Update on Joint ASCAC-BERAC Panel on Modeling for GTL

Rick Stevens

John Wooley

Participants

Panel Members:

Michael Banda — LBNL

Thomas Zacharia — ORNL

David Galas — ISB/Battelle

Rick Stevens — ANL/UChicago

*John Wooley — UCSD

*David Kingsbury — Moore Foundation

*Keith Hodgson — Stanford

*Barbara Wold — Caltech

*Chris Somerville — Stanford

Invited Presenters:

Nitin Baliga — ISB/U Washington

Rich Bonneau — NYU/Courant

Paramvir Dehal — LBNL/UCB

Justin Donato — U Wisconsin

Thierry Emonet — Yale University

Adam Feist — UCSD

Mick Follows — MIT

Peter Karp — SRI

Harley McAdams — Stanford

Sue Rhee — Carnegie Institution

Nagiza Samatova — ORNL

* BERAC members

The Subcommittee Charge

- Convene a joint panel with BERAC to examine the issue of computational models for GTL, including:
- How progress could be accelerated through targeted investments in applied mathematics, and computer science and how these can be incorporated to meet the needs of computational biology.
- The joint panel should consider whether the current ASCR long-term goal is too ambitious, given the status and level of buy-in from the community.

  > "By 2015, demonstrated progress toward developing through the Genomes to Life partnership with the Biological and Environmental Research program, the computational science capability to model a complete microbe and simple microbial community."

- It needs to consider what is happening in the computational-science and life-sciences communities. It should discuss possible intermediate goals that might be more relevant to the two programs.
- And it should identify the key computational obstacles to developing computer models of the major biological understandings necessary to characterize and engineer microbes for DOE missions, such as biofuels and bioremediation.

Status of the “Modeling in GTL” Report

- Preliminary findings and recommendations
  - These are being revised by the joint panel over the next few weeks
- Preparing background material for the report from the panel presentations
  - 10-15 page summary to provide context for the findings and recommendations
- Generating linkages to two important NRC reports that impact the modeling charge
  - The role of theory in advancing 21st Century Biology (Galas et. al.)
  - Catalyzing Inquiry at the Interface of Computing and Biology (Wooley, Lin et. al.)

Computational Modeling and Simulation as Enablers for Biological Discovery

Some Ways Models are Useful in Biology

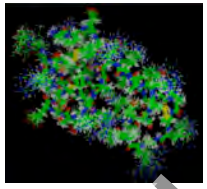
- Models Provide a Coherent Framework for Interpreting Data
- Models Highlight Basic Concepts of Wide Applicability
- Models Uncover New Phenomena or Concepts to Explore
- Models Identify Key Factors or Components of a System
- Models Can Link Levels of Detail (Individual to Population)
- Models Enable the Formalization of Intuitive Understandings
- Models Can Be Used as a Tool for Helping to Screen Unpromising Hypotheses
- Models Inform Experimental Design
- Models Can Predict Variables Inaccessible to Measurement
- Models Can Link What Is Known to What Is Yet Unknown
- Models Can Be Used to Generate Accurate Quantitative Predictions
- Models Expand the Range of Questions That Can Meaningfully Be Asked

From the NRC report "Catalyzing Inquiry at the Interface of Computing and Biology"

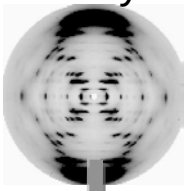
Data-driven Predictive Model Building

Genomics

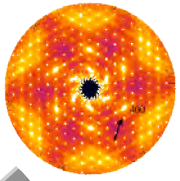
NMR



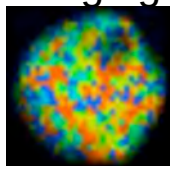
X-ray



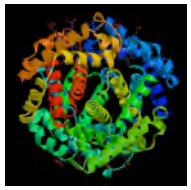
Neutron Scattering



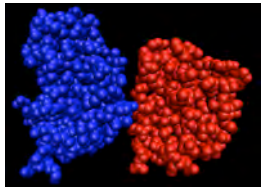
Imaging



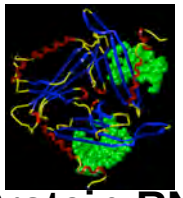
Structural Models



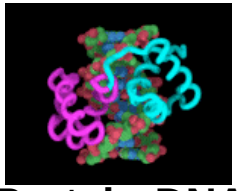
3-d Structure



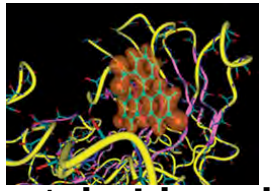
Protein Docking



Protein-RNA

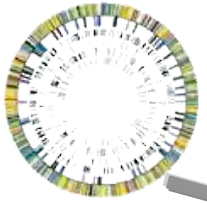


Protein-DNA

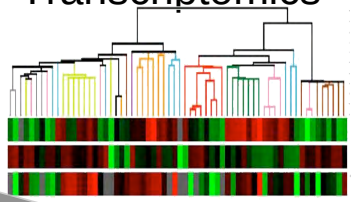


Protein-Ligand

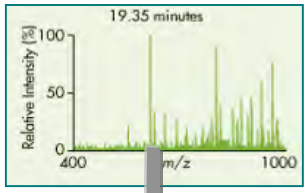
Genomics



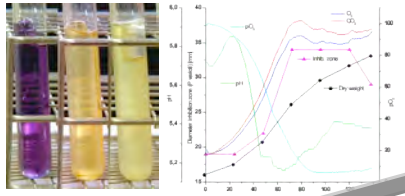
Transcriptomics



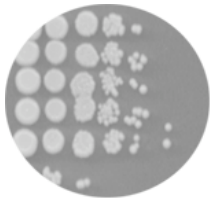
Quantitative Proteomics



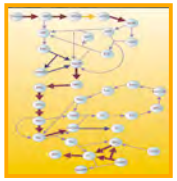
Metabolomics



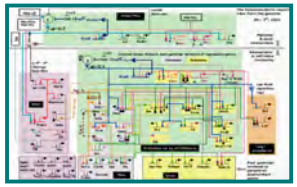
Interactomics



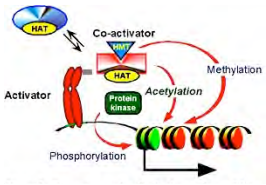
Network/Pathway Models



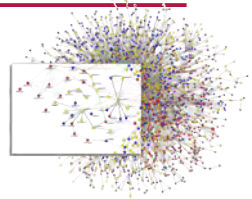
Metabolic



Regulatory



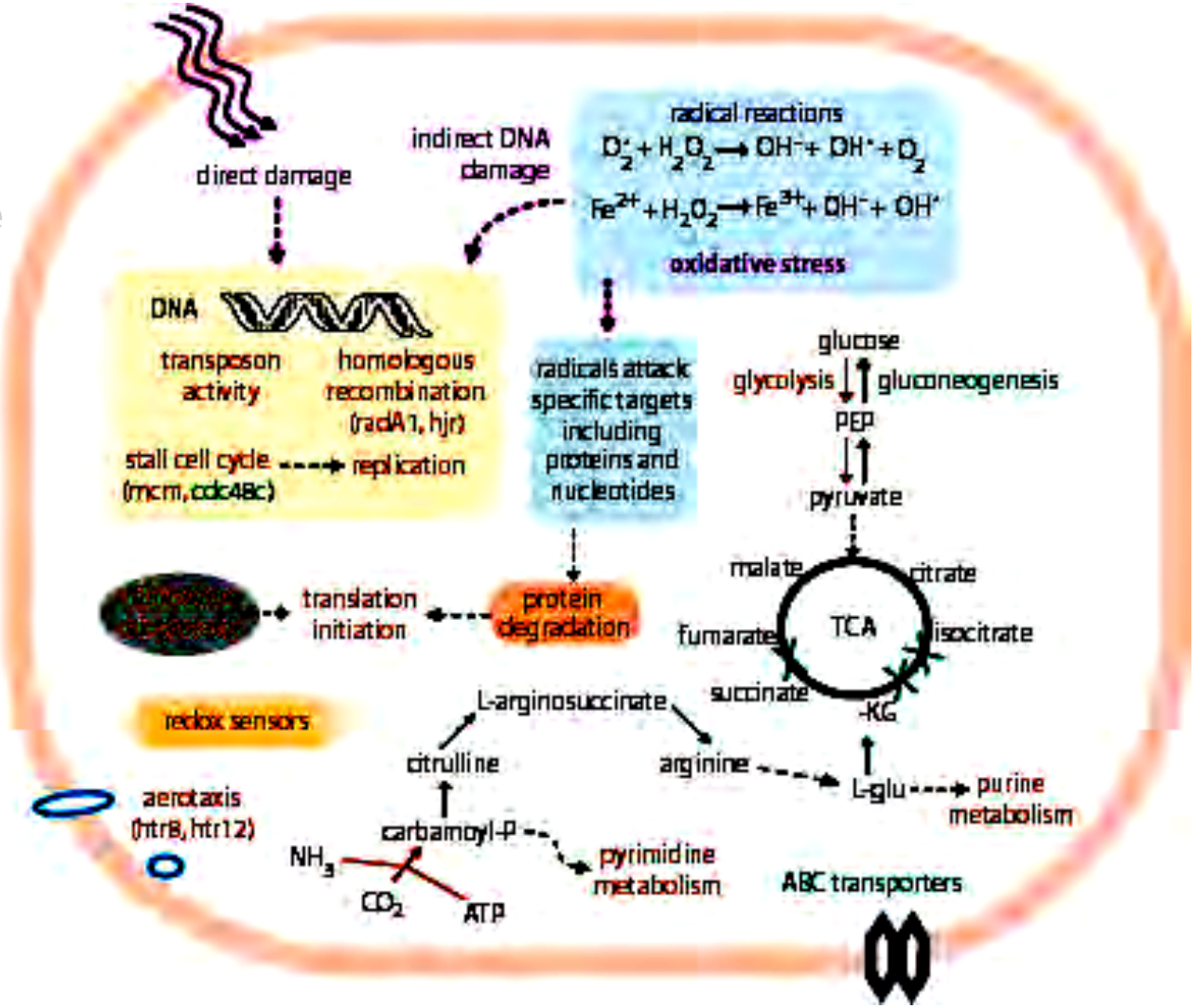
Signaling



Protein Interaction

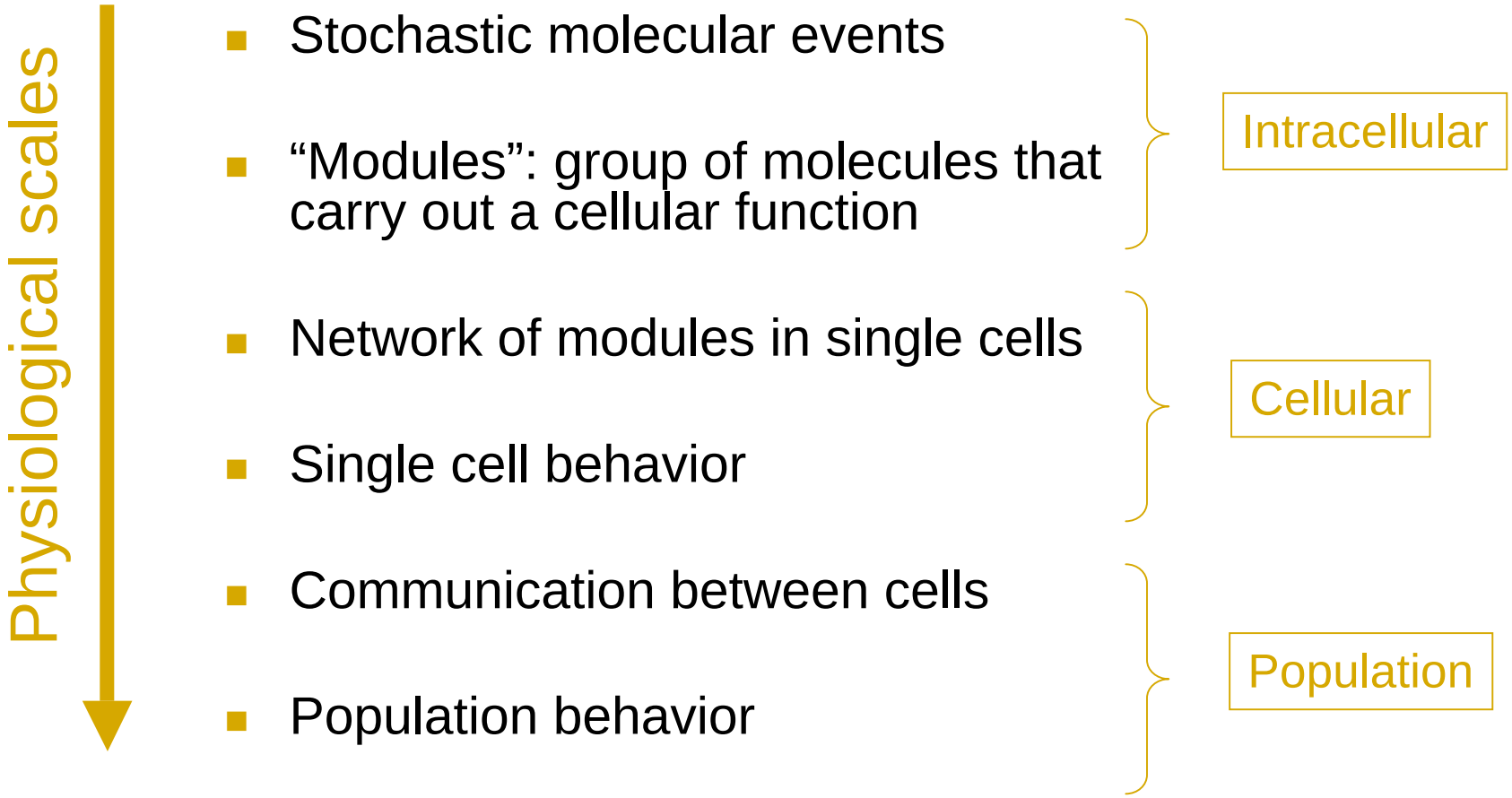
Response to ^{60}Co-γ rays: a static model

