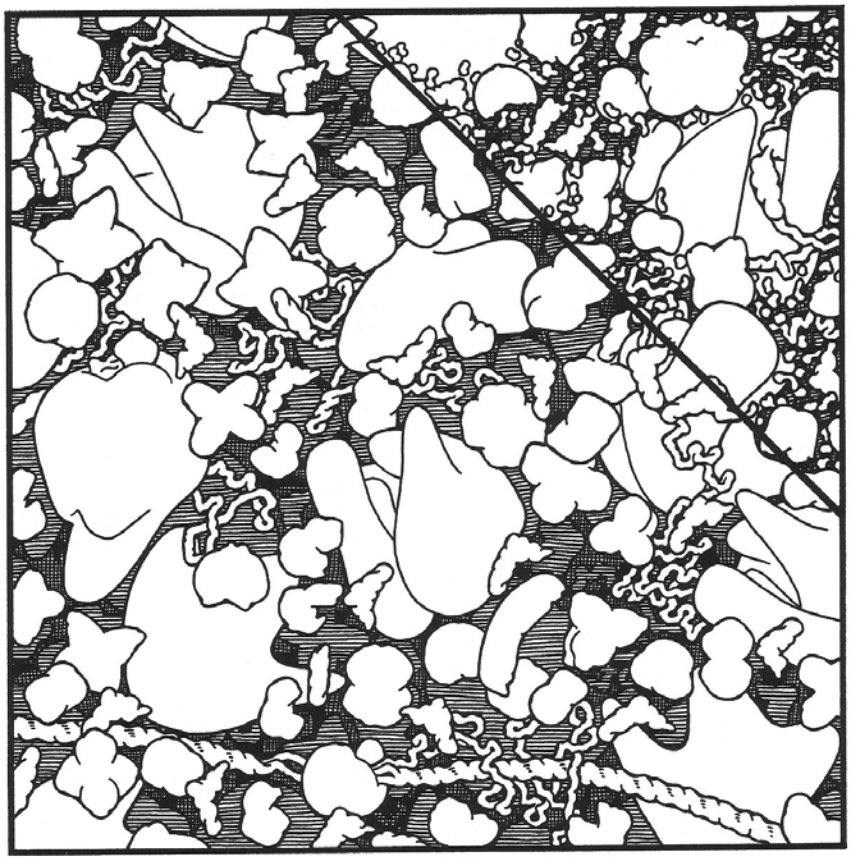


Figure 4.2 Cytoplasm

- 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

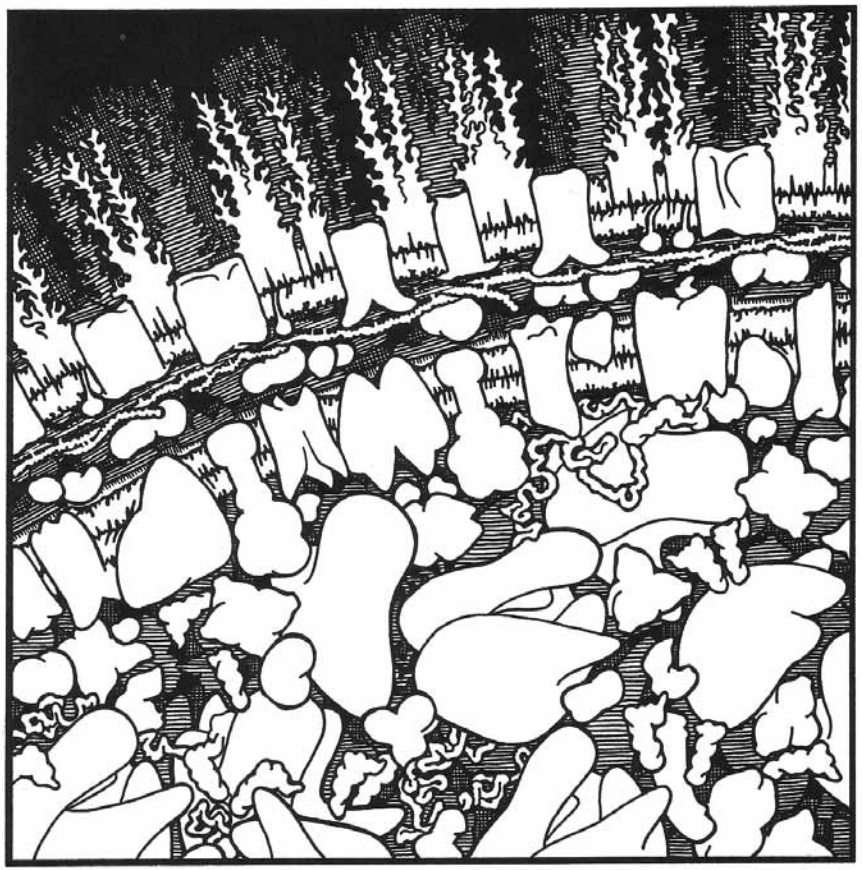


Figure 4.3 Cell Wall

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

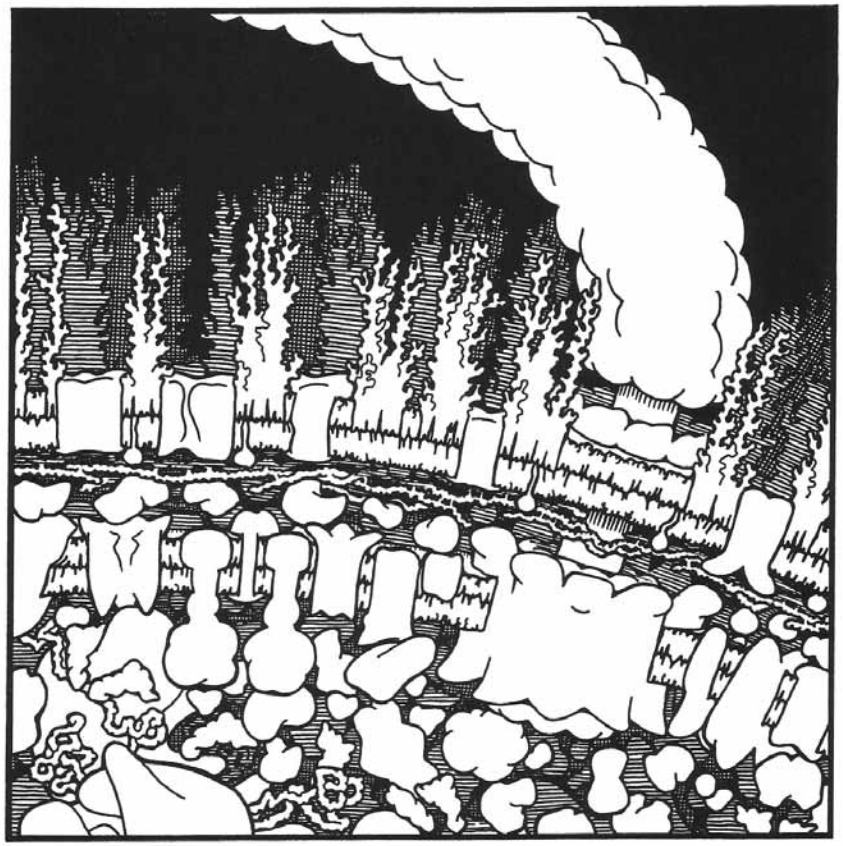


Figure 4.4 Flagellum and Flagellar Motor

- 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

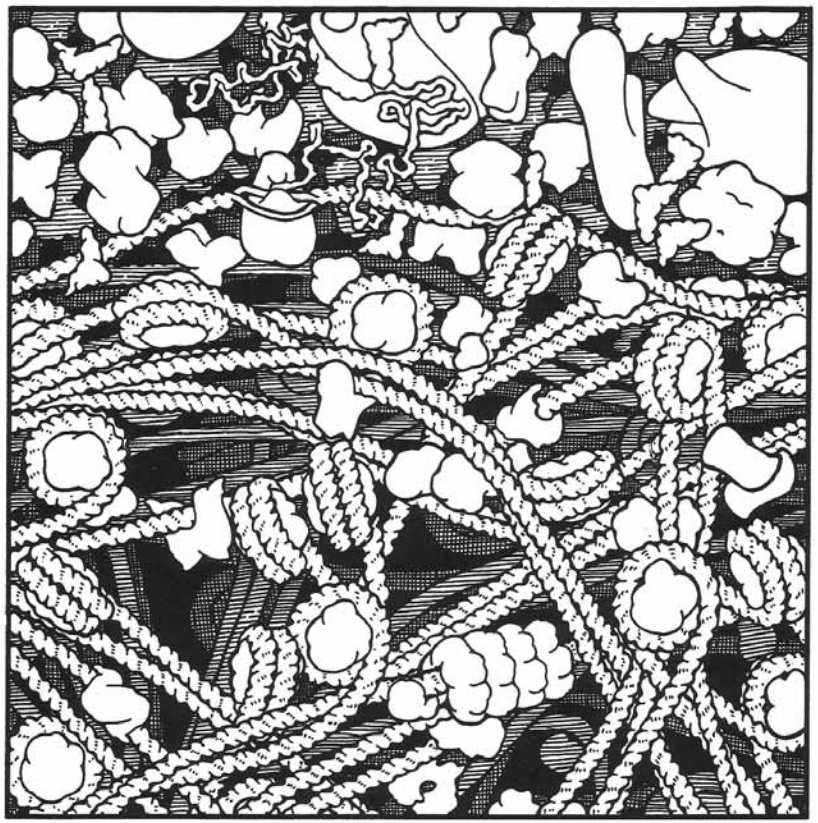
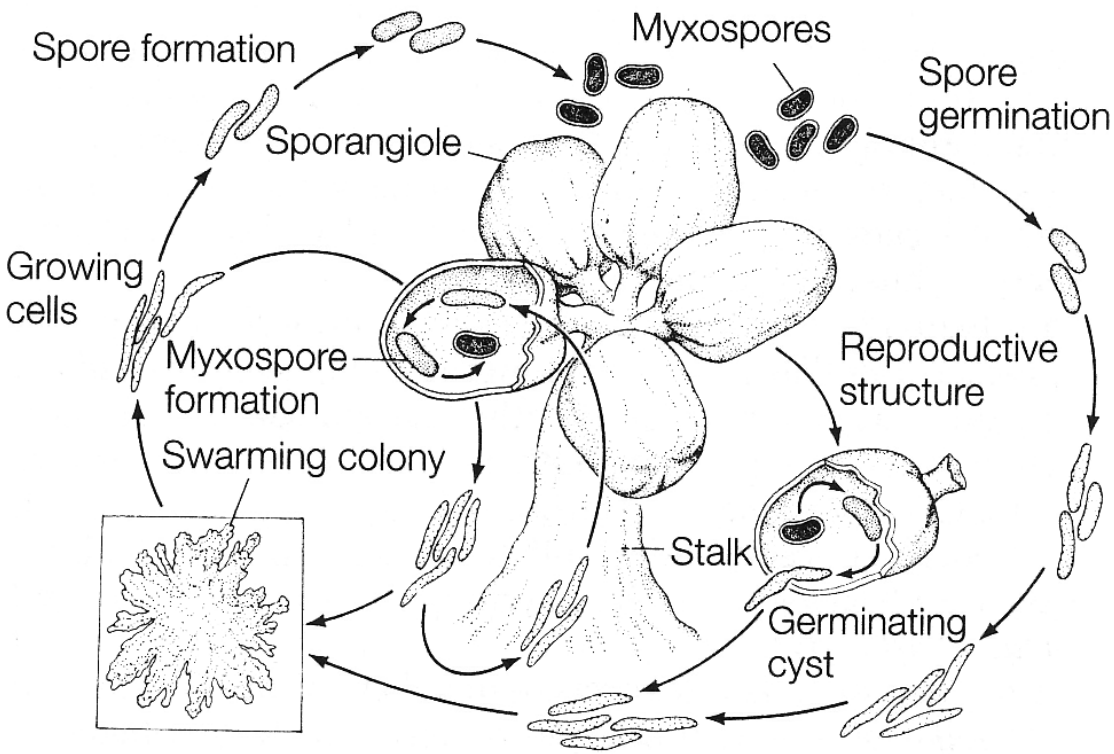


Figure 4.5 Nuclear Region

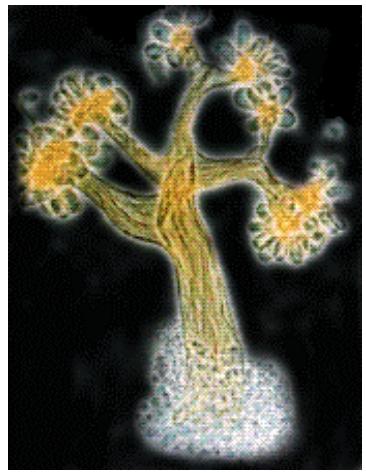
- DNA replication about 800 new nucleotides per second
- In circular DNA both directions at once
- 50 minute to duplicate entire circle of 4,700,000 nucleotides
- With cell replication ~30 minutes
- DNA replication is pipelined!!

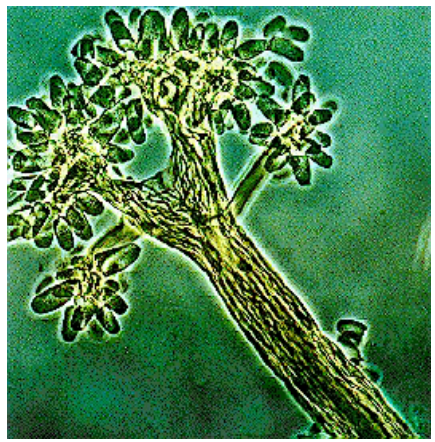
From: David Goodsell, The Machinery of Life

Understanding Bacterial Life Cycles



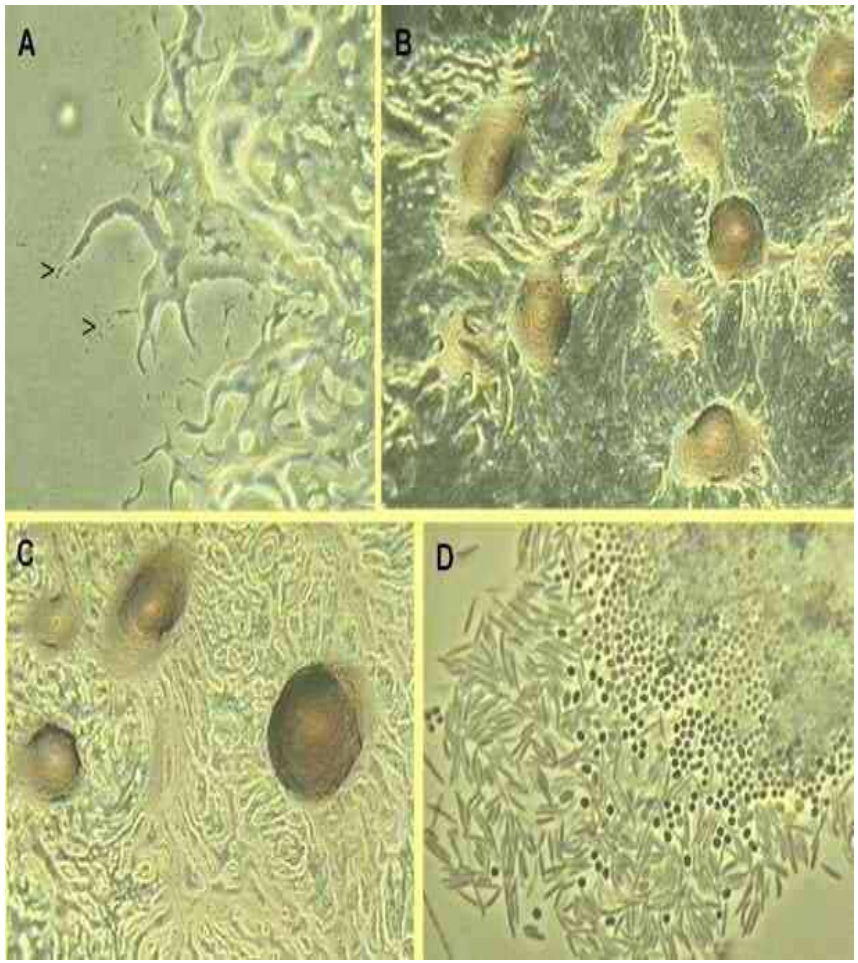
F Life cycle of *Stigmatella aurantiaca.* [Drawing by L. Meszoly; labeled by M. Dworkin.]





