Systems Engineering Challenges and Opportunities in Computational Biology

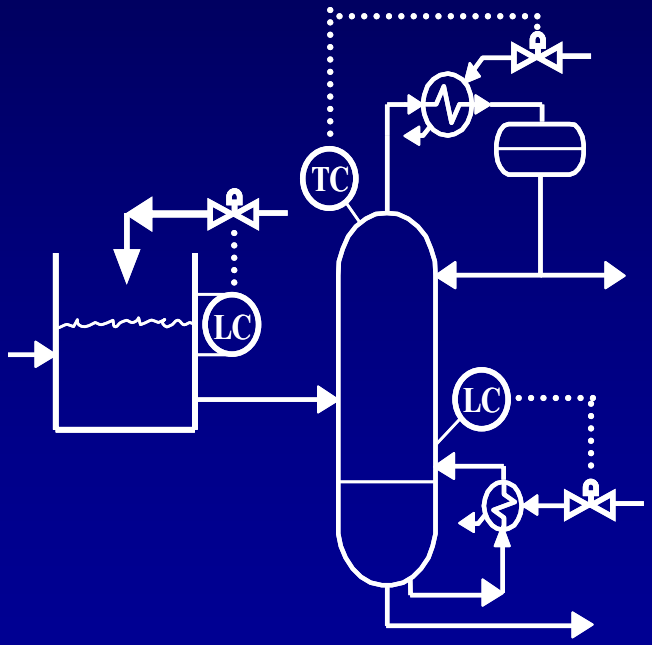
Costas D. Maranas

Penn State University
University Park, PA 16802

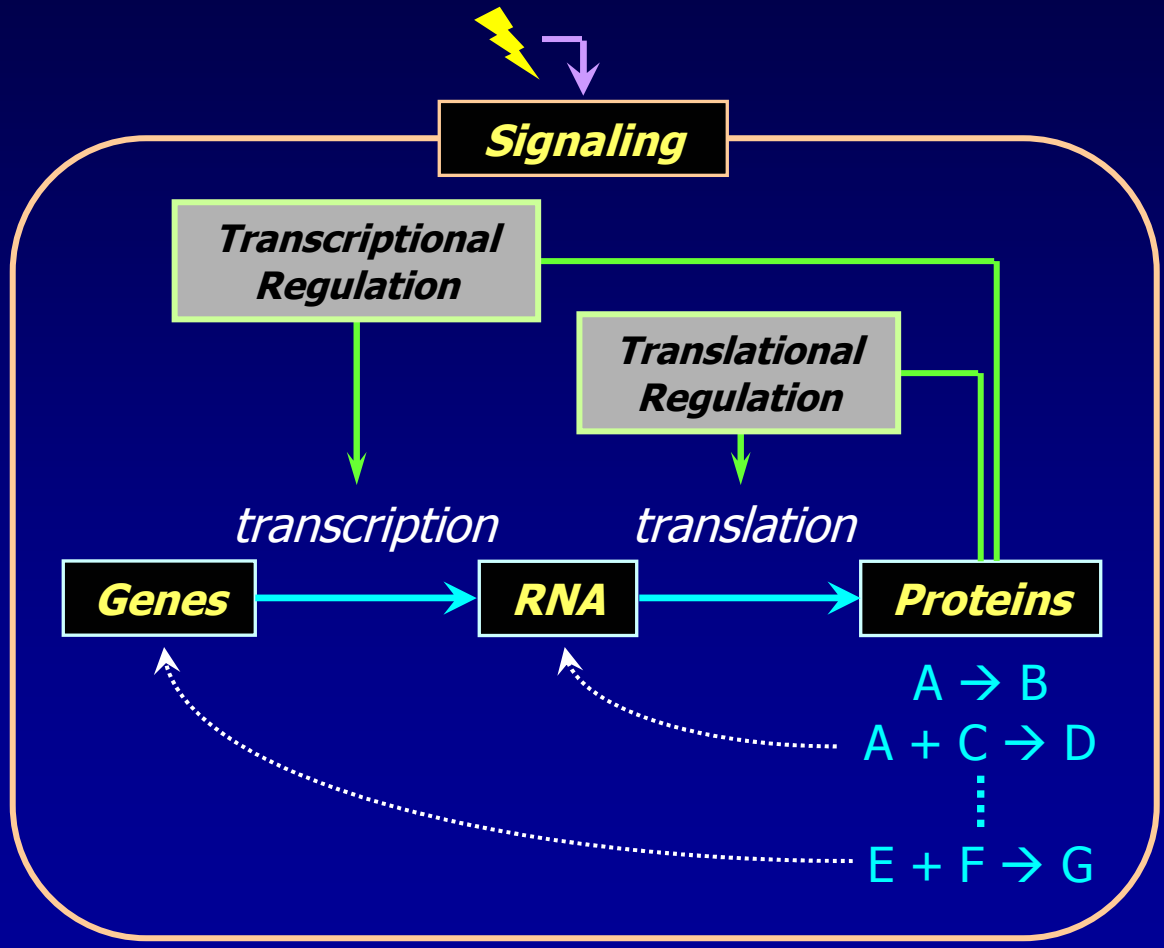
E-mail: costas@psu.edu
Web page: fenske.che.psu.edu/faculty/cmaranas

Networks…

Process system



Biological system

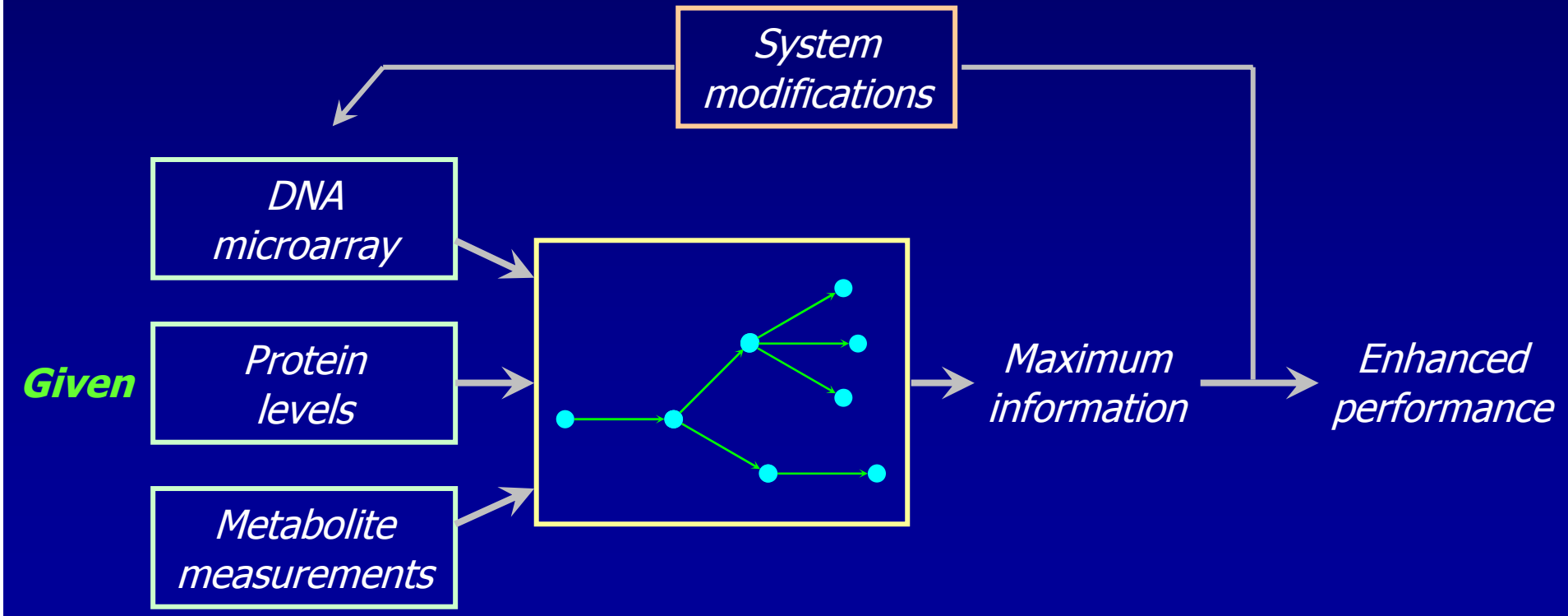


Systems Engineering Challenges in Biological Networks

(1) Component and interactions identification

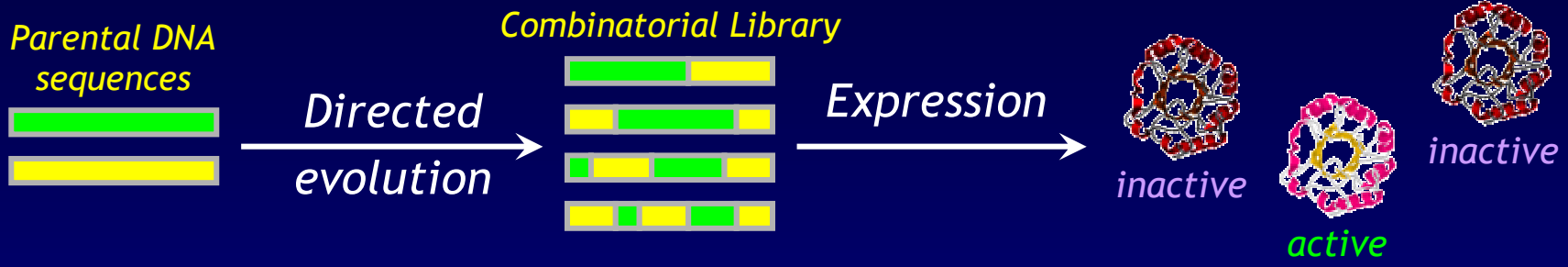
(2) Multiplexing different experimental techniques

