

# Some Interesting Problems in Systems Biology

Rick Stevens

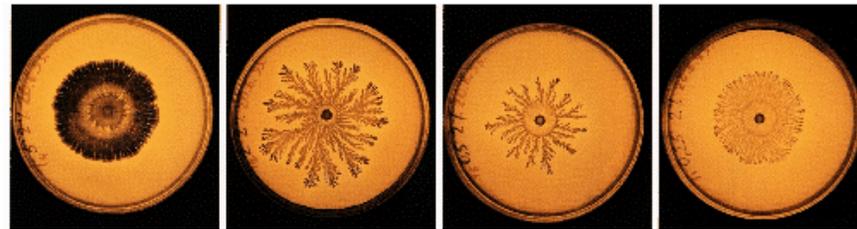
Argonne National Laboratory

The University of Chicago

Stevens@cs.uchicago.edu

Describe

Explain



Predict

Control

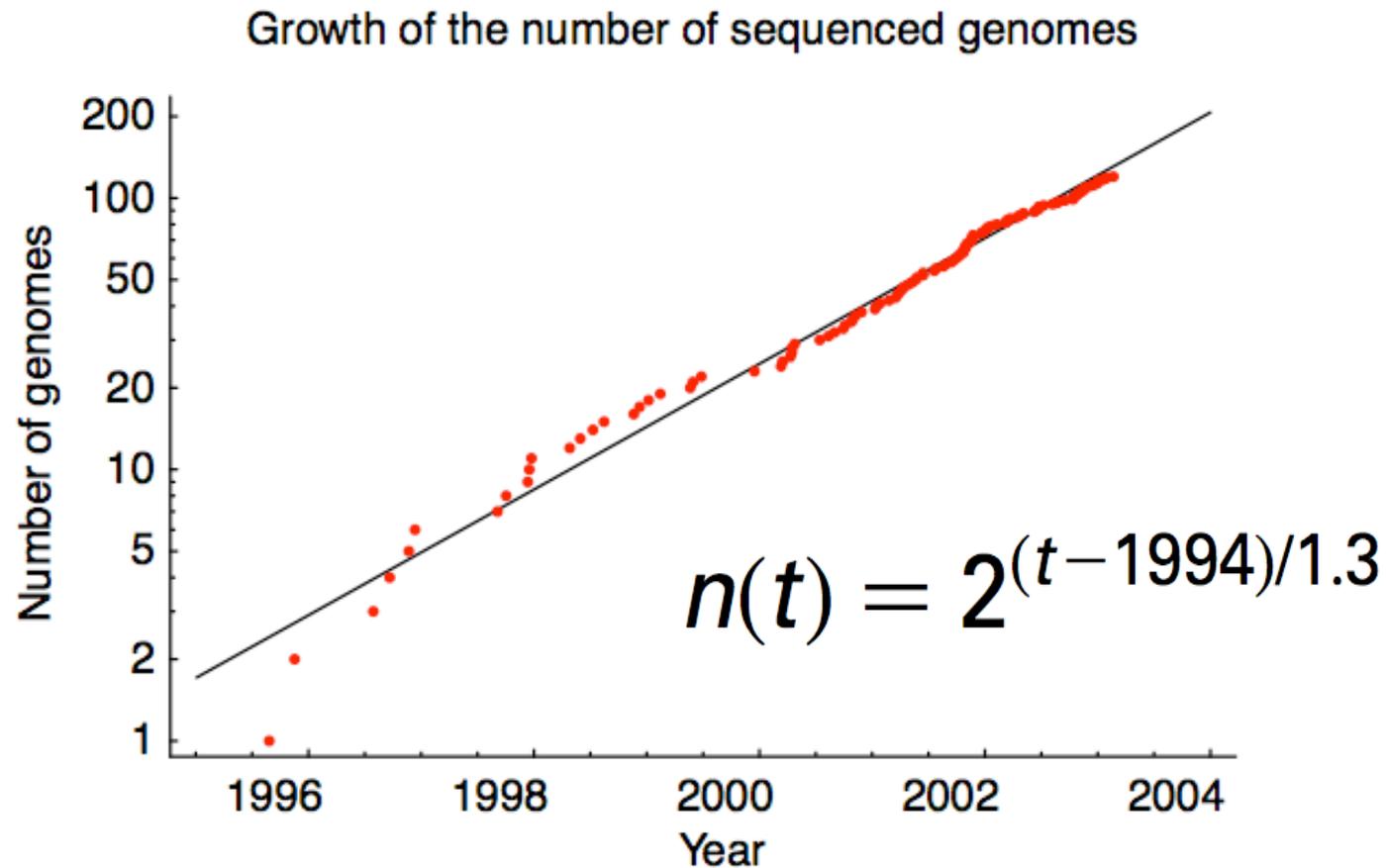
# The Outline of the Biology Revolution

---

- There will be many many genomes sequenced.
- Comparative analysis is key to improved genome annotation (assignment of function to genes).
- Evolution is the basis for gaining insight from comparative analysis.
- Advances in computing has made the advent of rigorous systems biology possible.
- Computing will enable biology to become a true theory driven predictive science.
- Strong theory and computing capability will enable the design and engineering of biological systems.

# Growth in the Number of Genome Sequences

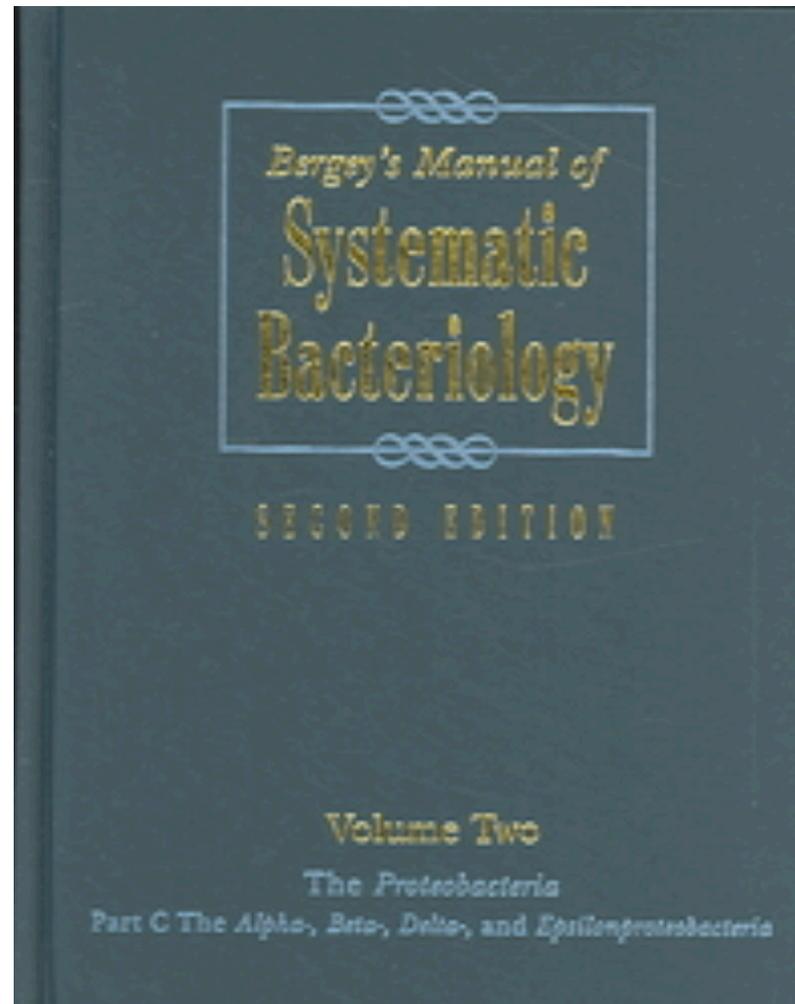
---



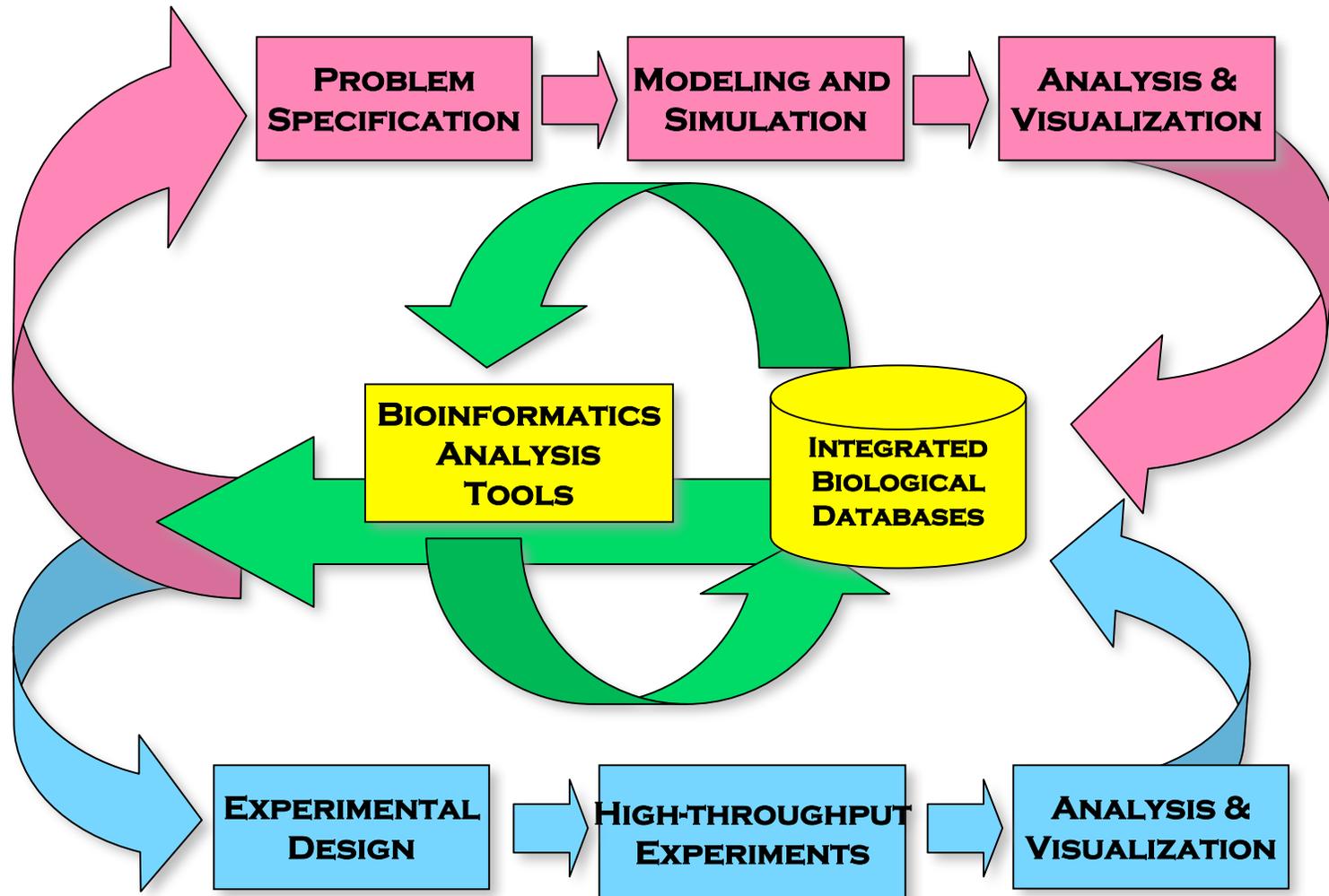
# Sequencing the Bergey's Manual

---

- Argonne and DOE's JGI are in discussions to launch a project to sequence all of the Bergey's Manual (all culturable prokaryotic organisms)
- 4900 taxa
- \$10M per year for about five years
- Motivating the high-throughput annotation of genomes (~24 hours turnaround time)



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



# Why Is Computational Biology Important to Me?

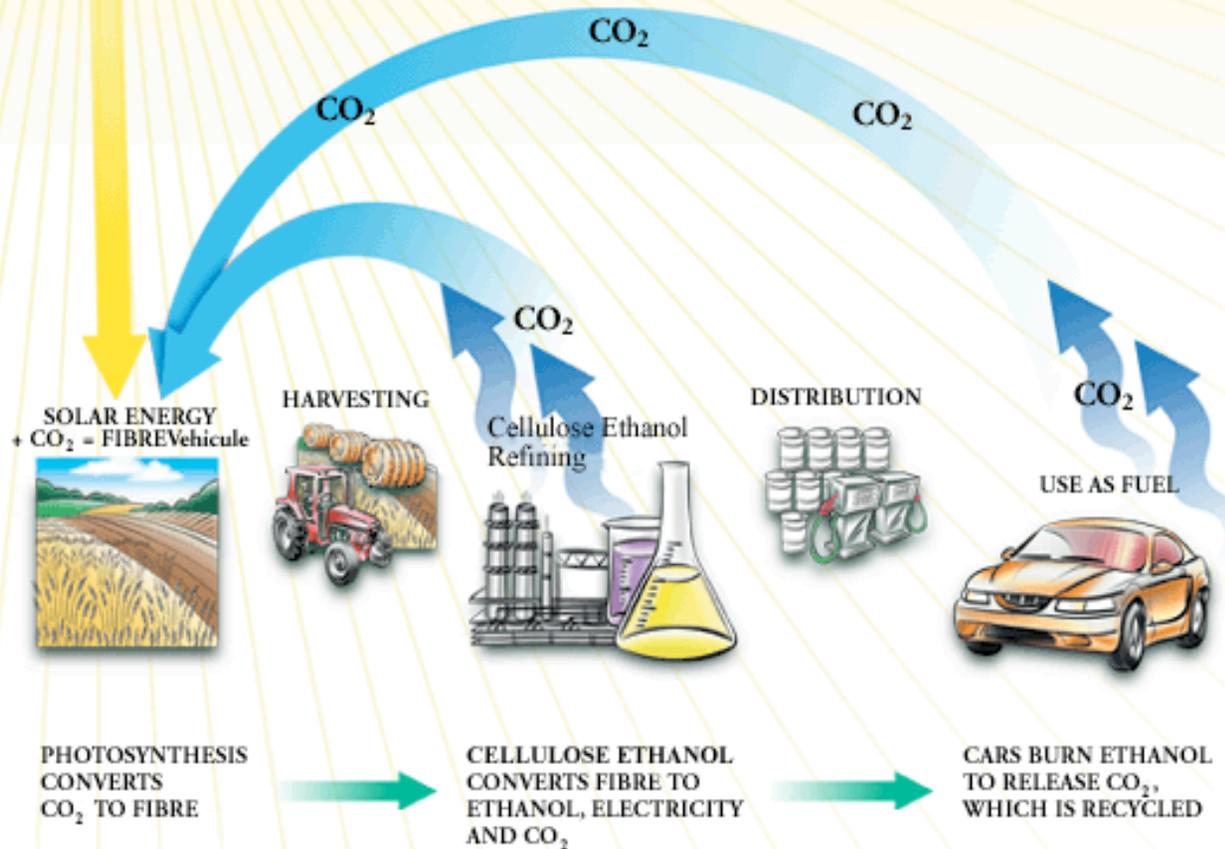
---

- Safe and abundant food supplies
  - Natural insecticides from *Photobacterium luminescens*
- Sustainable and benign energy sources
  - Cellulose to Butanol from *Clostridium acetobutylicum*
- Effective management of disease and aging
  - New antimicrobial drugs targets from *in silico* modeling
- Novel materials and renewable industrial feedstocks
  - Biological production of Hydrogen from biomass
- Advanced computational devices beyond silicon
  - Synthetic biological circuits
- Wide variety of molecular scale machinery
  - Proton powered rotary motors and ion pumps
- Self-assembly and self-reproduction technologies
  - Self perpetuating and environmentally friendly infrastructure

# The Clean Fuel Cycle

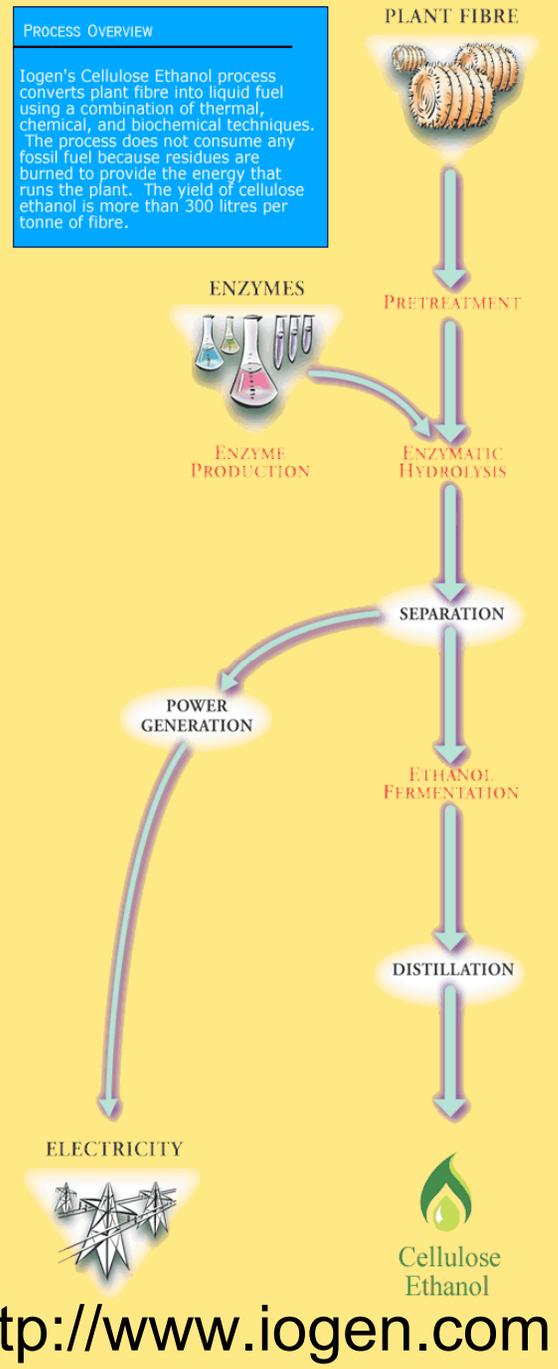
SUSTAINABLE ENERGY WITH NO GREENHOUSE EFFECT

Plants use the energy of the sun to grow. Plant fibre, called cellulose, is the most abundant organic molecule on earth. Iogen's EcoEthanol™ process takes cellulose and, using enzymes, turns it into fermentable sugars and subsequently into ethanol. Using CO<sub>2</sub> absorbing plant material as an ethanol feedstock offers environmental advantages unequalled by other feedstocks or fuels.



## PROCESS OVERVIEW

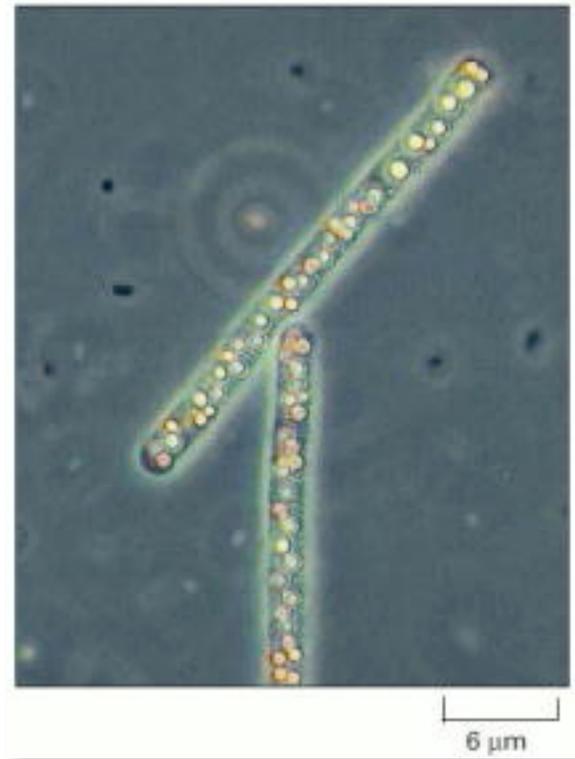
Iogen's Cellulose Ethanol process converts plant fibre into liquid fuel using a combination of thermal, chemical, and biochemical techniques. The process does not consume any fossil fuel because residues are burned to provide the energy that runs the plant. The yield of cellulose ethanol is more than 300 litres per tonne of fibre.



<http://www.iogen.com>

# Microbial Organisms are Important to Study

- Extremely Diverse Metabolisms
- Window on Biodiversity
- Ancient Origins
- Foundation of the Biosphere
- Agents of Symbiogenesis
- Infectious Disease
- Human Microbiota and Metagenomes
- Complex Community Structures
- Industrial and Agricultural Applications
- Biotechnology Applications
- Biofuels and Alternative Feed stocks
- Inexpensive and Experimentally Tractable

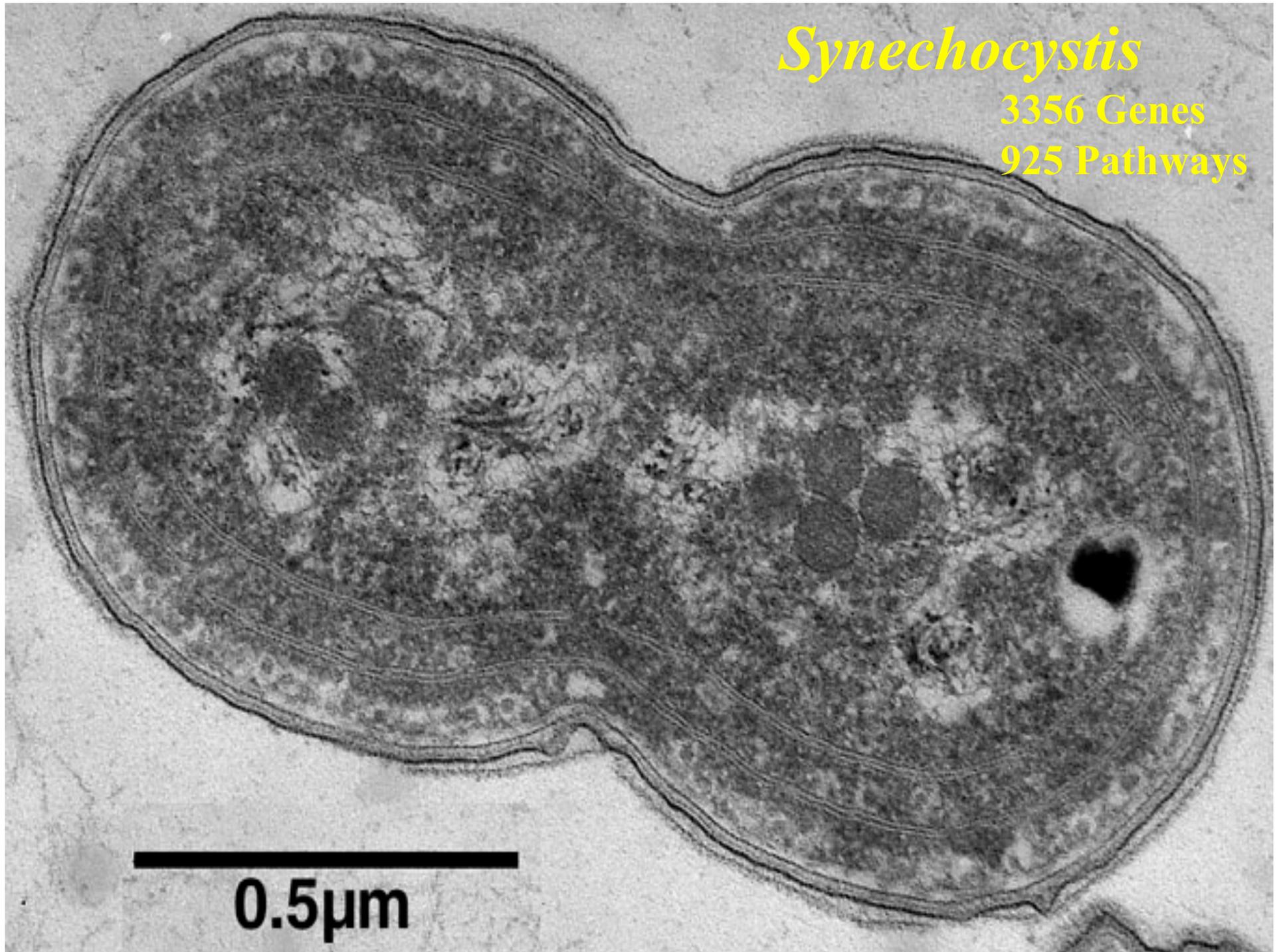


*Beggiatoa*, which lives in sulfurous environments, gets its energy by oxidizing  $H_2S$  and can fix carbon even in the dark. Note the yellow deposits of sulfur inside the cells. (Courtesy of Ralph W. Wolfe.)

