Constraint-based Modeling: Part I

Biological Constraints, Network Reconstruction, and FBA

Thursday, April 29, 2004

Timothy E. Allen / Bernhard Ø. Palsson

BE 203 Lecture

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Outline

- Constraints in biology
- Reconstructions and applying constraints
- Constraint-based modeling (CBM): philosophy and overview
- Basics of flux balance analysis (FBA)
- Lessons learned
- CBM: an expanding field

Constraints in Biology

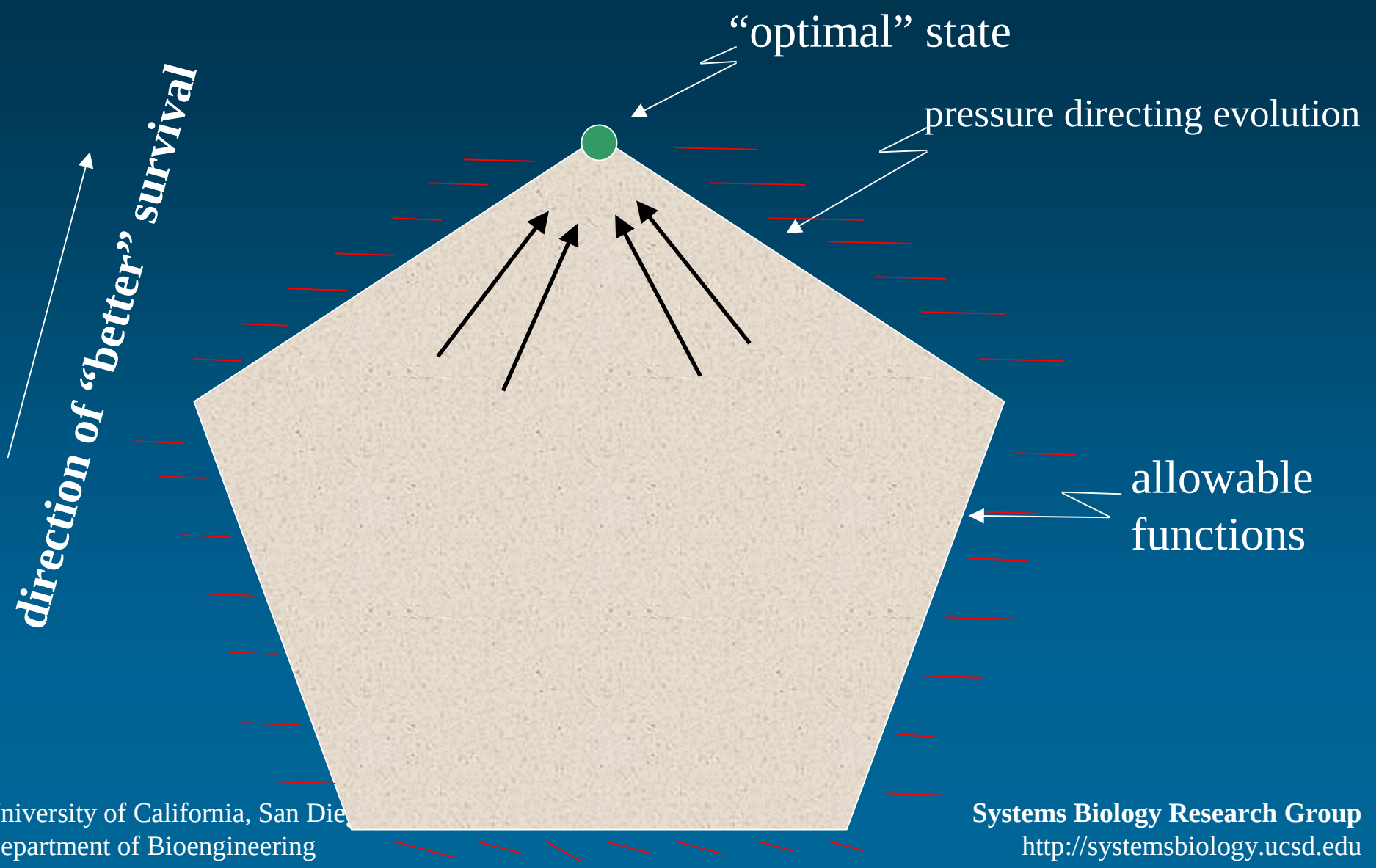
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Constraints Govern Possible Biological Functions

- Evolution: Organisms exist in resource-scarce environment
  - The more "fit" organisms survive with a higher probability than the less "fit"
  - Fitness requires satisfying a myriad of constraints which limit the range of available phenotypes
- Survival thus depends on best utilization of resources to survive & grow, subject to constraints
- All expressed phenotypes must satisfy imposed constraints → constraints therefore enable us to eliminate impossible cellular behaviors

Evolution and governing constraints



University of California, San Die
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

What kinds of constraints do cells have to abide by?

- Physico-chemical constraints
  - Conservation of mass, energy, & momentum
  - Maximal reaction/transport rates
  - Thermodynamic constraints
- Topobiological constraints
  - Macromolecular crowding constrains possible interactions & diffusion of large molecules
  - DNA, e.g., must be both tightly packed and yet easily accessible to the transcriptional machinery → two competing needs constrain the physical arrangement of DNA within the cell

What kinds of constraints do cells have to abide by?

- Environmental constraints
  - Condition-dependent → variable constraints
  - pH, temperature, osmolarity, availability of electron receptors, etc.
  - Availability of carbon, oxygen, sulfur, nitrogen, and phosphate sources in surrounding media
- Regulatory constraints
  - Self-imposed “restraints”
  - Subject to evolutionary change
  - Allow cells to eliminate suboptimal phenotypes and confine themselves to behaviors of increased fitness

Network Reconstruction:

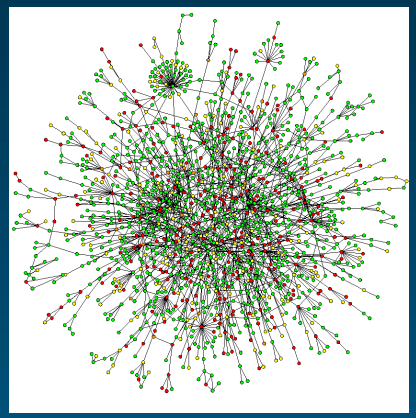
The Key to Systems Biology

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Reconstructing Networks

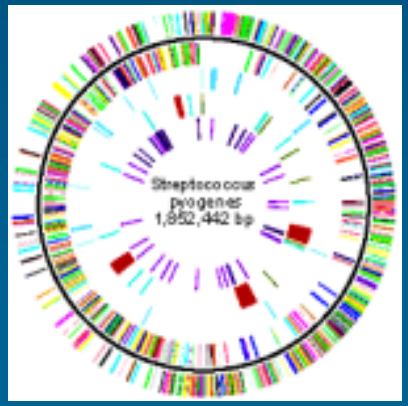
The challenge of integrating heterogeneous data types



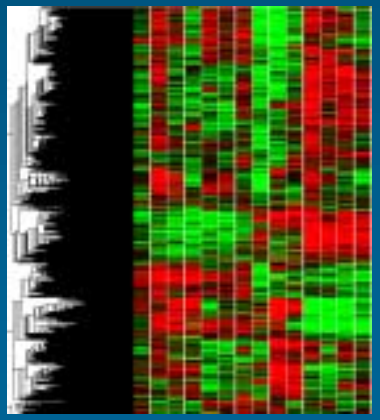
Proteomics

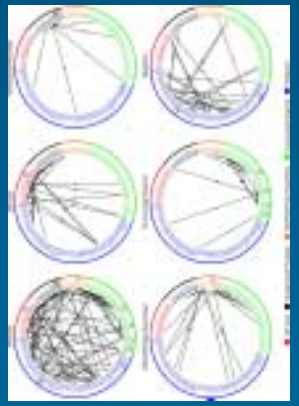
Phenomics

Metabolomics



Genomics





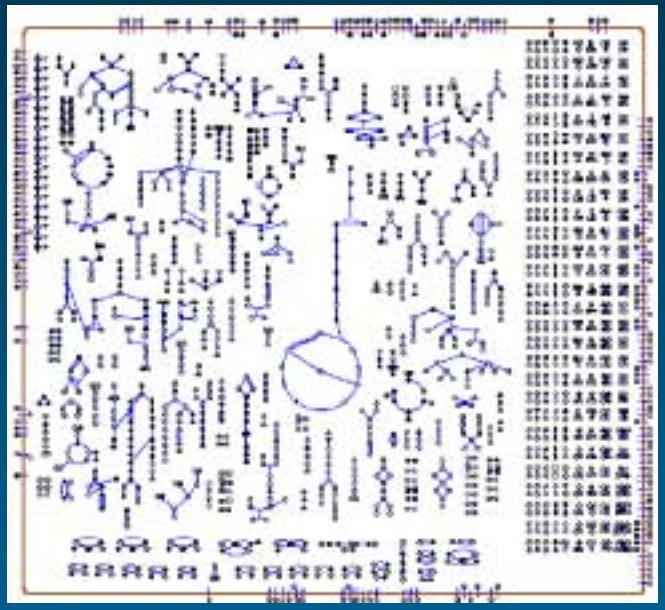
Transcriptomics

University of California, San Diego
Department of Bioengineering

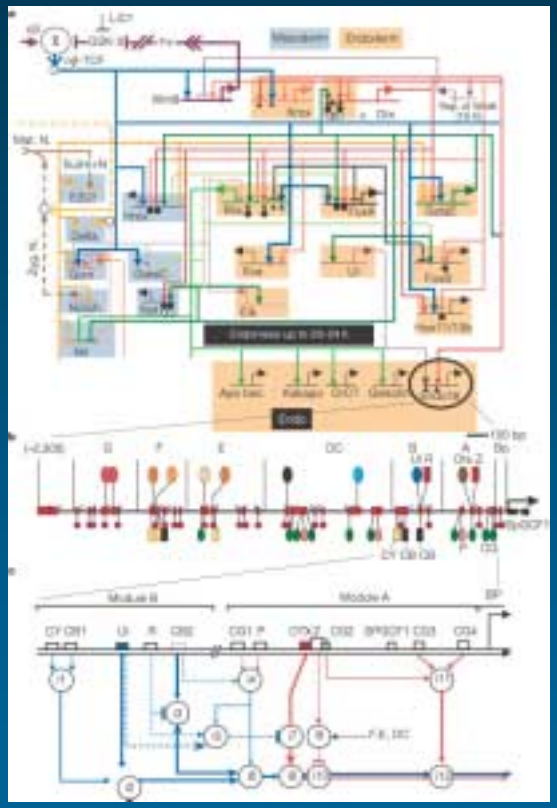
Systems Biology Research Group
http://systemsbiology.ucsd.edu