(3) Design and optimization of modifications

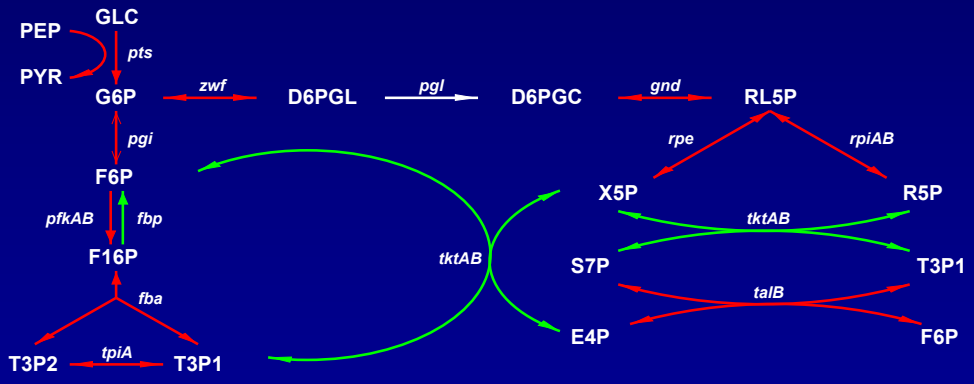


Research

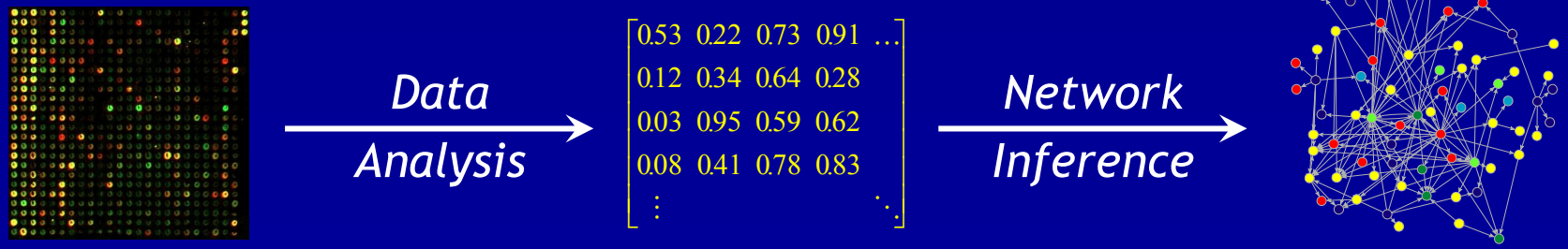
(1) Modeling and optimization in protein engineering



(2) Analysis and design of biochemical pathways



(3) Inference of regulatory networks



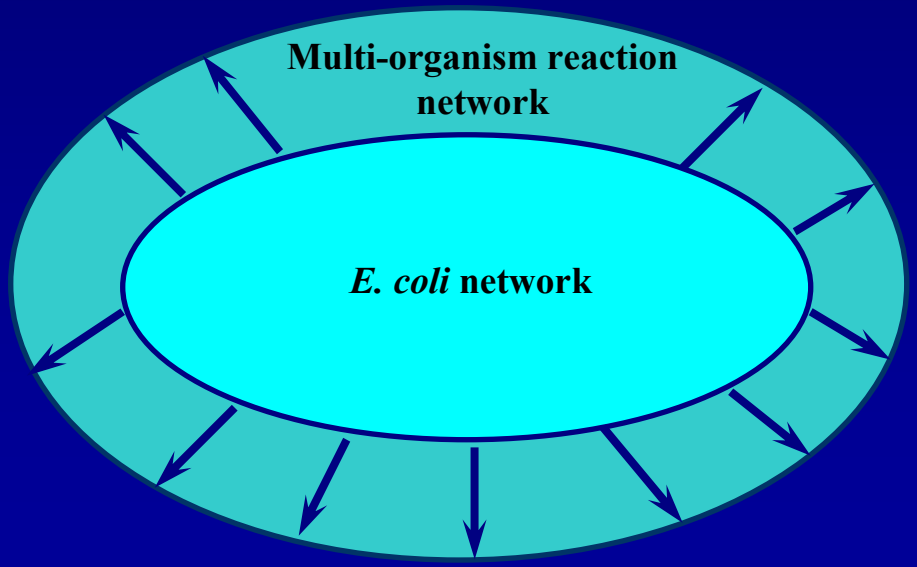
Pathway Design and Analysis

E. coli Stoichiometric Models:

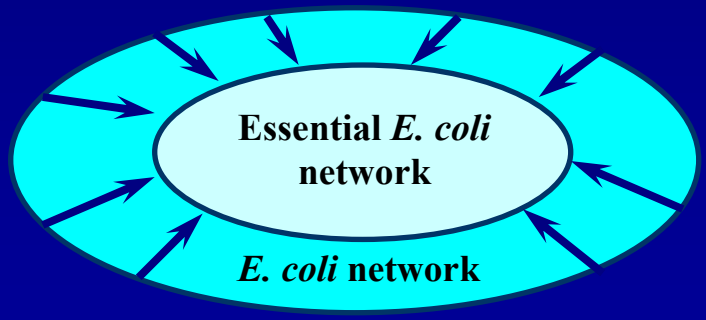
- Schmidt & Nielsen (1999)
  (52 reactions, 31 metabolites)
- Pramanik & Keasling (1997)
  (300 reactions, 289 metabolites)
- Edwards & Palsson (2000)
  (720 reactions, 436 metabolites)

\+ 3,400 reactions EcoCyc KEGG → Universal Network

Gene Addition Study:



Minimum Reaction Network Study:



Mathematical Description

Flux Balance Analysis:

Maximize $\sum_j c_j v_j$

subject **to** $\sum_j S_{ij} v_j = b_i$

$$y_j = \begin{cases} 1 & \text{if reaction flux } v_j \text{ is functional} \\ 0 & \text{otherwise} \end{cases} \longrightarrow 0 \le v_j \le v_j^{\max} \cdot y_j$$

Gene Addition Study:

(Burgard & Maranas, Biotechnol. Bioeng., 2001)

- **Maintain all *E. coli* reactions:**

$$y_j = 1, \quad \forall \; j \in E.\ coli$$

- **Allow up to *m* non-*E. coli* gene additions:**

$$\sum_{j \in \text{ non } E.\ coli} y_j \le m$$

Minimal Reaction Network Study:

(Burgard *et al.*, Biotechnol. Prog., 2001)

- **Minimize total number of reactions in *E. coli* network:**

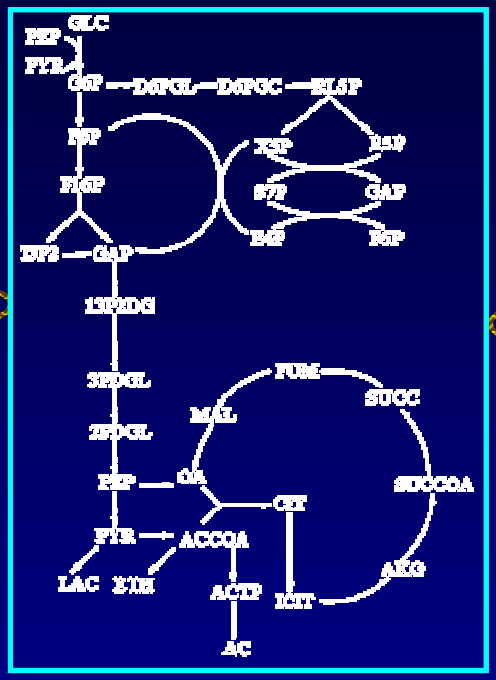
$$\text{Minimize} \quad \sum_{j \in E.\ coli} y_j$$

- **Maintain biomass production requirement:**

$$v_{biomass} \ge v_{biomass}^{target}$$

Motivating Challenge

Cellular Objective (*e.g.*, biomass yield)



Bioengineering Objective (*e.g.*, biochemical yield)

	Maximum Theoretical Yields (mol/mol glucose)	Experimental Yields (mol/mol glucose)
Acetate	2.40	0.81
Ethanol	2.00	0.84
Formate	2.78	1.16
Lactate	2.00	0.10
Succinate	1.71	0.34

(Stokes, *J. Bacteriology*, 1949)

Bilevel Optimization Framework

Outer Problem:

adjust y_j → maximize biotech. objective
(*e.g.*, ethanol, glycerol overproduction)

Inner Problem:

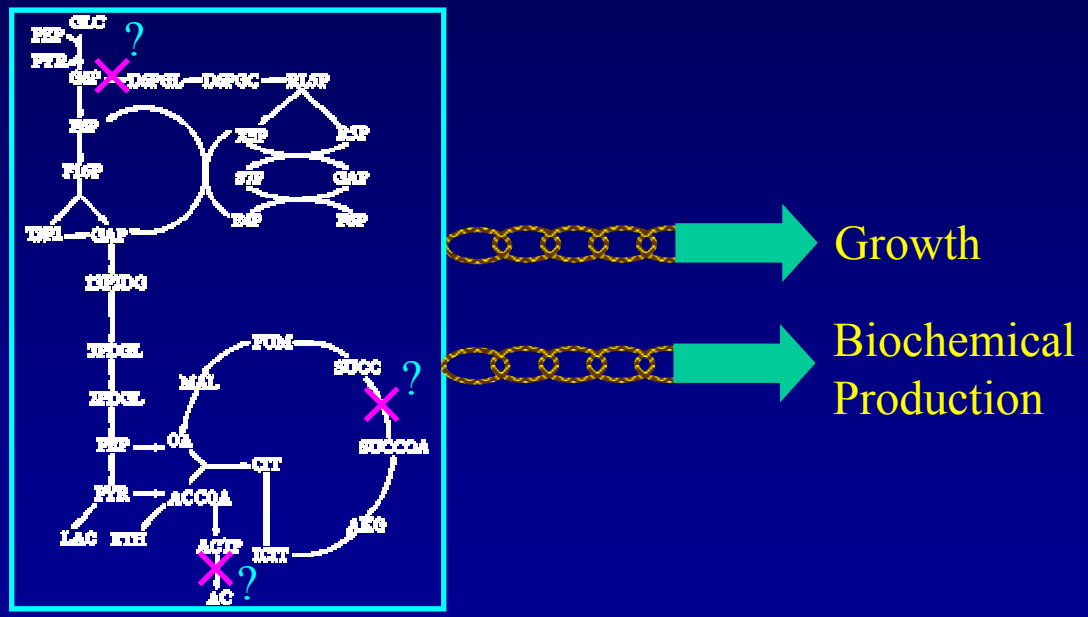
adjust v_j → maximize cellular objective
(*e.g.*, biomass production)

$$\left(\begin{array}{l} \underset{y_j}{\text{Maximize}} \quad v_{Product} \\ \text{s.t.} \quad \left(\begin{array}{l} \underset{v_j}{\text{Maximize}} \quad v_{Biomass} \\ \text{s.t.} \quad \sum_j S_{ij} v_j = 0 \\ \qquad\quad v_{GLC} = uptake \\ \qquad\quad v_j \geq 0 \end{array}\right) \\ 0 \leq v_j \leq \nu_j^{\max} \cdot y_j \\ \sum_j (1 - y_j) \leq \#\ of\ knockouts \\ y_j \in \{0,1\} \end{array}\right)$$

Optimal Gene Knockout Identification

Tasks:

- Identify single, double, triple, and quadruple knockout strategies
- Characterize allowable envelope of biomass vs. biochemical production