*Synechocystis*

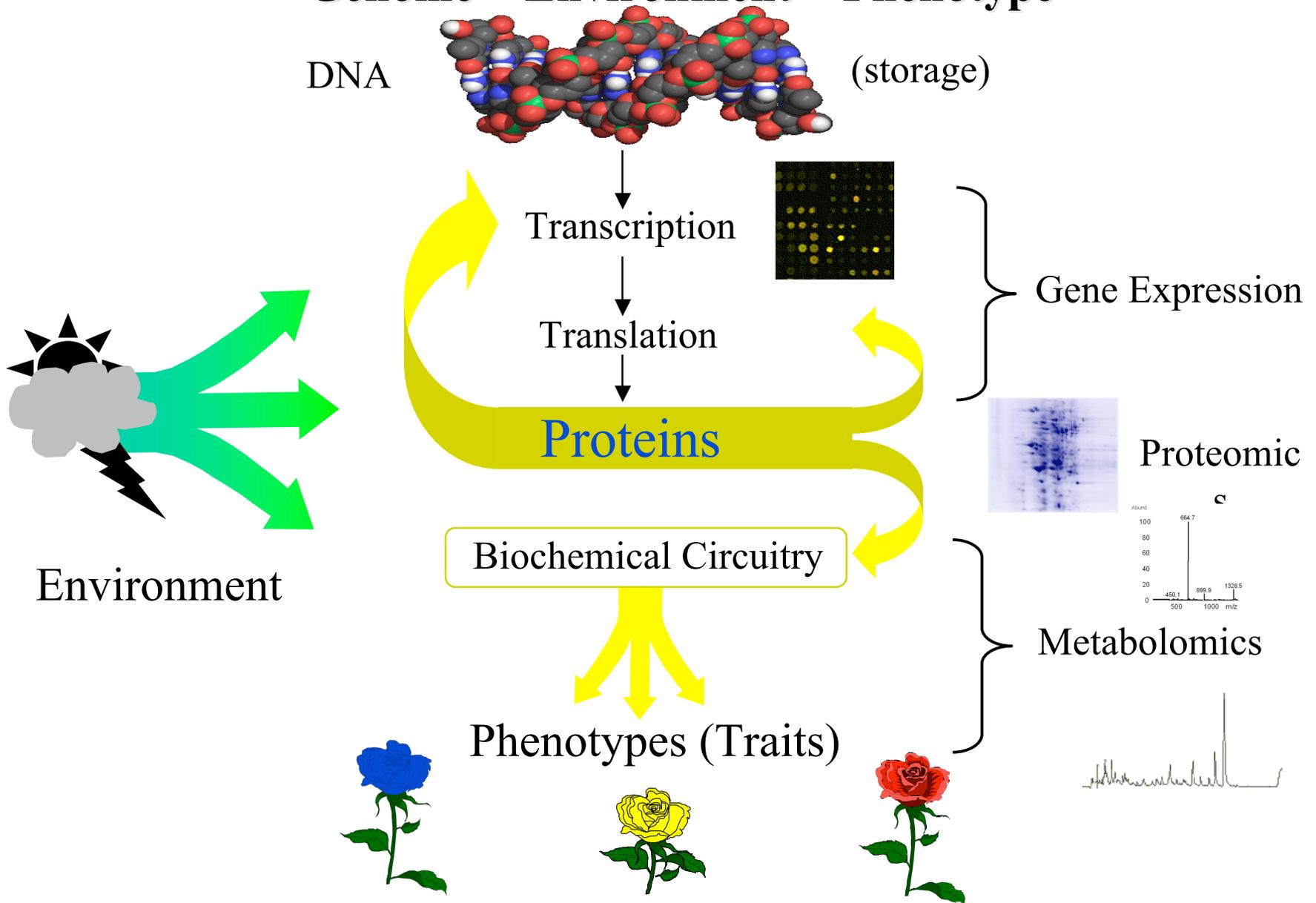
3356 Genes

925 Pathways



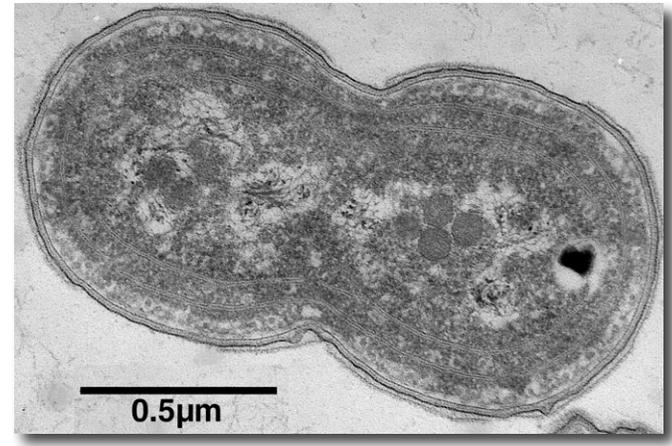
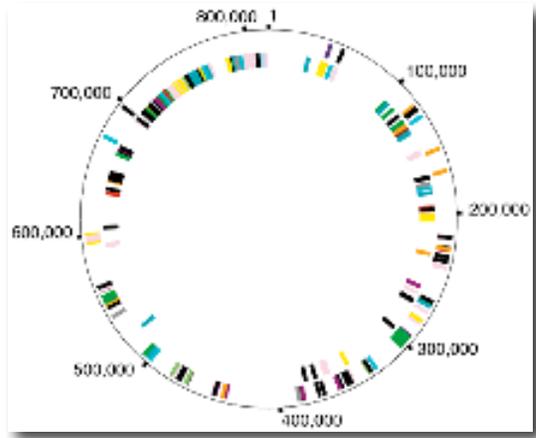
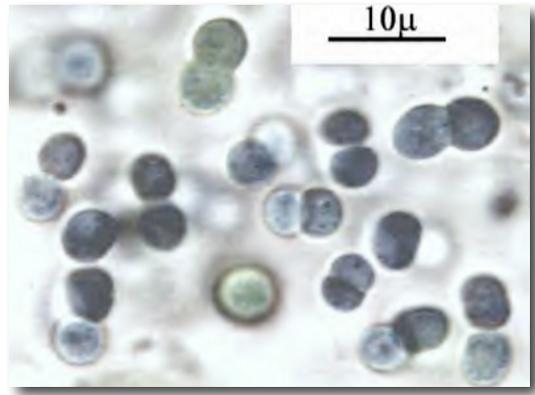
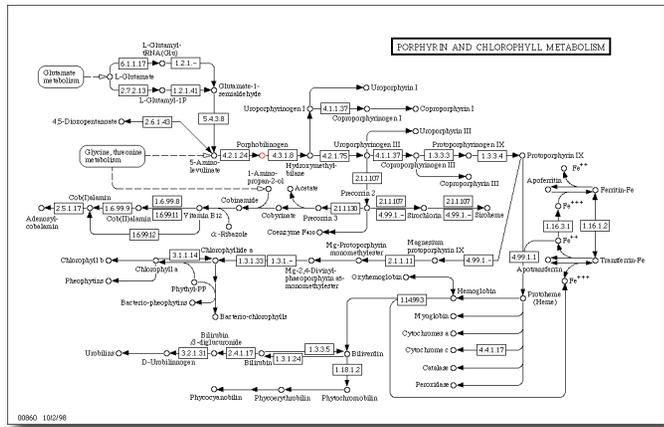
0.5 $\mu$ m

# Reverse Engineering Living Systems: Genome + Environment = Phenotype

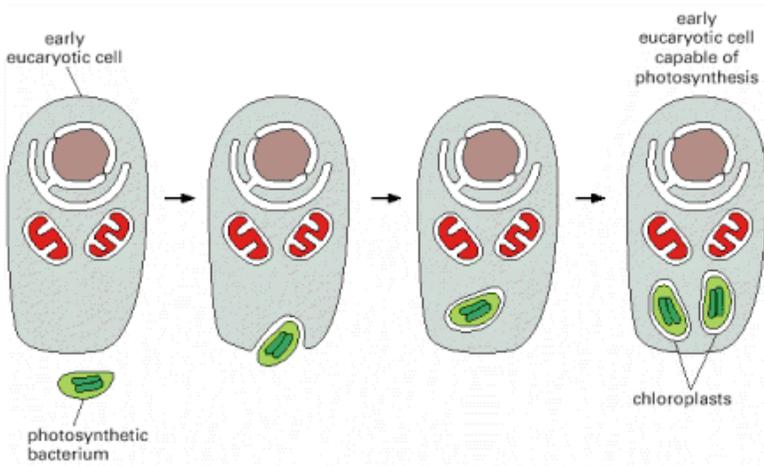
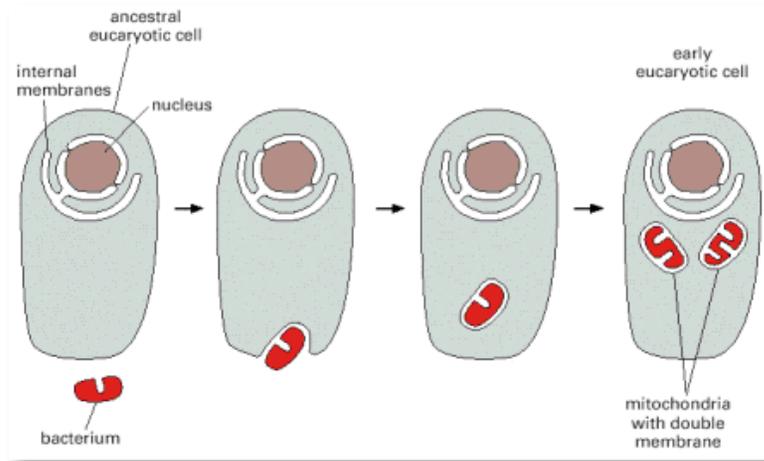


Adapted From Bruno Sobral VBI

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

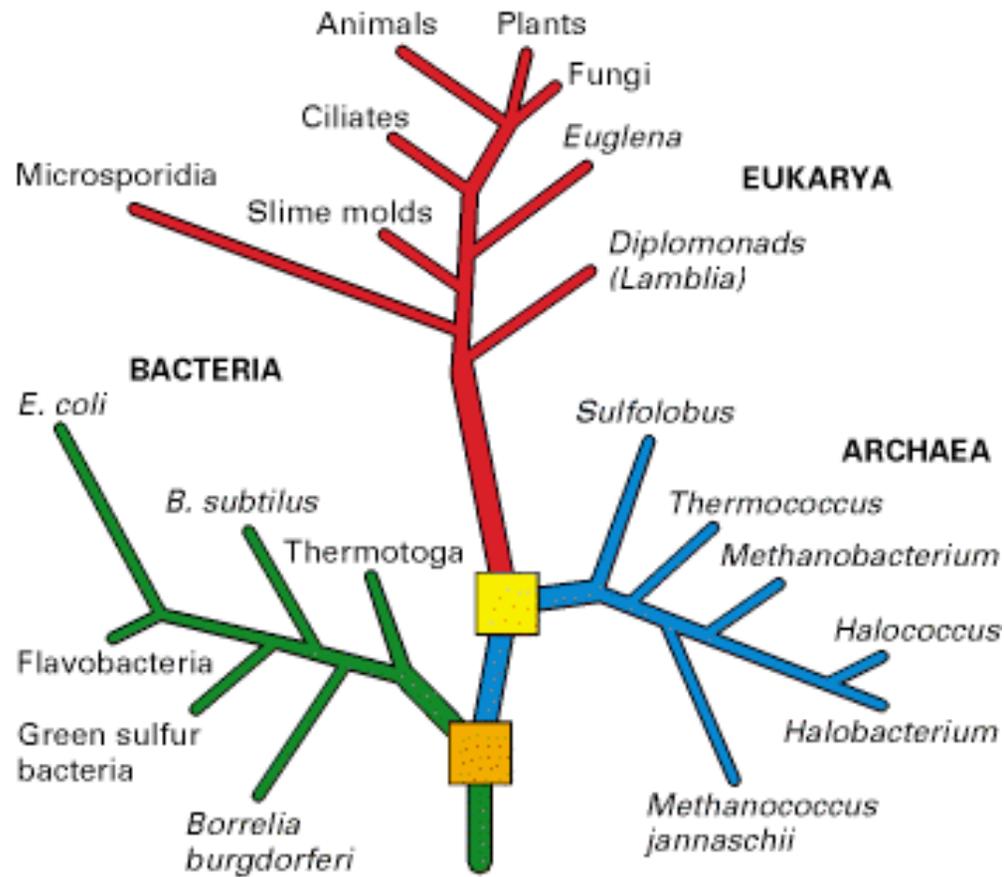


# Deep Motivating Problems

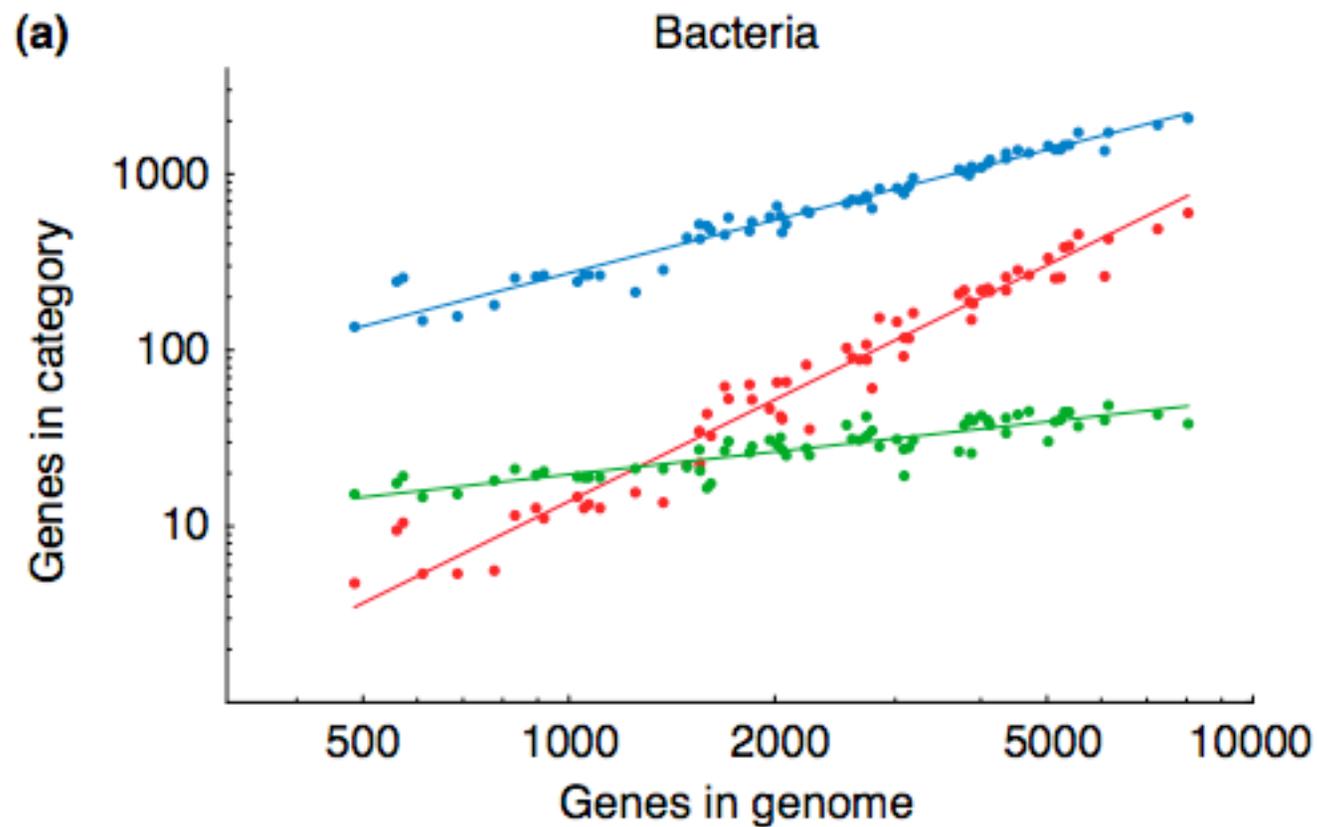


- Understanding the Evolution of Cellular Functions and the Role of Symbiogenesis
- Developing the Concept of Minimum Organisms as a Platform for Bioengineering
- Resolving Details of the Last Universal Common Ancestor(s)
- Designing Hypothetical RNA World Organisms
- Developing Quantitative Understanding of “Possible Biologies”

# Looking for LUCA



# Genome Size v. Protein Family (Function)



**Table 1. Estimates for the exponents of a selection of functional categories<sup>a</sup>**

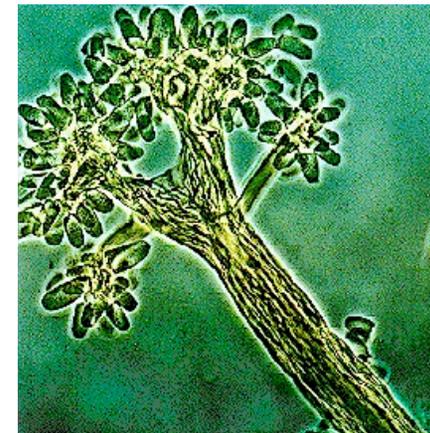
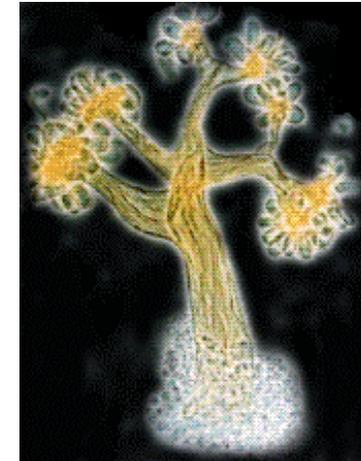
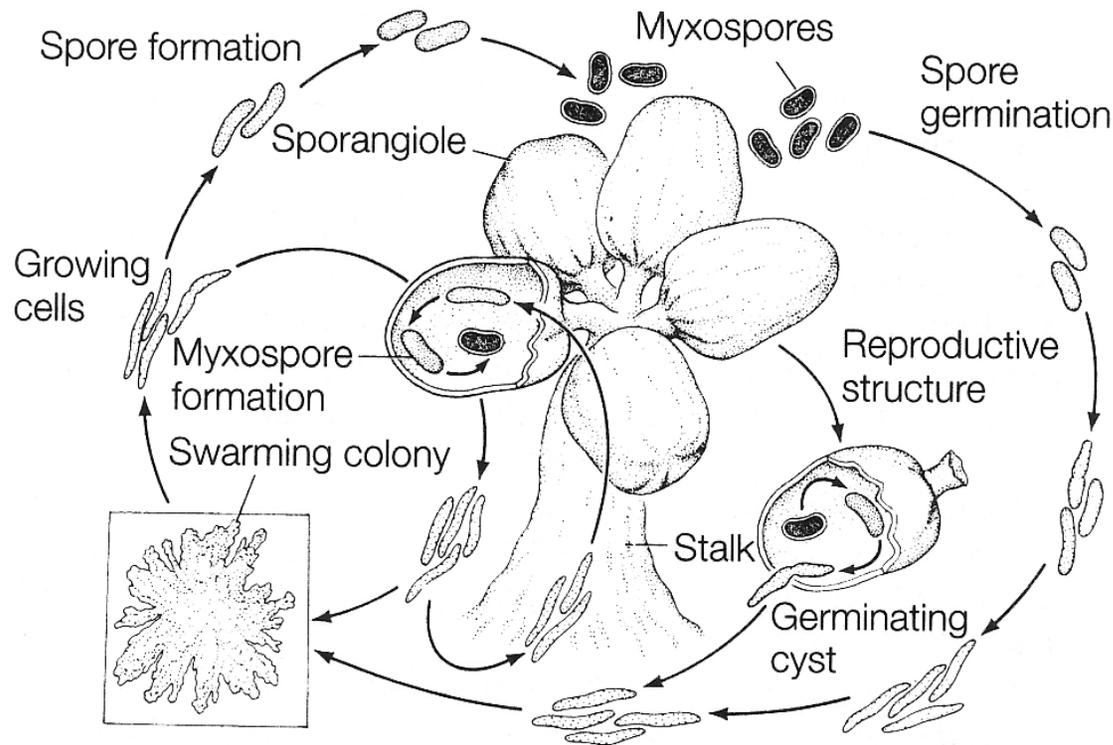
Category	Bacteria	Eukaryotes
Transcription regulation	1.87 ± 0.13	1.26 ± 0.10
Metabolism	1.01 ± 0.06	1.01 ± 0.08
Cell cycle	0.47 ± 0.08	0.79 ± 0.16
Signal transduction	1.72 ± 0.18	1.48 ± 0.39
DNA repair	0.64 ± 0.08	0.83 ± 0.31
DNA replication	0.43 ± 0.08	0.72 ± 0.23
Protein biosynthesis	0.13 ± 0.02	0.41 ± 0.15
Protein degradation	0.97 ± 0.09	0.90 ± 0.11
Ion transport	1.42 ± 0.28	1.43 ± 0.20
Catabolism	0.88 ± 0.07	0.92 ± 0.08
Carbohydrate metabolism	1.01 ± 0.11	1.36 ± 0.36
Two-component systems	2.07 ± 0.21	NA <sup>b</sup>
Cell communication	1.81 ± 0.19	1.58 ± 0.34
Defense response	NA <sup>b</sup>	3.35 ± 1.41

<sup>a</sup>The first number gives the maximum likelihood estimate of the exponent and the second number indicates the boundaries of the 99% posterior probability interval.

<sup>b</sup>NA, not analyzed.