Network Reconstruction

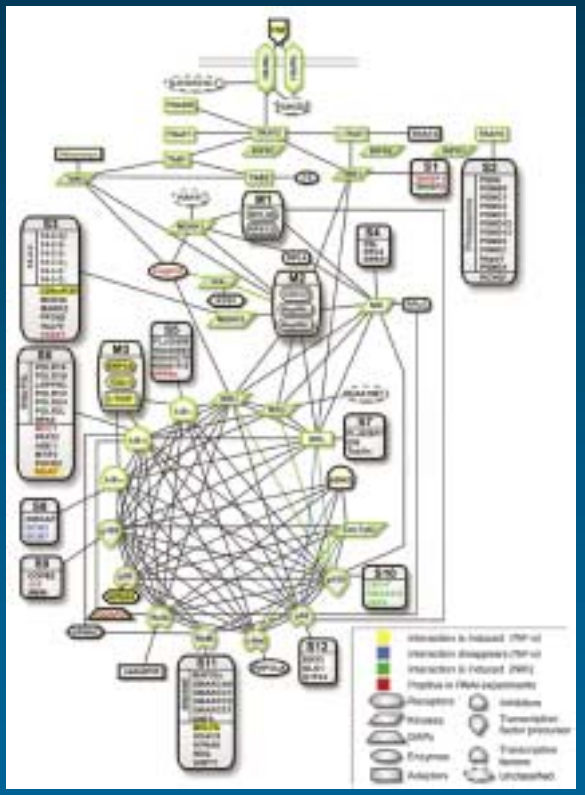
Metabolic



Regulatory



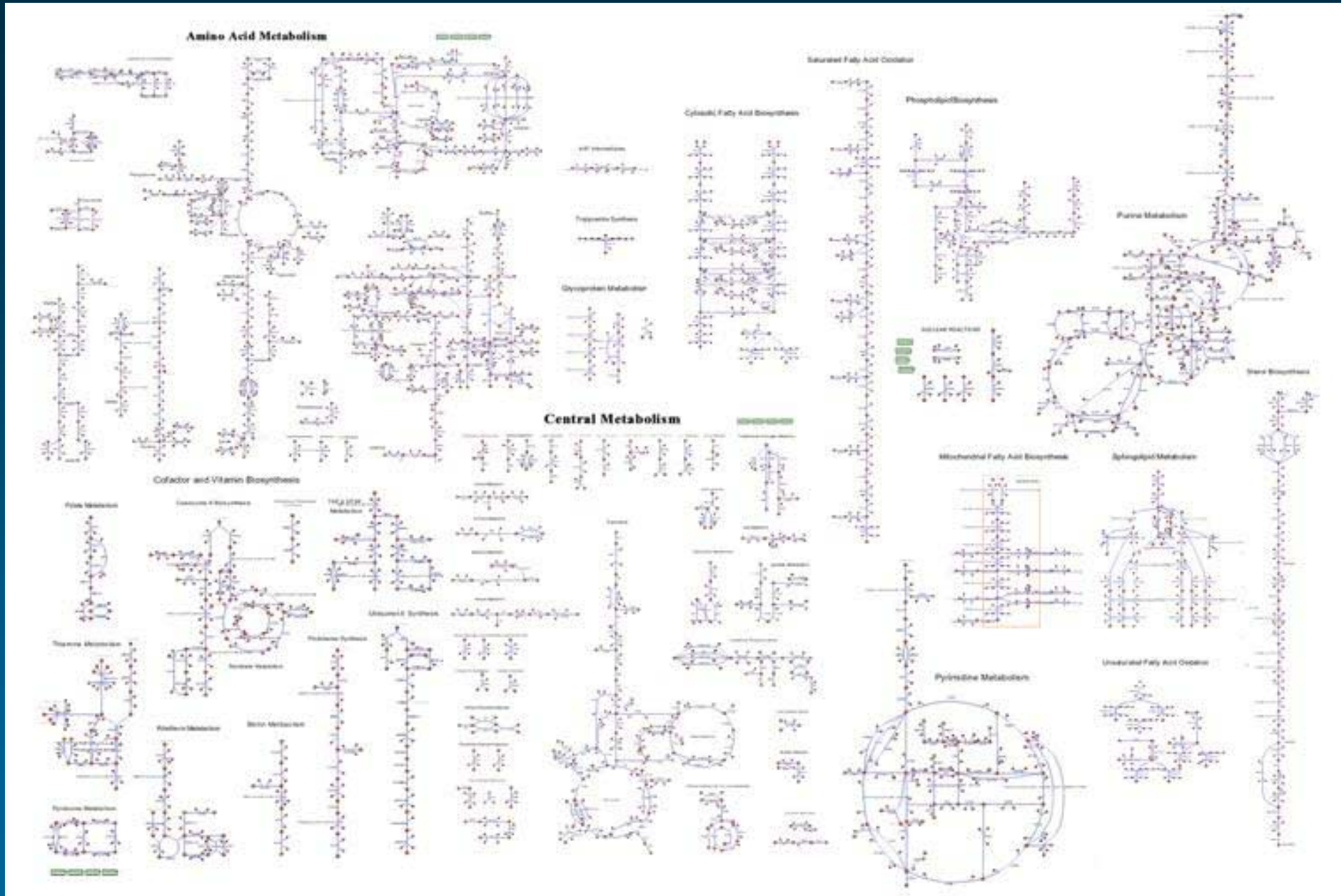
Signaling



Integrated but incomplete

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu



University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

How are metabolic networks reconstructed?

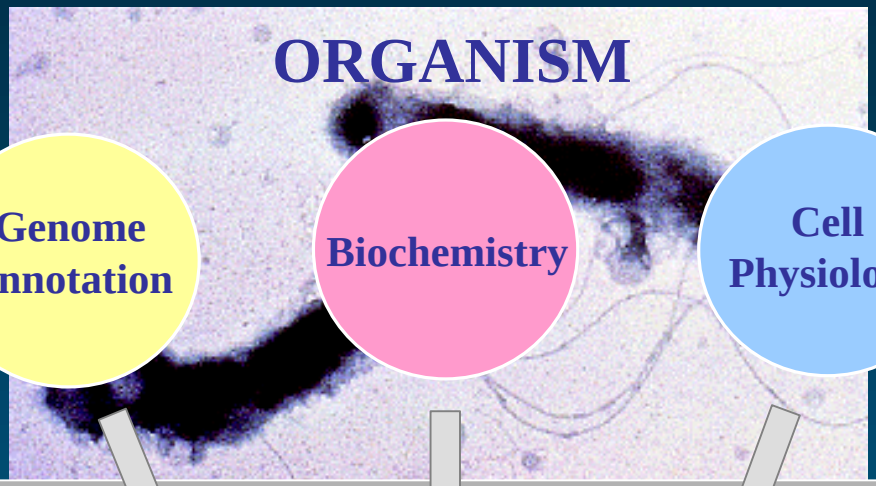
Genome Annotation
- by homology, location

Biochemical Data
- protein characterized

Physiological Data
- indirect, pathway known

Inferred Reactions
- indirect, inferred from biomass requirements

Quantitative Analysis
- simulate cell behavior
- drive experimental studies



Genome Annotation

Biochemistry

Cell Physiology

Inferred Reactions

Network Reconstruction

Metabolic Model

Quantitative Analytical Methods

New Predictions
Emergent Properties

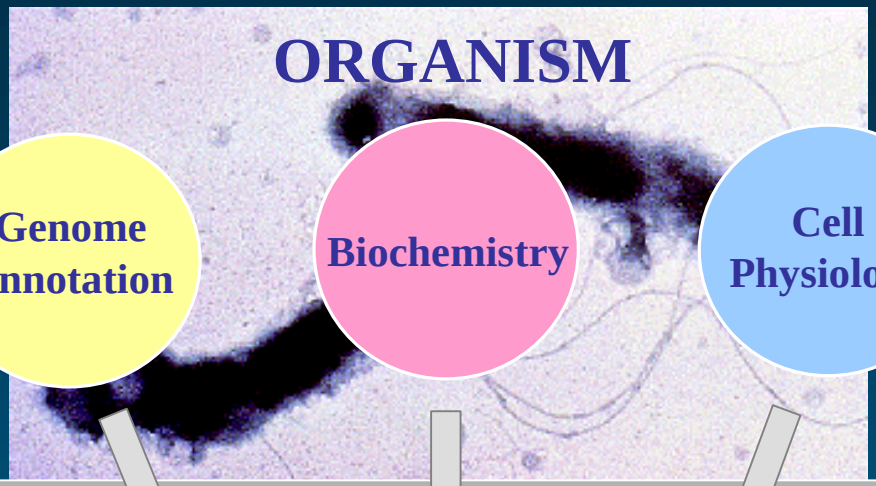
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

How are metabolic networks reconstructed?