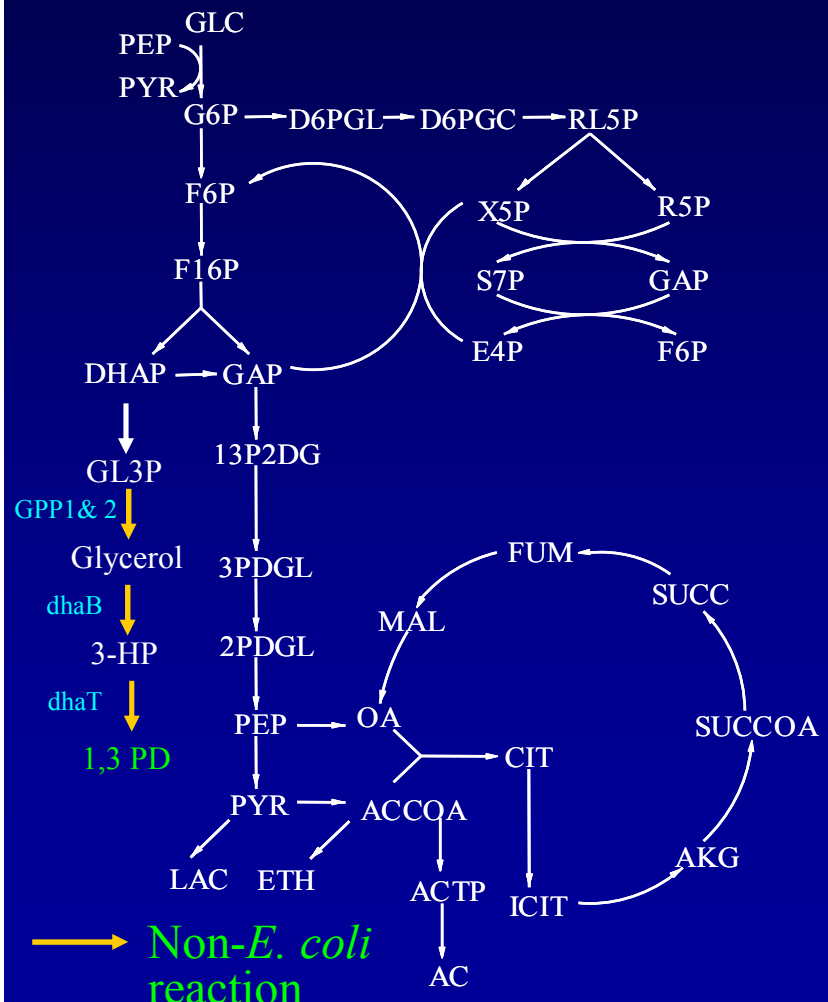


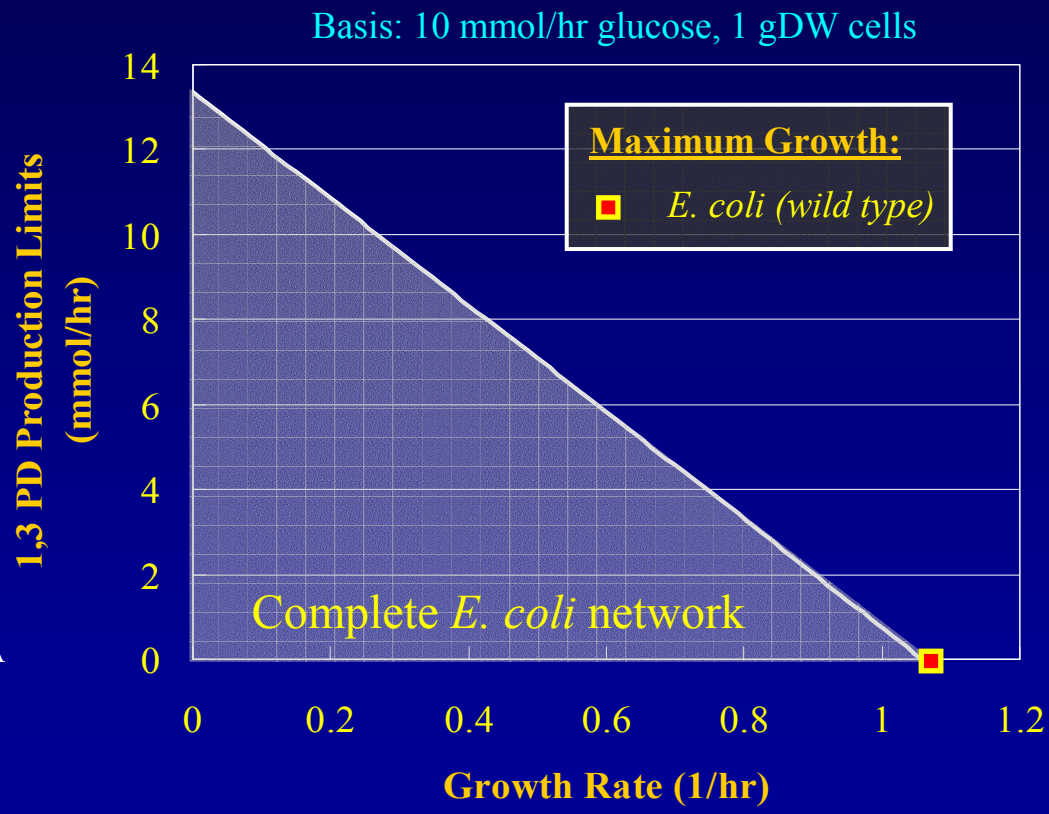
Case Studies:

1,3 Propanediol, succinate, lactate, …

1,3 PD Overproduction

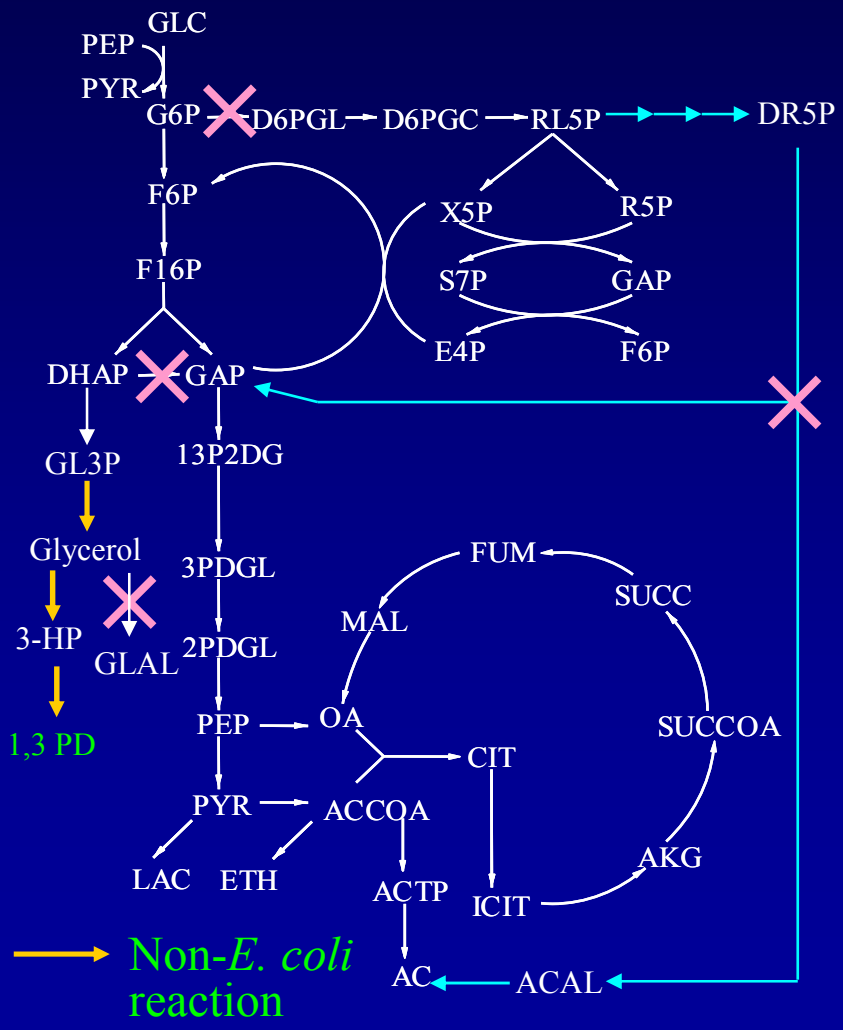
Non-*E. coli* genes/enzymes:			
	GPP1&2:	glycerol-3-phosphatase	*Saccharomyces cerevisiae*
	dhaB:	glycerol dehydratase	*Klebsiella pneumoniae*
	dhaT:	1,3 PD oxidoreductase	*Klebsiella pneumoniae*





1,3 PD Overproducing Mutants

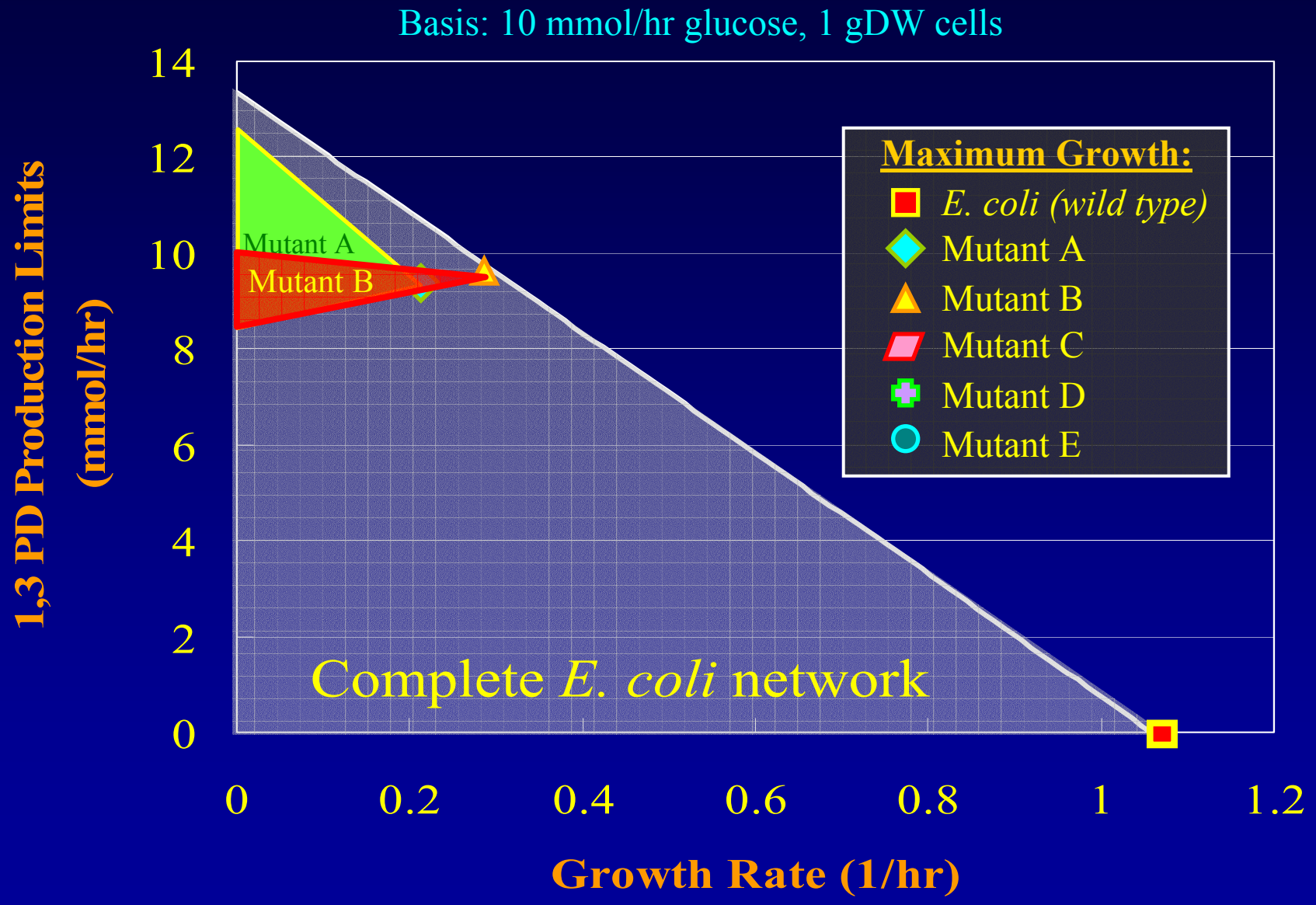
Mutant B:
(1) Aldehyde dehydrogenase (adhC)
(2) Triose phosphate isomerase (tpiA)
(3) Glucose 6-phosphate-1-dehydrogenase (zwf) or 6-Phosphogluconolactonase (pgl)
(4) Deoxyribose-phosphate aldolase (deoC)