- **Descriptive**
  - **›Static**
  - ›Dynamic
- Quantitative
- Mechanistic

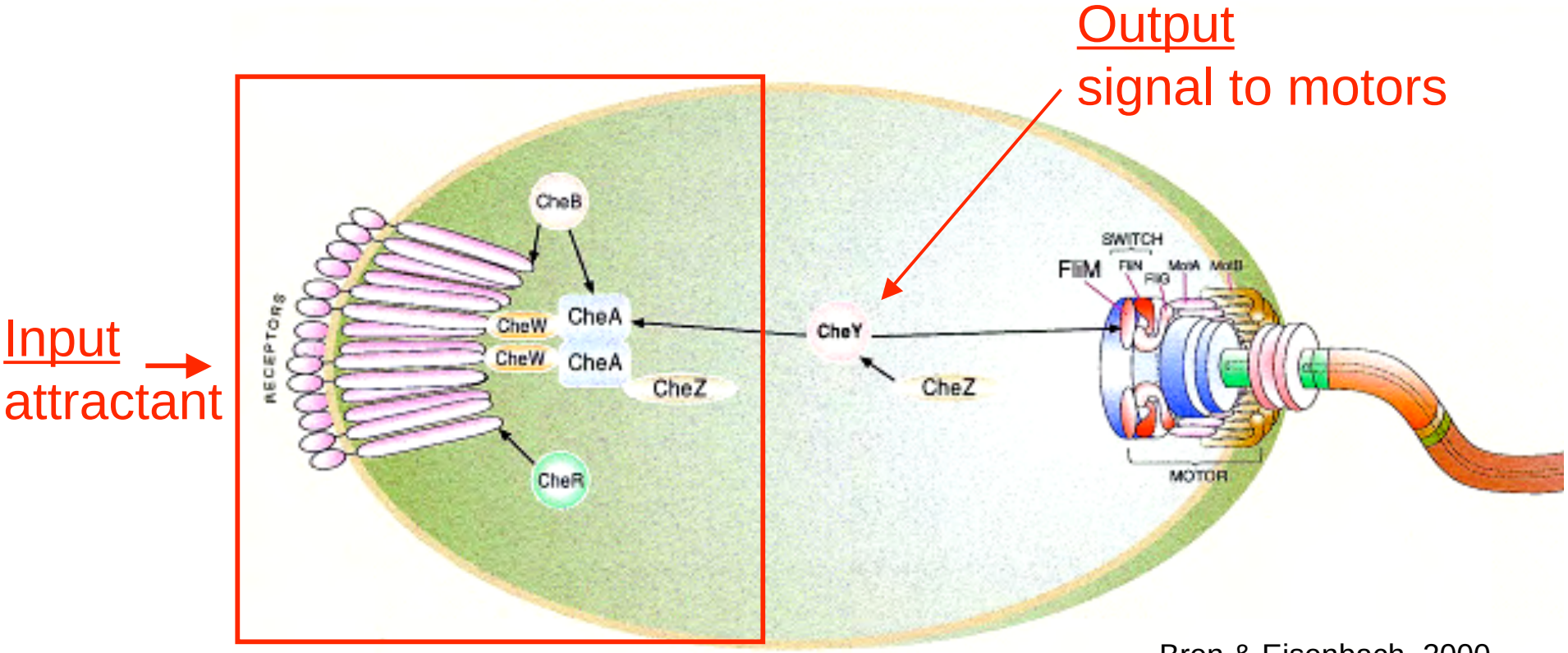


Whitehead et al. Nature MSB (2006)

From molecules to population behavior



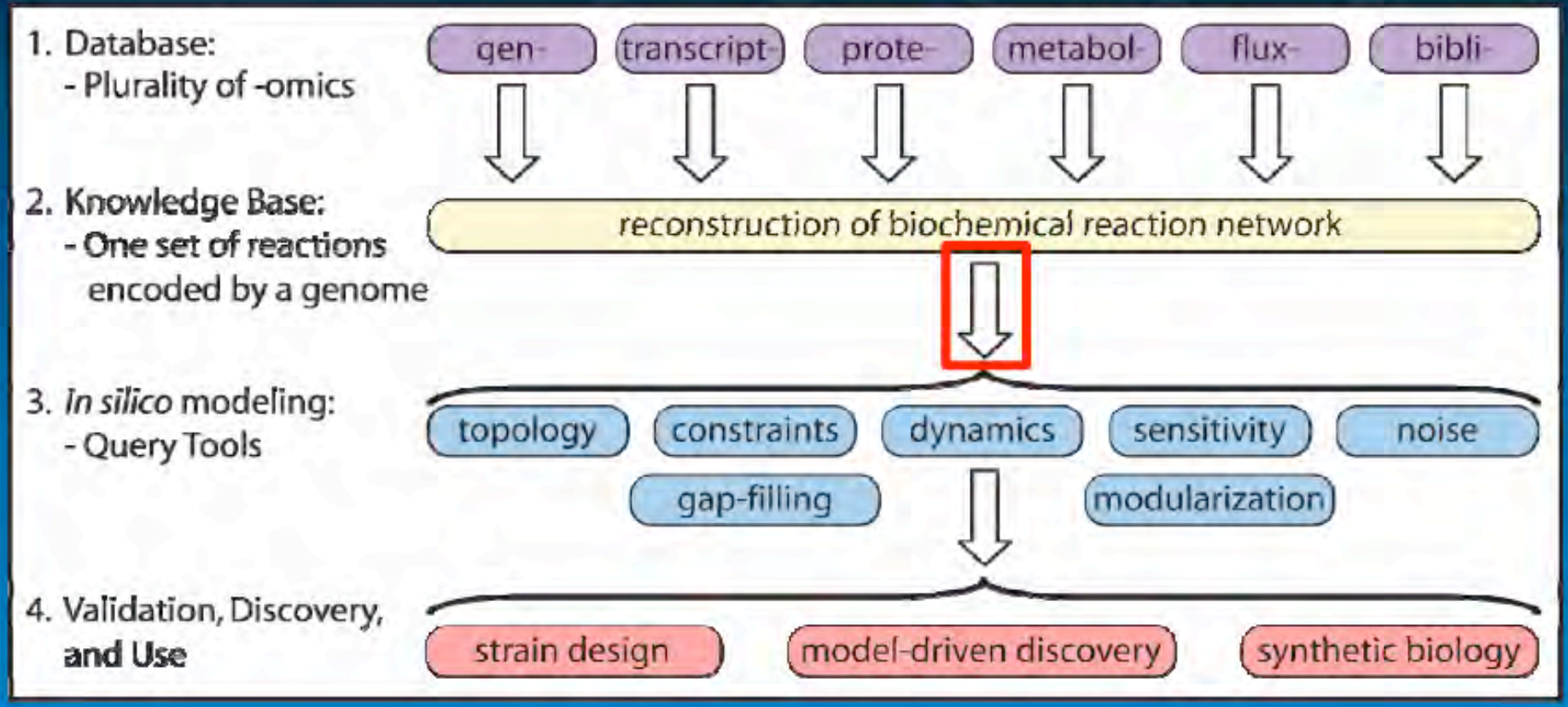
Bacterial chemotaxis sensory system

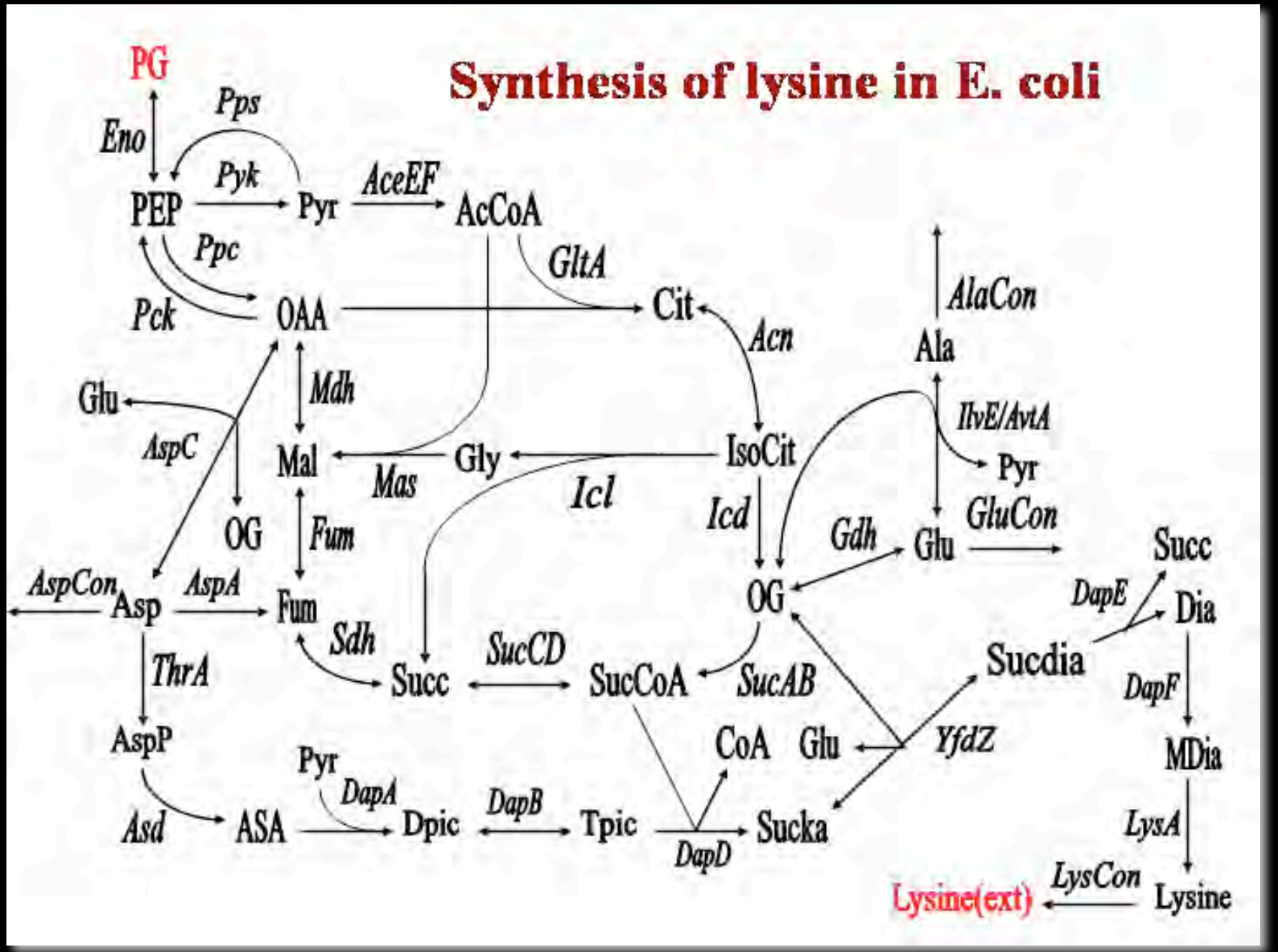


Bren & Eisenbach, 2000

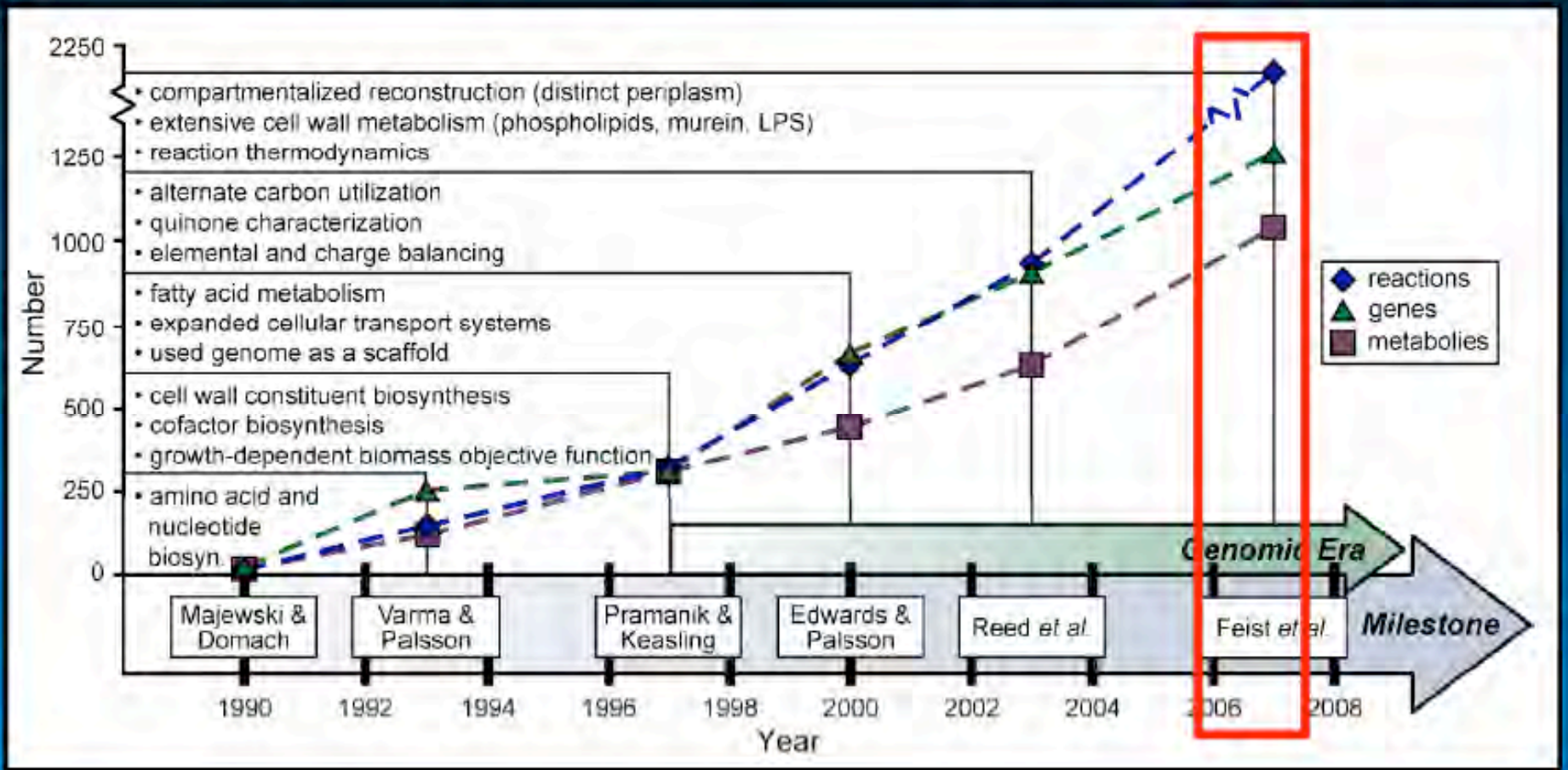
Information processing unit
Chemotaxis network