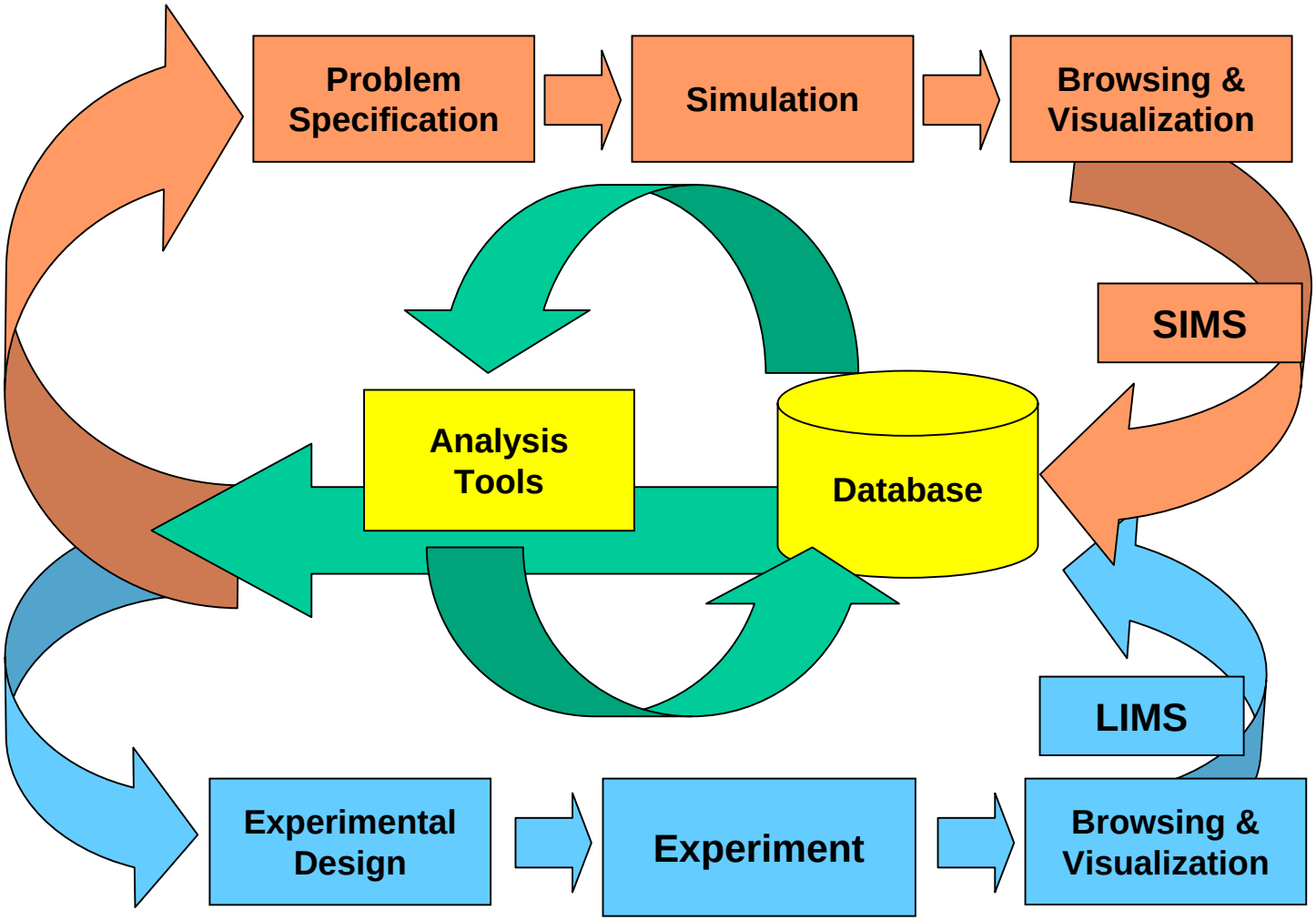
From Lynn Margulis and Karlene Schwartz

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

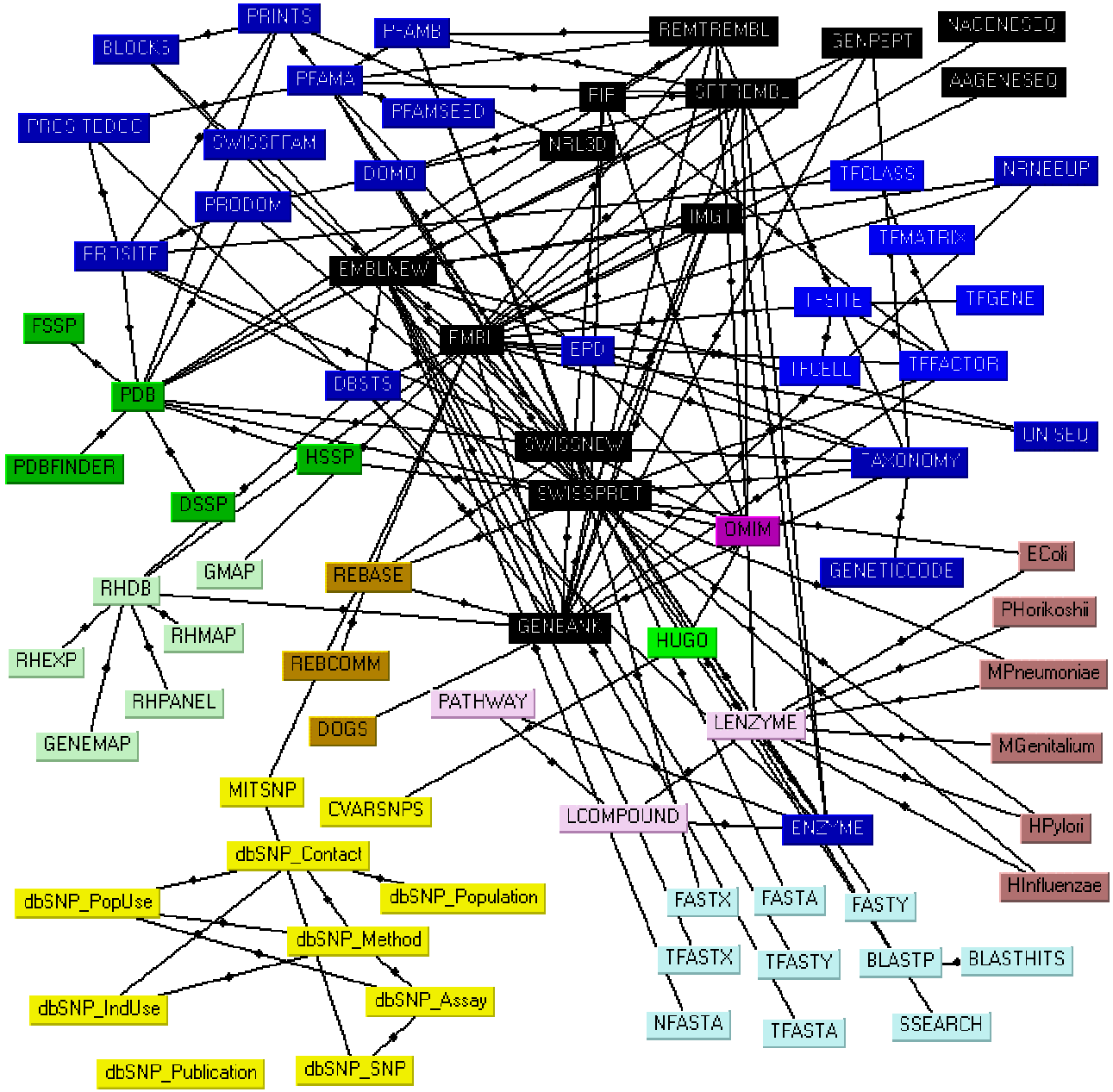
An Integrated View of Simulation, Experiment, and Bioinformatics



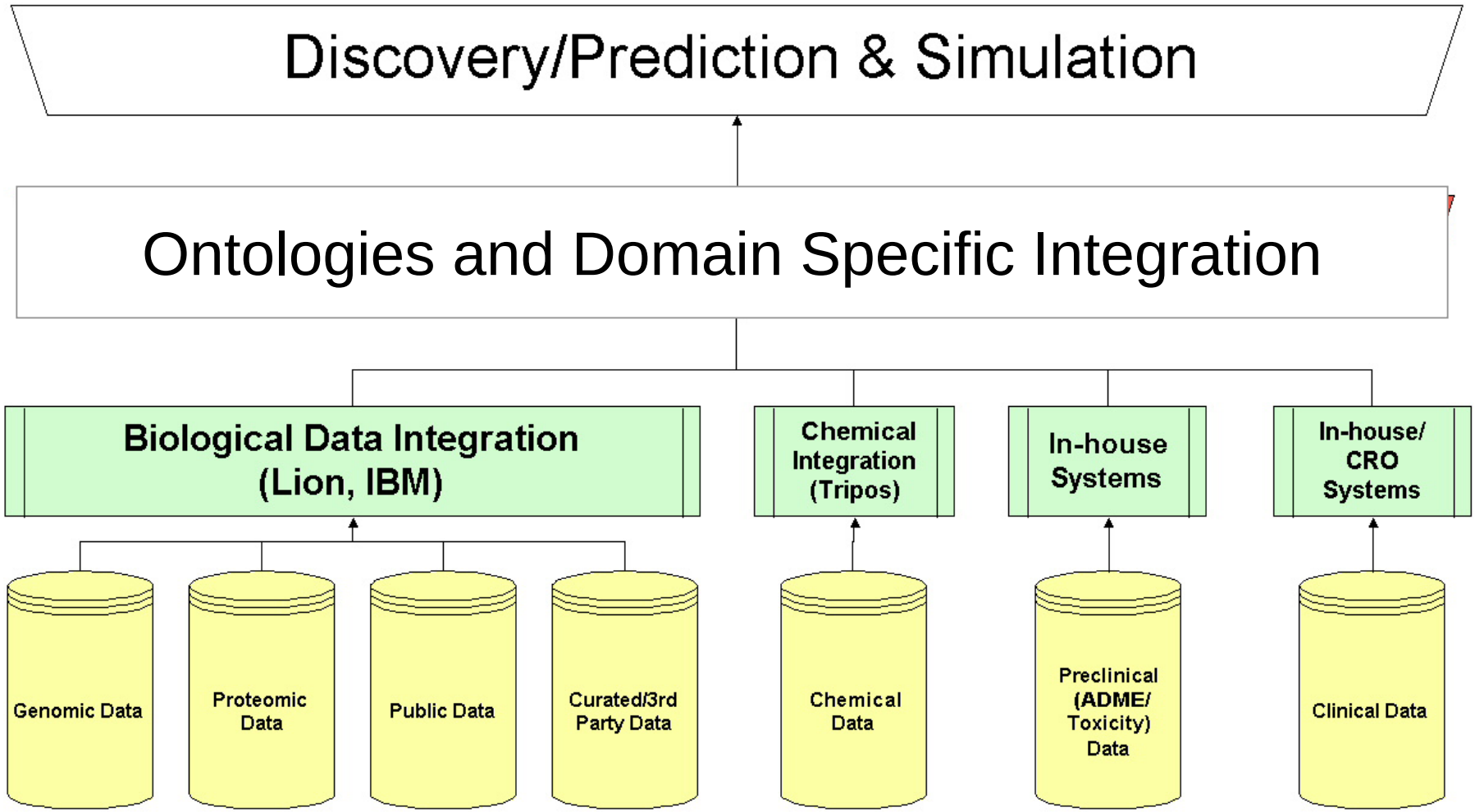
Biology Databases (335 in 2001)

- 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
PFAMB
REMTREMBL
GENPEPT
NACCNESCO
PFAMA
PIR
SPTTREMBL
AAGENESEQ
PFAMSEED
PROSITEDOC
SWISSPFAM
NRL3D
DOMO
TRCLASS
NRNEEUP
PRODOM
IMGT
TRMATRIX
PROSITE
EMBLNEW
TRSITE
TRGENE
FSSP
EMBL
EPD
TRCELL
TRFACTOR
PDB
DBSTS
UNSEQ
SWISSNEW
HSSP
TAXONOMY
PDBFINDER
SWISSPROT
DSSP
OMIM
EColi
GMAP
REBASE
GENETICCODE
RHDB
GENEANK
PHorikoshii
RHMAP
HUGO
RHEXP
REBCOMM
MPneumoniae
RHPANEL
PATHWAY
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

Software Infrastructure in Drug Discovery



From Richard Gardner(InCellico)