Basis: 10 mmol/hr glucose, 1 gDW cells

"Wild" type:	Maximum Biomass :	1.05 hr^{-1}
	1,3 PD :	0.00 mmol/hr
Mutant A:	Maximum Biomass :	0.21 hr^{-1}
	1,3 PD :	9.66 mmol/hr
Mutant B:	Maximum Biomass :	0.29 hr^{-1}
	1,3 PD :	9.67 mmol/hr
Mutant C:	Maximum Biomass :	0.11 hr^{-1}
	1,3 PD :	9.84 mmol/hr
Mutant D:	Maximum Biomass :	0.14 hr^{-1}
	1,3 PD :	9.78 mmol/hr
Mutant E:	Maximum Biomass :	0.16 hr^{-1}
	1,3 PD :	9.75 mmol/hr

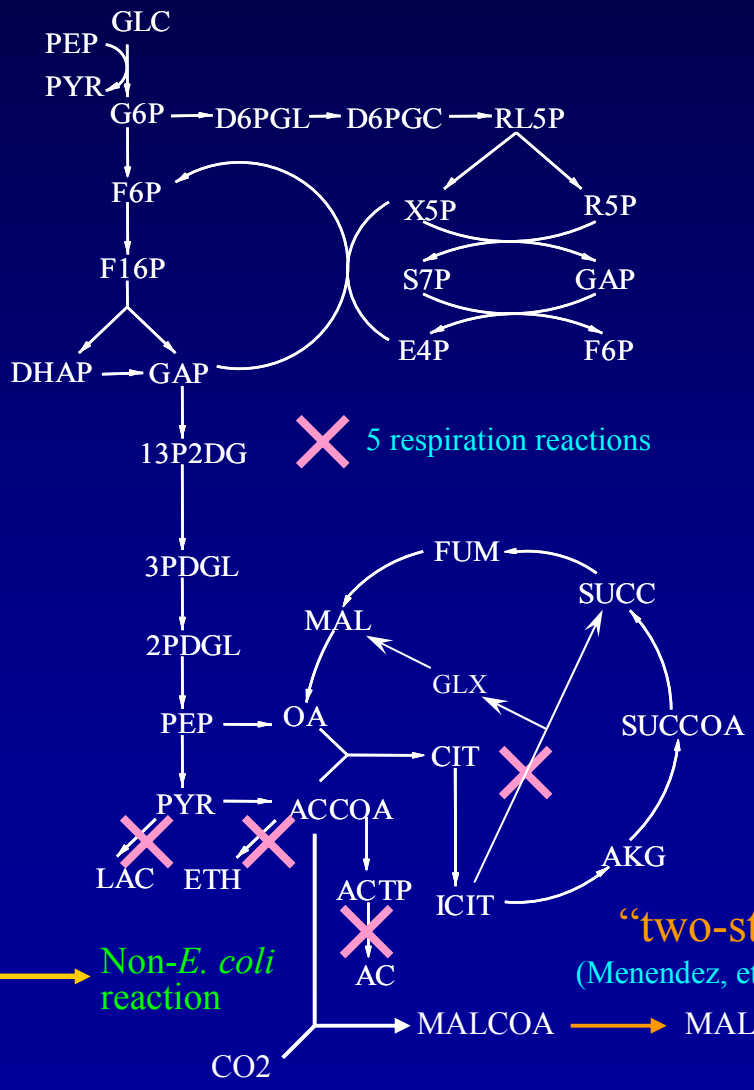
1,3 PD Mutant Characterization



Alternative Route to 1,3 PD

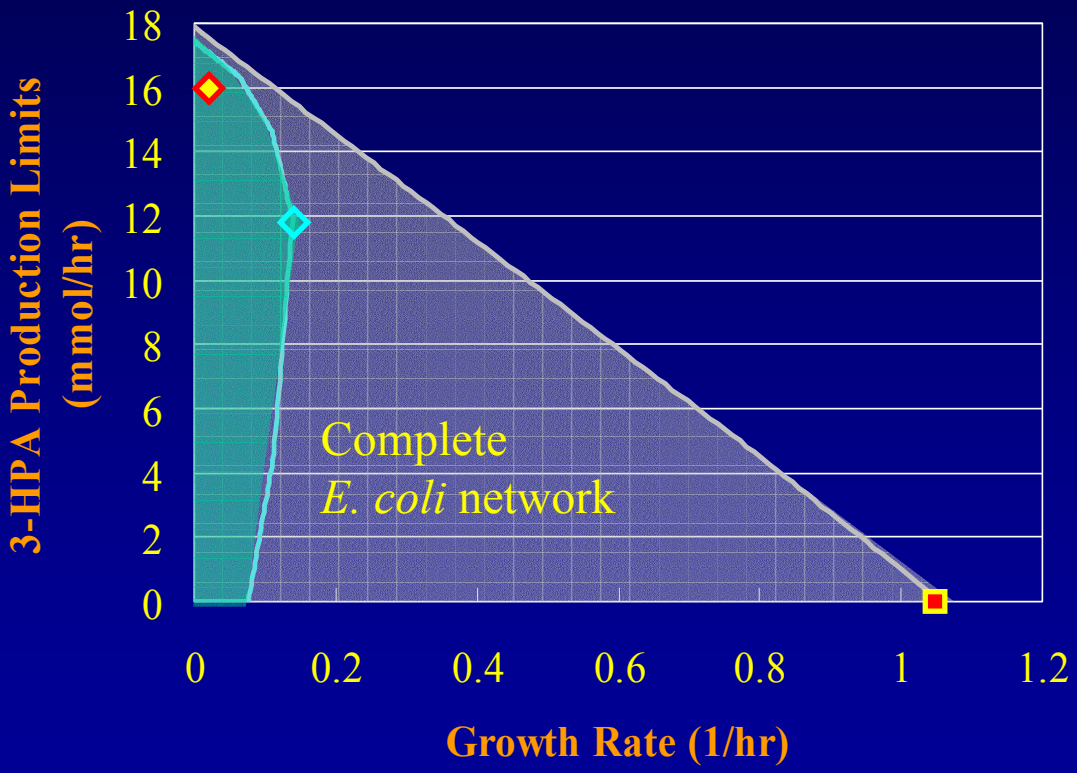
Non-*E. coli* enzyme: malonyl-CoA reductase *Chloroflexus aurantiacus*

Max. Theoretical Yield: 1,3 PD - 1.34 mol / mol glucose (Glycerol route)
1,3 PD - 1.79 mol / mol glucose (3-HPA route)



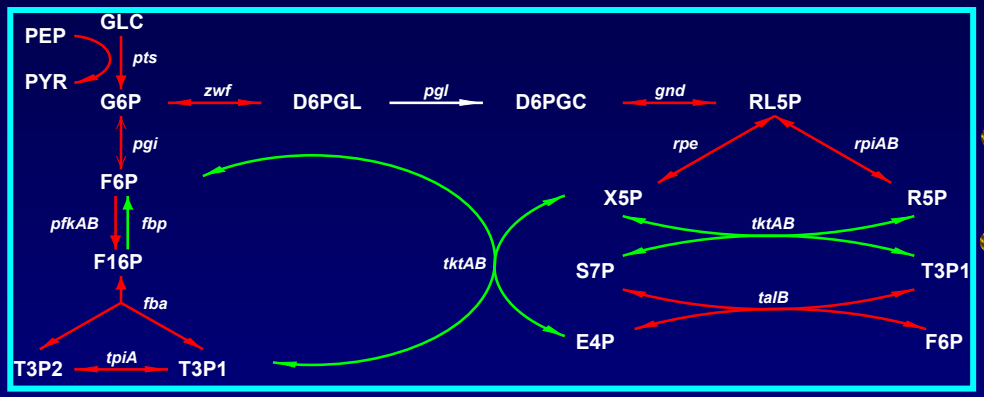
"two-step reduction"
(Menendez, et al., *J. Bacteriol.*, 2002)

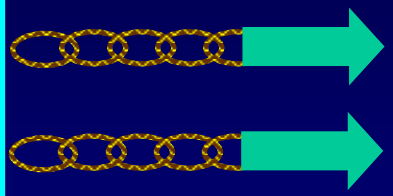
Basis: 10 mmol/hr glucose, 1 gDW cells



Summary

- Method for coupling growth with overproduction





Growth

Biochemical Production

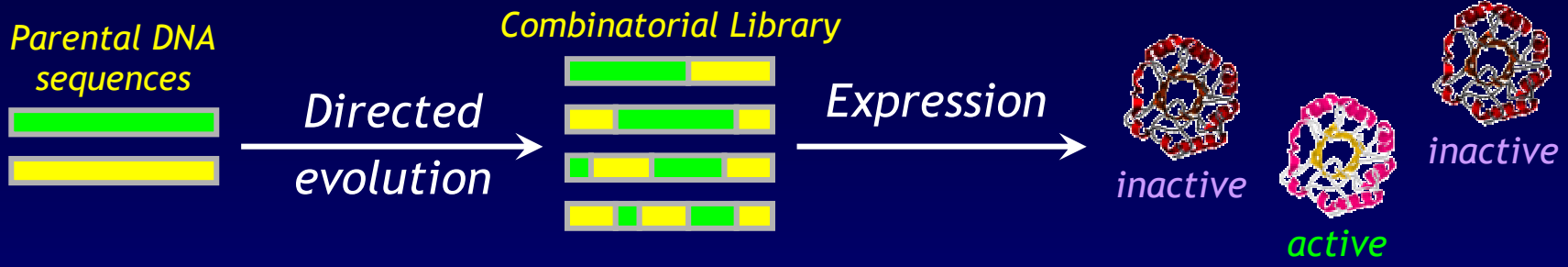
Ongoing work:

- Regulatory network manipulation
- Alternate cellular objectives
- Prioritization of gene knockouts

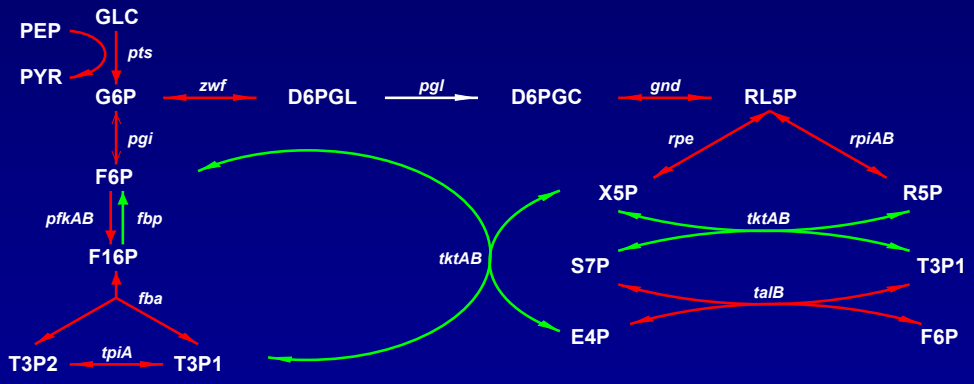
- ***B. Palsson, UCSD***
- ***J. Keasling, U. Berkeley***
- ***F. Blattner, U. Wisconsin***
- ***C. Schilling, Genomatica***

Research

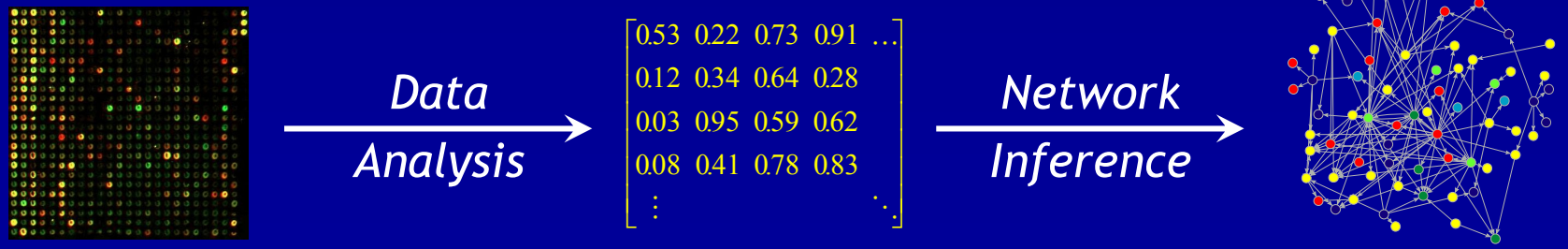
(1) Modeling and optimization in protein engineering



(2) Design and analysis of biochemical pathways



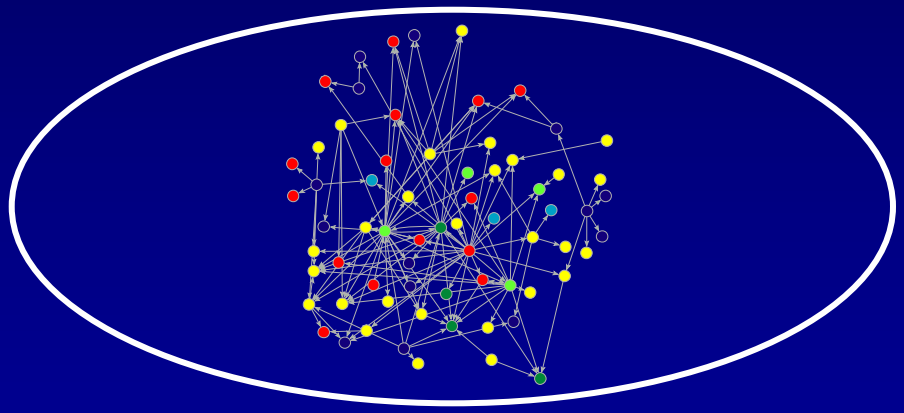
(3) Inference of regulatory networks



Gene Regulatory Network Inference

DNA Microarray Experiments
- Time evolution
- Perturbation

Computational Methods
- Cluster Analysis
- Boolean model
- Continuous model
- Bayesian approach

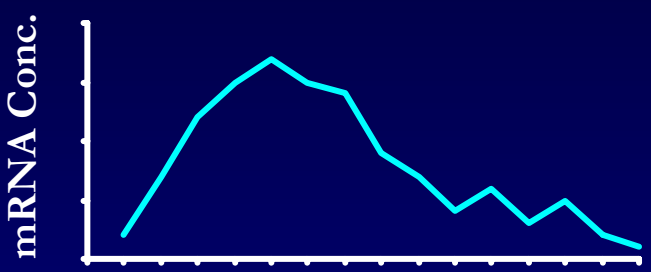


Redirect Cell Metabolism

Identify Therapeutic Targets

Existing Approaches

- **Continuous vs. Boolean Gene Expression**



Botstein *et al.* (1999,2000)
Fedoroff *et al.* (2000,2001)
D'Haeseleer *et al.* (1999)

Somogyi *et al.* (1998)
Akutsu *et al.* (1999)
Ideker *et al.* (2000)

- **Deterministic vs. Stochastic Models**

→ $X_i = F\left(X_j, j = 1,2,\ldots,N\right) \quad i = 1,2,\ldots,N$

Savageau (1998); Weaver *et al.* (1999); Church *et al.* (1999)
(Power Law) (Log-Linear) (Linear)

→ $\Pr\left(X_1, X_2, \ldots, X_N\right) = \prod_{i=1}^{N} \Pr\left(X_i \middle| Pa\left(X_i\right)\right)$

Hartemink (2001); Friedman *et al.* (2000,2001)
(Bayesian Networks)

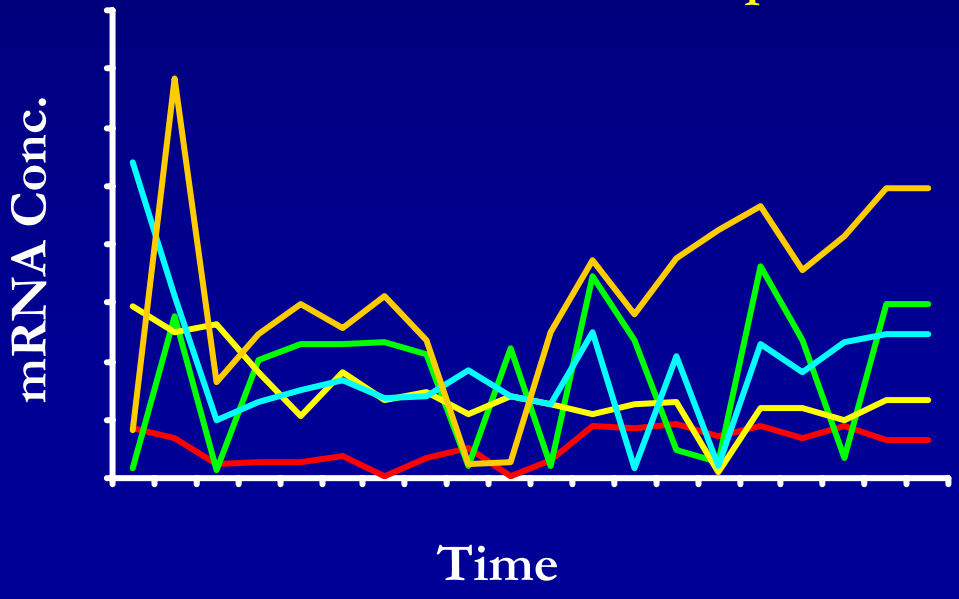
Time Series Experiments

- Affymetrix array data: ***Bacillus Subtilis*** **(~4,100 ORFs)**

  Dr. J. Varner, Genencor Inc., (747 gene subset)

- **Time series studies :**

  - **Exponential Growth Phase:** **5 time-points** **(Experiment T5)**
  - **Amino Acid Pulse:** **9 time-points** **(Experiment T9)**
  - **Cradle-to-Grave:** **20 time-points** **(Experiment T20)**