Model Development: an iterative process

- Biochemical data
- Revised ORF assignments



Genome Annotation

Biochemistry

Cell Physiology

Inferred Reactions

Network Reconstruction

Metabolic Model

Quantitative Analytical Methods

Computational, Biochemical Investigation

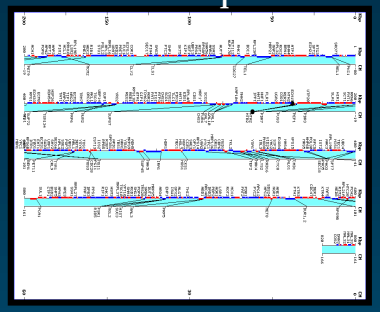
New Predictions
Emergent Properties

University of California, San Diego
Department of Bioengineering

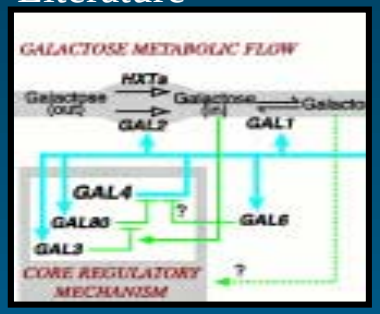
Systems Biology Research Group
http://systemsbiology.ucsd.edu

Reconstruction of Regulatory Networks

Genome sequence

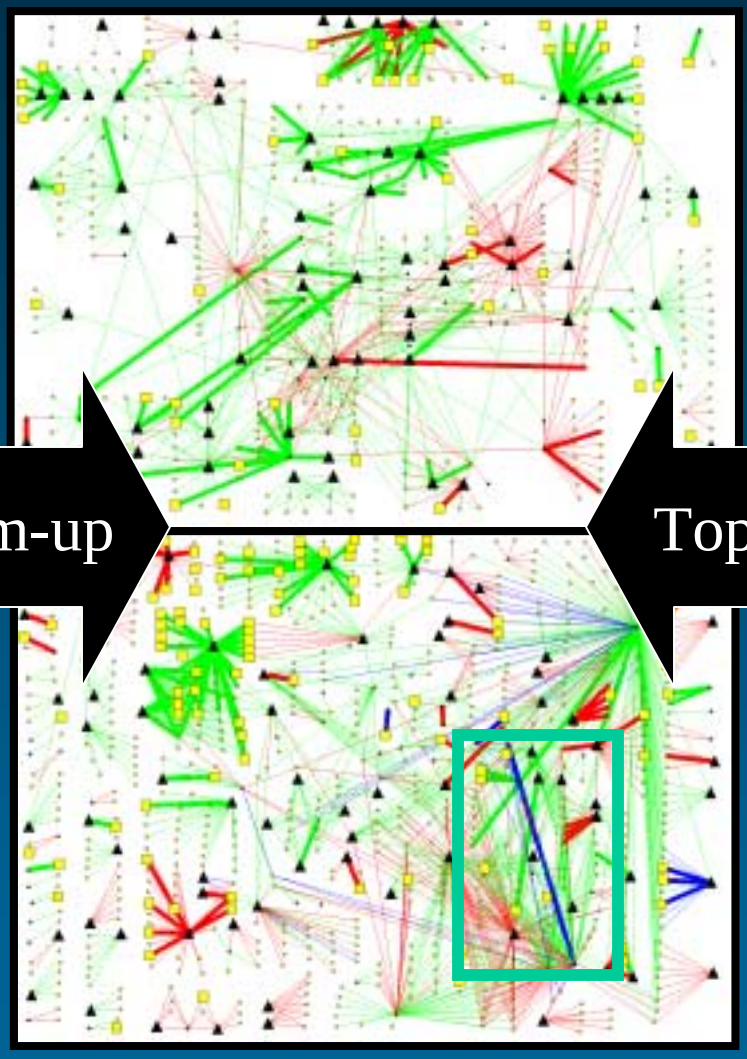


Literature



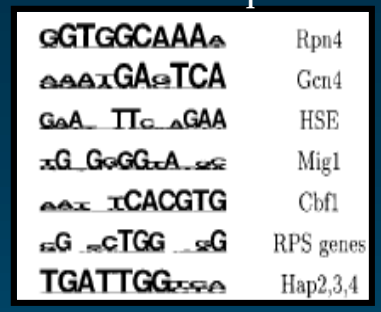
Databases

Bottom-up

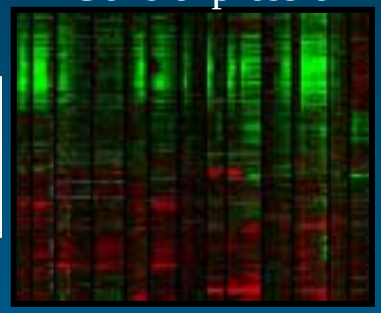


Top-down

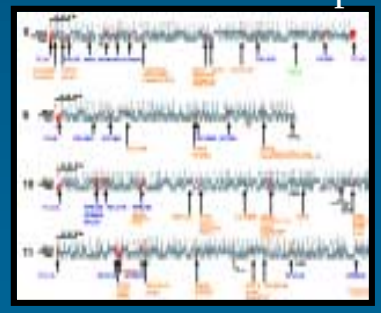
Promoter sequence



Gene expression



ChIP-Chip



Herrgard et al, *Curr Opin Biotechnol*, 2004

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

What is in a whole-cell reconstruction?

Genome:

- Annotated genes
- Gene location
- Regulatory regions
- Wobble base pairs

Biochemistry:

- Stereochemistry
- pH and pKa (charge)
- Elemental balance
- Charge balance
- Multiple reactions/enzyme
- Multiple enzymes/reaction

Transcription/translation:

- Gene to transcript to protein to reaction association
- Transcript half-lives
- tRNA abundances
- Ribosomal capacities

Physiology:

- Flux data
- Knock-outs
- Balanced functions
- Overall phenotypic behavior
- Location of gene product compartmentalization

Constraint-based Modeling:

Eliminating Impossible Phenotypes

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Criteria for Modern Biological Models

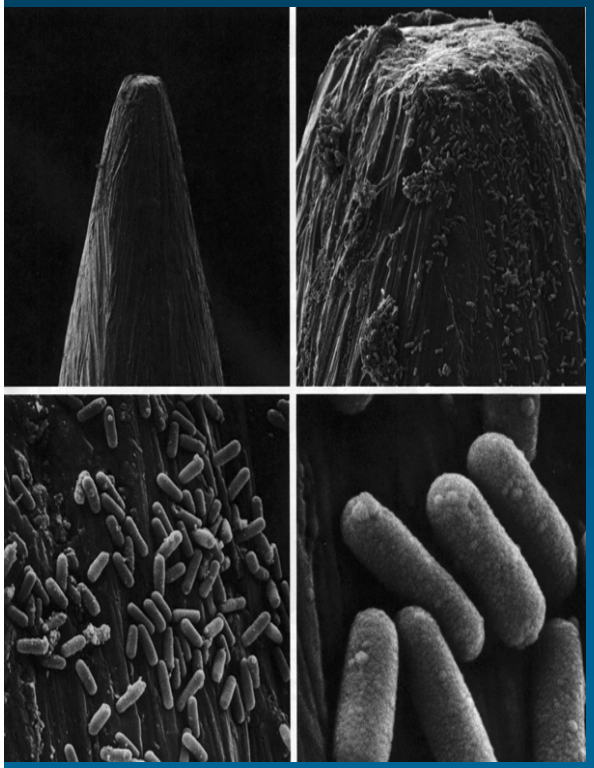
1. Must be data-driven
2. Based on large organism-specific databases (i.e. genome-scale)
3. Need to integrate diverse data types
4. Must be readily scalable to cell or genome-scale
5. Must account for inherent biological uncertainty

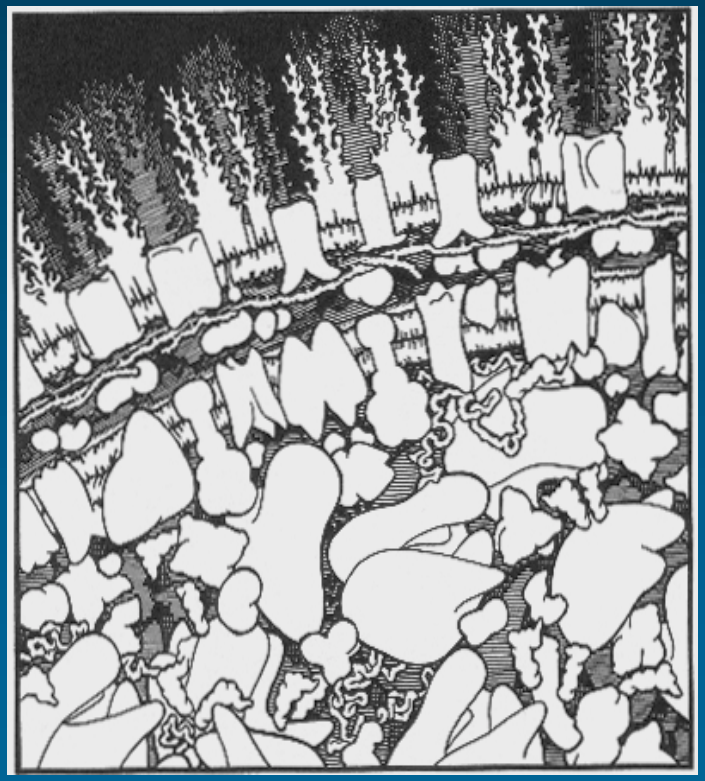
Challenges of Building Theory-based Models for Intracellular Functions at Genome-scale