The Cytoscape Tool for Cell Network Visualization

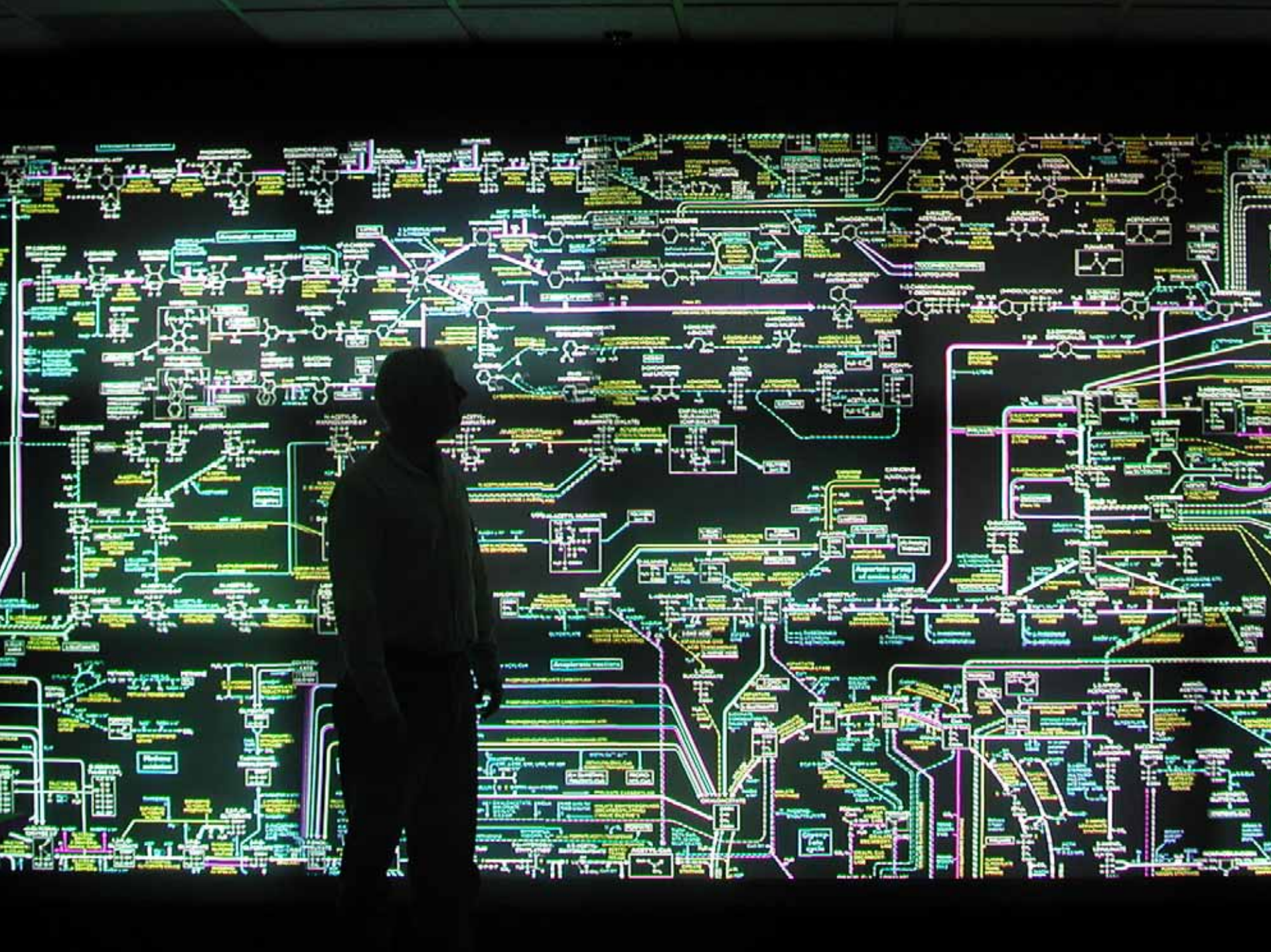
- Tools for understanding molecular interaction networks
- GML support
- Expression Profiles
- Functional annotations via GO database

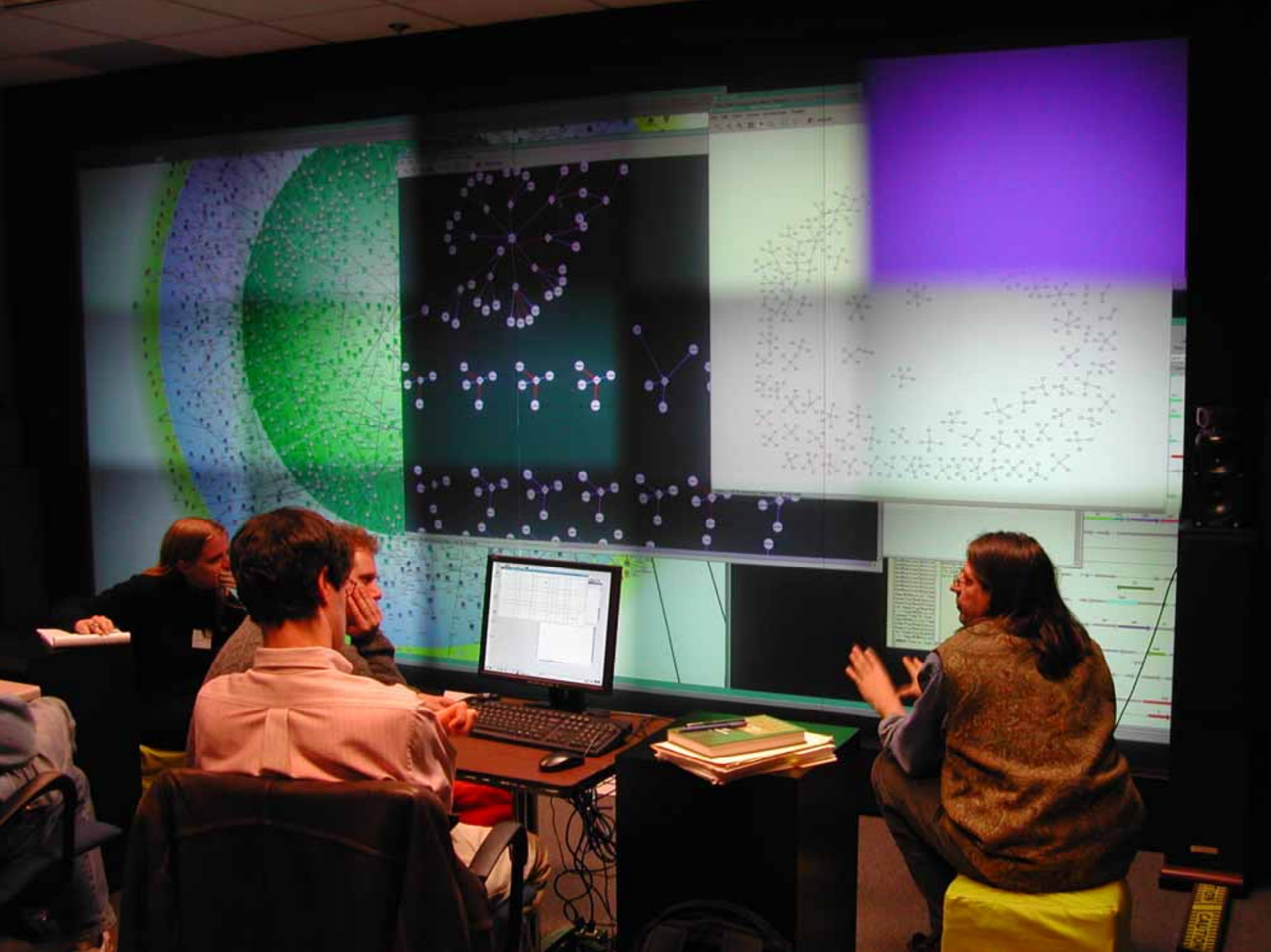
QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.

Source: Institute for Systems Biology

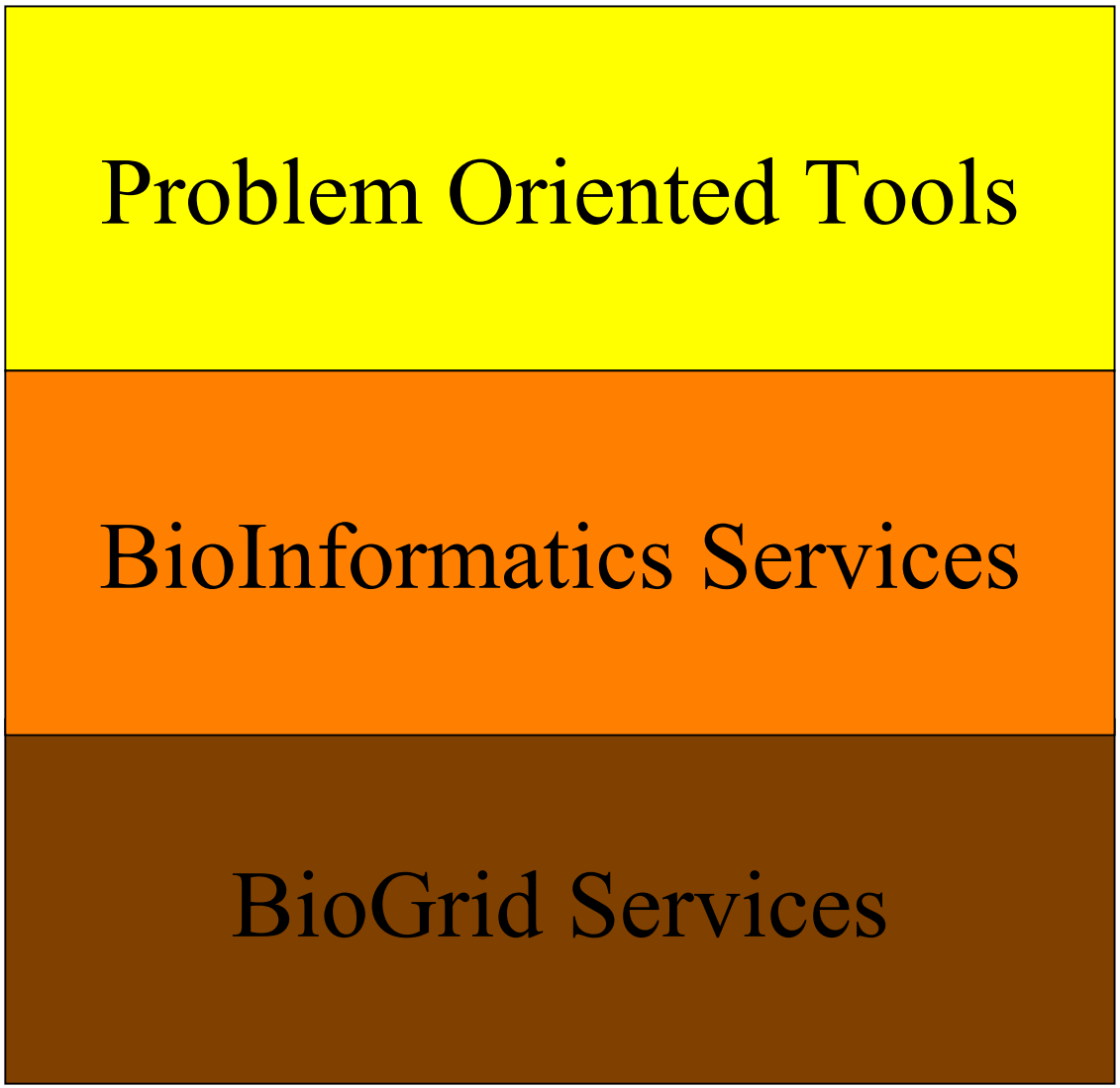
CellDesigner from www.systems-biology.org

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.





An Example BioGrid Services Model

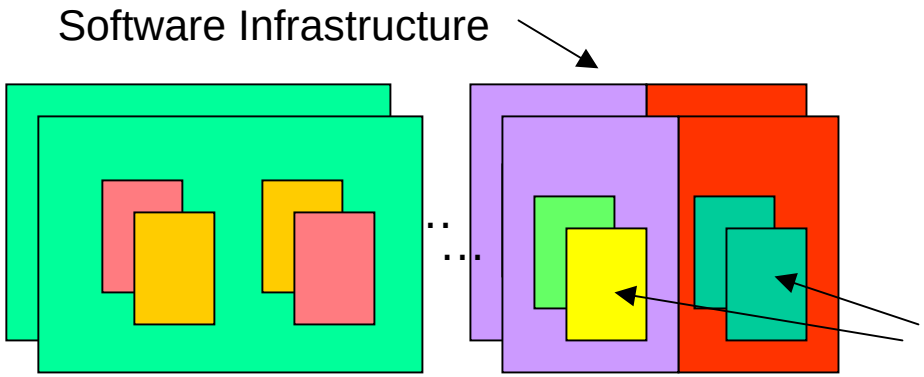


- Drug Discovery
- Microbial Engineering
- Molecular Ecology
- Oncology Research

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

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

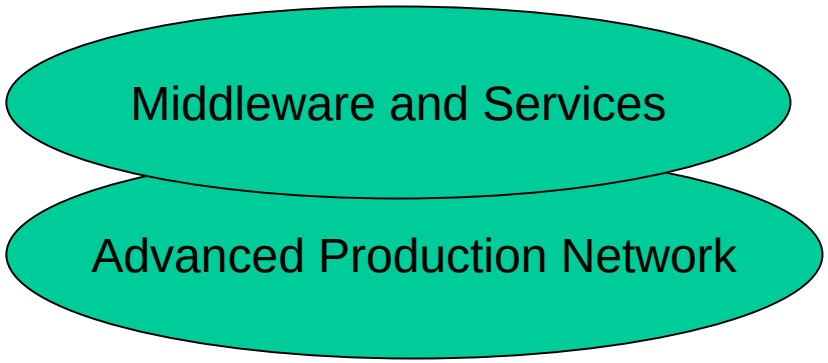
Attributes of Model



Applications and Community Tools Layer
(e.g. sensor toolkits, simulation codes, etc.)

Loosely Coordinated
- Competition
- Cooperation
- Collaboration

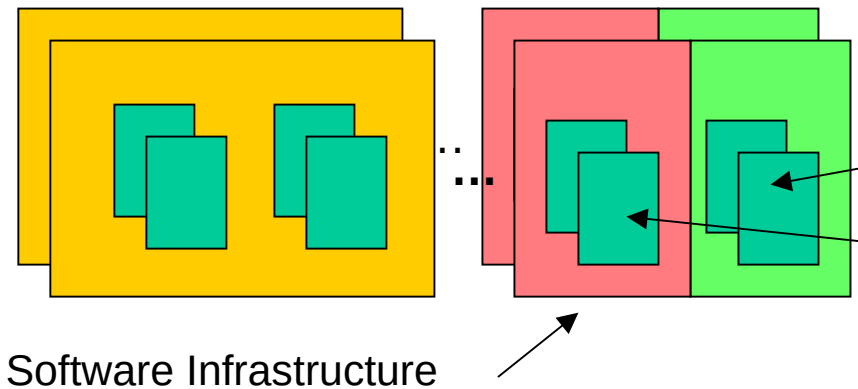
Value in Diversity



Integration Connectivity Communication Layer
(e.g. security, transport, brokers, CAs, user support, services, hosting)

