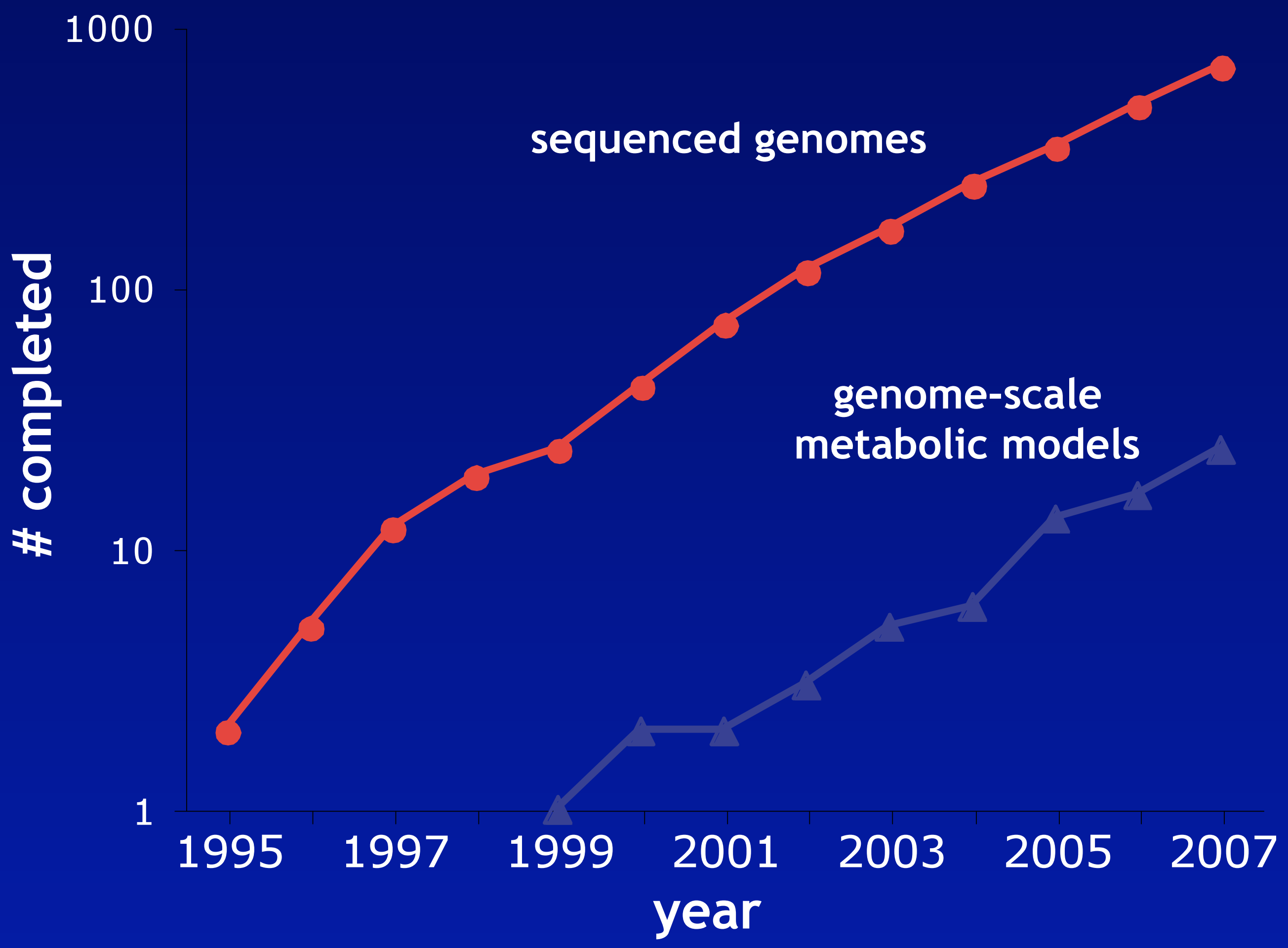
1. *Analysis and Redesign of Proteins and Biological Networks*

Costas Maranas / The Pennsylvania State University

- Biological Networks

- Development of computational workflows for reconstructing the complete metabolic repertoire of microbial and plant systems (i.e., *Mycoplasma genitalium*, *Methanosarcina acetivorans*, etc.)
- Automated testing/curation of metabolic models for completeness and correctness by using multiple types of data (i.e., network connectivity, gene essentiality experiments, metabolomic and transcriptomic data).
- Construction of algorithmic tools and mapping databases that allow for metabolic flux analysis (MFA) by tracking the fate of labeled atoms through metabolic networks.
- Development of computational tools for identifying all possible engineering strategies (i.e., knock in/out/up/down's) leading to increased production of a targeted molecule (e.g., a biofuel) using a microbial or plant production system.

Genome-scale metabolic models vs. sequenced genomes



Metabolic Reconstruction Technology

Genome Annotation:

DNA sequence

↓ ORFs identification

Genes

↓ ORFs assignment

Genes Products

↓

Function

Genome Database

ORF = open reading frame, a short fragment of DNA that is translated into RNA message

Metabolic Reconstruction:

List of reactions

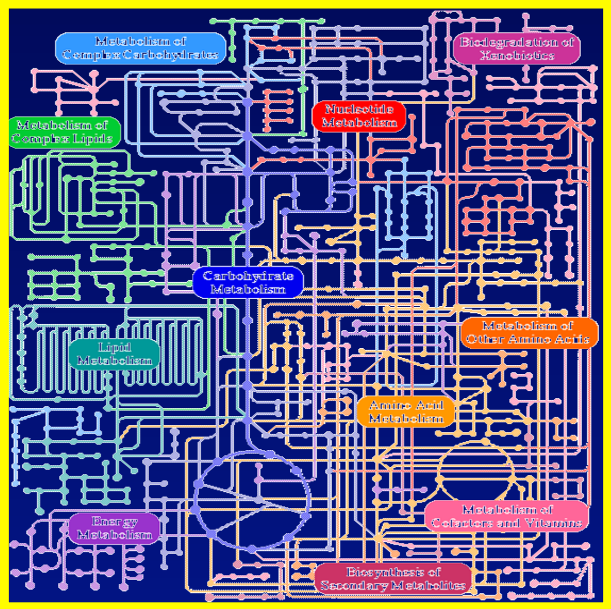
Pathway Database → Organism's metabolism

Organism-Specific Model Construction:

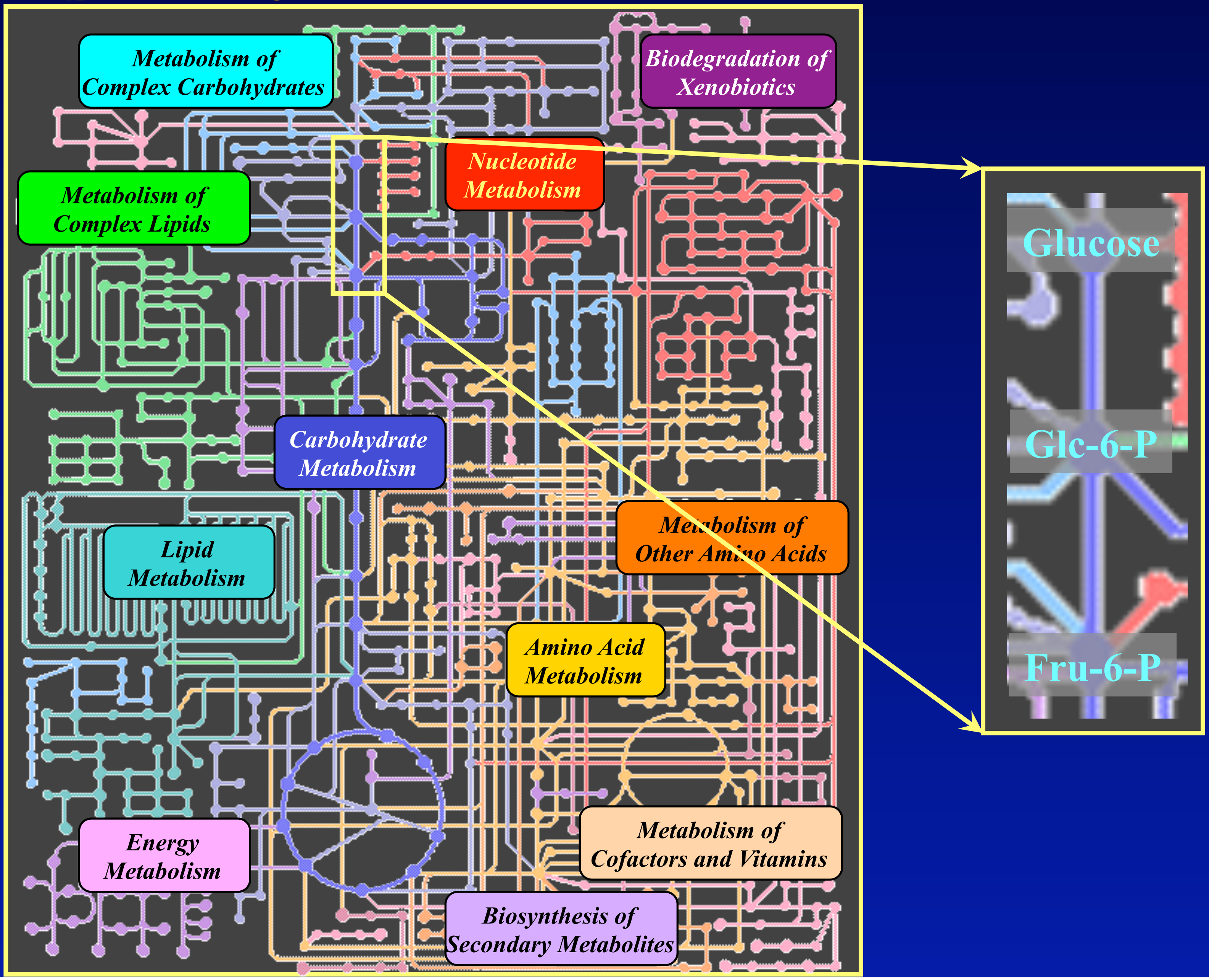
Literature Review ↔ Manual curation

Literature Review ↔ Wet Lab

Manual curation ↔ Wet Lab

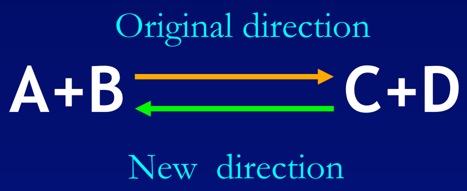

Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Metabolism of Complex Lipids
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Metabolism of Cofactors and Vitamins
Energy Metabolism
Biosynthesis of Secondary Metabolites

Complexity of Metabolic Networks

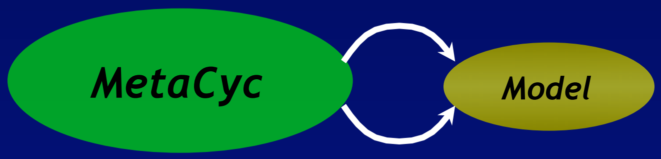

Metabolism of Complex Carbohydrates
Biodegradation of Xenobiotics
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites
Glucose
Glc-6-P
Fru-6-P

GapFill: Filling Connectivity gaps in model

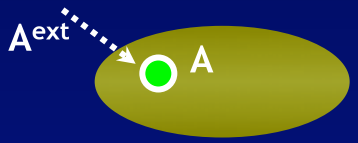
Reversing Directionality

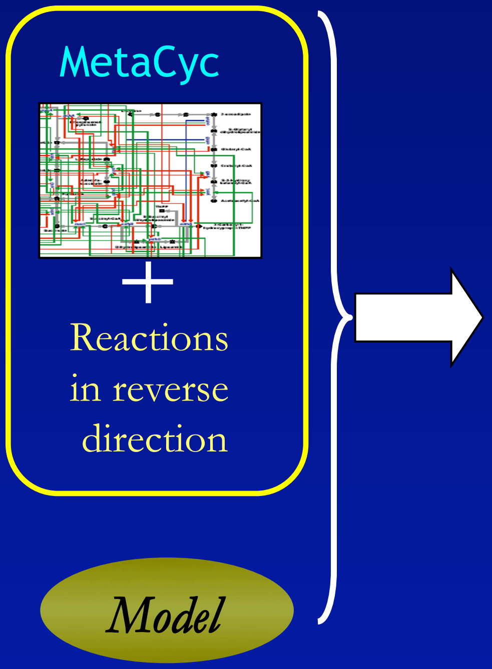


Addition of
Missing reactions



Addition of
Uptake route





minimize (# of rxn additions
and direction reversals)

subject to

- Network stoichiometry
- Net production term > 0, for each NPM
- Bounds on fluxes

GrowMatch: Restore consistency with G/NG experiments

Model Testing: Contrast model (*in silico*) predictions vs. experimental (*in vivo*) gene deletion data (e.g., Keio Collection (Baba et al. 2006))

Model Prediction

Experimental data	Growth (G)	No Growth (NG)
Growth	G/G=GG	NG/G =NGG
No Growth	G/NG=GNG	NG/NG=NGNG

Model under-predicts metabolic capabilities

- Absence of relevant reactions

Model over-predicts metabolic capabilities

- Presence of extra/wrong reactions
- Down-regulation of rxns in exp. conditions

Fix NGG $\xrightarrow{\text{add rxns}}$ GG

Fix GNG $\xrightarrow{\text{suppress rxns}}$ NGNG

NGNG ➡ GNG

GG ➡ NGG

Resolution of inconsistent experiments… while avoiding changing others that are already correct

➡ **Model modifications must be performed while taking into account entire model and all experimental data**

M. genitalium *model* *iPS189*

“Minimal gene”
model organism

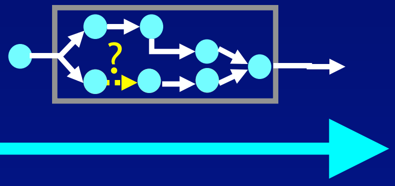
- Comparison of *M. genitalium with H. influenzae*
  overlap of 256 genes
  (Mushegian & Koonin, 2000)

(Collaboration with J.C. Venter Inst.)

- Global transposon mutagenesis
  382 essential genes our of 482 ORFs
  (Glass et al. PNAS 2006)
- (Genome transplantation)
  (Lartigue, et al. Science 2007)
- Synthetic genome construction
  (Gibson, et al. Science 2008)

GapFind & GapFill

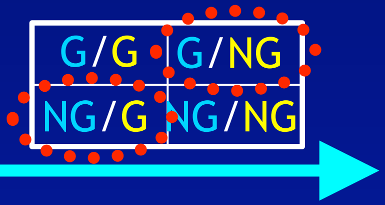
(Satish Kumar, et al., *BMC Bioinformatics*, 2008)



reconnected 25 metabolites
added 22 rxns & GPR for 8 genes

GrowMatch

(Satish Kumar and Maranas PLoS Comp Biol, accepted)



G/G	G/NG
NG/G	NG/NG

increased agreement with *in vivo* gene essentiality data from 79% to 87%

*i*PS189 contains:

189 (39%) ORFs

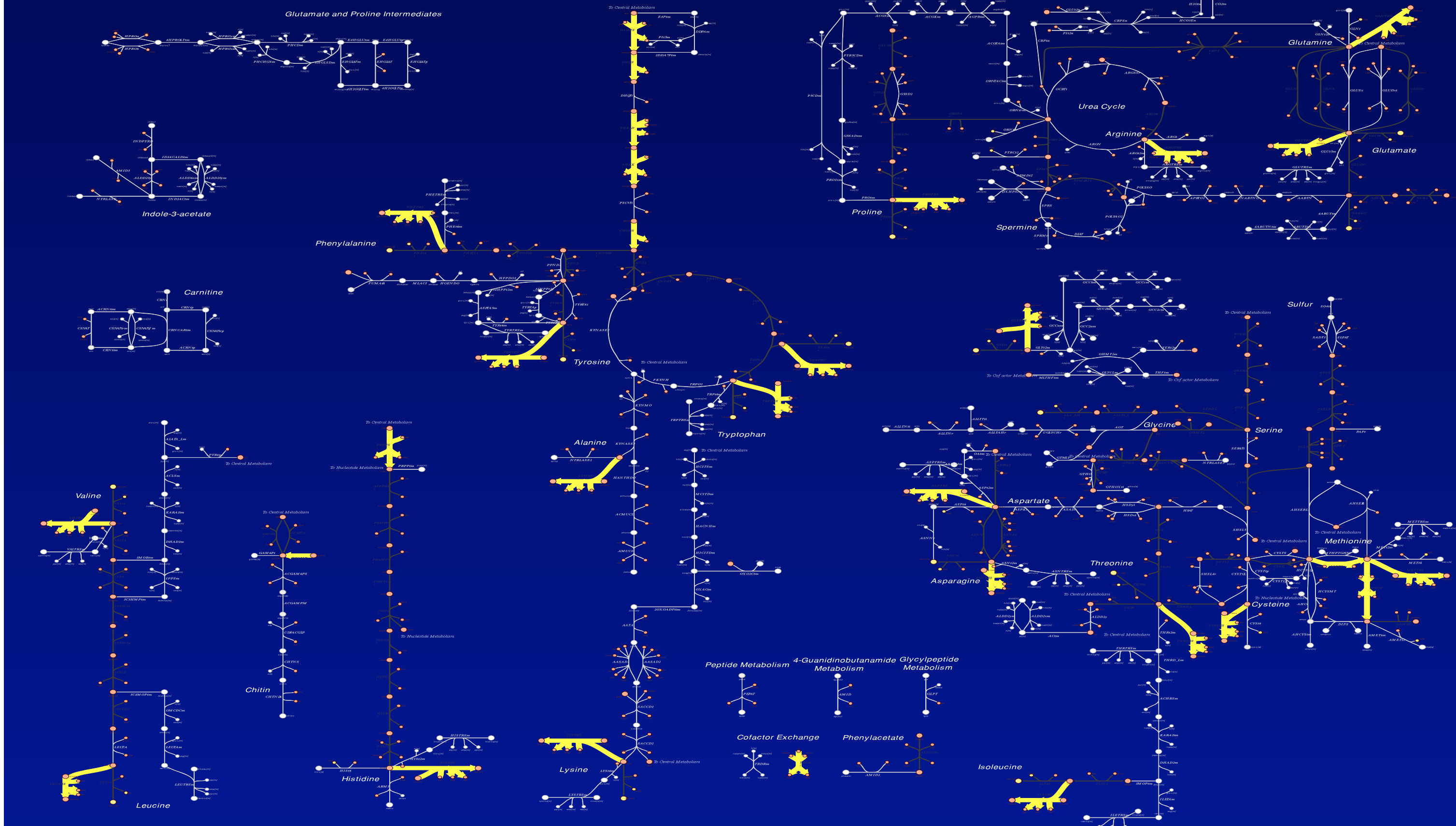
262 reactions

274 metabolites

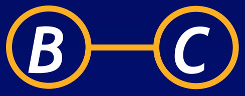
(Suthers, et al, PLoS Comp Biol)

Synthetic lethality- Definition

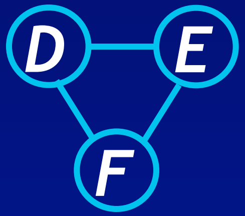
Amino acid metabolism



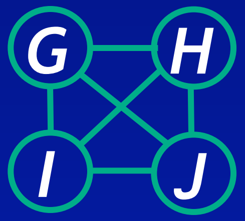
Essential genes



Synthetic lethal (SL) pairs



SL triples

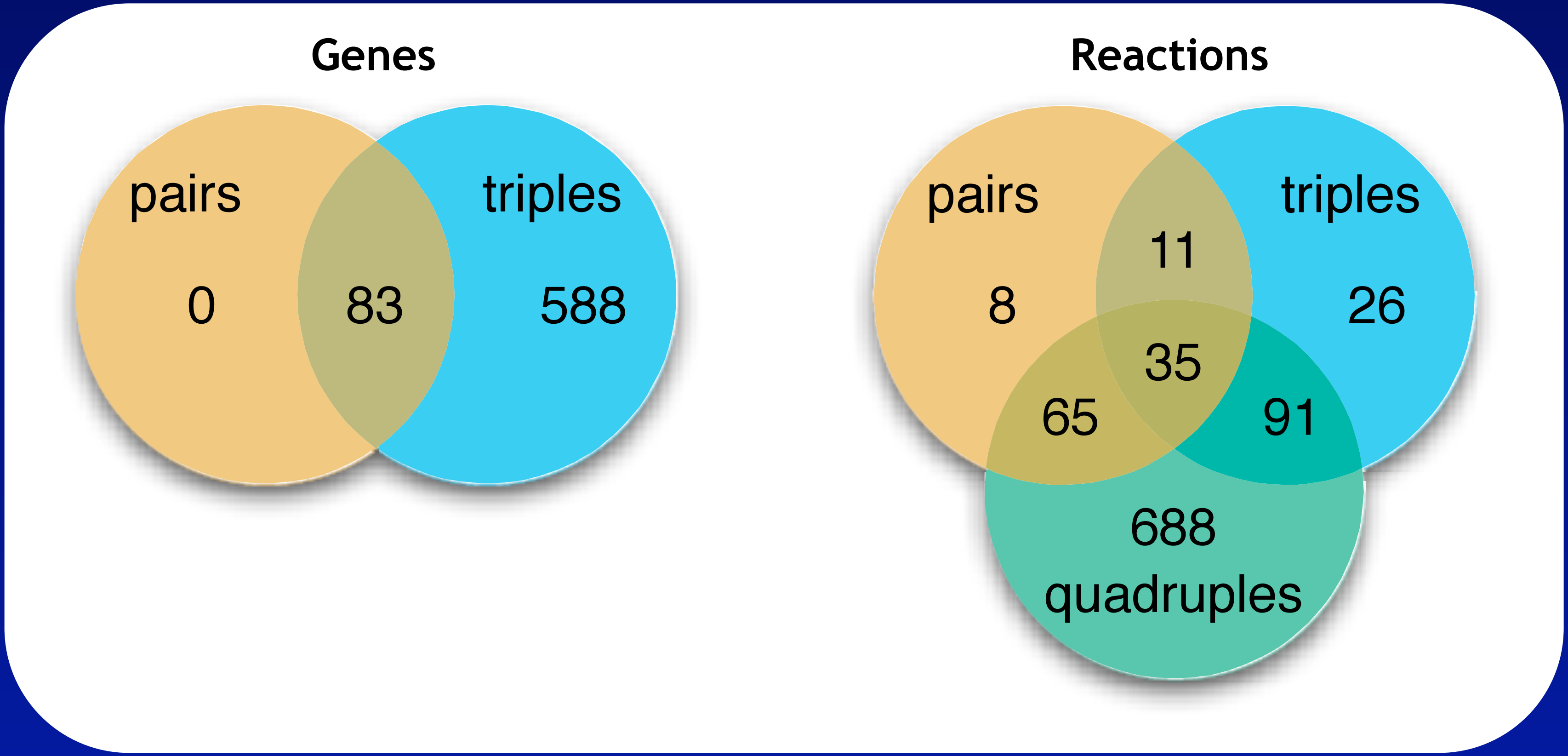


SL quadruples

- Reveal organizing principles of metabolism & patterns of dispensability
- Characterize genes/rxns w.r.t. their degree of essentiality
- Provide additional layer for curating metabolic models

Participation in higher order SLs

A gene/reaction involved in a SL pair can also participate in SL triples or even higher order SLs



Targeted enumeration of SLs

Direct Enumeration: Chose order of synthetic lethals = n

(e.g., n=2 synthetic lethal pairs, n=3 synthetic lethal triplets, etc.)

Outer Problem

Find synthetic rxn eliminations negating biomass formation

Inner Problem

Adjust fluxes to find the max biomass production potential of the network

Minimize Biomass flux

(over sum of rxn eliminations = n)

s.t.