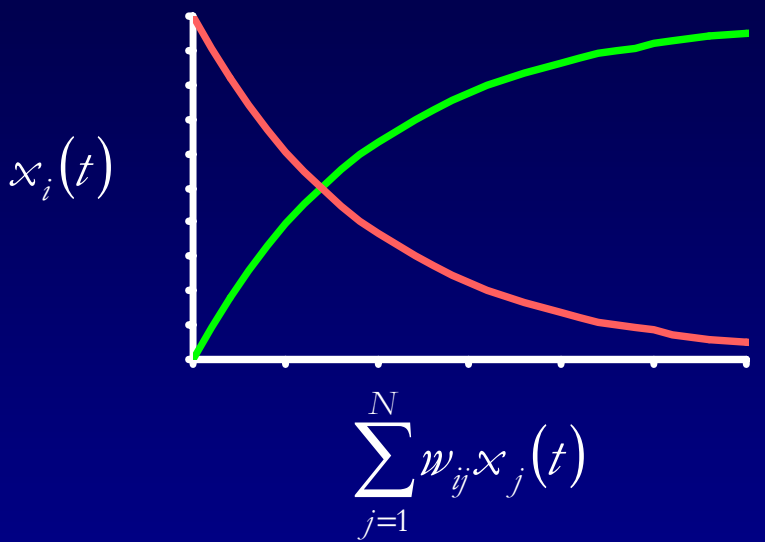


Linear Regulatory Model

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{N} w_{ij} x_j(t)$$

where

$$\frac{dx_i(t)}{dt} \approx \frac{x_i(t+1) - x_i(t)}{\Delta t}$$



w_{ij} = regulatory impact of gene *j* on gene *i*

$w_{ij} \geq 0$: j activates i

$w_{ij} \leq 0$: j inhibits i

Methodology

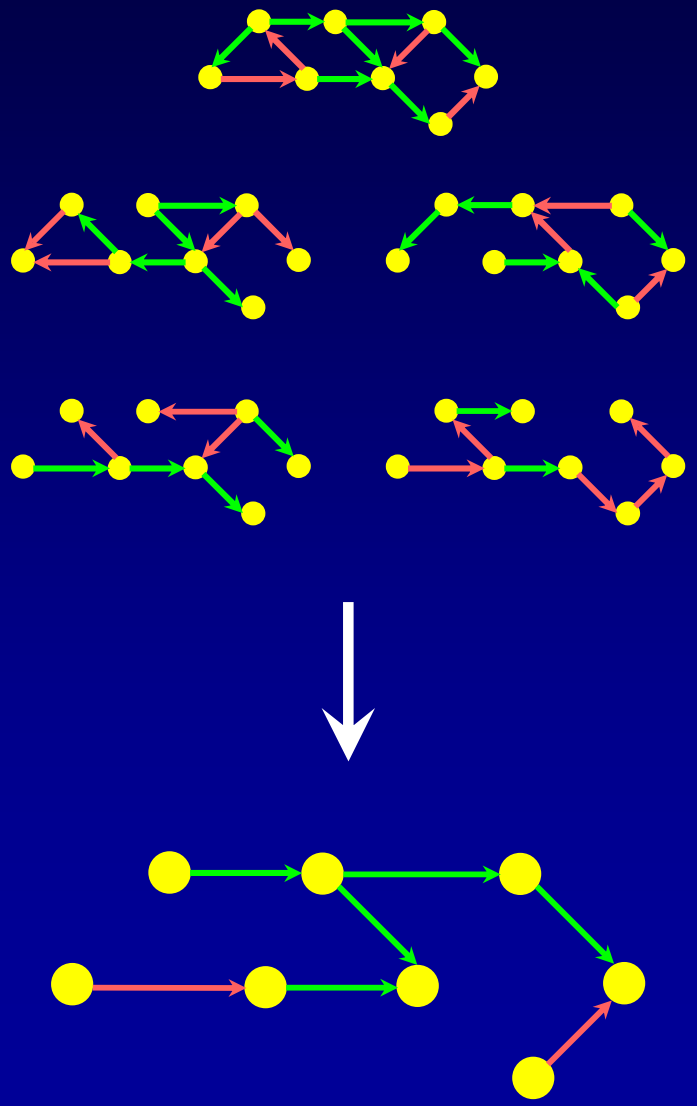
Linear Model

SVD

Family of Feasible Networks

LP-based Optimization

Sparsest Network



Singular Value Decomposition (SVD)

(Yeung, Tegner & Collins, PNAS, 2002)

Network connectivity matrix

$$\dot{X}_{(N\times(T-1))} = W_{(N\times N)} X_{(N\times(T-1))}$$

Expression rate-of-change matrix

Expression matrix

- Underdetermined system of linear equations since N>>T
- Multiple alternative network configurations feasible
- SVD used to represent entire family of potential networks

SVD

$$X^T_{(T-1)\times N} = U_{(T-1)\times(T-1)} \Sigma_{(T-1)\times N} V^T_{N\times N}$$

↓

Particular Solution

$$\hat{W}^T_{N\times N} = V_{N\times N} \Sigma^{-1}{}_{N\times(T-1)} U^T_{(T-1)\times(T-1)} \dot{X}^T_{(T-1)\times N}$$

↓

General Solution

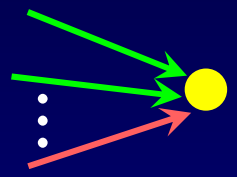
$$W^T_{N\times N} = \hat{W}^T_{N\times N} + C_{N\times(N-T+1)} \hat{V}^T_{(N-T+1)\times N}$$

Null-space matrix

Arbitrary scalar matrix

Maximizing Sparseness (LP)

- **Determine coefficient matrix C such that sparseness of connectivity matrix W maximized**
- **(T-1) incident arcs per gene**

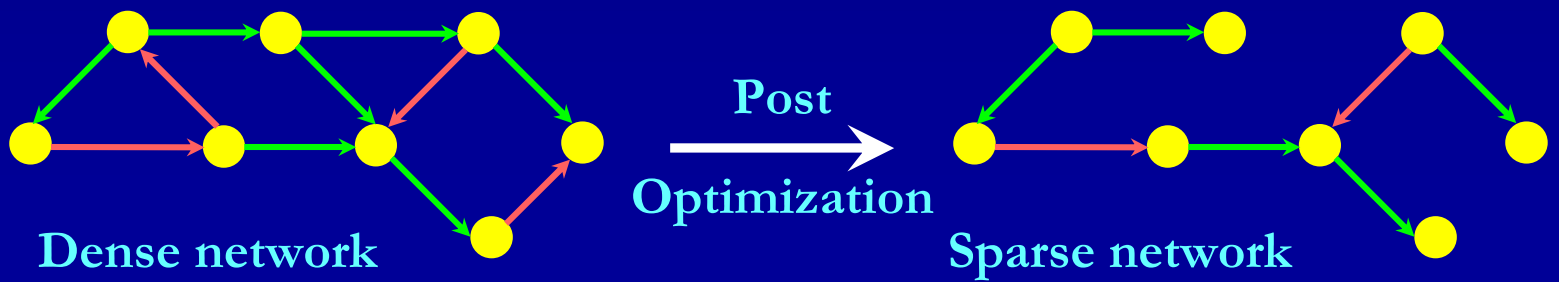


$$\underset{c_{jk},w_{ij}^{+},w_{ij}^{-}}{\text{minimize}} \sum_{i,j} \left(w_{ij}^{+} + w_{ij}^{-} \right)$$

subject to

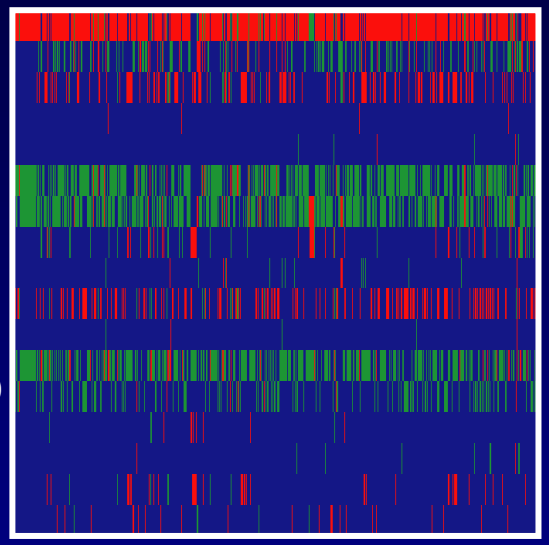
$$\hat{w}_{ij} + \sum_{k=1}^{N-T+1} c_{jk} \hat{v}_{ki} = w_{ij}^{+} - w_{ij}^{-} \qquad \forall i, j = 1,2,\ldots,N$$