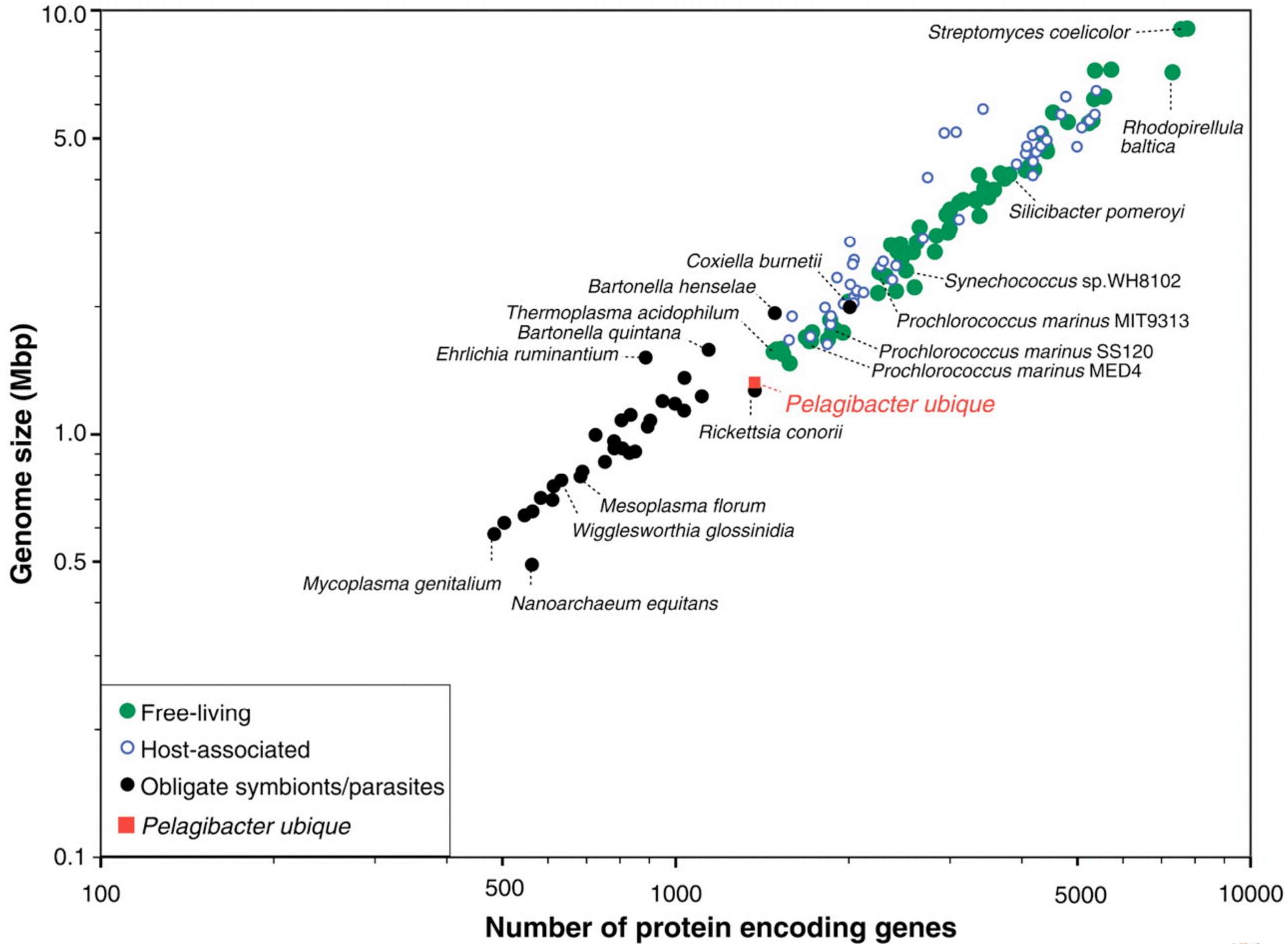
# Complex Bacterial Life Cycles

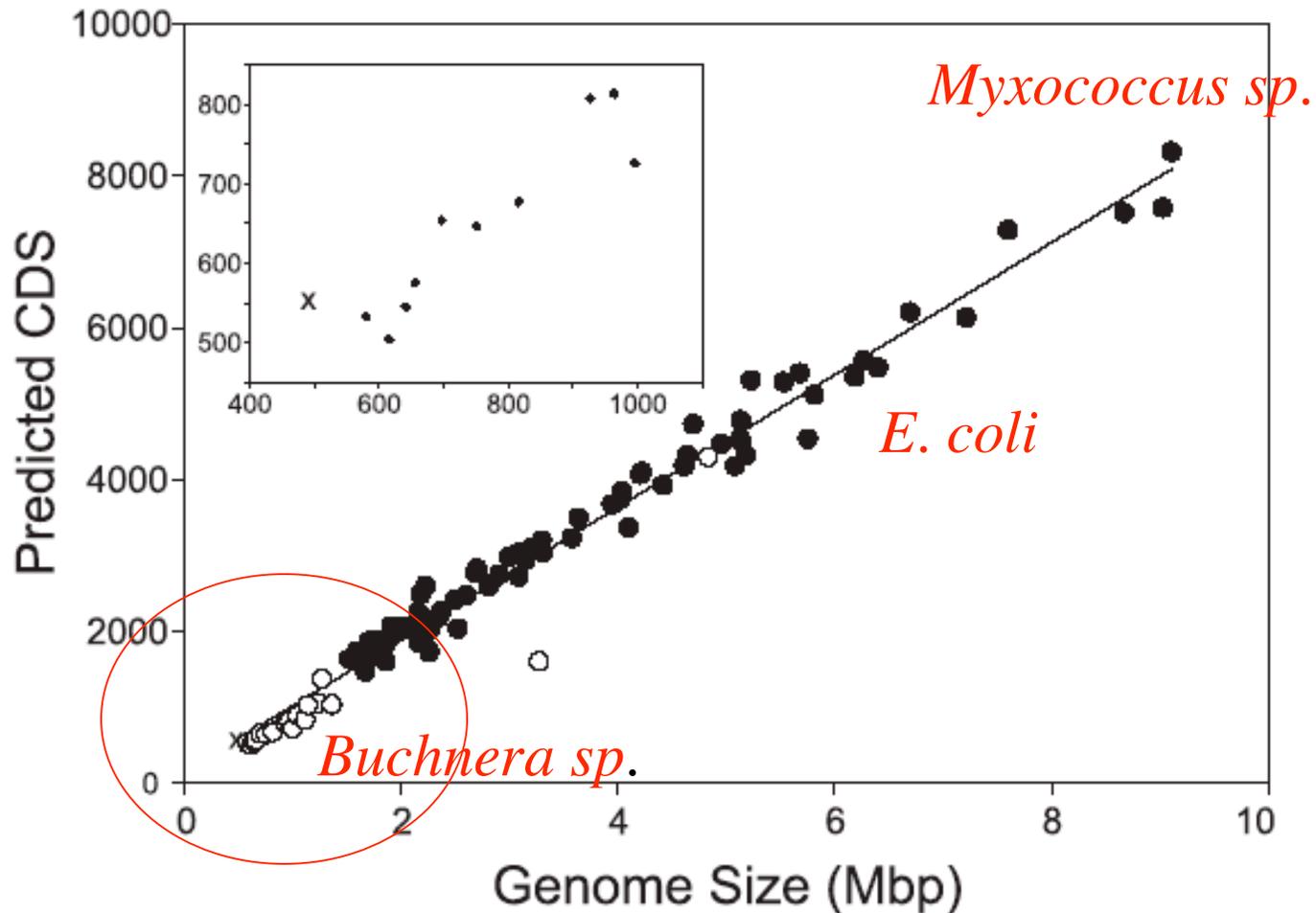


**F** Life cycle of *Stigmatella aurantiaca*. [Drawing by L. Meszoly; labeled by M. Dworkin.]

From Lynn Margulis and Karlene Schwartz



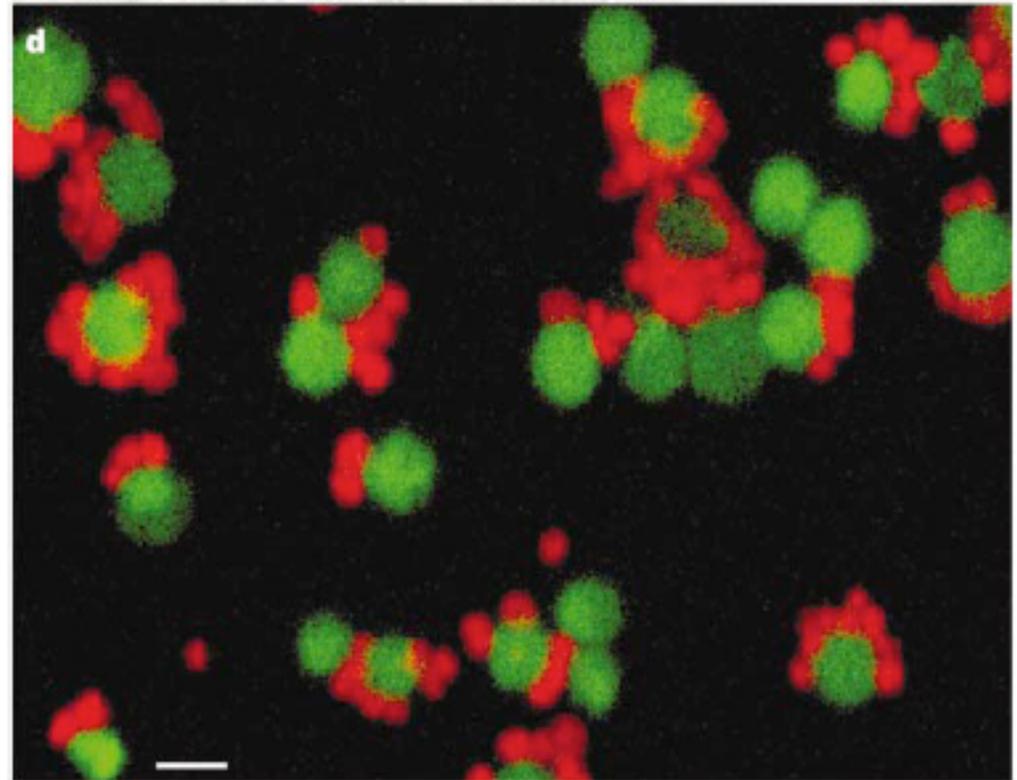
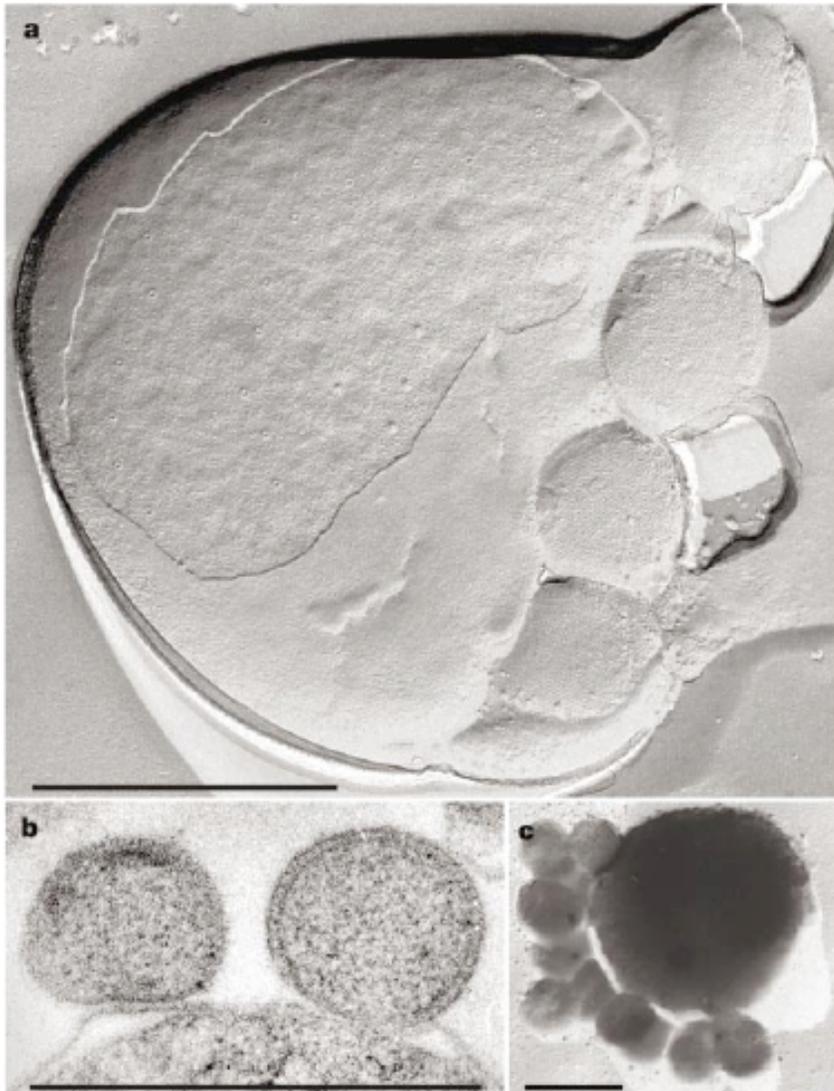




How small  
can a  
functional  
microbial  
genome be ?

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.

# 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.

# 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

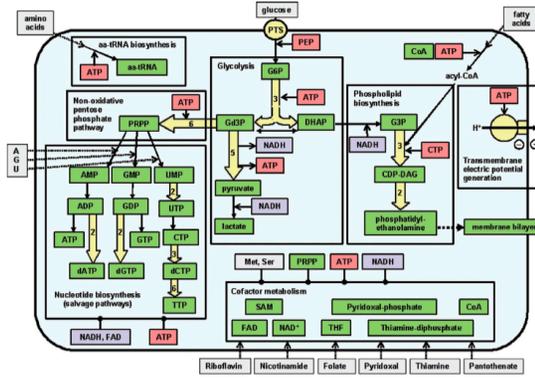


FIG. 1. A minimal metabolism. The minimal cell can obtain its more basic components from the environment: glucose, fatty acids, amino acids, adenine, guanine, uracil, and coenzyme precursors (nicotinamide, riboflavin, biotin, pantoic acid, and pyridoxal). Each box includes the metabolic transformations classified in major groups of pathways: glycolysis, phospholipid biosynthesis, nonoxidative pentose-phosphate pathway, nucleotide biosynthesis, synthesis of enzymatic cofactors. Lines represent incorporation from the environment several enzymatic steps (the number within the boxes) for some of the transformations in Metabolites acting as a source of chemical energy are indicated by a red box. Metabolic precursors of external or

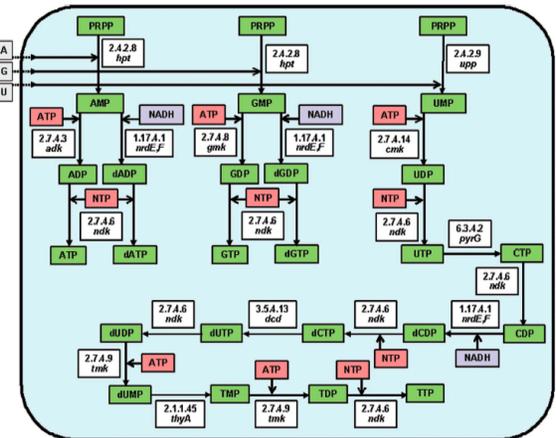


FIG. 2. A minimal nucleotide metabolism based on salvage pathways. Activated ribonucleotides and deoxyribonucleotides are obtained from free bases (A, G, and U), PRPP, ATP-dependent phosphorylating reactions, and NADH-dependent reduction. White boxes indicate the individual enzymatic activity (EC number and coding gene). Other colors are used as in Fig. 1.

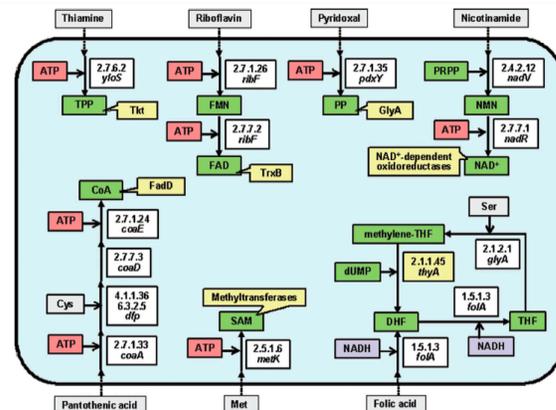
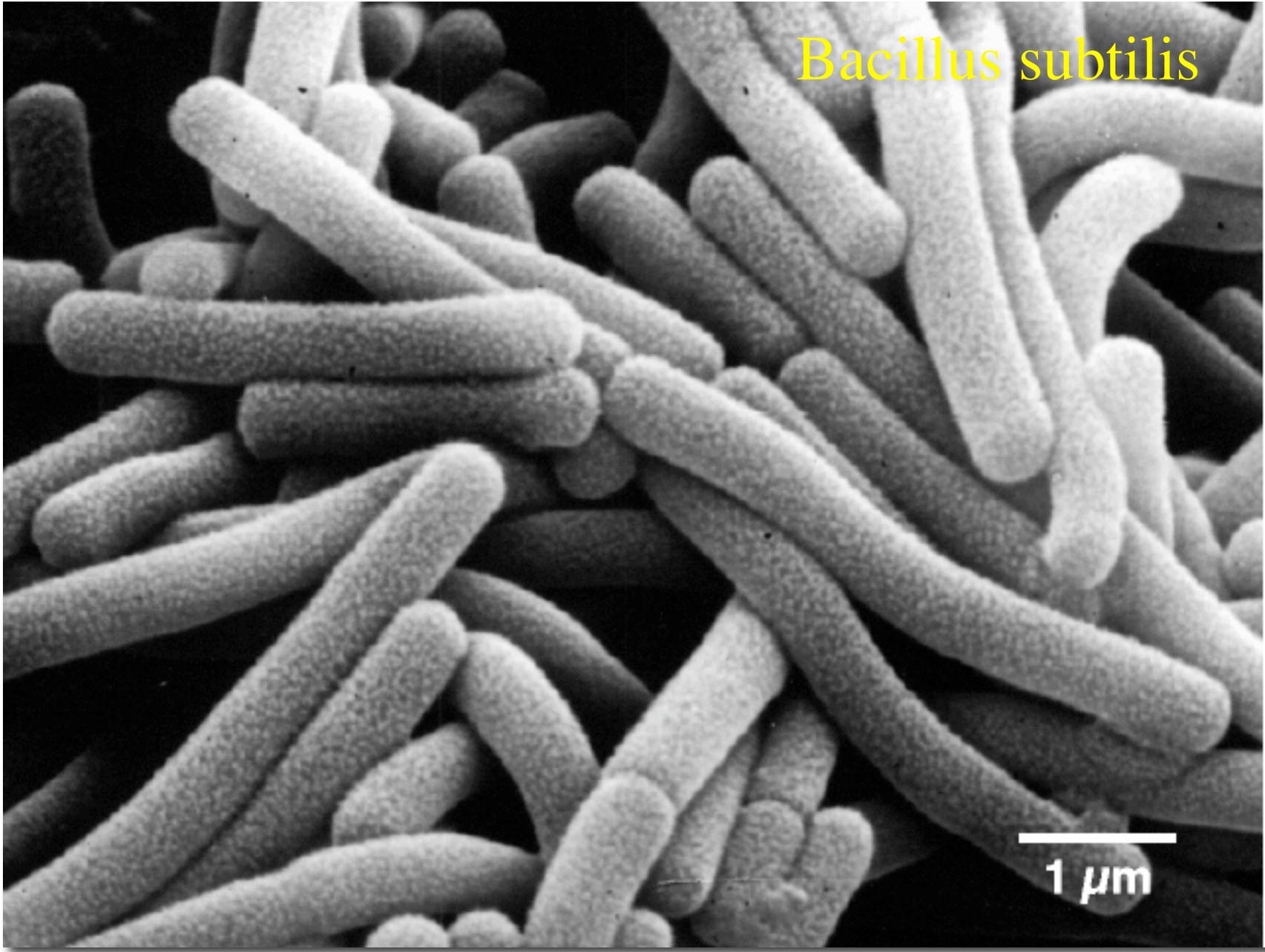
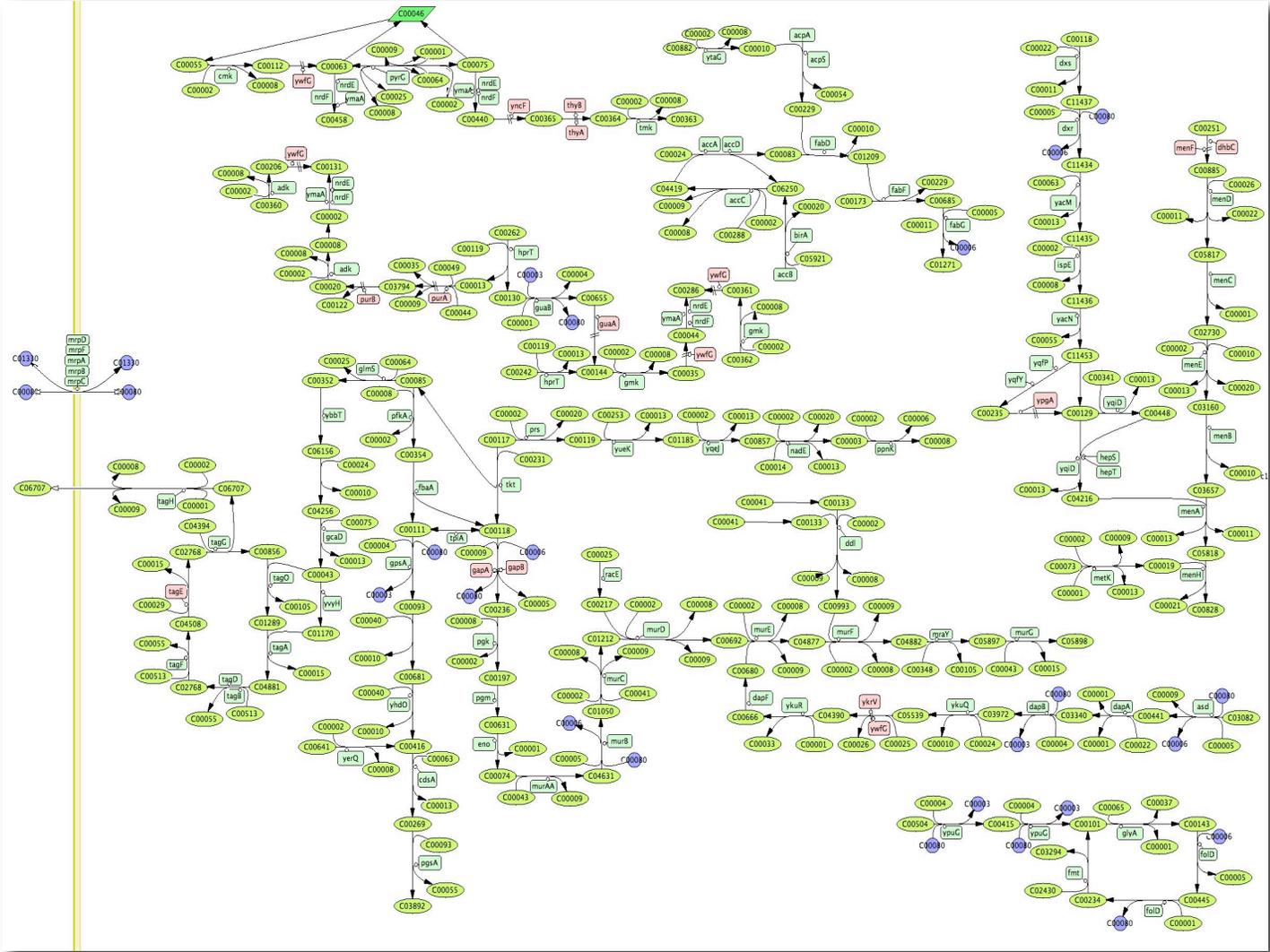