Maximize Biomass flux

(over fluxes)

s.t. Network connectivity

Uptake/secretion conditions

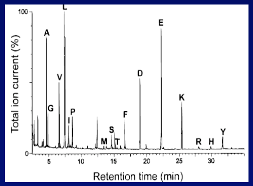
No flow in eliminated rxns by outer problem

If max biomass < cutoff ⇒ Report synthetic lethal

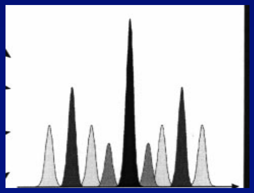
Elucidating fluxes in metabolic models (MFA)

Principle: Deconvolute fluxes in metabolic networks based on distribution of labels in measured metabolites

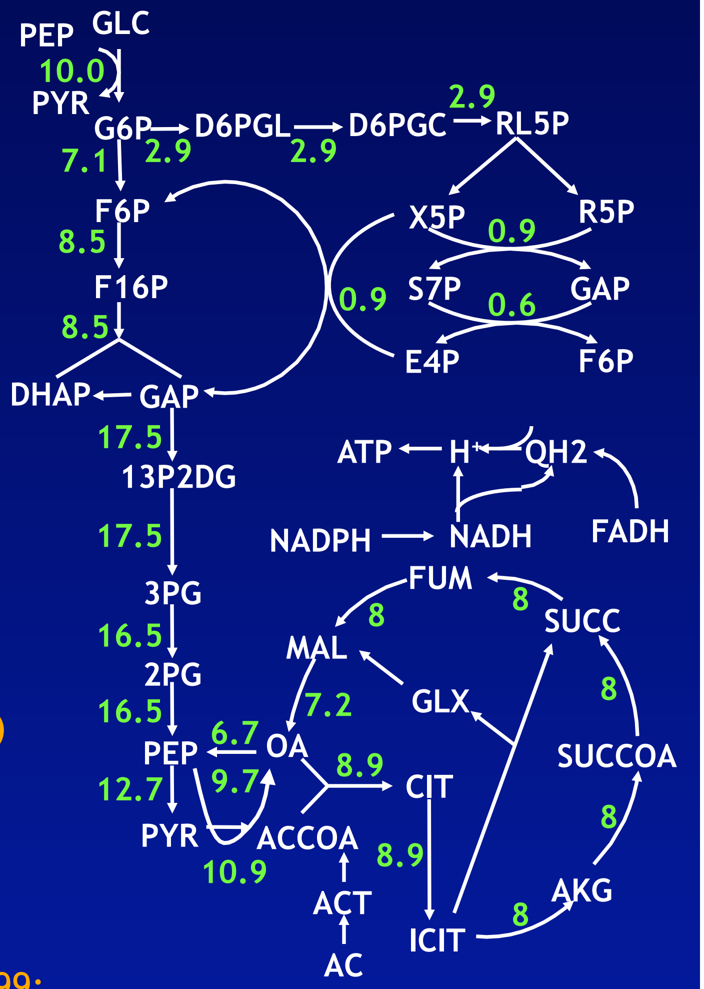
labeled isotopes



GC-MS

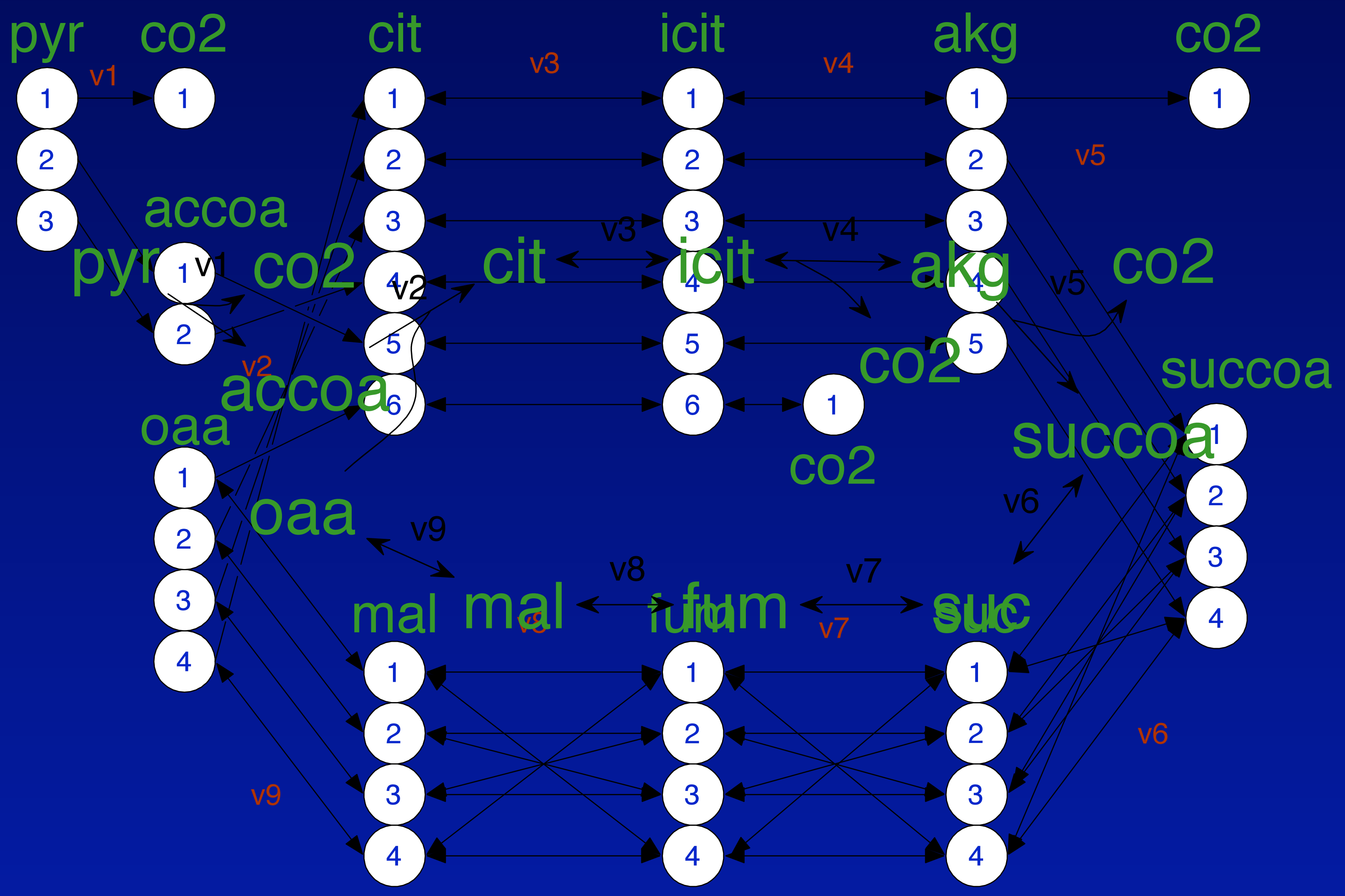


NMR



- ❑ Isotopomer analysis using GC/MS
  (Park *et al*. 1997; Christensen & Nielsen 1999; Fischer & Sauer 2003)
- ❑ Isotopomer analysis using NMR spectra
  (Marx *et al*. 1996; Schmidt *et.al*. 1999)
- ❑ Computational models for flux elucidation
  (Zupke *et al*. 1994; Wiechert & Graff 1996; Wiechert *et.al*. 1996,1999; Mollney *et al*. 1999; van Winden *et al*. 2002; Antoniewicz *et al*. 2006,2007)
- ❑ Optimization algorithms
  (Ghosh *et al*. 2004; Fiascos *et al*. 2004; Phalakornkule *et al*. 2001)

