Systems Biology

Nicolás Loira
Center for Mathematical Modeling
November 2014
nloira@gmail.com

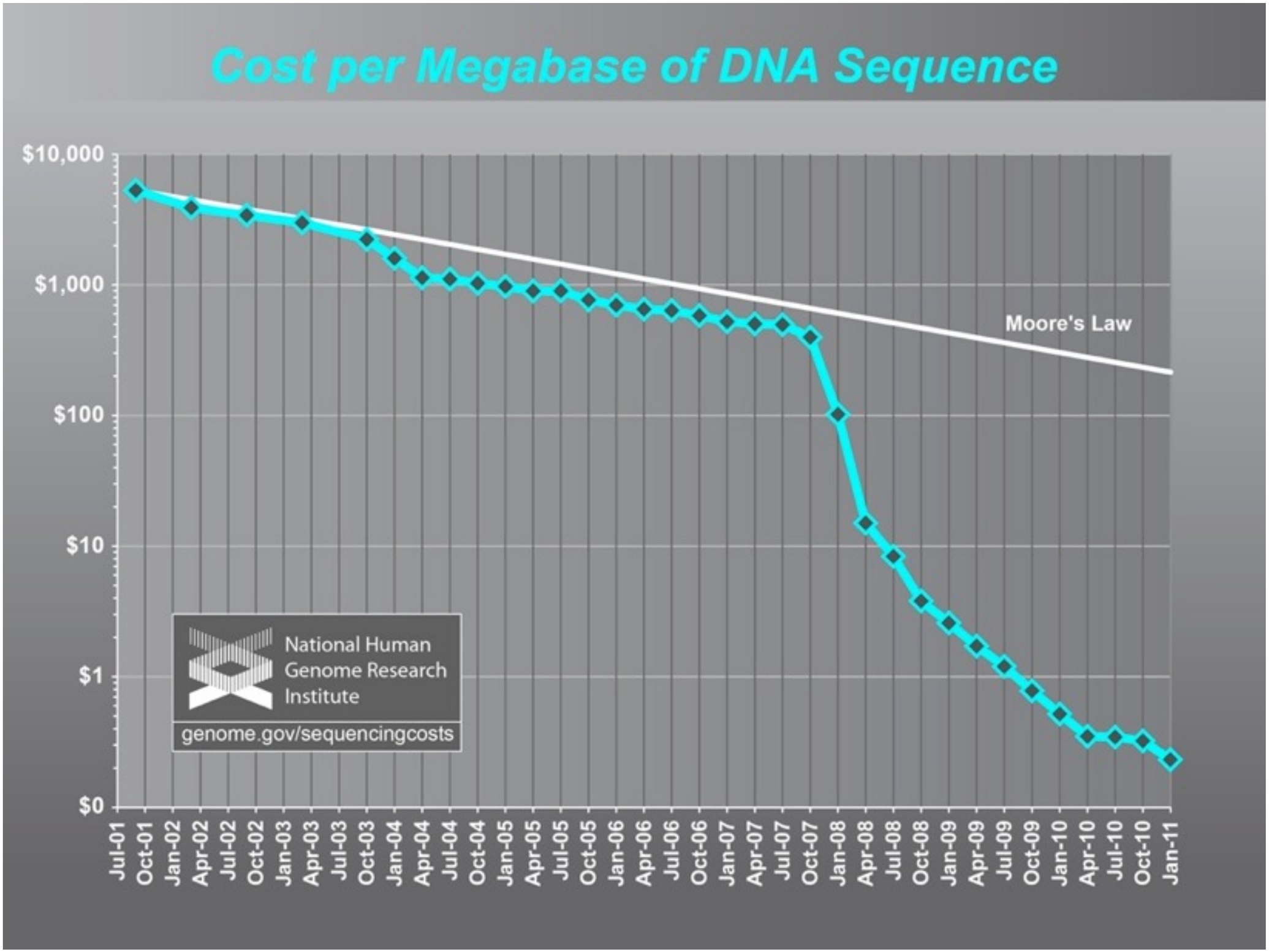
Overview

Intro to Systems Biology

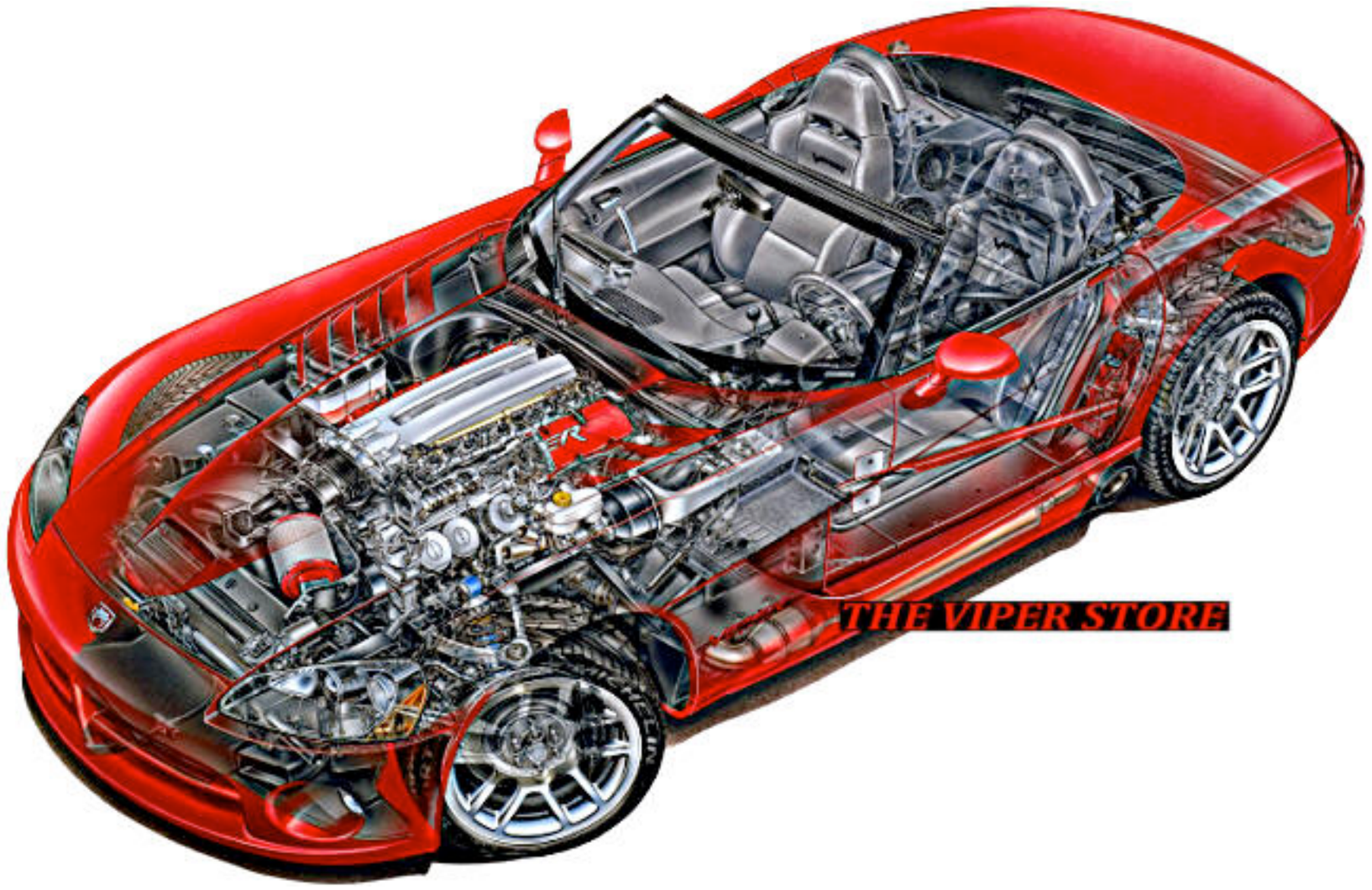
Metabolic Models

Systems Biology

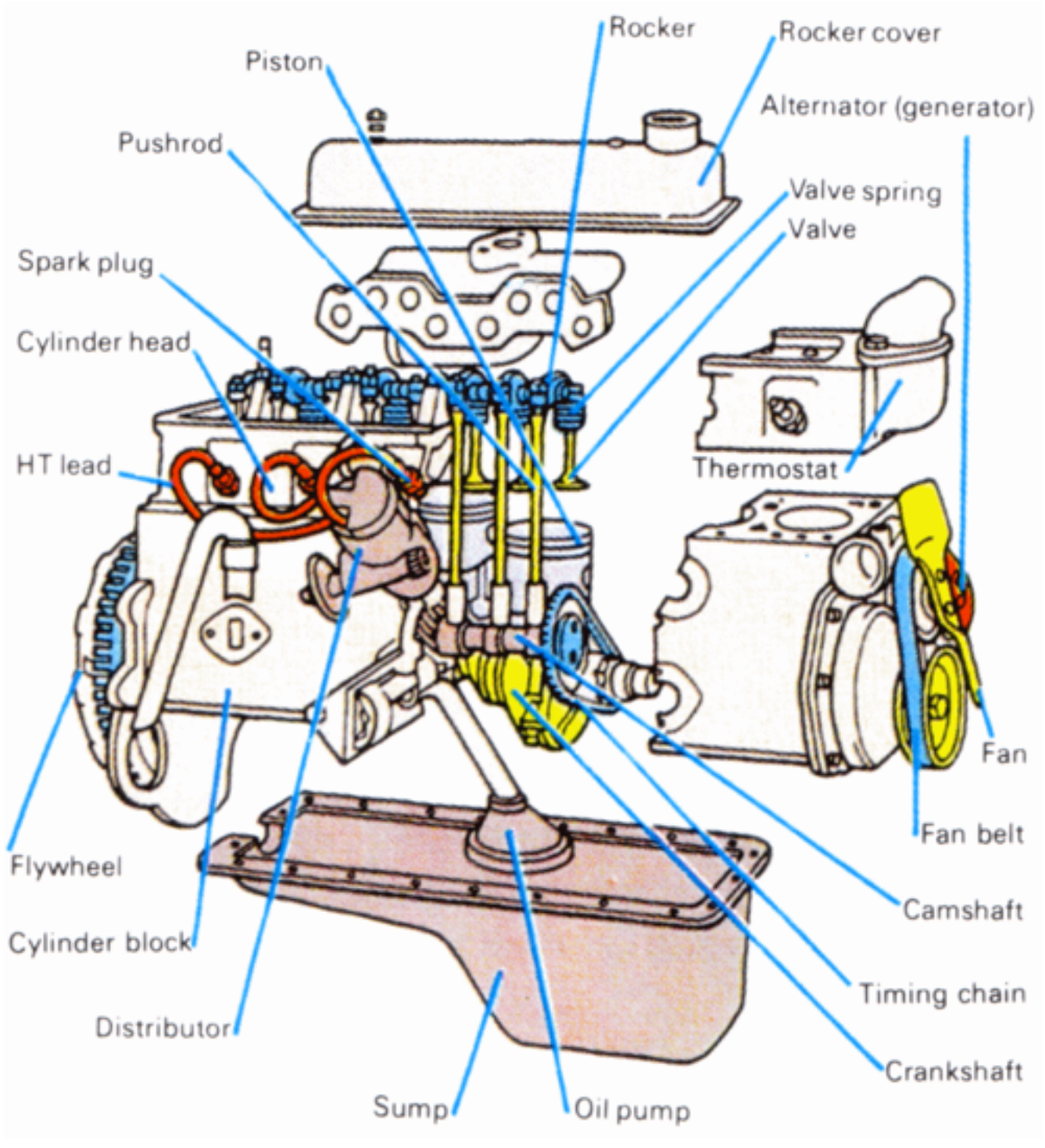
Parts and pieces

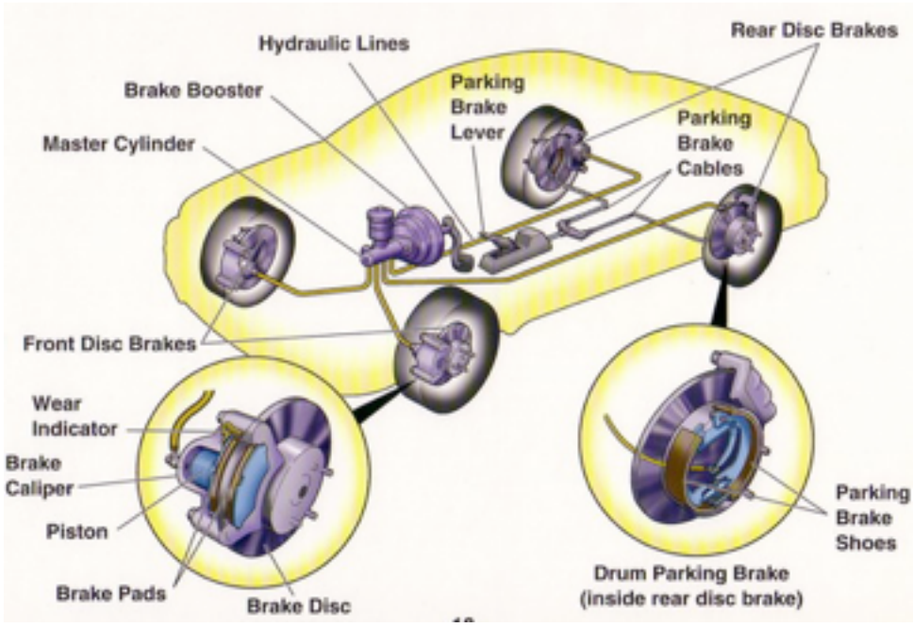

Cost per Megabase of DNA Sequence
$10,000
$1,000
$100
$10
$1
$0
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11

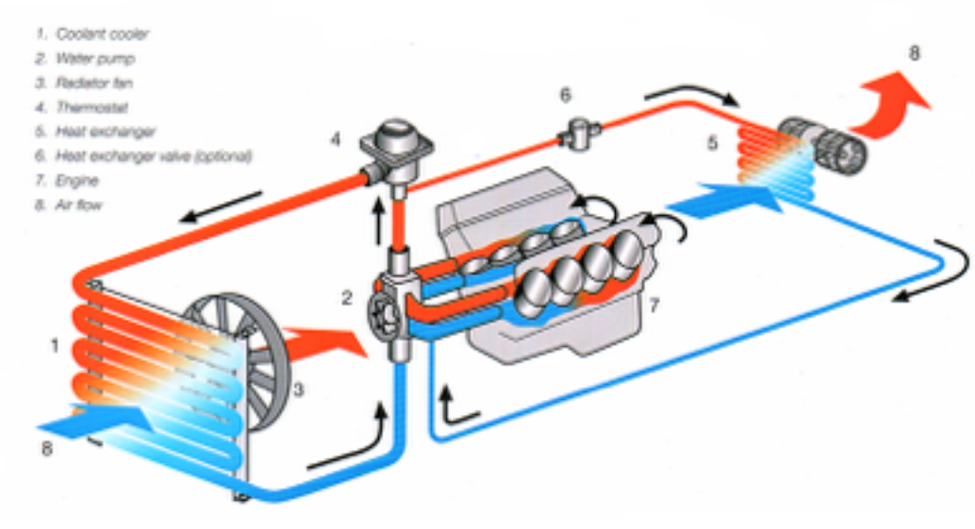
Systems