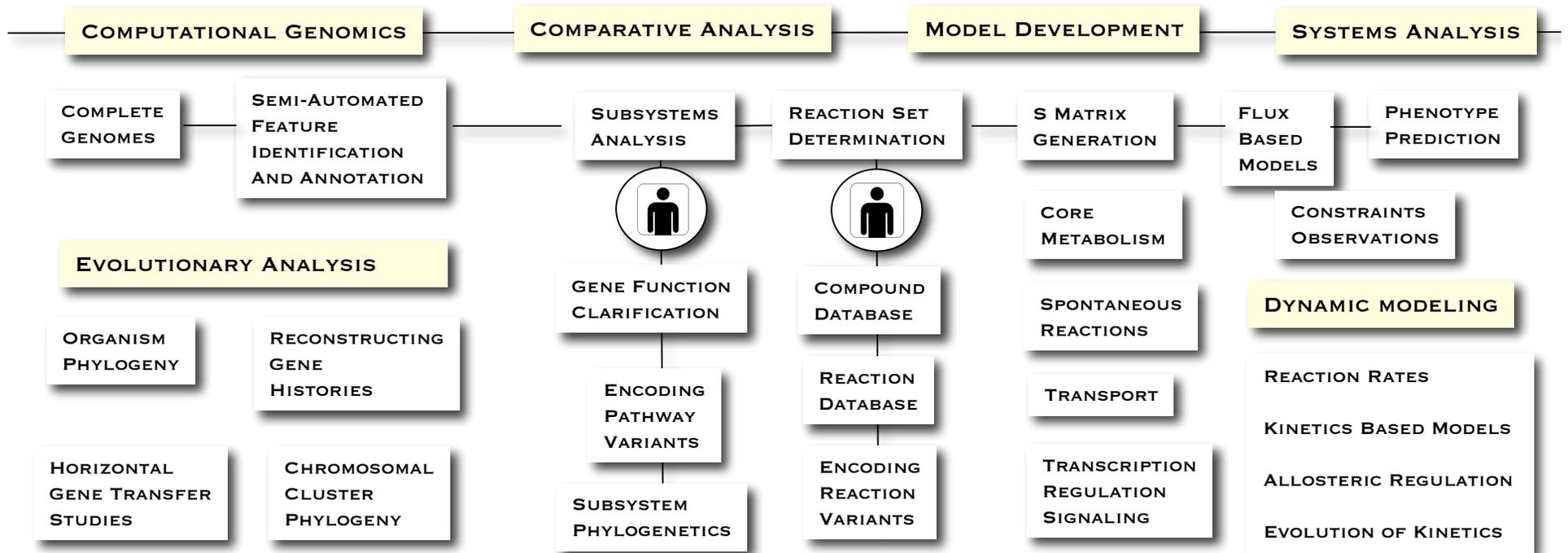
FIG. 3. Biosynthesis of cofactors. A metabolism of essential cofactors used by the minimal cell starting with precursors (i.e., vitamins) nicotinamide, riboflavin, thiamine, pyridoxal, pantoic acid, methionine, and folic acid. Yellow boxes indicate the enzymes that need each cofactor for their correct function. Other colors and symbols are as in Fig. 1 and 2. PP, pyridoxal-phosphate.

*Bacillus subtilis*



# Model of *Bacillus subtilis* Essential Core





### SEARCHING FOR CANDIDATES OF HORIZONTAL GENE TRANSFER

$G$  IS A GENE/PROTEIN INDEX (THINK COHERENT PROTEIN FAMILY OR COG)  
 $O$  IS SET OF ORGANISMS (WE HAVE ABOUT 300 NOW - WILL HAVE 1,000 IN THREE YEARS)  
 $X$  IS A SET OF CONSERVED GENES/PROTEINS COMMON TO MANY  $O$  (ORDER HUNDREDS)

$dnaSeq[G]$  ← EXTRACT CODING DNA SEQUENCES FOR  $X$  FROM SOME  $O$   
 $proteinSeq[G]$  ← TRANSLATE DNA TO PROTEIN FOR EACH  $dnaSeq[G]$   
 $Bbhs[G]$  ← COMPUTE BIDIRECTIONAL BEST HITS FOR EACH  $proteinSeq[G]$   
 $TC[G]$  ← COMPUTE TRANSITIVE CLOSURE OF  $Bbhs[G]$   
 $MSA[G]$  ← COMPUTE MULTIPLE SEQUENCE ALIGNMENT OF EACH  $TC[G]$   
 $GeneTree[G]$  ← COMPUTE MAXIMUM LIKELIHOOD TREE FOR EACH  $MSA[G]$   
 $OrgTree$  ← COMPUTE ORGANISM TREE\* FOR  $O$   
 $ReconTrees[G]$  ← RECONCILE  $OrgTree$  WITH EACH  $GeneTree[G]$   
 $Candidates[G]$  ← SCORE AND SORT  $ReconTrees[G]$   
 $TreeGraphs[G]$  ← COMPUTE READABLE RECONCILED TREES FOR  $Candidates[G]$

Each functional area admits a complex workflow, similar to the one on the left

# Large-Scale Computing and the Hunt for Horizontal Gene Transfer



- 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

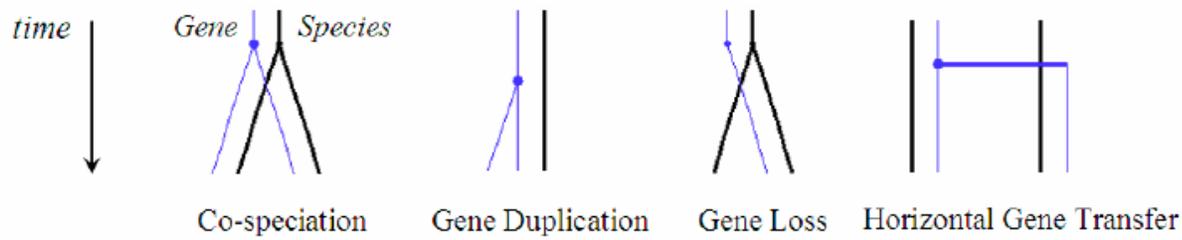


Figure 1. Four basic types of evolutionary events

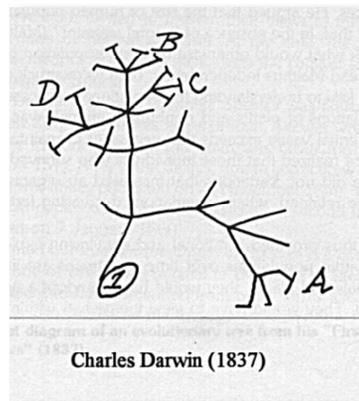
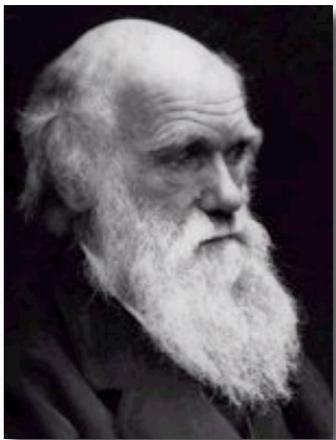
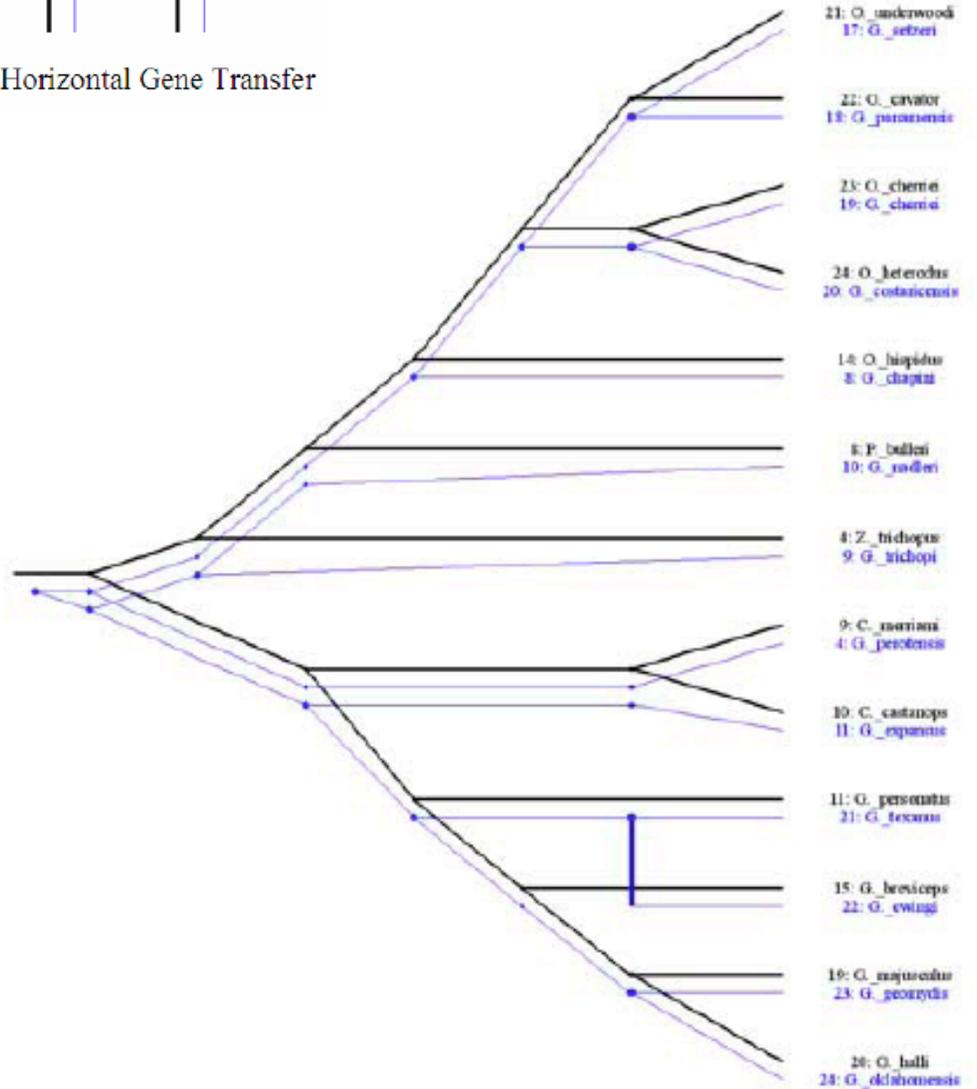
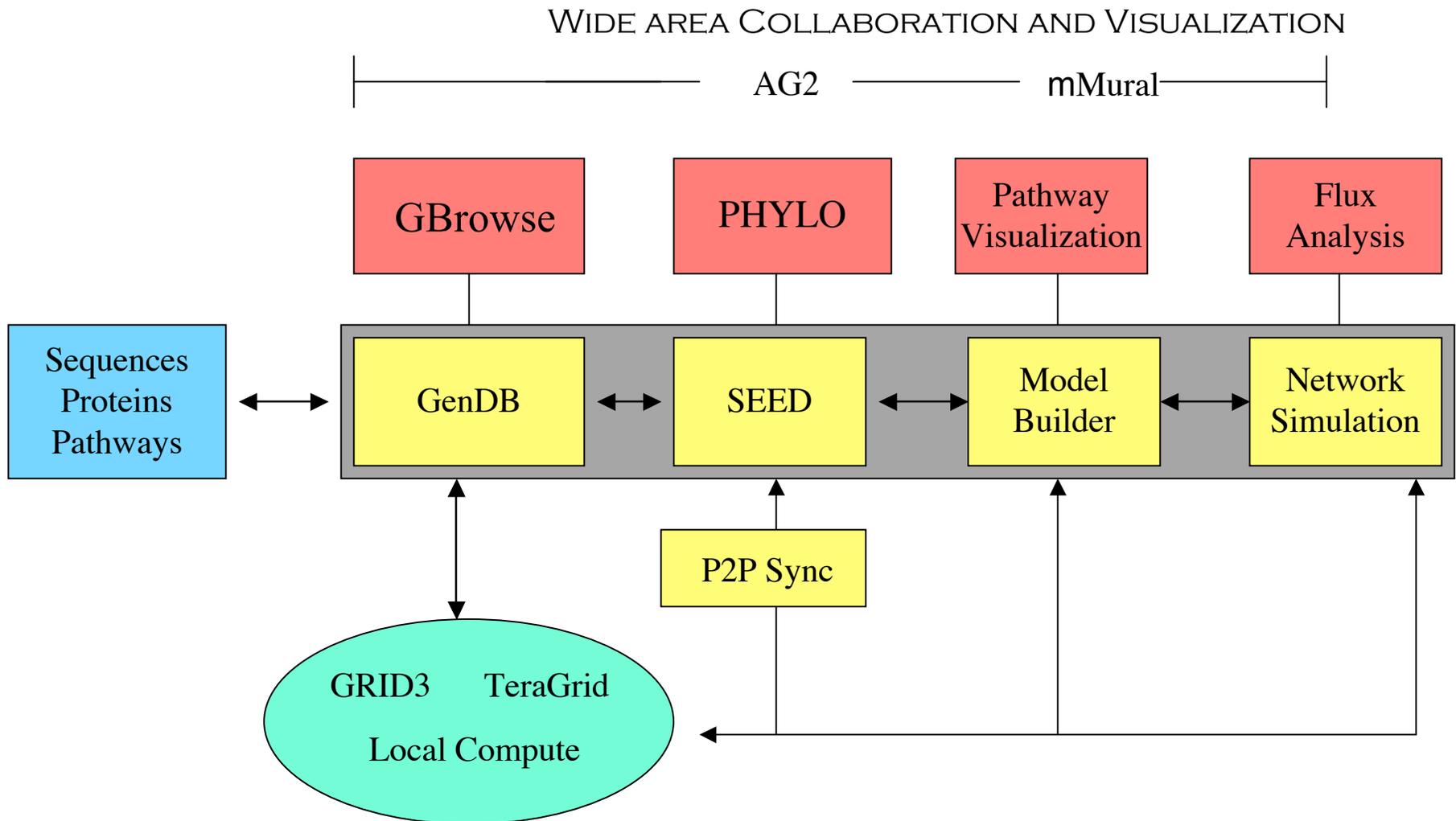


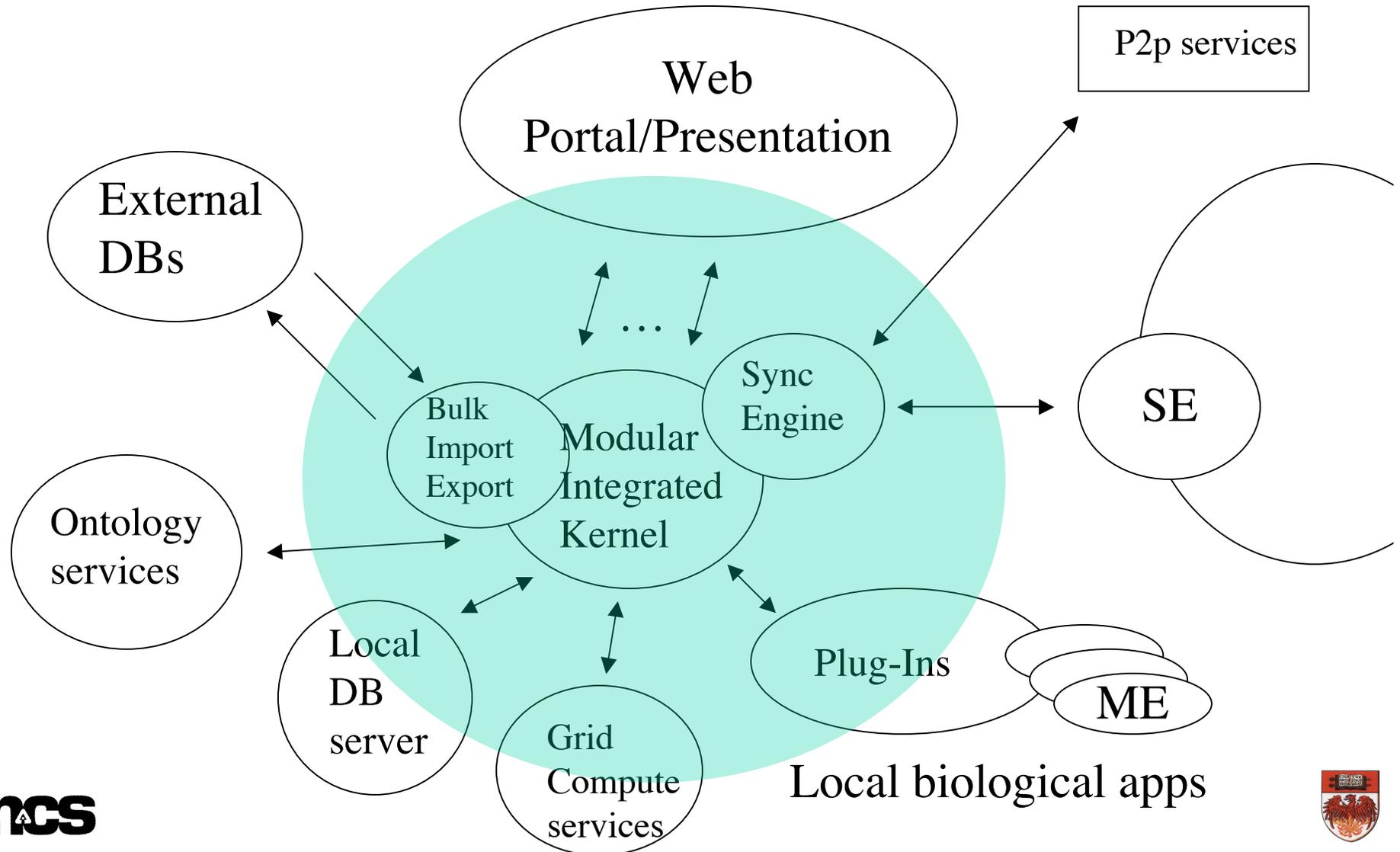
Figure 4. Visualized reconciled trees: HKY85 data

# The Chicago Systems Biology Infrastructure



# Peer-to-Peer Bioinformatics

## *the prototype SEED*



# Subsystems and The SEED

## Subsystem: NAD and NADP cofactor biosynthesis global

Functional Roles

Column	Abbrev	Functional Role
1	TDO	Tryptophan 2,3-dioxygenase (EC 1.13.11.11)
2	IDO	Indoleamine 2,3-dioxygenase (EC 1.13.11.42)
3	KFA_e	Kynurenine formamidase (EC 3.5.1.9)
4	KFA_b	Kynurenine formamidase, bacterial (EC 3.5.1.9)
5	KMO	Kynurenine 3-monooxygenase (EC 1.14.13.9)
6	KYN	Kynureninase (EC 3.7.1.3)
7	HAD	3-hydroxyanthranilate 3,4-dioxygenase (EC 1.13.11.6)
8	ASPOX	L-aspartate oxidase (EC 1.4.3.16)
9	ASPDH	Aspartate dehydrogenase [same functional role as] (EC 1.4.3.16)
10	QSYN	Quinolinate synthetase (EC 4.1.99.-)
11	QAPRT	Quinolinate phosphoribosyltransferase [decarboxylating] (EC 2.4.2.19)
12	NAMNAT	Nicotinate-nucleotide adenylyltransferase (EC 2.7.7.18)
13	NMNAT	Nicotinamide-nucleotide adenylyltransferase (EC 2.7.7.1)
14	NADS	NAD synthetase (EC 6.3.1.5)
15	GAT	Glutamine amidotransferase chain of NAD synthetase
16	NADK	NAD kinase (EC 2.7.1.23)

Subsets of Roles

Subset	Includes These Roles
*ASPOX	8,9
*KFA	3,4
*PNAT	12,13
*RNK	21,22
*TDO	1,2
Bacterial_type	8,9,10,11,12,13,14,15,16,17,18,19,20,21
Eukaryotic_type	1,2,3,4,5,6,7,11,12,13,14,15,16,17,18,19,22

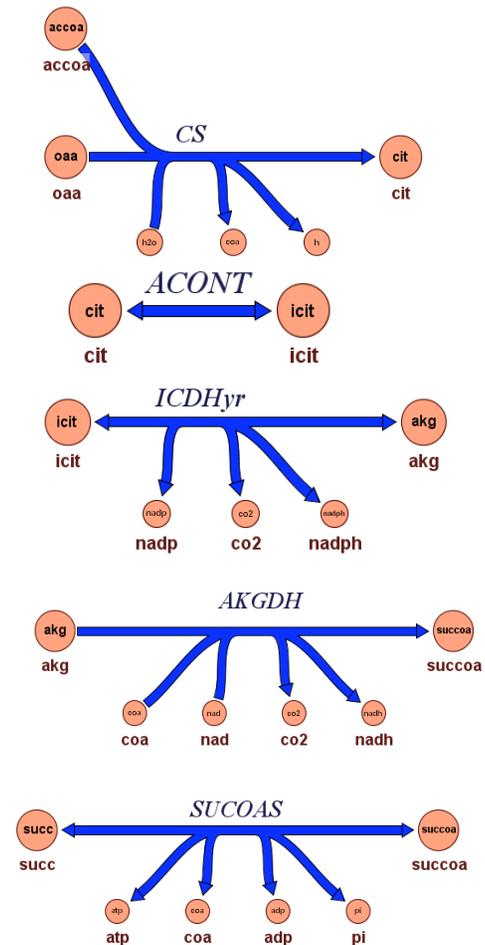
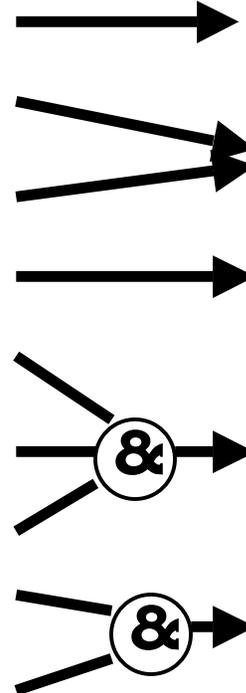
Genome ID	Organism	Variant Code	*TDO	*KFA	KMO	KYN	HAD	QAPRT	*PNAT	NADS	GAT	NADK	NAM	NAPRT	NMPRT	*RNK
28450.1	Burkholderia pseudomallei K96243 [B]	1	<a href="#">5488-1</a>	<a href="#">5486-4</a>		<a href="#">5487</a>		<a href="#">5411</a>	<a href="#">3781-12</a>	<a href="#">3924-6430</a>	<a href="#">3924</a>	<a href="#">3300</a>	<a href="#">1439</a>	<a href="#">4058-5625</a>		
227377.1	Coxiella burnetii RSA 493 [B]	1						<a href="#">94</a>	<a href="#">531-12</a>	<a href="#">821</a>	<a href="#">821</a>	<a href="#">1232</a>		<a href="#">983</a>		
165597.1	Crocospaera watsonii WH 8501 [B]	4						<a href="#">4296</a>	<a href="#">600-12</a>			<a href="#">2279-6136</a>	<a href="#">5122</a>	<a href="#">601</a>		
985.1	Cytophaga hutchinsonii [B]	8	<a href="#">2355-1</a>		<a href="#">839</a>	<a href="#">840</a>	<a href="#">841</a>		<a href="#">262-12</a>			<a href="#">867</a>			<a href="#">1073</a>	<a href="#">2710-22</a>
211586.1	Shewanella oneidensis MR-1 [B]	4				<a href="#">4007</a>			<a href="#">1087-12</a>	<a href="#">1842</a>		<a href="#">1411</a>			<a href="#">1809</a>	