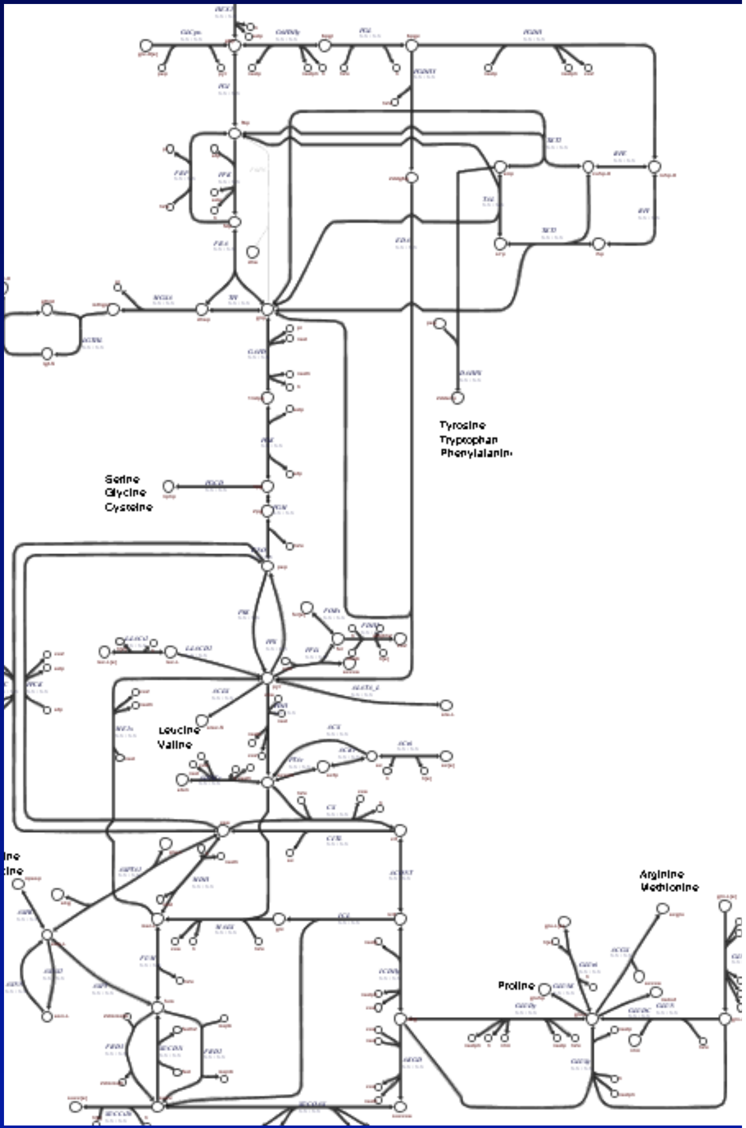
Atom transition view of the TCA cycle



Isotopomer Mapping Model

Suthers *et.al.*, *Metab Eng* 2007

- *E. coli* metabolism
- Includes 238 reactions, 184 metabolites, 17,346 isotopomers
  - Glycolysis
  - TCA cycle
  - Pentose phosphate pathway
  - Anaplerotic reactions
  - Amino acid biosynthesis/ degradation
  - Oxidative phosphorylation
- Balances on cofactors such as NADH, NADPH, and ATP
- Detailed biomass flux that drains the proper proportion of precursor metabolites



Genome-scale isototope mapping model

Atom mapping procedure

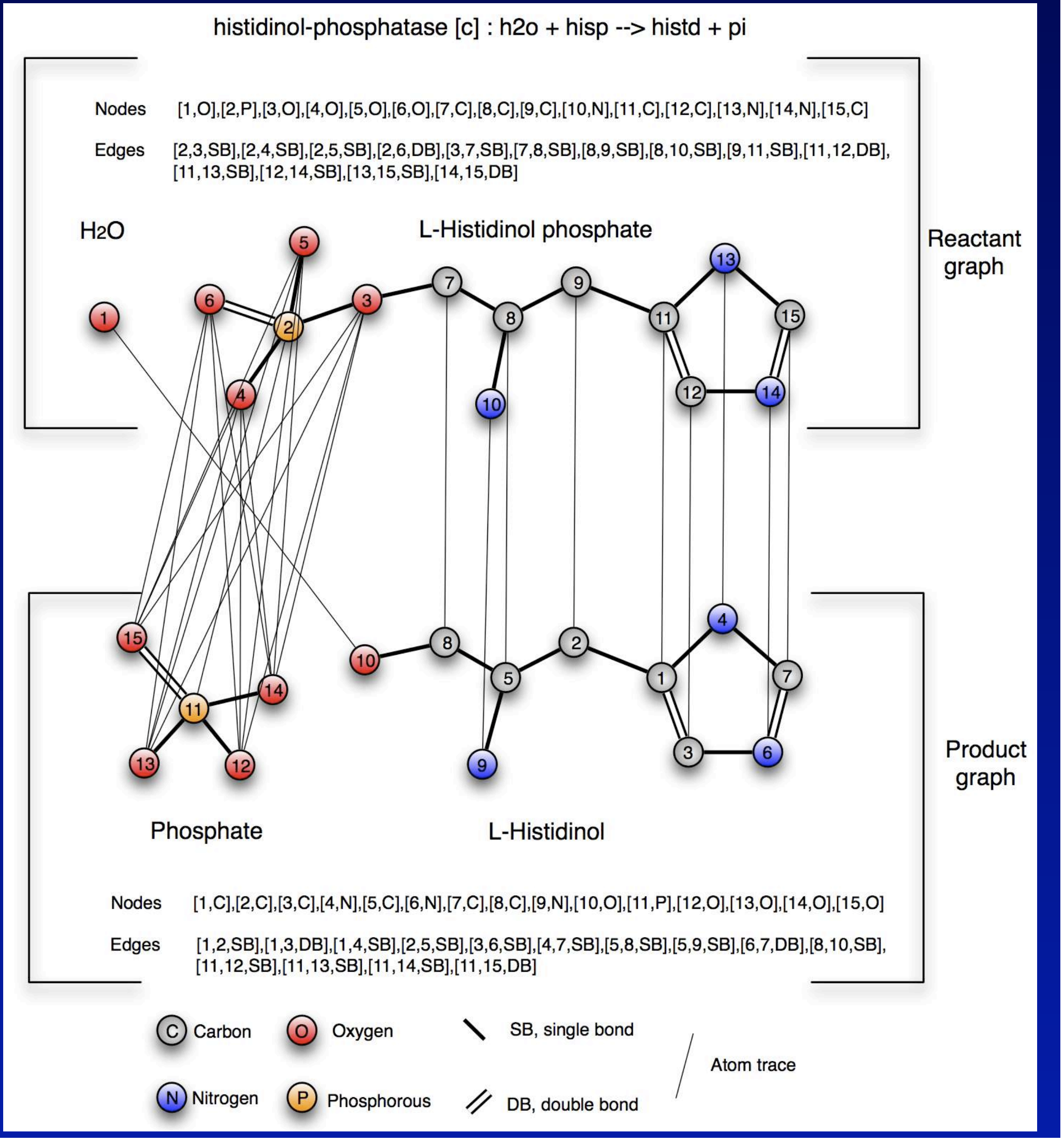
1. Identification of metabolites with unchanging labelings and elucidation of recurring reaction motifs

2. Generation of reactant and product molecular graphs

3. Construction of atom mappings between reactant and product graphs

4. Automated curation to retain correct atom mappings based on reaction chemistry

“Seamless” integration with genome scale models



histidinol-phosphatase [c] : h2o + hisp --> histd + pi

Nodes [1,O],[2,P],[3,O],[4,O],[5,O],[6,O],[7,C],[8,C],[9,C],[10,N],[11,C],[12,C],[13,N],[14,N],[15,C]

Edges [2,3,SB],[2,4,SB],[2,5,SB],[2,6,DB],[3,7,SB],[7,8,SB],[8,9,SB],[8,10,SB],[9,11,SB],[11,12,DB],[11,13,SB],[12,14,SB],[13,15,SB],[14,15,DB]

Nodes [1,C],[2,C],[3,C],[4,N],[5,C],[6,N],[7,C],[8,C],[9,N],[10,O],[11,P],[12,O],[13,O],[14,O],[15,O]

Edges [1,2,SB],[1,3,DB],[1,4,SB],[2,5,SB],[3,6,SB],[4,7,SB],[5,8,SB],[5,9,SB],[6,7,DB],[8,10,SB],[11,12,SB],[11,13,SB],[11,14,SB],[11,15,DB]

Genome-scale isotope mapping model

(based on iAF1260 E. coli model)

90,068 atoms traced in 2,077 reactions

Atoms traced:

C:	49,539
O:	29,061
P:	3,280
N:	2,386
S:	409

5,393other non-hydrogen atoms
(Ag, As, Ca, Cd, Cl, Co, Cu, Fe, Hg, K, Mg, Mn, Na, Ni, Se, W, Zn, and halogens)

