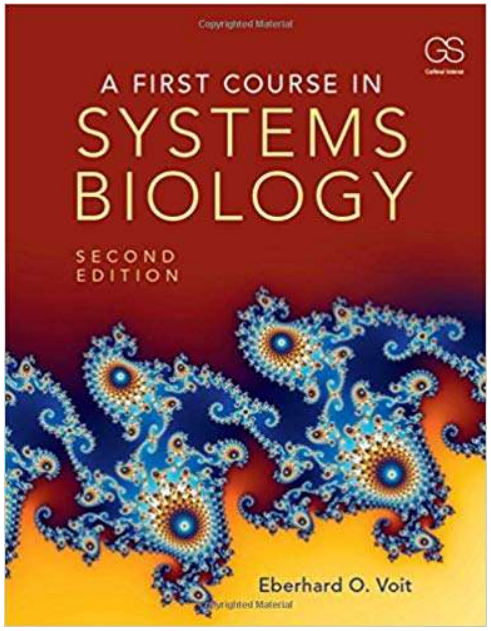
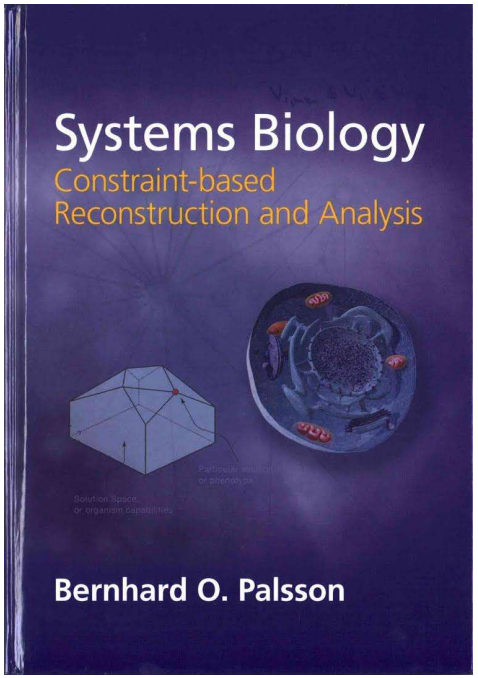
COBRA Optimization Methods for Analysis of Genome-scale Models and *in-silico* Strain Design

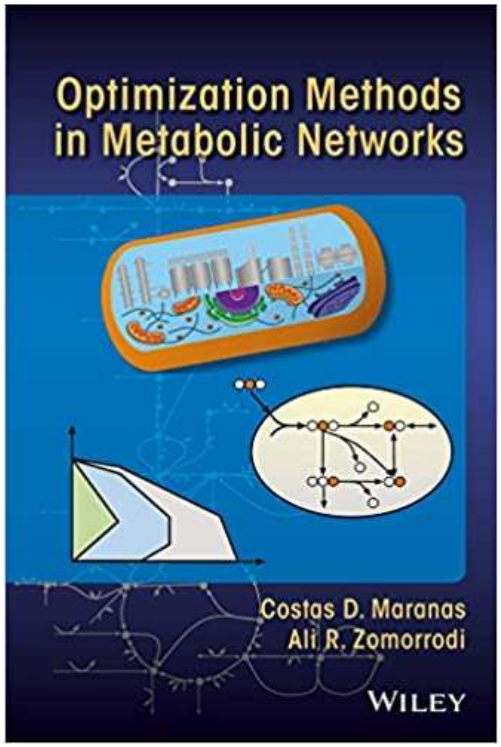
Payam Setoodeh

School of Chemical and Petroleum Engineering, Shiraz University

References

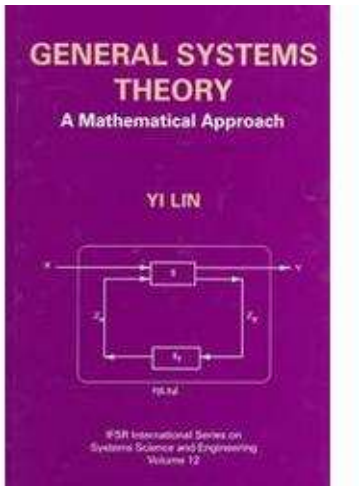

Copyrighted Material
A FIRST COURSE IN
SYSTEMS
BIOLOGY
SECOND
EDITION
Eberhard O. Voit
Copyrighted Material


Systems Biology
Constraint-based
Reconstruction and Analysis
Bernhard O. Palsson

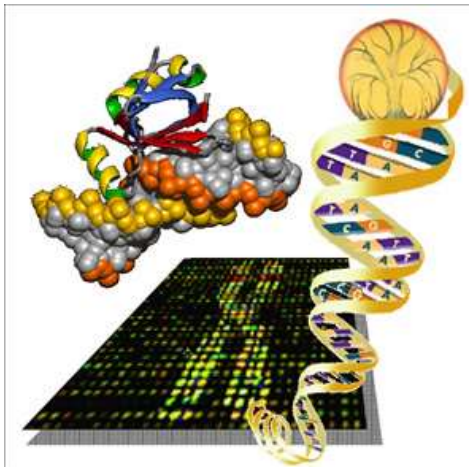

Optimization Methods
in Metabolic Networks
Costas D. Maranas
Ali R. Zomorrodi
WILEY

Preface

Systems and Control Theory



Holistic Biology



Systems Biology

Systems Analysis

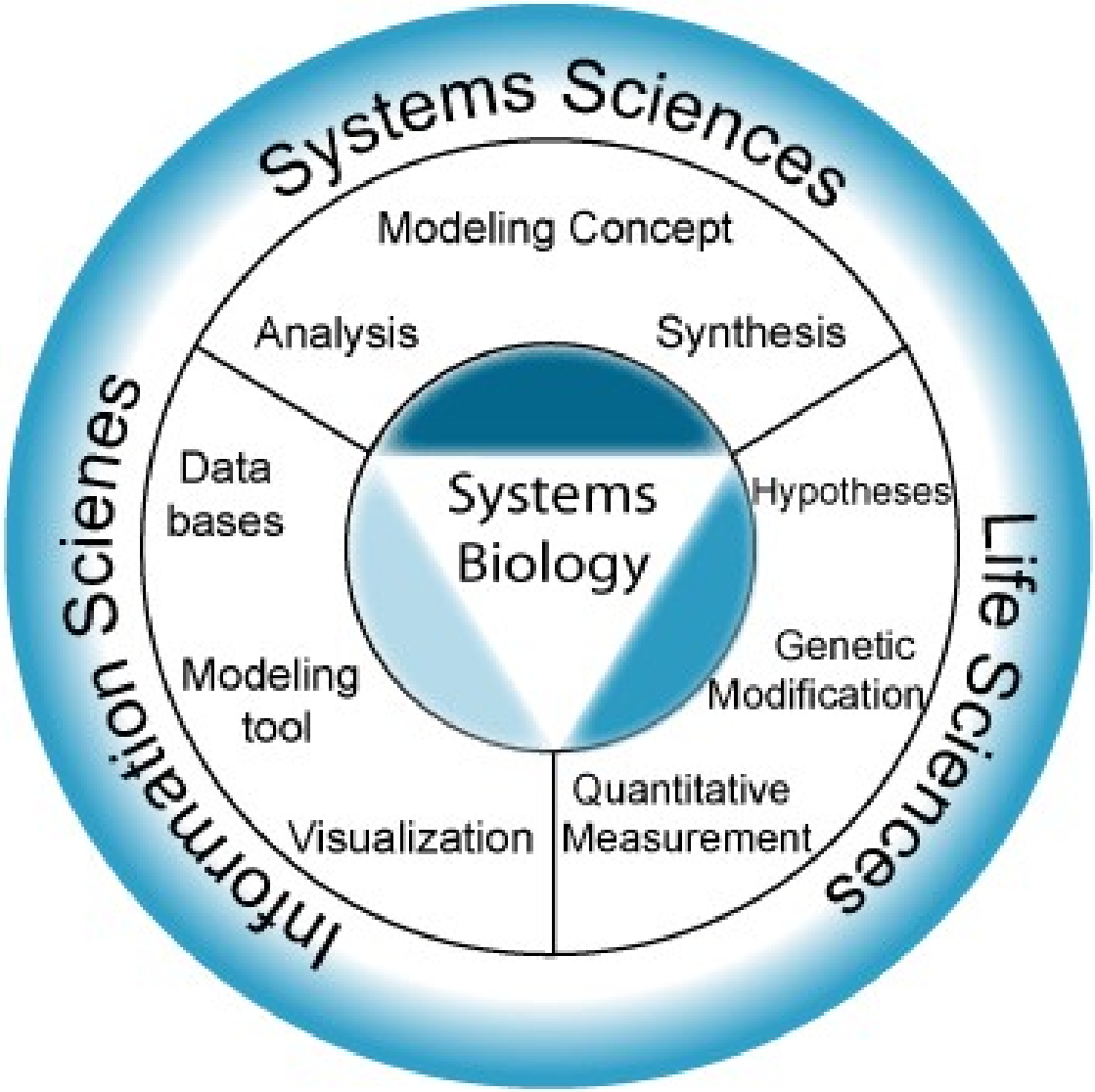
***In silico* Experimentation**

Conceptual Analysis

Modeling and Simulation

Understanding The Life

Preface


Systems Sciences
Modeling Concept
Analysis
Synthesis
Data bases
Systems Biology
Hypotheses
Modeling tool
Genetic Modification
Visualization
Quantitative Measurement
Information Sciences
Life Sciences

Systems Biology

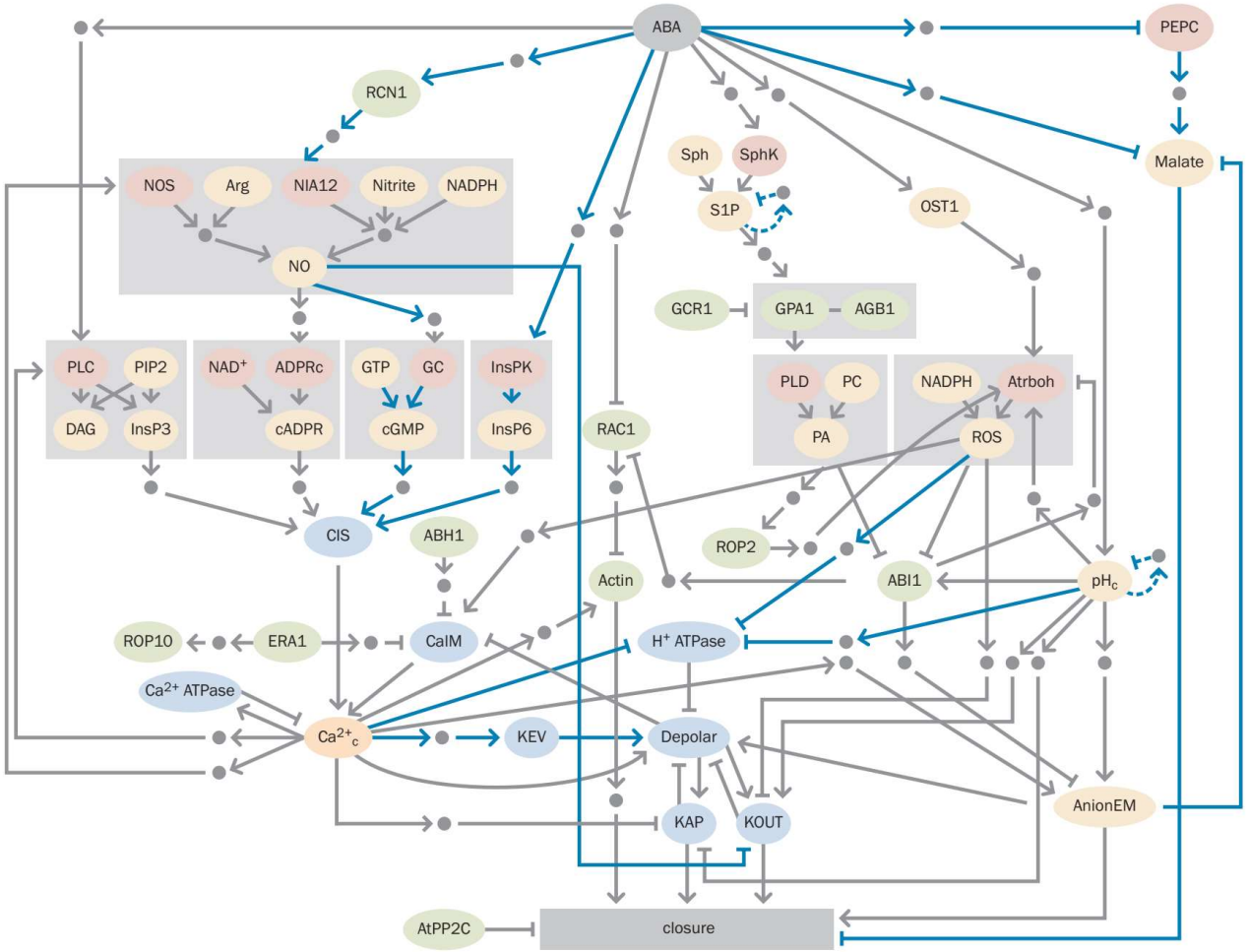
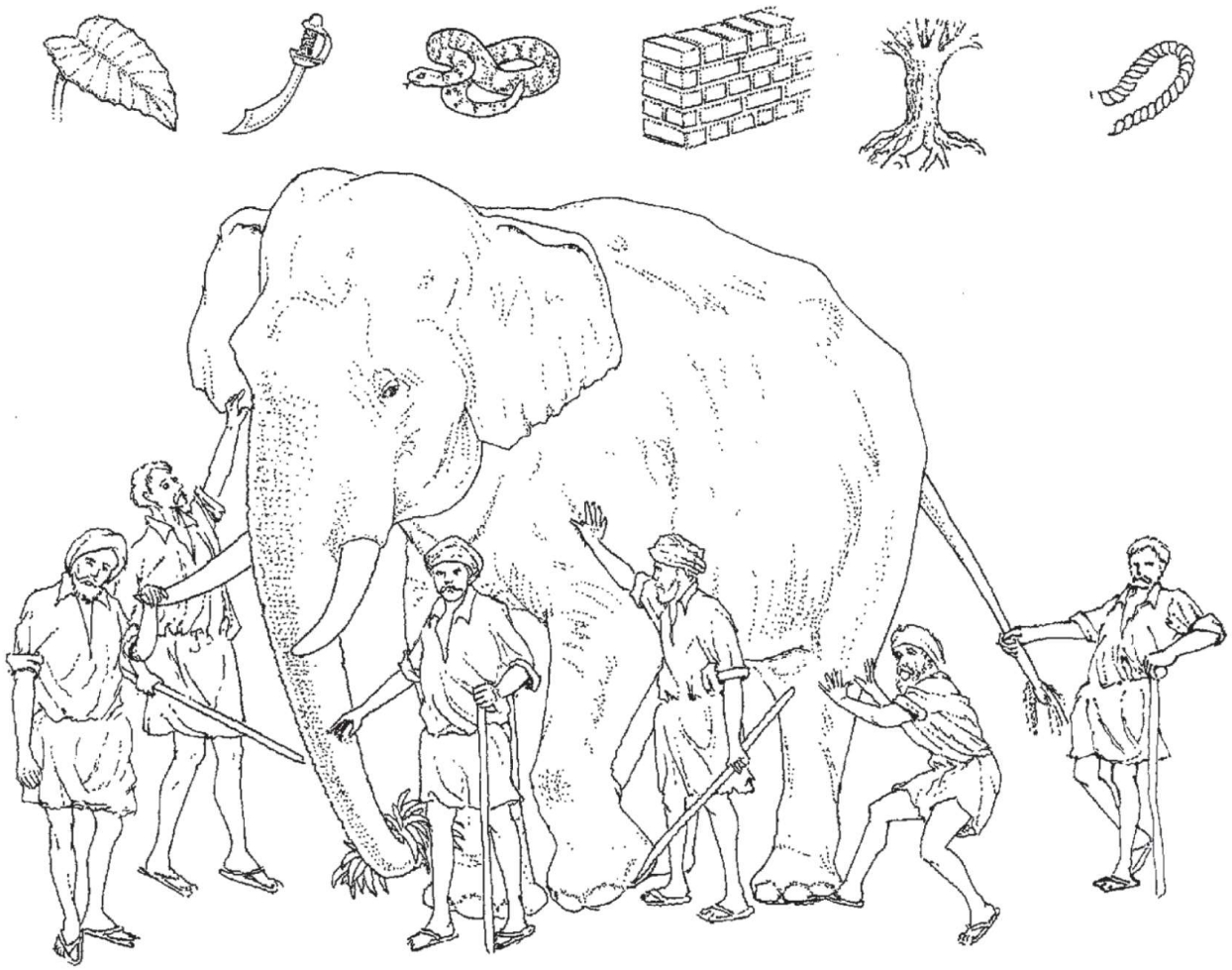


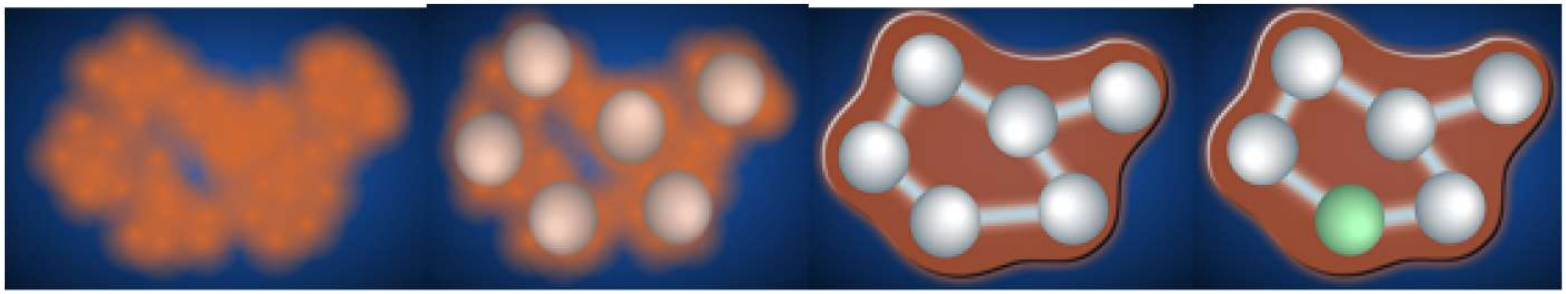
Diagram of a complicated system of molecules that coordinate the response of plants to drought. While the details are not important here, we can see that a key hormone, called abscisic acid (ABA), triggers a cascade of reactions that ultimately promote the closure of stomata and thereby reduce water evaporation [1]. Even a narrowly defined response like this closure process involves a complicated control system that contains a multitude of molecules and their interactions. In turn, this system is just one component within a much larger, physiological stress response system

Systems Biology



Information about isolated parts of a system alone does not always reveal the true nature of the system. An old story of six blind Indian men trying to determine what they touch is a parable for the dangers of scientific silos and the lack of good communication.

Holistic Thinking



- Discrete biological function can only rarely be attributed to an individual molecule.
- *Emergent phenomena* in complex cellular networks are not detectable by reductionist approaches.
- The mission is to understand the *interactions* that contribute to the structure and function of a living cell.