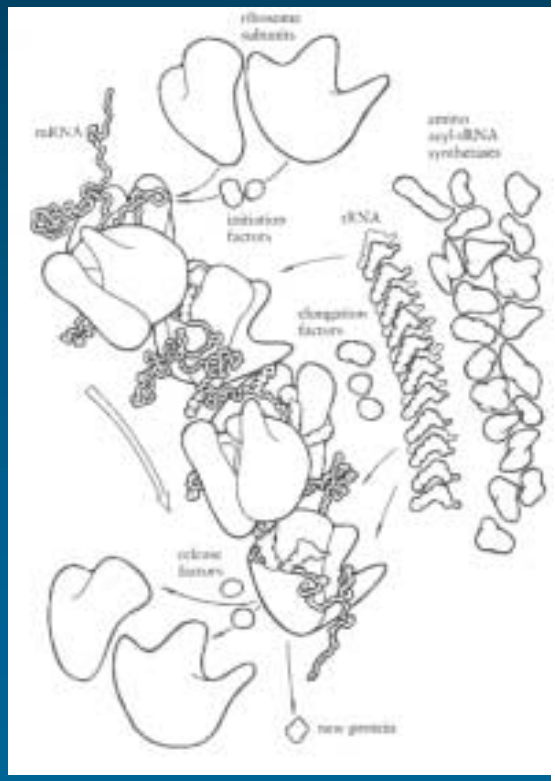
- Uncertain whether the physico-chemical laws apply
  - Crowding, small number of molecules, diffusion limitations
- Impossible to get the numerical values for the thousands of physical constants
- Parameters vary with:
  - Time (i.e. evolution)
  - Between individuals (i.e. polymorphism)

Functional states of networks

The constraint-based approach to analysis of complex biological systems







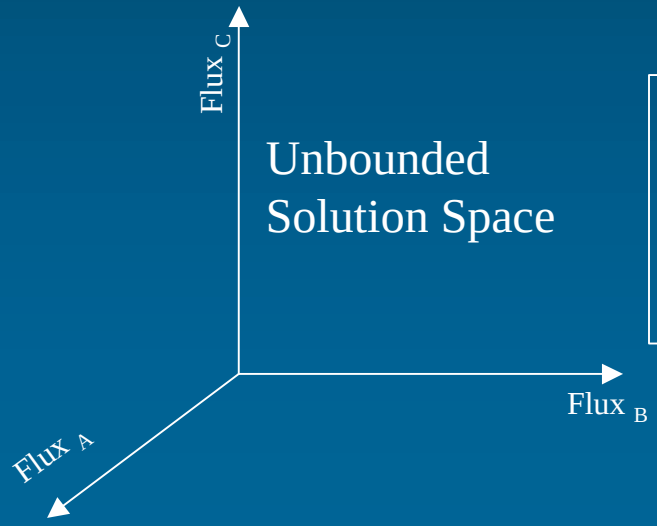
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Constraint-based Analysis

How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?

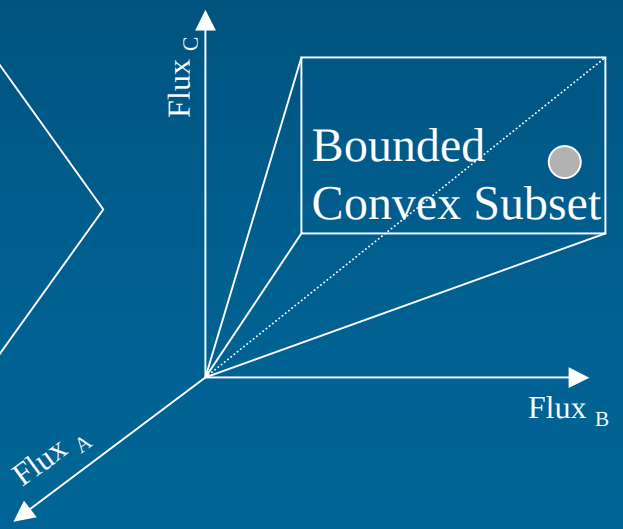
–Sherlock Holmes, A Study in Scarlet



Unbounded Solution Space

Constraints

(i) Stoichiometric
(ii) Thermodynamic
(iii) Capacity



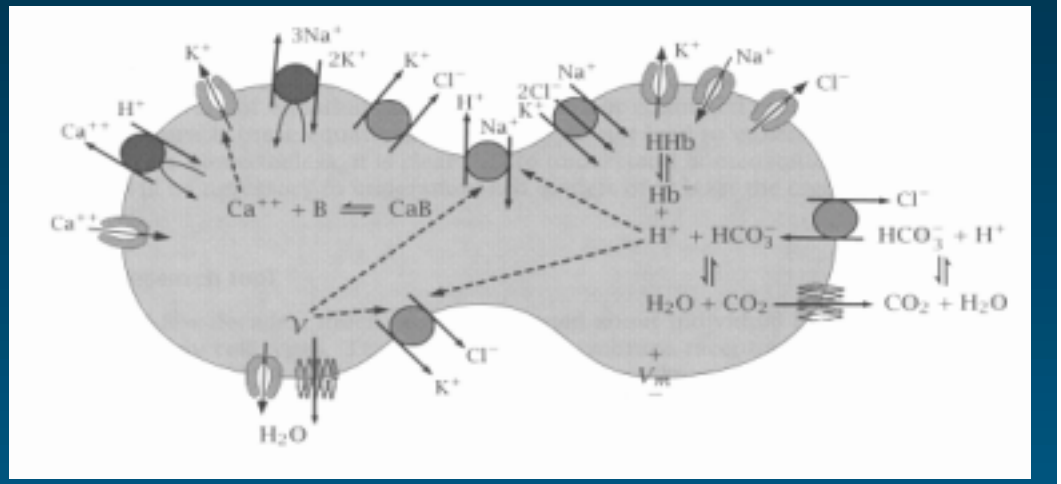
Bounded Convex Subset

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Factors Constraining Metabolic Function

- Connectivity:
  - Systemic stoichiometry
  - $\mathbf{S}\boldsymbol{v} = \mathbf{0}$
- Capacity:
  - Maximum fluxes
  - $v_i <$ maximum value
- P/C factors:
  - osmotic pressure, electro-neutrality, solvent capacity, molecular diffusion
- Rates:
  - Mass action, Enzyme kinetics, Regulation



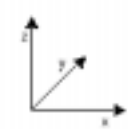
$$\frac{d[X_i]}{dt} = \sum V_{produce} - \sum V_{consume}$$

$$\frac{d\mathbf{X}}{dt} = \mathbf{S} \cdot \mathbf{v}$$

Approach:
application of successive constraints

Physico-Chemical and System-Specific Constraints:

- **Connectivity:** systemic stoichiometry
- **Thermodynamics:** directionality of the reactions
- **Capacity:** maximum flux rates
- **Kinetics:** time constants, mass action
- **Genetic Regulation**



R^n

1 ↓ Stoichiometry and linear algebra $\mathbf{S \cdot v = 0}$

Subspace of R^n

2 ↓ Reaction directions and convex analysis $v_i \geq 0$

Convex cone

3 ↓ Capacity constraints $v_i \leq v_{i,\max}$

Bounded convex subset

4 ↓ Relative saturation levels $k_M^i << k_M^j$

Union of convex subsets

Palsson, B.Ø., *Nature Biotechnology*, 2000 Nov, 18(11):1147-50.

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Mathematical Representation of Constraints

- Balances
  - Mass
  - Energy
  - Solvent capacity
- Bounds
  - Thermodynamics
  - Enzyme/transporter capacity
- Non-linear P/C phenomena

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

$$\Delta E = 0$$

$$\sum_i c_i \leq c_{\max}$$

$$0 \leq v_j \leq \infty$$

$$\alpha_j \leq v_j \leq \beta_j$$

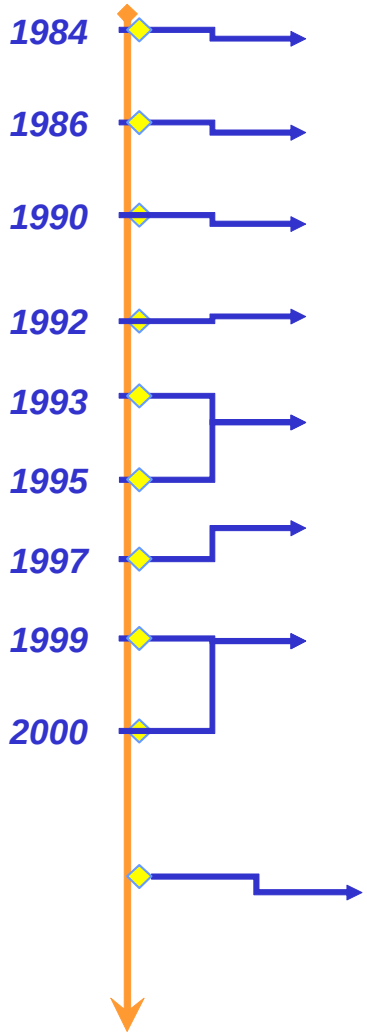
$$\pi = RT\left(\frac{c_i}{M_i} + Bc_i^2 + \text{☹}\right)$$

Flux Balance Analysis (FBA):
Interrogation of Genome-scale Network Reconstructions

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

History of Flux Balance Analysis (1984-2000)



Papoutsakis (1984) & Papoutsakis and Meyer (1985):
Used LP to calculate maximal theoretical yields

Fell and Small:
Used LP to study lipogenesis

Majewski and Domach:
Acetate overflow during aerobic growth

Savinell and Palsson:
Comprehensive assessment of FBA

Varma and Palsson (1993-1995):
Studied metabolic characteristics and capabilities of *E. coli*

Pramanik and Keasling:
Growth rate dependence on biomass composition

Edwards and Palsson (1999-2000):
Gene deletion studies of *E. coli*.
Analysis of *H. Influenzae* metabolic system.
Phenotypic Phase Plane Analysis