Isotopomers

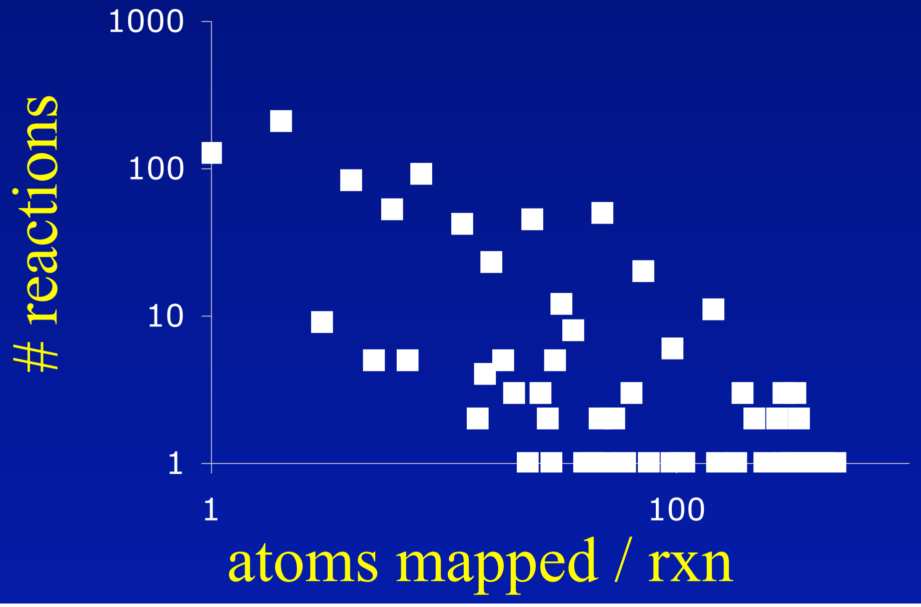
C: 8.34 x 10^{93}

O: 1.61 x 10^{60}

N: 2.58 x 10^{7}

P: 10,006

S: 4,091



Total complexity:
~10^{180} isotopomers

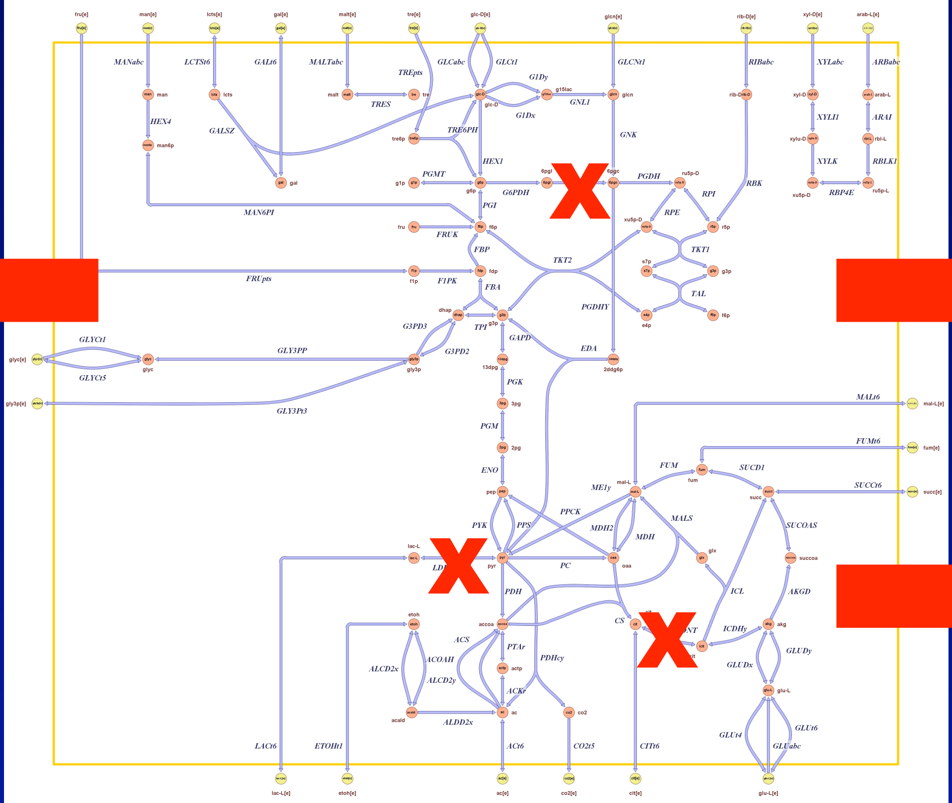
➔ Elemental Metabolite Units (EMU) to activate only part of model consistent with labeling

OptKnock methodology

(Burgard *et al*, *Biotech Bioeng*, 84, 647-657, 2003)

Computational Method to design biocatalysts that try to **couple** biomass with product formation

Cellular Objective
(*ex:* biomass yield)

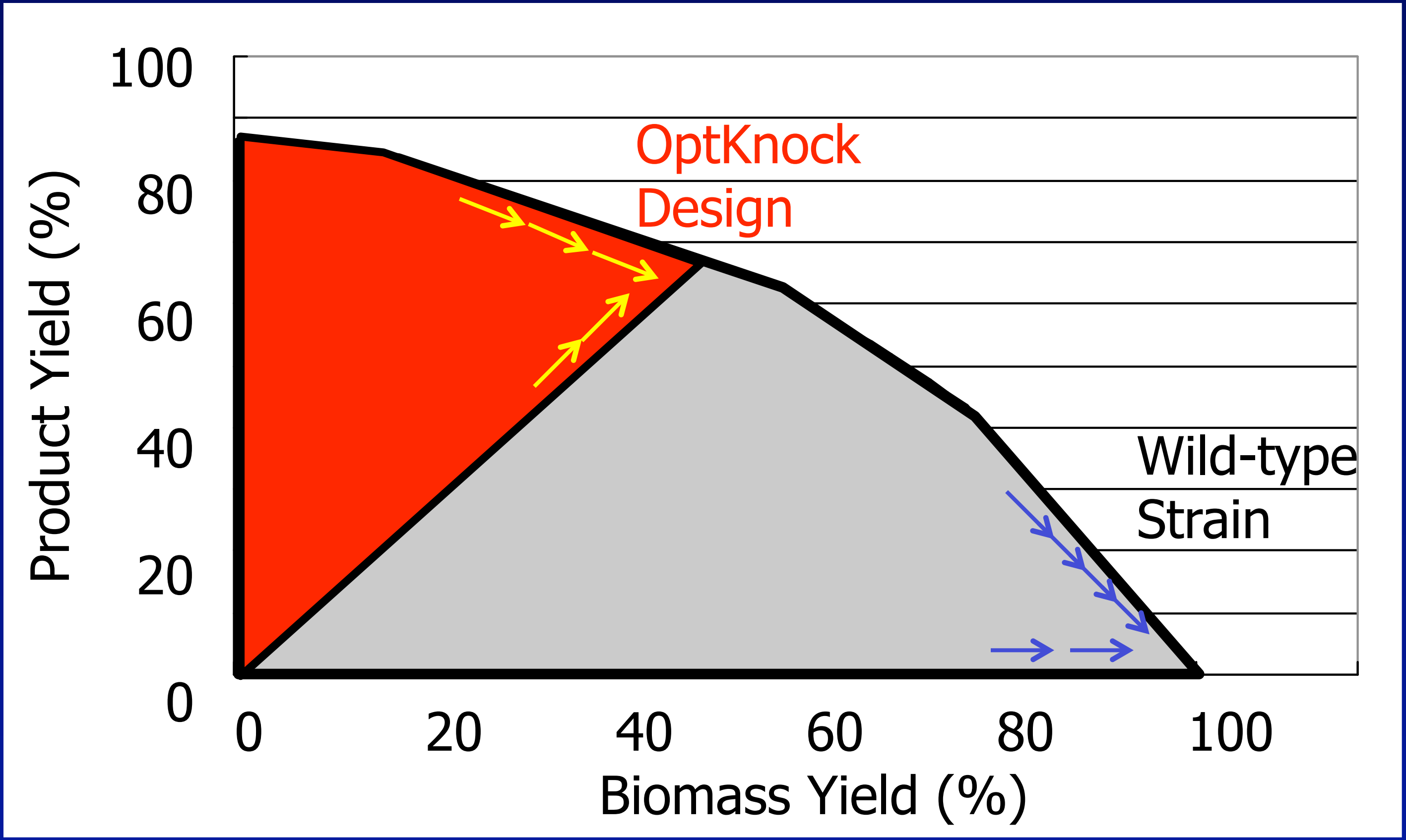


Bioengineering Objective
(*ex:* product yield)

Cellular Objective
(*ex:* biomass yield)

Graphical Illustration

Trade-off plot between biomass and product formation



Idea: Constraint phenotypic space so as max biomass yield
brings about a high product yield

OptKnock bilevel optimization framework

Outer Problem:

adjust ***knockouts***

→ optimize bioeng. objective

- Max. 1,3-propanediol yield
- Max. lactate yield

Inner Problem:

adjust ***reaction fluxes***

→ optimize cellular objective

- Max. biomass yield
- Min. metabolic adjustment
- Max. ATP yield

Maximize Biochemical Yield
(over gene knockouts)

s.t.

Maximize Biomass Yield
(over fluxes)

s.t.

□ **Fixed substrate uptake rate**

□ **Network connectivity**

□ **Blocked reactions identified by outer problem**

□ **Minimum biomass yield**

□ **# Knockouts ≤ limit**

- Burgard, A.P., Pharkya, P., and C.D. Maranas (2003), "OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization," *Biotechnology and Bioengineering*, 84, 647-657.
- Pharkya, P., Burgard, A.P., and C.D. Maranas (2003), "Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock," *Biotechnology and Bioengineering*, 84, 887-899.