Sub-systems

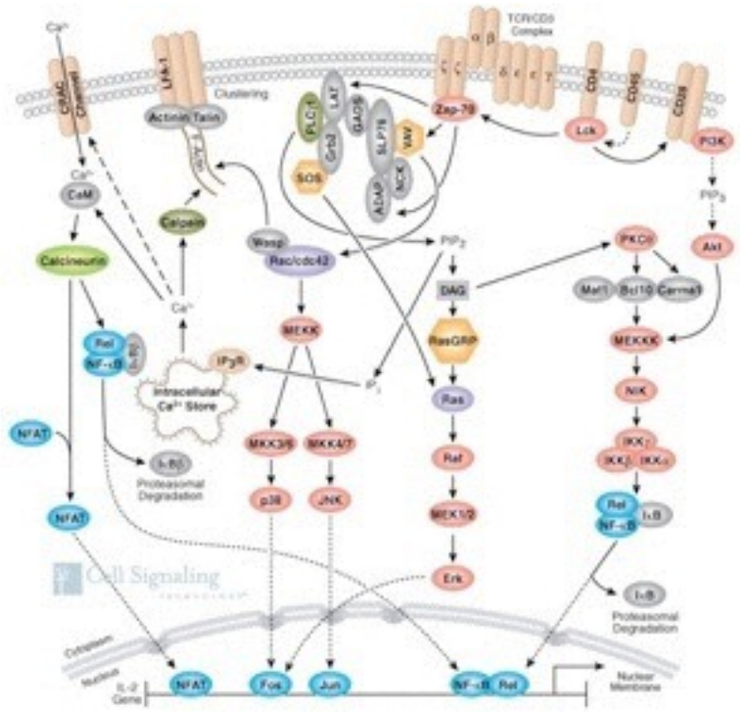

Piston
Rocker
Rocker cover
Alternator (generator)
Pushrod
Valve spring
Valve
Spark plug
Cylinder head
HT lead
Thermostat
Fan
Fan belt
Flywheel
Camshaft
Cylinder block
Timing chain
Distributor
Crankshaft
Sump
Oil pump


Hydraulic Lines
Rear Disc Brakes
Brake Booster
Parking Brake Lever
Master Cylinder
Parking Brake Cables
Front Disc Brakes
Wear Indicator
Brake Caliper
Piston
Parking Brake Shoes
Drum Parking Brake (inside rear disc brake)
Brake Pads
Brake Disc