Highly Coordinated
- Cooperation
- Collaboration

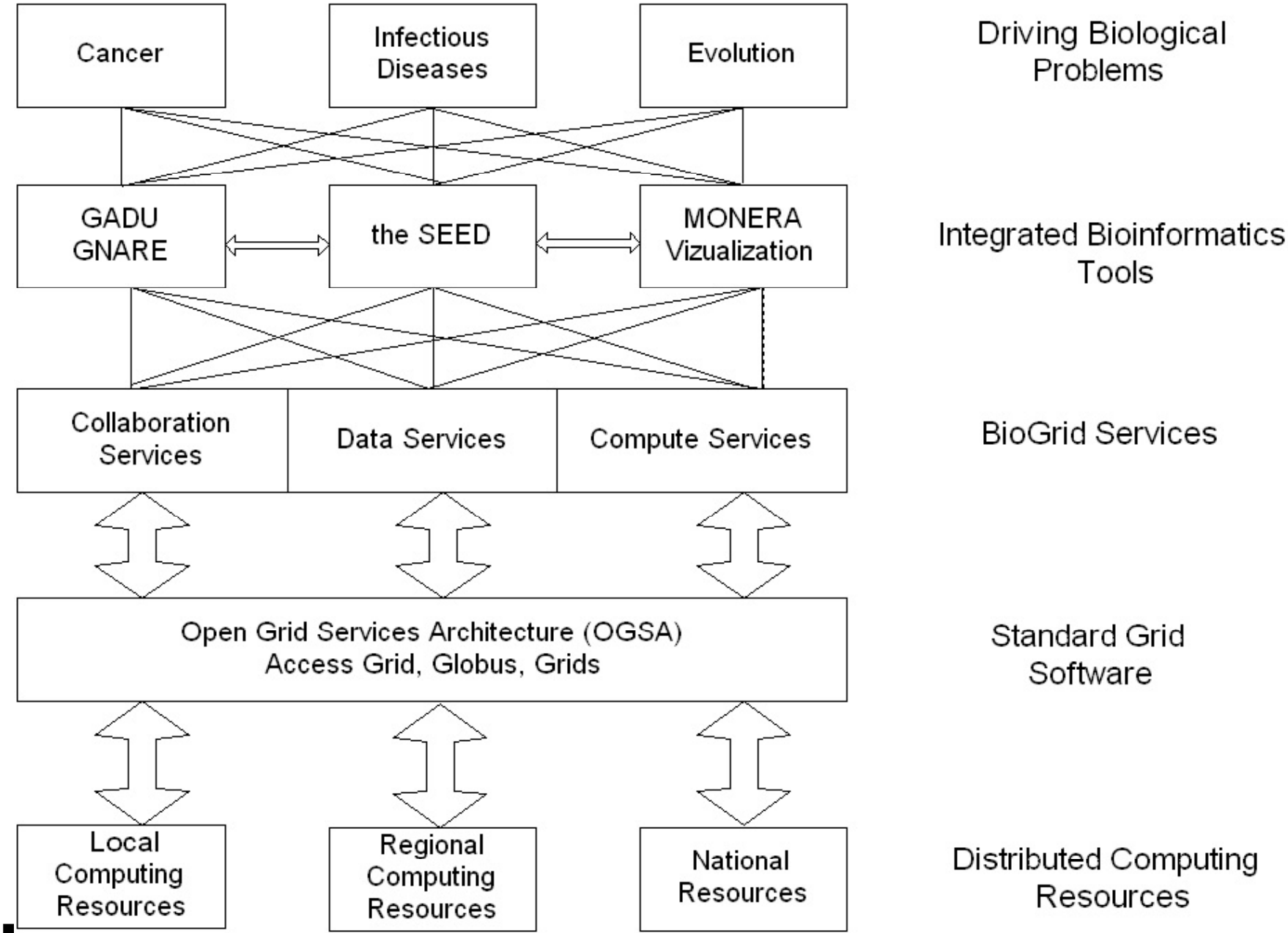
Value in Commonality and Stability, Ubiquity



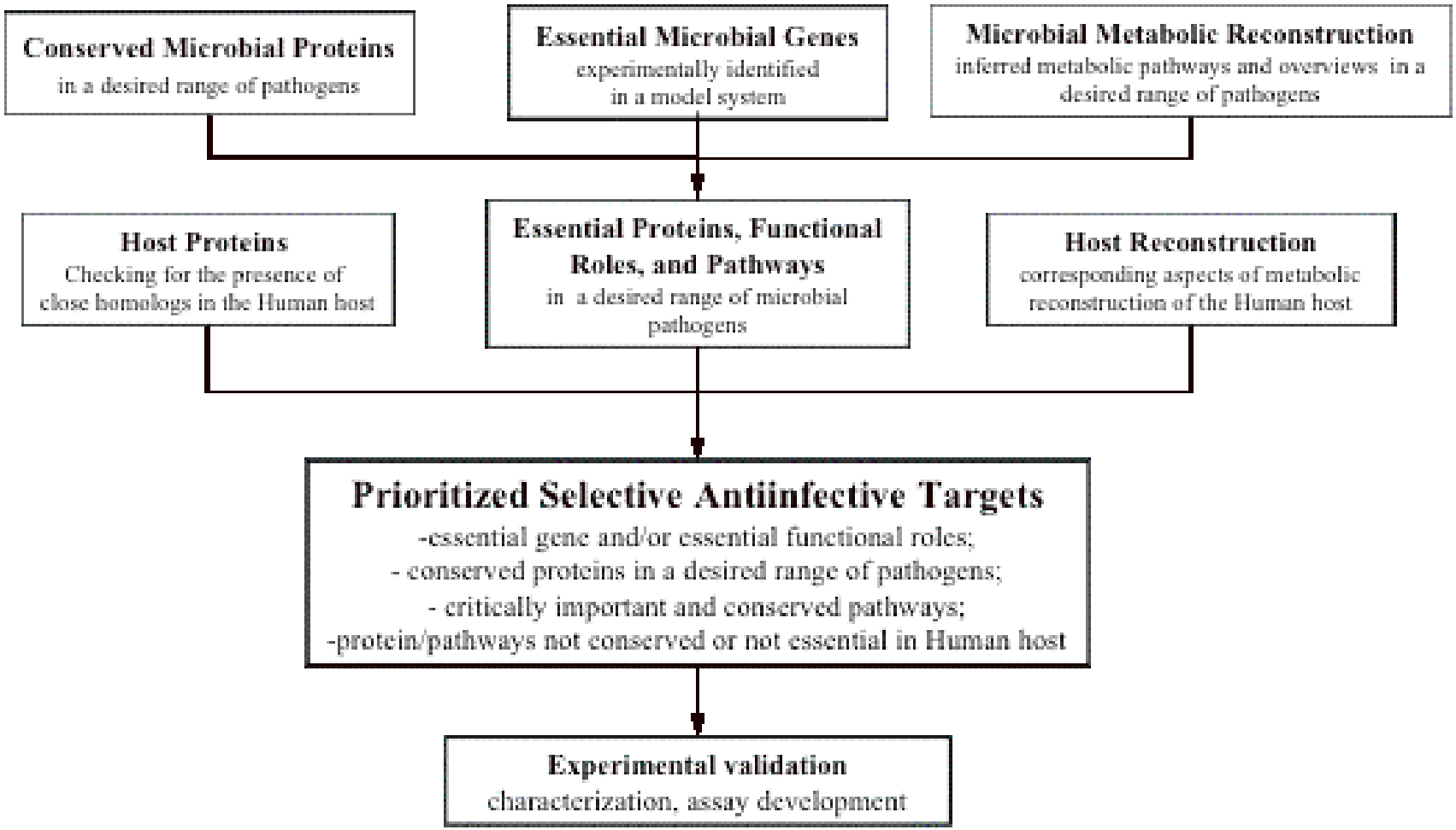
Resources Layer
(e.g. computers, data systems,
sensor networks,
instruments)

Loosely Coordinated
- Competition
- Cooperation
- Collaboration

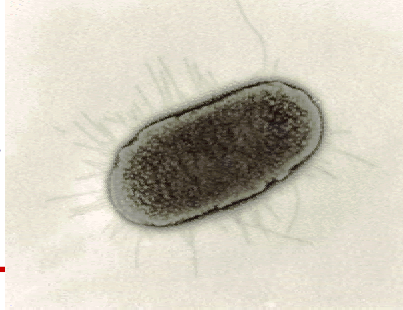
Value in Innovation and Scale

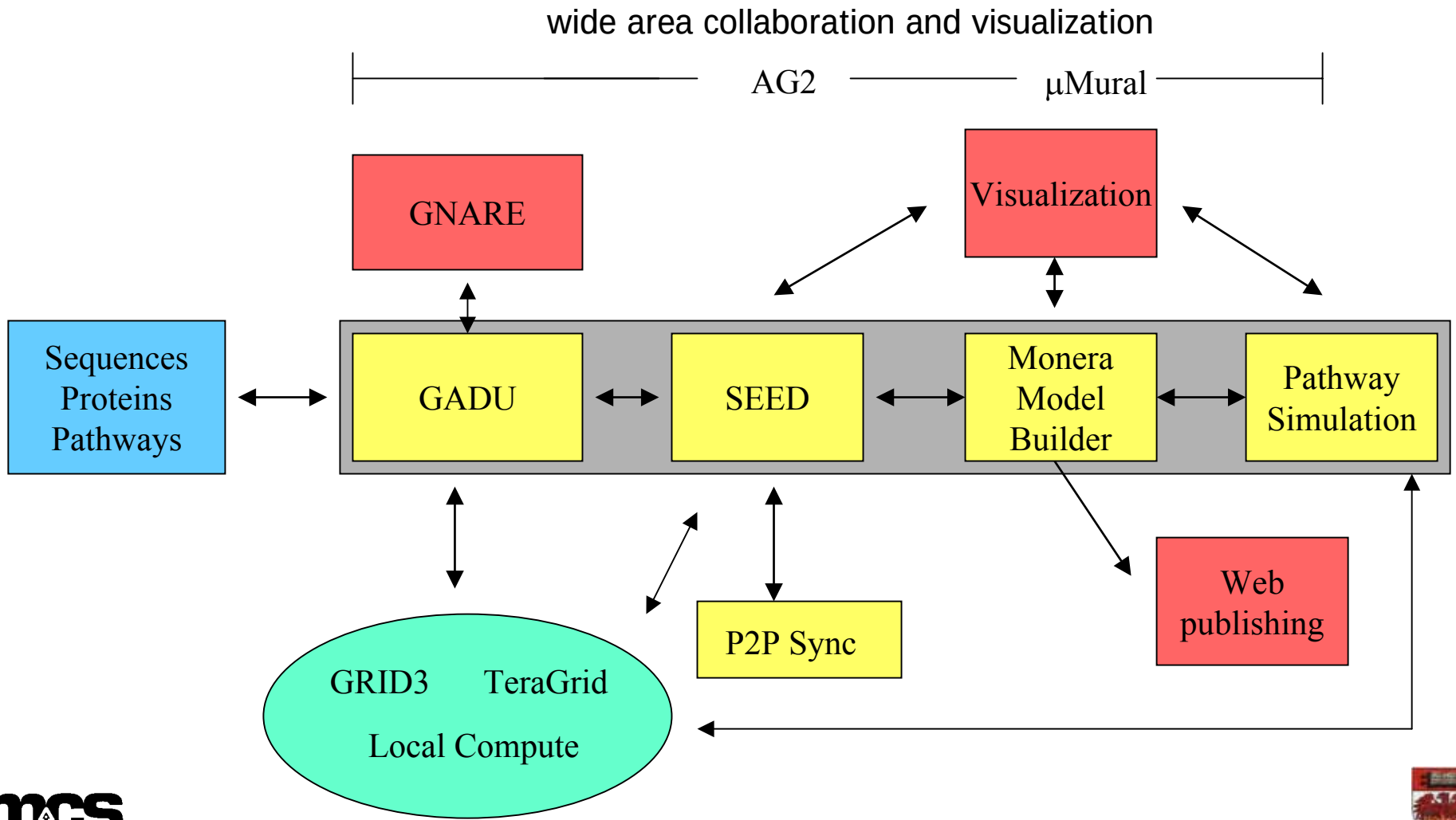

Cancer
Infectious Diseases
Evolution
Driving Biological Problems
GADU GNARE
the SEED
MONERA Vizualization
Integrated Bioinformatics Tools
Collaboration Services
Data Services
Compute Services
BioGrid Services
Open Grid Services Architecture (OGSA)
Access Grid, Globus, Grids
Standard Grid Software
Local Computing Resources
Regional Computing Resources
National Resources
Distributed Computing Resources