Computational strain design

Existing Strategies:	METAOPT	OptKnock	OptStrain	OptGene	OptReg	MFSSCOF
	(Hatzimanikatis *et al.* 1996)	(Burgard *et al.* 2003)	(Pharkya *et al.* 2004)	(Patil *et al.* 2005)	(Pharkya *et al.* 2006)	(Lee *et al.* 2007)

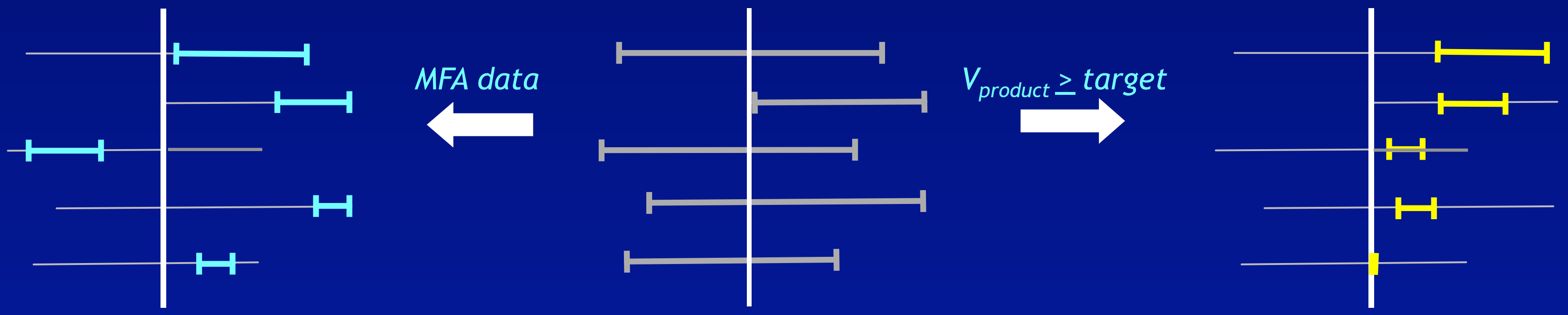
Limitations:

1. Generate one “redesign” at a time
2. Use of surrogate objective functions (e.g., max biomass or min MOMA)
3. No direct use of MFA or other flux data

Wild-type flux ranges (with MFA data)

Wild-type flux ranges (without MFA data)

Flux ranges required for overproduction



Min / Max v_j
s.t. *MFA data*
Stoichiometry
Uptake

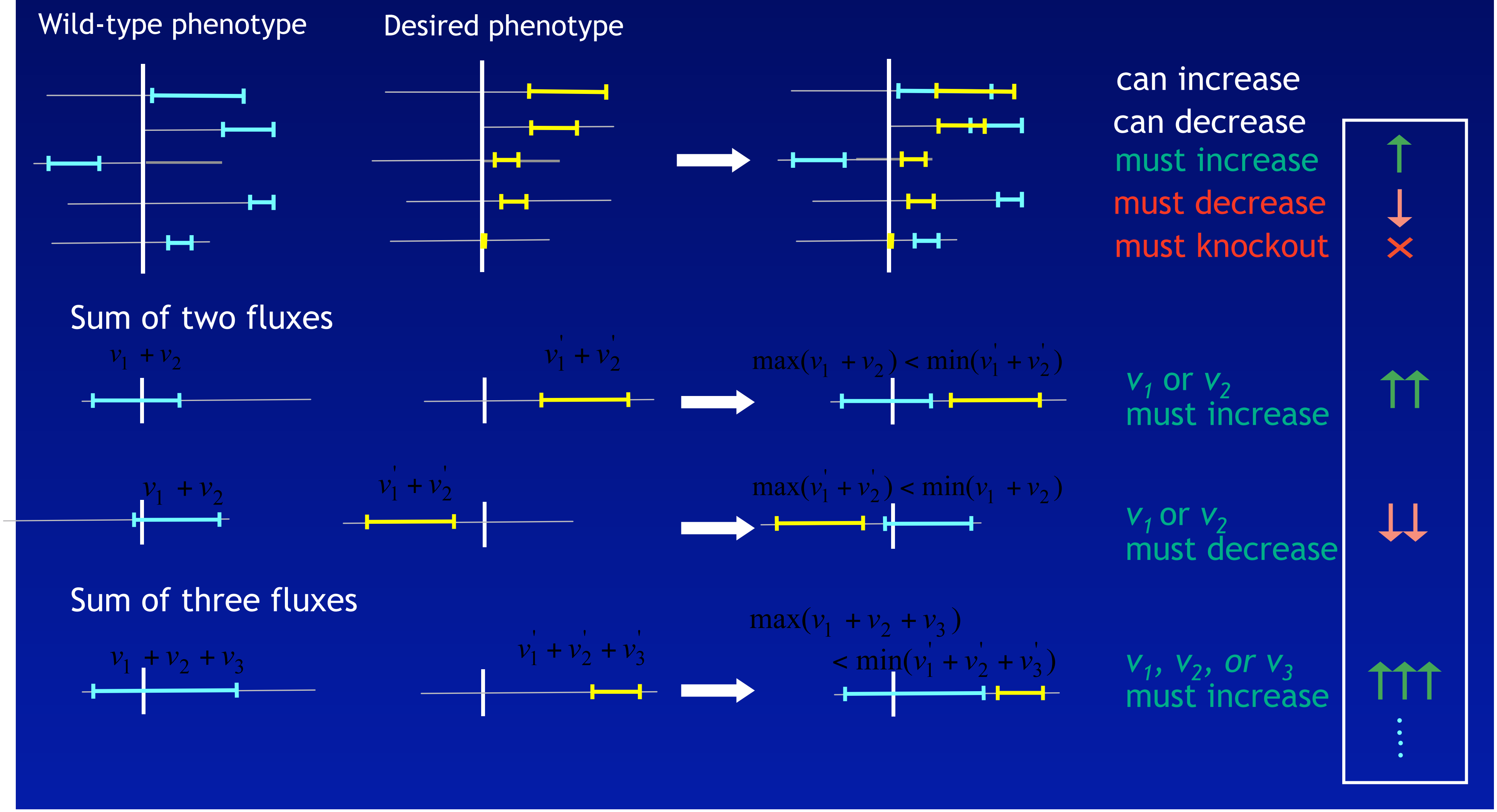
Min / Max v_j
s.t. *Stoichiometry*
Uptake

Min / Max v_j
s.t. *Stoichiometry*
Uptake
$V_{product} \geq target$

Suthers *et al.* Met. Eng. (2007)

Flux range classifications (MUST sets)

Key Idea: Identify all individual reactions and combinations thereof whose total flux value *MUST increase, decrease or be knocked out* to meet a newly imposed production target