Systems Biology as a 4 Step Process




Synthesis of lysine in E. coli
PG
Eno
Pps
Pyk
PEP
Pyr
AceEF
AcCoA
Ppc
Pck
OAA
GltA
Cit
Acn
AlaCon
Ala
Glu
Mdh
IlvE/AvtA
Pyr
AspC
Mal
Mas
Gly
Icl
IsoCit
Icd
OG
Fum
Gdh
Glu
GluCon
Succ
AspCon
Asp
AspA
Fum
OG
DapE
Dia
Sdh
SucCD
SucCoA
SucAB
Succ
Sucdia
ThrA
DapF
AspP
CoA
Glu
YfdZ
MDia
Pyr
DapA
DapB
Asd
ASA
Dpic
Tpic
Sucka
DapD
LysA
LysCon
Lysine(ext)
Lysine

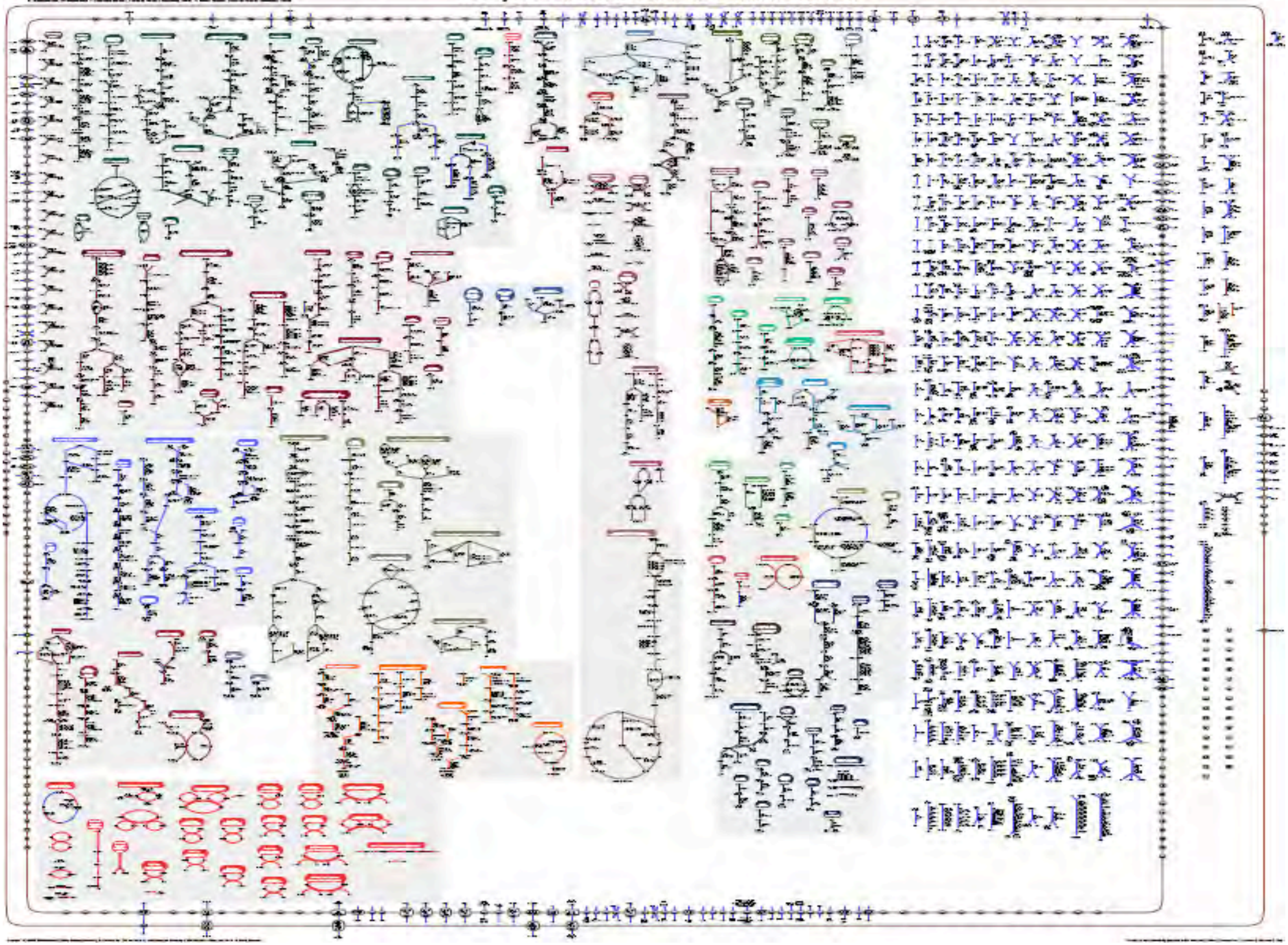
Reconstruction of *E. coli* Metabolism



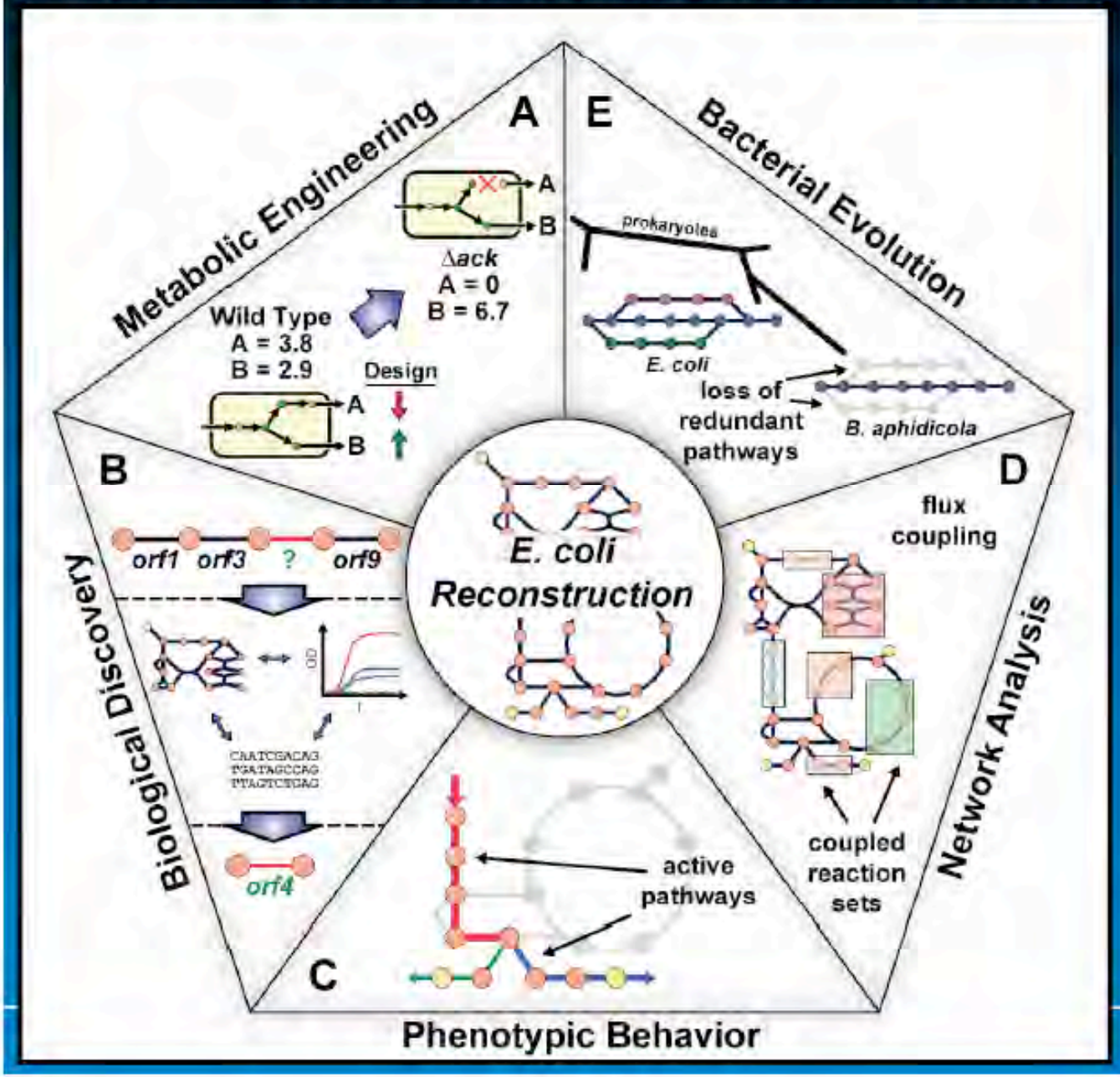
References:

R.A. Majewski and M.M. Domach, *Biotechnol Bioeng* **35**, 732 (1990)
A. Varma, B.W. Boesch, and B.O. Palsson, *Appl Environ Microbiol* **59** (8), 2465 (1993) & *Biotechnol Bioeng* **42** (1), 59 (1993)
J. Pramanik and J.D. Keasling, *Biotechnol Bioeng* **56** (4), 398 (1997) & **60** (2), 230 (1998)
J.S. Edwards and B.O. Palsson, *Proc Natl Acad Sci U S A.* **97** (10), 5528 (2000)
J.L. Reed, T.D. Vo, C.H. Schilling et al., *Genome Biology* **4** (9), R54.1 (2003)
A.M. Feist, C.S. Henry, P. Karp, V. Hatzimanikatis, B.O. Palsson et al., *Mol Sys Biol 3* (2007)

EcoCyc: *Escherichia coli* K-12 Cellular Overview



Uses of the *E. coli* Reconstruction



Metabolic Engineering:

1. *Biotechnol Bioeng* **84**, 647 (2003)
2. *Biotechnol Bioeng* **84**, 887 (2003)
3. *Genome Res* **14**, 2367 (2004)
4. *Metab Eng* **7**, 155 (2005)
5. *Nat Biotechnol* **23**, 612 (2005)
6. *Appl Environ Microbiol* **71**, 7880 (2005)
7. *Metab Eng* **8**, 1 (2006)
8. *Appl Microbiol Biotechnol* **V73**, 887 (2006)
9. *Biotechnol Bioeng* **91**, 643 (2005)
10. *Proc Natl Acad Sci U S A* **104** 7797(2007)

Biological Discovery:

11. *Genome Biology* **7**, R17 (2006)
12. *BMC Bioinformatics* **7**, 177 (2006)
13. *Proc Natl Acad Sci U S A* **103**, 17480 (2006)
14. *Nature* **429**, 92 (2004)
15. *PLoS Computational Biology* **2**, e72 (2006)

Phenotypic Behavior:

16. *Mol Syst Biol*, *3* 121 (2007)
17. *Biophys J* 83, 79 (2002)
18. *Proc Natl Acad Sci U S A* **99**, 15112 (2002)
19. *Bioinformatics* **21**, 2008 (2005)
20. *Proc Natl Acad Sci U S A* **102**, 7695 (2005)
21. *J theor Biol* **237**, 401 (2005)
22. *Biophys J* **90**, 1453 (2006)
23. *BMC Bioinformatics* **7**, 118 (2006)
24. *Mol Syst Biol* **2**, 2006 0034 (2006)
25. *Biophys J* **91**, 2304 (2006)
26. *BMC Bioinformatics* **7**, 512 (2006)
27. *Biophys J* **92**, 1792 (2007)
28. *Nature Genetics* **36**, 1056 (2004)
29. *Nat Biotechnol* **19**, 125 (2001)
30. *Nature* **420**, 186 (2002)
31. *Biochemistry (Mosc)* **71**, 1258 (2006)
32. *Curr Opin Biotechnol* **17**, 448 (2006)
33. *Biophys J* **92**, 1846 (2007)
34. *Biotechnol Prog* **17**, 791 (2001)
35. *BMC systems biology* **1** (2007).