Functional Genomics Approach to Anti-microbial Agent Development

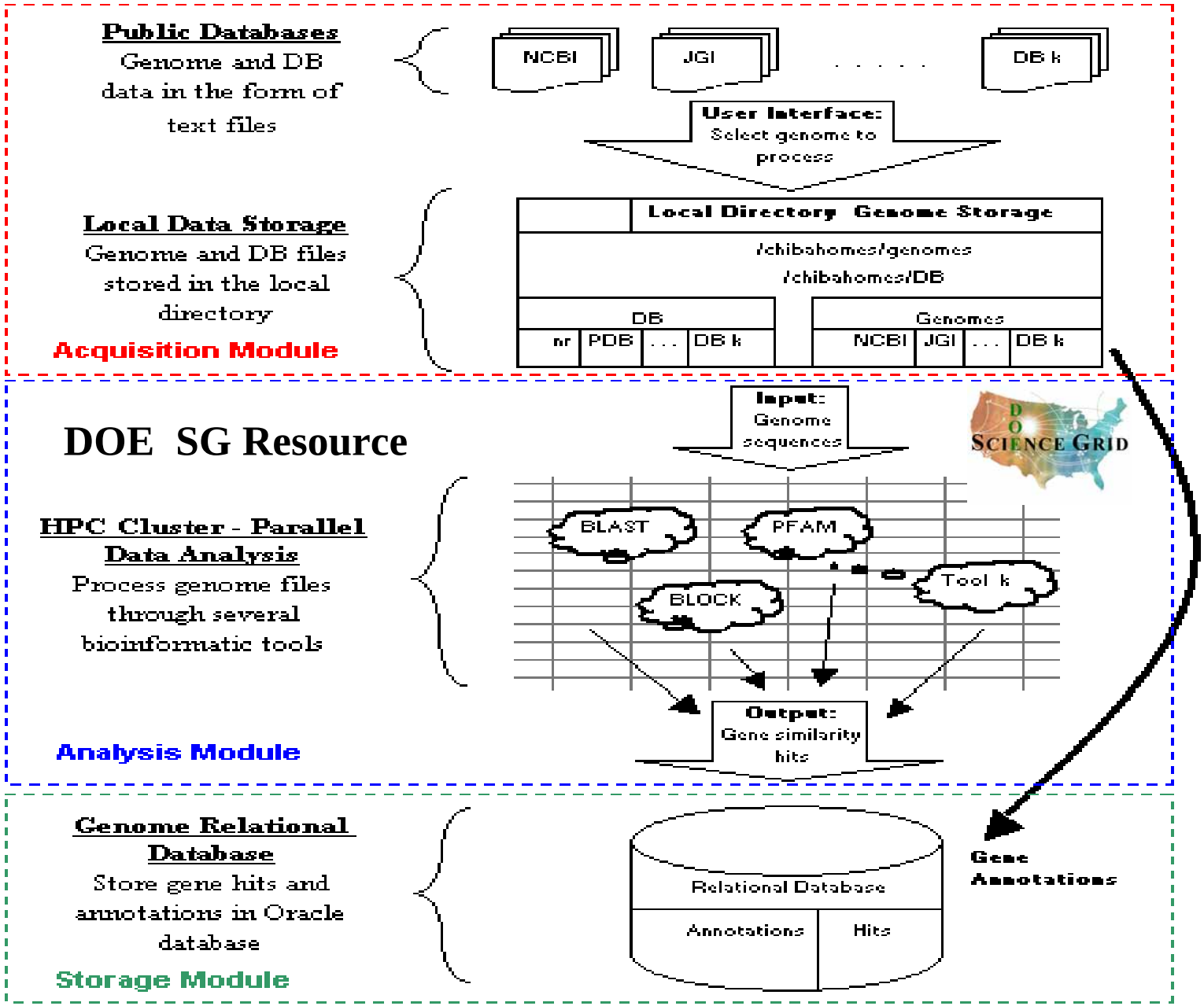


Argonne Systems Biology Workflow



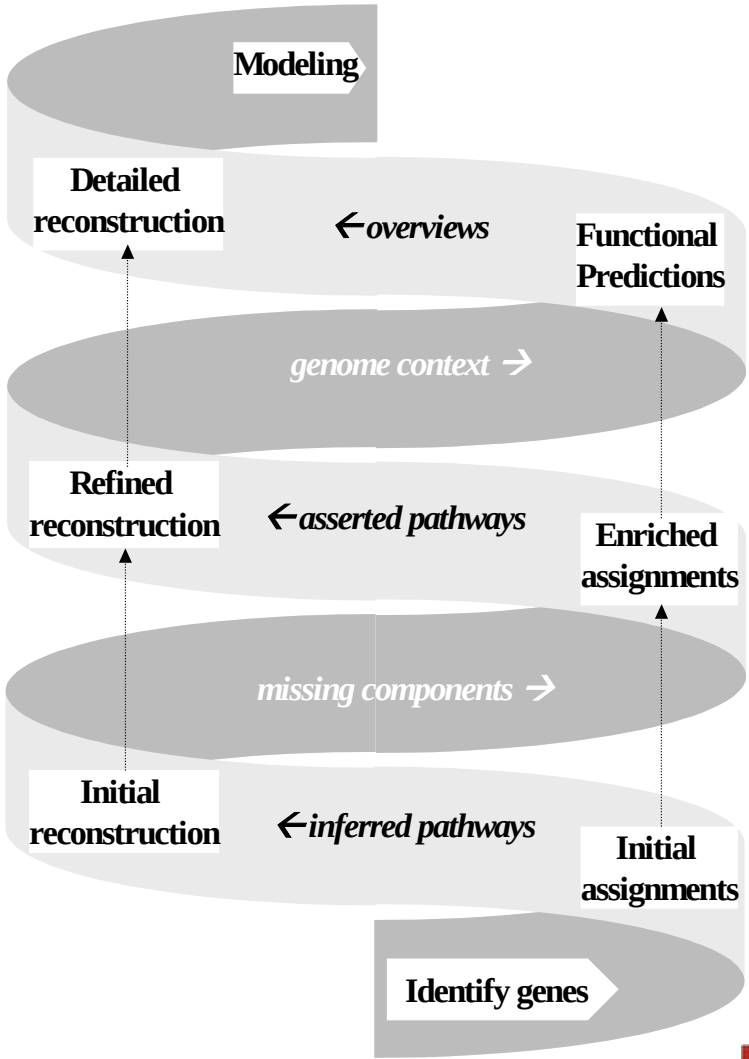


GADU Data Flow

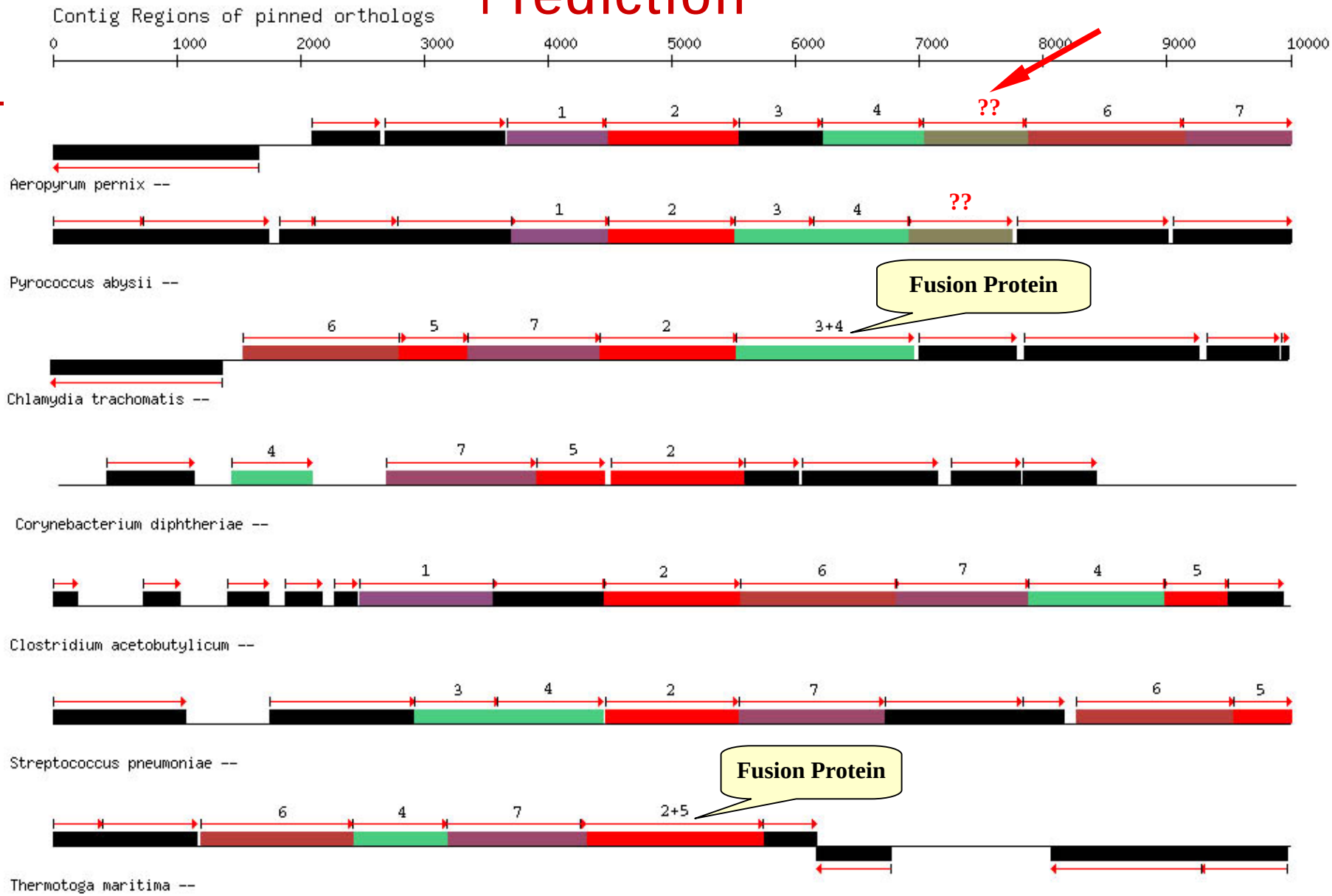


What Is Annotation All About?

- We identify the genes
- We identify clear functions and make tentative guesses
- We build an initial metabolic reconstruction
- We refine all estimates iteratively, seeking maximal consistency
- We identify "missing genes"
- We build "portfolios" to support identification of missing genes
- We confirm these assertions.
- We project the results



Chromosomal Clustering: Prediction



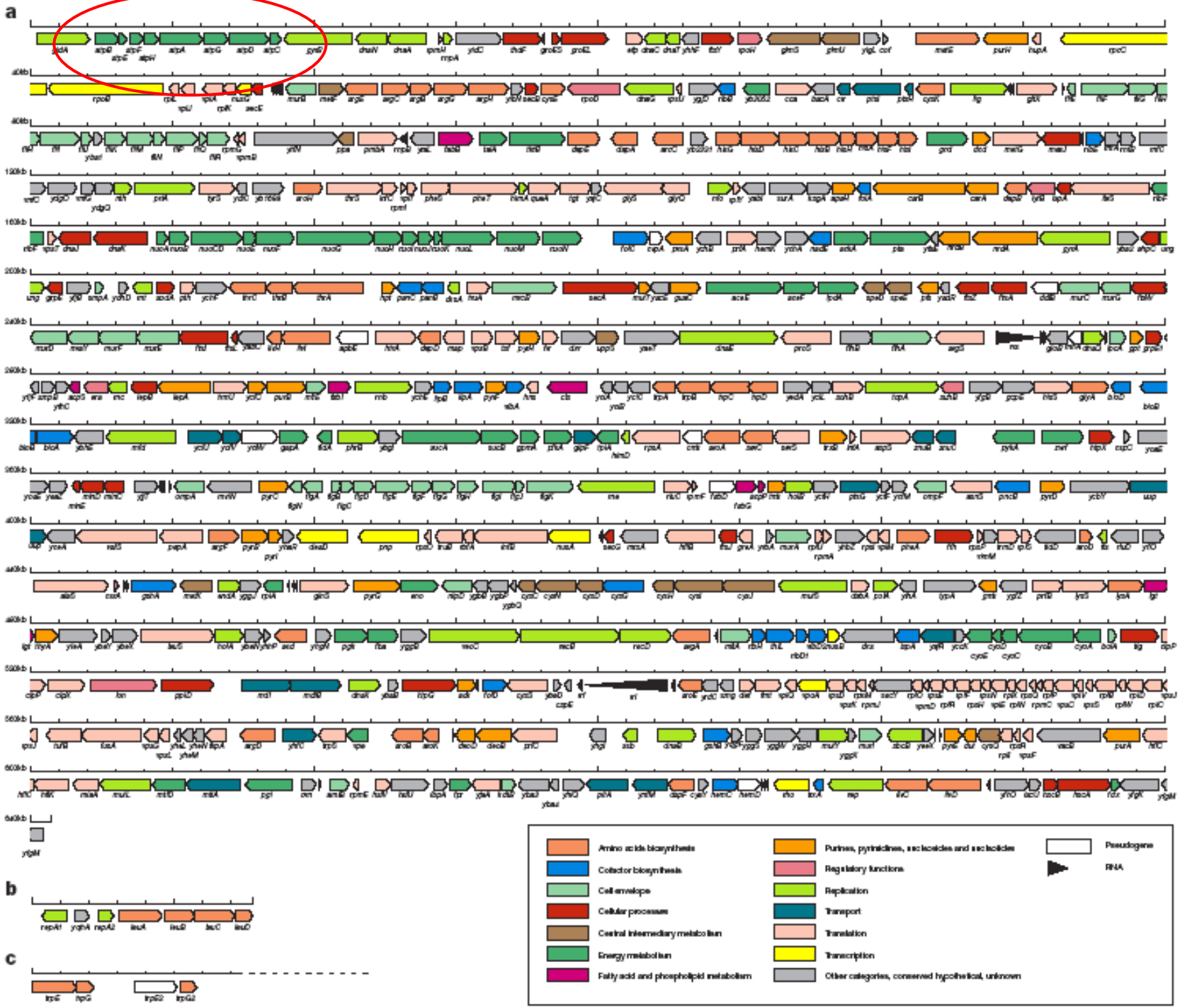
Buchnera sp.

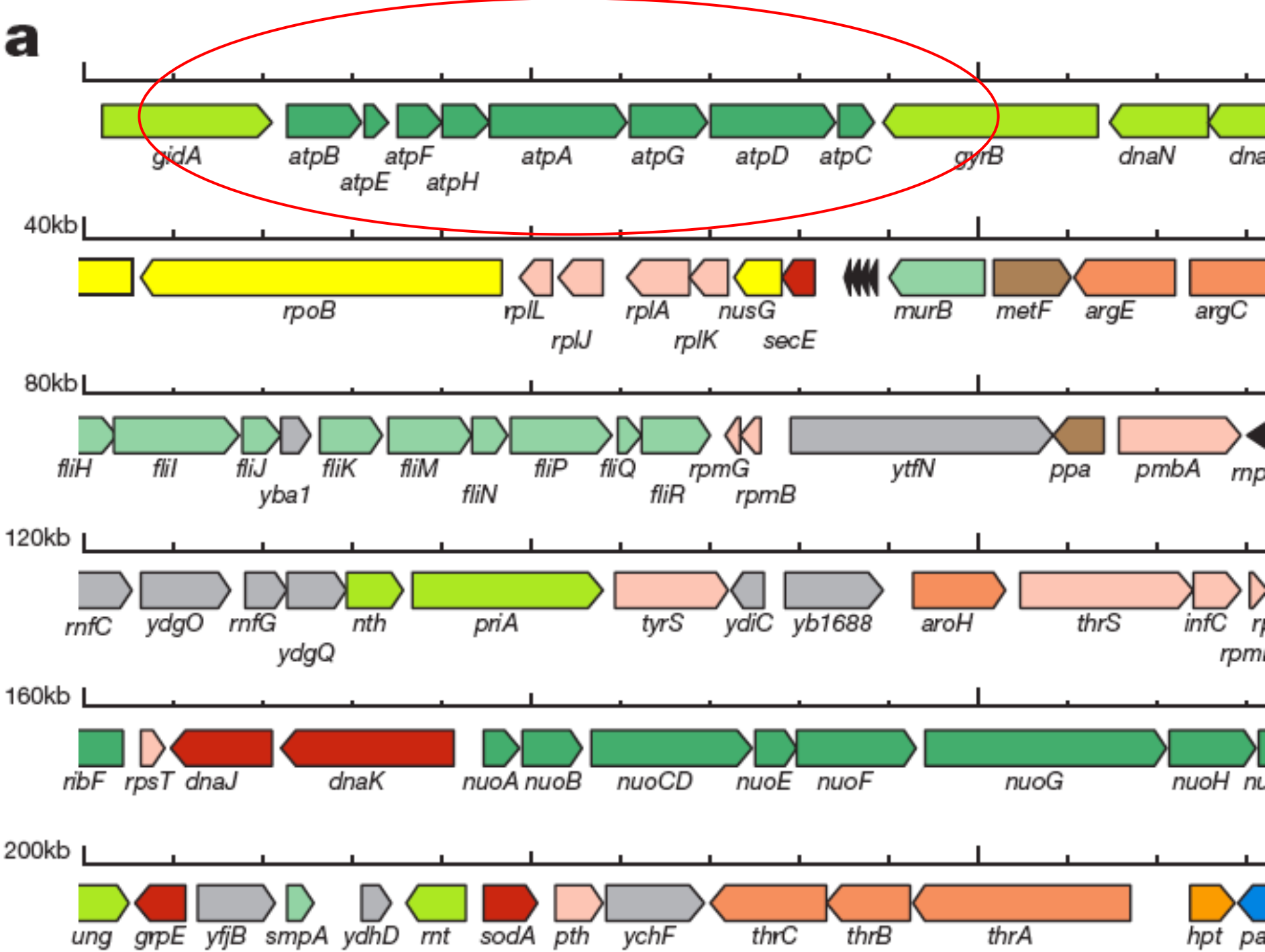
QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