Goal of Systems Biology

- To provide quantitative descriptions for *elegant complexities* observed in living organisms
  - Hidden layers of feedback
  - Cause-effect loops
  - Flow of information across the hierarchy of control levels
- The plan is to move from

$$\left[\begin{array}{c}\text{Description}\\ \Downarrow\\ \text{Representation}\\ \Downarrow\\ \text{A comprehensive understanding of bioprocesses}\end{array}\right]$$

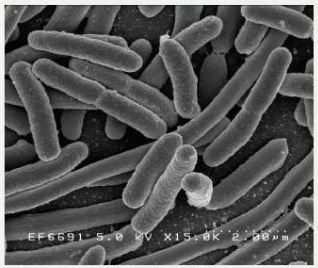
Molecular to Systems Biology

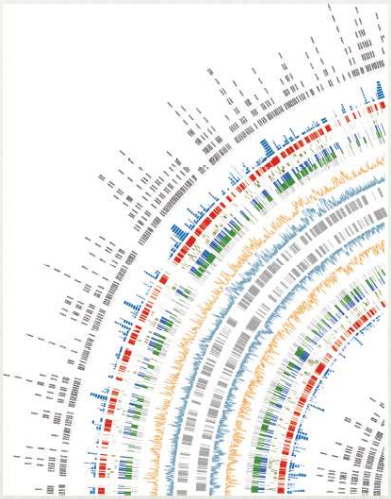
Reductionist approach:

Components Biology

High-throughput analytical chemistry:
genomics
transcriptomics
proteomics

20th century biology





Integrative approach:

Systems Biology

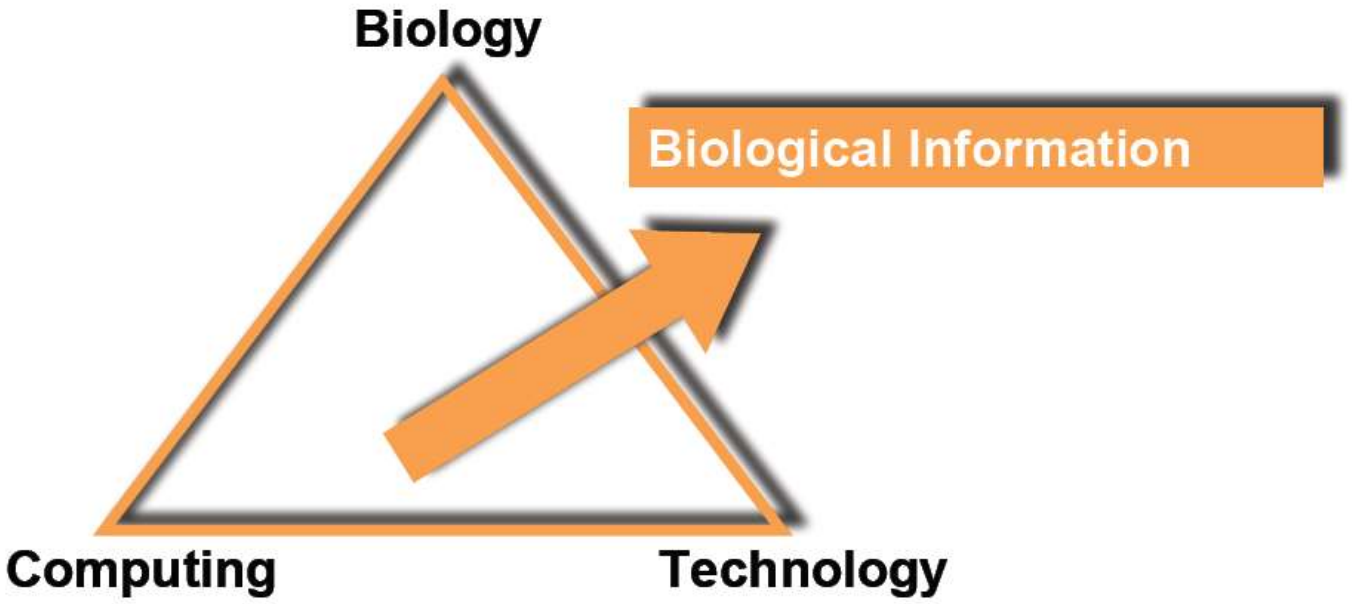
Integrative analysis:
bioinformatics
mathematical models
in silico simulation

21st century biology

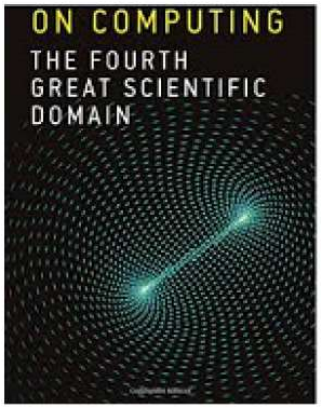
Nature Biotechnology, 18:1147, 2000

The Systems Biology Triangle

CC-SPMD
Competence Center for
Systems Physiology
and Metabolic Diseases



Computing as a Great Scientific Domain



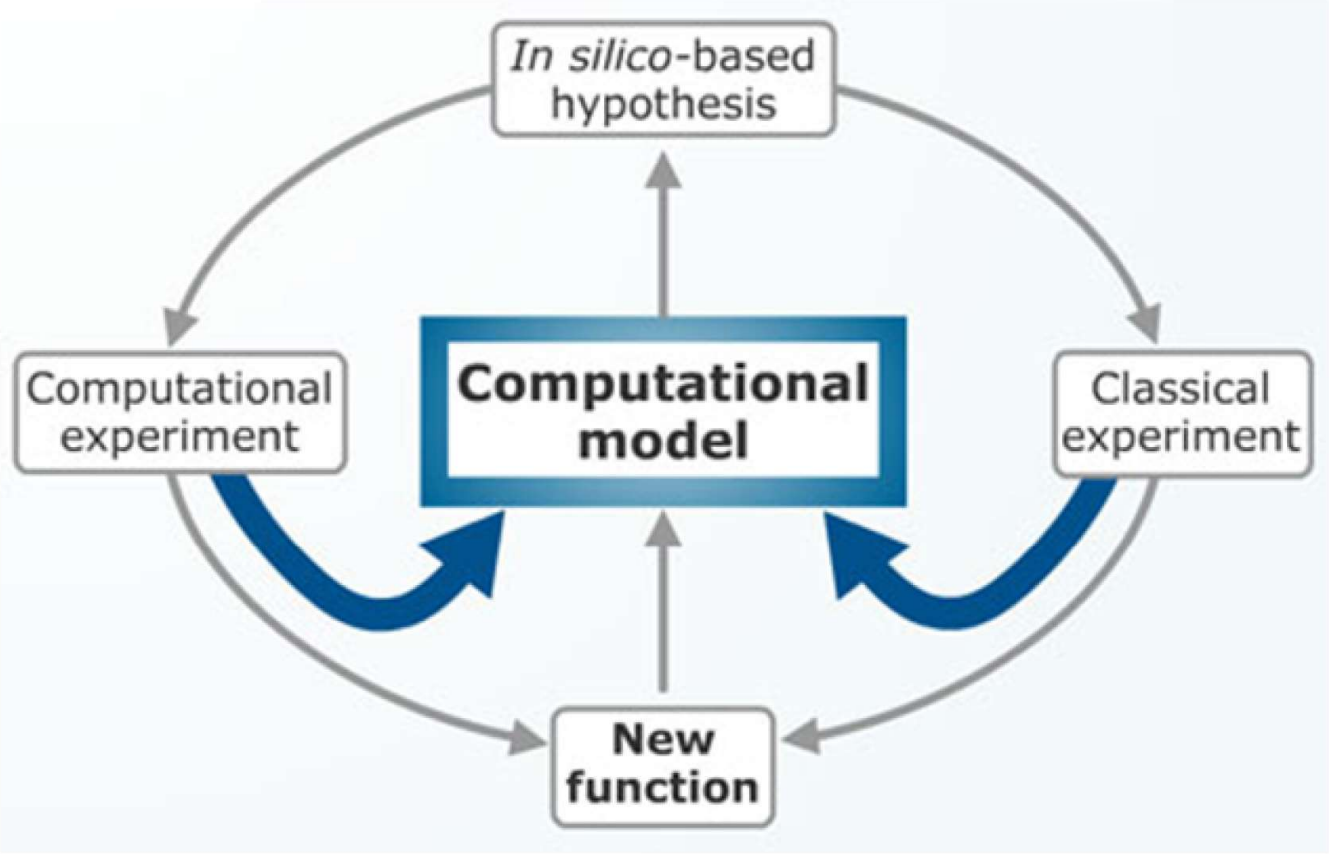
Paul Rosenbloom (2012)

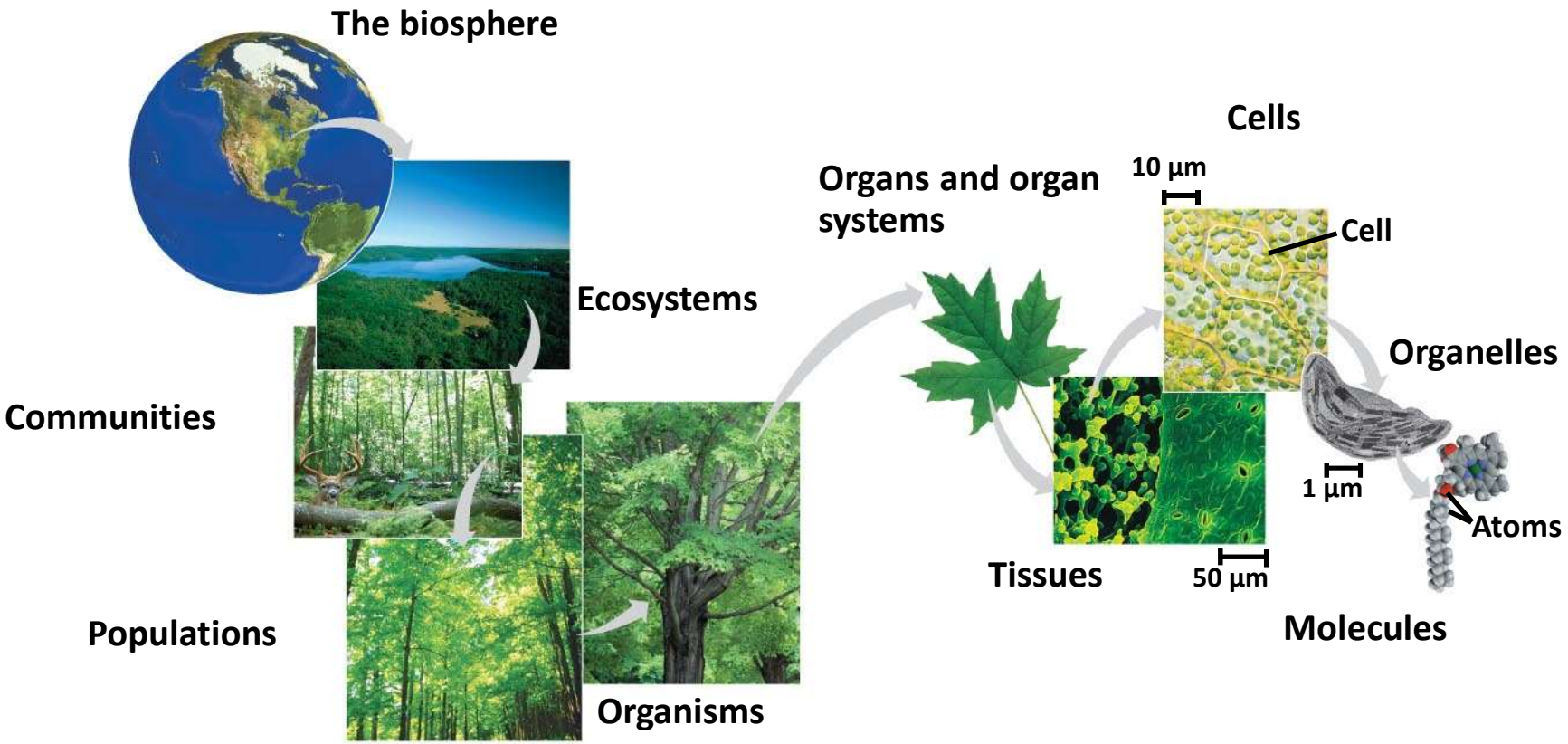
- On par with the physical, life, and social sciences.
- Computing as a great scientific domain concerns the *understanding* and *shaping* of the *interactions* among a coherent, distinctive and extensive body of *structures* and *processes*.
- Computing subsumes *mathematics*.

Computational Systems Biology (Rosenbloom, 2012)

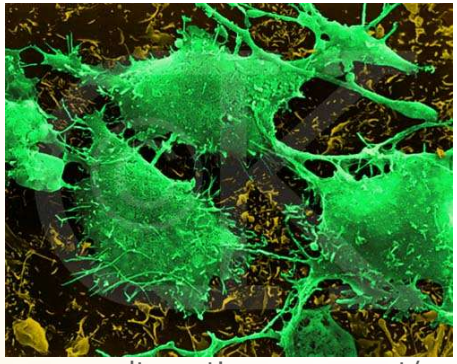
- Computing acts as a *full and equal partner* with biology in a symmetric relationship.
- *Beyond* just providing tools for use by biology.
- It aids in understanding both the discipline and the *disciplinary structure* within computing.
- It provides insight into how it might be possible to *rethink* the focus and boundaries.

Computational Systems Biology

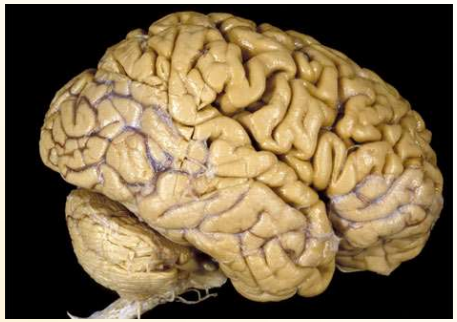




Recall: Generic Approach of CSB

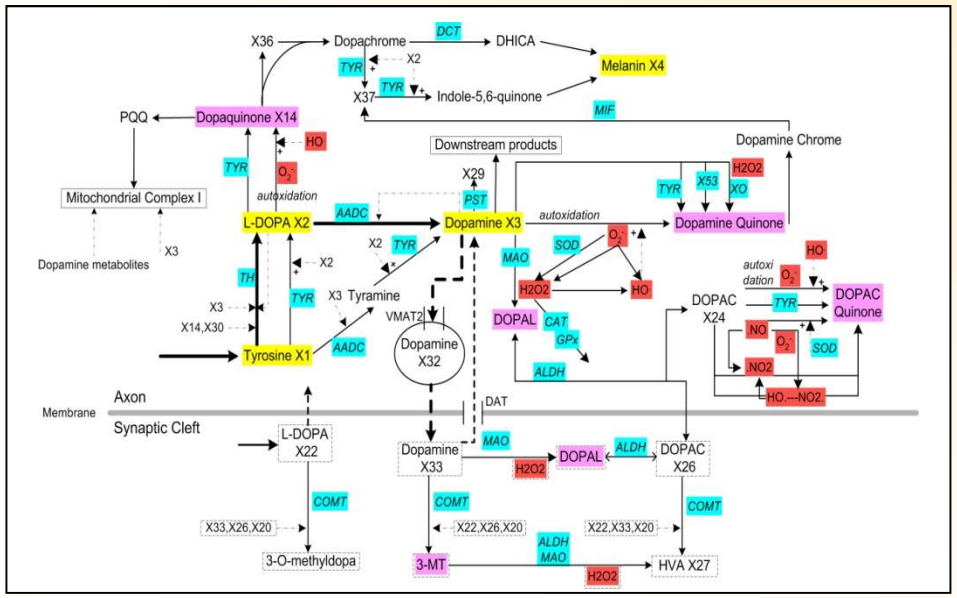


www.alternative-cancer.net/



science.nationalgeographic.com

e. o. voit

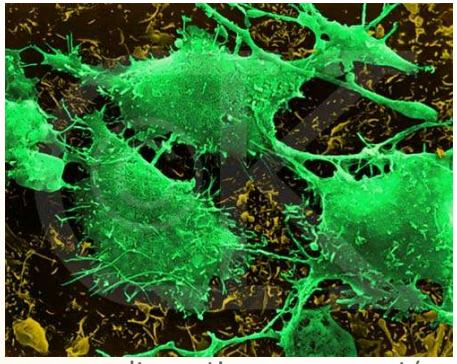


$$\frac{dX_{12}}{dt} = \gamma_{0,12} X_{24}^{f_{0,12,24}} X_{30}^{f_{0,12,30}} + \gamma_{51,12} X_{12}^{f_{51,12,12}} X_{28}^{f_{51,12,28}} X_{51}^{f_{51,12,51}} X_{52}^{f_{51,12,52}}$$

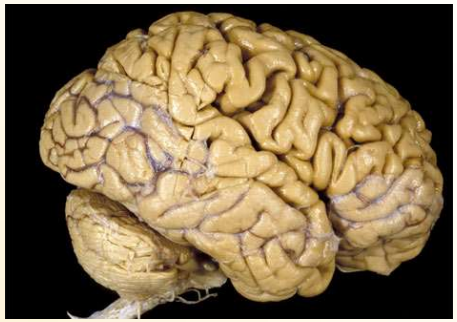
$$+ \gamma_{4,17} X_{4}^{f_{4,17,4}} X_{22}^{f_{41,17,22}} + \gamma_{6,17} X_{6}^{f_{6,17,6}} X_{22}^{f_{6,17,22}}$$

```
%ODEs
X12' = gam0_12 X24^f0_12_24 + gam51_12 X12^f51_12_12 >>
      >> * X28^f51_12_28 X51^f51_12_51 X52^f51_12_52 >>
      >> + gam4_17 X4^f4_17_4 X22^f41_17_22 >>
      >> + gam6_17 Xf6_17_6 X22^f6_17_22
&& X22 X30 X51 X52
!! X12
```

Recall: Generic Approach of CSB

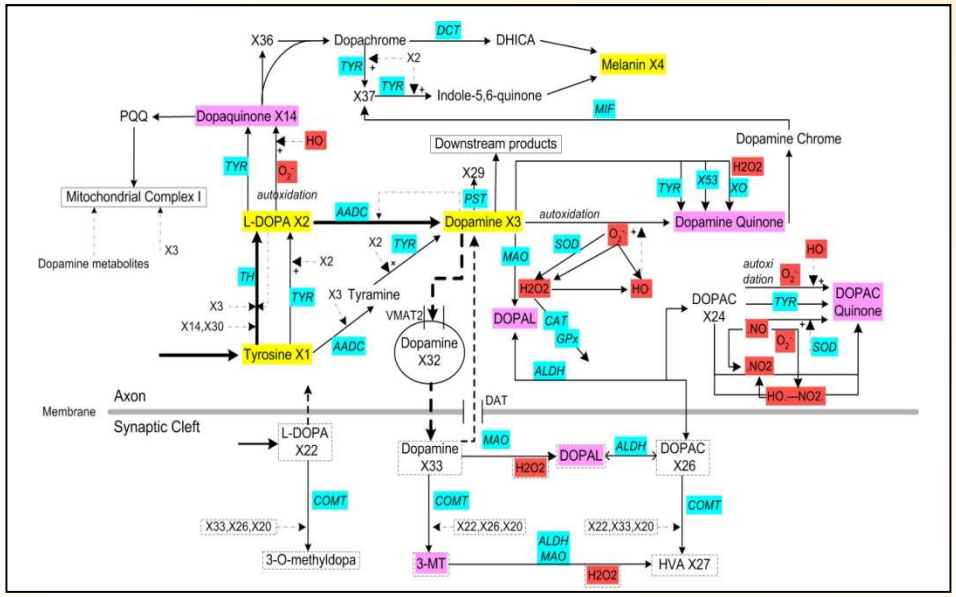


www.alternative-cancer.net/



science.nationalgeographic.com

e. o. voit



$$\frac{dX_{12}}{dt} = \gamma_{0,12} X_{24}^{f_{0,12,24}} X_{30}^{f_{0,12,30}} + \gamma_{51,12} X_{12}^{f_{51,12,12}} X_{28}^{f_{51,12,28}} X_{51}^{f_{51,12,51}} X_{52}^{f_{51,12,52}}$$

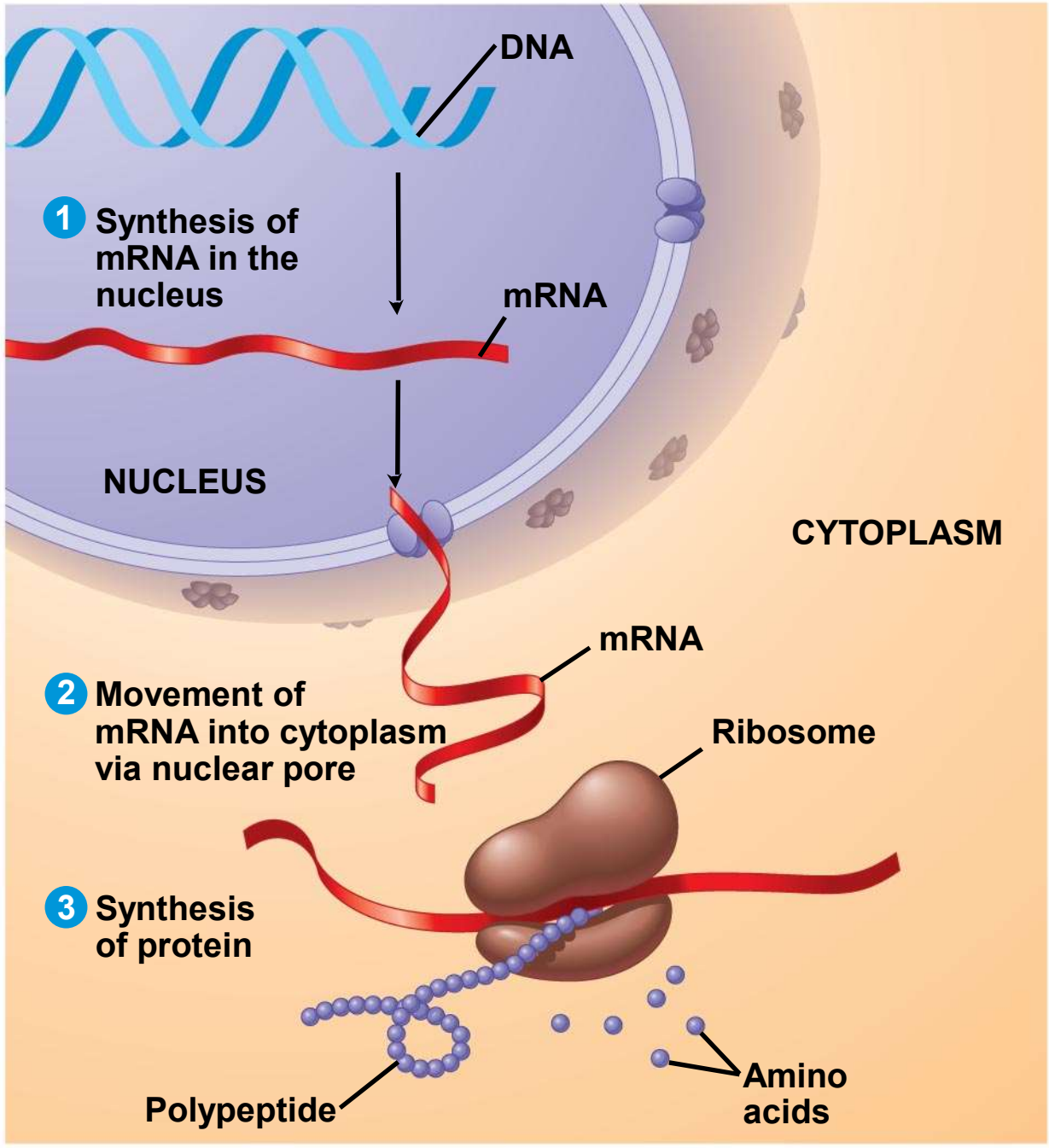
$$+ \gamma_{4,17} X_{4}^{f_{4,17,4}} X_{22}^{f_{41,17,22}} + \gamma_{6,17} X_{6}^{f_{6,17,6}} X_{22}^{f_{6,17,22}}$$

```
%ODEs
X12' = gam0_12 X24^f0_12_24 + gam51_12 X12^f51_12_12 >>
      >> * X28^f51_12_28 X51^f51_12_51 X52^f51_12_52 >>
      >> + gam4_17 X4^f4_17_4 X22^f41_17_22 >>
      >> + gam6_17 Xf6_17_6 X22^f6_17_22
&& X22 X30 X51 X52
!! X12
```