Network Analysis:

36. *Bioinformatics* **19**, 1027 (2003)
37. *Genome Res* **14**, 301 (2004)
38. *Nature* **427**, 839 (2004)
39. *Biophys J* **88**, 37 (2005)
40. *PLoS Comput Biol* **1**, e68 (2005)
41. *BMC Bioinformatics* **7**, 111 (2006)
42. *Proc Natl Acad Sci U S A* 102, 19103 (2005)
43. *Biophys J* **90**, 2659 (2006)
44. *Bioinformatics* **23**, 92-98 (2007)
45. *Mol Syst Biol* **3**, 101 (2007)
46. Samal, A & Jain, S. *Under Review* (2007)

Bacterial Evolution:

47. *Bioinformatics* **21 Suppl 2**, ii222 (2005)
48. *Nat Genet* **37**, 1372 (2005)
49. *Nature* **440**, 667 (2006)

Regulation of transcription factors in E. coli

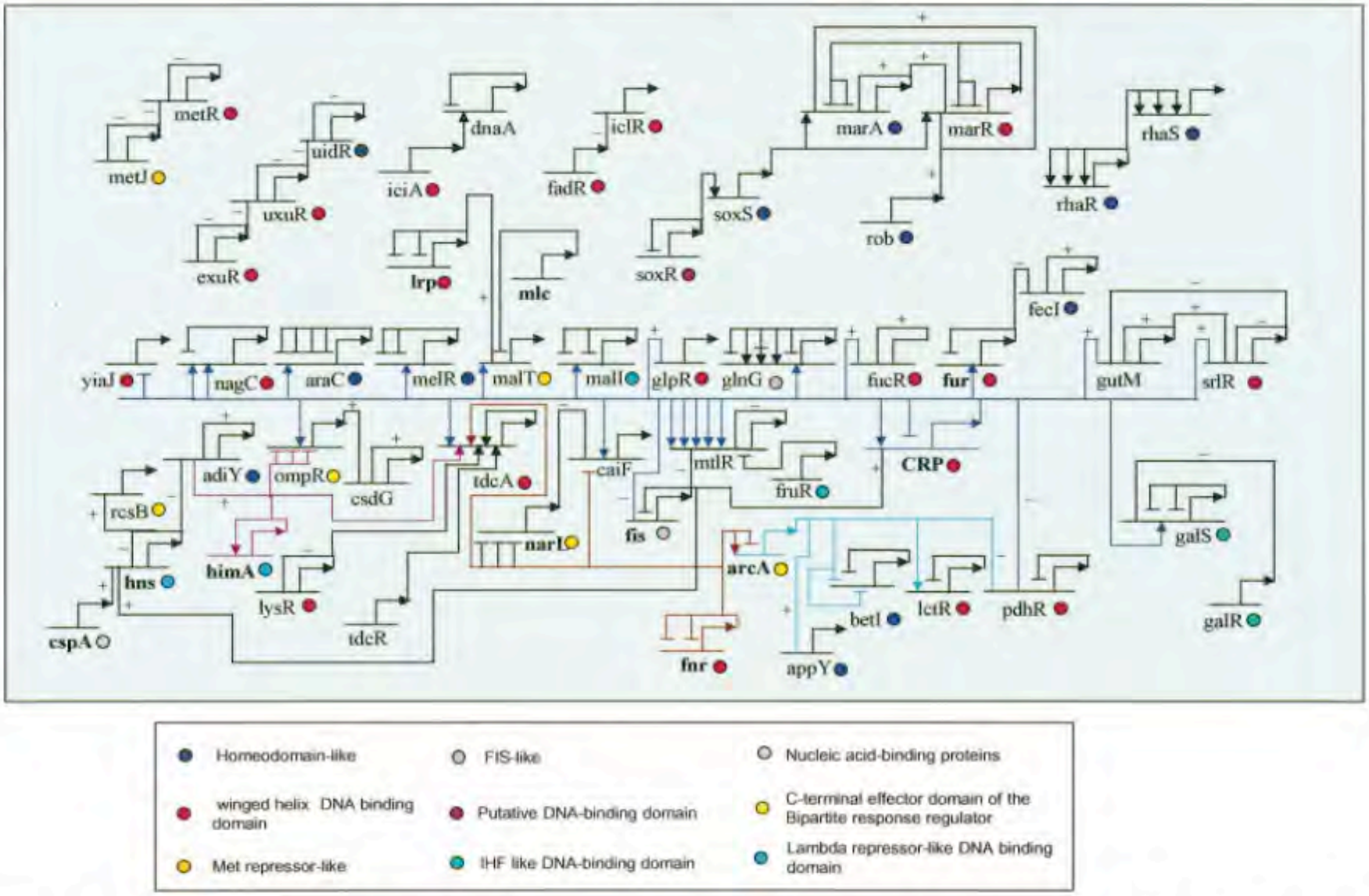
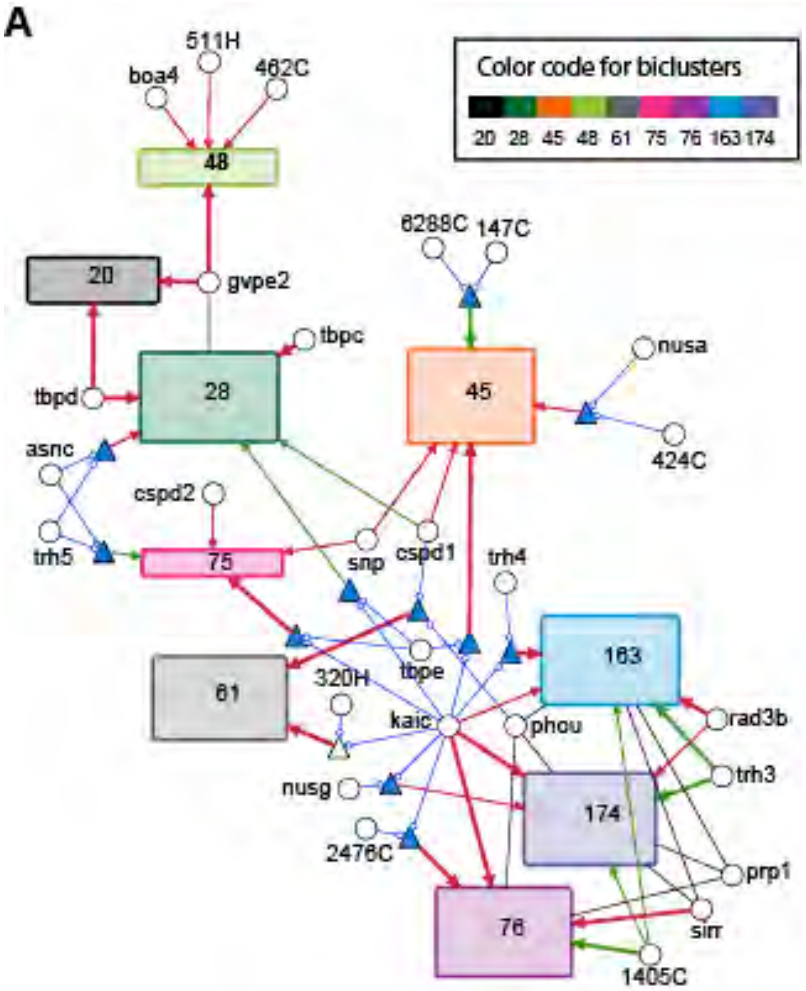
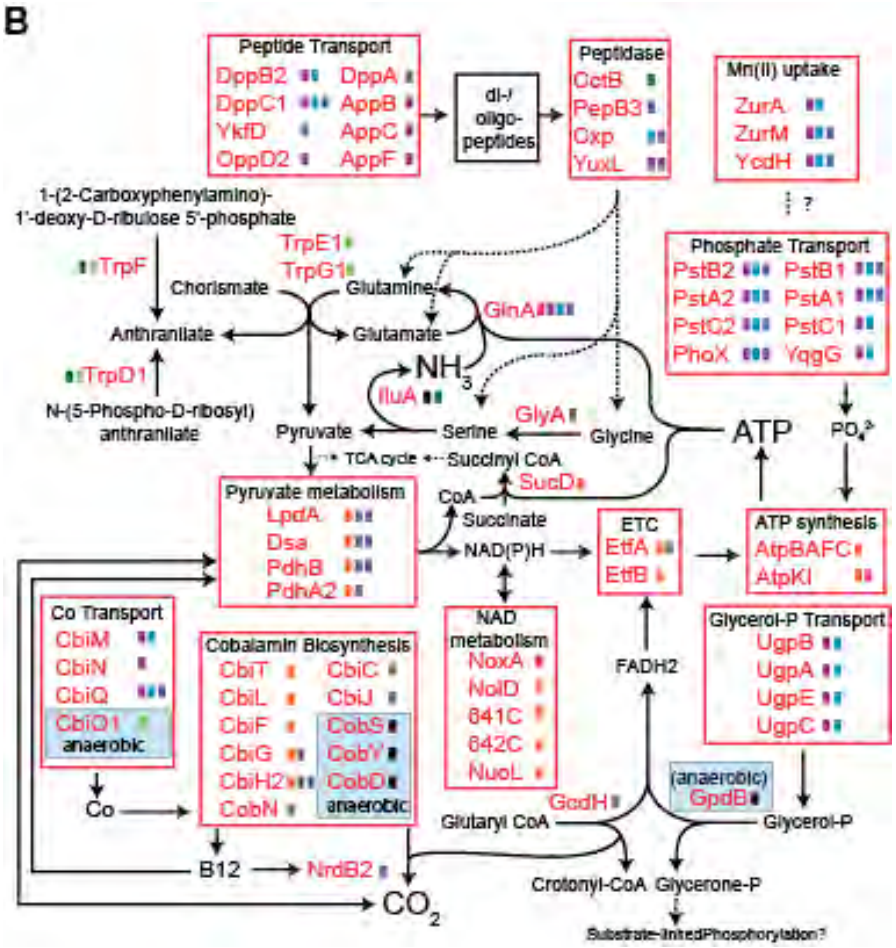


Figure 3. The transcription factor regulatory network in *E.coli*. When more than one transcription factor regulates a gene, the order of their binding sites is as given in the figure. An arrowhead is used to indicate positive regulation when the position of the binding site is known. A horizontal bar is used to indicate negative regulation when the position of the binding site is known. In cases where only the nature of regulation is known, without binding site information, + and – are used to indicate positive and negative regulation, respectively. These examples may be indirect rather than direct regulation. The DBD families are indicated by circles of different colours as given in the key. The names of global regulators are in bold.

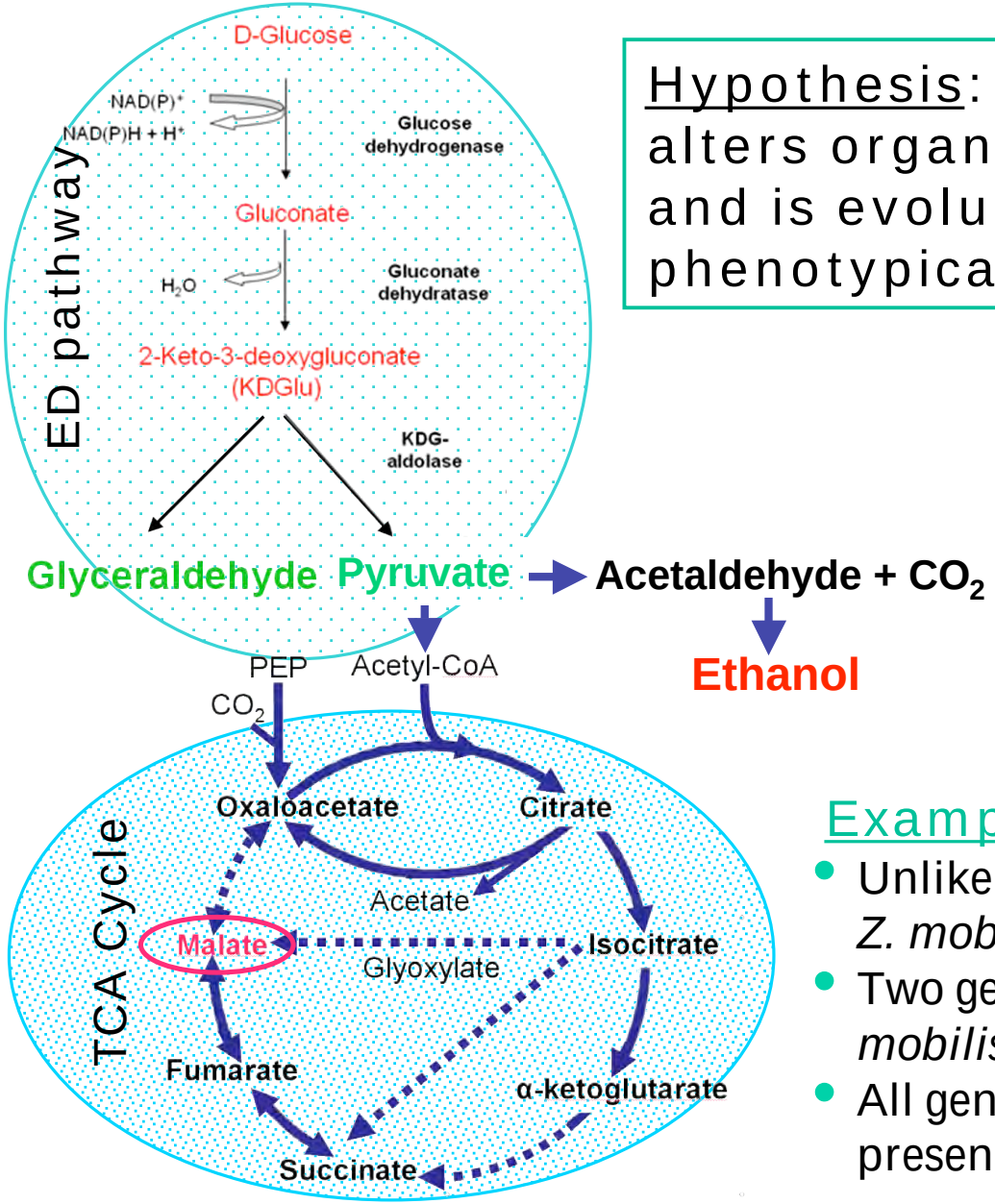
EGRIN models relationships among diverse cellular processes





Baliga Lab, submitted

What is the relationship between the structure of a pathway and its function?