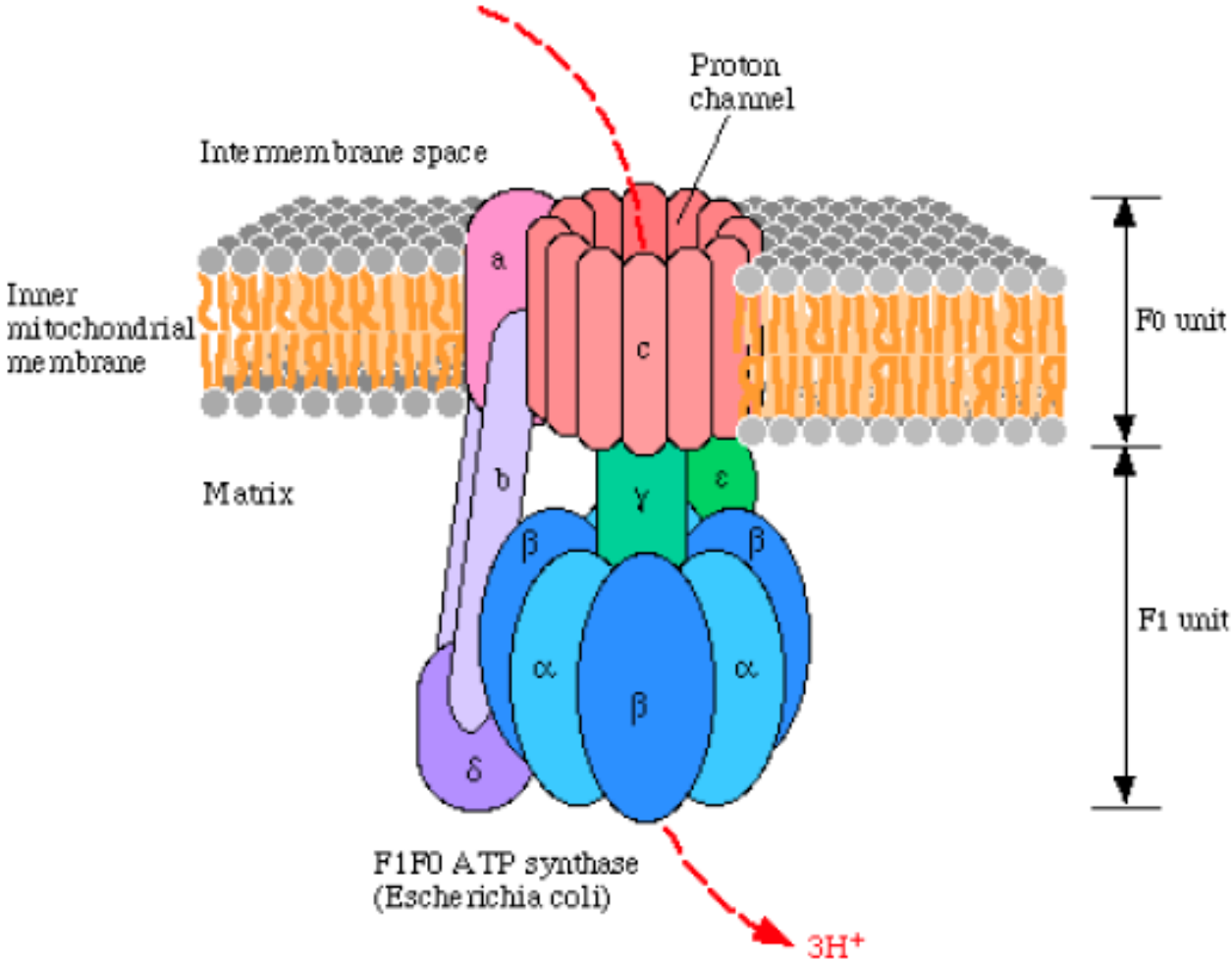
QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.


a
b
c
Amino acids biosynthesis
Cofactor biosynthesis
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid and phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Transport
Translation
Transcription
Other categories, conserved hypothetical, unknown
Pseudogene
RNA


a
gidA
atpB
atpE
atpF
atpH
atpA
atpG
atpD
atpC
gyrB
dnaN
40kb
rpoB
rplL
rplJ
rplA
rplK
nusG
secE
murB
metF
argE
argC
80kb
fliH
fliI
fliJ
yba1
fliK
fliM
fliN
fliP
fliQ
fliR
rpmG
rpmB
ytfN
ppa
pmbA
120kb
rnfC
ydgO
rnfG
ydgQ
nth
priA
tyrS
ydiC
yb1688
aroH
thrS
infC
160kb
ribF
rpsT
dnaJ
dnaK
nuoA
nuoB
nuoCD
nuoE
nuoF
nuoG
nuoH
200kb
ung
grpE
yfjB
smpA
ydhD
rnt
sodA
pth
ychF
thrC
thrB
thrA
hpt

ATP SYNTHESIS



F-type ATPase (Bacteria)

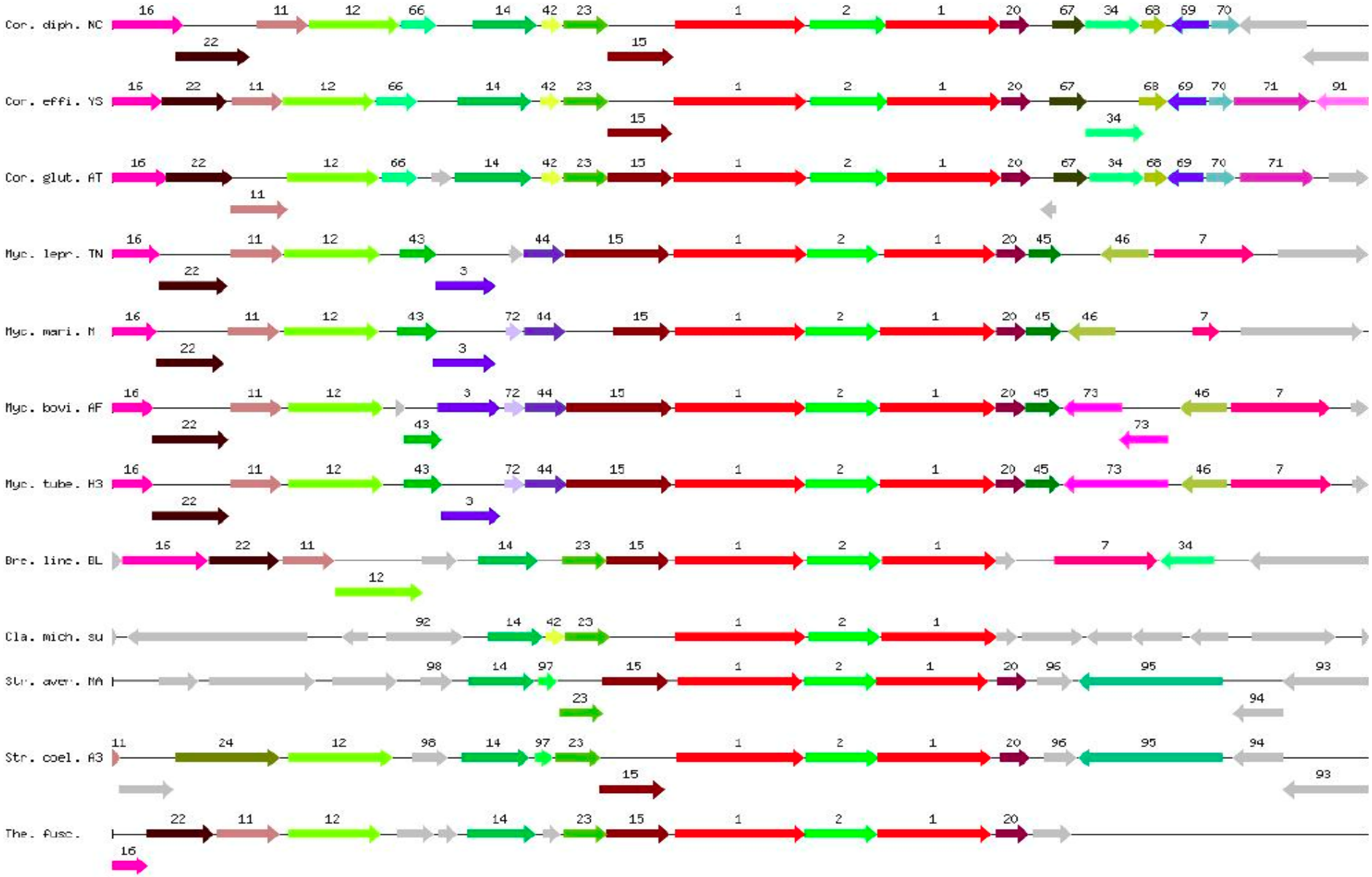
beta	alpha	gamma	delta	epsilon	c	a	b

Multiple Sequence Alignments

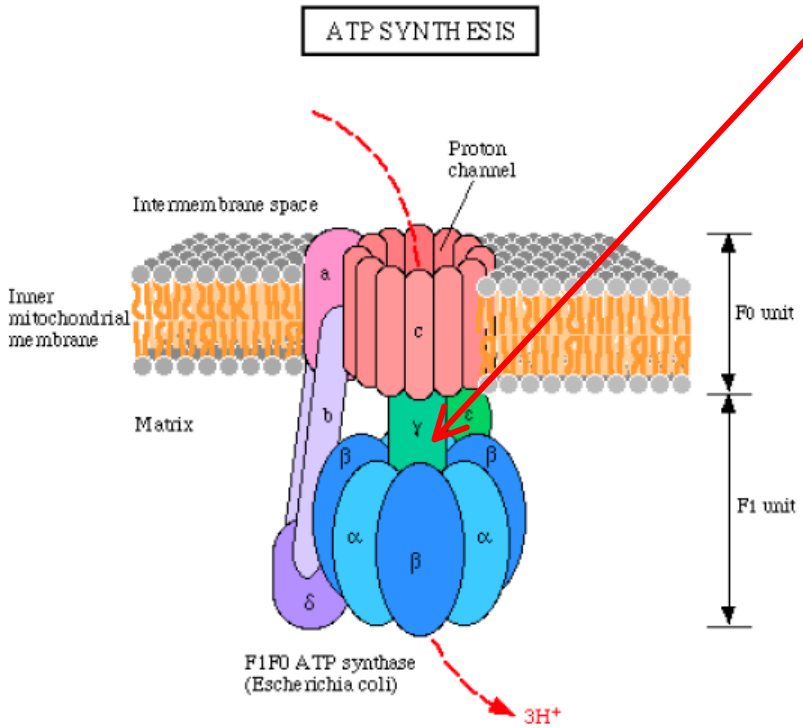
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