$$w_{ij}^{+} \geq 0, w_{ij}^{-} \geq 0 \qquad \forall i, j = 1,2,\ldots,N$$

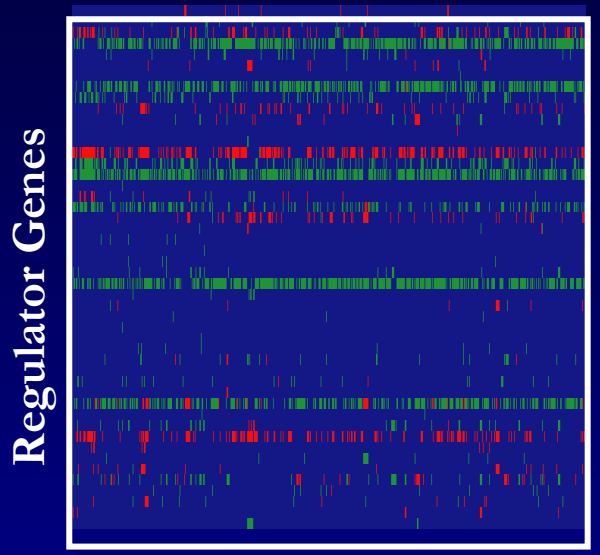


Inferred Regulatory Connections

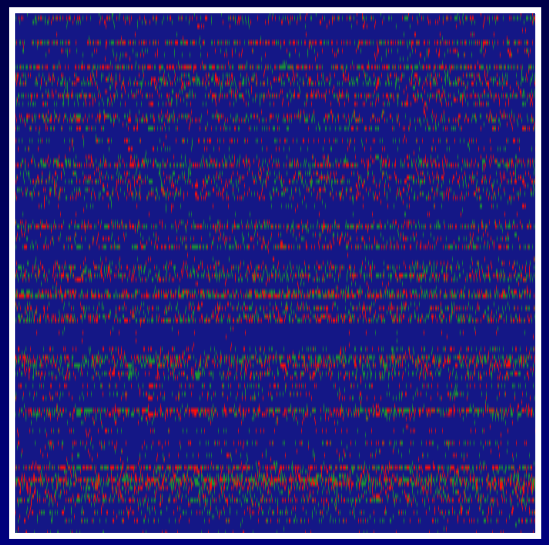
Expo. Growth Phase

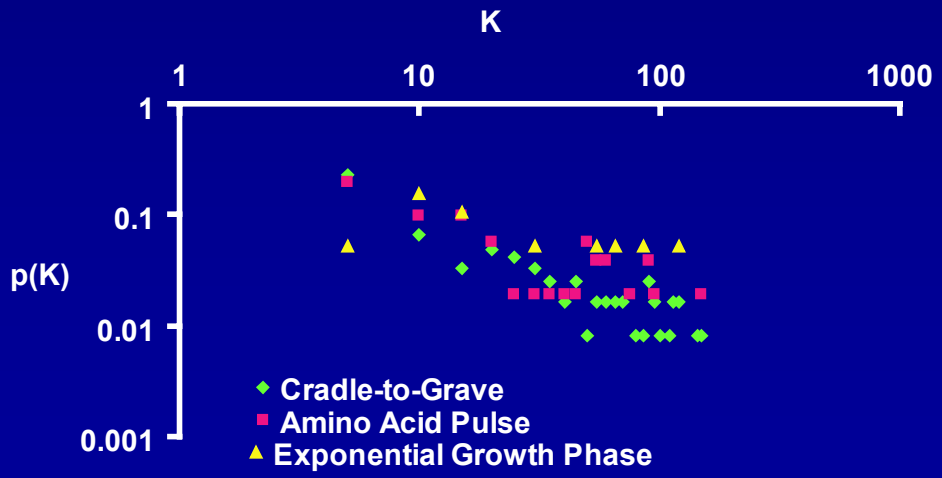


Amino Acid Pulse



Cradle-to-Grave





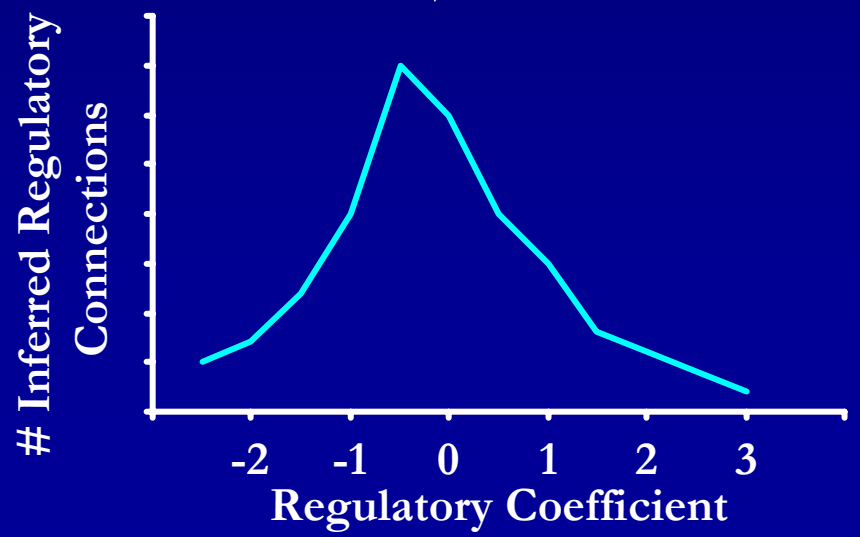
Scale-free
"hub-and-spoke" topology

$$p(K) \approx K^{-\gamma}$$

Robustness Analysis

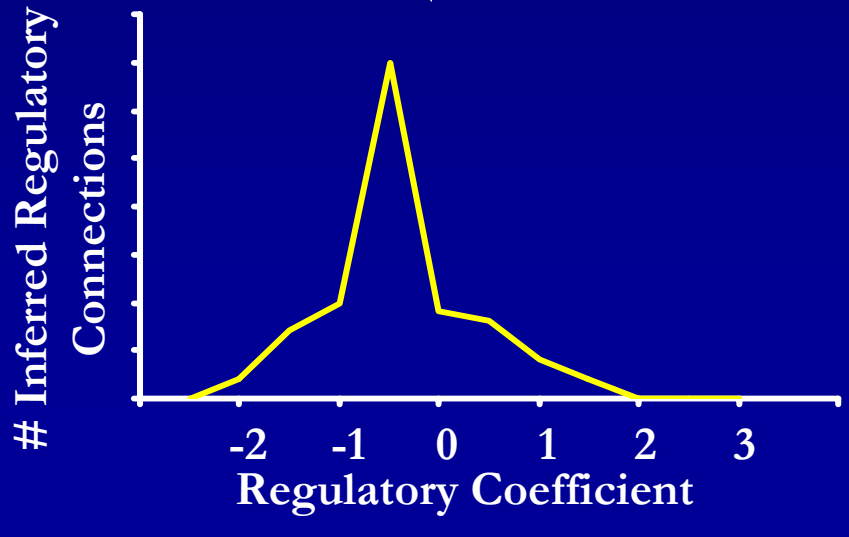
Original Expression Data

2.63	12.52	0. 36	1.59
9.78	8.12	5.47	6.63
⋮	⋮	⋮	⋮
2.92	1.56	1.23	0.83

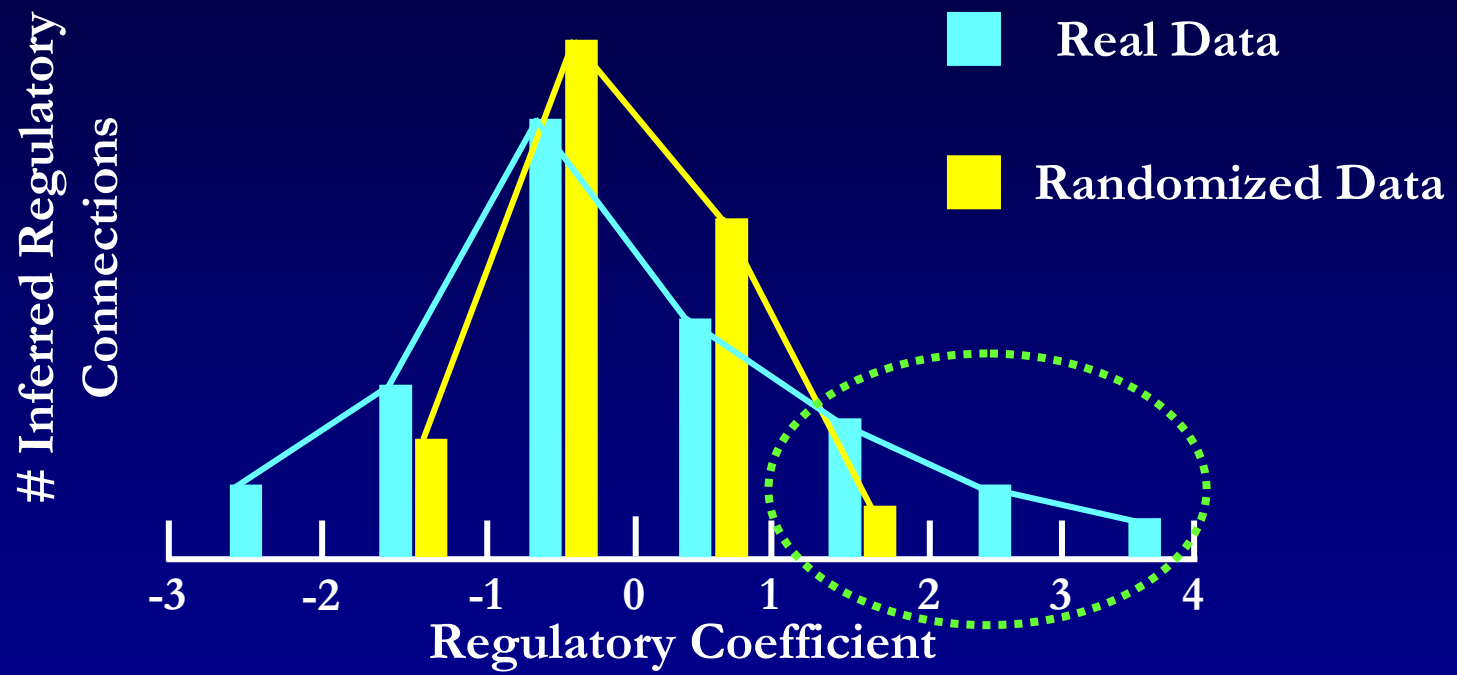


Randomized Expression Data

2.63	12.52	0. 36	1.59
9.78	8.12	5.47	6.63
⋮	⋮	⋮	⋮
2.92	1.56	1.23	0.83

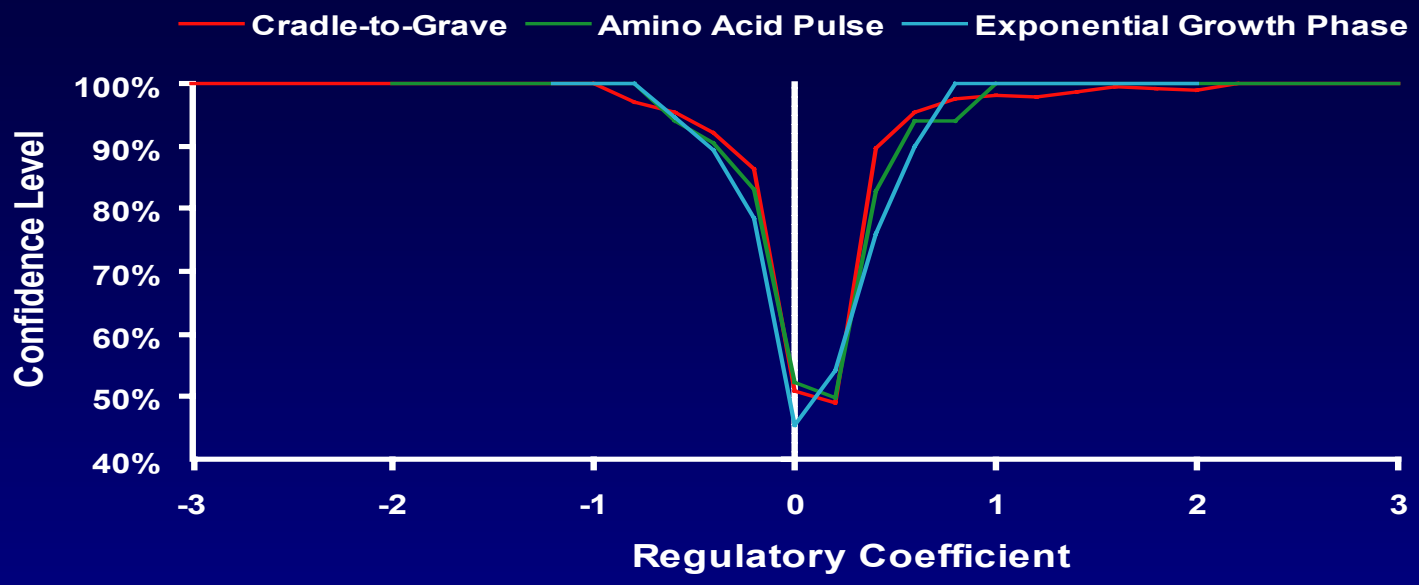


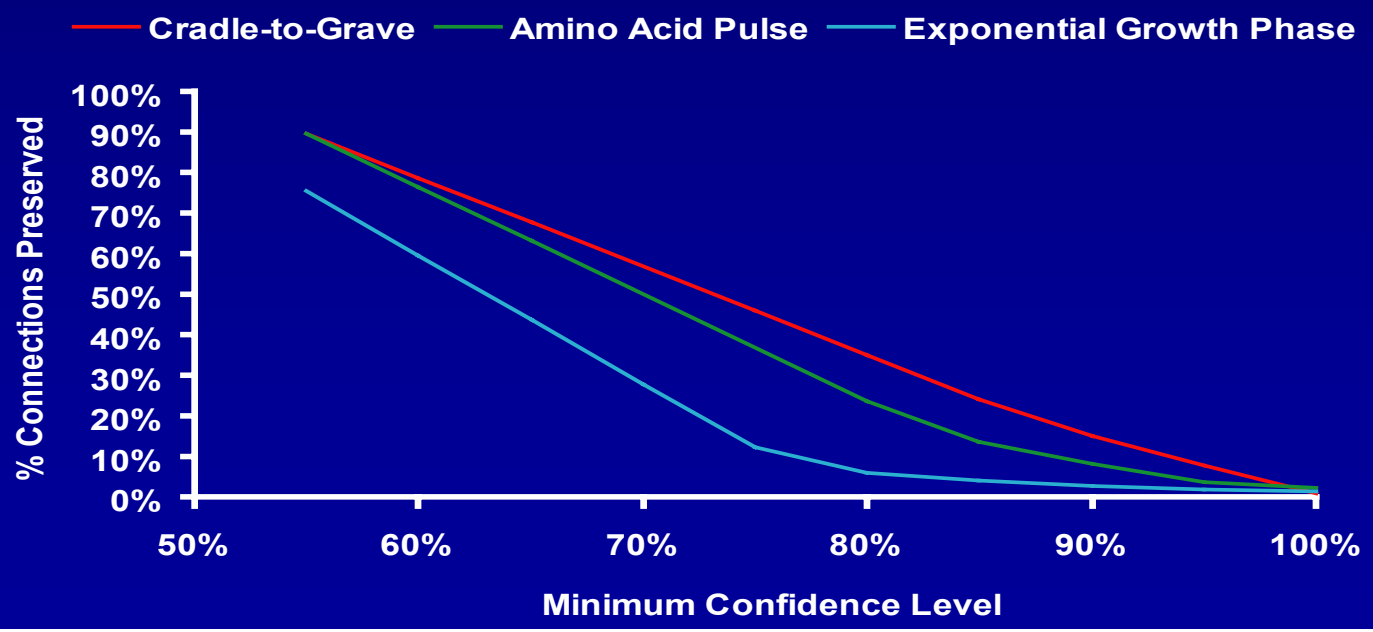
Regulatory Coef. Distribution



$$\text{Confidence Level of } w_{ij} \geq 1 = \frac{\#\text{ of regulatory connections (Real)}}{\#\text{ of regulatory connections (Real + Randomized)}}$$

Regulatory Coefficient Thresholds

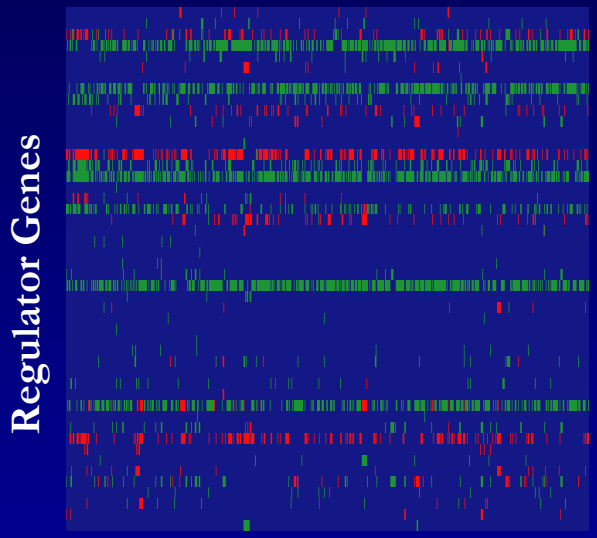




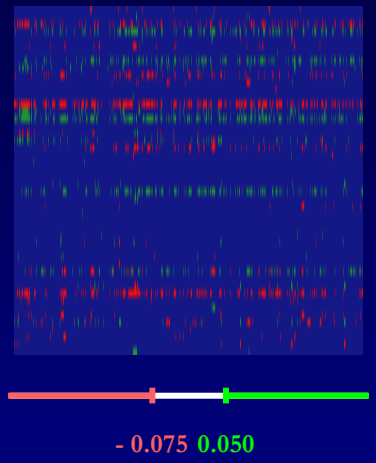
Network Topology

Experiment T9 (Amino Acid Pulse)