# Functional Roles and Reaction Sets

Reactions in the context of the pathway should be associated with each functional gene annotation

## FUNCTIONAL ROLES

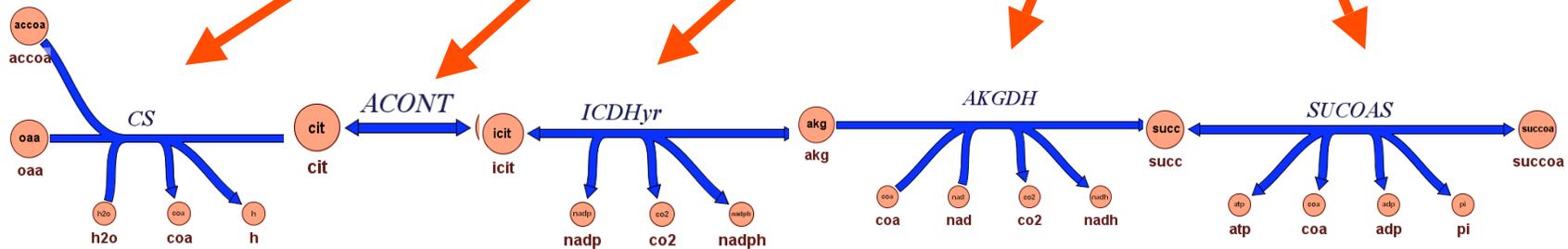
Col	Abr	Functional Role
1	gltA	Citrate synthase (EC 2.3.3.1)
2	acnB	Aconitate hydratase 2 (EC 4.2.1.3)
3	acnA	Aconitate hydratase (EC 4.2.1.3)
4	icd	Isocitrate dehydrogenase (EC 1.1.1.42)
5	sucA	2-oxoglutarate dehydrogenase E1 component (EC 1.2.4.2)
6	sucB	2-oxoglutarate dehydrogenase E2 component (EC 2.3.1.61)
7	lpdA	Dihydrolipoamide dehydrogenase (EC 1.8.1.4)
8	sucD	Succinyl-CoA synthetase alpha chain (EC 6.2.1.5)
9	sucC	Succinyl-CoA synthetase beta chain (EC 6.2.1.5)



# Subsystems to Reaction Networks

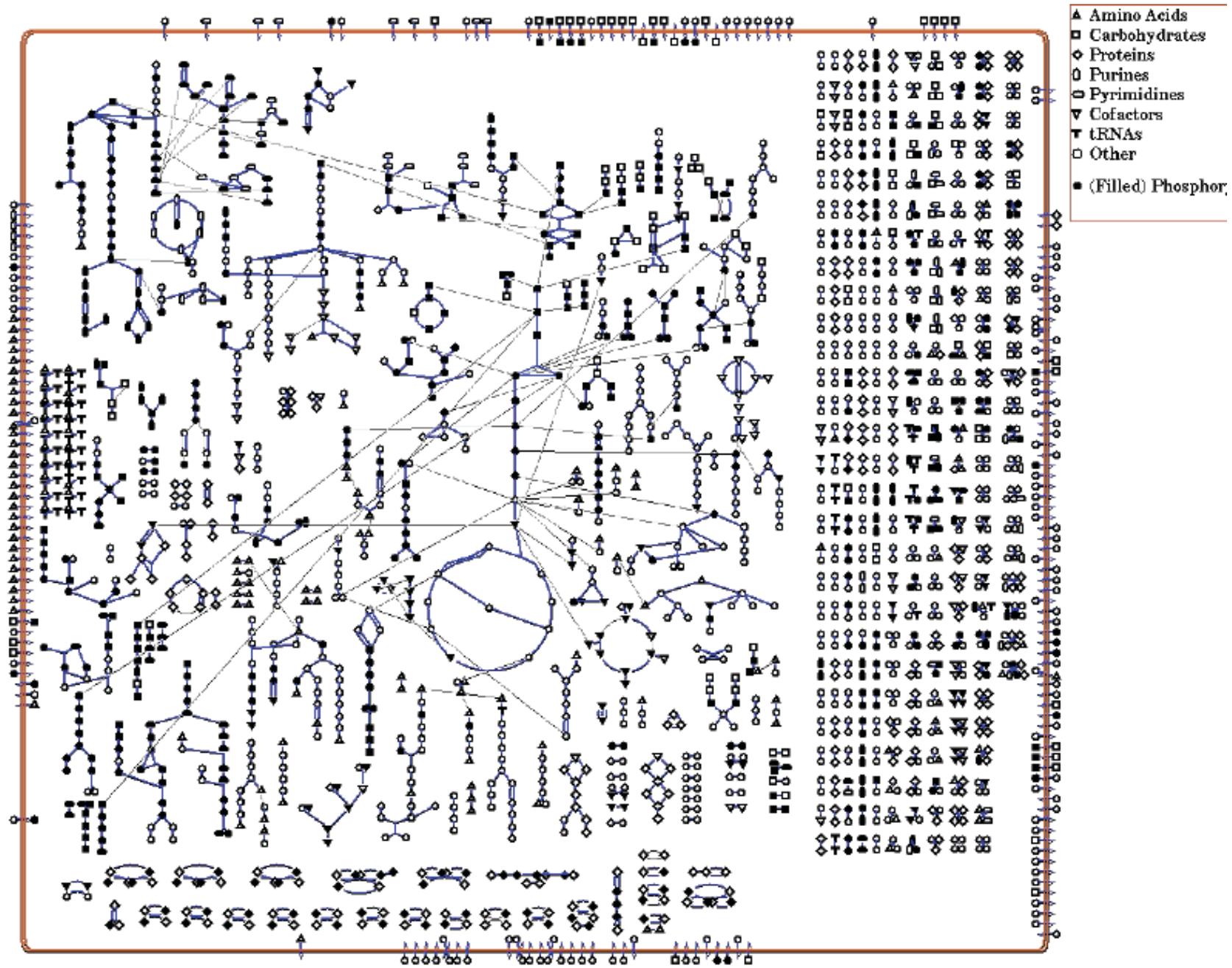
The reactions that are associated with an organism are combined to form a network based on the subsystems

Genome ID	Organism	Variant Code	*TDO	*KFA	KMO	KYN	HAD	QAPRT	*PNAT	NADS	GAT	NADK	NAM	NAPRT	NMPRT	*RNK
28450.1	Burkholderia pseudomallei K96243 [B]	1	<a href="#">5488-1</a>	<a href="#">5486-4</a>		<a href="#">5487</a>		<a href="#">5411</a>	<a href="#">3781-12</a>	<a href="#">3924-6430</a>	<a href="#">3924</a>	<a href="#">3300</a>	<a href="#">1439</a>	<a href="#">4058-5625</a>		
227377.1	Coxiella burnetii RSA 493 [B]	1						<a href="#">94</a>	<a href="#">531-12</a>	<a href="#">821</a>	<a href="#">821</a>	<a href="#">1232</a>		<a href="#">983</a>		
165597.1	Crocospaera watsonii WH 8501 [B]	4						<a href="#">4296</a>	<a href="#">600-12</a>			<a href="#">2279-6136</a>	<a href="#">5122</a>	<a href="#">601</a>		
985.1	Cytophaga hutchinsonii [B]	8	<a href="#">2355-1</a>		<a href="#">839</a>	<a href="#">840</a>	<a href="#">841</a>		<a href="#">262-12</a>			<a href="#">867</a>			<a href="#">1073</a>	<a href="#">2710-22</a>
211586.1	Shewanella oneidensis MR-1 [B]	4				<a href="#">4007</a>			<a href="#">1087-12</a>	<a href="#">1042</a>		<a href="#">1411</a>			<a href="#">1809</a>	



# ~200 SUBSYSTEMS CURRENTLY UNDER DEVELOPMENT

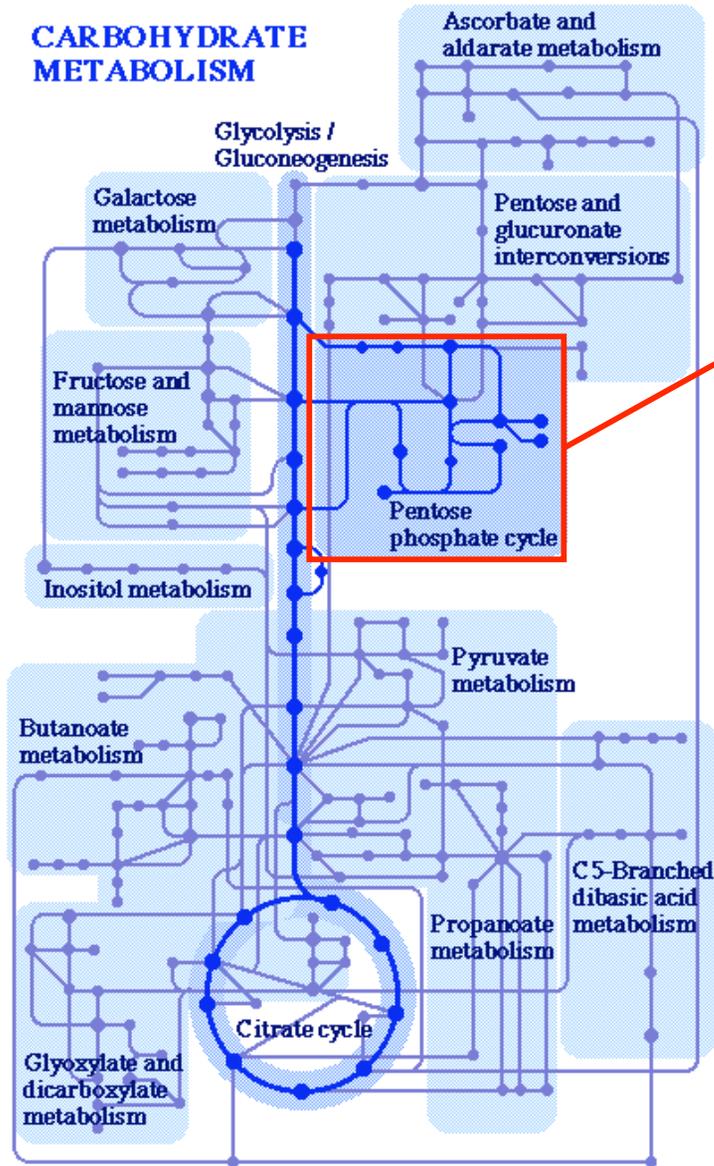
ABC TRANSPORTER ALKYLPHOSPHONATE (TC 3.A.1.9.1)  
ABC TRANSPORTER ARABINOSE (TC 3.A.1.2.2)  
ABC TRANSPORTER BRANCHED-CHAIN AMINO ACID (TC 3.A.1.4.1)  
ABC TRANSPORTER DIPEPTIDE (TC 3.A.1.5.2)  
ABC TRANSPORTER FERRIC ENTEROBACTIN (TC 3.A.1.14.2)  
ABC TRANSPORTER FERRIC IRON (TC 3.A.1.14.3)  
ABC TRANSPORTER GALACTOSE (TC 3.A.1.2.3)  
ABC TRANSPORTER GLUTAMATE ASPARTATE (TC 3.A.1.3.4)  
ABC TRANSPORTER GLUTAMINE (TC 3.A.1.3.2)  
ABC TRANSPORTER GLYCEROL (TC 3.A.1.1.3)  
ABC TRANSPORTER HEME (TC 3.A.1.107.1)  
ABC TRANSPORTER HISTIDINE LYSINE ARGININE ORNITHINE (TC 3.A.1.3.1)  
ABC TRANSPORTER IRON(III) DICITRATE (TC 3.A.1.14.1)  
ABC TRANSPORTER L-PROLINE GLYCINE BETAINE (TC 3.A.1.12.1)  
ABC TRANSPORTER MACROLE  
ABC TRANSPORTER MALTOSE  
ABC TRANSPORTER MOLYBDENUM (TC 3.A.1.8.1)  
ABC TRANSPORTER NICKEL (TC 3.A.1.5.3)  
ABC TRANSPORTER OLIGOPEPTIDE (TC 3.A.1.5.1)  
ABC TRANSPORTER PEPTIDE (TC 3.A.1.5.5)  
ABC TRANSPORTER PHOSPHATE (TC 3.A.1.7.1)  
ABC TRANSPORTER POLYAMINE PUTRESCINE SPERMIDINE (TC 3.A.1.11.1)  
ABC TRANSPORTER PUTRESCINE (TC 3.A.1.11.2)  
ABC TRANSPORTER RIBOSE (TC 3.A.1.2.1)  
ACETOGENESIS\_FROM\_PYRUVATE  
ADHESION\_TO\_EUKARYOTIC\_CELL  
AEROBIC\_RESPIRATORY\_DEHYDROGENASES  
ALANINE\_BIOSYNTHESIS  
ALLANTOIN\_DEGRADATION  
AMMONIA\_ASSIMILATION  
ANAEROBIC\_RESPIRATORY\_DEHYDROGENASES  
ANAEROBIC\_RESPIRATORY\_REDUCTASES  
ARGININE\_BIOSYNTHESIS  
ARGININE\_DEGRADATION  
ASP-GLU-TRNA(ASN-GLN)\_TRANSMIDATION  
BACTERIAL\_CELL\_DIVISION  
BETAINE\_BIOSYNTHESIS  
BILIN\_BIOSYNTHESIS  
BIOTIN\_BIOSYNTHESIS  
CALVIN-BENSON\_CYCLE  
CARNITINE\_METABOLISM  
CAROTENOIDS  
CHAPERONES  
CHLOROPHYLL\_BIOSYNTHESIS  
CHORISMATE\_SYNTHESIS  
CMP-N-ACETYLNEURAMINATE\_BIOSYNTHESIS  
COENZYME\_A\_BIOSYNTHESIS  
CYANOBACTERIAL\_CIRCADIAN\_CLOCK  
CYANOBACTERIAL\_CO2\_UPTAKE  
CYANOPHYCIN\_METABOLISM  
CYSTEINE\_BIOSYNTHESIS  
CYTOCHROME\_B6-F\_COMPLEX  
CYTOLETHAL\_DISTENDING\_TOXIN\_OF\_CAMPYLOBACTER\_JEJUNI  
D-ARABINOSE\_DEGRADATION  
D-GALACTARATE\_DEGRADATION  
D-GALACTURONATE\_DEGRADATION  
D-GLUCARATE\_DEGRADATION  
DE\_NOVO\_PURINE\_BIOSYNTHESIS  
DE\_NOVO\_PYRIMIDINE\_SYNTHESIS  
DENITRIFICATION  
DNA-REPLICATION  
DNA\_REPAIR\_BASE\_EXCISION  
DTDP-RHAMNOSE\_SYNTHESIS  
EMBDEN-MEYERHOF\_AND\_GLUCEONEGENESIS  
ENTEROBACTIN\_BIOSYNTHESIS  
ENTNER-DOUDOROFF\_PATHWAY  
FOF1-TYPE\_ATP\_SYNTHASE  
FATTY\_ACID\_BIOSYNTHESIS\_FASII  
FATTY\_ACID\_METABOLISM  
FATTY\_ACID\_OXIDATION\_PATHWAY  
FE-S\_CLUSTER\_ASSEMBLY  
FLAGELLUM  
FMN\_AND\_FAD\_BIOSYNTHESIS  
FOLATE\_BIOSYNTHESIS  
FORMATE\_HYDROGENASE  
FRUCTOSE\_AND\_MANNANOSE\_METABOLISM  
FUMARATE\_REDUCTASE  
GALACTITOL\_DEGRADATION  
GALACTOSE\_DEGRADATION  
GENERAL\_SECRETORY\_PATHWAY\_(SEC-SRP)\_COMPLEX\_(TC\_3.A.5.1.1)  
GLUTAMATE\_BIOSYNTHESIS  
GLUTATHIONE\_REDOX\_METABOLISM  
GLYCEROL\_METABOLISM  
GLYCEROLIPID\_METABOLISM  
GLYCINE\_SYNTHESIS  
GLYOXYLATE\_DEGRADATION  
GLYOXYLATE\_SYNTHESIS  
GROEL\_GROES  
HISTIDINE\_BIOSYNTHESIS  
HISTIDINE\_DEGRADATION  
HMG\_COA\_SYNTHESIS  
INORGANIC\_SULFUR\_ASSIMILATION  
INOSITOL\_CATABOLISM\_BY\_VV  
IRON\_ACQUISITION  
ISOPRENOID\_BIOSYNTHESIS  
KETOGLUCONATE\_METABOLISM  
L-ASCORBATE\_DEGRADATION  
LACTOSE\_DEGRADATION  
LEUCINE\_DEGRADATION\_AND\_HMG-COA\_METABOLISM  
LEUCINE\_SYNTHESIS  
LYSINE\_BIOSYNTHESIS\_DAP\_PATHWAY  
MANNANOSE-SENSITIVE\_HEMAGGLUTININ\_TYPE\_4\_PILUS  
MANNANOSE\_AND\_FRUCTOSE\_METABOLISM  
MANNANOSE\_AND\_GDP-MANNANOSE\_METABOLISM  
MENAQUINONE\_AND\_PHYLLOQUINONE\_BIOSYNTHESIS  
METHANOGENESIS  
METHIONINE\_BIOSYNTHESIS  
METHIONINE\_METABOLISM  
METHYLCITRATE\_CYCLE  
N-ACETYL-D-GLUCOSAMINE\_UTILIZATION  
N-LINKED\_GLYCOSYLATION\_IN\_BACTERIA  
NAD\_AND\_NADP\_COFACTOR\_BIOSYNTHESIS\_GLOBAL  
NADH-QUINONE\_OXIDOREDUCTASE\_(COMPLEX\_I)  
NADH-UBIQUINONE\_OXIDOREDUCTASE\_(COMPLEX\_II)  
NITRATE,NITRITE,\_NITROUS\_OXIDE\_REDUCTASES  
NITRATE\_AND\_NITRITE\_REDUCTION  
NITRATE\_ASSIMILATION  
NITRITE\_REDUCTION  
NITROSATIVE\_STRESS  
P-TYPE\_ATPASE\_TRANSPORTER\_POTASSIUM\_(TC\_3.A.3.7.1)  
PENTOSE\_PHOSPHATE\_PATHWAY\_(SG)  
PEPTIDOGLYCAN\_BIOSYNTHESIS  
PHENYLALANINE\_SYNTHESIS  
PHOTOSYSTEM\_I  
PHOTOSYSTEM\_II  
PHYCOBILISOME  
PLASTOQUINONE\_BIOSYNTHESIS  
POLYAMINE\_METABOLISM  
PORPHYRIN,\_HEME,\_AND\_SIROHEME\_BIOSYNTHESIS  
PPGPP\_BIOSYNTHESIS  
PROLINE\_SYNTHESIS  
PROPIONATE\_CATABOLISM\_VIA\_2-METHYLCITRATE\_CYCLE  
PROTEASOME\_ARCHAEL  
PROTEASOME\_EUKARYOTIC  
PTERIN\_BIOSYNTHESIS  
PURINE\_CONVERSIONS  
PURINE\_CONVERSIONS\_2  
PUTRESCINE\_AND\_4-AMINOBUTYRATE\_DEGRADATION  
PYRIMIDINE\_CONVERSIONS  
PYRUVATE,\_PEP\_AND\_ACETYL-COA\_(ANAPLEROTIC\_REACTIONS)  
PYRUVATE\_ALANINE\_SERINE\_INTERCONVERSIONS  
QUEUOSINE  
RESISTANCE\_TO\_FLUOROQUINOLONES  
RIBOFLAVIN\_METABOLISM  
RIBONUCLEOTIDE\_REDUCTION  
RIBOSOME\_BIOGENESIS\_BACTERIAL  
RIBOSOME\_LSU\_BACTERIAL  
RIBOSOME\_LSU\_EUKARYOTIC\_AND\_ARCHAEL  
RIBOSOME\_SSU\_BACTERIAL  
RIBOSOME\_SSU\_CHLOROPLAST  
RIBOSOME\_SSU\_EUKARYOTIC\_AND\_ARCHAEL  
RNA\_POLYMERASE\_ARCHAEL  
RNA\_POLYMERASE\_ARCHAEL\_INITIATION\_FACTORS  
RNA\_POLYMERASE\_BACTERIAL  
RNA\_POLYMERASE\_CHLOROPLAST  
RNA\_POLYMERASE\_I  
RNA\_POLYMERASE\_II  
RNA\_POLYMERASE\_II\_INITIATION\_FACTORS  
RNA\_POLYMERASE\_III  
SERINE\_BIOSYNTHESIS  
SOLUBLE\_CYTOCHROMES\_AND\_FUNCTIONALLY\_RELATED\_ELECTRON\_CARRIERS  
SUCCINATE\_DEHYDROGENASE  
SUCROSE\_METABOLISM  
SULFATE\_ASSIMILATION  
SULFUR\_METABOLISM  
SUPERPATHWAY\_OF\_FUCOSE\_AND\_RHAMNOSE\_DEGRADATION  
SUPERPATHWAY\_OF\_GLUTAMATE,\_ASPARTATE,\_ASPARAGINE\_BIOSYNTHESIS  
SUPERPATHWAY\_OF\_HEXITOL\_DEGRADATION  
SUPERPATHWAY\_OF\_RIBOSE\_AND\_DEOXYRIBOSE\_PHOSPHATE\_METABOLISM  
TCA\_CYCLE  
TERMINAL\_CYTOCHROME\_OXIDASES  
TERMINAL\_CYTOCHROME\_C\_OXIDASES  
THERMOTOGA\_ALANINE\_BIOSYNTHESIS  
THIAMIN\_BIOSYNTHESIS  
THREONINE\_SYNTHESIS  
THREONINE\_TO\_Isoleucine  
TOCOPHEROL\_BIOSYNTHESIS  
TRANSCRIPTION\_FACTORS\_ARCHAEL  
TRANSCRIPTION\_FACTORS\_BACTERIAL  
TRANSLATION\_ELONGATION\_FACTORS\_EUKARYOTIC\_AND\_ARCHAEL  
TRANSLATION\_FACTORS\_BACTERIAL  
TRANSLATION\_INITIATION\_FACTORS\_EUKARYOTIC\_AND\_ARCHAEL  
TRANSPORT\_OF\_NICKEL\_AND\_COBALT  
TREHALOSE\_BIOSYNTHESIS  
TRICARBALLYLATE\_UTILIZATION  
TRNA\_AMINOACYLATION  
TRNA\_PROCESSING  
TRNA\_SPLICING  
TRP\_SYNTHESIS  
TYPE\_II\_SECRETION\_SYSTEM  
TYPE\_III\_SECRETION\_SYSTEM  
TYPE\_IV\_SECRETION\_SYSTEM  
TYROSINE\_SYNTHESIS  
UBIQUINONE\_BIOSYNTHESIS  
UBIQUINONE\_MENAQUINONE-CYTOCHROME\_C\_REDUCTASE\_COMPLEXES  
UDP-N-ACETYLMURAMATE\_FROM\_FRUCTOSE-6-PHOSPHATE\_BIOSYNTHESIS  
UREA\_DECOMPOSITION  
V-TYPE\_ATP\_SYNTHASE