Hypothesis: The topology of a pathway alters organismal phenotypic functions and is evolutionarily conserved across phenotypically similar genomes.

Why *Z. mobilis*?

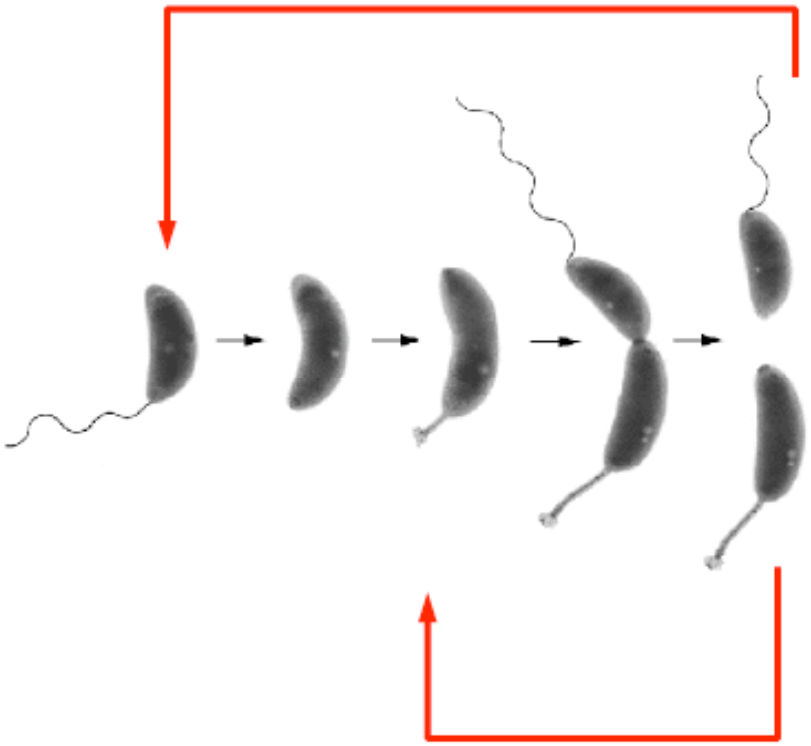
- Higher sugar uptake & ethanol yield
- Lower biomass production
- Higher ethanol tolerance
- Facultative anaerobic bacteria

Example Findings:

- Unlike EMP pathway in anaerobic bacteria, *Z. mobilis* utilizes ED pathway like aerobes
- Two genes (incl. *mdh*) are missing in *Z. mobilis* TCA cycle → low biomass.
- All genes except for 6-P-fructokinase are present in EMP pathway → inactive EMP.

Caulobacter crescentus . . . .

- Aquatic bacterium
- Lives in lakes and streams
- Feeds on plant degradation products
- ~4,000 genes



Swarmer daughter cell

Asymmetric cell division

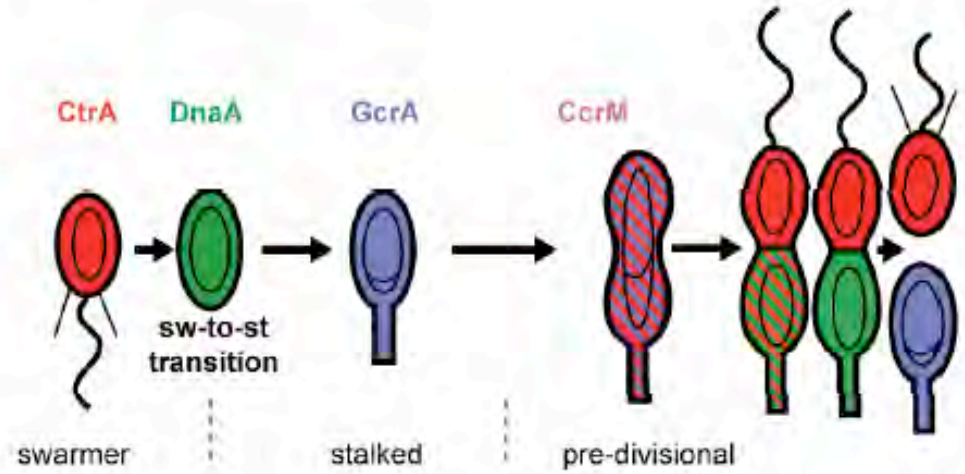
Stalked daughter cell

The daughter cells have distinctive, but coupled, cell cycles

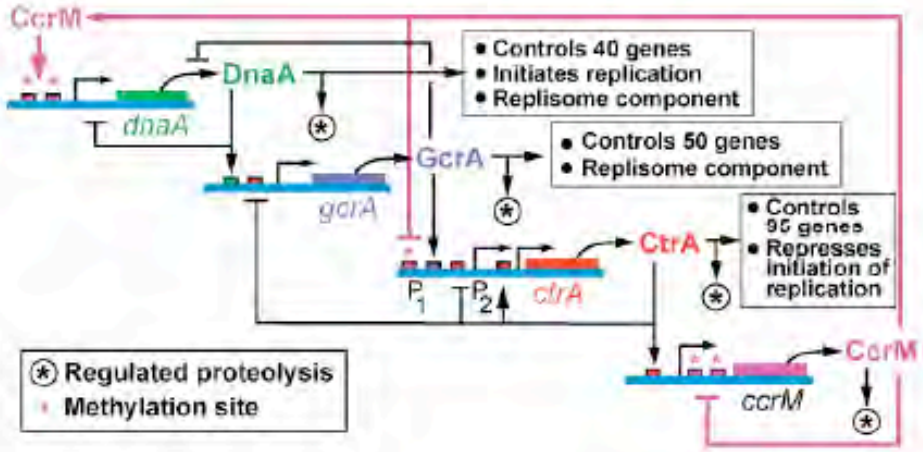
Model system for studying cell cycle control system and asymmetric cell division

More complete version of core circuitry

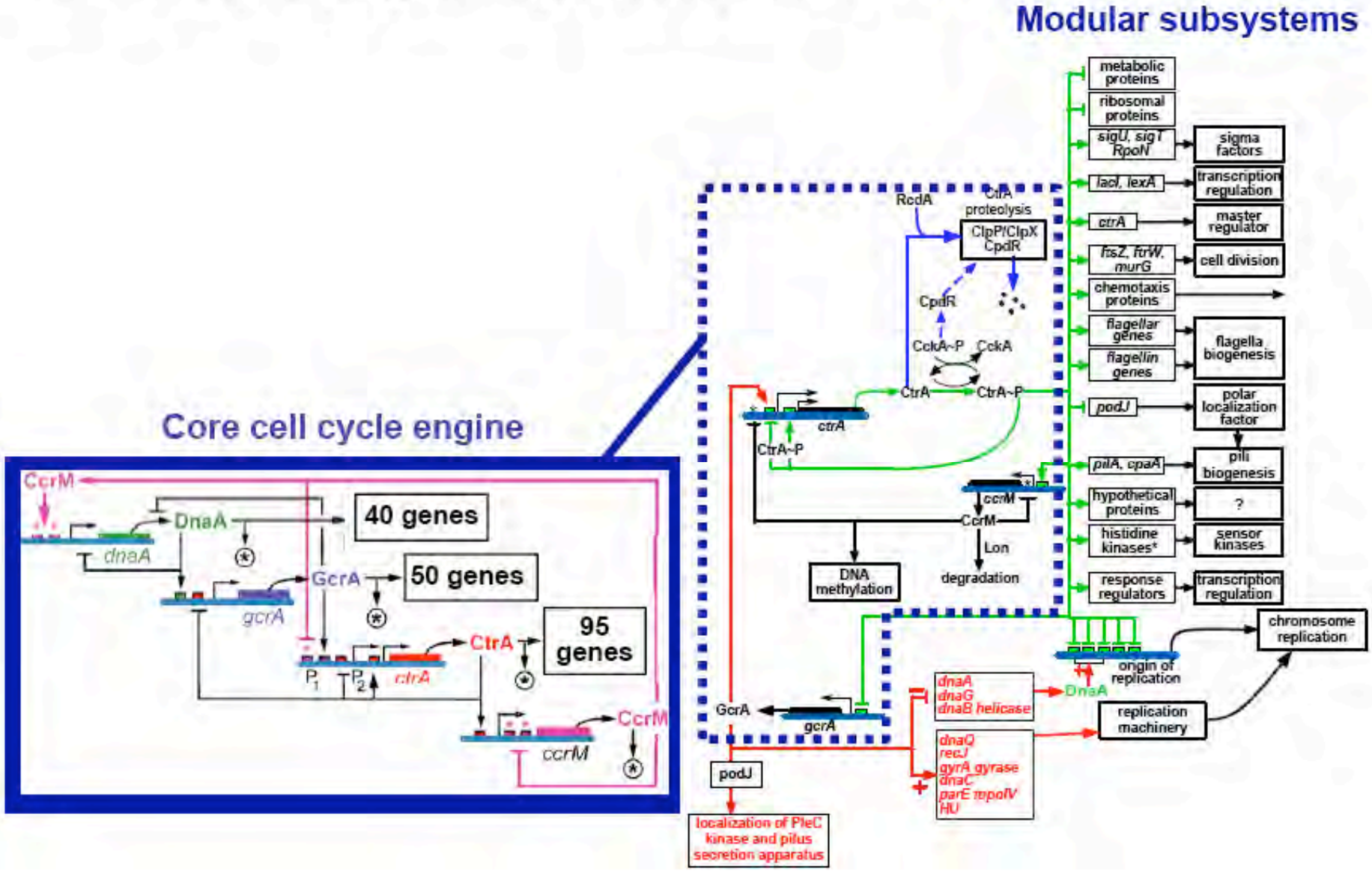
Four master regulator proteins



Complications:
- **Regulated proteolysis**
- **Methylation mechanism**

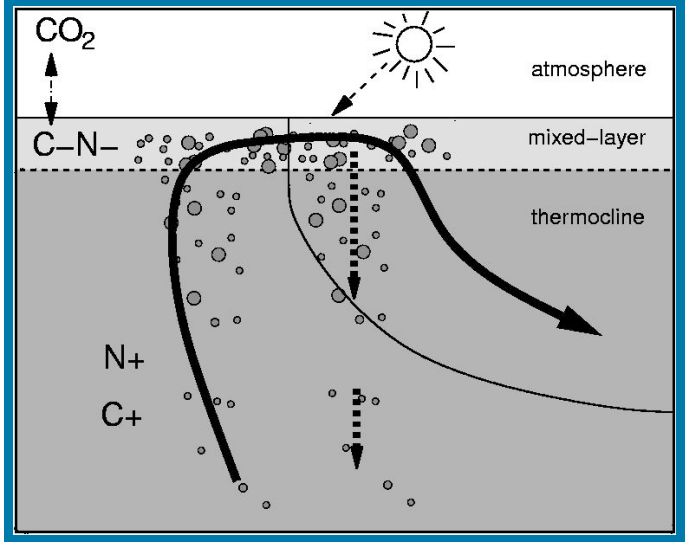


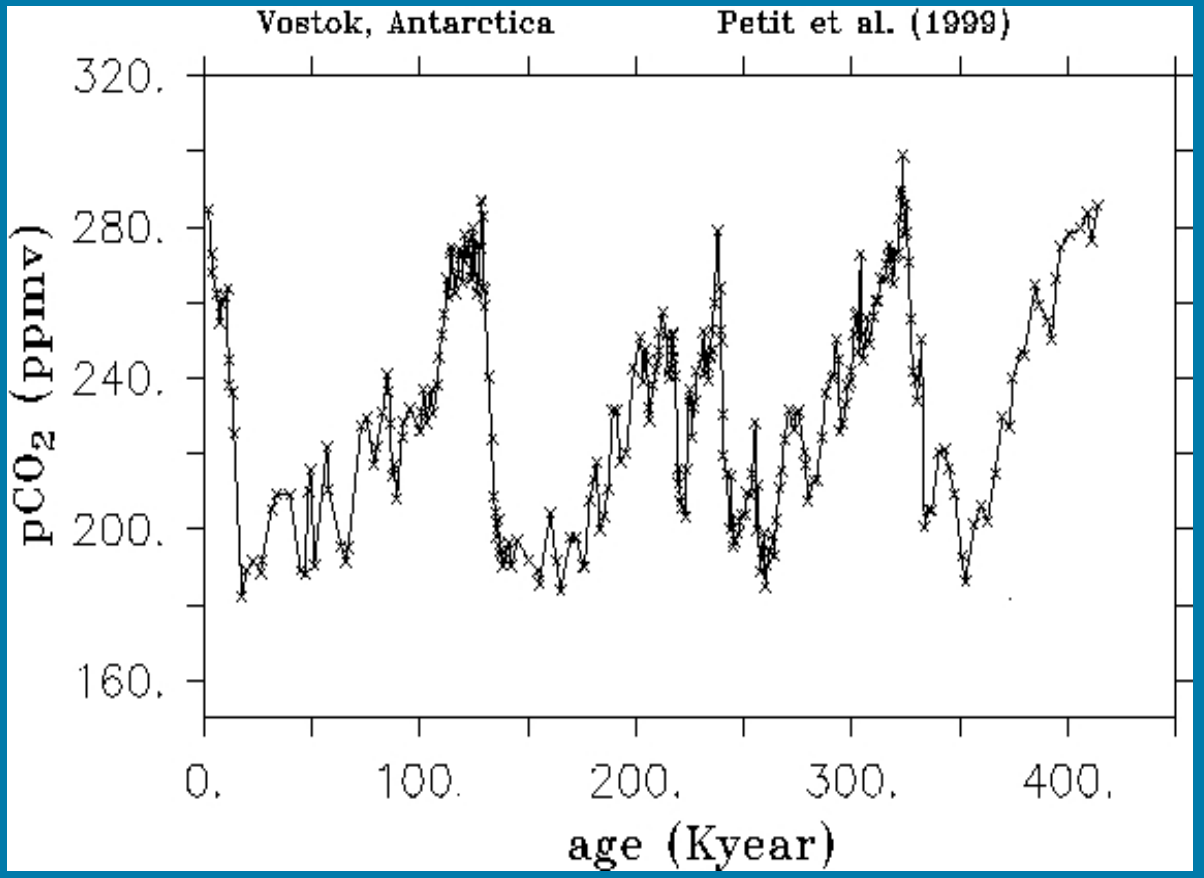
Hierarchical regulatory network controls the cell cycle