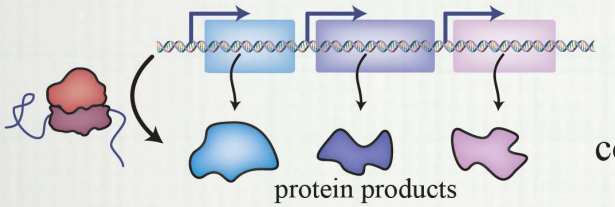
Mathematical Models

All models are abstractions of real systems and processes

Nevertheless, they serve as tools for engineers and scientists to develop an understanding of important systems and processes using mathematical equations

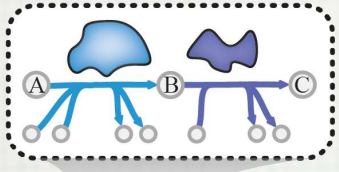


Hierarchy and systems biology: from genes to phenotypes

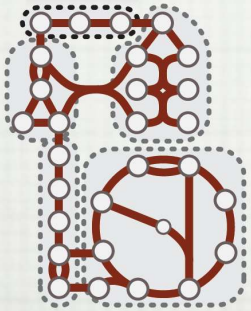


genotype

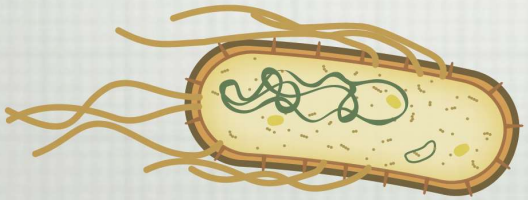
components



small-scale modules



large-scale modules

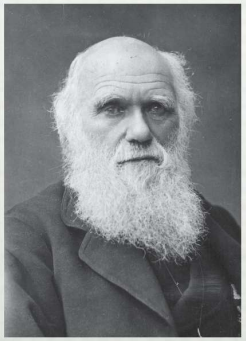


phenotype (physiology)

Ludwig Boltzmann

Chemical causation:
Can apply P/C laws and get causality on a small scale

Systems biology: emphasis on modules and understanding of how coherent physiological functions arise from the totality of molecular components



Charles Darwin

Biological causation:
genome-scale changes and description of 1000s of variables. Network and econometric type

Studies at higher system levels are likely to inform those at the simpler level of the cell and vice versa--Sir Paul Nurse (Nature, 2008)

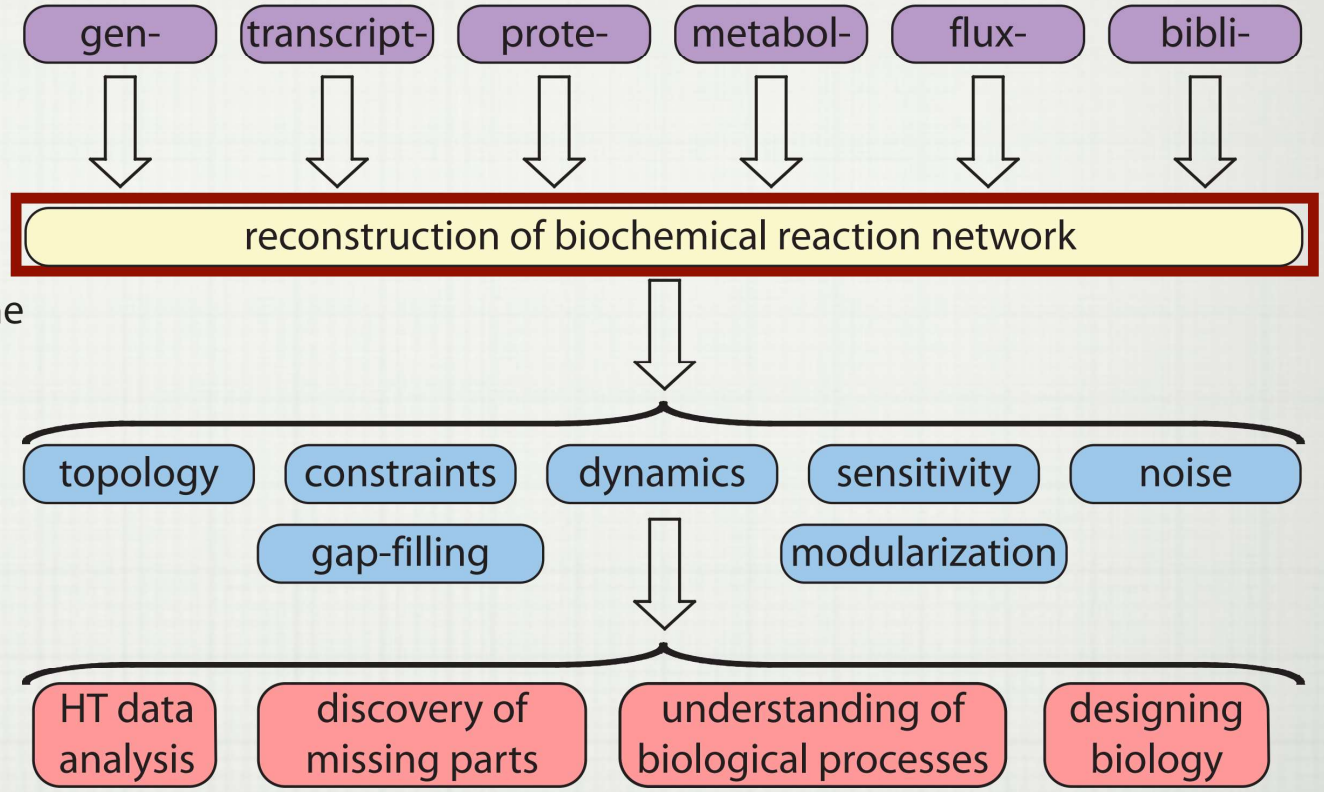
Applications of Systems Biology (Oberhardt et al, 2009)

- *Model* building and analysis
- *Contextualization* of high-throughput data
- Guidance of *metabolic engineering*
- Directing *hypothesis-driven* discovery
- Interrogation of *multi-species relationships*
- *Network* property discovery

Systems Biology Paradigm:

components -> networks -> computational models -> phenotypes

1. Database:
   - Plurality of -omics
2. Knowledge Base:
   - One set of reactions encoded by a genome
3. *In silico* modeling:
   - Query Tools
4. Validation, Discovery, and Use



Palsson,BO; Systems Biology, Cambridge University Press 2006

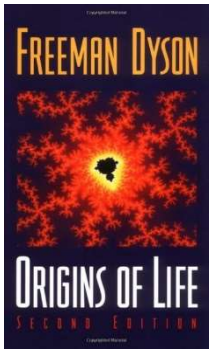
Why Networks?

- Focus on the organization of the system (rather than on its components)
- Simple representation
- Visualization of complex systems
- Networks as tools
- Underlying diffusion model (e.g. evolution on networks)
- **The structure and topology of the system affect (determine) its function**

Networks in Biology

- Molecular networks:
  - Protein-Protein Interaction (PPI) networks
  - Metabolic Networks
  - Regulatory Network
  - Synthetic lethality Network
  - Gene Interaction Network
  - More …

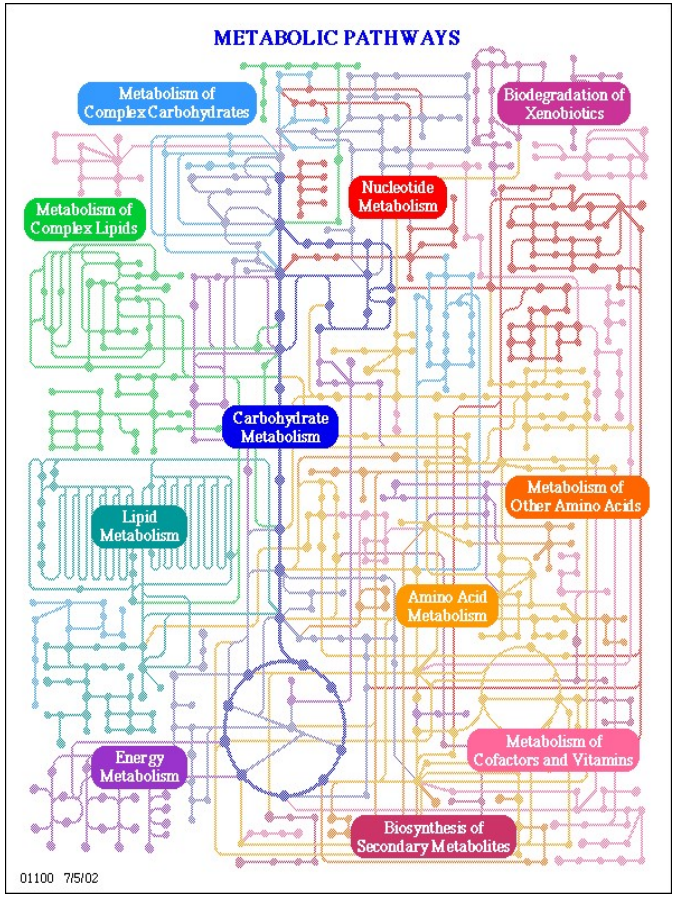
Why study metabolism?



- **It's the essence of life**
  (and maybe its origins)
- Tremendous importance in Medicine
  - Inborn errors of metabolism cause acute symptoms
  - Metabolic diseases (obesity, diabetes) are on the rise (and are major sources of morbidity and mortality)
  - Metabolic enzymes becoming viable drug targets
- Bioengineering applications
  - Design strains for production of biological products
  - Generation of bio-fuels
- **The best understood of all cellular networks**

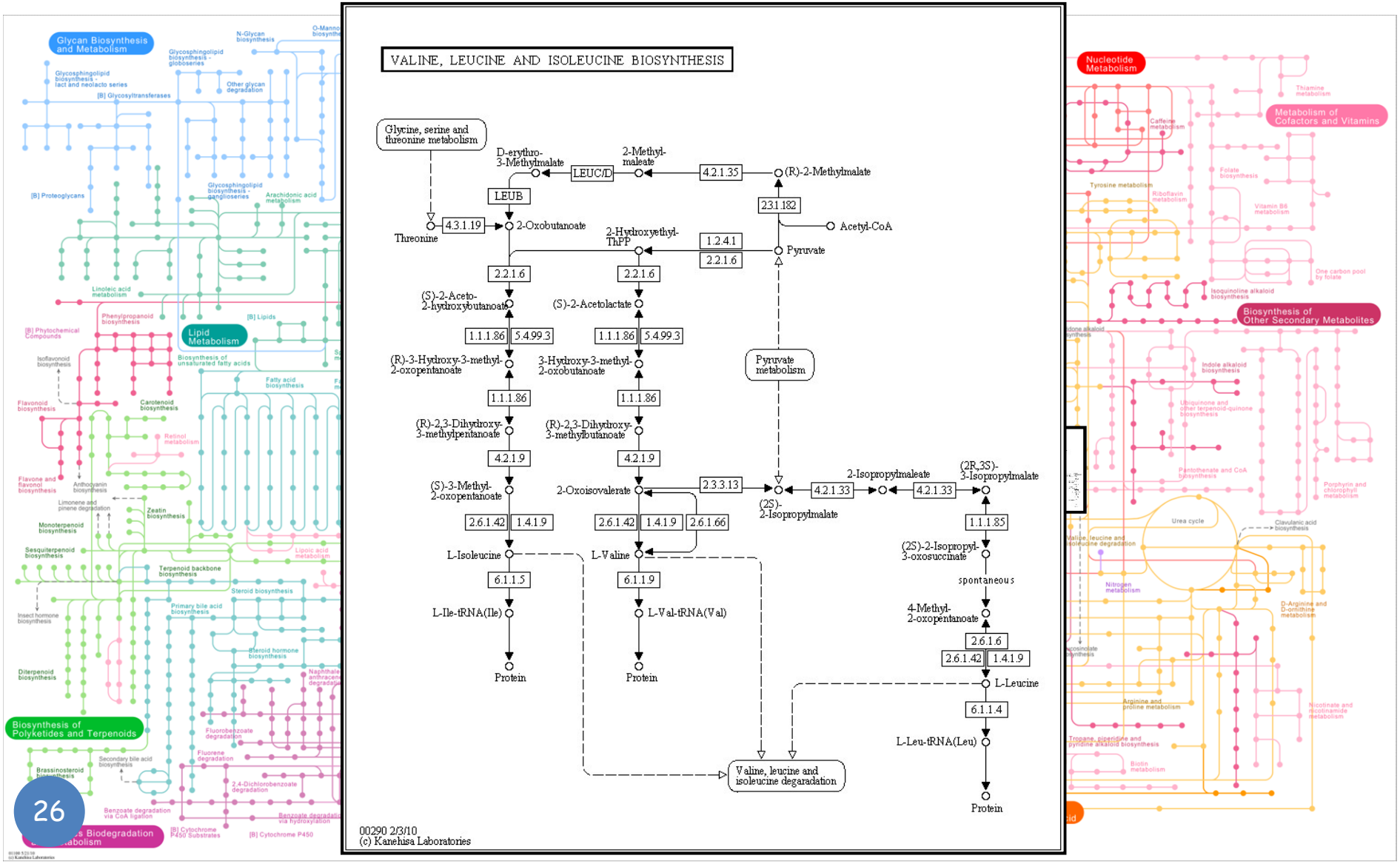
Metabolic Networks

- Reflect the set of biochemical reactions in a cell
  - Nodes: metabolites
  - Edges: biochemical reactions
  - **Additional representations!**
- Derived through:
  - Knowledge of biochemistry
  - Metabolic flux measurements