r

Gene Cluster Alignments



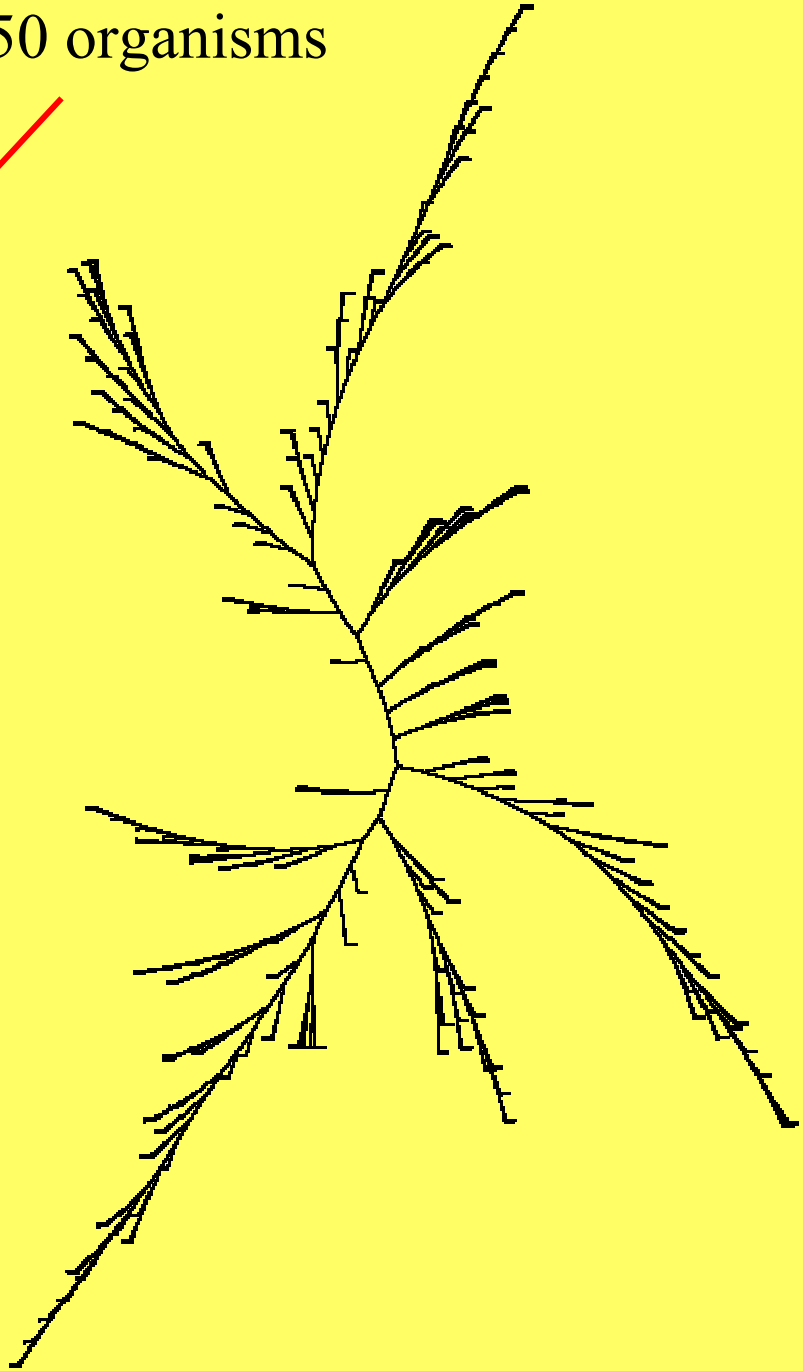
ATP gamma subunit



F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b

250 organisms

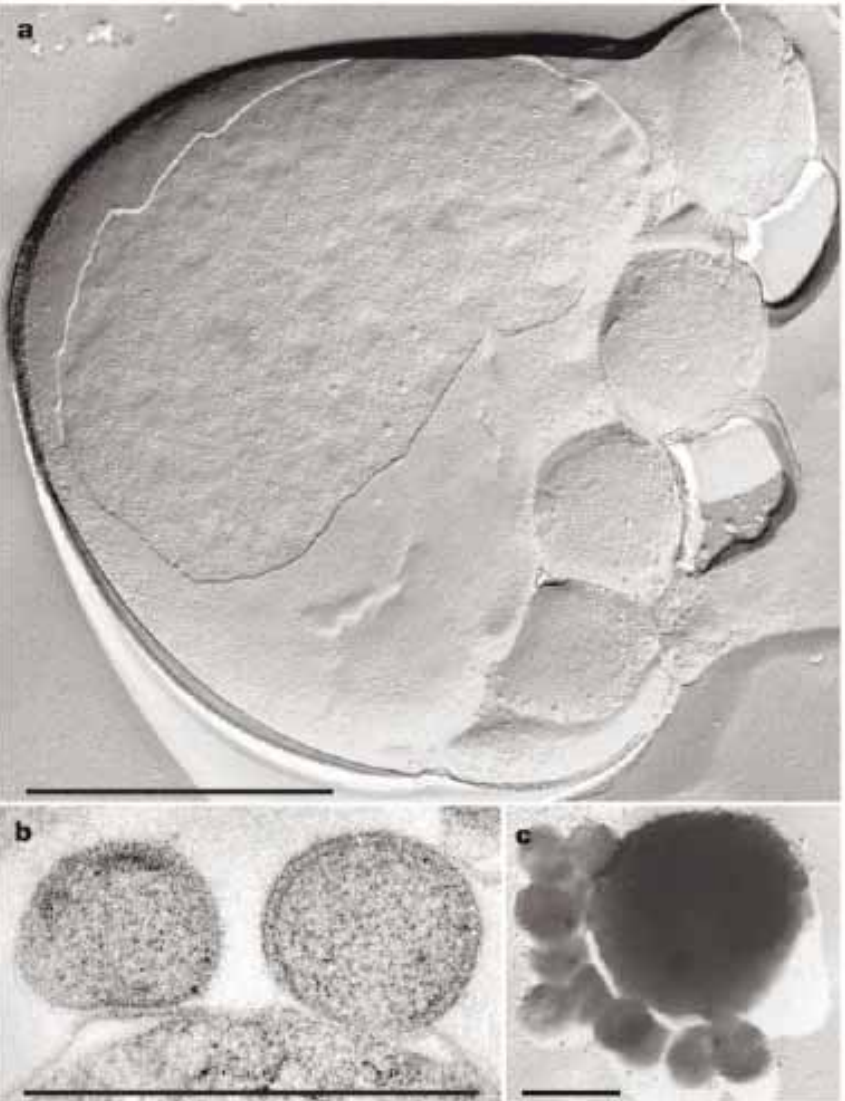


Large-Scale Computing and the Hunt for Horizontal Gene Transfer

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

- Organismal Trees
  - 16S rRNA subunit
  - Consensus trees from multiple genes
- Gene Trees
  - Phylogeny for ~1000 most conserved genes x ~200 organisms
  - Core metabolism (TCA, EMP, PPP), DNA replication and repair, nucleotide synthesis, amino acid biosynthesis, etc.
- Operon Trees
  - Phylogeny for ~50 more conserved operons (gene clusters)
- Reconciliation
  - Sorting events, co-speciation, duplication, horizontal transfer
  - About 1 million reconciled trees will be generated

Nanoarchaeum equitans



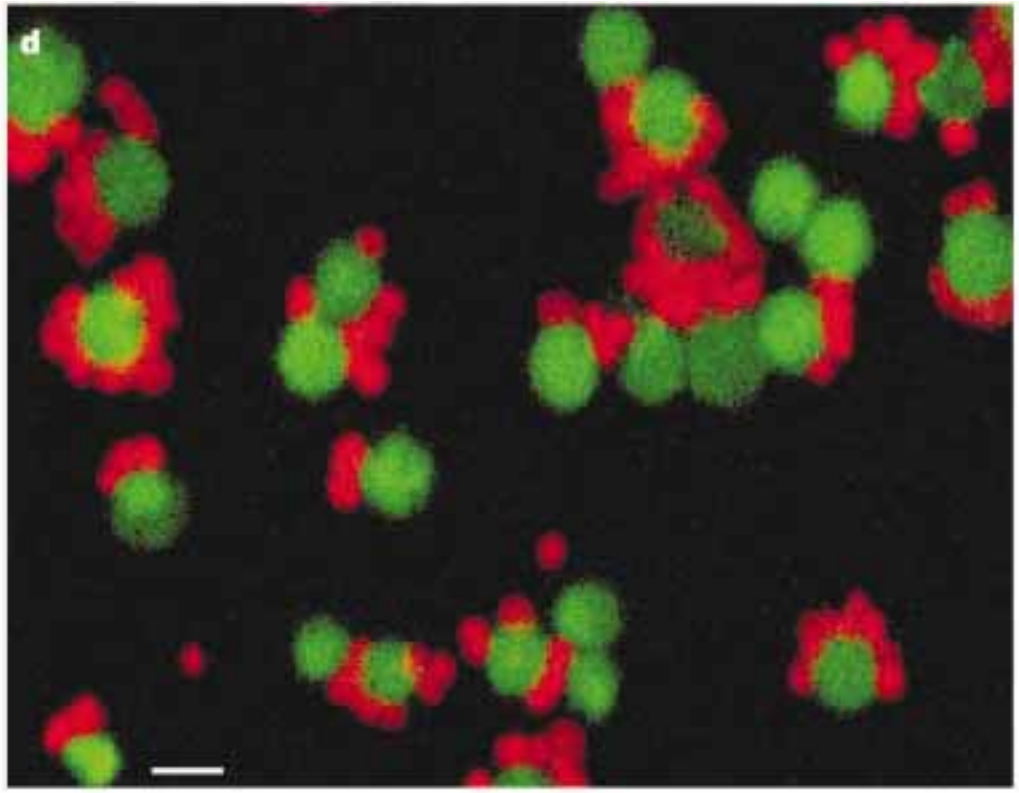
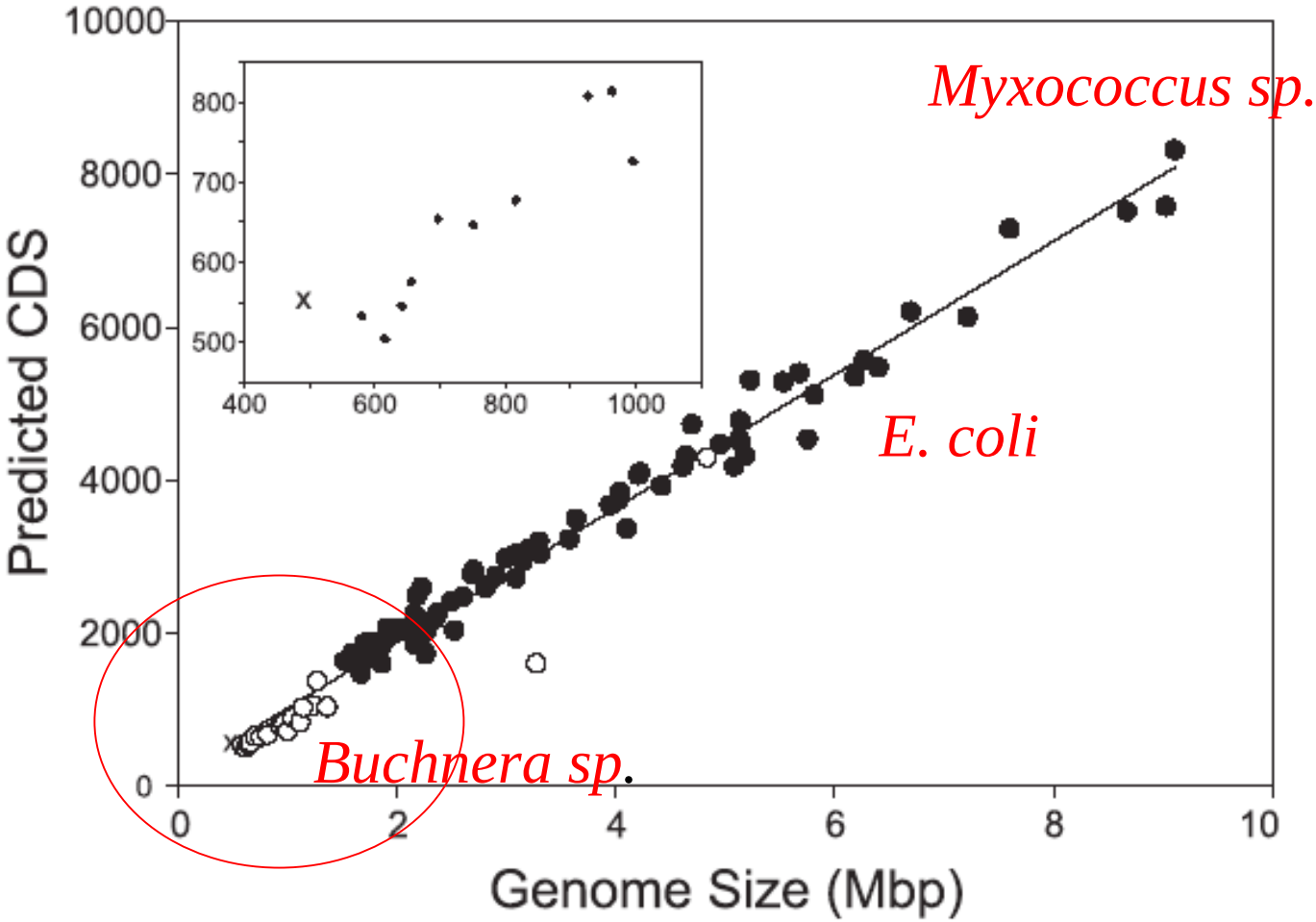
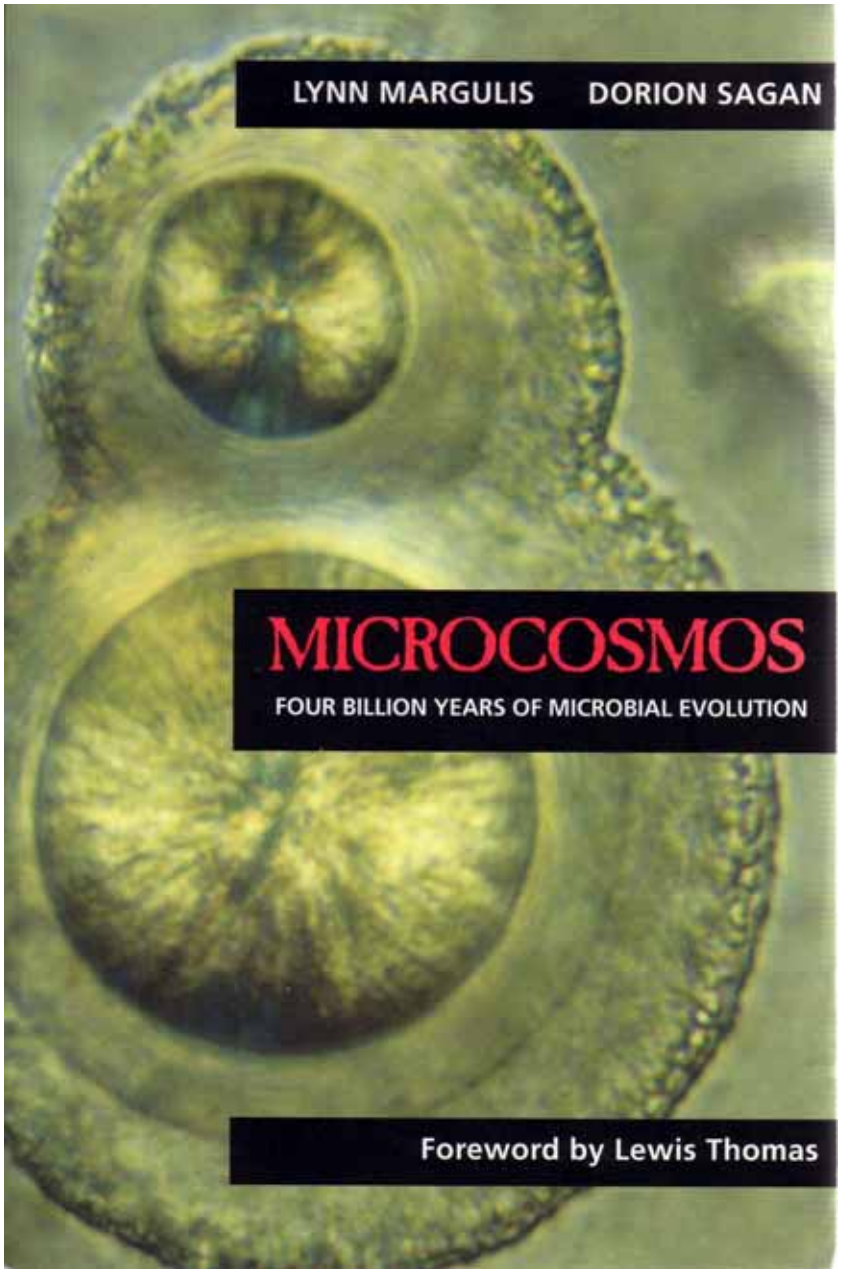


Figure 1 Electron microscopy and fluorescence light microscopy of the '*Nanoarchaeum equitans*'–*Ignicoccus* sp. coculture. **a**, Freeze-etched cell of *Ignicoccus* and four attached cells of '*Nanoarchaeum*', showing their crystalline S-layer with sixfold symmetry. **b**, Ultrathin section of two cells of '*Nanoarchaeum*' attached to the outer membrane of *Ignicoccus*. **c**, Cell of *Ignicoccus*, with several cells of '*Nanoarchaeum*' attached on the left side; platinum-shadowed. **d**, Confocal laser scanning micrograph after hybridization with the CY3-labelled probe 515mcR ('*Nanoarchaeum*') and rhodamine-green-labelled probe CREN499R (*Ignicoccus*). **a–d**, Scale bar, 1.0 μm.



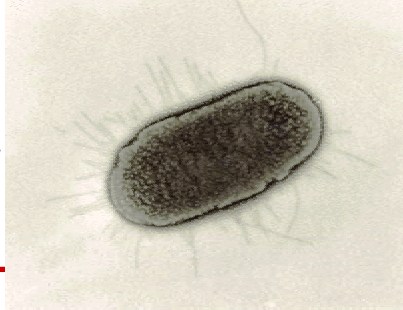
A Little Mystery?