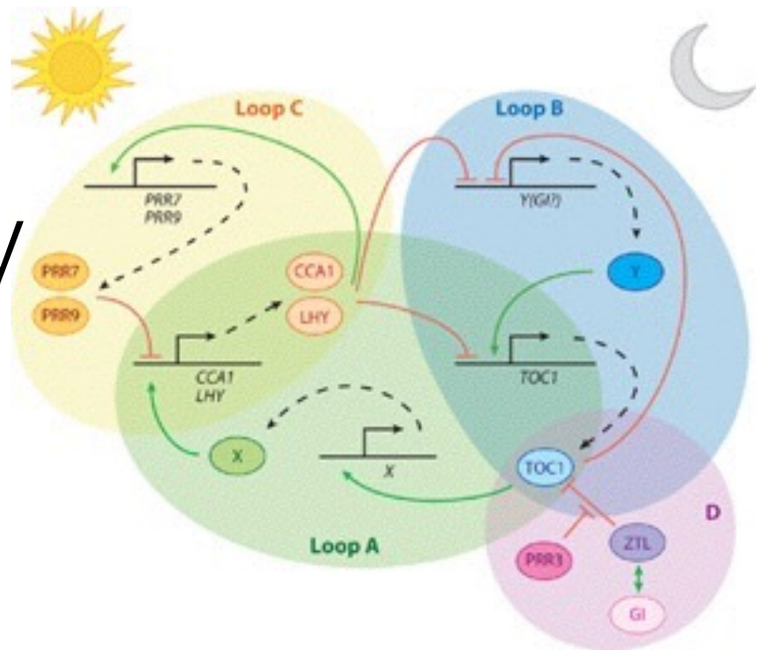

1. Coolant cooler
2. Water pump
3. Radiator fan
4. Thermostat
5. Heat exchanger
6. Heat exchanger valve (optional)
7. Engine
8. Air flow