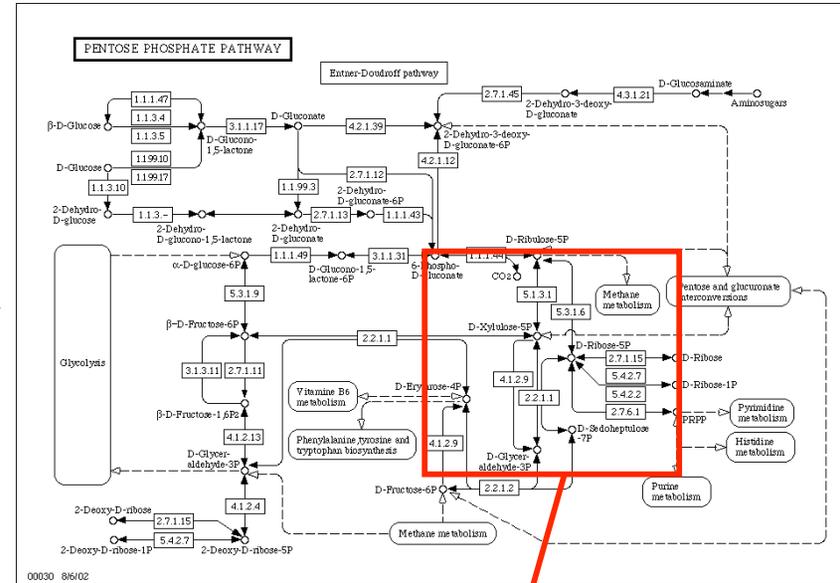
***E. coli* K-12 Metabolic Overview**

Source: EcoCyc

# CARBOHYDRATE METABOLISM

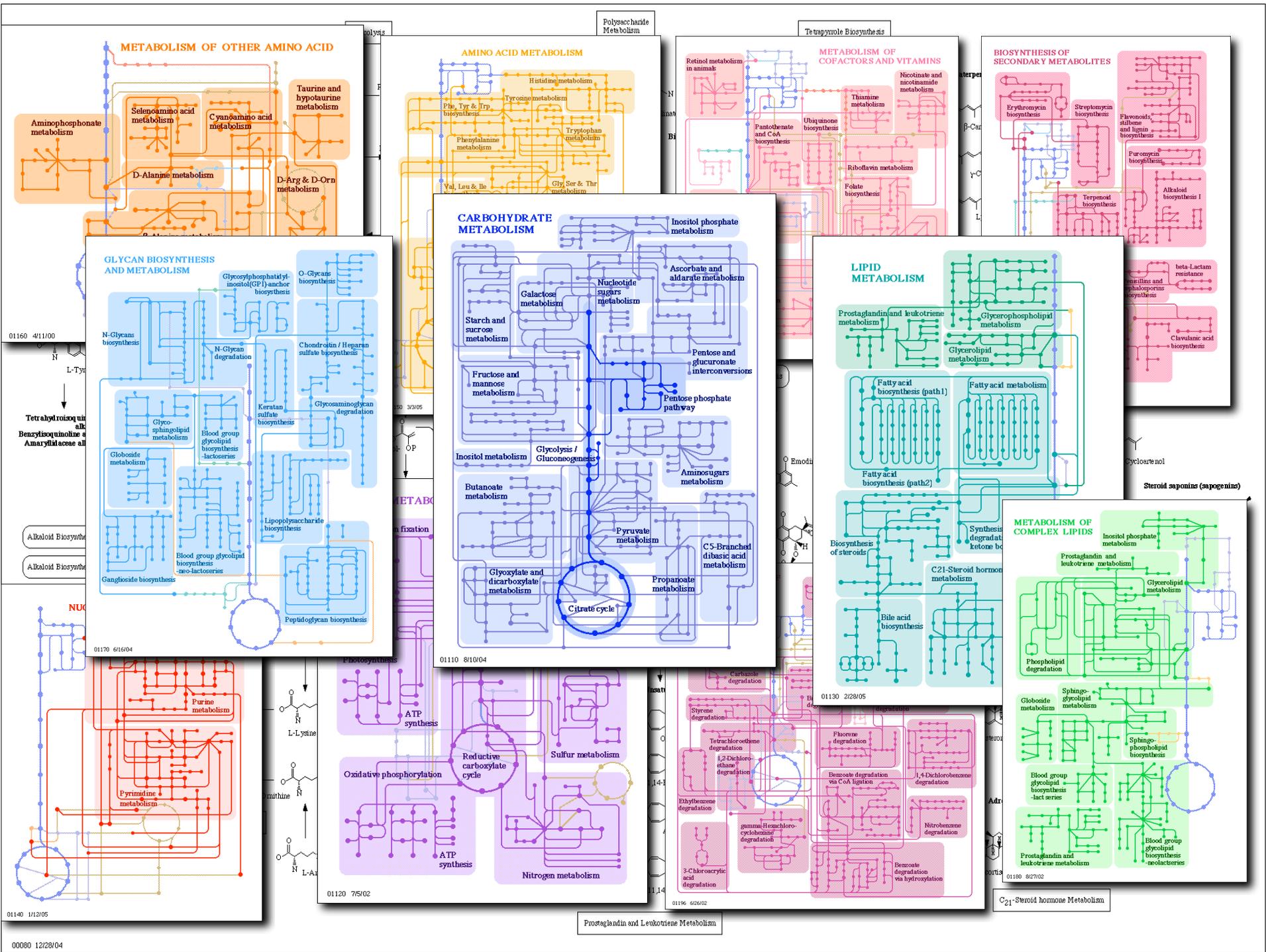


01110 7/8/02



```

SBMLNODES → {}, SBMLIC → {S1[0] == 1, S2[0] == 1, S3[0] == 1, S4[0] == 1, S5[0] == 1},
SBMLParameters → {}, SBMLSpecies → {S1[t], S2[t], S3[t], S4[t], S5[t]},
SBMLAlgebraicRules → {}, SBMLAssignmentRules → {},
SBMLReactions → {S1 → S2, S2 + 5S3 → 2S2 + 4S3, S3 → S4, S4 → S5},
SBMLStoichiometryMatrix →
{{-1, 0, 0, 0}, {1, 1, 0, 0}, {0, -1, -1, 0}, {0, 0, 1, -1}, {0, 0, 0, 1}},
SBMLMassBalanceEquations → {S1'[t] == -v[1], S2'[t] == v[1] + v[2], S3'[t] == -v[2] - v[3],
S4'[t] == v[3] - v[4], S5'[t] == v[4]}, SBMLMassActionEquations → {S1'[t] == -S1[t] v[1],
S2'[t] == S1[t] v[1] + S2[t] S3[t]^5 v[2], S3'[t] == -S2[t] S3[t]^5 v[2] - S3[t] v[3],
S4'[t] == S3[t] v[3] - S4[t] v[4], S5'[t] == S4[t] v[4]},
SBMLMassActionVariables → {S1[t], S2[t], S3[t], S4[t], S5[t]}, SBMLFunctions → {},
SBMLUnitDefinitions → {Units`substance → Units`mole, Units`volume → Units`litre,
Units`time → Units`second, Units`area → Units`metre^2, Units`length → Units`metre},
SBMLUnitAssociations → {C → Units`litre, S1 → Units`mole / Units`litre, S2 → Units`mole / Units`litre,
S3 → Units`mole / Units`litre, S4 → Units`mole / Units`litre, S5 → Units`mole / Units`litre}, SBMLNameIDAssociations → {},
SBMLEvents → {}, SBMLModelName → stoichiometry_matrix_example,
SBMLModelid → stoichiometry_matrix_example, SBMLCompartments → {C},
SBMLSpeciesCompartmentAssociations → {S1 → C, S2 → C, S3 → C, S4 → C, S5 → C}
    
```



# MODEL GENERATION FROM DATABASES

## Database Tables

## Model Cores

## Simulation Environments

Genes

Cis regions

Regulons

Roles

Ecs and TCs

Kinetics

Reactions

Interactions

Compounds

Locations and Compartments

Constraints

Objective Functions

Mass and Charge  
Balance Equations

Regulatory Networks

Signaling Networks

Reaction Rate Equations

Protein Interaction networks

Boundary Fluxes

Organizational Structures

Flux Balance Models

- Mathematica
- Octave
- GAMS

Stochastic Models

- StochSim
- Stock 2

Kinetics Models

- e-Cell
- SBML/mathML

Logical Models

- Mathematica
- Prolog

# Genotype → Phenotype

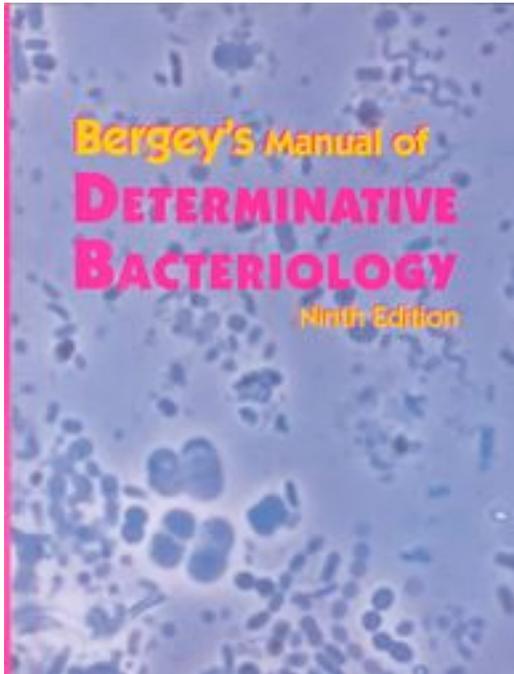
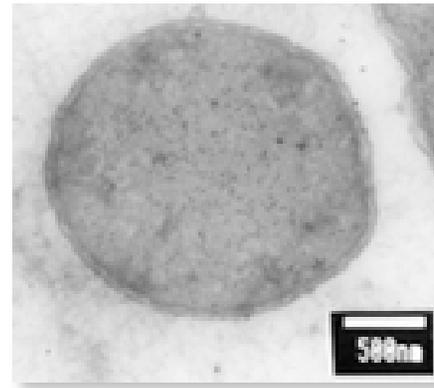
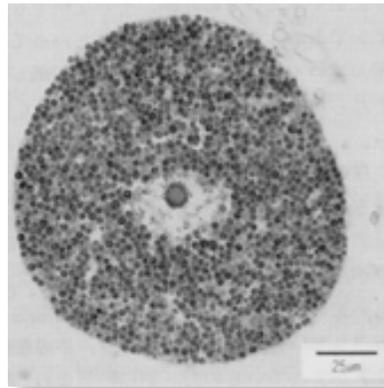
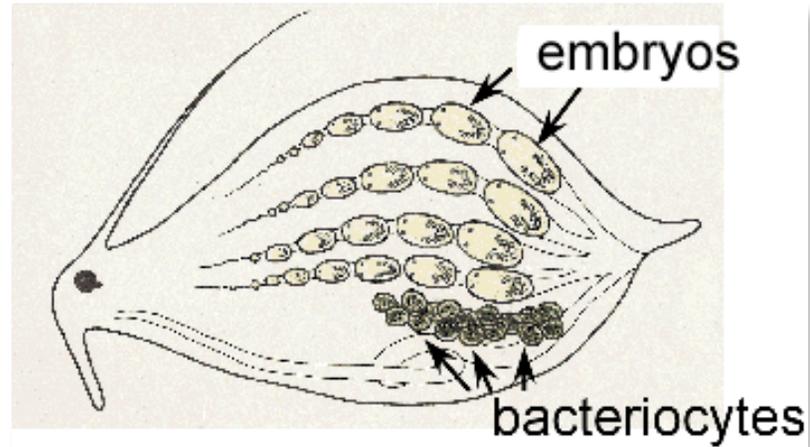


Table 13.4.  
Differential characteristics of the species of the genus *Bacillus*<sup>a,b</sup>

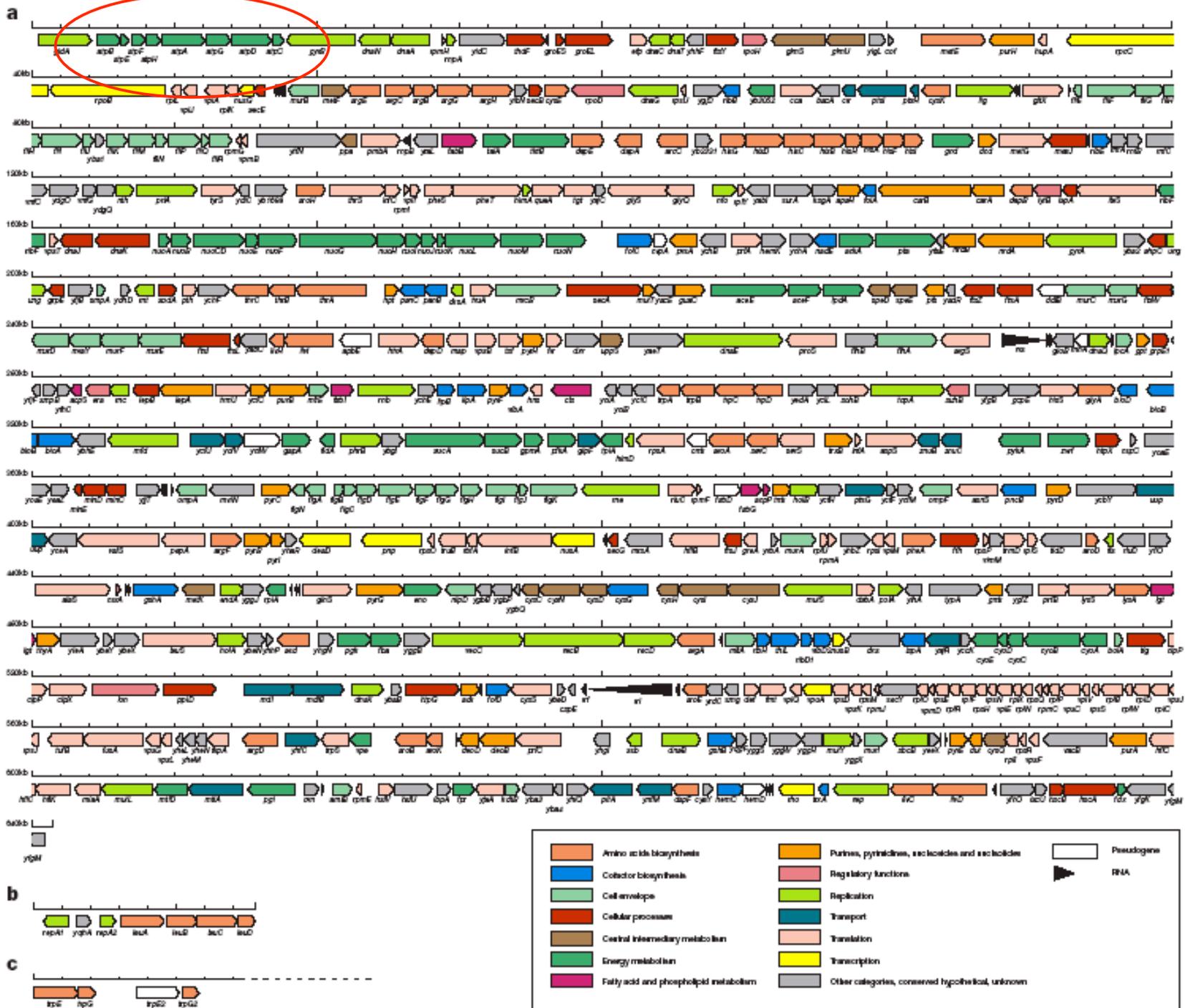
Characteristics	1. <i>B. subtilis</i>	2. <i>B. acidocaldarius</i> <sup>c</sup>	3. <i>B. alcalophilus</i> <sup>d</sup>	4. <i>B. alvei</i>	5. <i>B. anthracis</i>	6. <i>B. azotoformans</i> <sup>e</sup>	7. <i>B. bedius</i>	8. <i>B. brevis</i>	9. <i>B. cereus</i>	10. <i>B. circulans</i>	11. <i>B. coagulans</i>	12. <i>B. fastidiosus</i> <sup>f</sup>	13. <i>B. firmus</i> <sup>g</sup>	14. <i>B. globisporus</i> <sup>h</sup>	15. <i>B. insolitus</i> <sup>a</sup>	16. <i>B. lareüe</i>	17. <i>B. laterosporus</i>
Cell diameter >1.0 μm	-	ND	-	-	+	-	-	-	+	-	-	+	-	-	-	-	-
Spores round	-	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+
Sporangium swollen	-	ND	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Parasporal crystals	-	ND	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Catalase	+	ND	+	+	+	-	-	+	+	d	+	+	+	+	+	+	+
Anaerobic growth	+	-	-	+	+	-	-	-	+	-	+	+	+	+	+	+	+
Voges-Proskauer test	+	-	-	+	+	-	-	-	+	-	+	+	+	+	+	+	+
pH in V-P broth																	
<6	d		ND	+	+	ND	-	-	+	+	+	+	+	+	+	+	+
>7	-		ND	-	-	ND	+	+	-	-	-	-	-	-	-	-	-
Acid from																	
D-Glucose	+	ND	+	+	+	-	-	d	+	+	+	+	+	+	+	+	+
L-Arabinose	+	ND	+	+	+	ND	-	-	-	-	-	-	-	-	-	-	-
D-Xylose	+	ND	+	-	-	-	-	d	+	+	+	+	+	+	+	+	+
D-Mannitol	+	ND	+	-	-	-	-	d	+	+	+	+	+	+	+	+	+
Gas from glucose	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+
Hydrolysis of																	
Casein	+	ND	+	+	+	ND	+	d	+	d	d	+	+	+	+	+	+
Gelatin	+	ND	+	+	+	-	ND	d	+	d	+	+	+	+	+	+	+
Starch	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+
Utilization of																	
Citrate	+	-	-	-	d	+	-	d	+	d	d	-	-	-	-	-	-
Propionate	+	ND	-	ND	ND	-	ND	ND	+	+	+	+	+	+	+	+	+
Degradation of tyrosine	-	ND	-	d	d	-	-	+	+	+	+	+	+	+	+	+	+
Deamination of phenylalanine	-	ND	ND	-	-	-	-	-	+	+	+	+	+	+	+	+	+
Egg-yolk lecithinase	-	ND	-	-	+	ND	-	d	+	d	d	-	-	-	-	-	-
Nitrate reduced to nitrite	+	ND	-	-	+	ND	-	d	+	+	+	+	+	+	+	+	+
Formation of																	
Indole	-	ND	-	+	ND	-	-	-	+	+	+	+	+	+	+	+	+
Dihydroxyacetone	ND	ND	-	-	ND	-	-	-	+	+	+	+	+	+	+	+	+
NaCl and KCl required	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allantoin or urate required	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Growth at pH																	
6.8, nutrient broth	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5.7	+	d	-	-	+	-	-	d	+	d	+	-	-	-	-	-	-
Growth in NaCl																	
2%	+	ND	ND	ND	+	ND	ND	ND	ND	ND	+	+	+	+	+	+	+
5%	+	ND	-	d	+	-	-	ND	+	d	d	-	-	-	-	-	-
7%	+	ND	-	-	+	-	-	-	+	d	d	-	-	-	-	-	-
10%	ND	ND	-	-	ND	-	-	-	+	+	+	+	+	+	+	+	+
Growth at																	
5°C	-	-	ND	-	-	ND	-	-	-	-	-	-	-	-	-	-	-
10°C	d	-	ND	-	-	ND	-	-	d	d	+	+	+	+	+	+	+
30°C	+	-	+	+	+	+	+	+	d	d	+	+	+	+	+	+	+
40°C	+	-	+	+	+	+	+	+	d	d	+	+	+	+	+	+	+
50°C	d	+	-	-	-	-	-	d	-	-	-	-	-	-	-	-	-
55°C	-	+	-	-	-	-	-	d	-	-	-	-	-	-	-	-	-
65°C	-	+	-	-	-	-	-	d	+	d	-	-	-	-	-	-	-
Growth with lysozyme present	d	ND	-	+	+	-	-	-	+	+	-	-	-	-	-	-	-
Autotrophic with H <sub>2</sub> + CO <sub>2</sub> or CO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

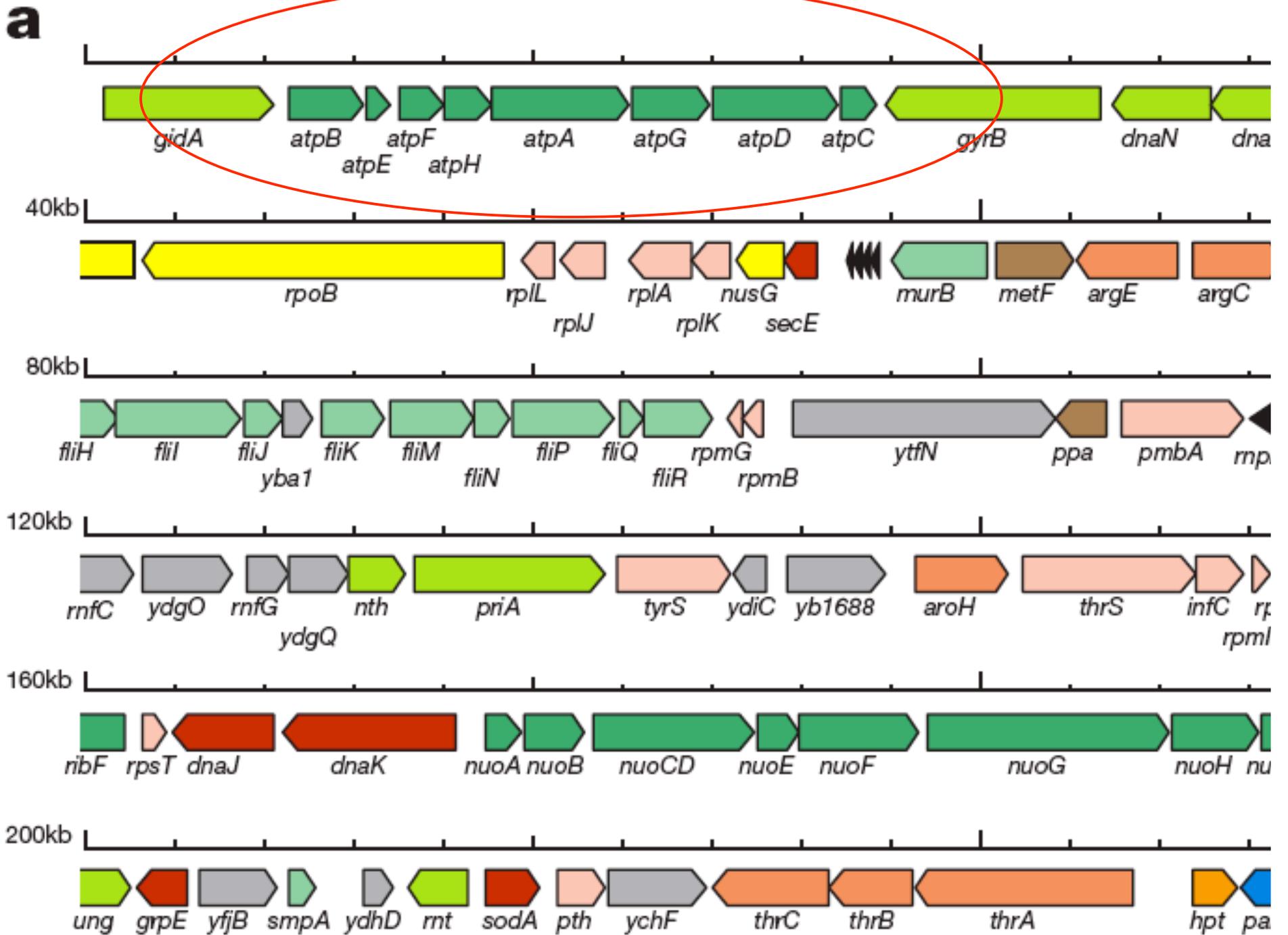
<sup>a</sup> Symbols: -, 90% or more of strains are negative; +, 90% or more of strains are positive; v, strain instability (not equivalent to "d"); d, 11-89% of strains are positive; ND, no data available; NG, no growth; <sup>b</sup> compiled from Smith et al. (1952), Gordon et al. (1973) and Knight and Proom (1950), except "v"; <sup>c</sup> data from Dariand and Brock (1971); <sup>d</sup> data from Boyer et al. (1973); <sup>e</sup> data from Pichinoty et al. (1978; 1983); <sup>f</sup> data from Fahmy (personal communication); <sup>g</sup> data from Gordon et al. (1977) and Gordon and Hyde (1982); <sup>h</sup> data from Larkin and Stokes (1967); <sup>i</sup> data from Hanáková-Bauerová et al. (1965); <sup>j</sup> data from McCaughey and Chu (1948); <sup>k</sup> data from Hanáková-Bauerová et al. (1966); <sup>l</sup> data from Gordon et al. (1977) and Gordon and Hyde (1982); <sup>m</sup> data from Marshall and Ohye (1966); <sup>n</sup> data from Rüger (1983); <sup>o</sup> data from Schenk and Aragno (1979) and Krüger and Meyer (1984); and <sup>p</sup> (-), few gas bubbles may be formed.