Carbon Cycle

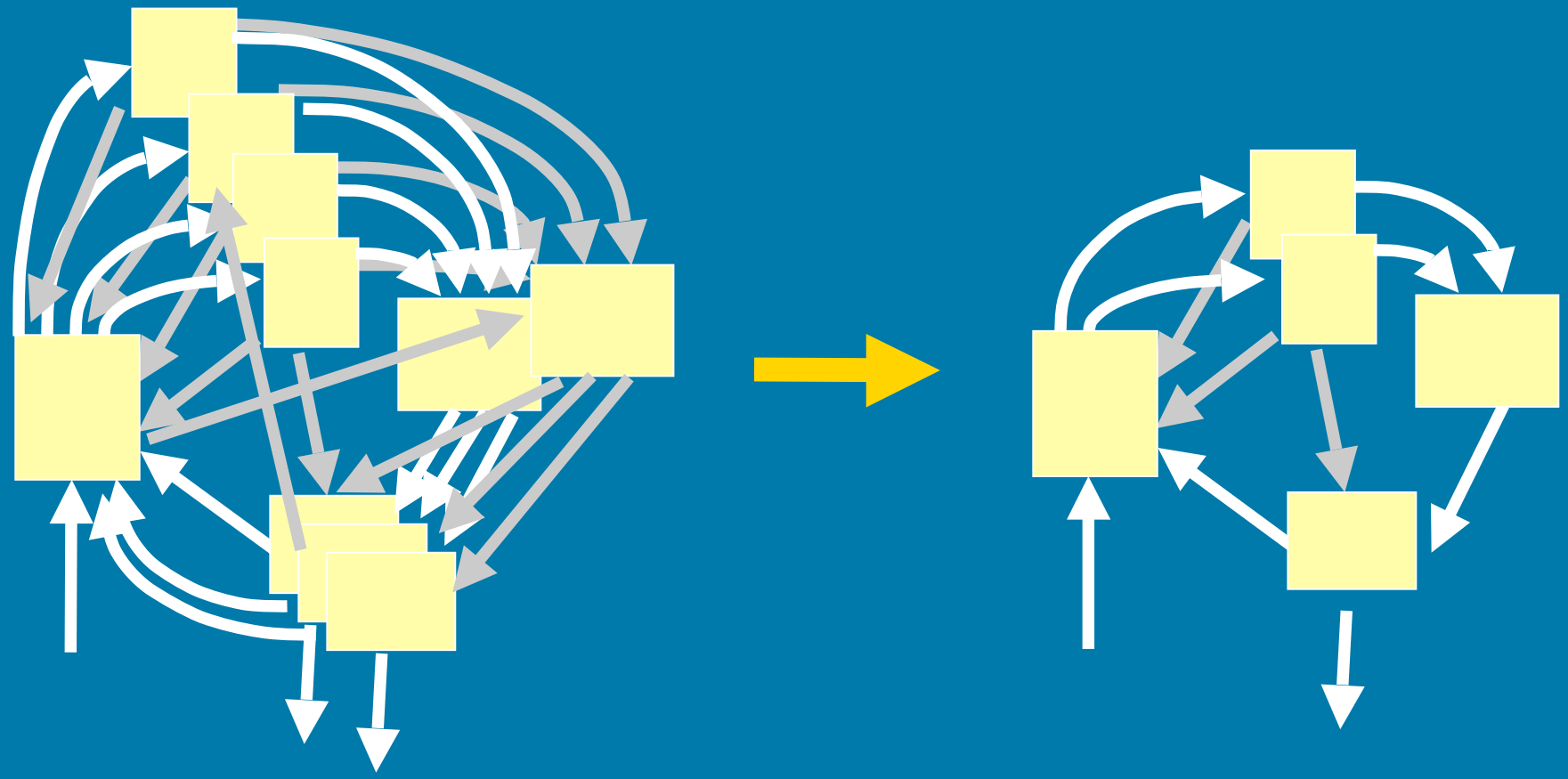
Efficiency of biological nutrient export regulates atmospheric CO_2





“Natural Selection” approach:

- initialize many potentially viable types
- allow system to self-organize …
- fittest physiologies (parameter combinations) succeed
- less fit physiologies “excluded”

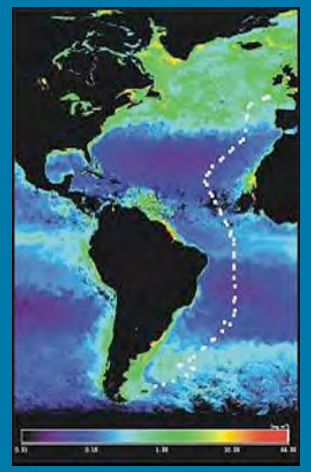


Complex initialized food web

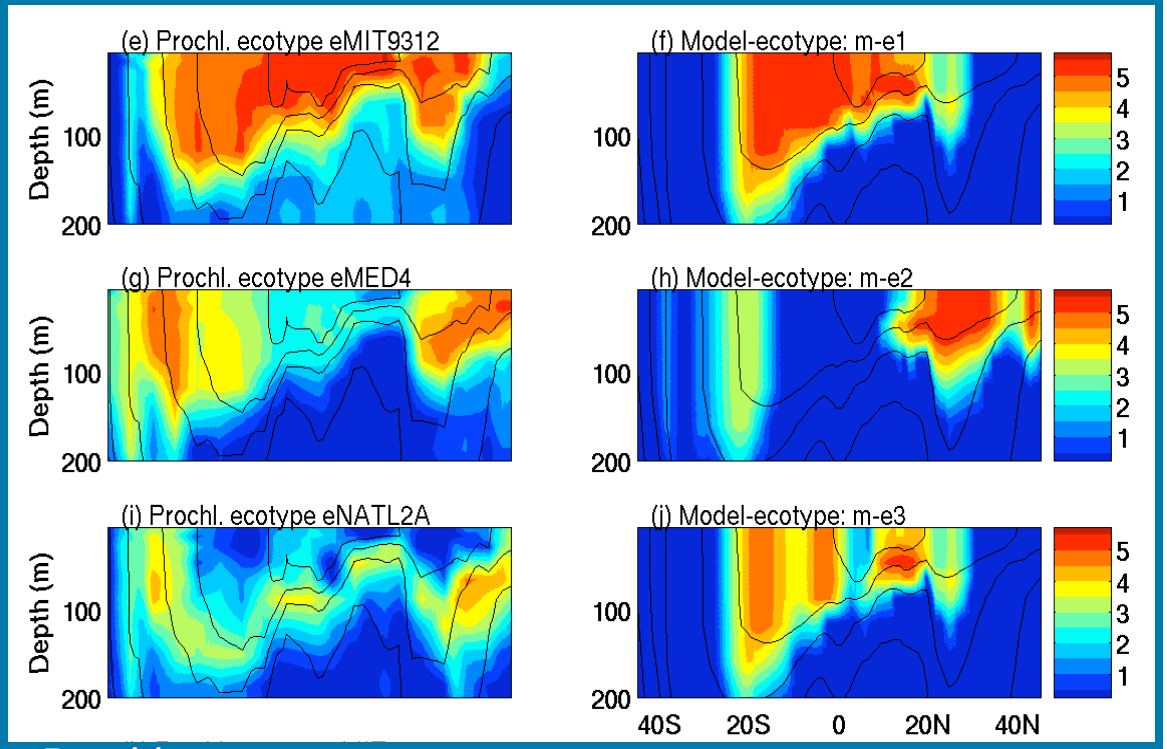
self-organized state

Emergent Biogeography of Microbial Communities in a Model Ocean

Michael J. Follows,[1*] Stephanie Dutkiewicz,[1] Scott Grant,[1,2] Sallie W. Chisholm[3]



- Plausible analogs of *Prochlorococcus* ecotypes present in solutions
- *Prochlorococcus* analogs defined by:
  - high surface area:vol
  - inability to utilize nitrate
  - appropriate light, T sensitivities selected



Prochlorococcus ecotypes AMT13 (Johnson et al., 2006)

model-ecotypes

Petascale Impact on Biological Theory

- Potential high impact on theory development
  - The ability to run large-scale simulations that can capture non-trivial variation in an evolutionary process could have a dramatic impact on our ability to move from qualitative to quantitative theory in biology
- Software readiness for petascale systems
  - While physical process oriented software is on a trajectory to achieve scalable performance on petascale systems, agent based evolution and ecosystem modeling environments are lagging far behind
  - Data analysis and bioinformatics environments are in the middle, hindered in part by the lack of data intensive infrastructure
- Capability and capacity computing estimates
  - First principles MD and QM simulations have enormous computing requirements, but perhaps limited impact on large-scale theory
  - Agent based simulations have not been effective scoped
- Related experimental support is needed
  - Validation experiments driven by the simulation and modeling will be required

Example Applications Ported to BG/L

- The following lists codes ported to date on BG/L evidencing the strong community interest and potential scientific ROI.

General Domain	Code	Institution
Astro Physics	**Enzo**	UCSD/SDSC
Astro Physics	**Flash**	UC/Argonne
Basic Physics	**CPS**	Columbia
Basic Physics	**QCD kernel**	IBM
Basic Physics	**QCD**	Argonne
Basic Physics	**QMC**	CalTech
Basic Physics	**QMC**	Argonne
BioChemistry	**BGC.5.0**	NCAR
BioChemistry	**BOB**	NCAR
CAE/FEM Stucture	**PAM-CRASH**	ESI
CFD	**Miranda**	LLNL
CFD	**Raptor**	LLNL
CFD	**SAGE**	LLNL
CFD	**TLBE**	LLNL
CFD	**sPPM**	LLNL
CFD	**mpcugles**	LLNL
CFD	**Nek5**	Argonne
CFD	**Enzo**	Argonne
CFD	**TLBE**	LLNL
Financial	**KOJAK**	NIC, Juelich
Financial	**Nissei**	NIWS
Finite Element Solvers	**HPCMW**	RIST
Fusion	**GTC**	PPPL
Fusion	**Nimrod**	Argonne
Fusion	**Gyro**	GA

General Doman	Code	Institution
Material Sciences	**ALE3D**	LLNL
Material Sciences	**LSMS**	LLNL
Molecular Biology	**mpiBLAST**	Argonne
Molecular dynamics	**MDCASK**	LLNL
Molecular Dynamics	**Amber**	UCSF
Molecular dynamics	**APBS**	UCSD
Molecular Dynamics	**Blue Matter**	IBM
Molecular Dynamics	**Charmm**	Harvard
Molecular dynamics	**LJMD**	CalTech
Molecular Dynamics	**NAMD**	UIUC/NCSA
Molecular Dynamics	**Qbox**	LLNL
Molecular Dynamics	**Shake & Bake**	Buffalo
Molecular Dynamics	**MDCASK**	LLNL
Molecular dynamics	**Paradis**	LLNL
Nano-Chemistry	**DL_POLY**	Argonne
Neuroscience	**pNEO**	Argonne
neutron transport	**SWEEP3D**	LArgonne
Nuclear Physics	**QMC**	Argonne
Quantum Chemistry	**CPMD**	IBM
Quantum Chemistry	**GAMESS**	Ames/Iowa State
Seismic wave propogatio	**SPECFEM3D**	GEOFRAMEWORK.org
Transport	**SPHOT**	LLNL
Transport	**UMT2K**	LLNL
Weather & Climate	**MM5**	NCAR
Weather & Climate	**POP**	Argonne

Example Petascale Biological Computations

- Simulating the assembly of the cellulosome
  - 14 enzyme extracellular complex responsible for efficient degradation of cellulose
    - *Initial modeling effort focused on understand the assembly and conformation of the complex in relation to the cellulose pseudo crystal*
- 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 Prokayotes
    - *~3000 protein families ⇒ for each one we want to build detailed gene phylogeny and reconcile with species tree*
    - *Thousands of phylogenies and tree reconciliations*