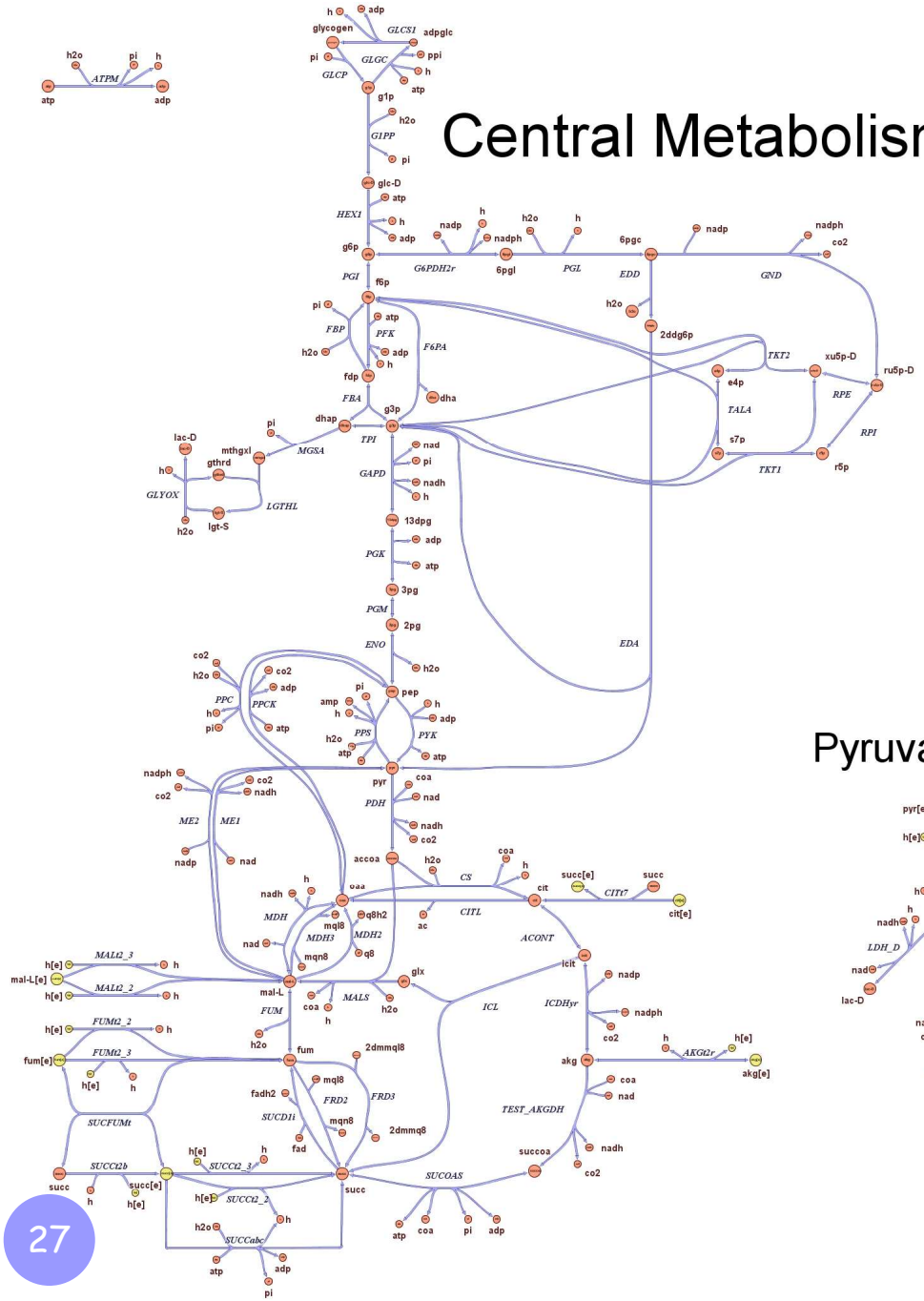


http://www.genome.jp/kegg/pathway/map/map01100.html

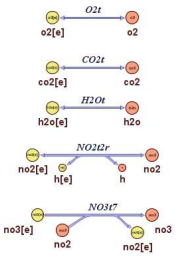
From Pathways to a Network



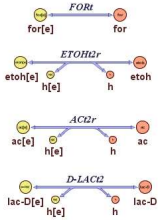
Central Metabolism



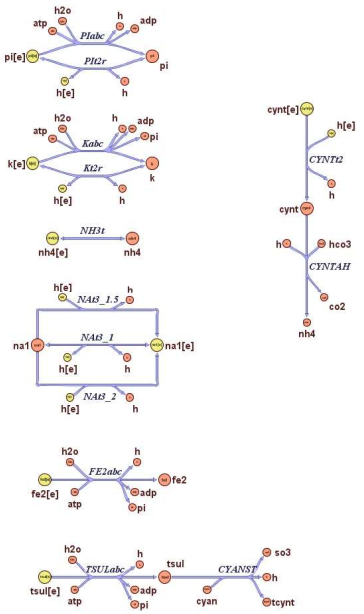
Misc Transporters



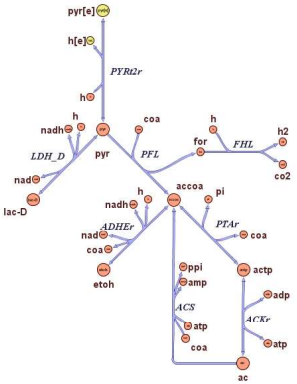
Common Byproducts



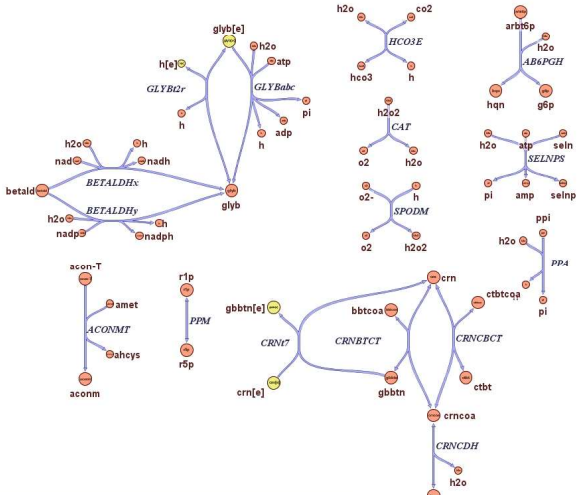
Small Compounds/Elements



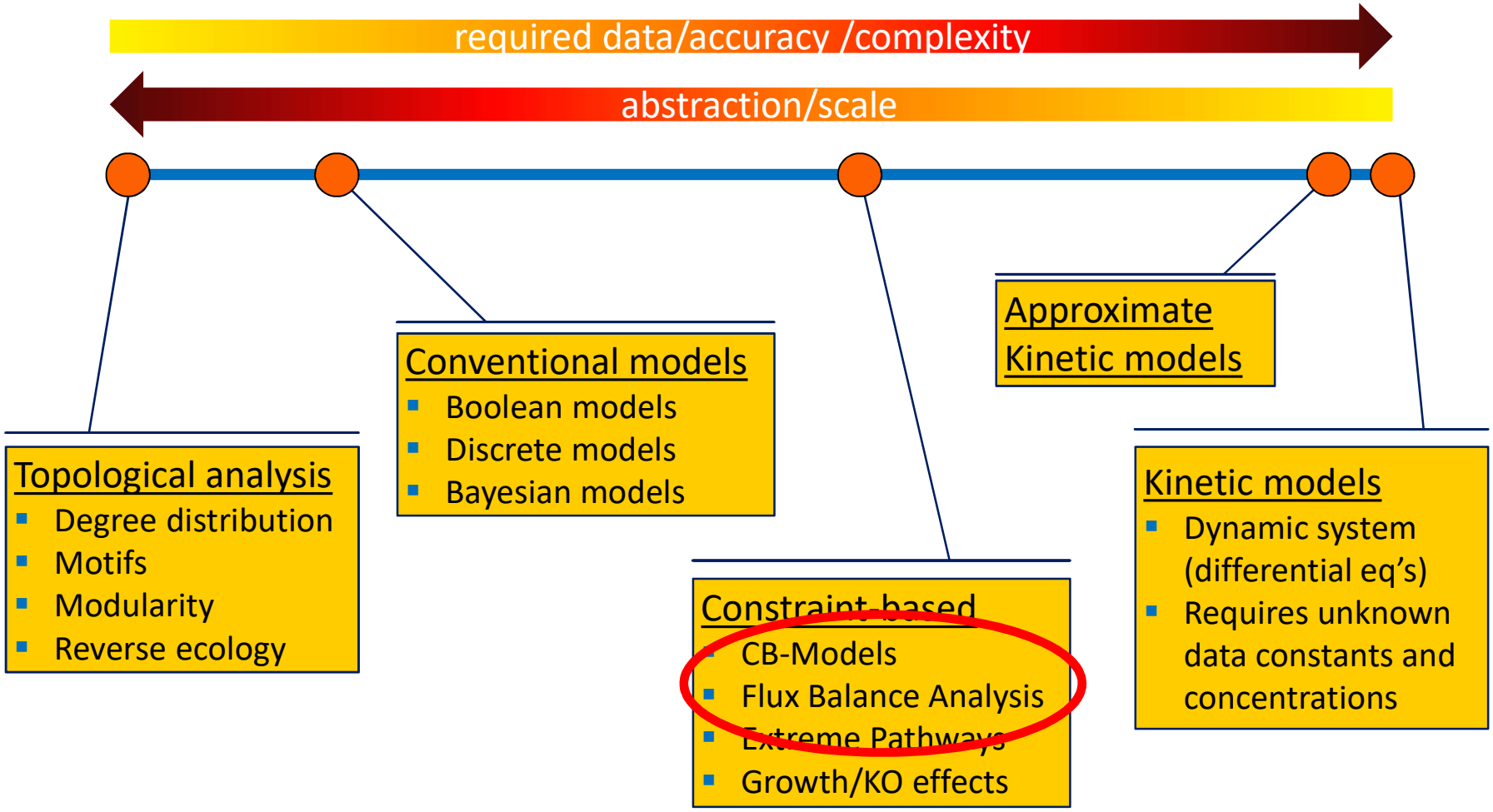
Pyruvate Metabolism



Orphan Reactions

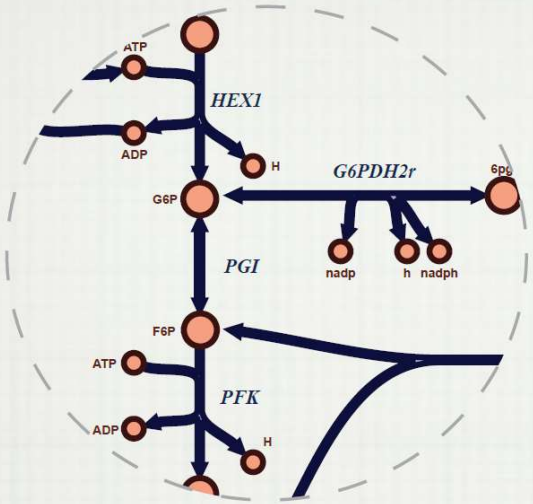


Metabolic Network Models

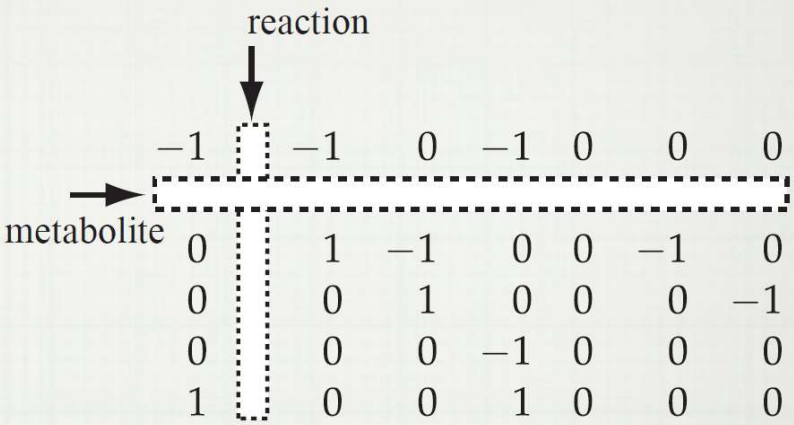


COnstraint-Based Reconstruction and Analysis (COBRA)

Network map

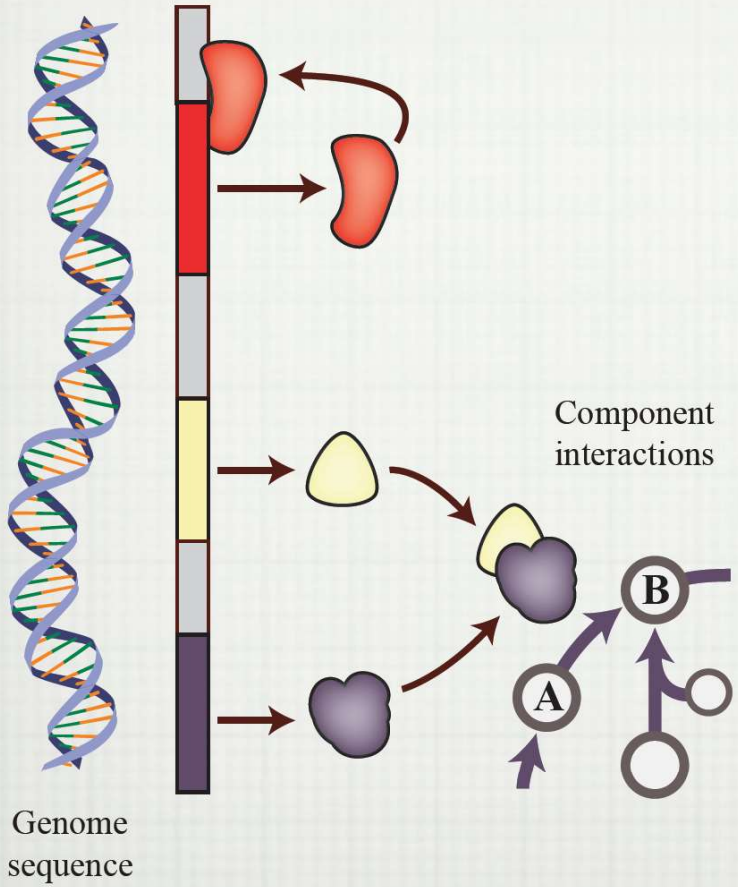


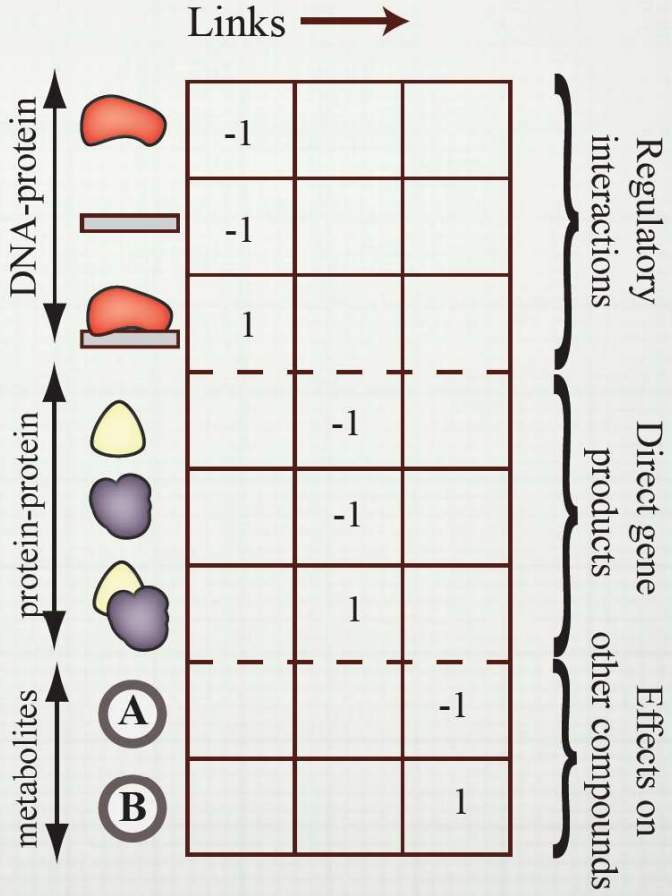
Mathematical representation



- Network reconstruction is a BiGG knowledge base
- Can convert structured knowledge into mathematical format
- Birth of genome-scale (metabolic) systems biology
- Provides a mechanistic basis for the genotype-phenotype relationship
- Dual causality needs to be accounted for
  - different treatment than physics a 100 years ago

Network Reconstruction:
Effectively a 2D Genome Annotation





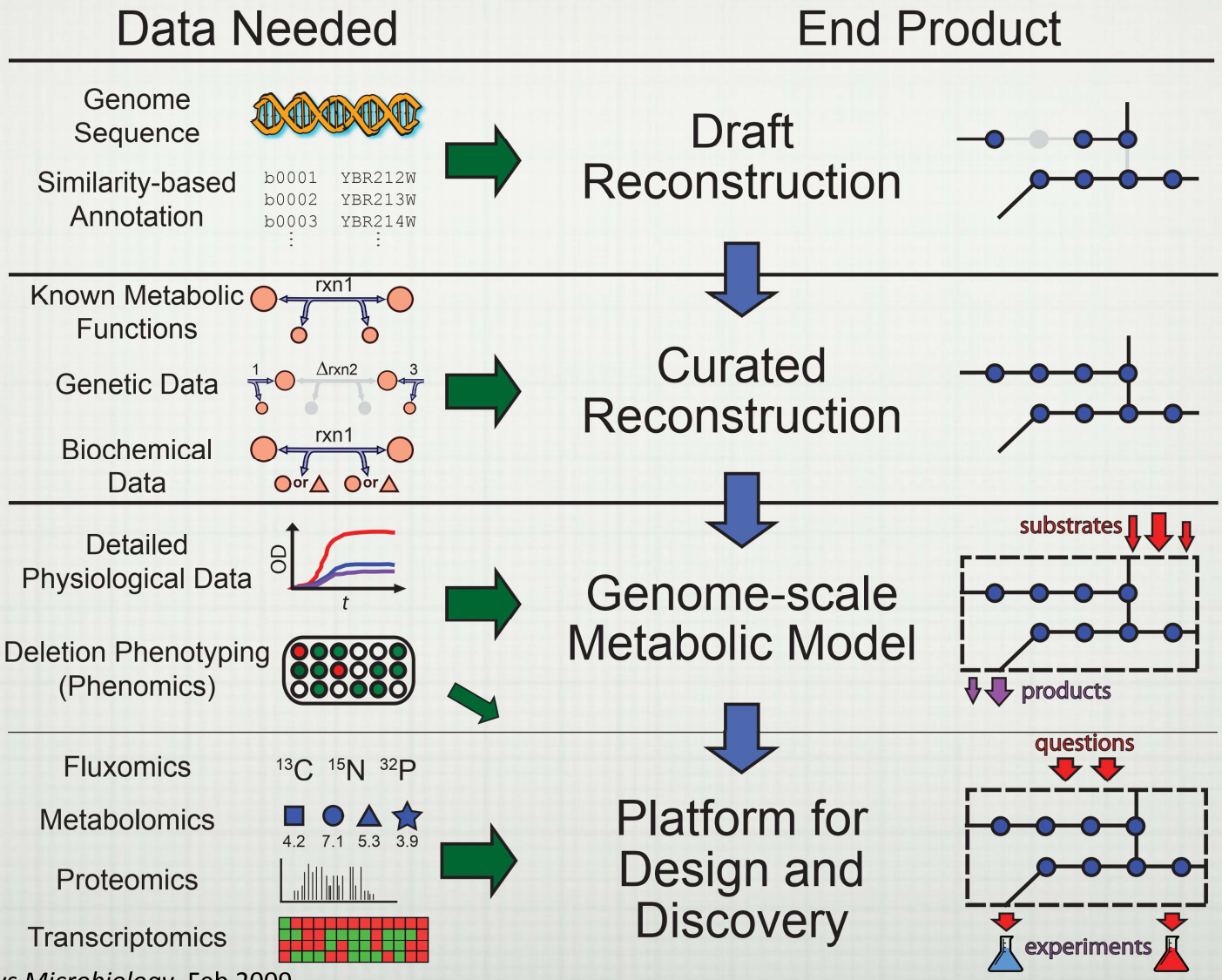
Component (1D) Annotation

Systemic (2D) Annotation

Nature Biotechnology, 22: 1218, 2004

Bottom-up Network Reconstruction:

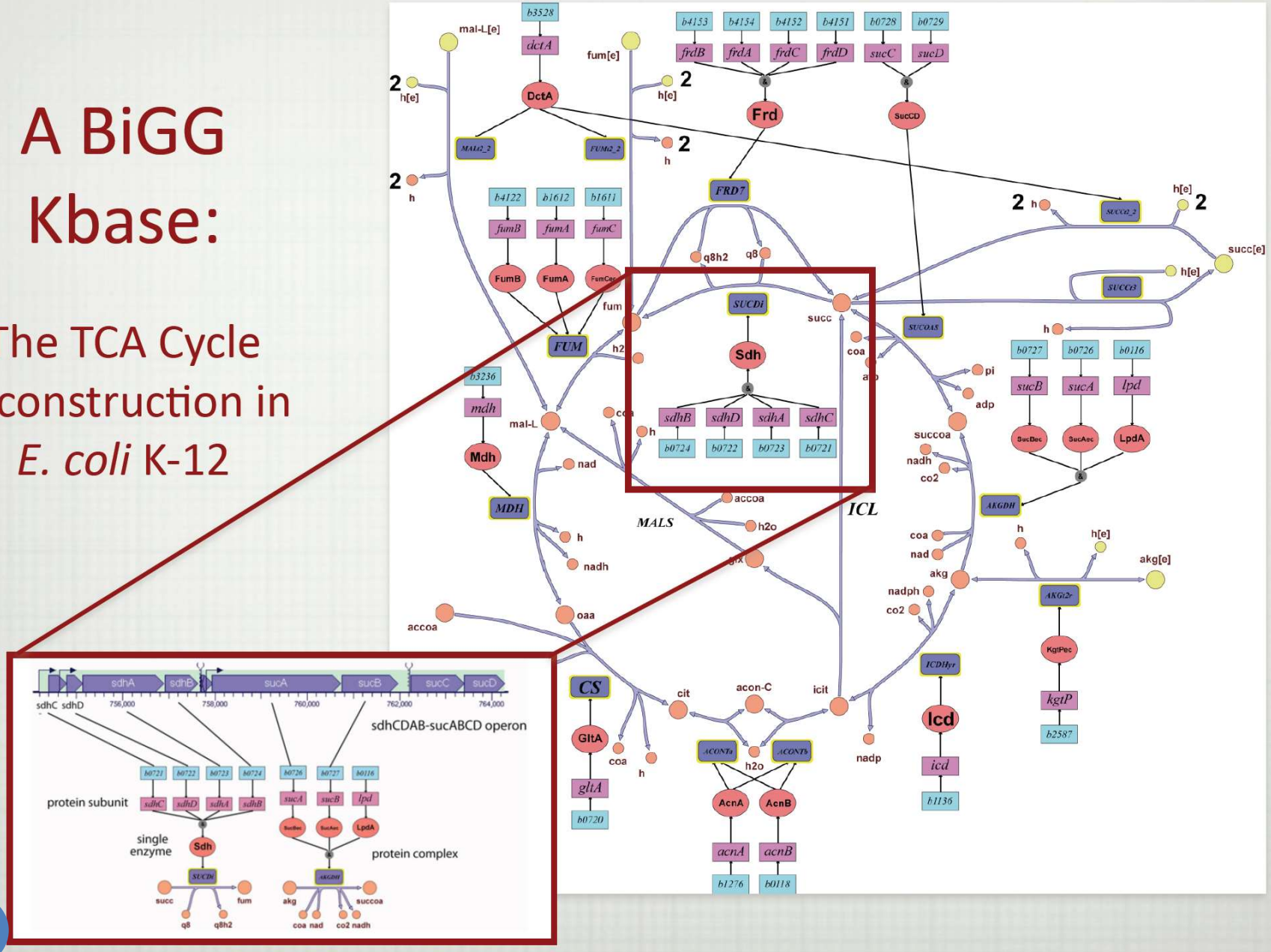
A four step process



ure Reviews Microbiology, Feb 2009

A BiGG Kbase:

The TCA Cycle reconstruction in *E. coli* K-12



Nat Protoc. Author manuscript; available in PMC 2011 Jun 28.
Published in final edited form as:
Nat Protoc. 2010; 5(1): 93–121.
Published online 2010 Jan 7. doi: [10.1038/nprot.2009.203]

PMCID: PMC3125167
NIHMSID: NIHMS251754
PMID: 20057383

A protocol for generating a high-quality genome-scale metabolic reconstruction

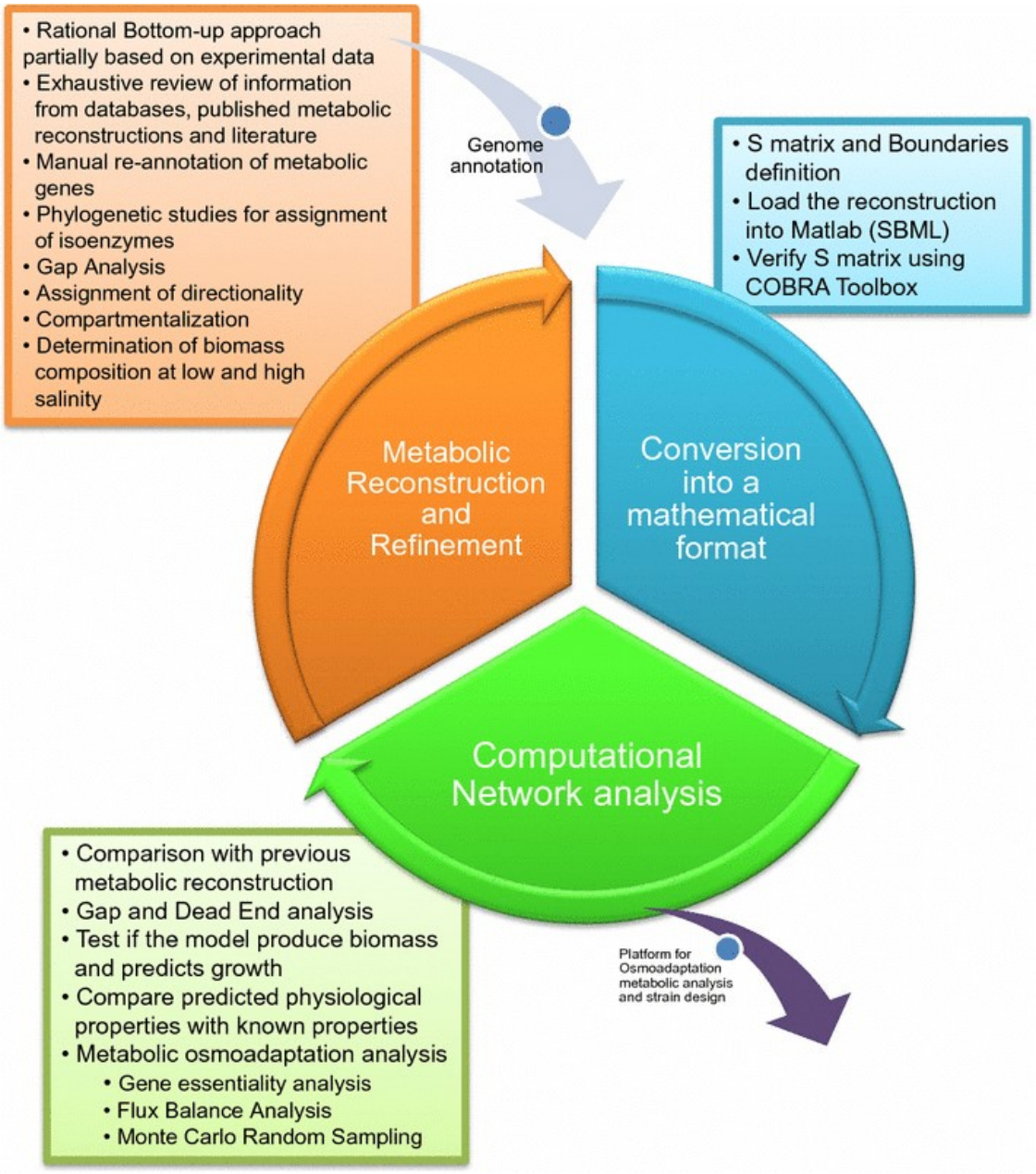
Ines Thiele[1,2] and Bernhard Ø. Palsson[1,*]

Author information ► Copyright and License information ► Disclaimer

Abstract

Go to:

Network reconstructions are a common denominator in systems biology. Bottom-up metabolic network reconstructions have developed over the past 10 years. These reconstructions represent structured knowledge-bases that abstract pertinent information on the biochemical transformations taking place within specific target organisms. The conversion of a reconstruction into a mathematical format facilitates myriad computational biological studies including evaluation of network content, hypothesis testing and generation, analysis of phenotypic characteristics, and metabolic engineering. To date, genome-scale

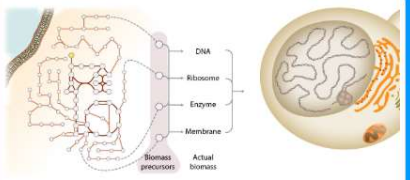

• Rational Bottom-up approach partially based on experimental data
• Exhaustive review of information from databases, published metabolic reconstructions and literature
• Manual re-annotation of metabolic genes
• Phylogenetic studies for assignment of isoenzymes
• Gap Analysis
• Assignment of directionality
• Compartmentalization
• Determination of biomass composition at low and high salinity
Genome annotation
• S matrix and Boundaries definition
• Load the reconstruction into Matlab (SBML)
• Verify S matrix using COBRA Toolbox
Metabolic Reconstruction and Refinement
Conversion into a mathematical format
Computational Network analysis
• Comparison with previous metabolic reconstruction
• Gap and Dead End analysis
• Test if the model produce biomass and predicts growth
• Compare predicted physiological properties with known properties
• Metabolic osmoadaptation analysis
• Gene essentiality analysis
• Flux Balance Analysis
• Monte Carlo Random Sampling
Platform for Osmoadaptation metabolic analysis and strain design

BiGG Models

Search the database by model, reaction, metabolite, or gene ?