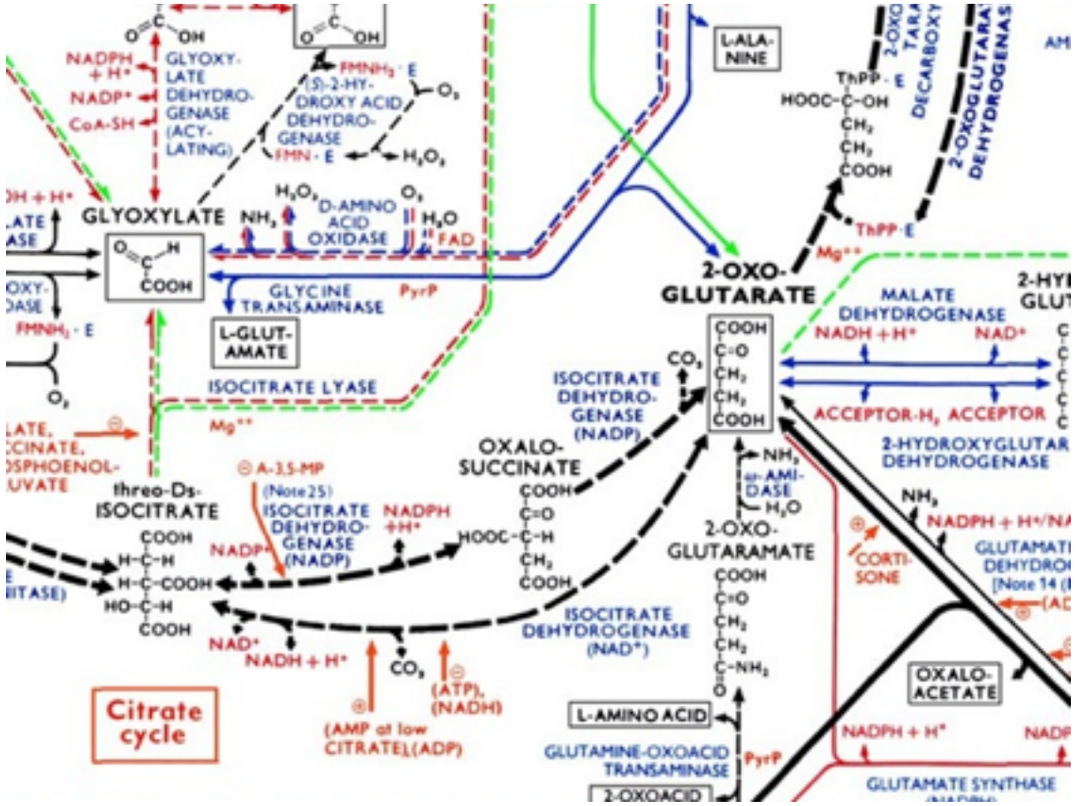
Systemic view of a Cell

Signaling networks



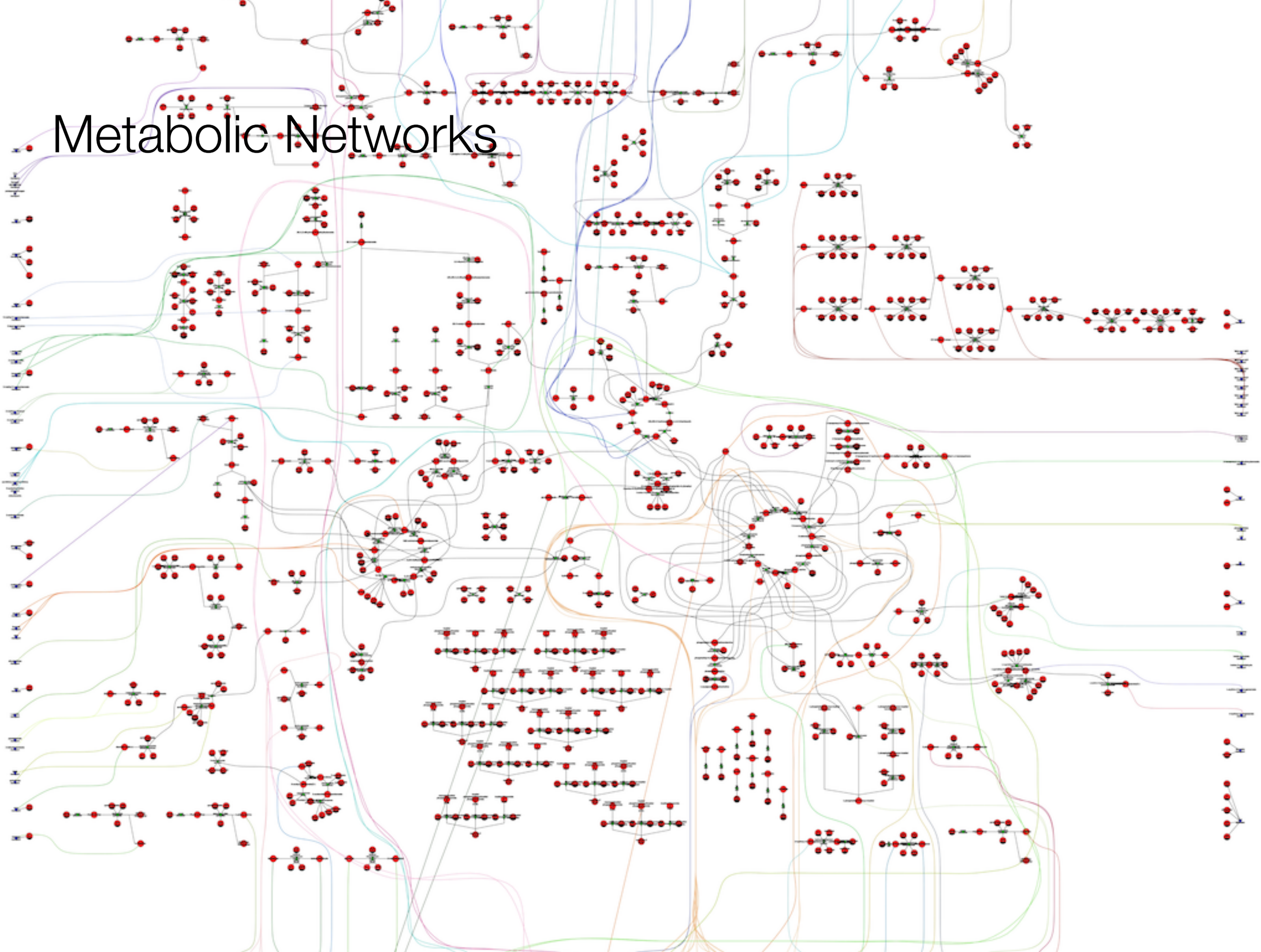
Regulatory networks



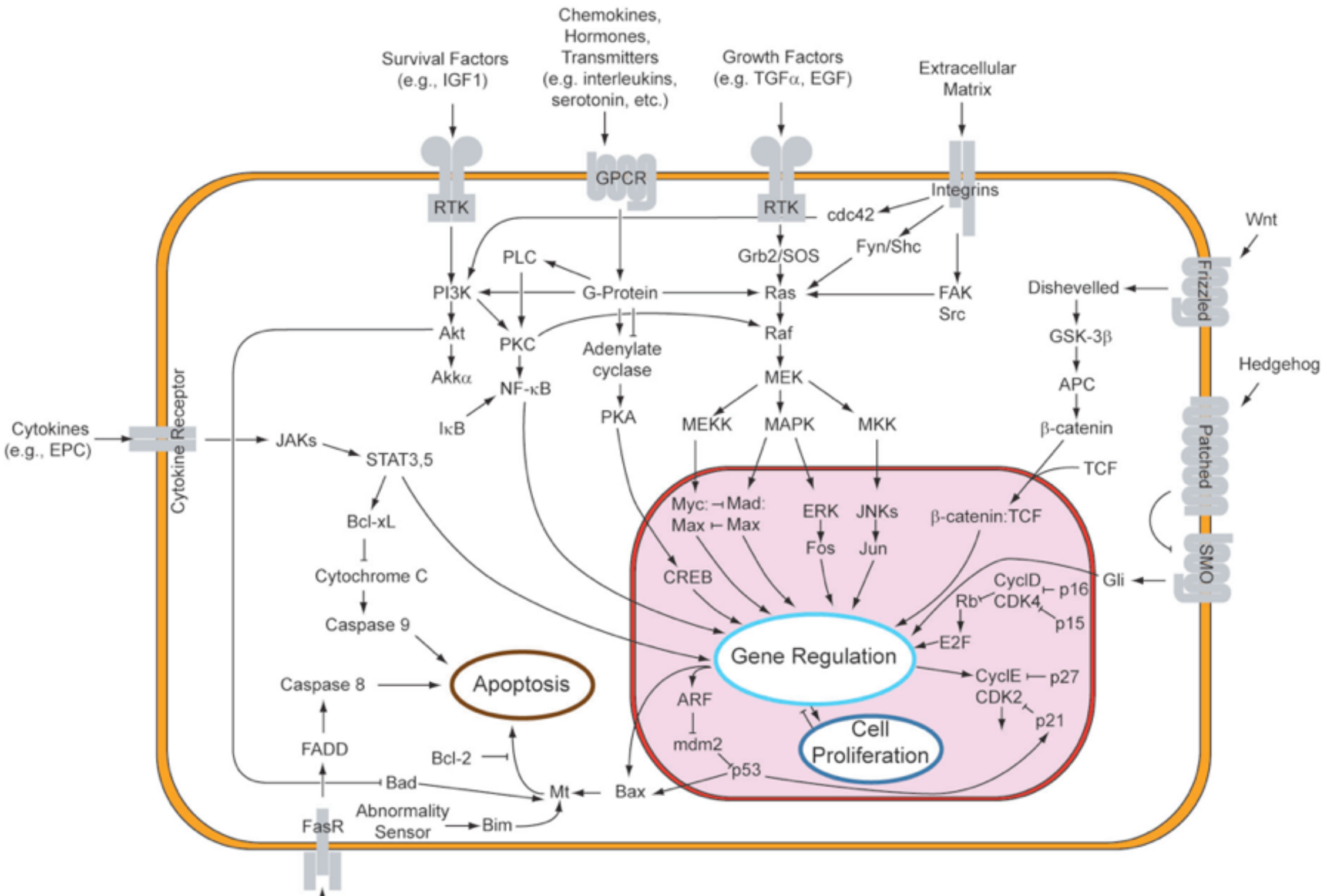


Metabolic networks

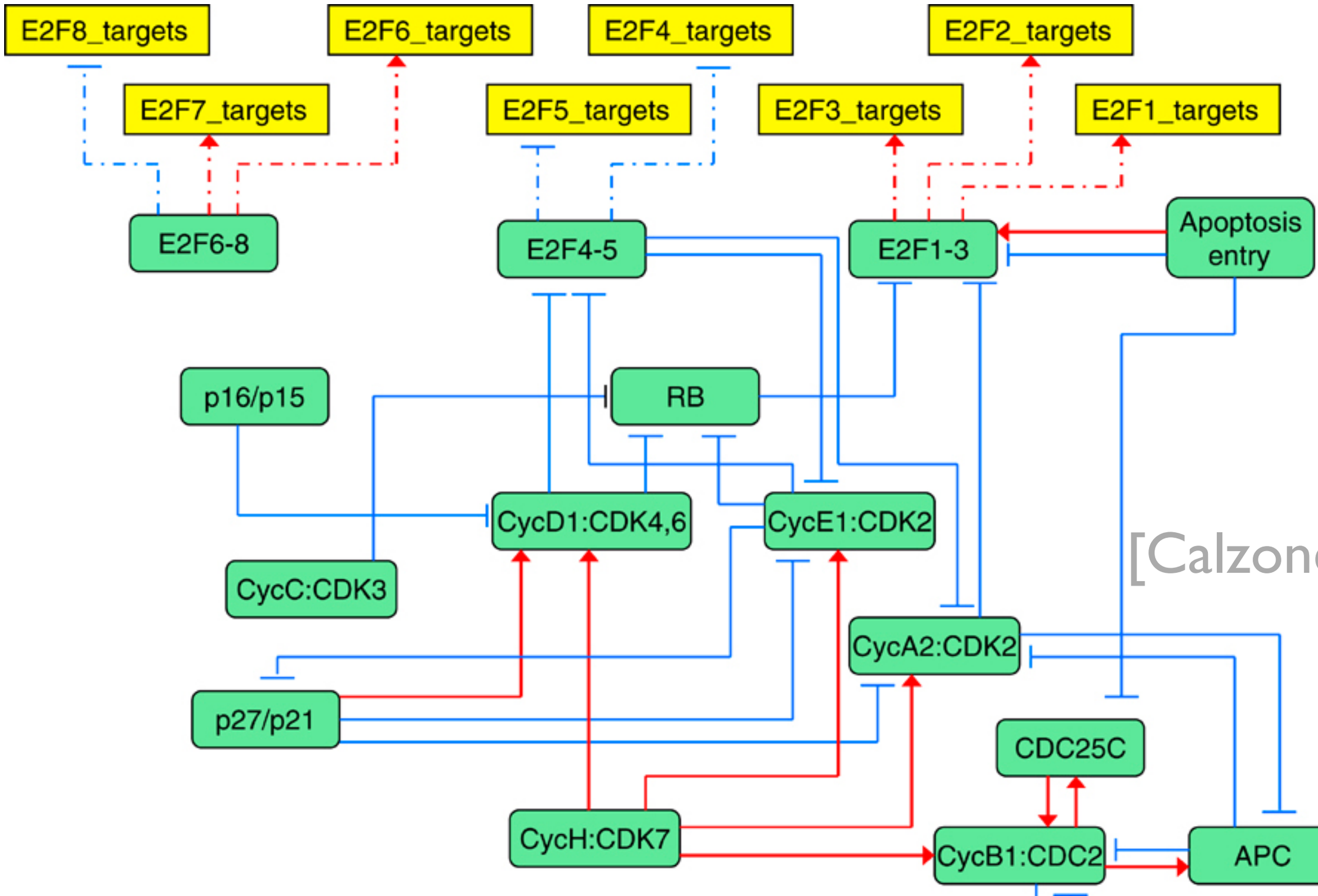
Metabolic Networks



Signaling networks

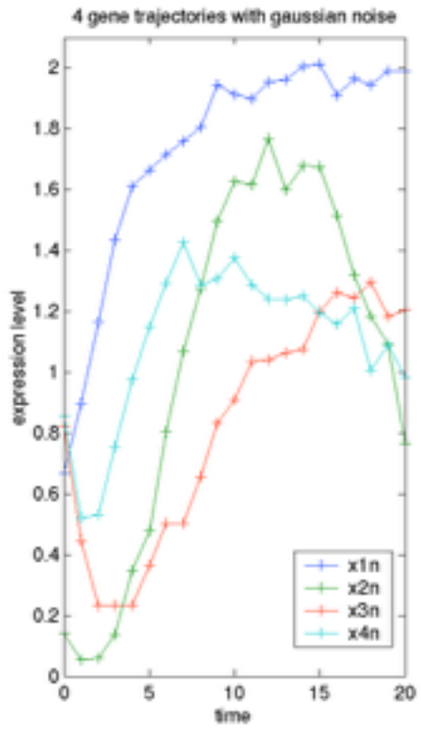


Regulatory networks



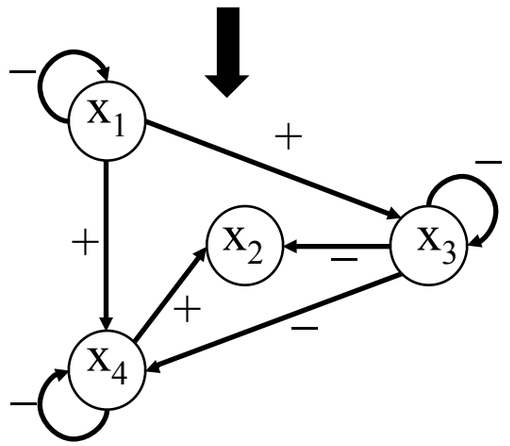
[Calzone, 2008]