C. thermocellum ***Cellulosome***

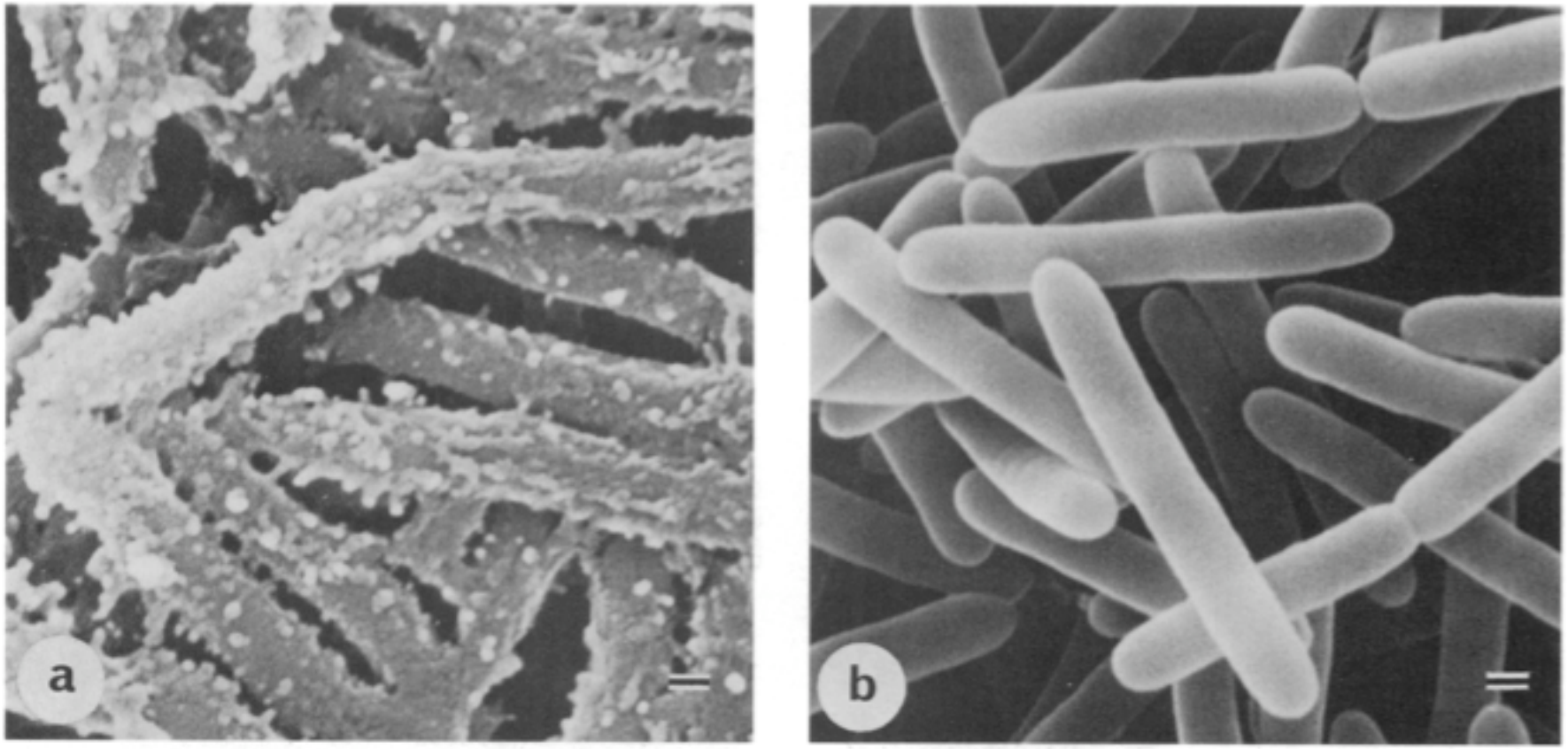
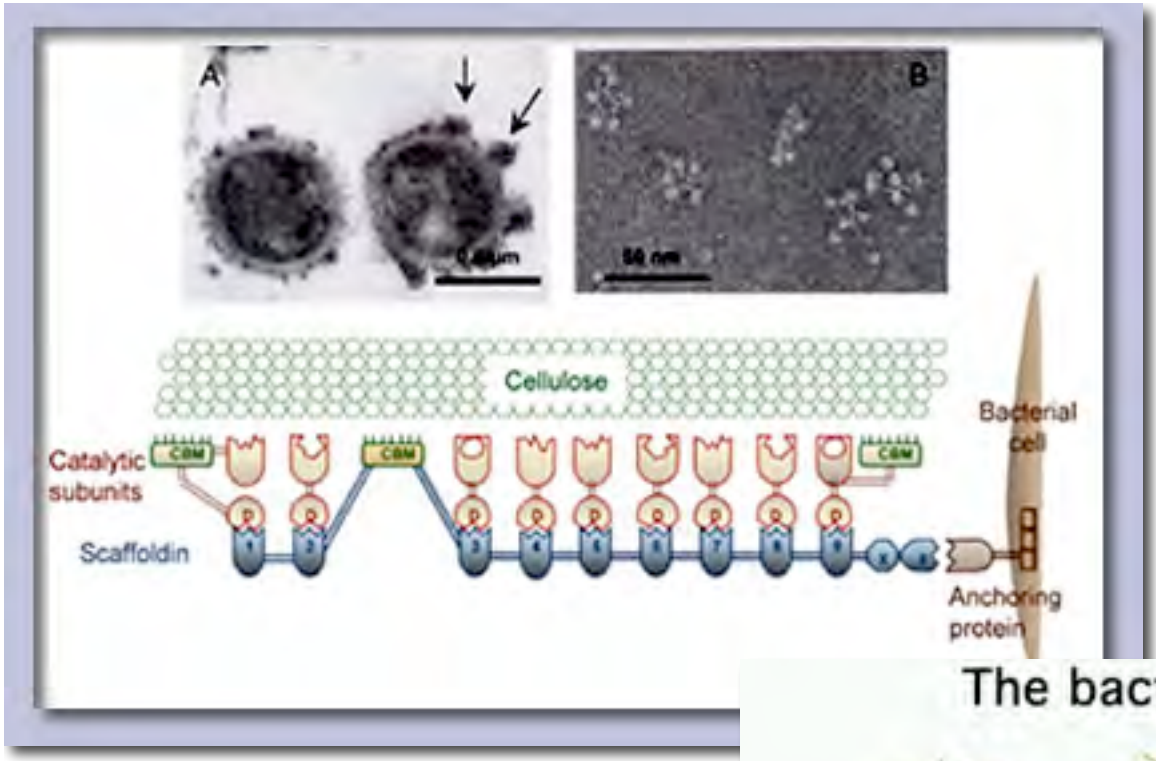
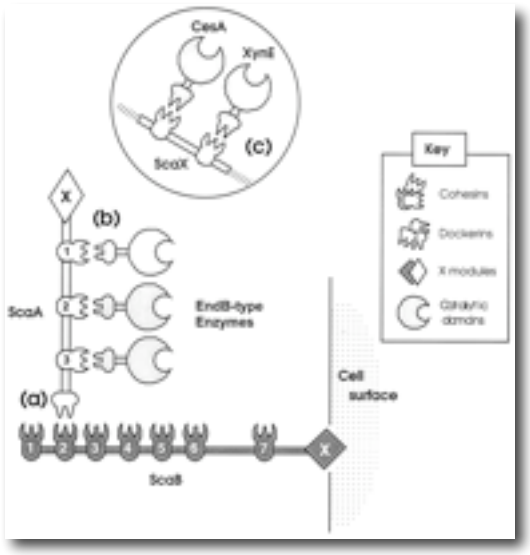
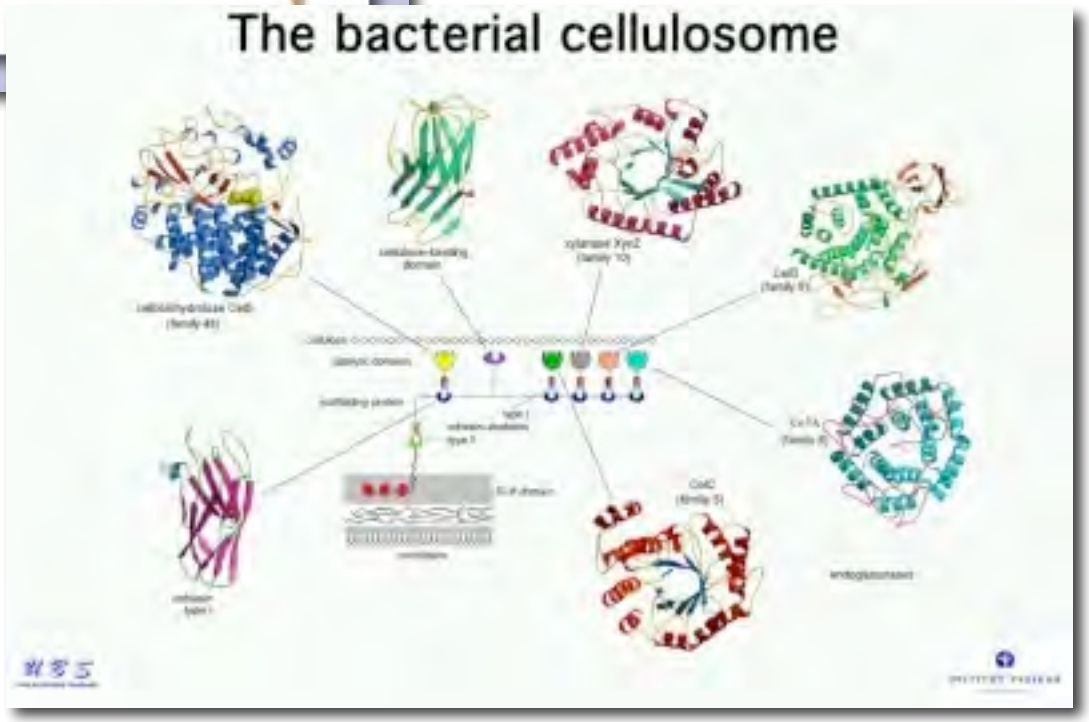


FIG. 4. SEM of CF-labeled cellobiose-grown cells of *C. thermocellum* YS. Cells were grown and labeled as described in the legend to Fig. 3 prior to processing. Wild-type YS cells (a) are easily distinguishable from mutant AD2 cells (b) by the appearance of protuberances which inundate the entire cell surface of the wild type. Bars, 200 nm.


A
B
50 nm
Cellulose
Catalytic subunits
CBM
Scaffoldin
Bacterial cell
Anchoring protein


CesA
XynE
ScaX
(c)
X
(b)
ScaA
EndB-type Enzymes
(a)
ScaB
Cell surface
Key
Cohesins
Dockerins
X modules
Catalytic domains

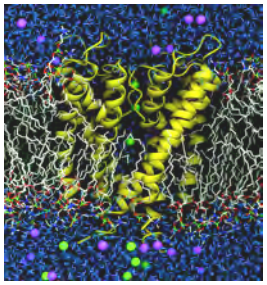

The bacterial cellulosome

Large Scale Simulations of Biomolecular Systems

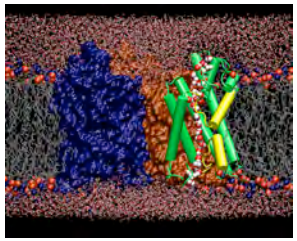
- **Integration of theory, modeling and simulations (TMS)**
  - Visualize the “workings” of biological molecular machines
  - Component of DOE Genomics: GTL Roadmap
- **All-atom classical molecular dynamics simulations**
  - CHARMM force field, PME, NPT ensemble
  - Modeling with Rosetta Approach
  - NAMD code, 2002 Gordon Bell Award
  - Fully scalable and parallel, Charm++
- **Benchmark system: Voltage-activated K+ channel**
  - Substantial function-altering conformational change of the protein triggered by externally-controlled voltage
  - Design of artificial switches in various nanotechnologies
  - Collaborators B. Roux (UC, ANL), K. Schulten, E. Tajkhorshid, J. Phillips (UIUC), V. Yarov (U of W)

Achievements and Current Situation

- ***Simulation studies of membrane proteins***
  - *Berneche & Roux (Nature 2001), Tajkhorshid et al (Science 2002), Noskov et al (Nature 2004), Tornroth-Horsefield et al (Nature 2006)*
  - *Review by Roux & Schulten (Structure 2004)*

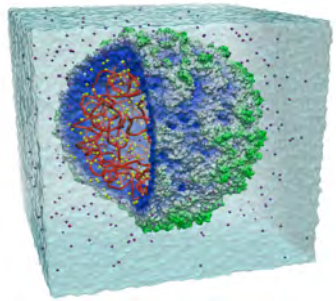


KcsA

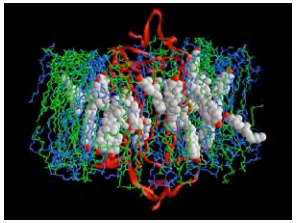


aquaporin

- **Breakthroughs with NAMD**
  - *Microsecond MD simulation of an entire life form*
  - *Mechanics of molecular machines*
  - *Microsecond simulations, millions of atoms*