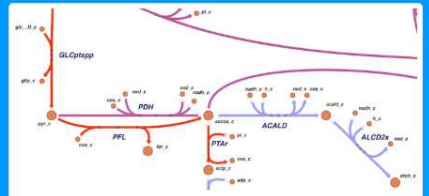
Search

Latest update Version 1.5: Introducing Recon3D



View Models

View Metabolites

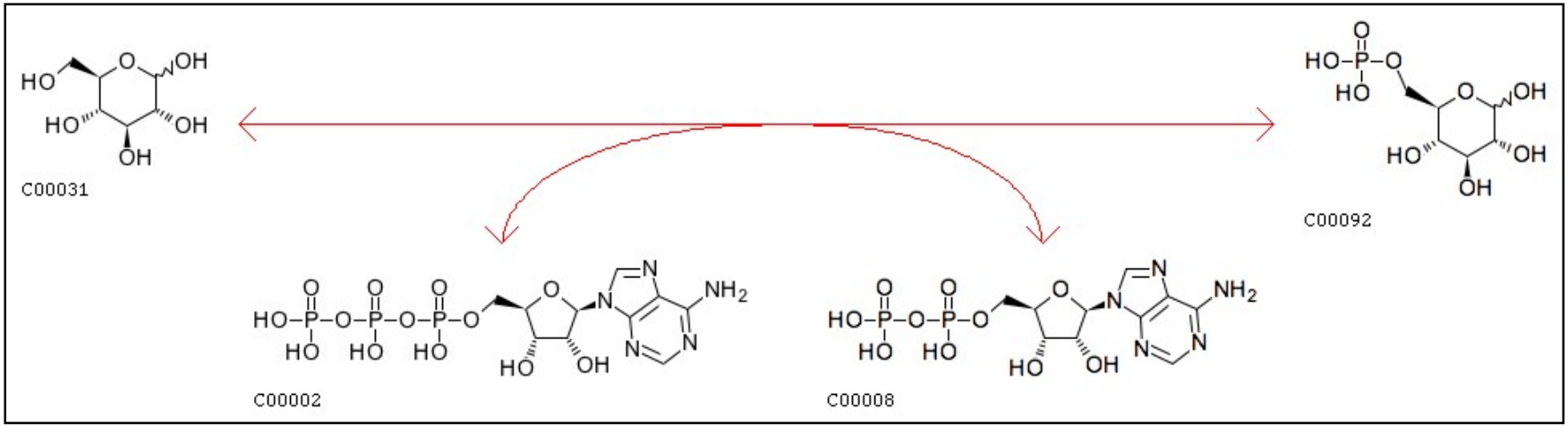


View Reactions

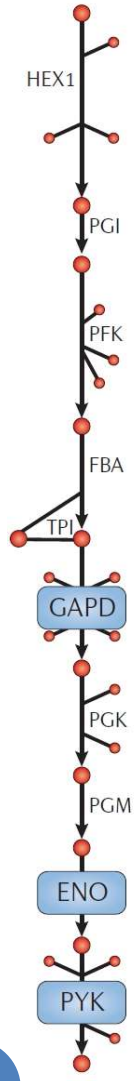
Reaction Stoichiometry

- ***Stoichiometry*** **- the quantitative relationships of the reactants and products in reactions**

1 Glucose + 1 ATP ⟷ 1 Glucose-6-Phosphate + 1 ADP



Stoichiometric Matrix

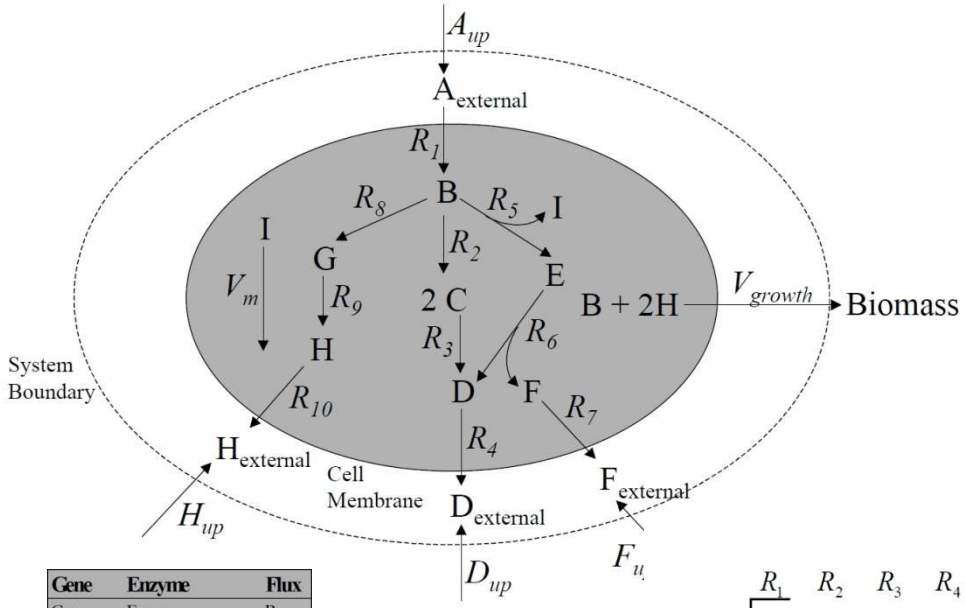


Abbreviation	Glycolytic reactions	Genes
HEX1	[c]GLC + ATP → G6P + ADP + H	*glk*
PGI	[c]G6P ↔ F6P	*pgi*
PFK	[c]ATP + F6P → ADP + FDP + H	*pfkA*, *pfkB*
FBA	[c]FDP ↔ DHAP + G3P	*fbaA*, *fbaB*
TPI	[c]DHAP ↔ G3P	*tpiA*
GAPD	[c]G3P + NAD + PI ↔ 13DPG + H + NADH	*gapA*, *gapC1*, *gapC2*
PGK	[c]13DPG + ADP ↔ 3PG + ATP	*pgk*
PGM	[c]3PG ↔ 2PG	*gpmA*, *gpmB*
ENO	[c]2PG ↔ H_2O + PEP	*eno*
PYK	[c]ADP + H + PEP → ATP + PYR	*pykA*, *pykF*

	HEX1	PGI	PFK	FBA	TPI	GAPD	PGK	PGM	ENO	PYK
ATP	−1	0	−1	0	0	0	1	0	0	1
GLC	−1	0	0	0	0	0	0	0	0	0
ADP	1	0	1	0	0	0	−1	0	0	−1
G6P	1	−1	0	0	0	0	0	0	0	0
H	1	0	1	0	0	1	0	0	0	−1
F6P	0	1	−1	0	0	0	0	0	0	0
FDP	0	0	1	−1	0	0	0	0	0	0
DHAP	0	0	0	1	−1	0	0	0	0	0
G3P	0	0	0	1	1	−1	0	0	0	0
NAD	0	0	0	0	0	−1	0	0	0	0
PI	0	0	0	0	0	−1	0	0	0	0
13DPG	0	0	0	0	0	1	−1	0	0	0
NADH	0	0	0	0	0	1	0	0	0	0
3PG	0	0	0	0	0	0	1	−1	0	0
2PG	0	0	0	0	0	0	0	1	−1	0
PEP	0	0	0	0	0	0	0	0	1	−1
H_2O	0	0	0	0	0	0	0	0	1	0
PYR	0	0	0	0	0	0	0	0	0	1

S

Stoichiometric Matrix and Fluxes



$$\frac{d\bar{m}}{dt} = S \cdot \bar{v}$$

- **m**: *metabolite concentrations vector (mol/mg)*
- **S**: *stoichiometric matrix*
- **v**: *reaction rates vector*

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}

	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
B	1	−1	0	0	−1	0	0	−1	0	0	0	−1	0	0	0	0
C	0	2	−1	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	1	−1	0	1	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	1	−1	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	1	−1	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	1	−1	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	1	−1	0	−2	0	0	0	0
I	0	0	0	0	1	0	0	0	0	0	−1	0	0	0	0	0
$\mathbf{A}_{external}$	−1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
$\mathbf{D}_{external}$	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
$\mathbf{F}_{external}$	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
$\mathbf{H}_{external}$	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \\ R_6 \\ R_7 \\ R_8 \\ R_9 \\ R_{10} \\ V_m \\ V_{growth} \\ A_{up} \\ D_{up} \\ F_{up} \\ H_{up} \end{bmatrix}$$

A Full Model? Not Really

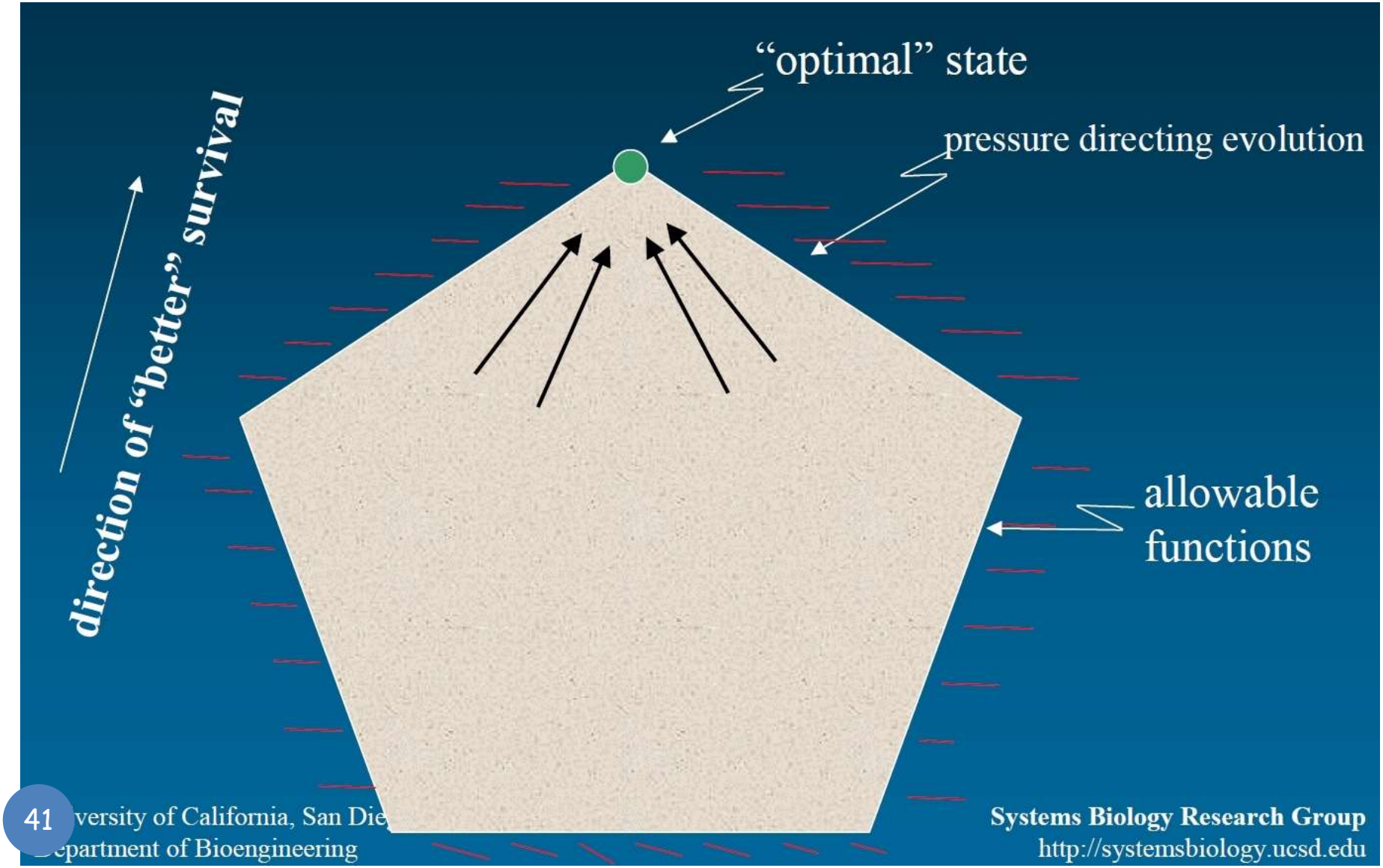
A set of Ordinary Differential Equations (ODE)

$$\frac{d\overline{m}}{dt} = S \cdot \overline{v} = S \cdot f(\overline{m}, k)$$

Reaction rate equation Kinetic parameters

Requires knowledge of m, f and k!

Evolution Under Constraints



Constraint-Based Modeling

- Living systems obey physical and chemical laws
- These can be used to constrain the space of possible behaviors of the network

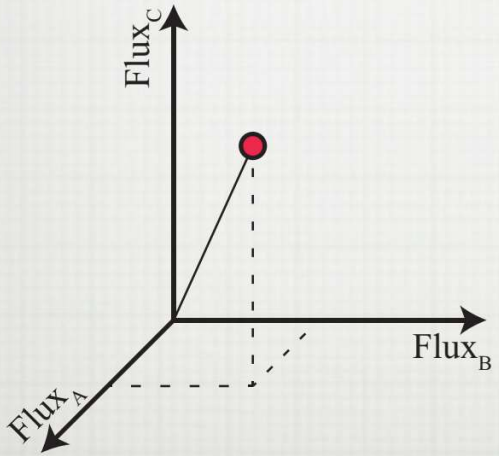
How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?

– Sherlock Holmes (A Study of Scarlet)

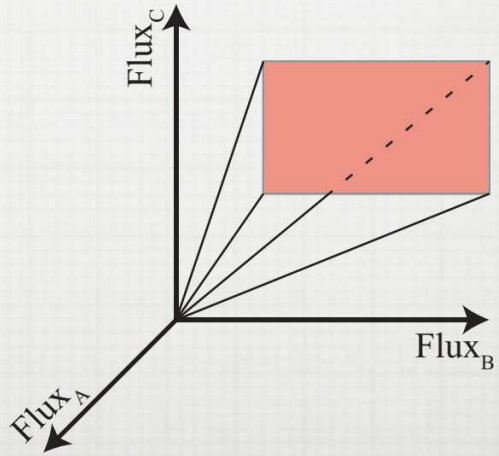
Constraint-Based Modeling

Contrasting Thinking

Theory-based
- Complete knowledge
- Solution is a single point



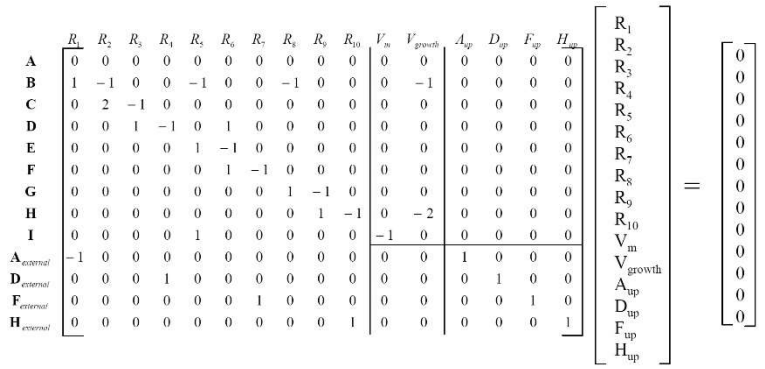
Constraint-based
- Incomplete knowledge
- Solution confined to a space



Constraint-Based Modeling

- Assumes a quasi steady-state!
  - No changes in metabolite concentrations
  - Metabolite production and consumption rates are equal

$$\frac{d\bar{m}}{dt} = S \cdot \bar{v} = 0$$



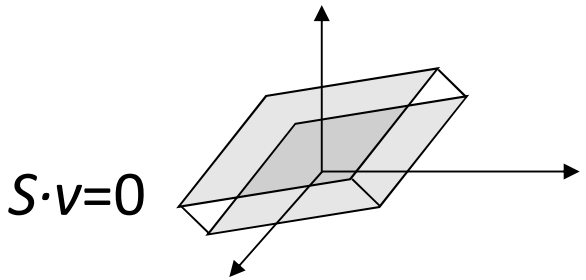
- No need for info on metabolite concentrations, reaction rate functions, or kinetic parameters

Constraint-Based Modeling

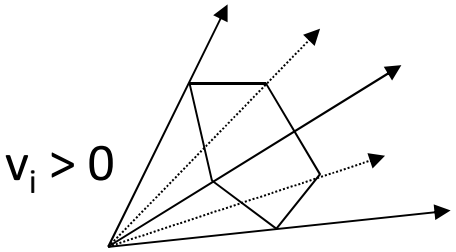
- In most cases, *S* is underdetermined:
  → a subspace of R^n (possible flux distributions)

$$\begin{pmatrix}0\\0\\0\end{pmatrix} = \begin{bmatrix}-1 & -1 & 1 & 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 1 & 0 & -1 & 0\\ 0 & 1 & -1 & -1 & 0 & 0 & -1\end{bmatrix} \begin{bmatrix}v_1\\v_2\\v_3\\v_4\\b_1\\b_2\\b_3\end{bmatrix}$$

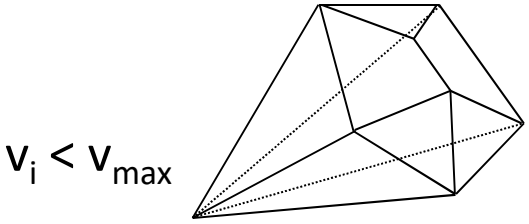
S (columns) · V (rows)