Network inference

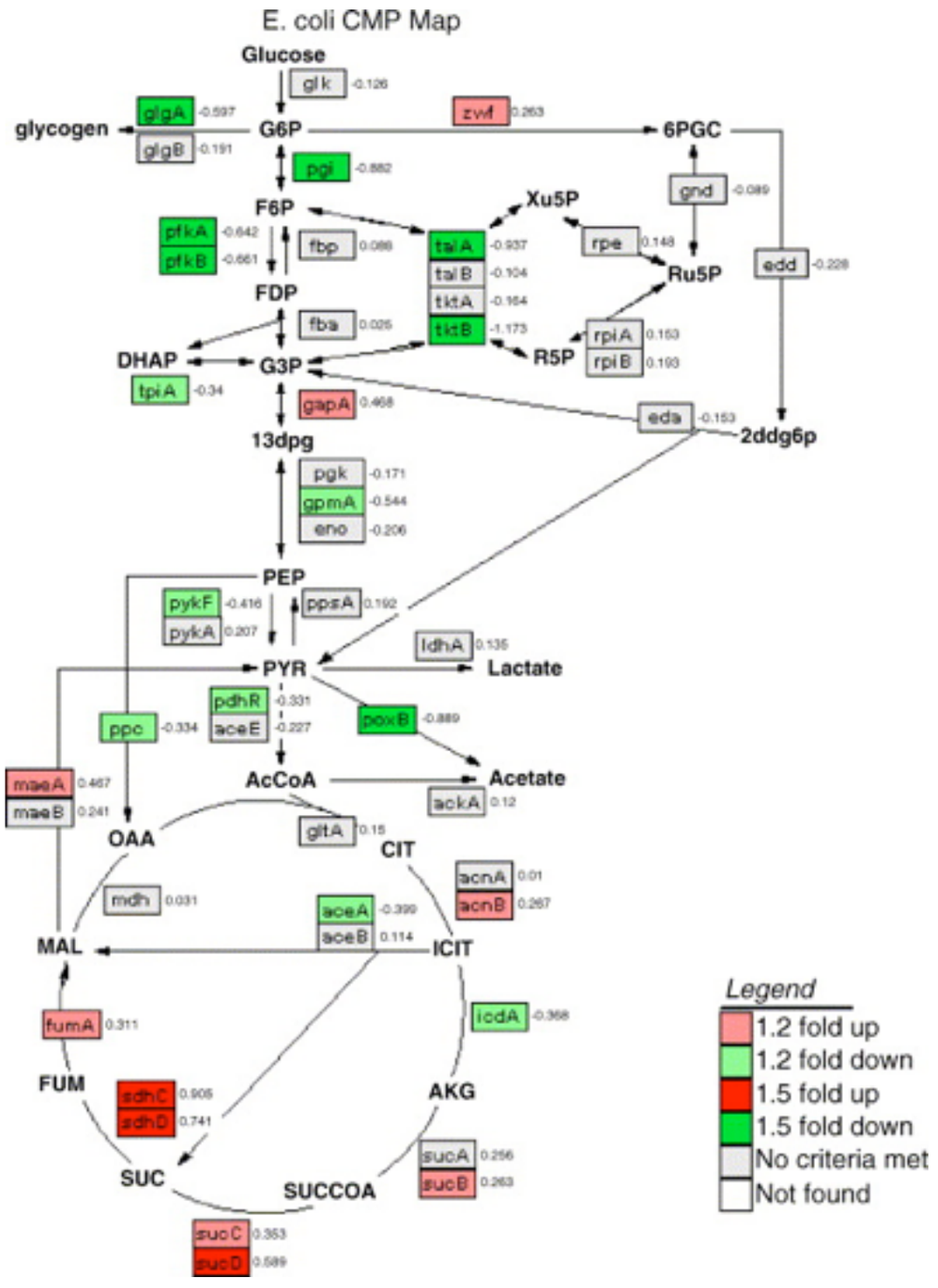


regression procedure

	$a_{0,i}$	$a_{1,i}$	$a_{2,i}$	$a_{3,i}$	$a_{4,i}$
x_1	.431	-.248	0	0	0
x_2	0	0	0	-.473	.374
x_3	-.427	.376	0	-.241	0
x_4	0	.435	0	-.315	-.437



Scientific & Biotech Applications



1. Contextualization of data

2. Guidance of metabolic engineering

3. Directing hypothesis-driven discovery

4. Discovery of multi-species relationships

5. Network property discovery

Bioinfo. sub-fields

-- Sequence analysis

-- Genome annotation

-- Computational evolutionary biology

-- Literature analysis

-- Analysis of gene expression

-- Analysis of regulation

-- Analysis of protein expression

-- Analysis of molecules (metabolomics)

-- Comparative genomics

-- Modeling biological systems

-- High-throughput image analysis

-- Structural Bioinformatics

Tools from Math, CS, IT

---- algorithms, computational biology
---- databases & information system
---- web, web services
---- software engineering
---- HPC (High Performance Computing)
---- data mining
---- image processing
---- modeling and simulation
---- discrete mathematics (eg: graphs, logic)
---- control and system theory
---- statistics

- engineering vs. reverse engineering
- biology: reverse engineering
- systems ⟷ knowledge

- Reverse engineering is the process of discovering the technological principles of a device, object, or system through analysis of its structure, function, and operation.

- Engineering is the discipline, art, skill and profession of acquiring and applying scientific, mathematical,economic, social, and practical knowledge, in order to design and build structures, machines, devices, systems, materials and processes

Biology and models

-- biology: reverse engineering

systems ⟶ models

-- text, symbols, standards, language, math

-- models: static vs dynamic

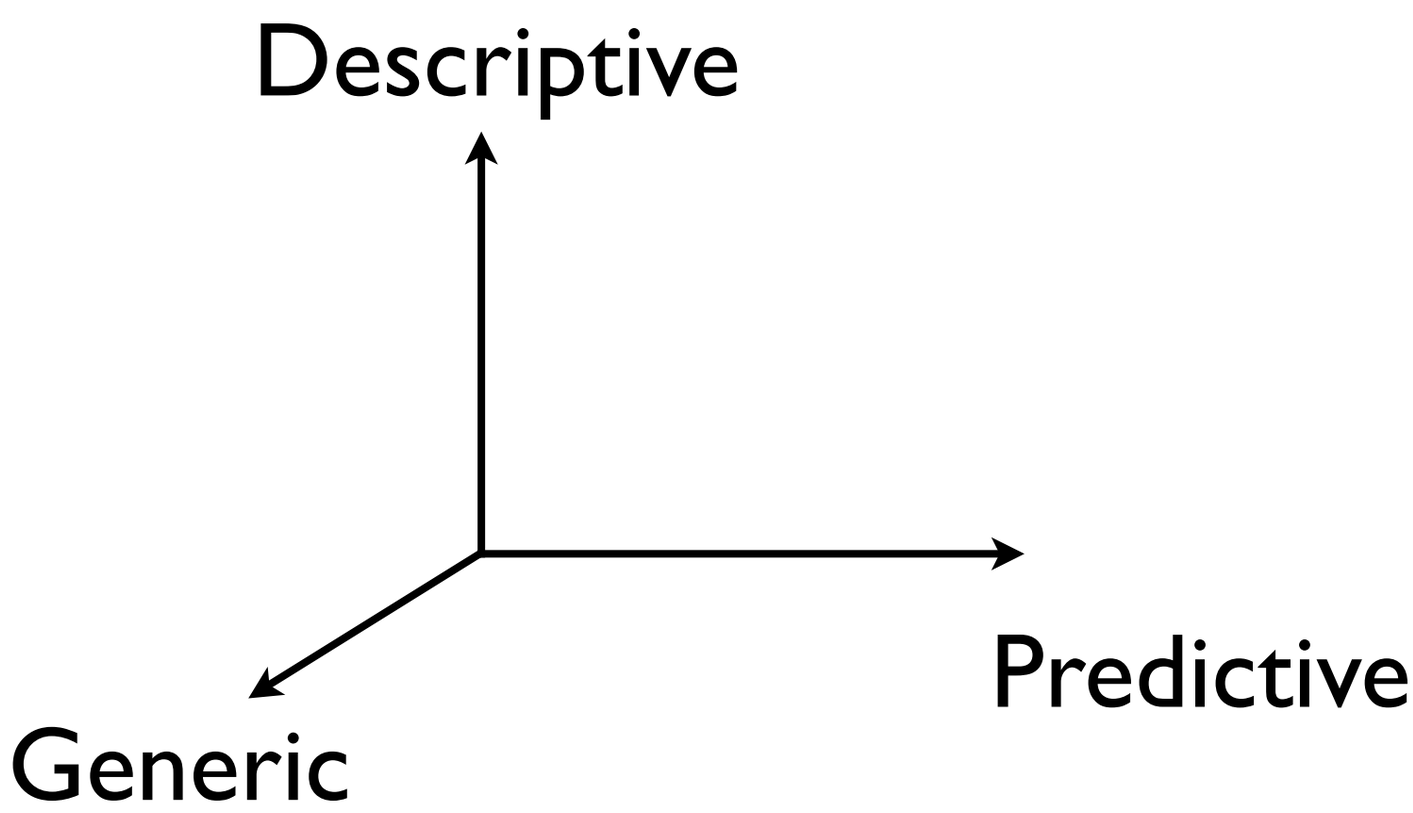
---- static: data and their relationships