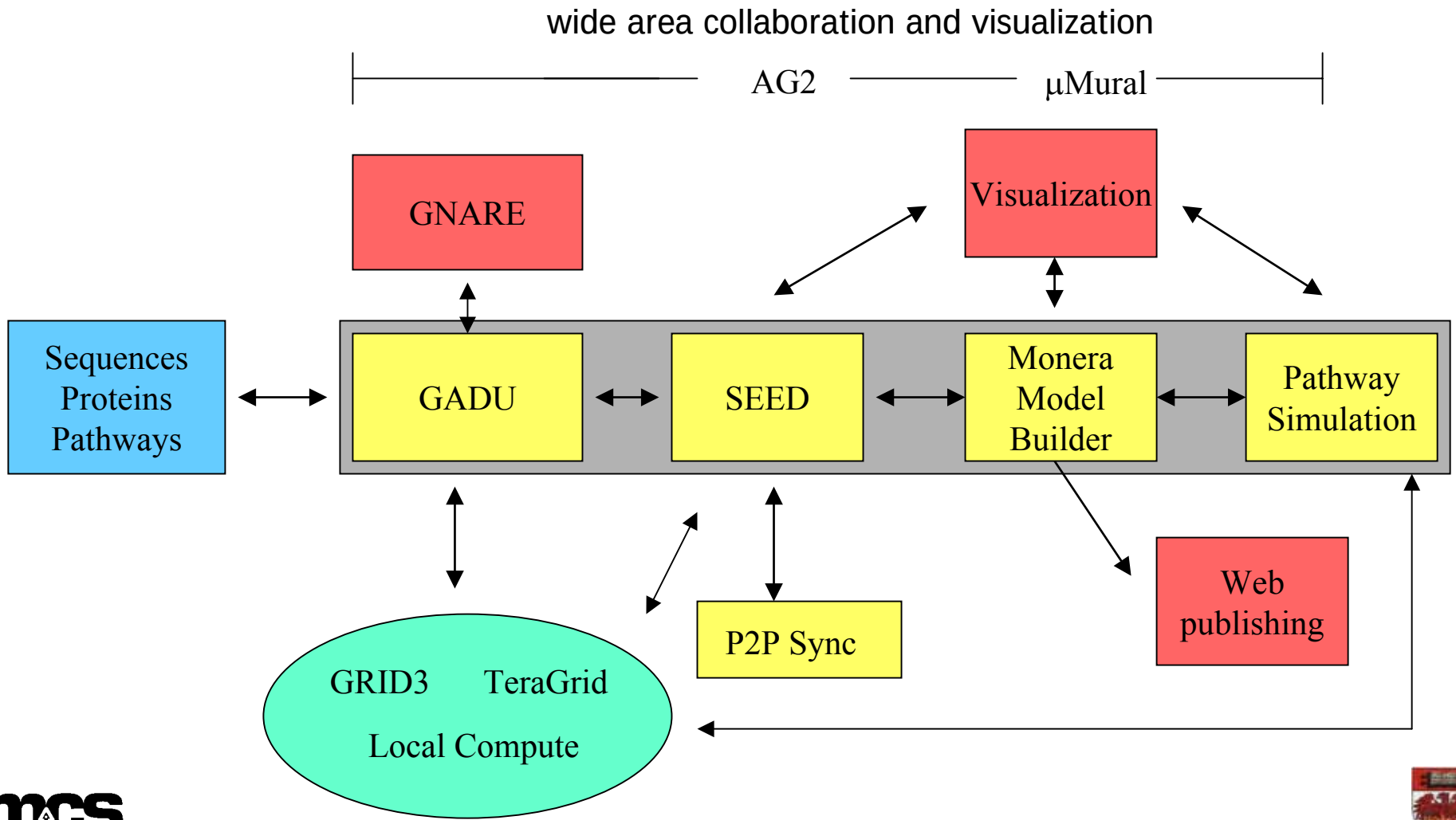
Fig. 1. Correlation between microbial genome size and the number of predicted coding DNA sequences CDS. Bacterial genomes predicted to be undergoing reductive evolution are indicated by open circles, whereas other genomes are indicated by filled circles. The *N. equitans* genome is marked by "x". (*Inset*) An expansion of the data from small microbial genomes with the abscissa shown in genome size units of kbp.



- Lynn Margulis and Dorion Sagan
- Story of microbial evolution
- Introduction to early life on earth

Argonne Systems Biology Workflow





The SEED: Peer-to-Peer Software for Distributed Curation of Biological Data

- Support community-wide annotations and analysis of biological data
- Maintain an up-to-date collection of publicly-available datasets within the SEED framework
- Peer-to-peer synchronization is ideally suited for community-wide annotation of data collections
- Enable the SEED to be gracefully extended both via plug-in modules, but also through new forms of data integration

What do users need to be able to do?

- Install and share new genomes
  - Publish to a limited set of colleagues
- Share gene function assignments
  - Locally in a collaboration and remotely
- Share gene and subsystem annotations (notes)
  - Can be arbitrary annotations (xml, html,etc.)
- Share naming rules (translations)
  - Publish dictionaries (ontologies) as views
- Lightweight code update
  - Propagate code to peers to continuously update the algorithms

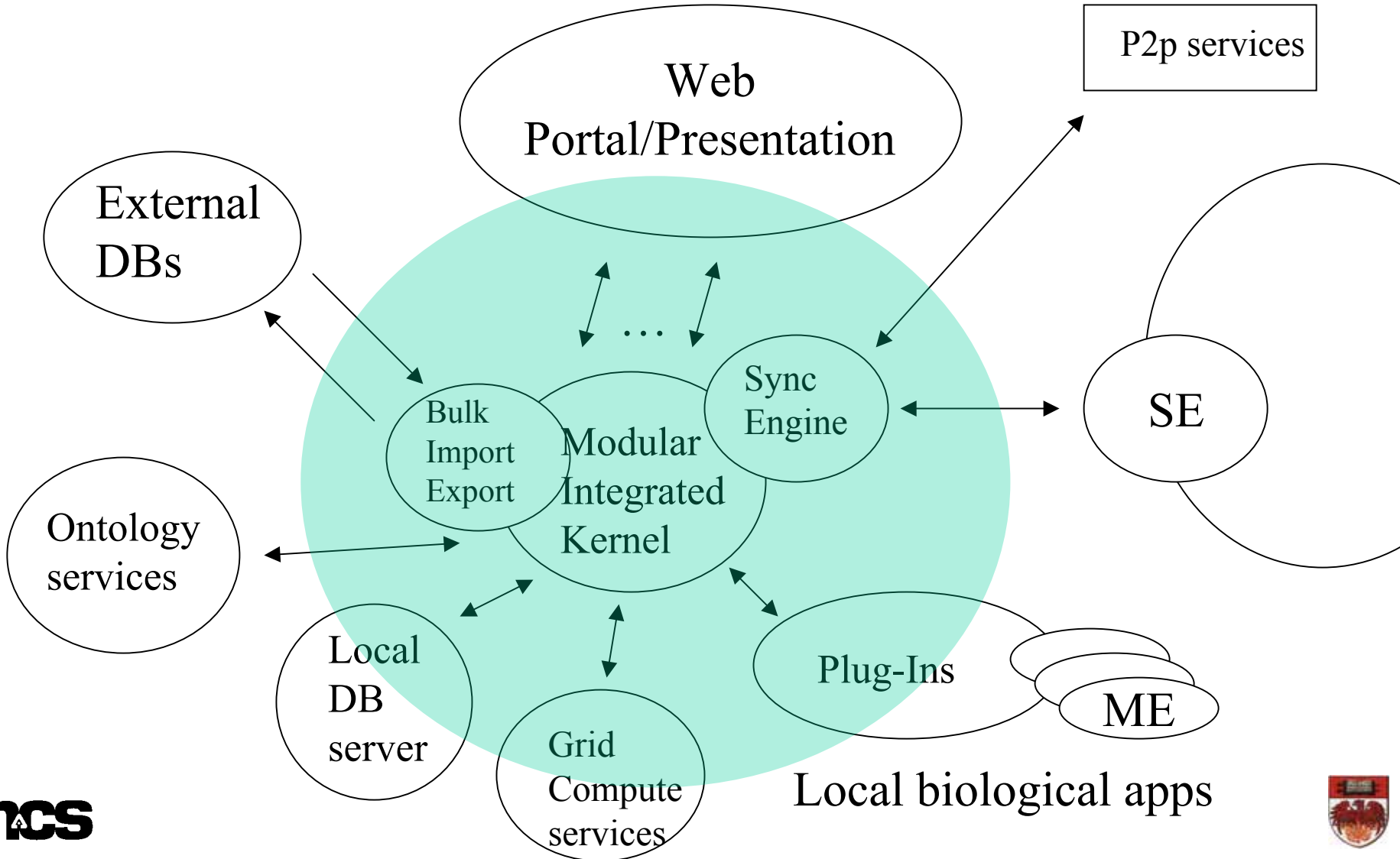
- Environment update (heavyweight, tools etc.)
- Clone the system for teams and peers

User Centered Grids

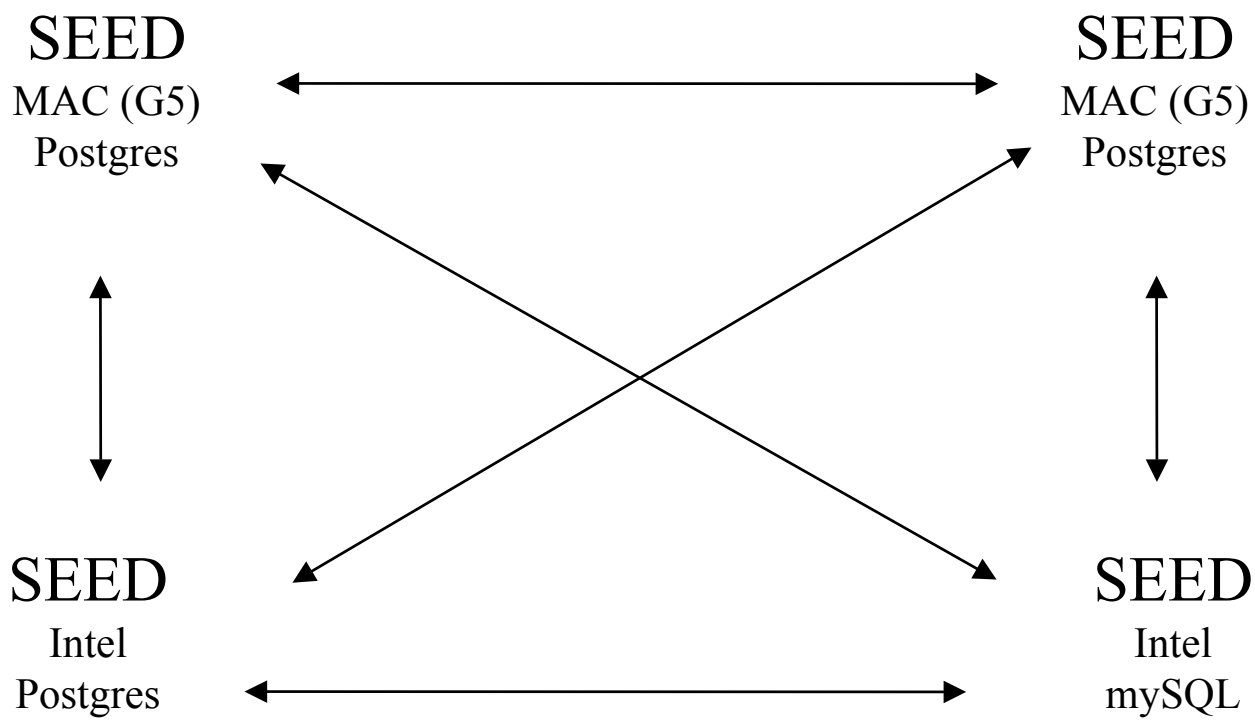
- Enabling end users to assemble collections of resources and share them within virtual communities
  - Applications Sharing
  - Data (Content) Sharing
  - CPU and Storage Sharing
- Transparent integration with existing user environments
  - Virtual installation of applications in parallel with personal installs
  - Virtual resources appear local
    - iMac storage for example
- Integration with dominate end user technologies
  - Microsoft XP etc.
  - Globus, etc.

Peer-to-Peer Open Life Sciences Grid

the prototype SEED



Peer-to-Peer Demonstration at SC'03



Scalability of the SEED P2P Design

QuickTime™ and a TIFF (LZW) decompressor are needed to see this pictu

- 100,000s of genomes (organisms)
  - Including support of many close variants
- Billions of genes and gene products
  - GBs-TBs of annotation per gene or pathway
- Ten's of Thousands of deployed instances
  - Thousands of cooperating sites
  - Gigabit class networks
- Many update channels (pub/sub)
  - Thousands (some private, some open)
- Loose synchrony
  - Hourly/Daily/Weekly/Monthly updates

Example Biological Use Cases

- Comparative analysis of gene clusters
- Looking for missing genes
- Comparing pathways between organisms
- Reconstructing core cellular machinery
- Extracting rules for model development
- Visualization of whole cell networks
- Studying horizontal gene transfer
- Studying evolution of metabolism
- Designing new pathways
- Designing new organisms

Grid Futures!

- Grid Services ≥ Web Services
  - at least for large-scale science applications
- Two big open problems
  - Simple Grid development environment for scientific end users (e.g. Grid version of csh)
  - Peer-to-Peer environments as Grid environments
    - Enable ad hoc deployment by end users
- Metric for Grid progress needs to be # of deployed production applications
  - Compare to other platforms (linux, mac, windows)

BioGrid Services - Collaboration

- Based on our 10+ years of work on collaboration tools and collaborative environments
- Core services are:
  - Event channels to provide application synchronization
  - Venues for sharing and storing state
  - Real-time media (audio, video, text)
  - Access to Grid services

BioGrid Services - Data

- Two primary needs
  - Peer-to-Peer data updates and synchronization
    - Updates to local databases (annotations, genomes, etc.)
    - Updates to local code bases (new functionality)
    - Support of data access restrictions/rights
  - Access to large-scale shared data resources
    - Databases too large for large-scale replication
      - Microarray data, similarity matrices, imaging, mass-spec
    - Real-time data feeds from instruments
      - High-throughput data
- Proposed data model for large-scale access
  - Attribute-value pairs (via scripting interfaces PERL, Python)

BioGrid Services - Compute

- Two Primary needs
  - Simple scripting interface to Grid resources
    - "Grid shell" enabling users to move existing pipelines to Grids
    - Prototype grid shell
      - Map grid resources into hierarchical (filesystem like) tree structures
  - Tools for workflow management
    - Enabling complex workflows to be scheduled against Grid resources
      - Moving from simple pipes to complex flows with human-in-loop
    - Virtualization of compute, data and collaboration (UI) resources
      - Enabling access to grid resources via simple scripting tools

Taiwan's NCHC's SARS Combat Task Force

sarsgrid.nchc.org.tw

- System consists of Access Grid, H.323 VTC, medical information management, emergency dispatch and network monitoring.
- Access Grid nodes installed:
  - Sanchung Hospital
  - Chang Gung Memorial Hospital (CGMH)
  - Taiwan's Center of Disease Control
  - Taipei Municipal Jen-Ai Hospital
  - Operational in 10 medical centers on the island

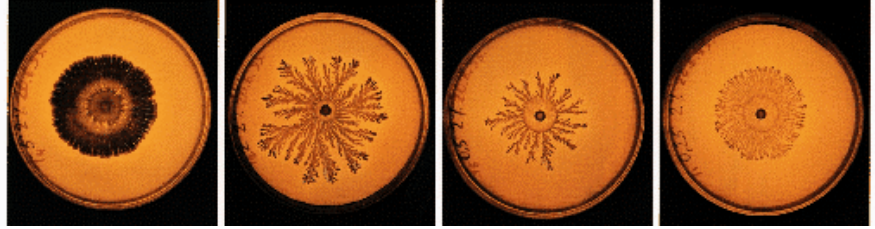
Intentional Engineering of Biological Systems

- Synthetic Biology is focused on the intentional design of artificial biological systems, rather than on the understanding of natural biology. It builds on our current understanding while simplifying some of the complex interactions characteristic of natural biology.

- First International meeting @ MIT June 10-12th in Cambridge MA

- Syntheticbiology.org

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, NIH, ANL, UC, Microsoft and IBM for supporting my research
- Ian Foster (ANL/UC), Ross Overbeek (ANL/FIG), Dan Reed (RENCI/UNC), John Wooley (UCSD), Mike Colvin(UCM), Richard Gardner (InCellico), and others contributed to this talk

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.