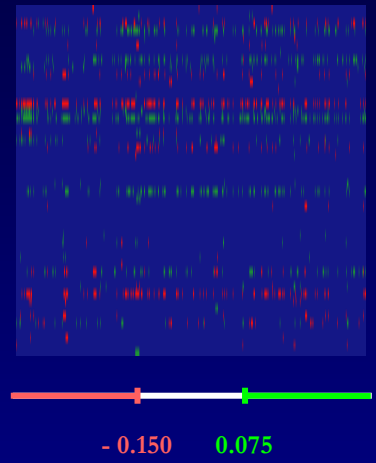
Base Case



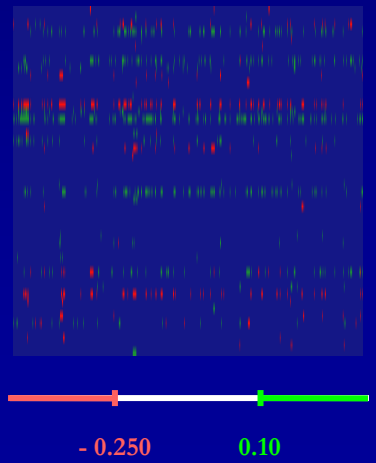
60% Confidence



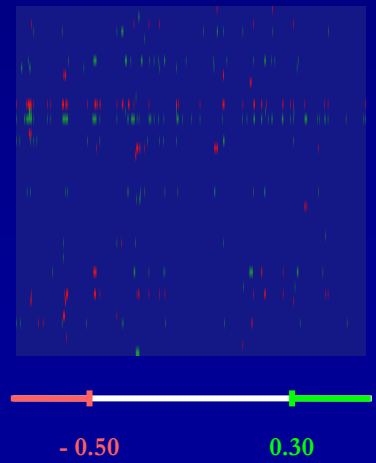
70% Confidence



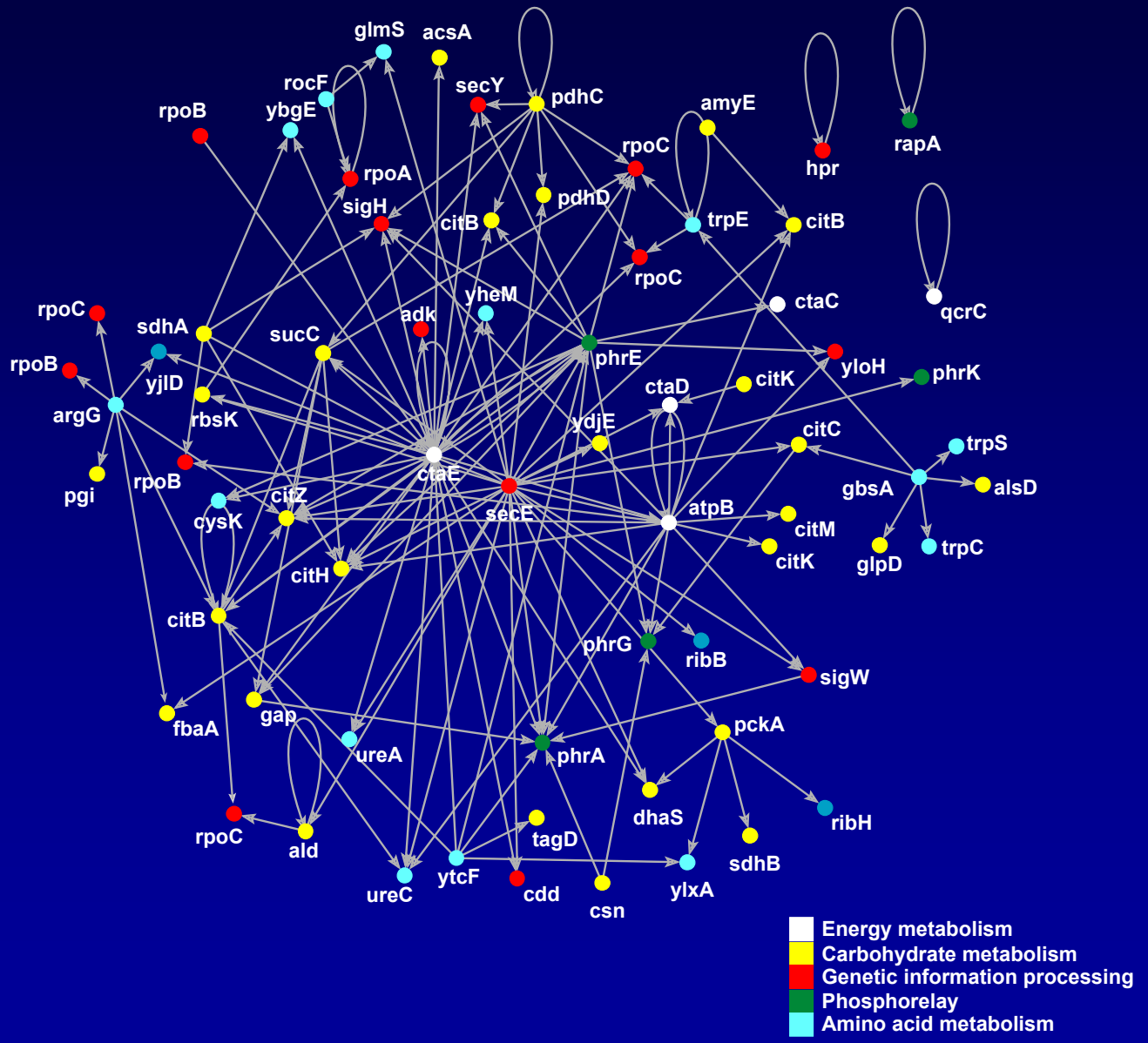
80% Confidence



90% Confidence



Network Connectivity

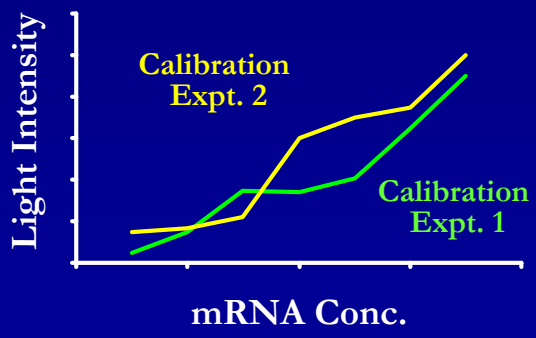


Summary

- Large-scale network inference
- Robustness analysis of inferred arcs
- Recover topological features of regulatory networks

Challenges

- Linear, log-linear or other model formalism ?
- Linking mRNA conc. to light intensity



- Validation with literature data

Concluding Thoughts

- *Convergence of "Biology" and "Systems Engineering"*
- *Inherently inter-disciplinary*
- *Highly goal-oriented*
- *Data-driven*

Acknowledgements

Metabolic Engineering

- Anthony Burgard, Priti Pharkya
- B. Palsson, C. Schilling, Genomatica, Inc.
- J. Keasling, F. Blattner

Protein Engineering

- Gregory Moore, Manish Saraf
- Prof. Stephen J. Benkovic
- Drs. Alexander Horswill, Stefan Lutz

Regulatory Network Inference

- Dr. Anshuman Gupta, Madhukar Dasika
- Jeff Varner, Genencor, Inc.

Funding:

NSF, DOE, IBM-SUR