---- dynamic: fluxes and multi-agent

-- genomics

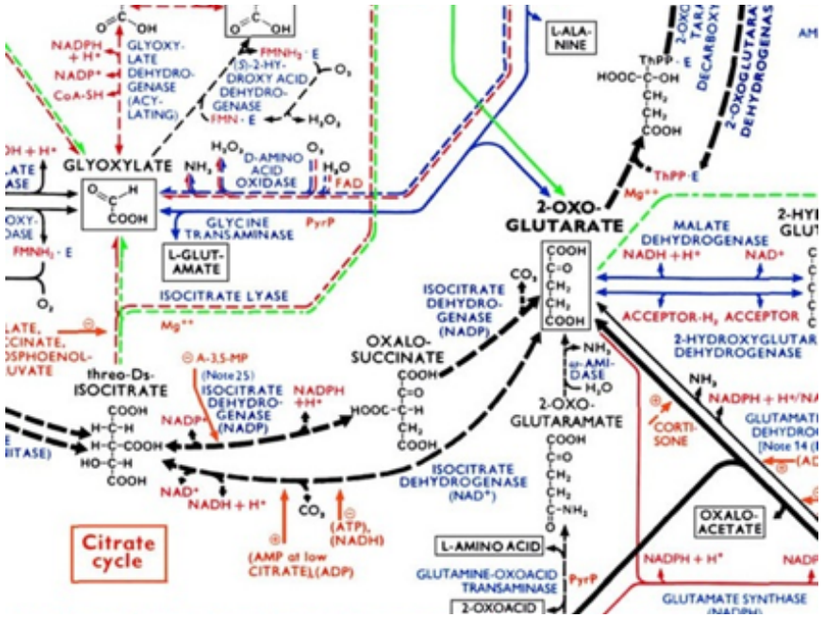
-- prokaryote vs eukaryote

Models



*choose two

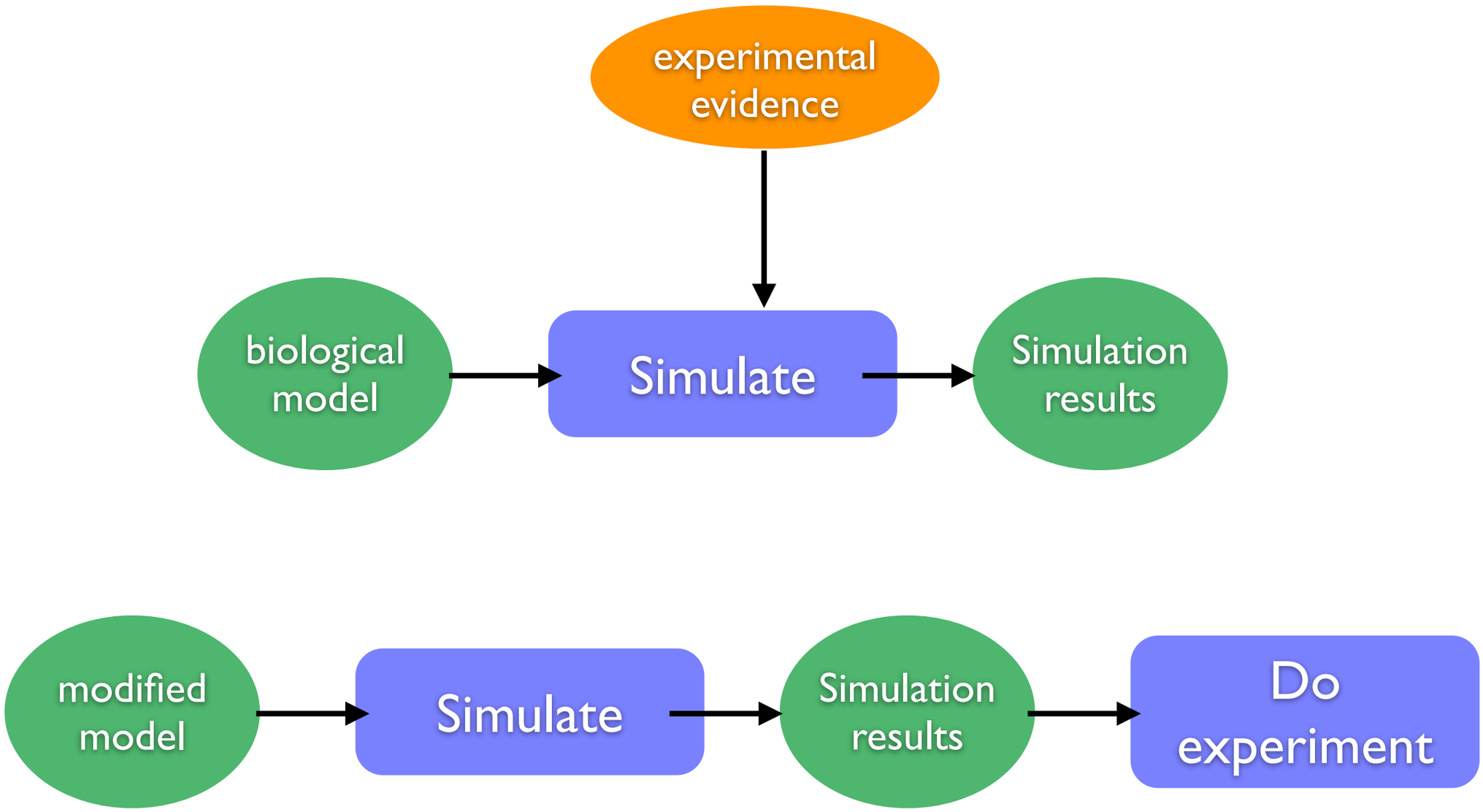
Descriptive Models



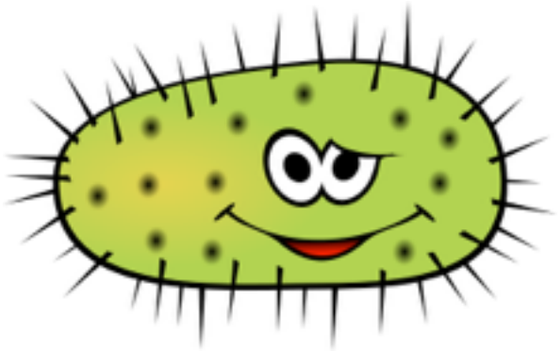
VS

Black Box

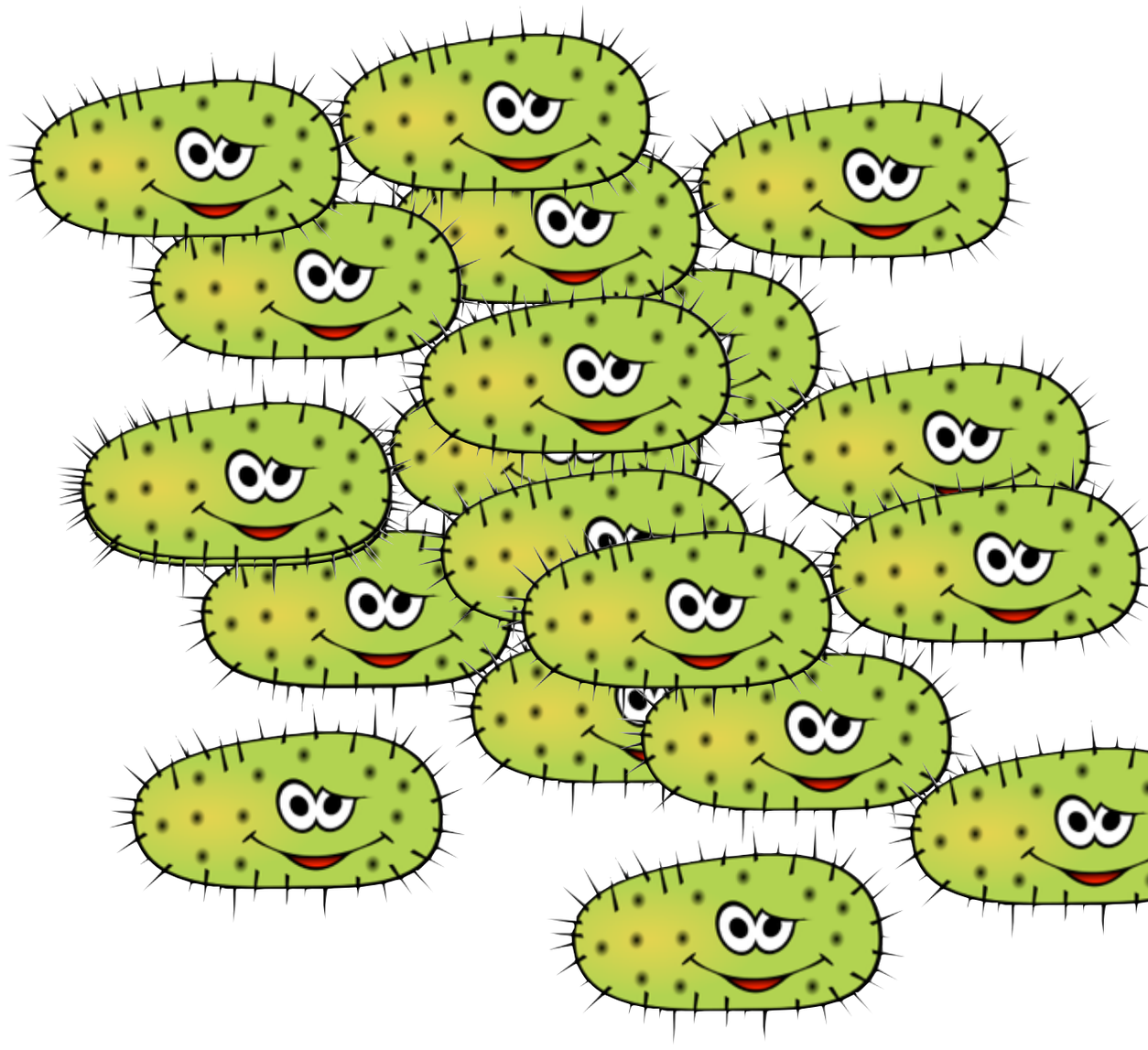
Predictive Models



Generic Models

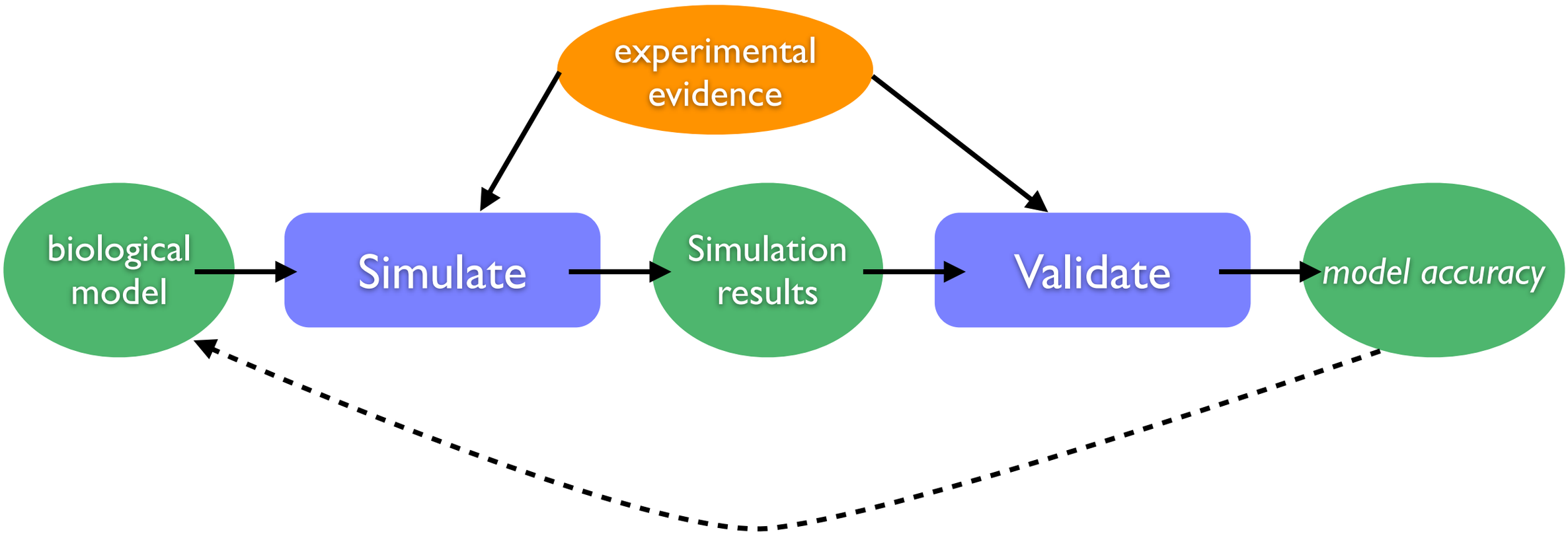


Specific model