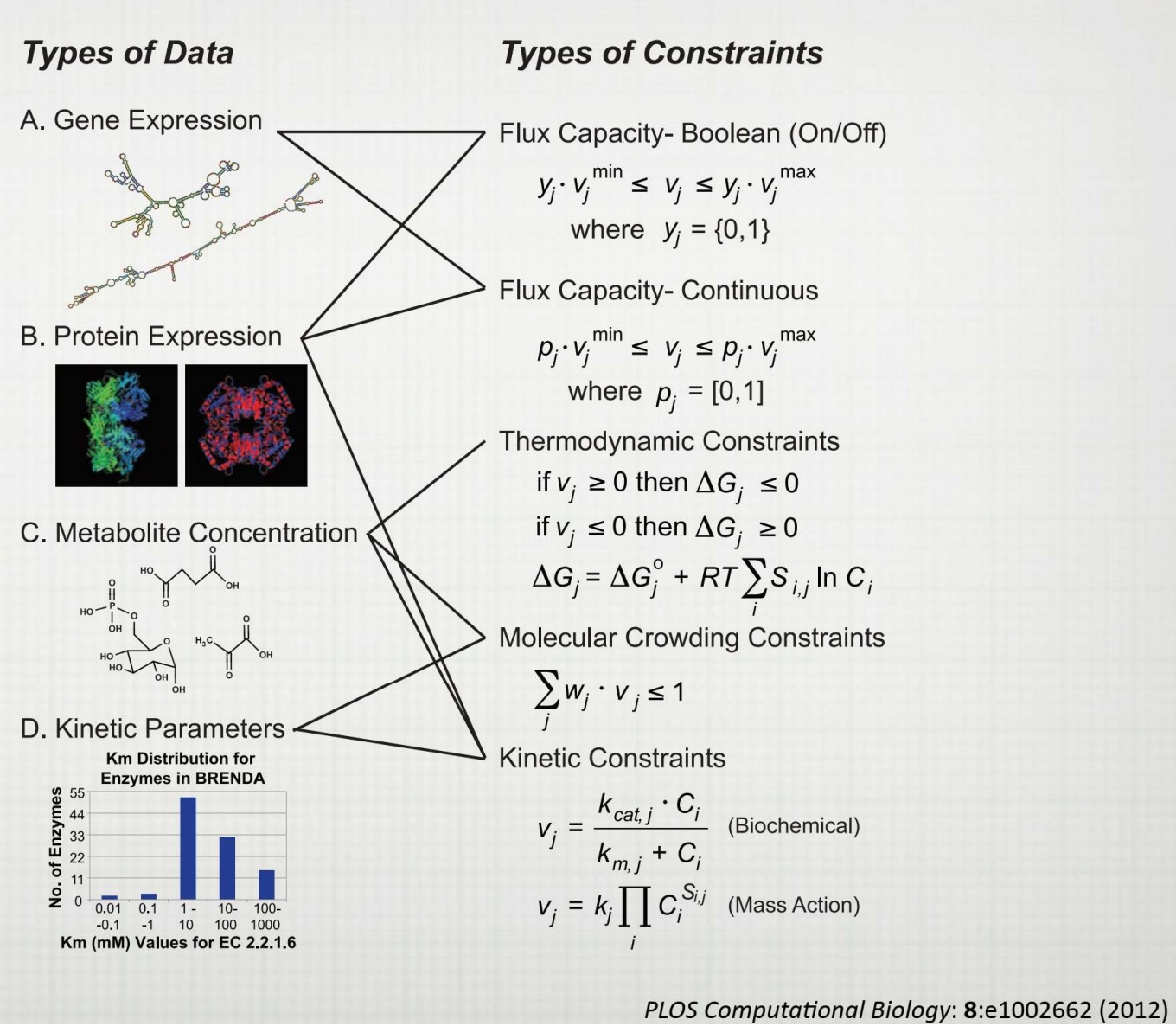
- *Thermodynamic constraints*:
  → a convex cone



- *Capacity constraints*:
  → a bounded convex cone



Constraint-Based Modeling



PLOS Computational Biology: **8**:e1002662 (2012)

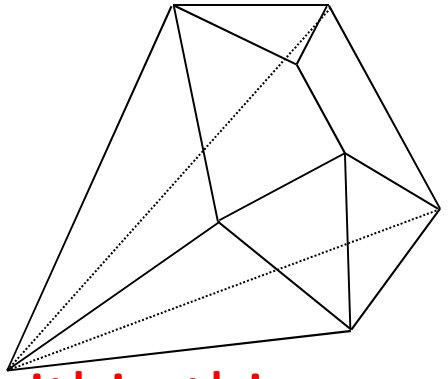
Constraint-Based Modeling

Comparing the Properties of Linear and Convex Bases

	Linear Spaces	Convex Spaces
	Described by linear equations	Described by linear equations and inequalities
1.	$\mathbf{S}\mathbf{v}_{ss} = 0$	$\mathbf{S}\mathbf{v}_{ss} = 0$ and $0 \leq v_i \leq v_{i,max}$
	Vector spaces defined by a set of linearly independent basis vectors ($\mathbf{b}_i$)	Convex polyhedral cone defined by a set of conically independent vectors ($\mathbf{p}_i$)
2.	$\mathbf{v} = \sum w_i \mathbf{b}_i \quad -\infty \leq w_i \leq +\infty$	$\mathbf{v} = \sum \alpha_i \mathbf{p}_i \quad 0 \leq \alpha_i \leq +\infty$ ← α_{max}
	Every point in the vector space is uniquely described by a linear combination of basis vectors (unique representation for a given basis)	Every point in the vector space is described as a non-negative linear combination of the convex basis vectors (non-unique representation)
3.	Number of basis vectors equals dimension of the null space	Number of convex basis vectors may exceed dimension of the null space
4.	Infinite number of bases that can be used to span the space	Unique set of basis convex vectors

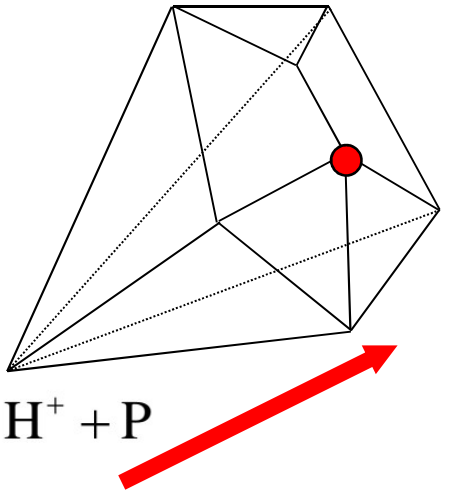
Flux Balance Analysis (FBA)

- But this still leaves a space of solutions



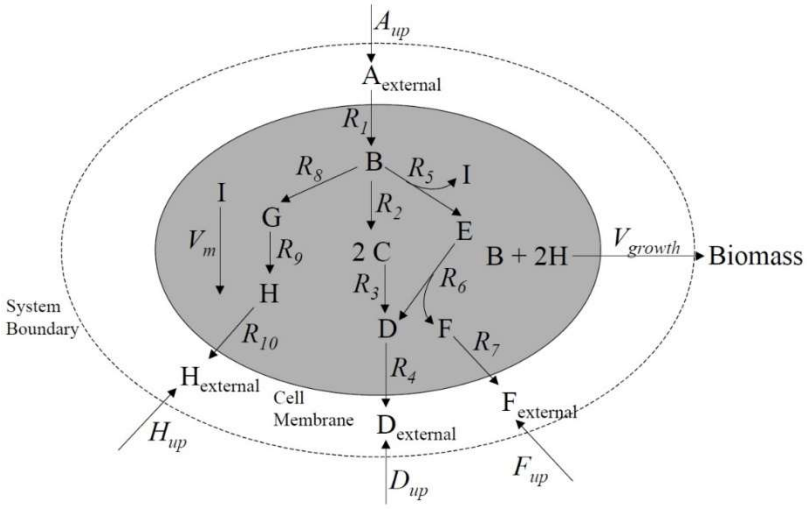
- How can we identify plausible solutions within this space?

- Optimize for maximum growth rate !!



$$\sum_{i \in I^{\text{biomass}}} S_{i,\text{biomass}} X_i + S_{\text{ATP,biomass}} \text{ATP} \xrightarrow{v_{\text{biomass}}} \text{biomass} + \text{ADP} + \text{H}^+ + \text{P}$$

Flux Balance Analysis (FBA)



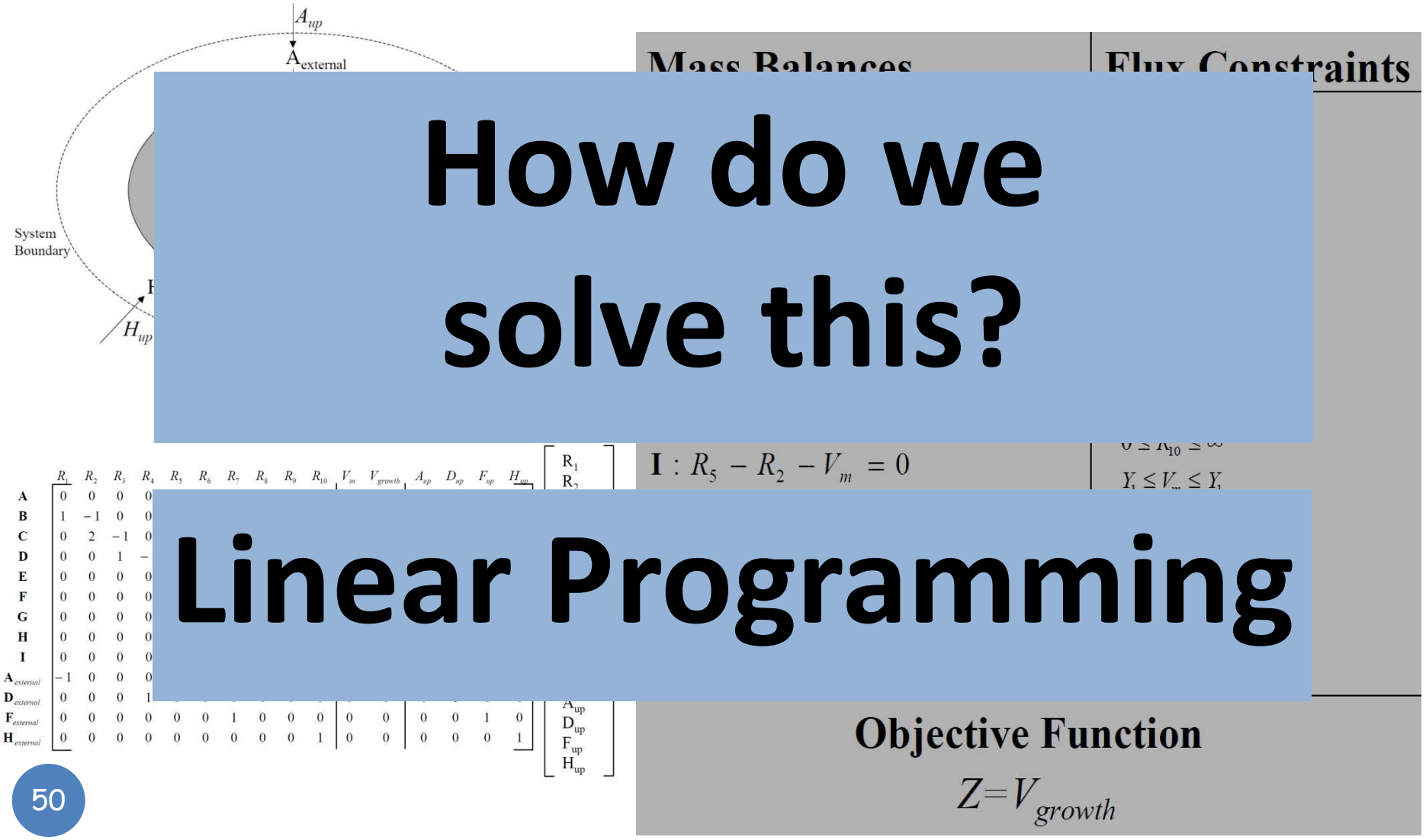
	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	V_m	V_{growth}	A_{up}	D_{up}	F_{up}	H_{up}	
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	R_1
B	1	−1	0	0	−1	0	0	−1	0	0	0	−1	0	0	0	0	R_2
C	0	2	−1	0	0	0	0	0	0	0	0	0	0	0	0	0	R_3
D	0	0	1	−1	0	1	0	0	0	0	0	0	0	0	0	0	R_4
E	0	0	0	0	1	−1	0	0	0	0	0	0	0	0	0	0	R_5
F	0	0	0	0	0	1	−1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

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

Objective Function

$$Z = V_{growth}$$

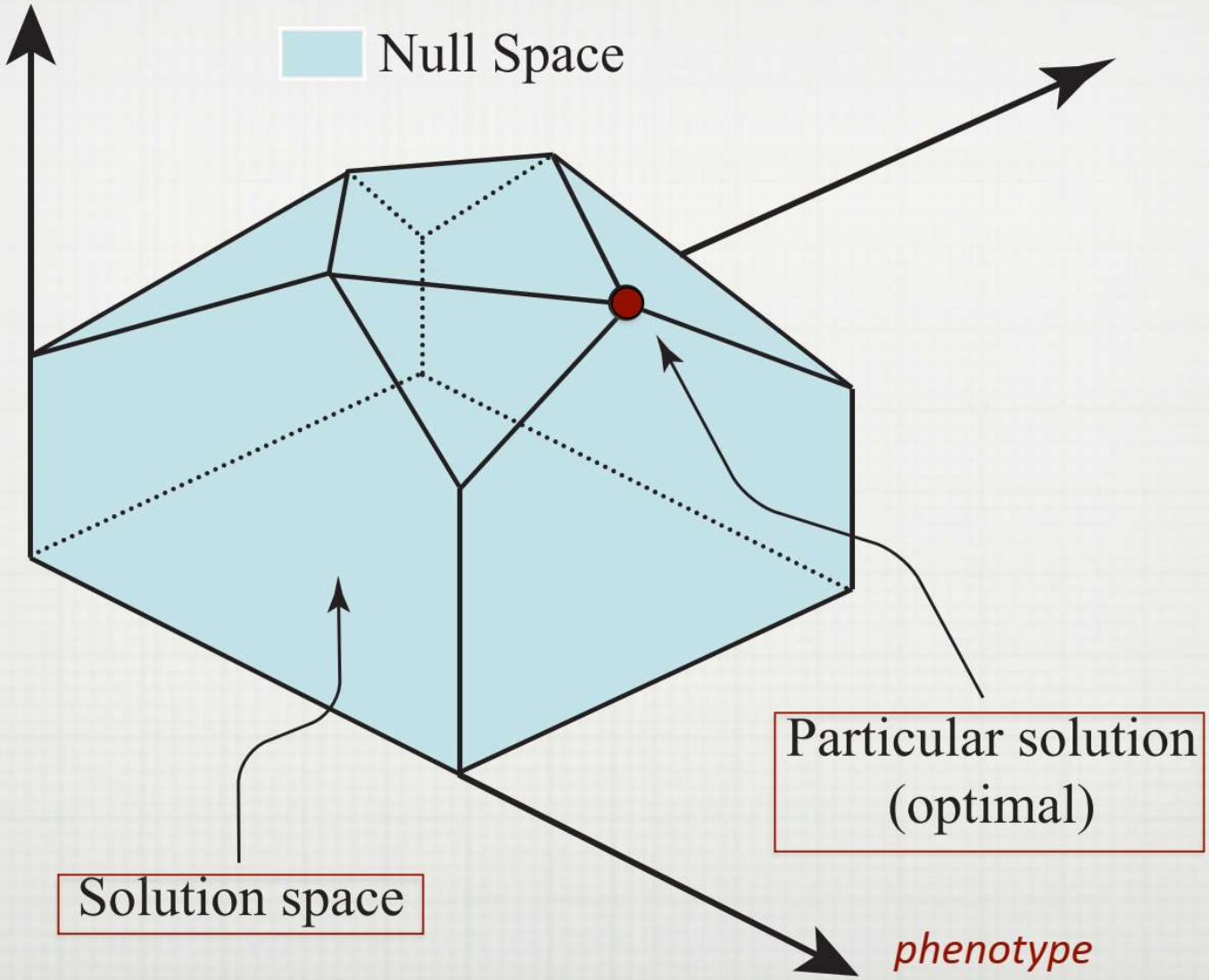
Flux Balance Analysis (FBA)



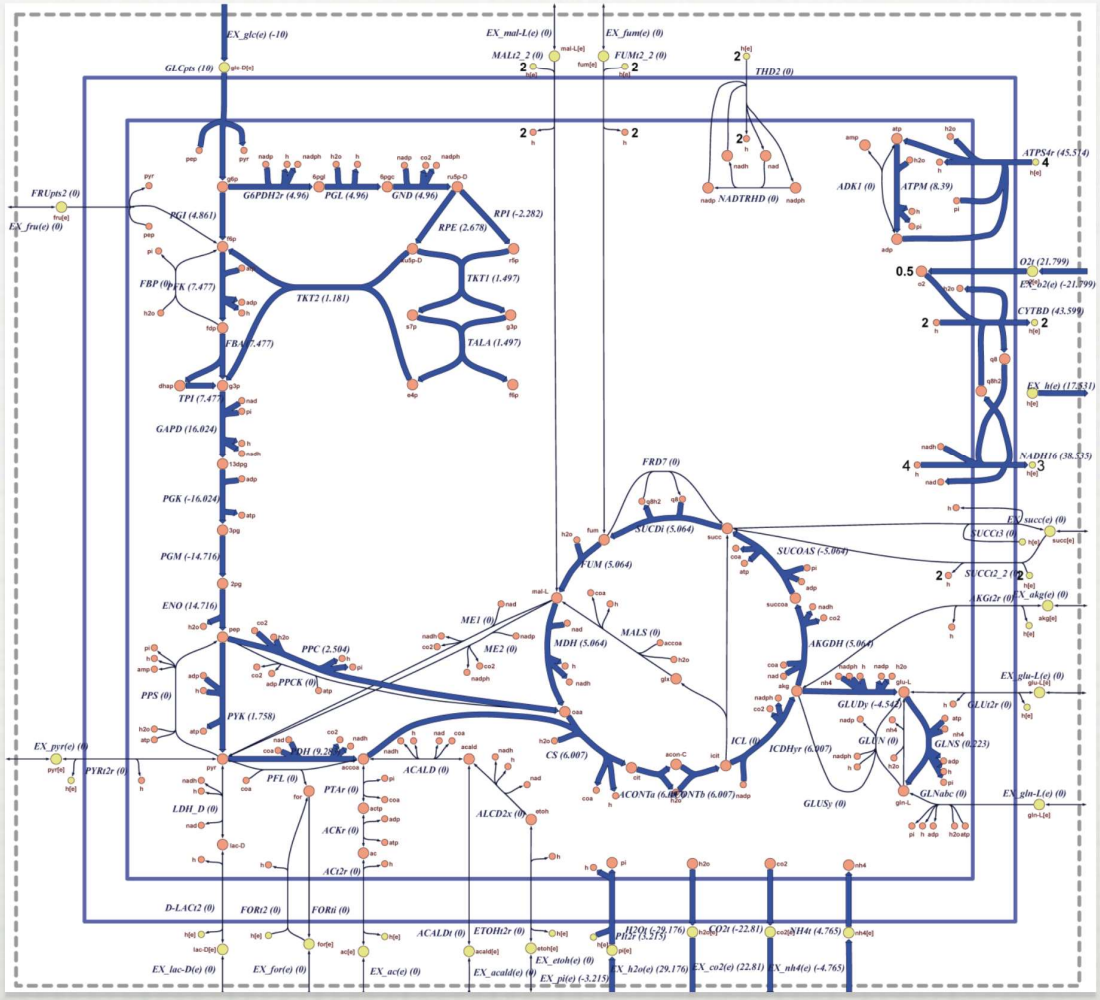
How do we solve this?