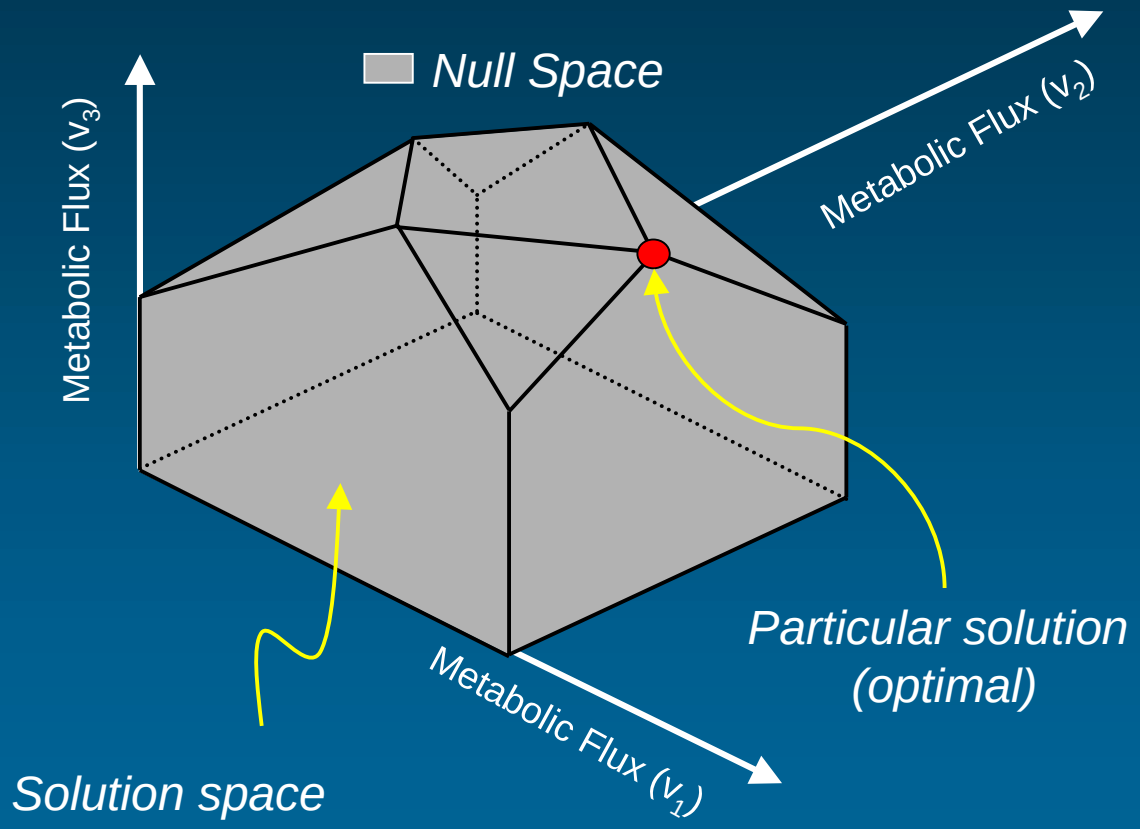
Schilling, Edwards and Palsson:
Integrating Flux-Balance Analysis with Metabolic Pathway Analysis

Burgard and Maranas:
Performance limits of *E. coli* subject to gene additions/deletions
Minimal reaction sets

Covert, Schilling and Palsson:
Incorporating regulatory constraints with flux-balance models

Edwards, J.S., et al. *Environmental Microbiology* (2002)

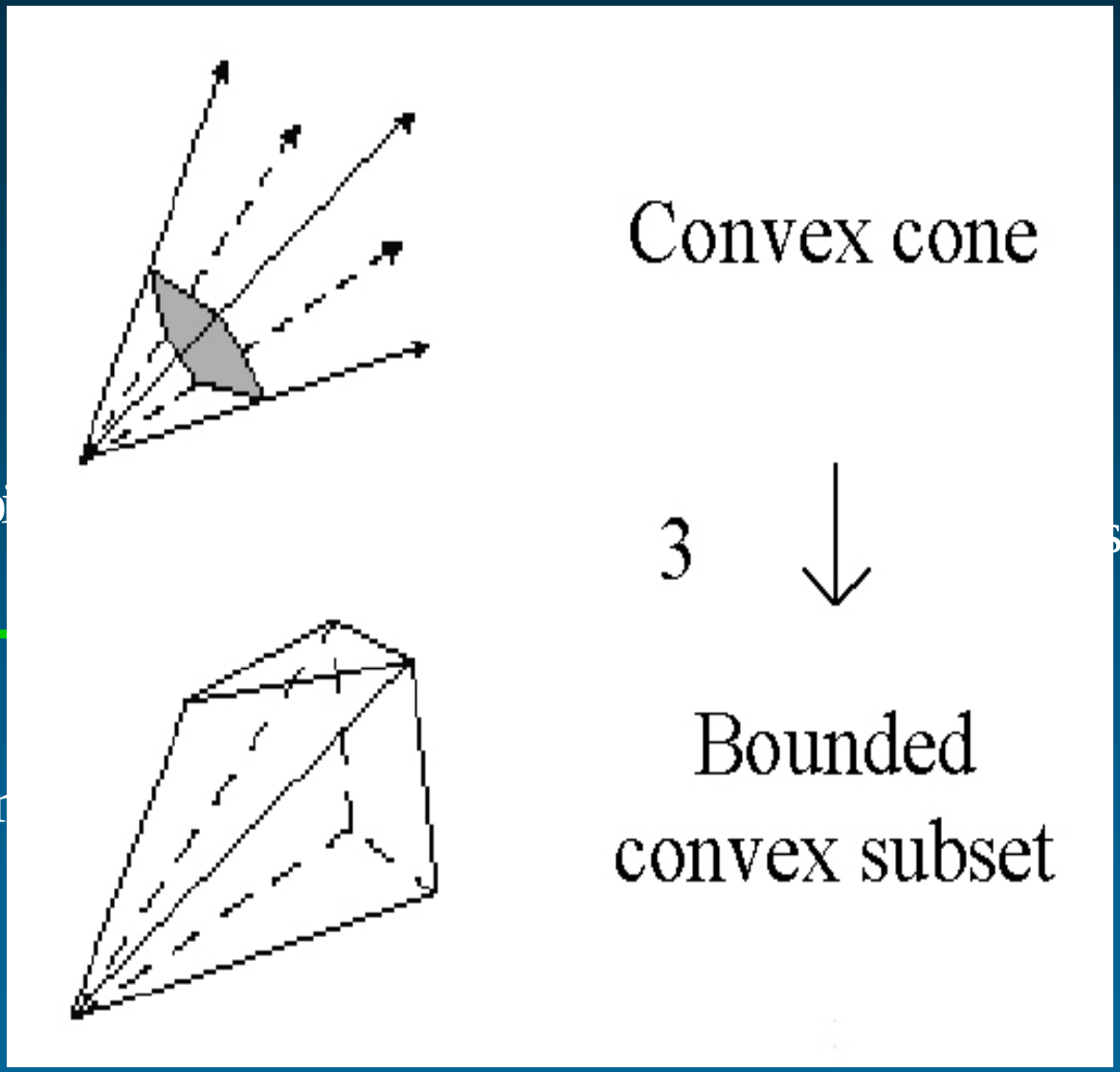
Linear Programming (LP): What is it?



University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Optimizing cellular growth (=max likelihood of survival?)



Mathematics

Maximize

$$Z = \sum_i c_i v_i = \mathbf{c} \cdot \mathbf{v}$$

Subject to

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

$$\alpha_j \le v_j \le \beta_j$$

Data-derived!

University of California, San Diego
Department of Bioengineering

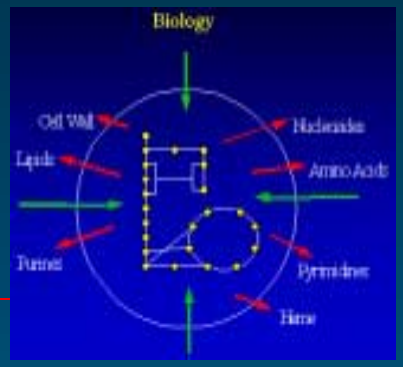
Systems Biology Research Group
http://systemsbiology.ucsd.edu

direction of 'better' survival

pressure directing evolution

"optimal" state (e.g. growth)