Generic model

Iterative model improvement



Metabolic Models

Definition of Metabolic Pathways

- A chemical reaction interconverts chemical compounds (analogous to a production rule)

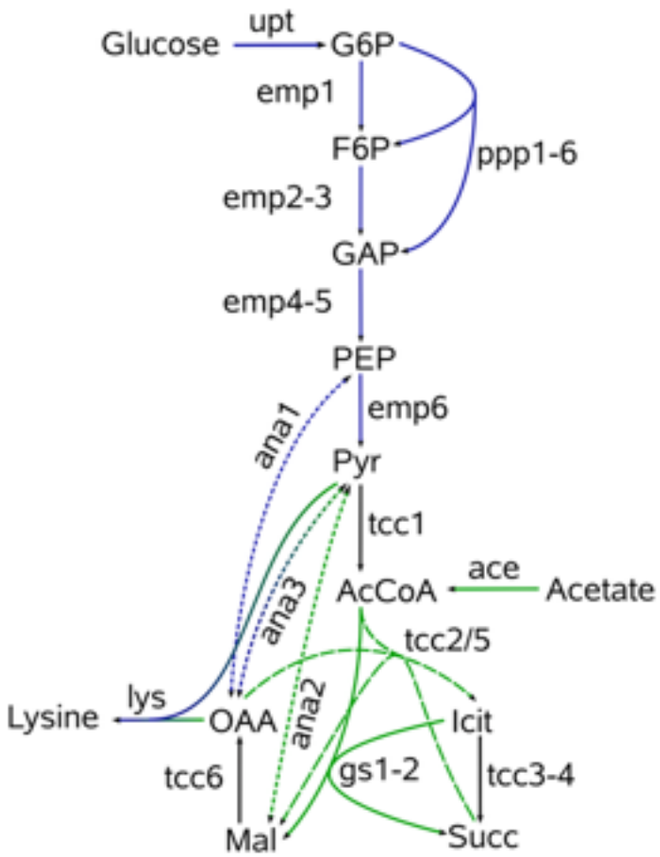
$$A + B = C + D$$

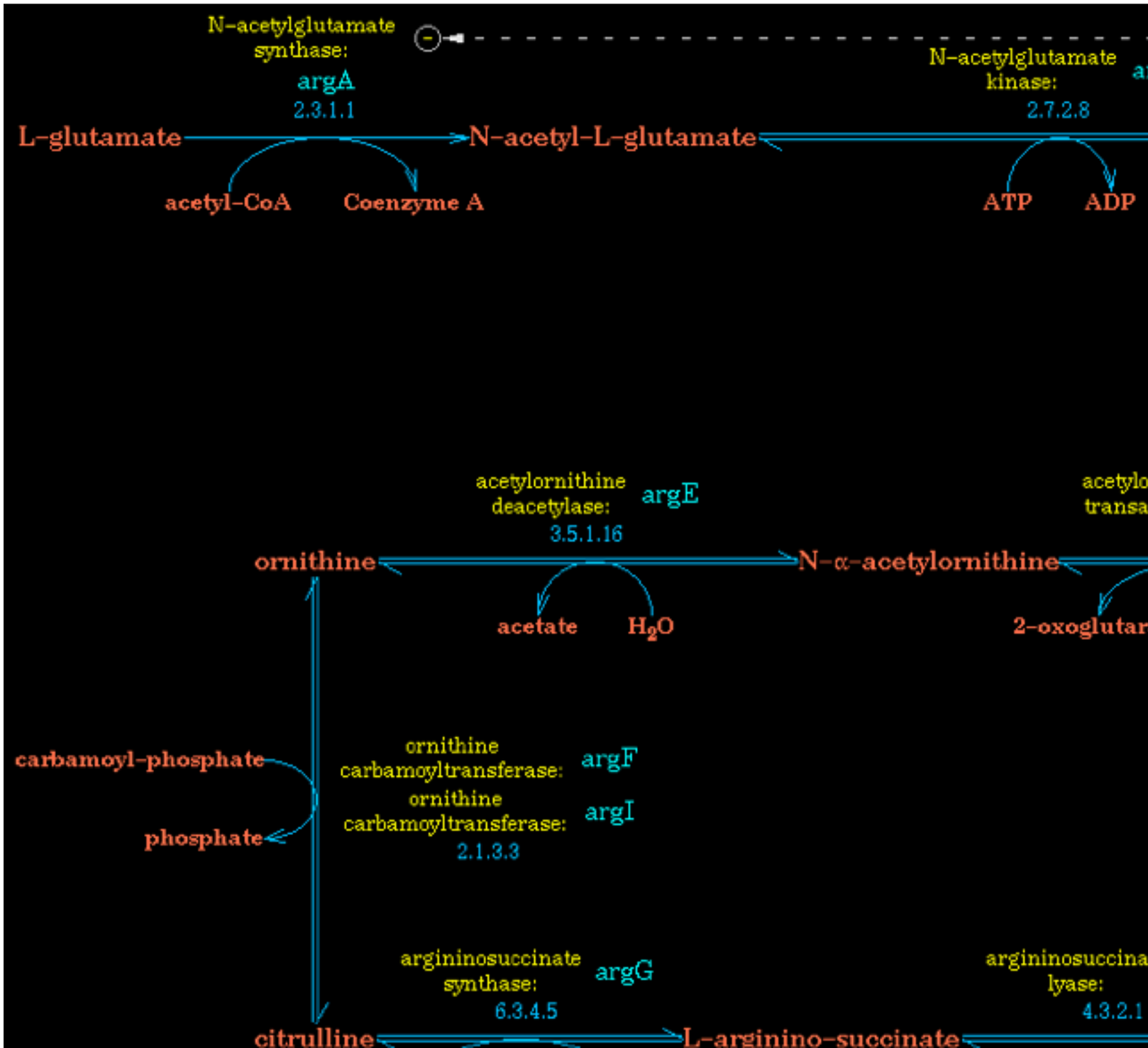
- An enzyme is a protein that accelerates chemical reactions. Each enzyme is encoded by one or more genes.