Linear Programming

Solution Spaces and Phenotypic States



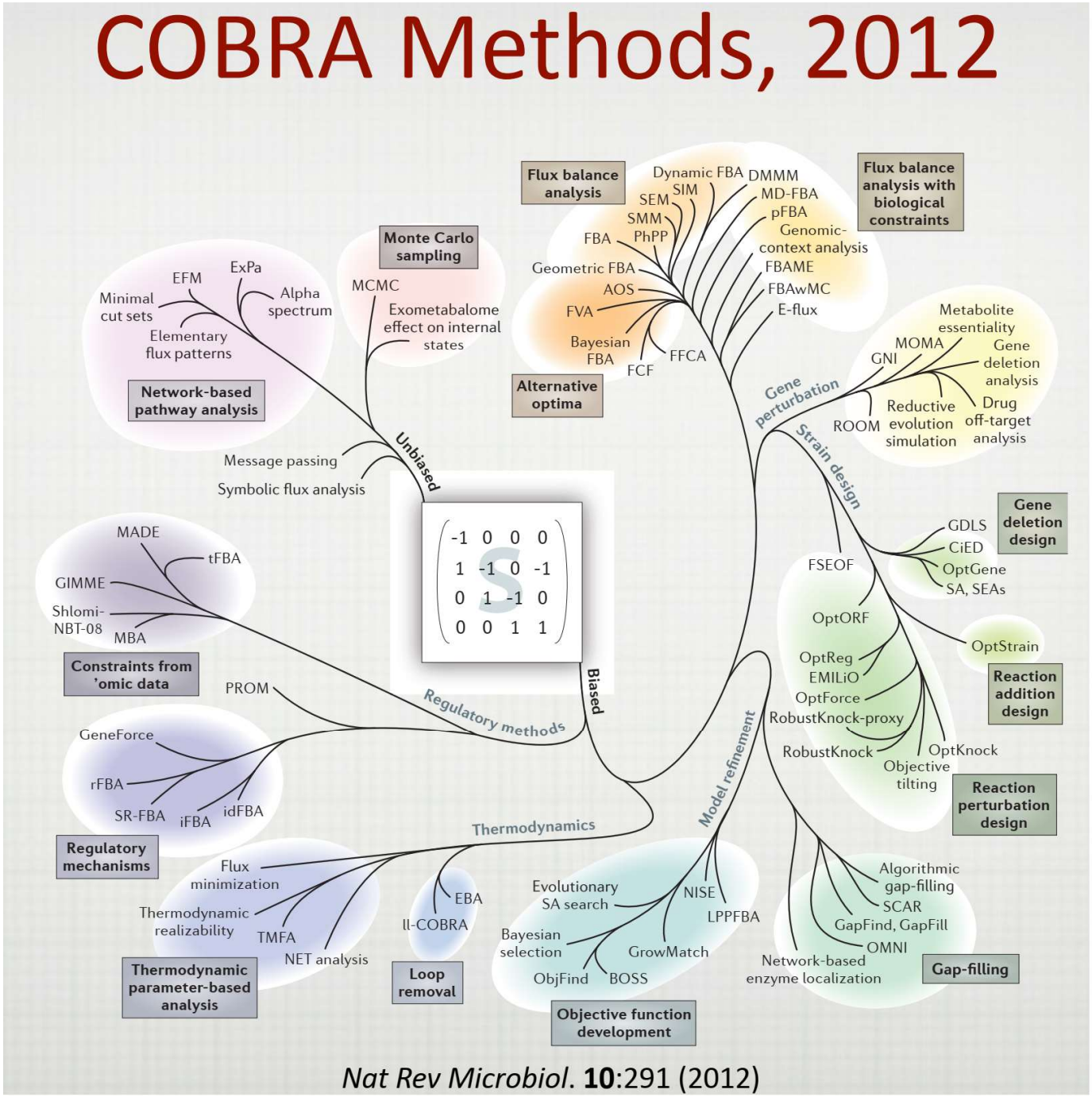
Max Growth Rate of *E. coli* on Glucose



Applications of CBM & FBA

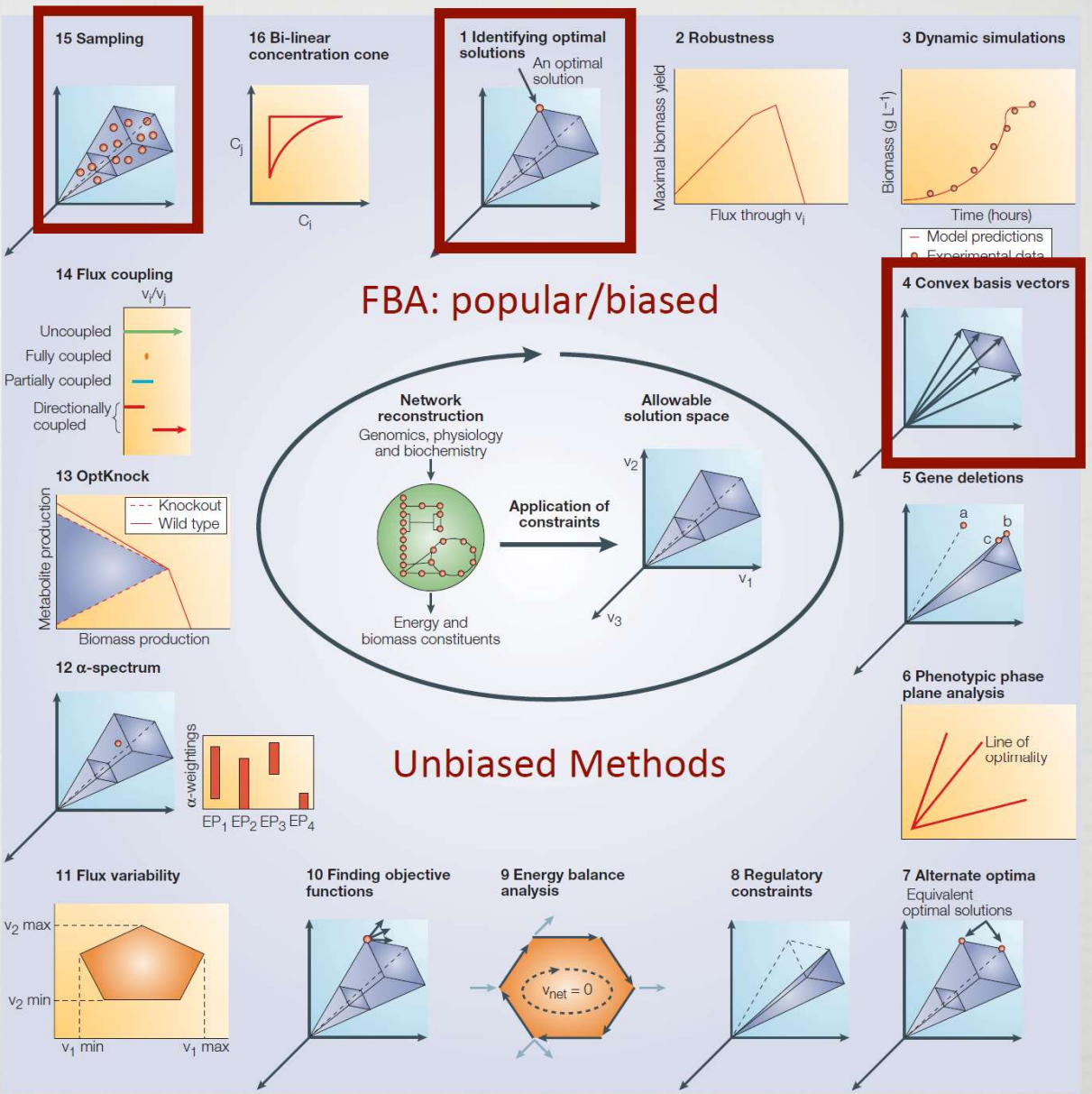
- Predict metabolic fluxes on various media
- Predict growth rate
- Predict gene knockout lethality
- Characterize solution space
- Many more …

COBRA Methods, 2012



Nat Rev Microbiol. **10**:291 (2012)

The Toolbox of Constraint-based Methods for Computational Modeling



Nat Rev Microbiol. **2**:886 (2004)

55

MENU

nature REVIEWS MICROBIOLOGY

Review Article | Published: 27 February 2012

Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods

Nathan E. Lewis, Harish Nagarajan & Bernhard O. Palsson

Nature Reviews Microbiology **10**, 291–305 (2012) | Download Citation

Abstract

Reconstructed microbial metabolic networks facilitate a mechanistic description of the genotype–phenotype relationship through the deployment of constraint-based reconstruction and analysis (COBRA)

Topological Methods

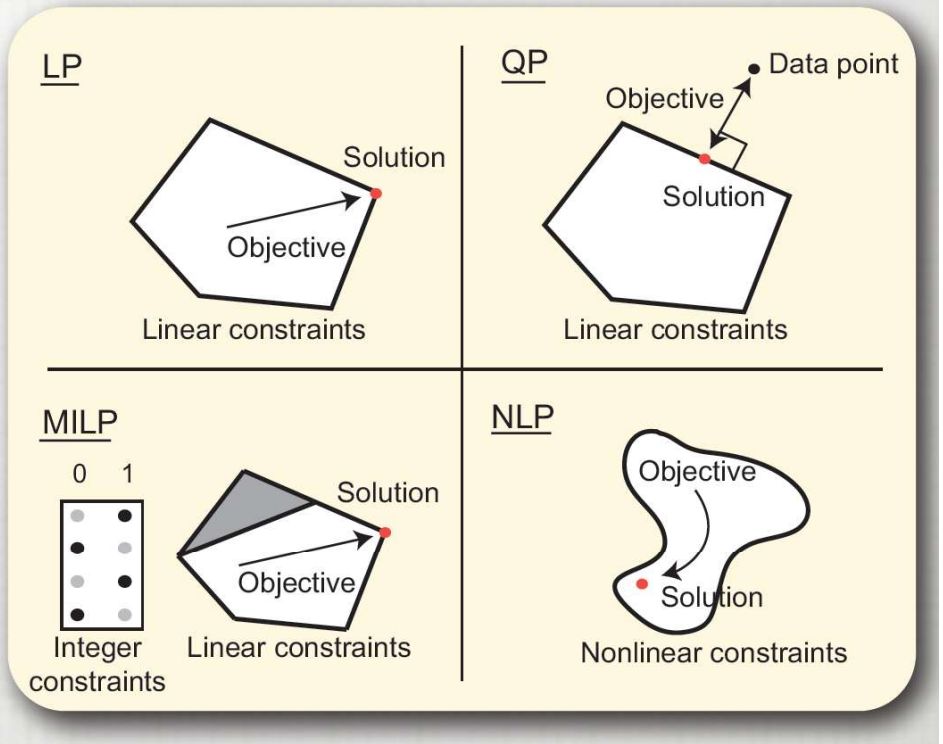
- Not biased by a statement of an objective

- Network based pathways:
  - Extreme Pathways
  - Elementary Flux Modes
- Decomposing flux distribution into extreme pathways
- Extreme pathways defining phenotypic phase planes
- Uniform random sampling

Biased Methods:

Constraint-based Optimization

- Linear programming (LP)
  - Linear constraints and linear objective function
- Quadratic programming (QP)
  - Linear constraint, 2nd order objective
  - Eucledian distance
- Mixed integer LP (MILP)
- Bi-level optimization (OptKnock)
- Non-linear programming (convex, concave)



58

Flux Balance Analysis (FBA) and LP Problems

FBA Optimization Formulation

Maximizing or minimizing a postulated objective function subject to a number of constraints gives rise to an optimization problem used to identify the metabolic flux distributions.

$$\text{maximize}\left(\text{or minimize}\right) z = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{v} \qquad \text{maximize}\left(\text{or minimize}\right)\ z = \sum_{j\in J} c_j v_j$$

$$\text{subject to} \qquad \text{subject to}$$

$$\boldsymbol{S}\boldsymbol{v} = 0 \qquad \sum_{j\in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$\text{LB} \le \boldsymbol{v} \le \text{UB} \qquad LB_j \le v_j \le UB_j, \quad \forall j \in J$$

$$\boldsymbol{v} \in \mathbb{R} \qquad v_j \in \mathbb{R}$$

Flux Balance Analysis (FBA) and LP Problems

$$\text{maximize} \quad z = v_{\text{biomass}} \quad [\text{FBA}]$$

$$\text{subject to}$$

$$\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$LB_j \leq v_j \leq UB_j, \quad \forall j \in J$$

$$v_j \in \mathbb{R}$$

Flux Balance Analysis (FBA) and LP Problems

Simulating Gene Knockouts

Metabolic gene knockouts can be captured by FBA using the GPR relationships that translate the effect of genetic interventions at the reaction level.

$$v_j = 0, \quad \forall j \in J^{\mathrm{KO}}$$

Flux Balance Analysis (FBA) and LP Problems

Gene/Reaction Essentiality and Synthetic Lethality

A gene or reaction whose deletion is lethal (i.e., arrests growth) is called an essential gene or reaction. “No growth” is captured in FBA as a maximum biomass flux of zero or less than a pre-specified viability threshold.

Gene non-essentiality implies that there may exist other genes that provide backup for the loss of function.

$$\begin{aligned}
&\text{maximize} \quad z = v_{\text{biomass}} \quad [\text{FBA-KO}] \\
&\text{subject to} \\
&\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I \\
&LB_j \le v_j \le UB_j, \quad \forall j \in J \\
&v_{j^*} = 0 \\
&v_j \in \mathbb{R}
\end{aligned}$$

Flux Balance Analysis (FBA) and LP Problems

Maximum Theoretical Yield of Product Formation

Many metabolic engineering applications of FBA involve the overproduction of a target metabolite. The most straightforward way for calculating the maximum product yield [MPY] capacity of the organism is to maximize the flux of the exchange reaction exporting that product metabolite.

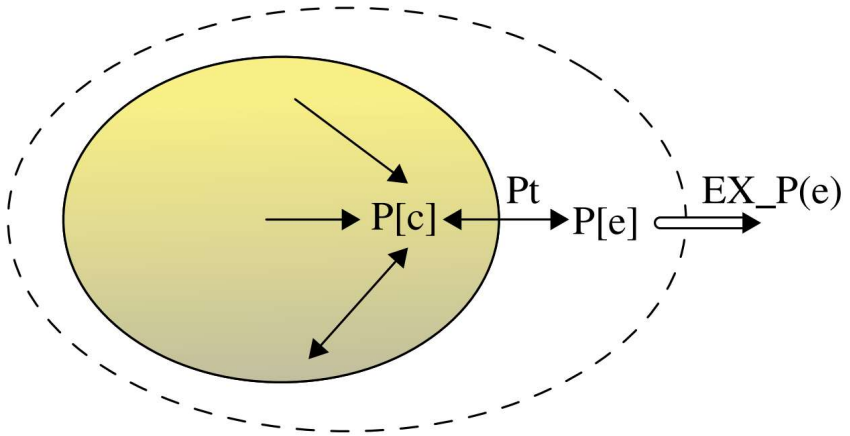
$$\text{maximize} \quad z = v_{\text{EX\_P(e)}} \quad [\text{MPY}]$$

$$\text{subject to}$$

$$\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$LB_j \le v_j \le UB_j, \quad \forall j \in J$$

$$v_j \in \mathbb{R}, \quad \forall j \in J$$



Flux Balance Analysis (FBA) and LP Problems

Biomass vs. Product Trade-Off

In most cases, the production of a chemical of interest is in direct competition with biomass formation. As a result, maximization of the product formation flux is associated with a zero biomass flux (i.e., no growth). This can be problematic for metabolic engineering applications where a minimum level of growth (biomass formation) is required.

$$\text{maximize / minimize} \quad z = v_{\text{EX\_P(e)}}$$

$$\text{subject to}$$

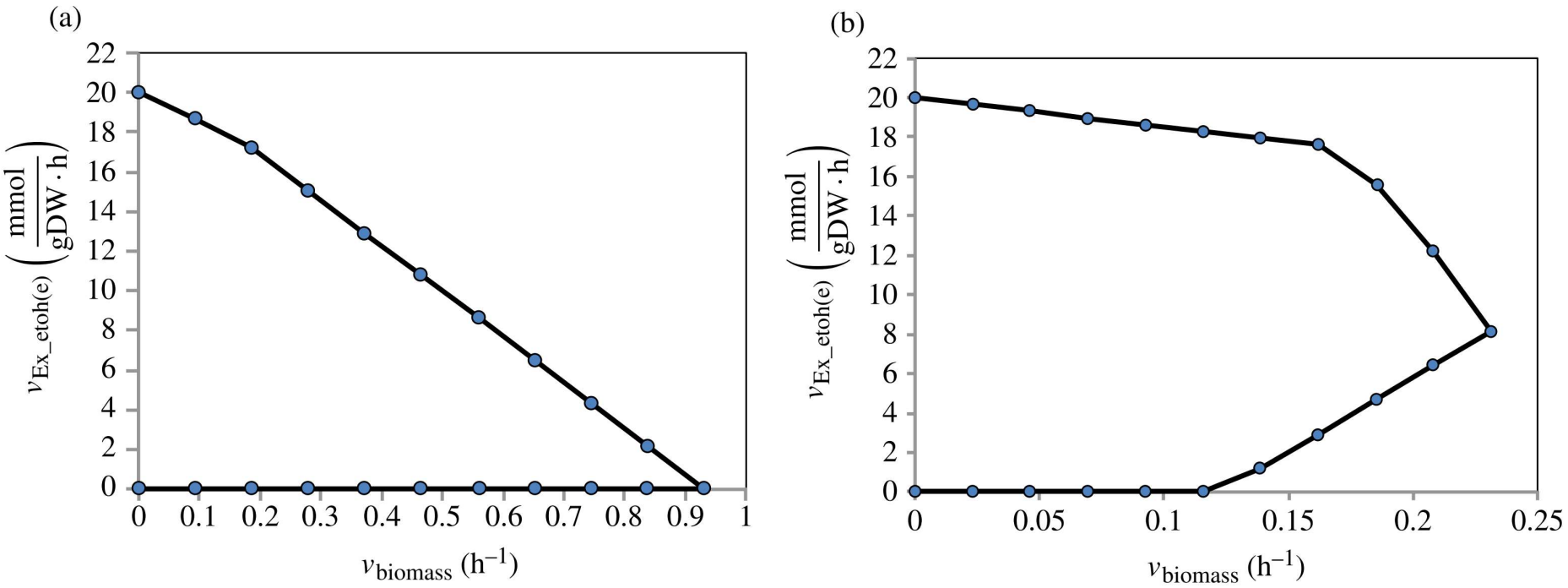
$$\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$LB_j \le v_j \le UB_j, \quad \forall j \in J$$

$$v_{\text{biomass}} \ge f\, v_{\text{biomass}}^{\max}$$

$$v_j \in \mathbb{R}$$

Flux Balance Analysis (FBA) and LP Problems



Trade-off plot for ethanol production in *Escherichia coli* under the (a) aerobic and (b) anaerobic condition in a minimal medium with glucose as the carbon source

Flux Balance Analysis (FBA) and LP Problems

Flux Variability Analysis (FVA)

Generally, LP problems arising in FBA involve alternate optima with many flux distributions leading to the same maximum biomass flux value. This means that different nonbasic variables can be brought into the basis without changing the optimal value of the objective function as they have a relative profit of zero.

$$\text{maximize}\left(\text{and minimize}\right) \quad z = v_j \quad [\text{FVA}]$$

$$\text{subject to}$$

$$\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$LB_j \le v_j \le UB_j, \quad \forall j \in J$$

$$v_{\text{biomass}} = f\, v_{\text{biomass}}^{\max}$$

$$v_j \in \mathbb{R}, \quad \forall j \in J$$

Modeling with Binary Variables and MILP Problems

Identifying Minimal Reaction Sets Supporting Growth

Generally, only a subset of genes is essential for cellular growth. The smallest set(s) of genes that can support growth is referred to as the minimal gene set. While many of these required genes have no (known) metabolic role, there exists a subset with distinct metabolic functions necessary for biomass formation. Identifying this subset provides insight into the minimal set of metabolic functions necessary for life.

Modeling with Binary Variables and MILP Problems

A binary variable is defined for each reaction in the network as follows

$$y_j = \begin{cases} 1, & \text{if reaction } j \text{ is active} \\ 0, & \text{otherwise} \end{cases}$$

The following MILP formulation identifies the minimal reaction set for the metabolic model of interest:

$$\text{maximize} \sum_{j \in J} y_j$$

$$\text{subject to}$$

$$\sum_{j \in J} S_{ij} v_j = 0, \quad \forall i \in I$$

$$\text{LB}_j y_j \le v_j \le \text{UB}_j y_j, \quad \forall j \in J$$

$$v_{\text{biomass}} \ge f v_{\text{biomass}}^{\max,\text{WT}}$$

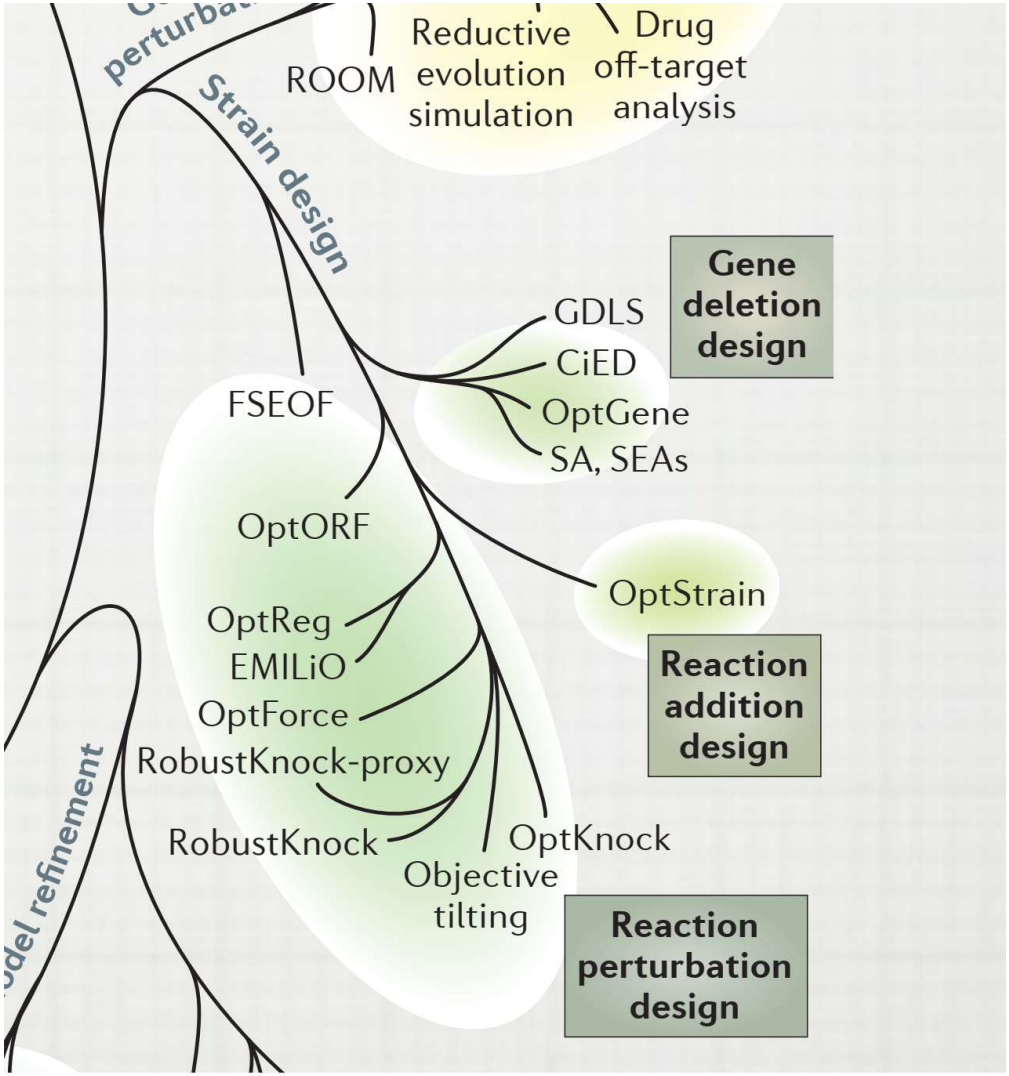
$$y_j \in \{0,1\}, \quad \forall j \in J$$

Computational (in-silico) Strain Design

- **Top-down approaches** (*OptKnock*, *OptReg*, *RobustKnock*, *OptGene*, *MoMAKnock*, *CiED*, *MCSEnumerator*, …): top-down strategies iteratively search for the metabolic reaction network until optimal solutions are identified

- **Bottom-up approaches** (*FSEOF*, *OptForce*, …): bottom-up approaches discover appropriate intervention strategies by comparing two flux distributions associated with the wild-type strain and the desired one.