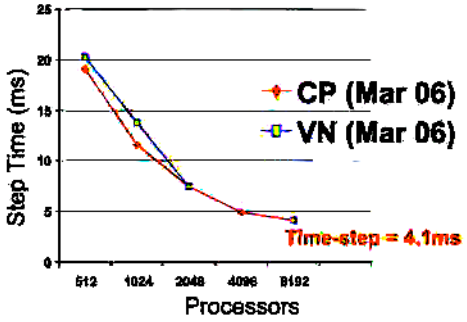


Satellite Tobacco Mosaic Virus



GPCR

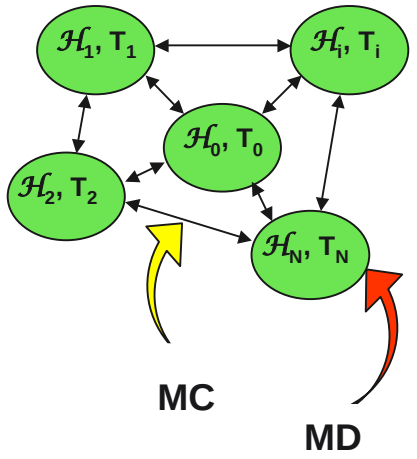
- **Progress with NAMD on BG/L-W**
  - *Lysosyme misfolding (B.J. Berne, Columbia & IBM, NAMD & BlueMatter, BG/W)*
  - *G Protein-Coupled Receptors (GPCR) in a membrane environment (R. Germaine, BlueMatter BG/W)*
  - *NAMD scales very well on BG*
  - *J. Phillips (UIUC), S. Kumar & C. Sosa (IBM)*



APoA1 benchmark (S. Kumar, IBM)

Future Science: Molecular Machines

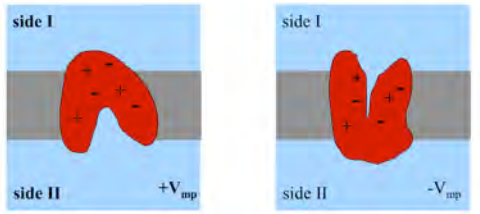
- **REMD: Replica-Exchange MD Sampling**
  - *Effective computing strategy with large scale parallel Blue Gene System*
  - *Consistency with the correct statistical mechanical probability density of the system*
  - *Designed to cover voltages from -500 mV to 500 mV and T from 300K to 400K*



Replica-Exchange MD Strategy

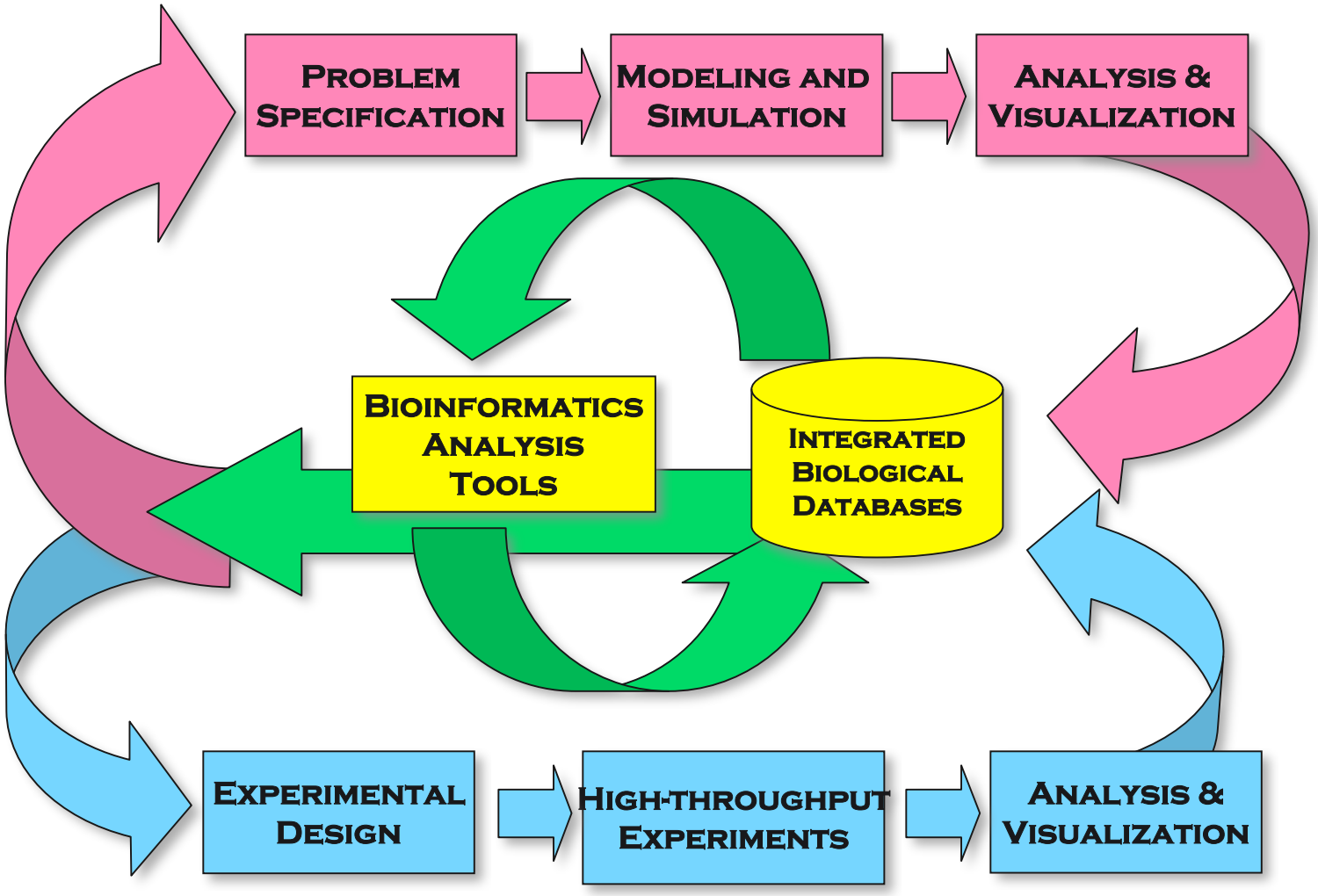
- **Function of Membrane Proteins**
  - *Channels, transporters, exchangers, energy production, receptors, cellular signaling*

- **Design of Artificial Nanodevices**
  - *Voltage-driven switch*
  - *ion pumps*
  - *Electro-mechanical -chemical driving forces*



Voltage-driven conformational switch

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



Finding. *Modeling and simulation is beginning to play a critical role in integrating our understanding of biological mechanisms at multiple levels, including specific cellular subsystems such as metabolism, motility, signaling, regulation, differentiation and development.*

These are critical areas of understanding that are relevant to advancing DOE mission areas.

The community is ready to take big steps in the direction of more complete models, models that incorporate more detailed biological mechanisms and to apply these models to more areas of biological science.

We note that integrative modeling of biological systems complements the relatively well developed field of atomistic modeling (e.g. molecular dynamics, etc.) which can contribute to DOE mission areas in biology, but which is not sufficient to meet the long-term bioengineering goals alone.

Finding. *While there has been considerable progress in advancing integrative modeling during the last decade (as witnessed in the high quality of presentations heard by the subcommittee) this progress has been largely driven by a relatively small number of research groups that have been successful at piecing together research support from a number of disparate sources (e.g. NIH, NSF, DOE, DAPRA).*

There is not currently a long-term research program of appropriate scale aimed explicitly at developing biological modeling and simulation capabilities relevant to DOE missions.

The lack of such a program is holding the field back and makes choosing to work in this area higher risk than the panel would like.

Finding. *The ASCR supported components of the GTL program and computational biology SciDAC activities are not currently supporting projects in applied mathematics or computer science* ***primarily*** *targeted at developing integrated modeling and simulation capabilities for microbes or plants.*

The vast majority of the current INCITE projects in computational biology are focused on atomistic simulations. Many of these systems are important to our understanding of biological mechanism (e.g. precise details of cellulose degradation by cellulase enzymes, ion channel mechanics, protein interactions in signaling pathways, and protein folding pathways).

Integrated multi-scale modeling of biological systems ranging from individual molecules to complete cellular networks are just now be contemplated by the community

Recommendation 1. The ten year OMB PART goal for ASCR the joint modeling and simulation activity of ASCR and BER be modified to read

> *"(ASCR) By 2018, demonstrate significant advances in the capability to predict an organisms' phenotype from its genome sequence, through advances in genome sequence annotation, whole genome scale modeling and simulation and integrated model driven experimentation"*

This PART goal should be accompanied by a specific set of metrics of progress, example metrics could include for a given organism: *the fraction of an organism's genes and gene products included in a model, number of correct metabolic phenotype measurements predicted, number of transcription regulatory elements in a model, number of correct gene expression experiments predicted, fraction of correct predictions of essential genes, number of organisms for which predictive models can be generated, etc.*

Recommendation 2. *DOE should develop an explicit research program aimed at achieving significant progress on the overarching goal of predictive modeling and simulation in DOE relevant biological systems.*

This program should be a joint effort between ASCR and BER and should include a diversity of modeling approaches.

The program should leverage existing experimental activities as well as support the development of new experimental activities that are directly tied to the needs of developing predictive models.

This new research program should be aimed at advancing the state-of-the-art of cell modeling directly, should include equal participation from biologists and mathematicians, computer scientists and engineers and should be indirectly coupled to the more applied goals of bioenergy, carbon cycle research or bioremediation.

This program will need to be supported at a large-enough scale that a multiple target approach can be pursued that will enable progress on many intermediate goals simultaneously by different research groups.

Recommendation 3. *DOE should establish an annual conference that focuses on highlighting the progress in predictive modeling in biological systems.*

This meeting should be an open meeting and separate from any programmatic PI meeting.

One goal of the meeting would be to establish a series of scientific "indicators" of progress in predictive modeling, similar to successful indicators associated with the competitive assessment of structure prediction (CASP).

These types of measures will enable the community to benchmark progress on methods and will be critical to assessing the impact of the research program on fundamentally advancing the state-of-the-art.

Example metrics could include predicting essentiality in microbial genomes, predicting gene expression patterns in novel environments, to predicting yields in metabolic engineering scenarios.

Finding. Integrative modeling and simulation efforts are highly dependent on the curation of genomics data and associated integrated pathway and protein databases that support metabolic reconstruction, interpretation of microarrays and other experimental data.

These databases are the foundation for the development of models and provide the critical biological context for a given organism or problem.

Through resources like NIH's NCBI and NIAID and the dozens of community lead database projects there is reasonable coverage of model organisms (e.g. *Escherichia coli*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, etc.) and pathogens, however there is not the same level of support for curating the data associated with organisms related to energy and the environment.

Finding. Modeling and simulation in microbial systems has advanced in many areas simultaneously.

Today for some systems we have useful and interesting predictive models for core metabolism, for global transcription regulation, for signaling and motility control and for life-cycle development and differentiation.

However we do not yet have many integrated models that include two or more of these capabilities.

Also the successful examples in each case are typically limited to a few model systems and have not be generally extended to the hundreds of organisms relevant to DOE whose genomes are now available.

Recommendation 4. *The modeling and simulation research program should be supported by an explicit series of investments in the modeling technology, database and algorithms and infrastructure needed to address the computational challenges.*

The appropriate early targets for a comprehensive attack on predictive biological modeling are specific functions of microbial organisms (e.g. cellular metabolism, motility, global transcription regulation and differentiation and life-cycle development).

The focus should include advancing the predictive skill on well studied models (e.g. *E. coli*, *B. subtilis*, etc.) but begin to push on to those organisms that stretch the capability beyond the existing well studied model systems (e.g. *Clostridium*, *Shewanella*, *Synechocystis*) and small consortia (communities) of microorganisms relevant to DOE missions such as those associated with bioremediation, carbon sequestration and nitrogen fixation and fermentation and degradation.

We also recommend that the lower eukaryotes (e.g. Diatoms, Coccolithopores, single cell fungi) and plants should be included as targets in longer-term modeling and simulation goals.

Finding. There are a number of obstacles to reaching the visionary goal of a predictive model useful for engineering of an organism derived largely from its genome and related data, here we describe four of the relevant ones.

First, we lack integrated genomics databases and the associated computational methods for supporting curation, extension and visualization of comparative data explicitly focused on supporting the development of modeling and simulations for DOE relevant organisms.

Second, we lack robust mathematical frameworks and software implementing those frameworks for integrating models of metabolism with those of gene regulation which are two of most highly developed areas of modeling and simulation at the whole cell level, but whose mathematical representations are quite different.

Third, we lack the multi-scale mathematics and associated software libraries and tools for integrating processes in cellular models of disparate scales (e.g. molecular scale to that of the whole cell and microbial community) that would enable the modeling community to begin the process of integrated whole cell scale models with atomistic simulations of specific mechanisms.

Fourth, all of computational biology should be framed in a computational and analytical theory that incorporates evolution as the basis for understanding and interpreting the results from comparative analysis. For example we have not yet developed the algorithms needed to make rapid progress on questions such as understanding the major forces governing the evolution of metabolism and regulatory networks. Understanding these forces will be critical to creating the stable engineered strains needed for large scale bioproduction of materials.

Recommendation 5. *DOE should establish a mechanism to support the long-term curation and integration of genomics and related datasets (annotations, metabolic reconstructions, expression data, whole genome screens, phenotype data, etc.) to support biological research in general and the needs of modeling and simulation in particular in areas of energy and the environment that are not well supported by NSF and NIH.*

This mechanism should target the creation of a state-of-the-art community resource for data of all forms that are relevant to organisms of interest to DOE.

This should be a joint activity of ASCR and BER with ASCR responsible for the database and computational infrastructure to enable community annotation and data sharing. It should also leverage the work of established groups.

Recommendation 6. *DOE should work with the community to identify novel scientific opportunities for connecting modeling and simulation at the pathway and organism level to modeling and simulation at other space and temporal scales.*

Examples that could be investigated include integration of atomistic models of protein substrate interactions and protein-protein complexes and their associated cellular pathways, and the integration of microbial models into ocean and terrestrial ecology models which in turn are coupled to global climate models, and models of bioremediation environments that can couple organism metabolic capabilities to external biogeochemistry.

This multi-scale model coupling is beginning to be explored, but much more can be done and it is likely to yield significant scientific insight.