$$Z = \sum_i c_i v_i = \mathbf{c} \cdot \mathbf{v}$$



allowable functions

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

$$\alpha_j \le v_j \le \beta_j$$

reduction in capability

University of California, San Diego
Department of Bioengineering

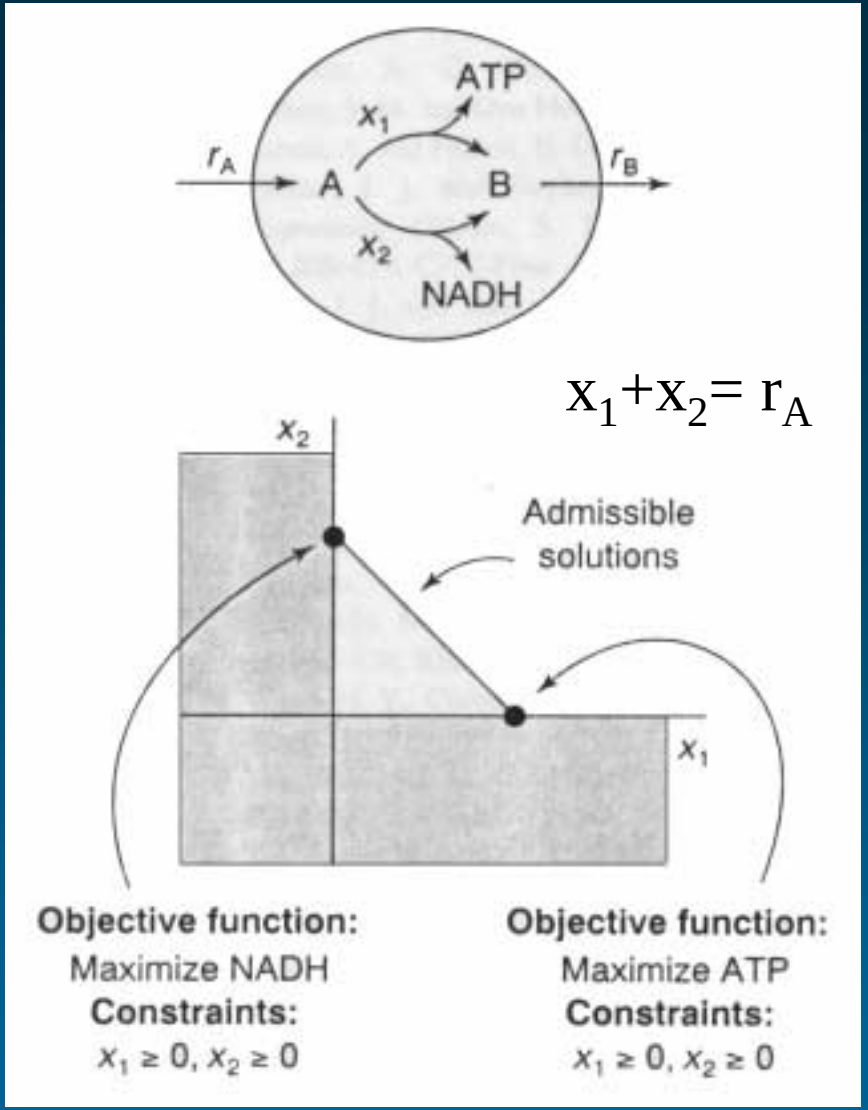
Systems Biology Research Group
http://systemsbiology.ucsd.edu

How does LP work? A very simple example

The solution space is the line of admissible in the positive orthant.

If we maximize ATP production the solution lies on the x-axis where all the flux would be through reaction x_1. Conversely, maximizing NADH production would give the point at the y-axis, where only reaction x_2 is active.

Note that the optimal solutions lie at the boundary of the admissible space.



Bonarius,et al TIBTECH vol 15:308 (1997)

Types of objective functions

- For basic exploration and probing of solution space
- To represent likely physiological objectives
- To represent bioengineering design objectives

Questions that can be addressed using LP: calculating optimal phenotypes

Minimize: *ATP production*
nutrient uptake
redox production
the Euclidean norm of the flux vector

Maximize: *biomass production (i.e. growth)*
metabolite production

Calculating Optimal States using LP: the objective function Z

Minimize Z, where

$$Z = \sum_i c_i v_i = \mathbf{c} \cdot \mathbf{v}$$

c is the vector that defines the weights for of each flux in the objective function, *Z*. The elements of **c** can be used to define a variety of metabolic objectives.

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Mathematical formulation of objective functions

$$\text{Minimize} Z = <\mathbf{c}\cdot\mathbf{v}> = \sum_i c_i v_i$$

Example: Minimize ATP production

$$\mathbf{v} = \begin{bmatrix} v_{G6P} \\ v_{F6P} \\ v_{ATP} \\ v_{NADH} \end{bmatrix} \rightarrow \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} \rightarrow \text{Minimize } Z$$

$$Z = 0\cdot v_{G6P} + 0\cdot v_{F6P} - 1\cdot v_{ATP} + 0\cdot v_{NADH}$$

The growth requirements

Metabolic demands of precursors and cofactors required for 1 g of biomass of *E. coli.*

These precursors are removed from the metabolic network in the corresponding ratios.

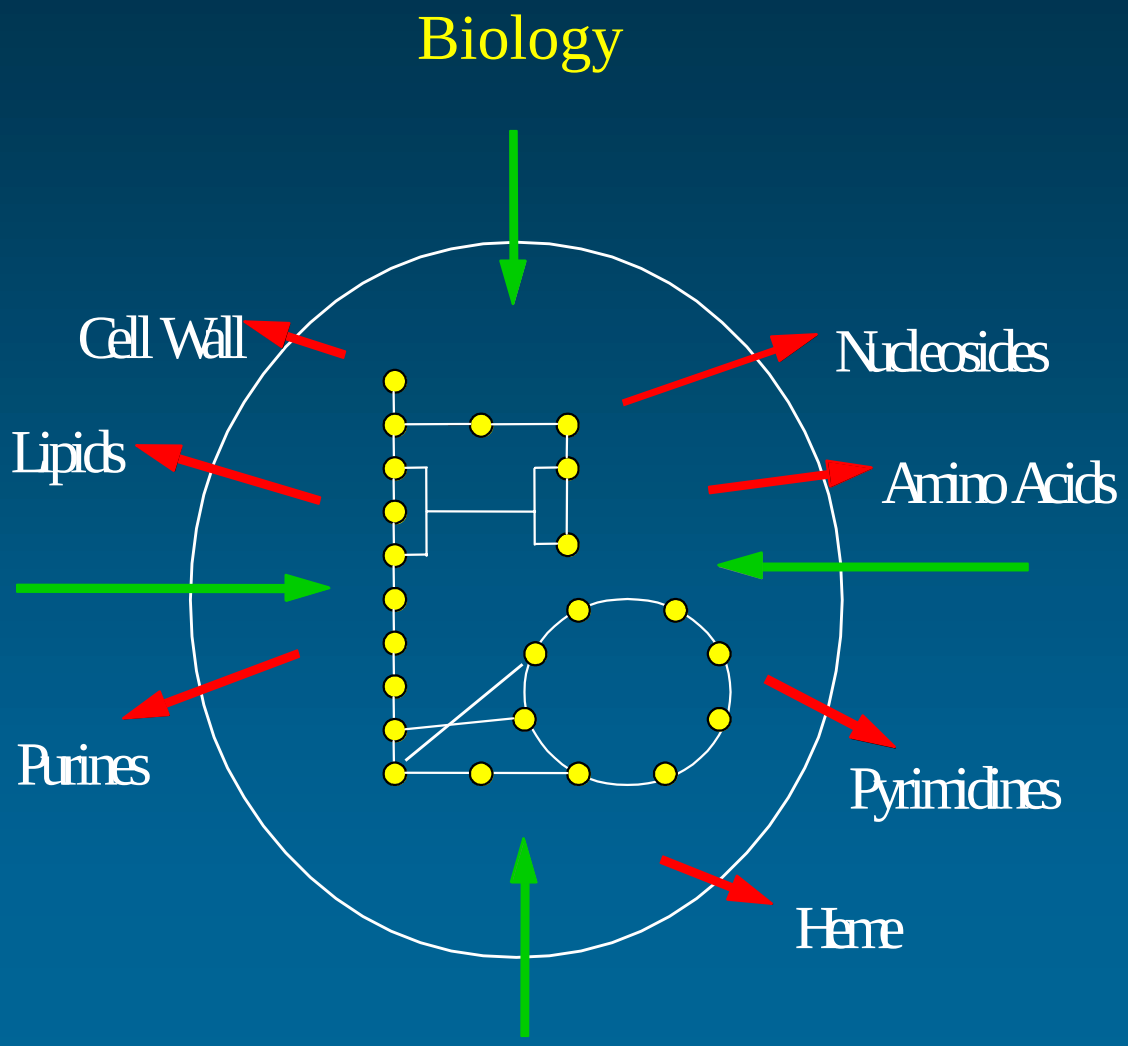
Thus, the objective function is:

$$Z = 41.2570\, v_{ATP} - 3.547 v_{NADH} + 18.225 v_{NADPH} + \ldots$$

Metabolite	Demand (mmol)
ATP	41.2570
NADH	-3.5470
NADPH	18.2250
G6P	0.2050
F6P	0.0709
R5P	0.8977
E4P	0.3610
T3P	0.1290
3PG	1.4960
PEP	0.5191
PYR	2.8328
AcCoA	3.7478
OAA	1.7867
AKG	1.0789

Neidhardt,et al. *Physiology of the Bacterial Cell* (1990)

Optimizing cellular growth (=max likelihood of survival?)



Mathematics

Maximize

$$Z = \sum_i c_i v_i = \mathbf{c} \cdot \mathbf{v}$$

Subject to

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

$$\alpha_j \le v_j \le \beta_j$$

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Biomass composition
Some issues

- Will vary from one organism to the next
- Will vary from one growth condition to another
- The optimum does not change much with changes in composition of a class of macromolecules, i.e. amino acid composition of protein
- The optimum does change if the relative composition of the major macromolecules changes, i.e. more protein relative to nucleic acids

The Constraints

Flux Balance Constraints

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

Capacity Constraints

$$0 \le v_i \le \infty$$

$$\alpha_j \le v_j \le \beta_j$$

$$v_k - \Delta v_k \le v_k \le v_k + \Delta v_k$$

All elementary reactions are irreversible, reversible reactions are defined as two separate strictly positive reactions

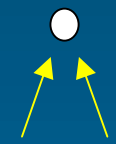
To constrain the upper and lower bound on specific fluxes. Used to set the maximal uptake rate if specific measurements are not available. i.e. maximal oxygen uptake

To set the flux level of a specific reaction. This constraint is used for fluxes that have been experimentally determined - typically the uptake rate of the carbon source