Computational (in-silico) Strain Design



Computational (in-silico) Strain Design

OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization

Computational (in-silico) Strain Design

OptKnock

OptKnock recognizes that organisms tend to counteract any externally imposed genetic perturbations through the redirection of metabolic flux to restore cellular growth. It aims to design reaction eliminations that reshape network connectivity in such a way that the production of the target metabolite is maximized while the organism still maintains maximum biomass production yield under the constraints imposed by gene knockouts. Often, this renders the targeted overproduction an obligatory by-product of biomass formation.

Computational (in-silico) Strain Design

maximize **bioengineering objective**
(through gene knockouts)
subject to
maximize **cellular objective**
(over fluxes)
subject to
- fixed substrate uptake
- network stoichiometry
- blocked reactions identified by outer problem

number of knockouts ≤ limit

Computational (in-silico) Strain Design

$$\text{maximize} \quad z = v_{\text{EX\_P(e)}} \quad [\text{OptKnock}]$$

subject to

$$\left[\begin{array}{ll}
\text{maximize } v_{\text{biomass}} & \\
\text{subject to} & \\
\sum_{j \in J} S_{ij} v_j = 0, & \forall i \in I \\
v_{\text{EX_glc(e)}} \geq -v_{\text{glc}}^{\text{uptake}} & \\
v_{\text{EX_O}_2\text{(e)}} \geq -v_{\text{O}_2}^{\text{uptake}} & \\
v_{\text{ATPM}} = v_{\text{ATP}}^{\text{maint.}} & \\
v_{\text{biomass}} \geq f v_{\text{biomass}}^{\text{max,WT}}, & (0 \leq f \leq 1) \\
LB_j y_j \leq v_j \leq UB_j y_j, & \forall j \in J
\end{array}\right]$$

$$\sum_{j \in J} \left(1 - y_j\right) \leq K$$

$$y_j \in \{0,1\}, \quad v_j \in \mathbb{R}, \quad \forall j \in J$$

Computational (in-silico) Strain Design

The single-level MILP representation of OptKnock upon pre-processing is as follows:

maximize $z = v_{\mathrm{EX\_P(e)}}$

subject to

$$\sum_{j\in J} S_{ij} v_j = 0, \quad \forall i \in I$$
$$LB_j y_j \le v_j \le UB_j y_j, \quad \forall j \in J$$

} Primal constraints

$$\sum_{i\in I} S_{ij} \lambda_i + \mu_j^{\mathrm{UB}} - \mu_j^{\mathrm{LB}} = 0, \quad \forall j \in J - \{\mathrm{biomass}\}$$
$$\sum_{i\in I} S_{i,\mathrm{biomass}} \lambda_i + \mu_{\mathrm{biomass}}^{\mathrm{UB}} - \mu_{\mathrm{biomass}}^{\mathrm{LB}} = 1$$
$$0 \le \mu_j^{\mathrm{LB}} \le \mu_j^{\mathrm{LB,max}}, \quad \forall j \in J$$
$$0 \le \mu_j^{\mathrm{UB}} \le \mu_j^{\mathrm{UB,max}}, \quad \forall j \in J$$
$$\mu_j^{\mathrm{LB}} \le \mu_j^{\mathrm{LB,max}} (1 - y_j), \quad \forall j \in J - \{\mathrm{biomass}, \mathrm{ATPM}, \mathrm{EX\_glc(e)}, \mathrm{EX\_O_2(e)}\} \cup J^{\mathrm{irrev}}$$
$$\mu_j^{\mathrm{UB}} \le \mu_j^{\mathrm{UB,max}} (1 - y_j), \quad \forall j \in J - \{\mathrm{ATPM}\}$$

} Dual constraints

$$v_{\mathrm{biomass}} = v_{\mathrm{ATP}}^{\mathrm{maint.}} \mu_{\mathrm{ATPM}}^{\mathrm{UB}} - \left[\left(f v_{\mathrm{biomass}}^{\mathrm{max,WT}} \right) \mu_{\mathrm{biomass}}^{\mathrm{LB}} + \left(v_{\mathrm{ATP}}^{\mathrm{maint.}} \right) \mu_{\mathrm{ATPM}}^{\mathrm{LB}} + \left(-v_{\mathrm{glc}}^{\mathrm{uptake}} \right) \mu_{\mathrm{EX\_glc(e)}}^{\mathrm{LB}} + \left(-v_{\mathrm{O_2}}^{\mathrm{uptake}} \right) \mu_{\mathrm{EX\_O_2(e)}}^{\mathrm{LB}} \right]$$

} Strong duality

$$\sum_{j\in J} (1 - y_j) \le K$$

} Allowable number of knockouts

$$y_j \in \{0,1\}, \quad v_j \in \mathbb{R}, \quad \forall j \in J$$
$$\lambda_i \in \mathbb{R}, \quad \forall i \in I$$

Computational (in-silico) Strain Design

OptKnock Modifications

RobustKnock

A potential limitation of OptKnock arises when the suggested mutant involves a range of product formation yields under maximum biomass conditions. This is because OptKnock always selects the most optimistic (i.e., the highest) flux for the product formation under the maximum biomass condition eventhough the actual product formation flux can be less than this maximum or even zero (i.e., uncoupled from growth). RobustKnock is a modified version of OptKnock that was developed to address this issue by optimizing the worst-case scenario for the product formation while maximizing the biomass flux. To this end, the objective function of the outer problem in OptKnock is modified as:

$$\text{maximize}\left(\text{minimize}\, v_{\text{EX\_P(e)}}\right)$$

Computational (in-silico) Strain Design

which leads to a trilevel optimization problem as follows:

$$\text{maximize}\left(\text{minimize}\right) \quad v_{\text{EX\_P(e)}} \quad \left[\text{RobustKnock}\right]$$

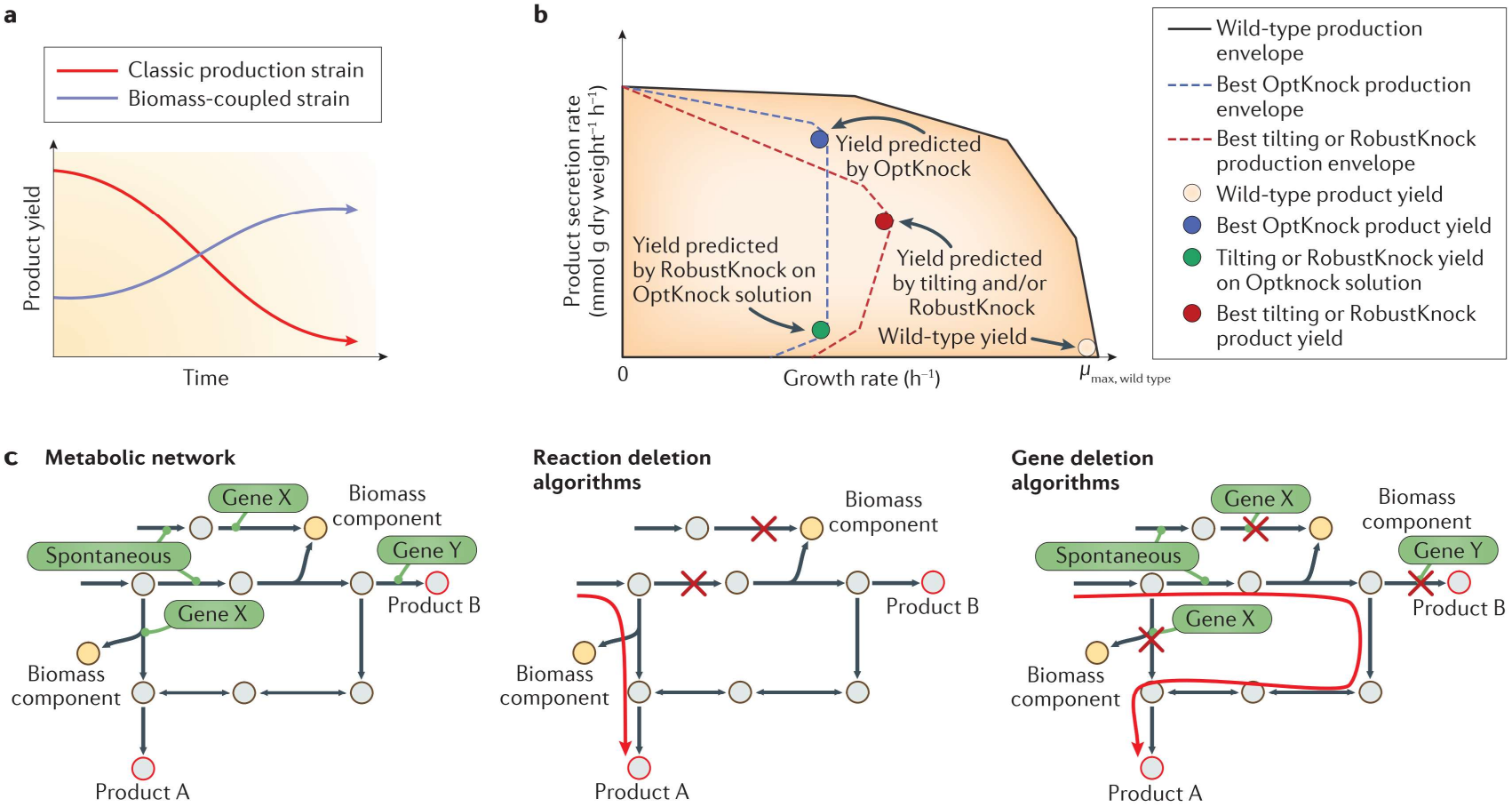
$$\text{subject to}$$

$$\left[\begin{array}{ll} \text{maximize } v_{\text{biomass}} & \\ \text{subject to} & \\ \sum_{j\in J} S_{ij} v_j = 0, & \forall i \in I \\ LB_j y_j \le v_j \le UB_j y_j, & \forall j \in J \end{array}\right]$$

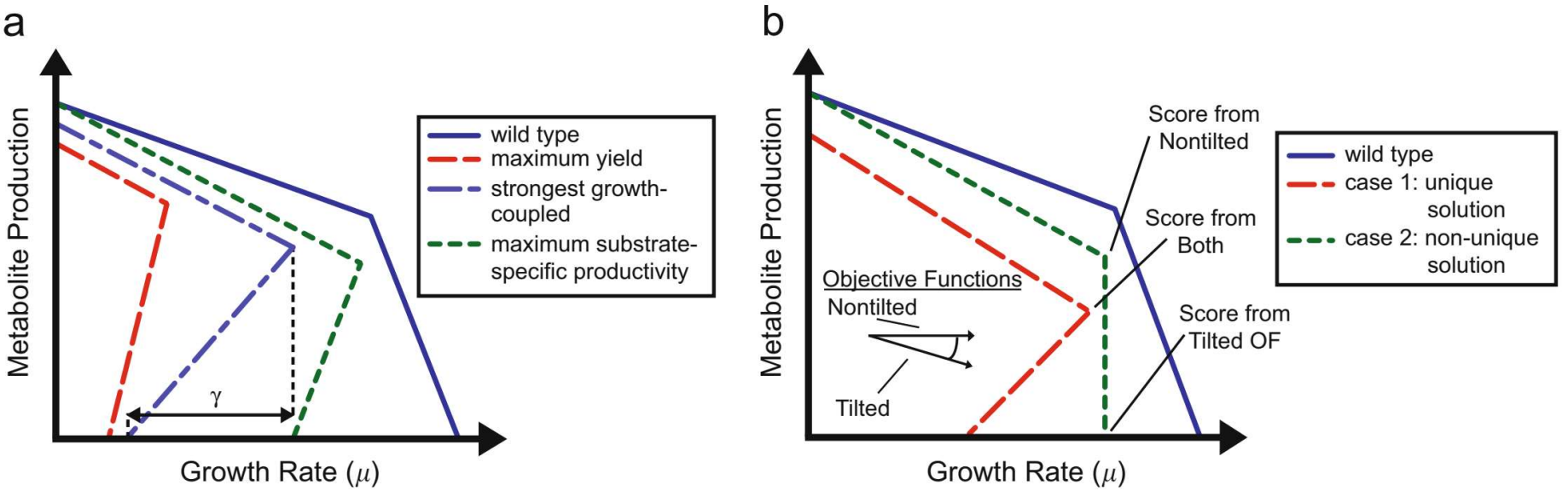
$$\sum_{j\in J}\left(1-y_j\right) \le K$$

$$y_j \in \{0,1\}, \quad v_j \in \mathbb{R}, \quad \forall j \in J$$

Computational (in-silico) Strain Design



Computational (in-silico) Strain Design



Computational (in-silico) Strain Design

Product yield ($Y_{p/s}$): Maximum amount of product that can be generated per unit of substrate

$$Y_{p/s} = \frac{production\ rate_{product}}{consumption\ rate_{substrate}} \left(\frac{mmol\ product}{mmol\ substrate} \right) \quad (1)$$

Substrate-specific productivity (SSP): Product yield per unit substrate multiplied by the growth rate

$$substrate\text{-}specific\ productivity = product\ yield \times growth\ rate$$

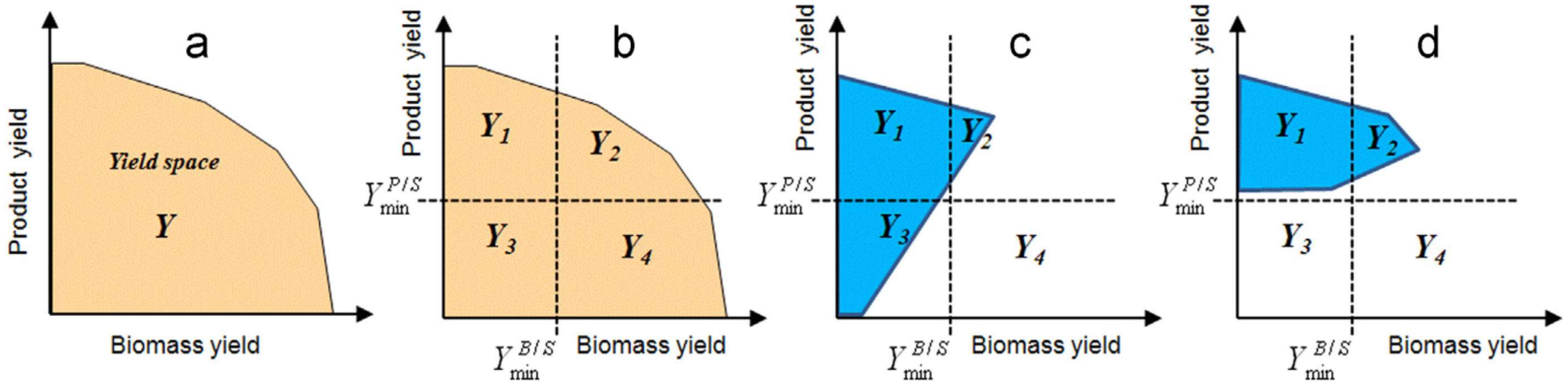
$$\left(\frac{mmol\ product}{mmol\ substrate \times hr} \right) \quad (2)$$

Strength of coupling: Product yield per unit substrate squared divided by the slope of the lower edge of the production curve

$$strength\ of\ growth\ coupling = \frac{(product\ yield)^2}{slope}$$

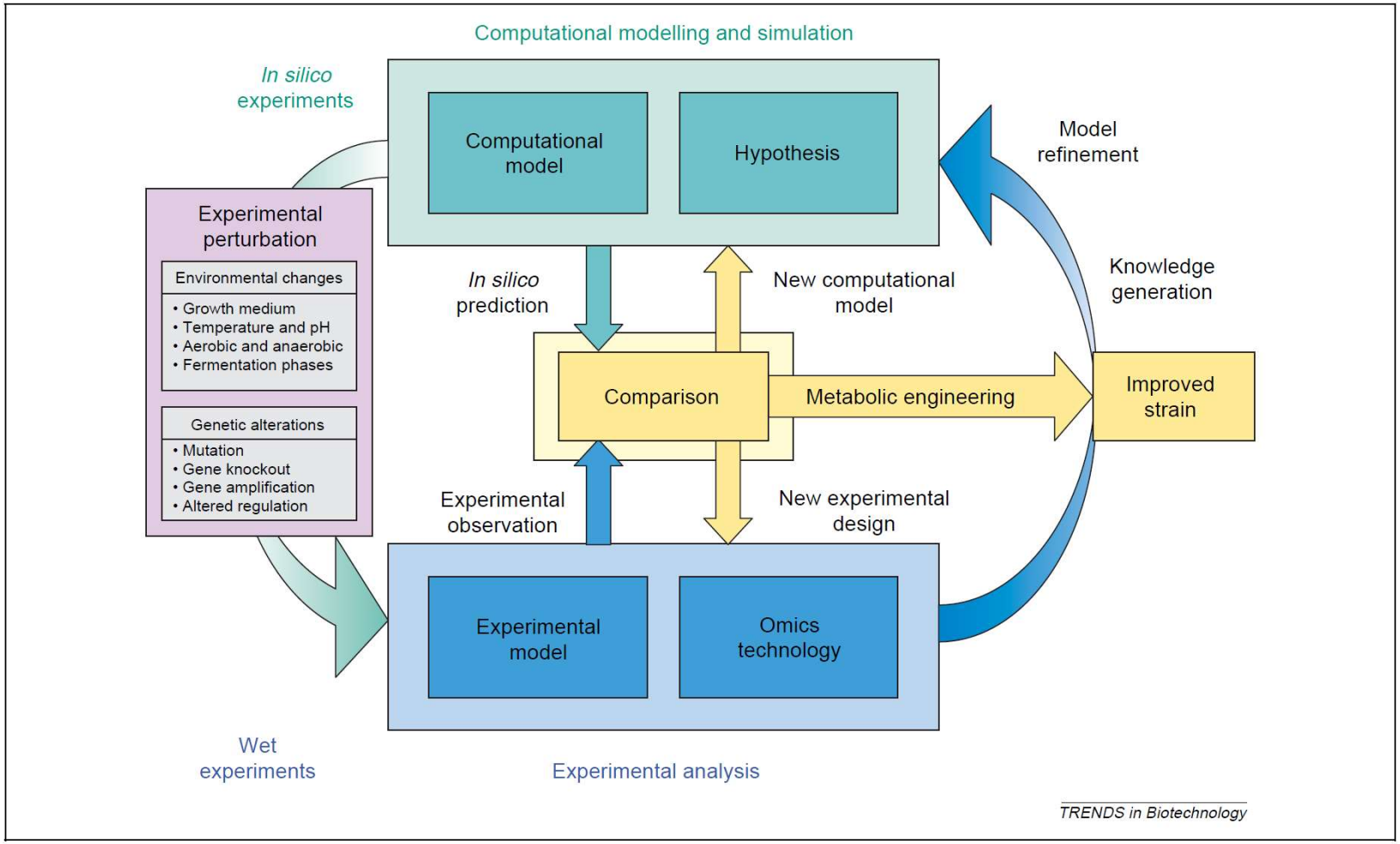
$$\left(\frac{mmol\ product}{mmol\ substrate \times hr} \right) \quad (3)$$

Computational (in-silico) Strain Design



Examples of yield spaces. (a) Typical shape of a yield space. (b) Introducing minimum thresholds for product and biomass yield dissects the yield space into four regions. (c) Example of a yield space with weak coupling of biomass and product synthesis. (d) Example of a yield space with strong coupling of biomass and product synthesis where high product synthesis is attained also under low biomass yields.

Concluding Remarks



Concluding Remarks

Expectations from systems biology

- *Health care*
  - Understanding diseases as malfunctions of normal cells or the interaction of cells with pathogens
  - Personalized medicine can take into account individual characteristics and conditions
- *Biotechnology*
  - Design and production of cells with desired properties
  - Production of cheap drugs
  - Energy

Escher Examples

Thank You for Your Attention!