# *Buchnera aphidicola* sp.



100 copies of the genome per cell, lacks cell defense genes





Alphaproteobacteria

### Basic Spreadsheet

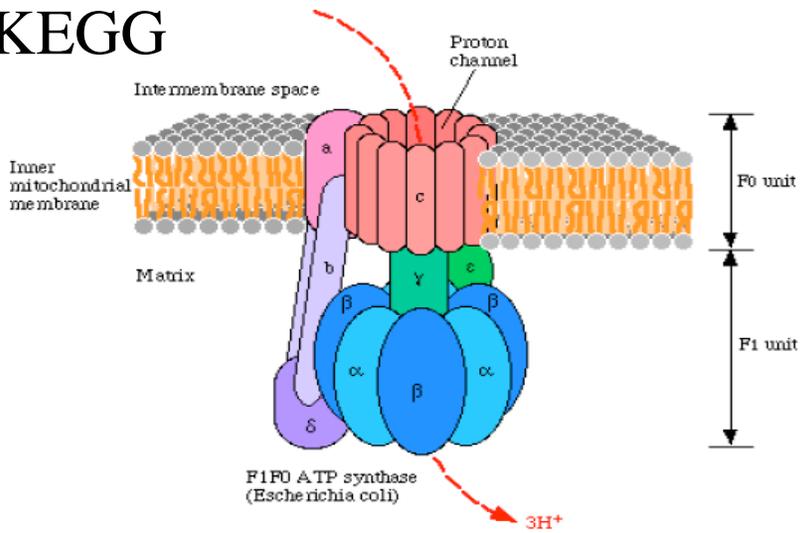
Genome ID	Organism	Variant Code	A	B	C	alpha	beta	gamma	delta	epsilon	I
1769.1	Mycobacterium leprae [B]	1	1141	1143	1142	1145	1147	1146	1144	1148	
216594.1	Mycobacterium marinum M [B]	1	2919	2921	2920	2923	2925	2924	2922	2926	
233413.1	Mycobacterium bovis subsp. bovis AF2122/97 [B]	1	1335	1337	1336	1339	1341	1340	1338	1342	
107806.2	Buchnera aphidicola str. APS (Acyrtosiphon pisum) [B]	1	4	4	3	6	8	7	5	9	
224915.1	Buchnera aphidicola str. Bp (Baizongia pistaciae) [B]	1	2	4	3	6	8	7	5	9	
198804.1	Buchnera aphidicola str. Sg (Schizaphis graminum) [B]	1	2	4	3	6	8	7	5	9	
83331.1	Mycobacterium tuberculosis CDC1551 [B]	1	1383	1385	1384	1387	1389	1388	1386	1390	
83332.1	Mycobacterium tuberculosis H37Rv [B]	1	1306	1308	1307	1310	1312	1311	1309	1313	
159288.1	Staphylococcus aureus EMRSA-16 (Str. 252) [B]	2	873	871	872	869	867	868	870	866	874
93061.1	Staphylococcus aureus NCTC 8325 [B]	2	1454	1456	1455	1458	143,1460	1459	1457	144	1453
Genome ID	Organism	Variant Code	A	B	C	alpha	beta	gamma	delta	epsilon	I
196620.1	Staphylococcus aureus subsp. aureus MW2 [B]	2	2033	2031	2032	2029	2027	2028	2030	2026	2034
158878.1	Staphylococcus aureus subsp. aureus Mu50 [B]	2	2109	2107	2108	2105	2103	2104	2106	2102	2110
176280.1	Staphylococcus epidermidis ATCC 12228 [B]	2	1706	1704	1705	1702	1700	1701	1703	1699	1707
208435.1	Streptococcus agalactiae 2603V/R [B]	1	836	837	835	839	841	840	838	842	
211110.1	Streptococcus agalactiae NEM316 [B]	1	875	876	874	878	880	879	877	881	
1336.1	Streptococcus equi [B]	1	1105	1106	1104	1108	1110	1109	1107	1111	
210007.1	Streptococcus mutans UA159 [B]	1	1390	1389	1391	1387	1385	1386	1388	1384	
714.2	Actinobacillus actinomycetemcomitans HK1651 [B]	1	1715	1717	1716	1719	1721	1720	1718	1722	
228399.1	Actinobacillus pleuropneumoniae serovar 1 str. 4074 [B]	1	818	820	819	822	824	823	821	825	
Genome ID	Organism	Variant Code	A	B	C	alpha	beta	gamma	delta	epsilon	I
181661.1	Agrobacterium tumefaciens str. C58 (Cereon) [B]	2	699	702	700	2520	2518	2519	2521	2517	698
180835.1	Agrobacterium tumefaciens str. C58 (U. Washington) [B]	1	699	702	700	2574	2572	2573	2575	2571	
224324.1	Aquifex aeolicus VF5 [B]	1	172	1136,1137	170	513	1443	1444	1138	509	
354.1	Azotobacter vinelandii [B]	2	1776	1778	1777,2560	1780	1782,2565	1781,2556	1779	1783	1775
191218.1	Bacillus anthracis str. A2012 [B]	2	716	714	715	712	710	711	713	709	717
198094.1	Bacillus anthracis str. Ames [B]	2	5139	5137	5138	5135	5133	5134	5136	5132	3133,5140
86665.1	Bacillus halodurans [B]	2	3760	3758	3759	3756	3754	3755	3757	3753	3761
224308.1	Bacillus subtilis subsp. subtilis str. 168 [B]	2	3694	3692	3693	3690	3688	3689	3691	3687	3695
817.2	Bacteroides fragilis NCTC9343 [B]	1	1638	1640	1639	1642	1635	1643	1641	1636	
Genome ID	Organism	Variant Code	A	B	C	alpha	beta	gamma	delta	epsilon	I
226186.1	Bacteroides thetaiotaomicron VPI-5482 [B]	1	714	716	715	718	711	719	717	712	
205913.1	Bifidobacterium longum DJO10A [B]	1	1550	1552	1551	1554	1556	1555	1553	1557	
206672.1	Bifidobacterium longum NCC2705 [B]	1	334	332	333	330	328	329	331	327	

Done



# From KEGG

## ATP SYNTHESIS



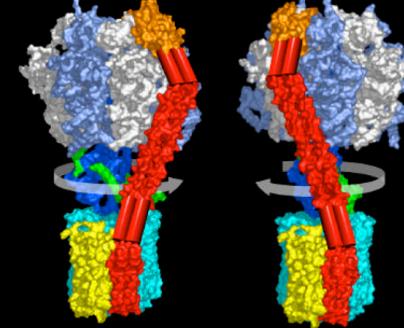
F-type ATPase (Bacteria)

beta	alpha	gamma	delta	epsilon	c	a	b
------	-------	-------	-------	---------	---	---	---

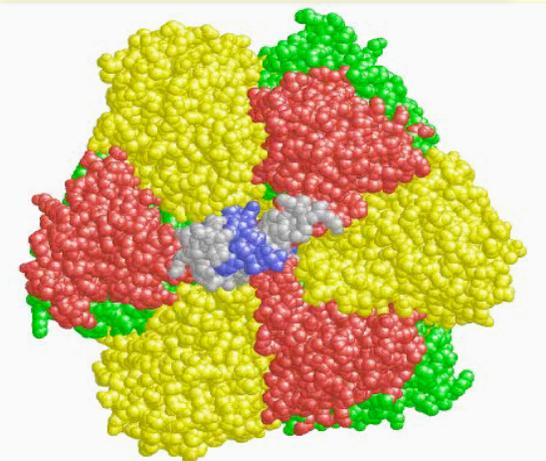
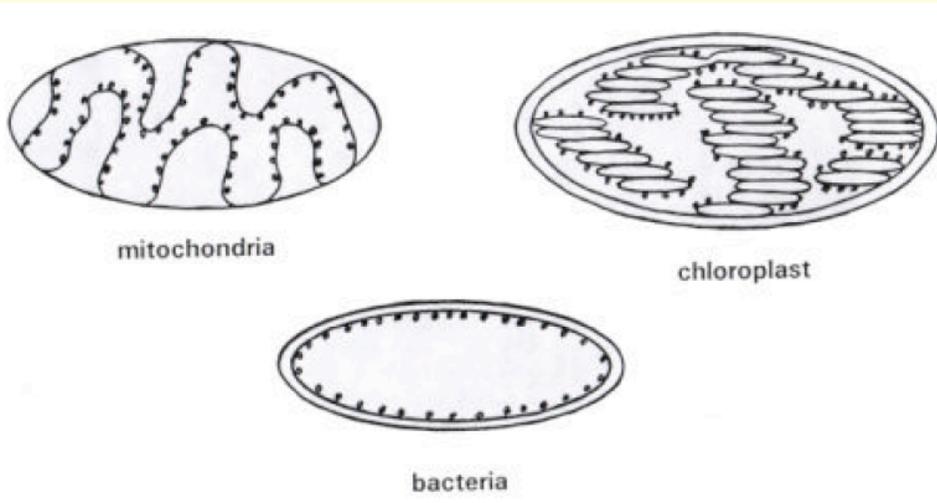


ATP Synthesis

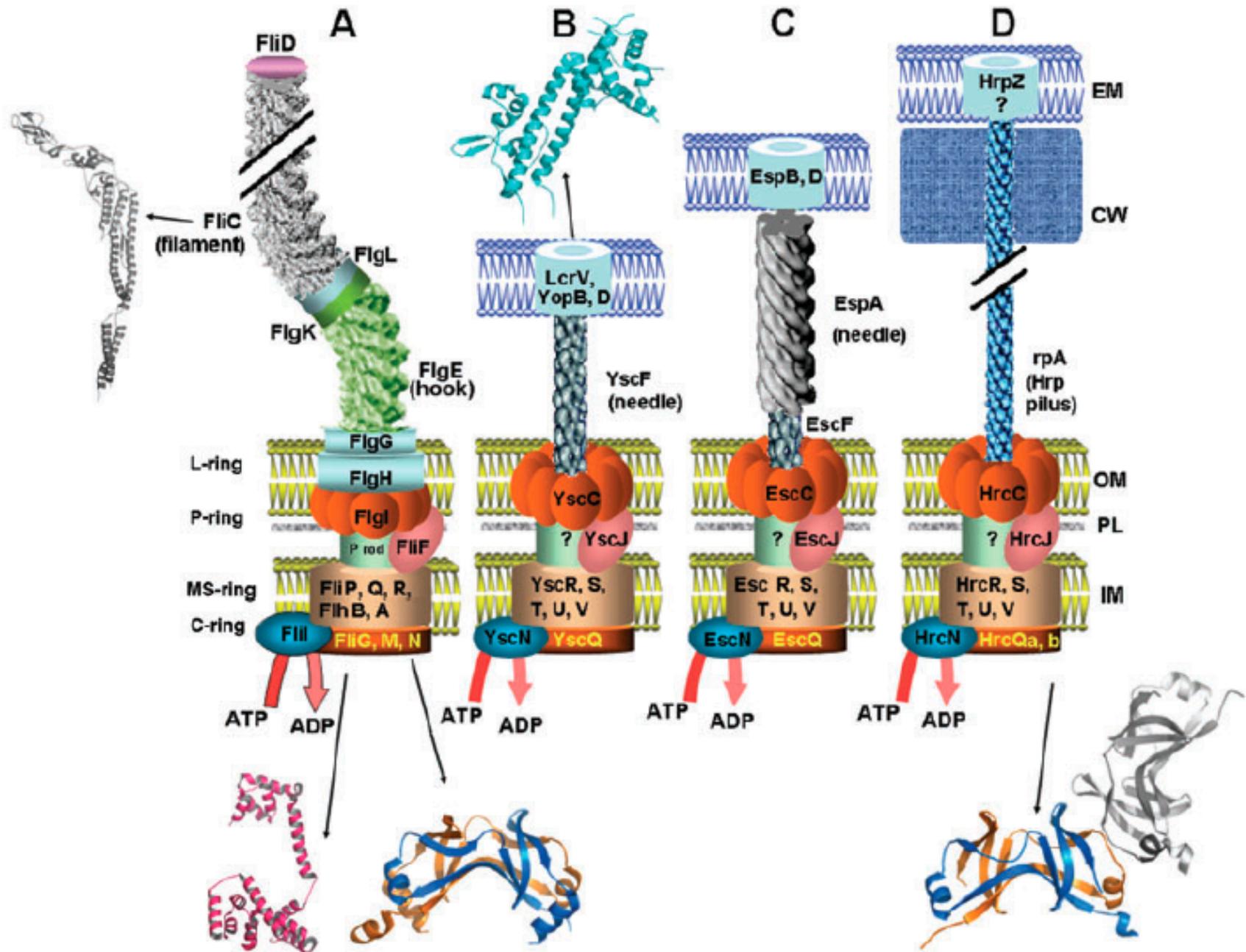
ATP-driven H<sup>+</sup> Pumping



Hardy et al (2003) J. Biomembr. Bioenerg. 35, 398-397



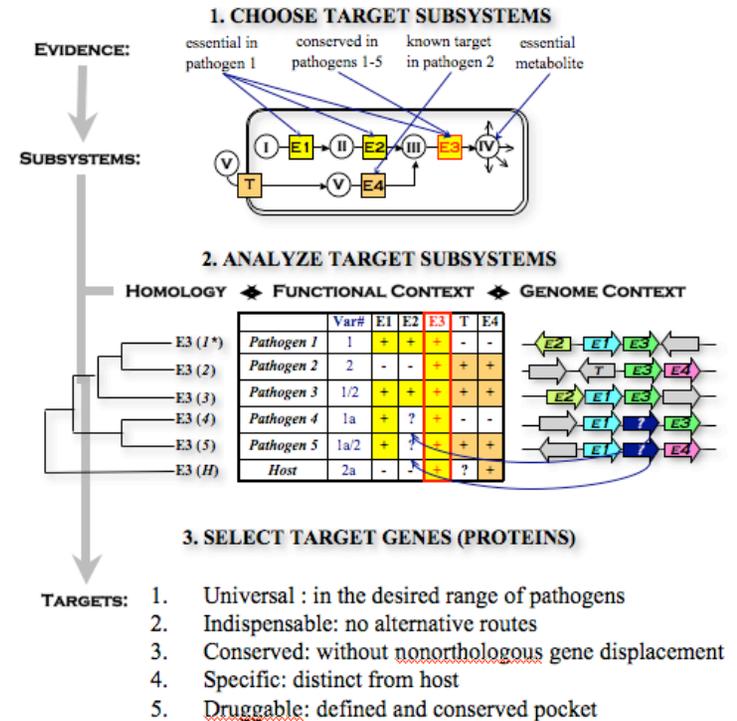
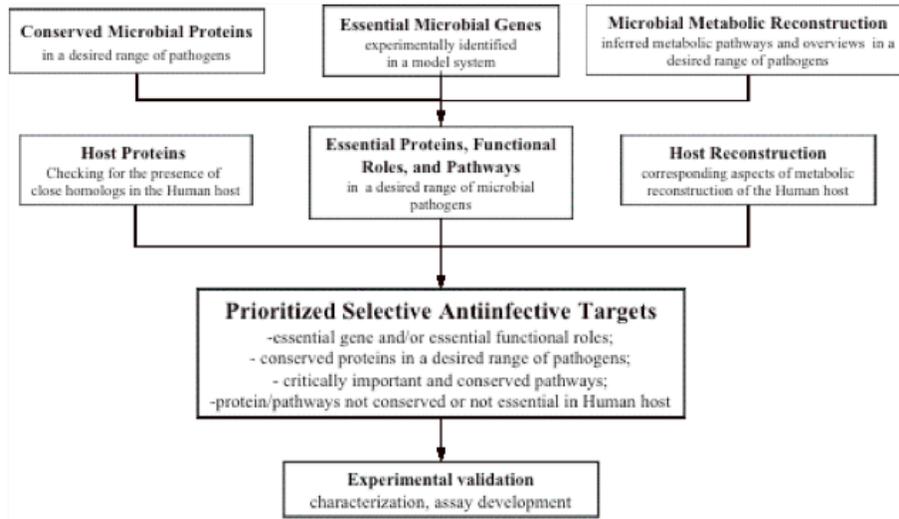
A top view of alpha3-beta3-gamma  
By Hongyun Wang & George Oster, U.C. Berkeley



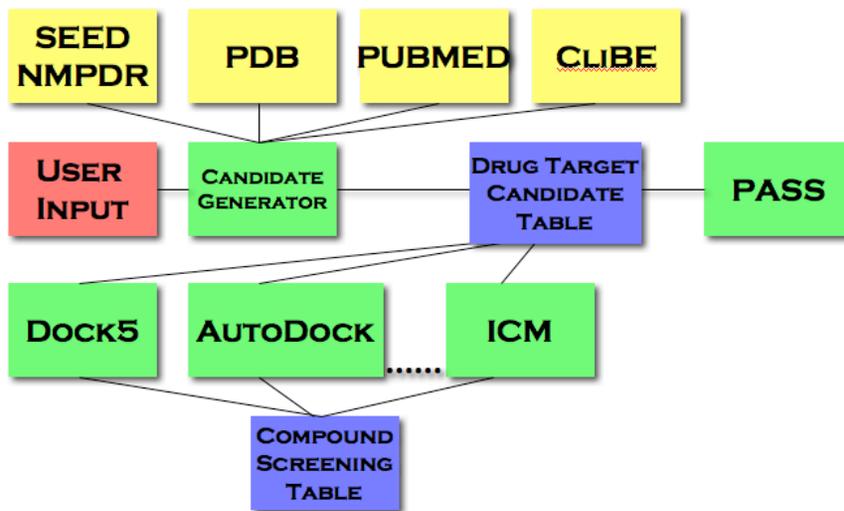
# DRUG DISCOVERY WORKFLOW {NEW ANTIBIOTICS}

1. IDENTIFY GENES/PROTEIN TARGETS FROM THE LITERATURE  
KNOWN ANTIBIOTIC TARGETS AND KNOWN ANTIBIOTIC RESISTANCE FACTORS  
ESSENTIAL GENES AND VIRULENCE ASSOCIATED FACTORS
2. DETERMINE FUNCTIONAL ROLE OF EACH TARGET  
COMPARATIVE ANALYSIS (CLUSTER ANALYSIS, SUBSYSTEM RECONSTRUCTION)
3. SEARCH FOR KNOWN INHIBITORS OF EACH TARGET  
LITERATURE MINING AND COMPUTATIONAL SCREENING (SEE STEP 8)
4. DETERMINE DEGREE OF CONSERVATION ACROSS SPECIES  
PHYLOGENY AND SEQUENCE ALIGNMENT  
CHARACTERIZATION OF THE ACTIVE SITE  
EARLY SCREEN IN HUMAN AND MODEL SYSTEMS
5. DETERMINE STRUCTURE OF EACH TARGET (PDB, COMPUTATION)  
DATABASE SEARCH/SIMILARITY AND STRUCTURAL MODELING
6. DETERMINE ACTIVE SITE OF EACH TARGET  
COMPUTATIONAL ANALYSIS OF EACH STRUCTURE
7. DETERMINE DRUGABILITY OF EACH TARGET  
SIZE OF POCKET, NUMBER OF POCKETS
8. SCREENING OF COMPOUNDS FOR BINDING AFFINITY ETC.  
COMPUTATIONAL AND HIGH-THROUGHPUT EXPERIMENTS
9. TOXICITY SCREENING IN HUMAN AND MODEL SYSTEMS  
COMPUTATIONAL AND HIGH-THROUGHPUT EXPERIMENTS

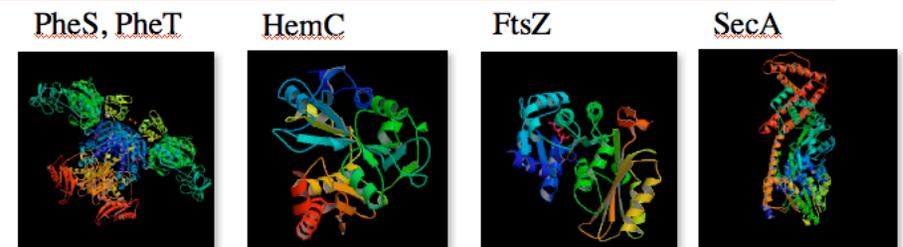
# Functional Genomics Approach to Anti-microbial Agent Development