Wild-type phenotype

Desired phenotype

can increase

can decrease

must increase ↑

must decrease ↓

must knockout ×

Sum of two fluxes

$v_1 + v_2$

$v_1' + v_2'$

$\max(v_1 + v_2) < \min(v_1' + v_2')$

v_1 or *v_2* must increase ↑↑

$v_1 + v_2$

$v_1' + v_2'$

$\max(v_1' + v_2') < \min(v_1 + v_2)$

v_1 or *v_2* must decrease ↓↓

Sum of three fluxes

$v_1 + v_2 + v_3$

$v_1' + v_2' + v_3'$

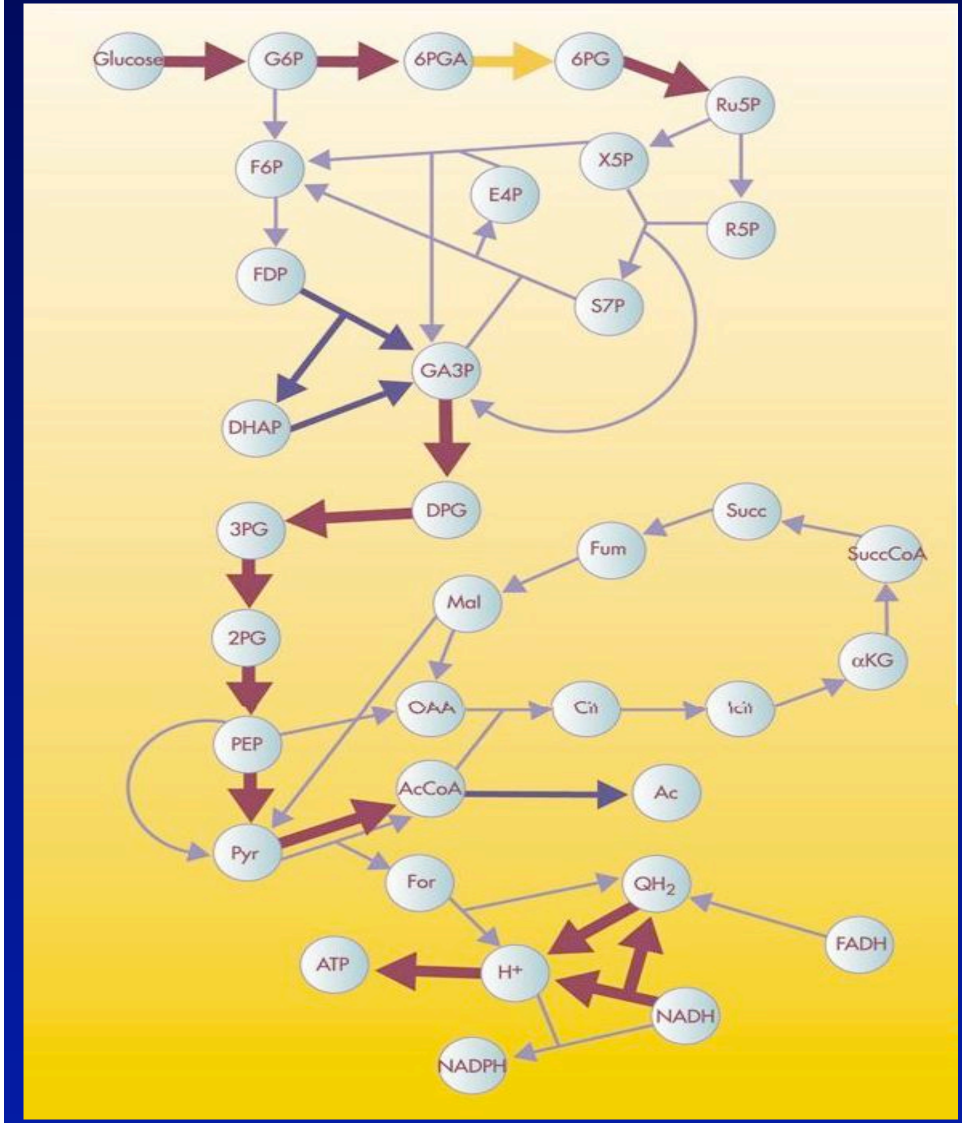
$\max(v_1 + v_2 + v_3) < \min(v_1' + v_2' + v_3')$

v_1, v_2, or v_3 must increase ↑↑↑

⋮

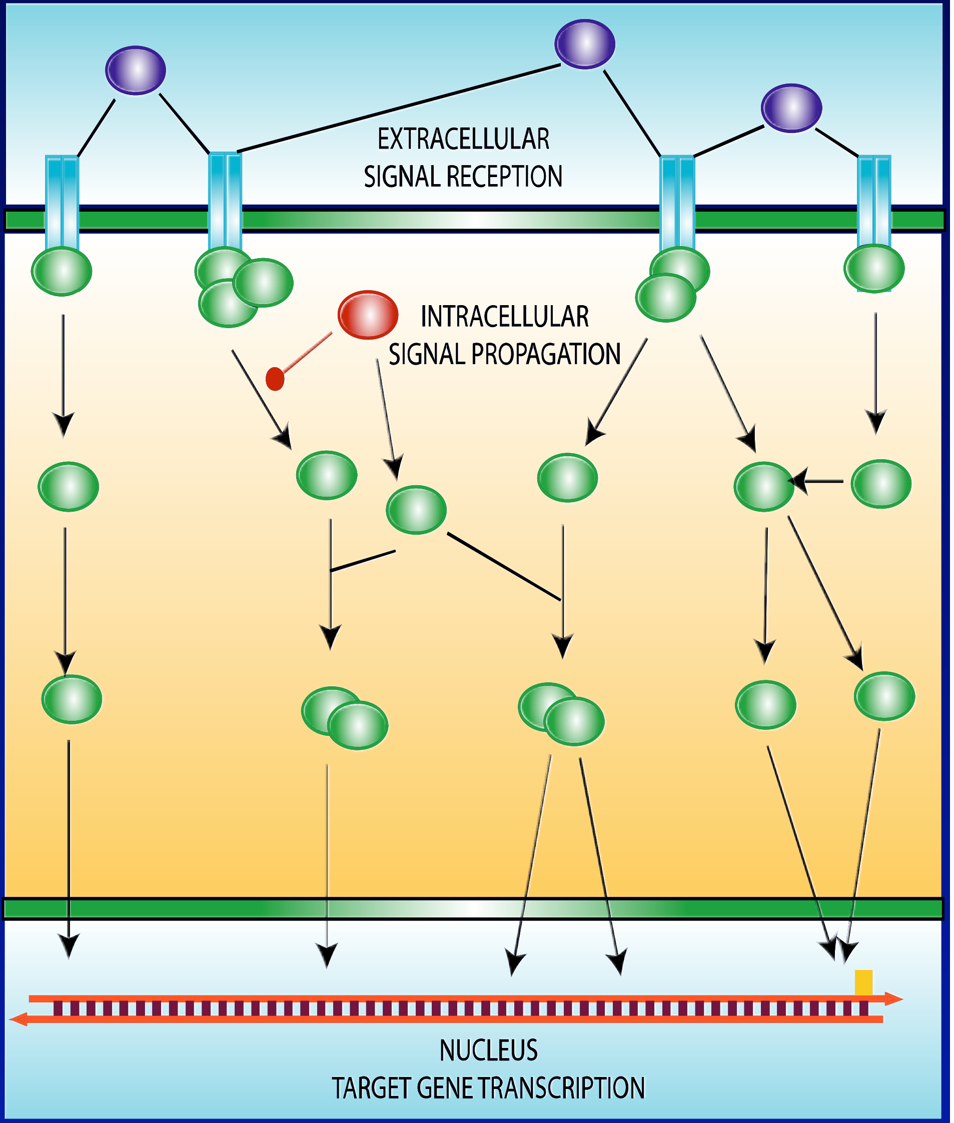
Networks…

Metabolic Networks



http://doegenomestolife.org

Signaling Networks



1. *Analysis and Redesign of Proteins and Biological Networks*

Costas Maranas / The Pennsylvania State University

- **-Protein Design**
- Computational identification of mutation(s) leading to improved enzymatic function (i.e., P450 small alkane oxidation, cellulases)

- →Substrate/cofactor binding calculations at the ground state

→ Estimation of energy barriers along reaction coordinate

→ Transfer of binding/active to a new protein scaffold

→ Derivation of scoring functions for protein library design

2. Current HPC Requirements

(see slide notes)

- Architectures: Linux cluster
- Compute/memory load: 4 to 200 hrs / up to 10 GB
- Data read/written: less than 1GB
- Necessary software, services or infrastructure: In-house developed software IPRO, OptGraft, GapFill, GrowMatch, OptKnock, etc. and commercially available codes include CPLEX, CONOPT, CHARMM, Gaussian03
- Current primary codes and their methods or algorithms: Primary codes rely on algorithms for solving MILP and NLP optimization and combinatorial graph algorithms. Parallelism is currently handled by manually seggregating computing tasks to different computing nodes
- Known limitations/obstacles/bottlenecks: NP-hard nature of underlying mathematical problems. Both compute time and memory can be limiting

3. HPC Usage and Methods for the Next 3-5 Years

(see slide notes)

- Upcoming changes to codes/methods/approaches: As size and complexity of biological networks increases this will tax the computational performance of the analysis, curation and redesign tools
- Changes to Compute/memory load: 300 hrs / 30 GB+
- Changes to Data read/written: increase, but remain < 1GB
- Changes to necessary software, services or infrastructure: use of decomposition methods; new parallelizable versions of solvers
- Anticipated limitations/obstacles/bottlenecks on 10K-1000K PE system.

4. Summary

- What new science results might be afforded by improvements in NERSC computing hardware, software and services?

  - -Ability to perform flux elucidation in genome-scale metabolic reconstructions including plant systems and communities
  - -Global identification of strain optimization strategies
  - -*De novo* protein design

- Recommendations on NERSC architecture, system configuration and the associated service requirements needed for your science

- General discussion