Determining constraints

- Experimental determination
- Estimation

Example: estimating oxygen uptake rates:



oxygen

Flux = $kDC = (2D/d)C_{sat}$

If Sh =2

Then the maximum oxygen uptake rate is

$N_{max} = 2\ (2.1 \times 10^{-5}\ cm^2/sec)(0.21mM)/1mm$

$= 8 \times 10^{-10} M/cm^2/sec$

If the area per cell is 12 $mm^2 = 12 \times 10^{-8}\ cm^2$

$N_{max} = 10^{-16} M/sec/cell$

Since one cell is about 1 fg = 10^{-12} mg

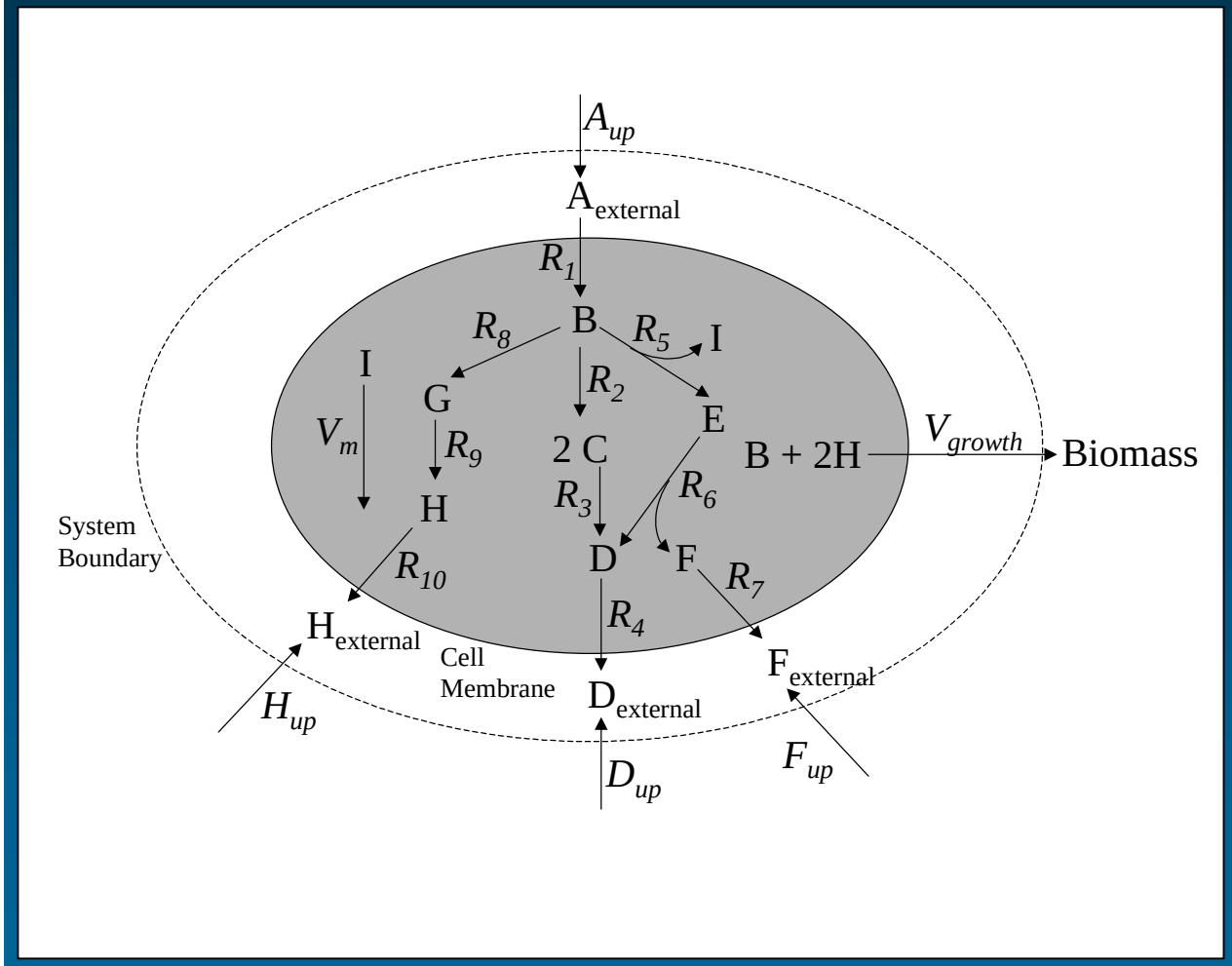
$N_{max} = 100$ mmol/cell/sec

Flux Balancing: an example of model formulation

University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

Example continued: the reconstructed network, its map and gene list



Gene	Enzyme	Flux
$Gene_1$	$Enzyme_1$	R_1
$Gene_2$	$Enzyme_2$	R_2
$Gene_3$	$Enzyme_3$	R_3
$Gene_4$	$Enzyme_4$	R_4
$Gene_5$	$Enzyme_5$	R_5
$Gene_6$	$Enzyme_6$	R_6
$Gene_7$	$Enzyme_7$	R_7
$Gene_8$	$Enzyme_8$	R_8
$Gene_9$	$Enzyme_9$	R_9
$Gene_{10}$	$Enzyme_{10}$	R_{10}
$Gene_A$	A Transporter	A_{up}
$Gene_D$	D Transporter	D_{up}
$Gene_F$	F Transporter	F_{up}
$Gene_H$	H Transporter	H_{up}

The flux balance equation for example network

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	V_m	V_{growth}	A_{up}	D_{up}	F_{up}	H_{up}			
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	R_1		0
B	1	−1	0	0	−1	0	0	−1	0	0	0	−1	0	0	0	0	R_2		0
C	0	2	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	R_3		0
D	0	0	1	−1	0	1	0	0	0	0	0	0	0	0	0	0	R_4		0
E	0	0	0	0	1	−1	0	0	0	0	0	0	0	0	0	0	R_5		0
F	0	0	0	0	0	1	−1	0	0	0	0	0	0	0	0	0	R_6		0
G	0	0	0	0	0	0	0	1	−1	0	0	0	0	0	0	0	R_7		0
H	0	0	0	0	0	0	0	0	1	−1	0	−2	0	0	0	0	R_8		0
I	0	0	0	0	1	0	0	0	0	0	−1	0	0	0	0	0	R_9		0
$\mathbf{A}_{external}$	−1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	R_{10}		0
$\mathbf{D}_{external}$	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	V_m		0
$\mathbf{F}_{external}$	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	V_{growth}		0
$\mathbf{H}_{external}$	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	A_{up}		0
																	D_{up}		0
																	F_{up}		0
																	H_{up}		0

Example continued:

The mass balances, the capacity constraints, and the objective function

Mass Balances	Flux Constraints
$\mathbf{B}: R_1 - R_2 - R_5 - R_8 - V_{growth} = 0$	$0 \le R_1 \le \infty$
$\mathbf{C}: 2R_2 - R_3 = 0$	$0 \le R_2 \le \infty$
$\mathbf{D}: R_3 + R_6 - R_4 = 0$	$0 \le R_3 \le \infty$
$\mathbf{E}: R_5 - R_6 = 0$	$0 \le R_4 \le \infty$
$\mathbf{F}: R_6 - R_7 = 0$	$0 \le R_5 \le \infty$
$\mathbf{G}: R_8 - R_9 = 0$	$0 \le R_6 \le \infty$
$\mathbf{H}: R_9 - R_{10} - 2V_{growth} = 0$	$0 \le R_7 \le \infty$
$\mathbf{I}: R_5 - R_2 - V_m = 0$	$0 \le R_8 \le \infty$
$\mathbf{A}_{external}: A_{up} - R_1 = 0$	$0 \le R_9 \le \infty$
$\mathbf{D}_{external}: D_{up} + R_4 = 0$	$0 \le R_{10} \le \infty$
$\mathbf{F}_{external}: F_{up} + R_7 = 0$	$Y_1 \le V_m \le Y_1$
$\mathbf{H}_{external}: H_{up} + R_{10} = 0$	$0 \le V_{growth} \le \infty$
	$Y_2 \le A_{up} \le Y_2$
	$-\infty \le D_{up} \le 0$
	$-\infty \le F_{up} \le 0$
	$-\infty \le H_{up} \le 0$

Objective Function

$$Z = V_{growth}$$

Illustrative example of basics of LP

Consider a system that has two metabolites A and B.

The production constraints on them are

$0 < A < 60$, and $0 < B < 50$

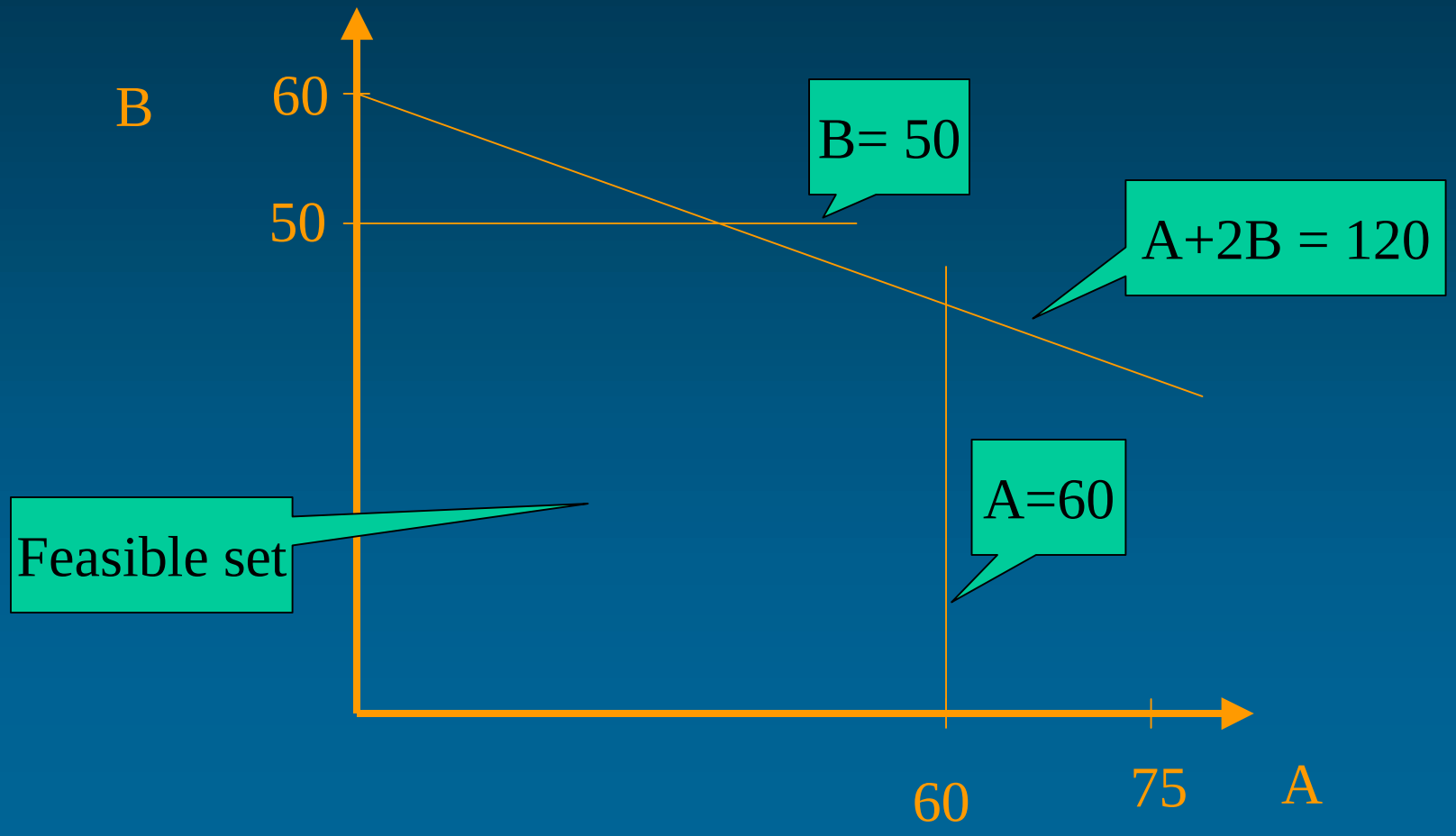
Additionally the capacity for producing them simultaneously is limited by:

$A + 2B < 120$

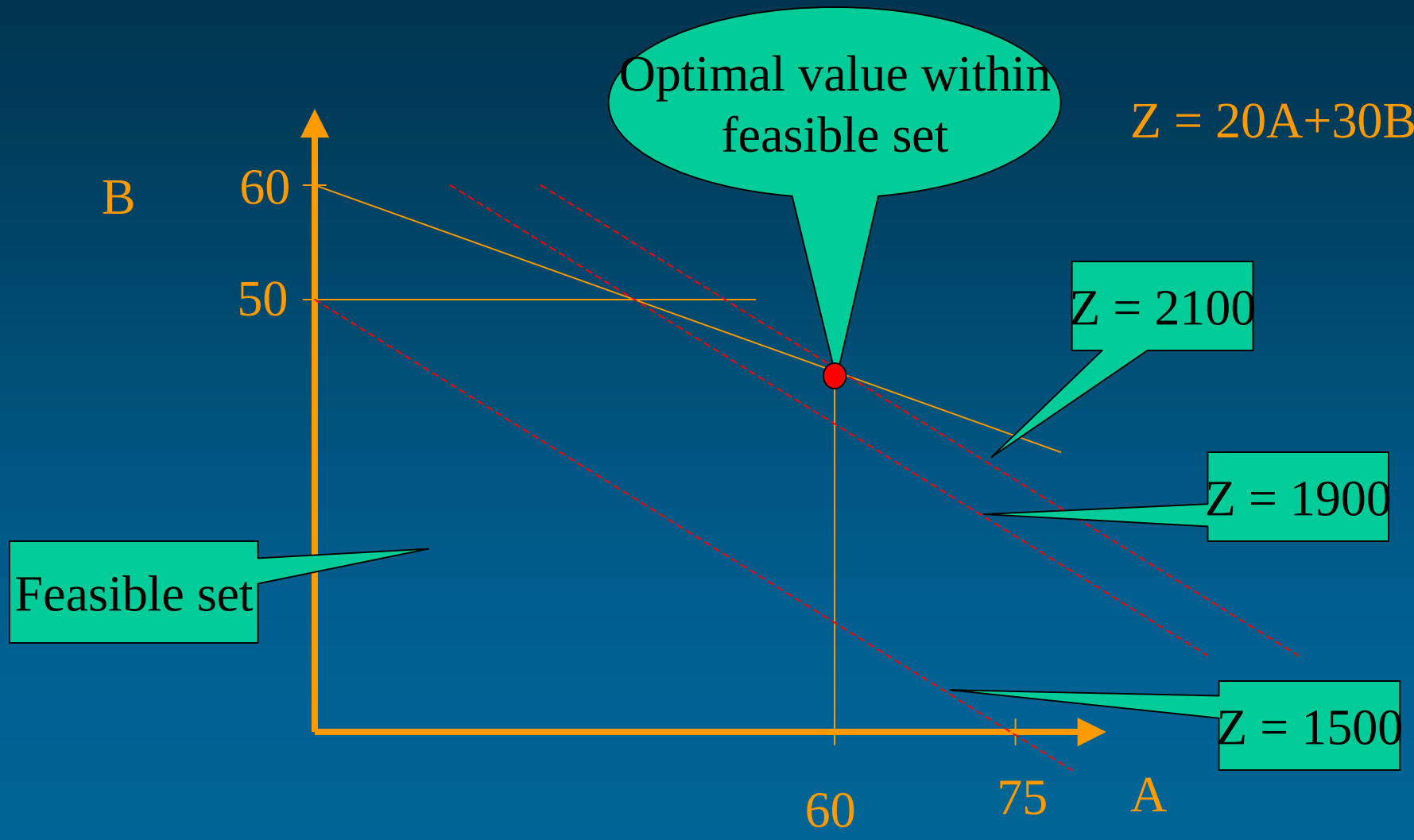
The objective function is

$Z = 20A + 30B$

LP ➔ Graphical representation of feasible set



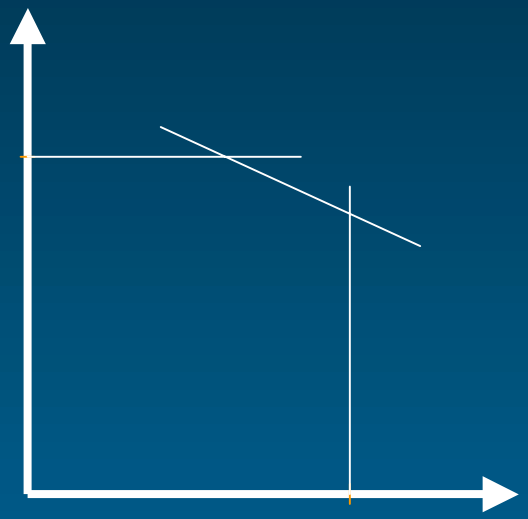
LP ➔ Graphical representation of the objective function



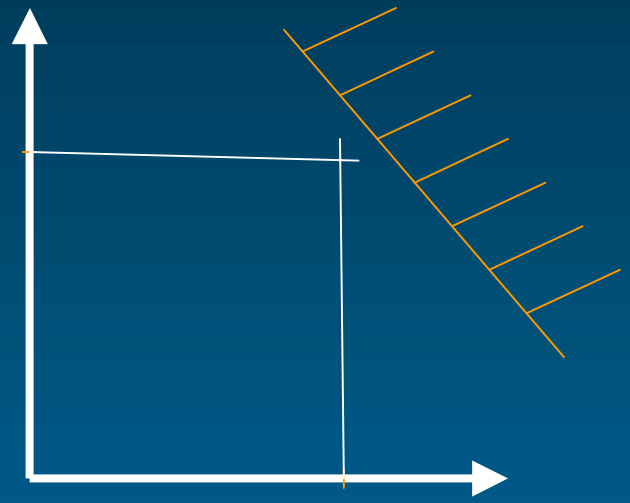
University of California, San Diego
Department of Bioengineering

Systems Biology Research Group
http://systemsbiology.ucsd.edu

LP ➔ Types of solutions: feasible and non-feasible solutions



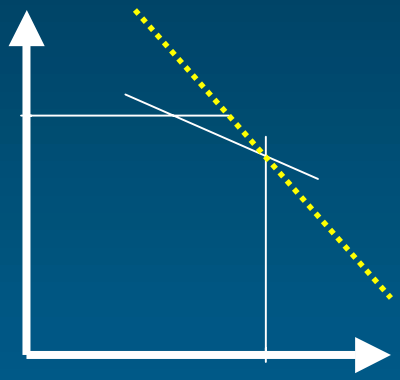
Feasible: solutions possible within all stated constrains



Not feasible: solutions not possible within all stated constrains

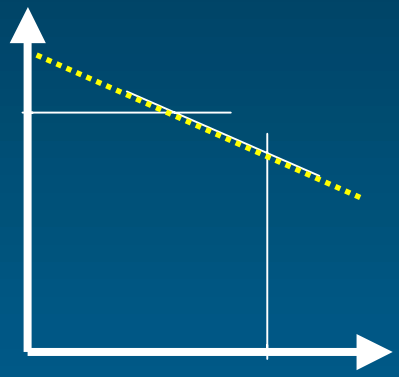
LP ➔ Types of solutions: the impact of the objective function

Single solution



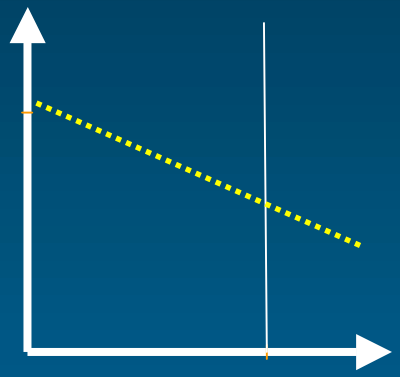
Optimal solution in a corner

Degenerate solution



Optimal solution along an edge

No solution



Optimal solution not found--region unbounded

..... Lines of constant Z

Next Lecture…

- *Lessons learned from genome-scale constraint-based models*
- *The future of constraint-based modeling and associated techniques*