## Drug Target Development and Screening

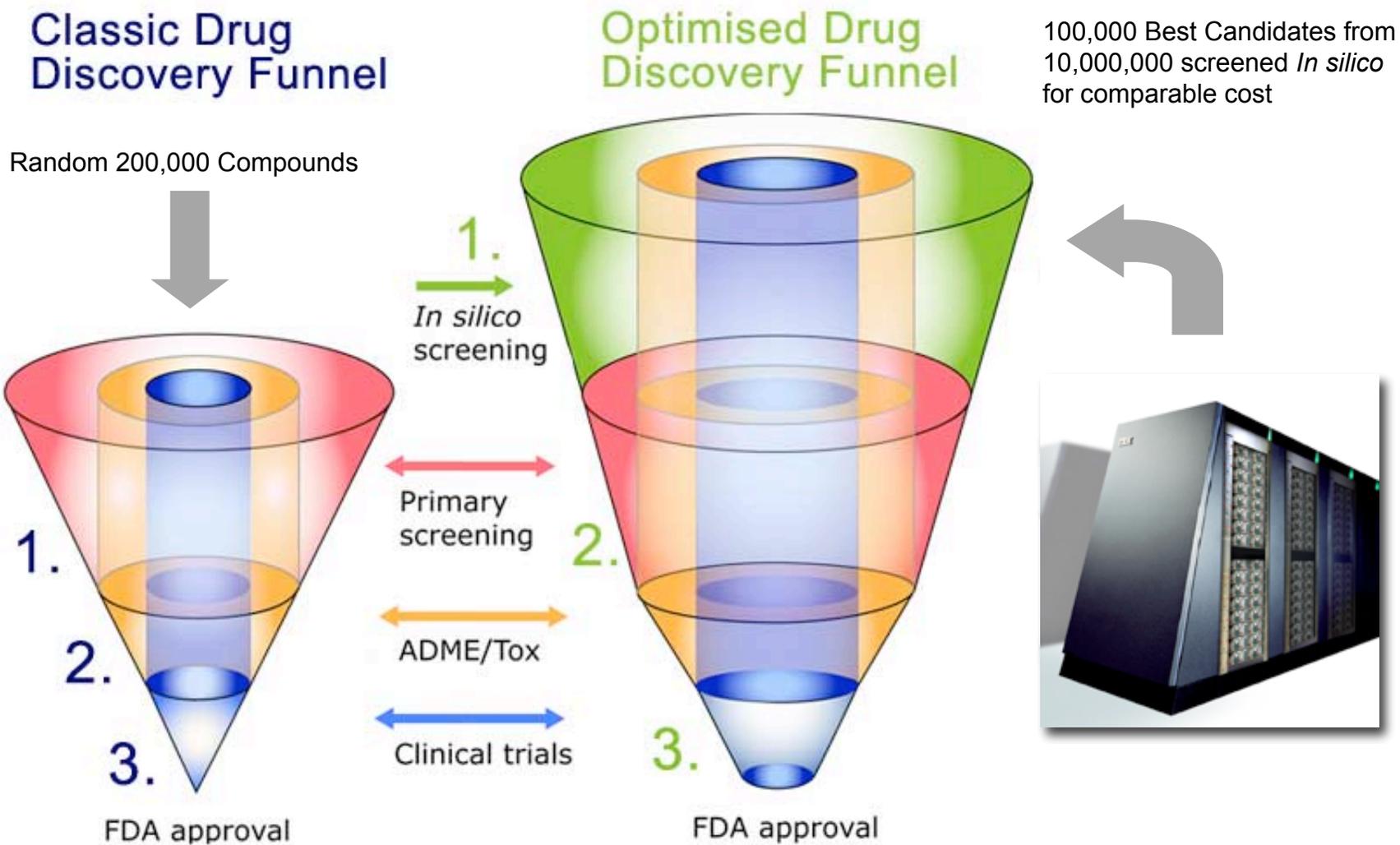


## Anti-Microbial Drug Targets on Blue Gene

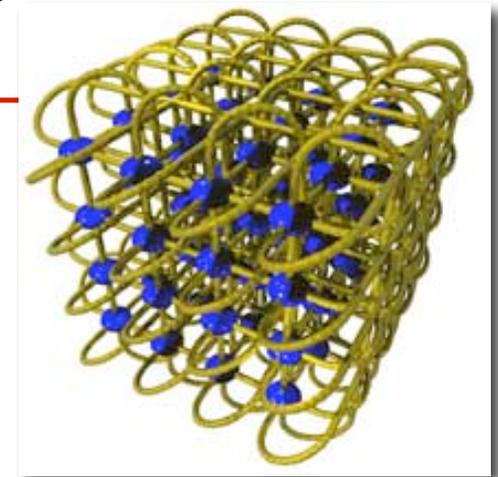
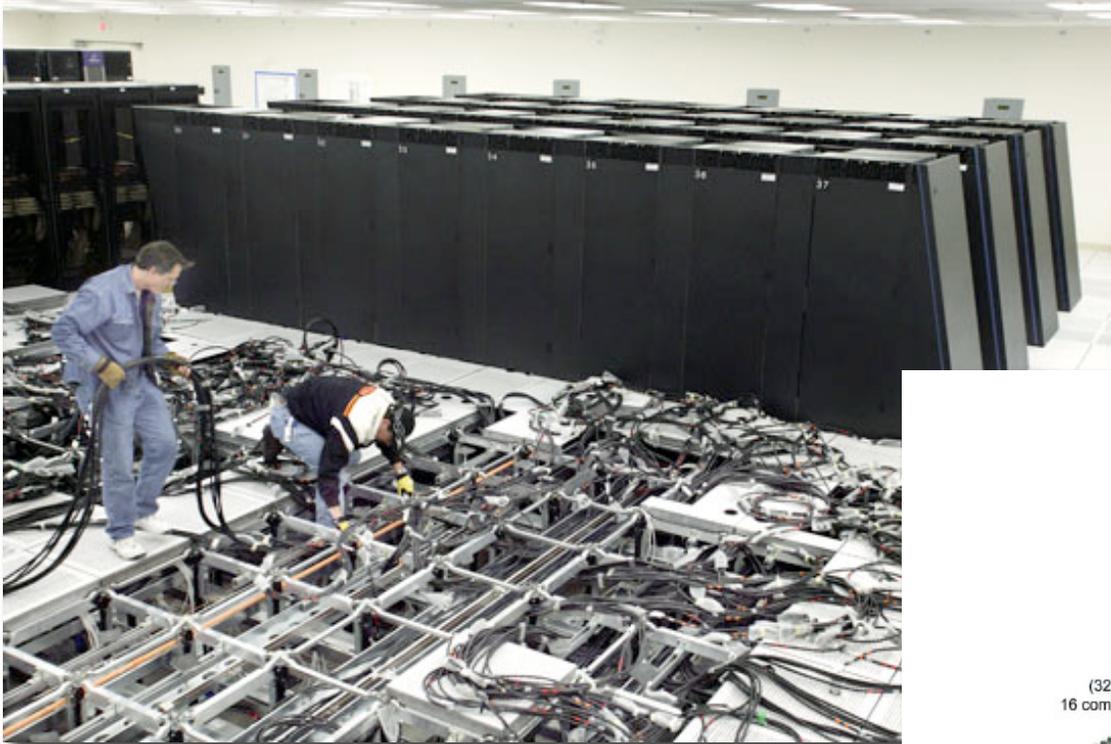


Our preliminary results indicate we can approach an *in silico* initial ligand screening throughput of approximately 100,000 screens per Blue Gene Rack (1024 nodes) per day.

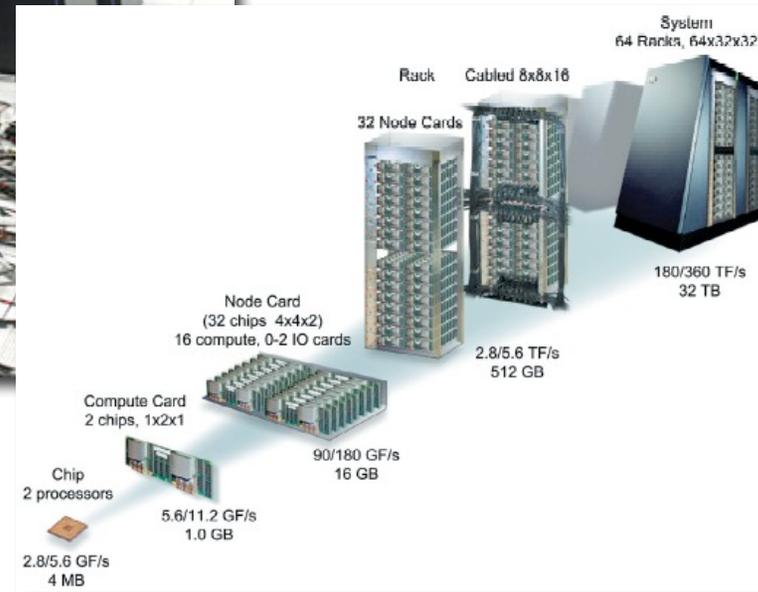
# The Economics of *in silico* Screening



# IBM's Blue Gene Supercomputer



Worlds fastest  
Supercomputer  
280 TeraFLOPS



# Petascale Biological Computations

---

- Searching for new antibiotics
  - 300 essential-gene-products x 3.3 million compounds
    - 990 million drug docking computations (each one involves about 20 different computations) ⇒ over 10 billion jobs
- Determining *in silico* essential genes in pathogens
  - Single, double and triple deletion *in silico* mutants
    - 1,000 gene models, 1M runs for double deletion mutants, 1B runs for triple deletion mutants
- Understanding the evolution of protein families
  - Searching horizontal gene transfers in early Prokaryotes
    - ~3000 protein families ⇒ for each one we want to build detailed gene phylogeny and reconcile with species tree
    - Thousands of phylogenies and tree reconciliations

# Computational Biology @ petascale

Biology Problem Area	@ 360 TF/s	@1000 TF/s	@ 5000 TF/s
Determining the detailed evolutionary history of each protein family ⇒ <i>This will enable rational planning for structural biology initiatives and will provide a foundation for assessing protein function and diversity</i>	3,000 hours to build reference database	300 hours to build reference database	60 hours to build reference database
Determining the frequency and detailed nature of horizontal gene transfers in prokaryotes ⇒ <i>This will shed light on the molecular and genetic mechanisms of evolution by means other than direct "Darwinian" descent and will contribute to our understanding of the acquisition of virulence and drug resistance in pathogens and the means by which prokaryotes adapt to the environment</i>	1,000 hours to study 200 gene families	1,000 hours to study 2000 gene families	1,000 hours to study 10,000 gene families
Automated construction of core metabolic models for all the sequenced DOE genomes ⇒ <i>This will enable dramatic acceleration of the promise of the GTL program and the use of microbial systems to address DOE mission needs in energy, environment, and science</i>	1 hour per organism, 100 hours per metagenome	10 organisms per hour, 10 hours per metagenome	50 organisms per hour, 2 hours per metagenome
Predict essential genes for all known sequenced micro-organisms ⇒ <i>This will enable a broader class of genes and gene products to be targeted for potential drugs and to predict culturability conditions for environmental microbes</i>	300 hours for 1,000 organisms 10 hours to predict culturability per organism	30 hours for 1,000 organisms, 1 hour to predict culturability per organism	30 hours for 5,000 orgs
Computational screening all known microbial drug targets against the public and private databases of chemical compounds to identify potential new inhibitors and potential drugs ⇒ <i>The resulting database would be a major national biological research resource that would have dramatic impact on worldwide health research and fundamental science of microbiology</i>	2 M ligands per day per target (1 year to screen all microbial targets)	20 M ligands per day per target (~1 month to screen all microbial targets)	1 machine year to screen all known human drug targets

Model and simulate the precise cellulose degradation and ethanol and butanol biosynthesis pathways at the protein/ligand level to identify opportunities for molecular optimization ⇒ <i>This would result in a set of model systems to be further developed for optimization of the production of biofuels</i>	Simulate in detail the directed evolution of individual enzymes	Simulate the co-evolution and optimization of a degradation or biosynthesis pathway of up five enzymes	Simulate the optimization of a complete cellulose to ethanol or butanol production system of over a dozen enzymatic steps
Model and simulate the replication of DNA to understand the origin of and the repair mechanisms of genetic mutations ⇒ <i>This would result in dramatic progress in the fundamental understanding of how nature manages mutations, which molecular factors determine the broad range of organism susceptibility to radiation and other mutagens</i>	30 ns simulation of DNA polymerase	10 ensembles of different DNA repair enzymes	Complete polymerase mediated base pair addition step
Model and simulate the process of DNA transcription and protein translation and assembly ⇒ <i>This would enable us to move forward on understanding post-transcription and post-translation modification and epigenetic regulation of protein synthesis</i>	Validate current understanding of ribosomal function	Explore splicing function and the evolution of intron/exon functions	Model the complete coupled processes of DNA transcription to Protein translation including regulatory processes
Model and simulate the interlinked metabolisms of microbial communities ⇒ <i>This project is relevant to understanding the biogeochemical cycles of extreme, natural and disturbed environments and will lead to the development of strategies for the production of bio-fuels and the development of new bio-engineered processes based on exploiting communities rather than individual organisms</i>	20 organisms in a linked metabolic network	100 organisms in a linked metabolic network	200 organisms in a linked metabolic network
<i>In silico</i> prediction of mutations and activity, conformational changes, active site alterations	One enzyme	Five-enzyme pathway	Eight enzyme pathway optimization
<b>Biology Problem Area</b>	<b>@ 360 TF/s</b>	<b>@1000 TF/s</b>	<b>@ 5000 TF/s</b>

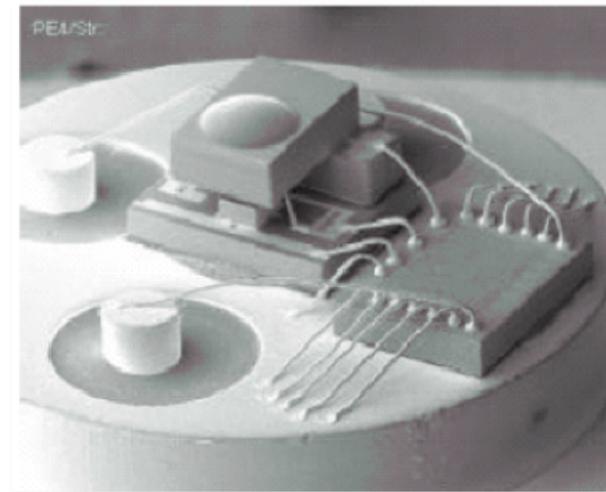
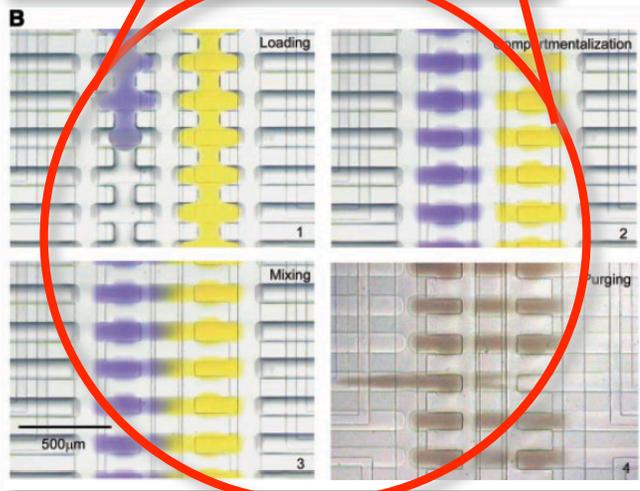
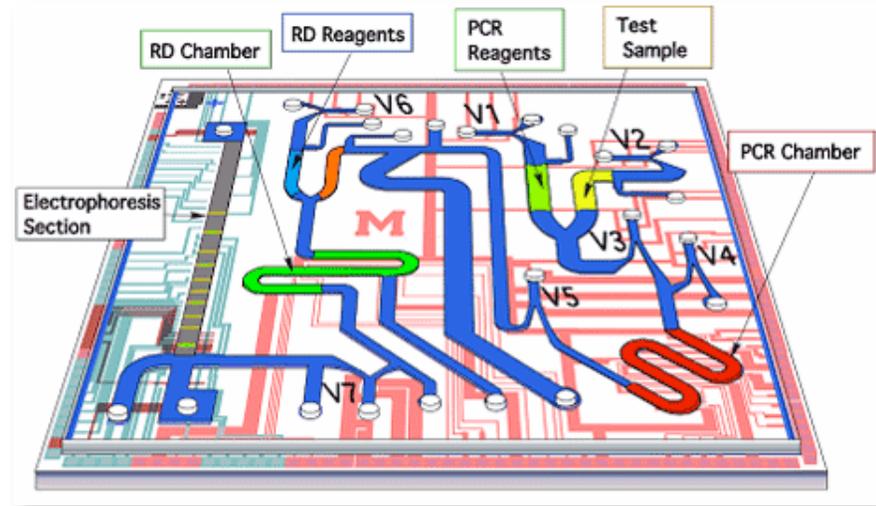
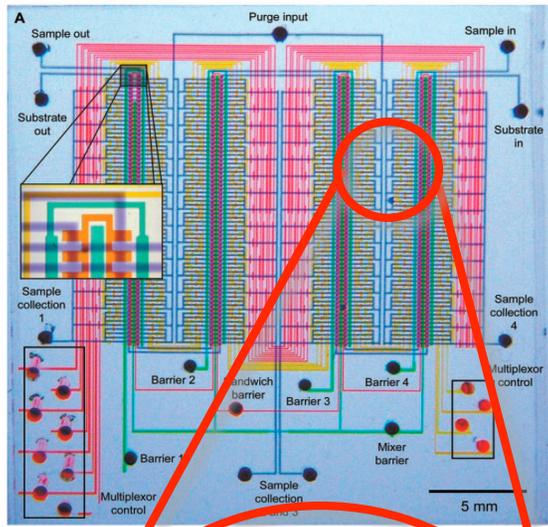
# Data Intensive Science: The Next Revolution

---

- HPC has been historically driven in part by the desire to create ever more realistic simulations of natural and man-made systems
- These systems are driven largely by and contribute to the success of “theoretical work”
- However, the availability of high-throughput experimental devices will generate high-quality datasets of unprecedented scale
  - Genome sequencers, gene chips, advanced detectors and imaging systems to large-scale networks of sensors
  - > Moore’s Law like improvements in data collection
  - New generations of mass storage devices and the computing to integrate them
  - New infrastructures, tools and techniques for data collection, archiving, integration, annotation and discovery will be needed
- Peer-to-peer analysis and data product publishing to development of systems for automated analysis, discovery and annotation
- Automated hypothesis creation tools for pattern detection, and capable of suggesting relationships
- In areas like Biology this will enable a new type of “data driven” theory to make powerful, testable predictions in areas previously impossible with analytical theory (e.g. protein structure prediction)

# Molecular Biology Laboratory-on-a-Chip

## What could you do with 10,000 of these?



# The Personal Laboratory

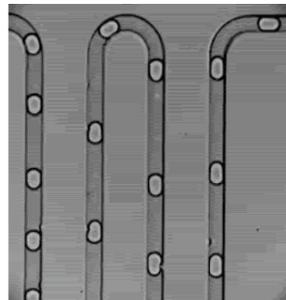
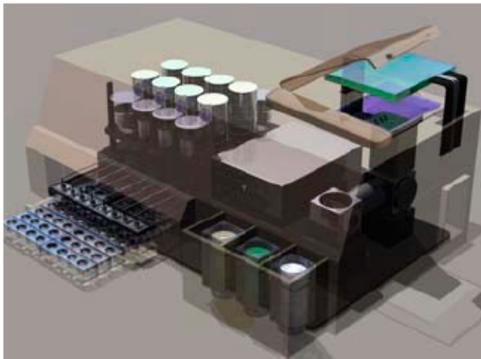


The **Personal Laboratory System™** (PLS) is the fluid handling equivalent of a liquid computer that can be programmed to create, test, and sort samples on inexpensive, disposable NanoReactor™ Chips. Using RDT's proprietary NanoReactor™ technology, the PLS™ performs a wide range of analytical functions of interest to the life science and chemical industries. In a compact, bench-top design, the PLS incorporates standard optical assays and read-outs for data collection used in academic and industrial laboratories.

**NanoReactor™ Chips** containing an array of on-chip fluid handling modules are "liquid circuit boards" that can be configured for a variety of biochemical and cell-based assays including:

- Enzyme & cell based screening
- RNAi & RNAi combination screening
- Protein & antibody engineering
- On-chip chemical synthesis & screening
- Pharmaceutical formulation

In addition to disposable Chips, RDT supplies users of the PLS with a variety of standard and custom reagent kits.



# Communities and Partners

- Communities

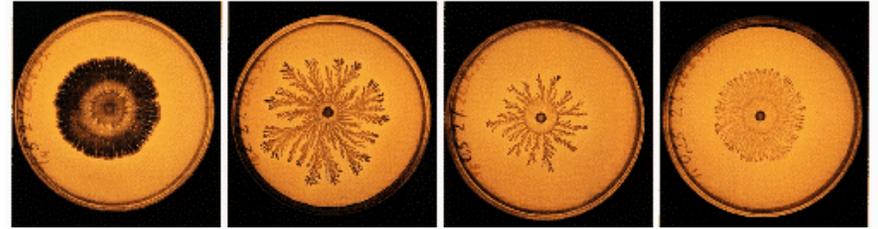
- Computational Genomics
  - NIH BRCs
  - NSF Genomics
  - DOE JGI
  - NIH NCBI
- Infectious Disease
  - NIH RCEs
- Molecular Evolution
  - NSF Tree of Life
- Bioinformatics Tools
  - NSF, and NIH NCRR
- Neuroscience
  - NIH NIMH

- Partners

- University of Chicago
- Argonne National Laboratory
- Fellowship for Interpretation of Genomes
- Universität Bielefeld
- INRA, France
- UCSD BIRN, NBCR
- VBI, Virginia Tech
- TIGR
- GGF - Life Science WG
- IBM
- Apple



# Conclusions



- Biology is well positioned to co-dominate computing applications for the next several decades
- Biological and Biomedical applications of computing 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 high-performance computing are well suited for each other and IBM's Blue Gene initiative is aimed at this intersection of interests

# Acknowledgements

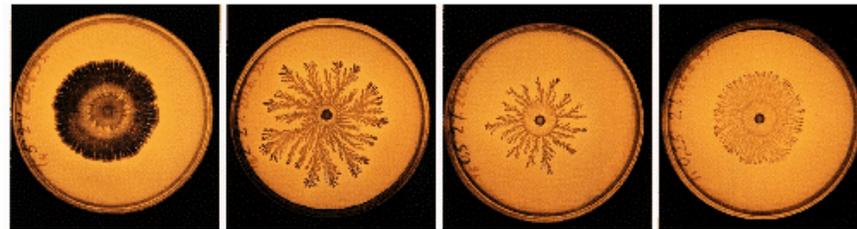
---

- Many thanks to DOE, NSF, NIH, DOD, ANL, UC, Apple, Microsoft and IBM for supporting my research group over the years



Describe

Explain



Predict

Control