- A pathway is a linked set of reactions (analogous to a chain of rules)

$$A \longrightarrow C \longrightarrow E$$

Pathways

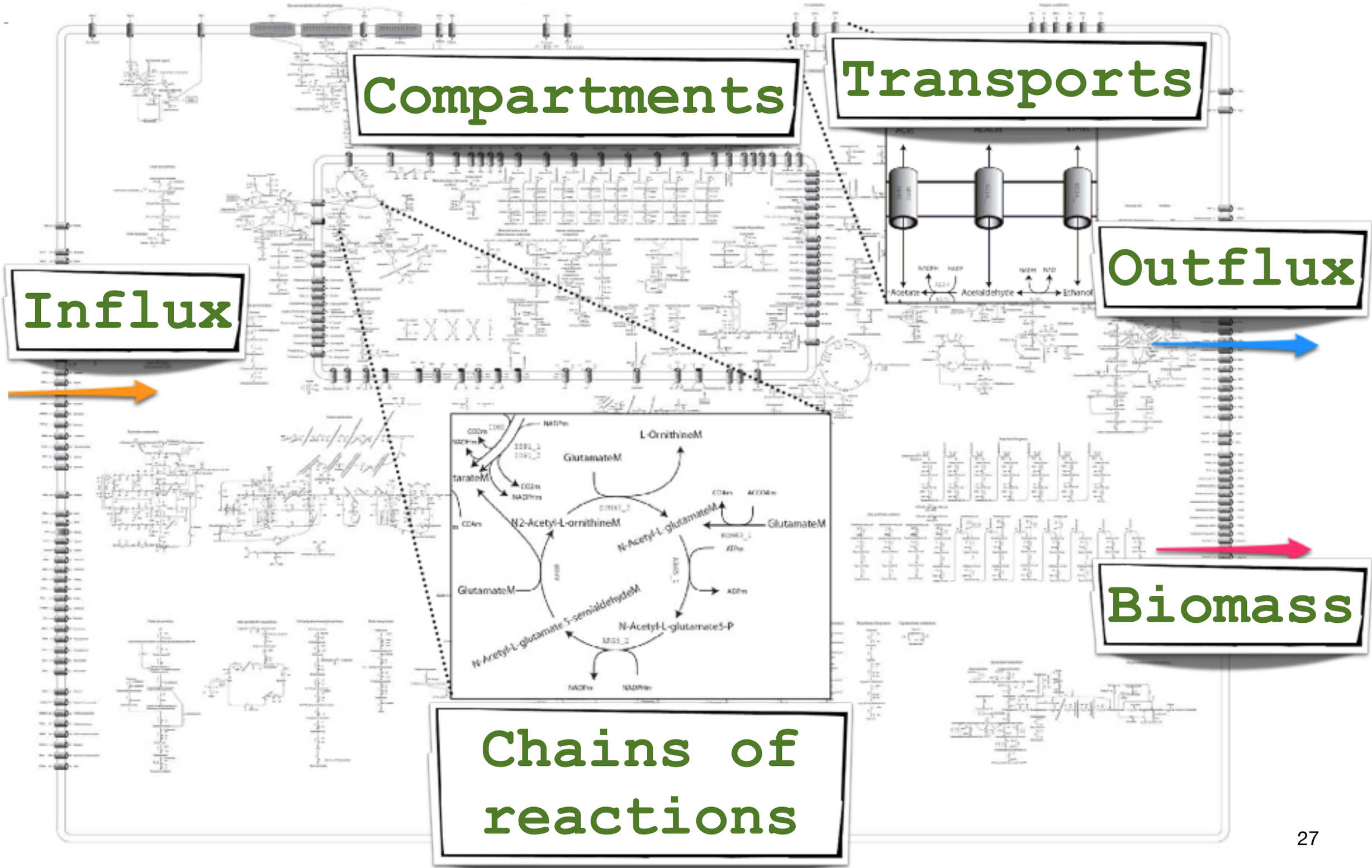

Glucose
upt
G6P
emp1
F6P
ppp1-6
emp2-3
GAP
emp4-5
PEP
emp6
ana1
Pyr
tcc1
ana3
AcCoA
ace
Acetate
tcc2/5
Lysine
lys
OAA
ana2
Icit
tcc6
gs1-2
tcc3-4
Mal
Succ


N-acetylglutamate synthase:
argA
2.3.1.1
N-acetylglutamate kinase:
2.7.2.8
L-glutamate
N-acetyl-L-glutamate
acetyl-CoA
Coenzyme A
ATP
ADP
acetylornithine deacetylase:
3.5.1.16
argE
ornithine
N-α-acetylornithine
acetate
H2O
carbamoyl-phosphate
phosphate
ornithine carbamoyltransferase:
argF
ornithine carbamoyltransferase:
2.1.3.3
argI
argininosuccinate synthase:
6.3.4.5
argG
citrulline
L-arginino-succinate

What is a Metabolic Pathway?

- A pathway is a conceptual unit of the metabolism
- An ordered set of interconnected, directed biochemical reactions
- A pathway forms a coherent unit:
  - Boundaries defined at high-connectivity substrates
  - Regulated as a single unit
  - Evolutionarily conserved across organisms as a single unit
  - Performs a single cellular function
  - Historically grouped together as a unit
  - All reactions in a single organism

Genome-scale Metabolic Networks



Elements of Met.Networks

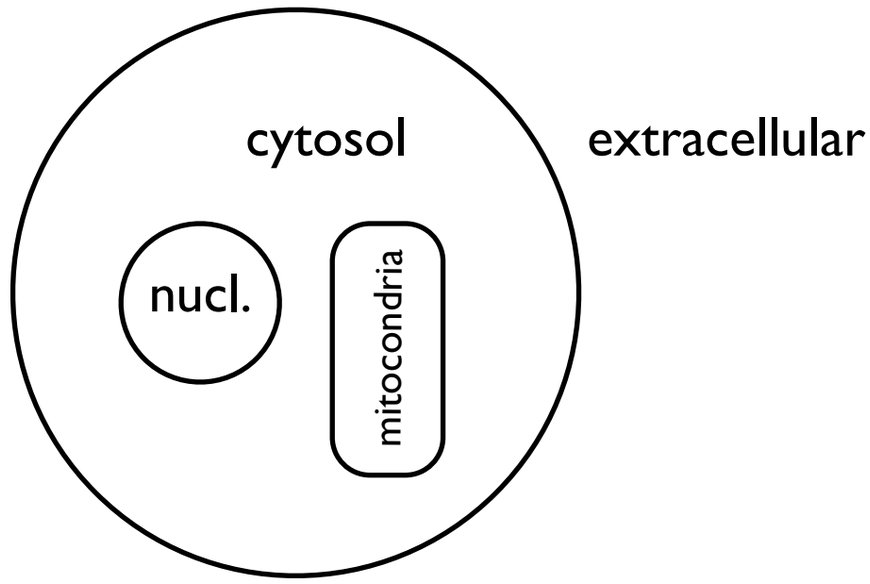
a Chemical Reactions

	GLCt1	HEX1	PGI	PFK	FBP	FBA	TPI	EX_glc
glc-D[e]	−1	0	0	0	0	0	0	−1
glc-D	1	−1	0	0	0	0	0	0
atp	0	−1	0	−1	0	0	0	0
H	0	1	0	1	0	0	0	0
adp	0	1	0	1	0	0	0	0
g6p	0	1	−1	0	0	0	0	0
f6p	0	0	1	−1	1	0	0	0

= **S**

b Genes->Proteins->Reactions

c



Stoichiometry is the measuring of metabolites in a chemical reaction

Reaction2: $2\ M_1 + 3\ M_2 \Rightarrow 1\ M_3 + 4\ M_4$

(-2 -3 +1 +4)

A network of reactions can be described with an Stoichiometric Matrix

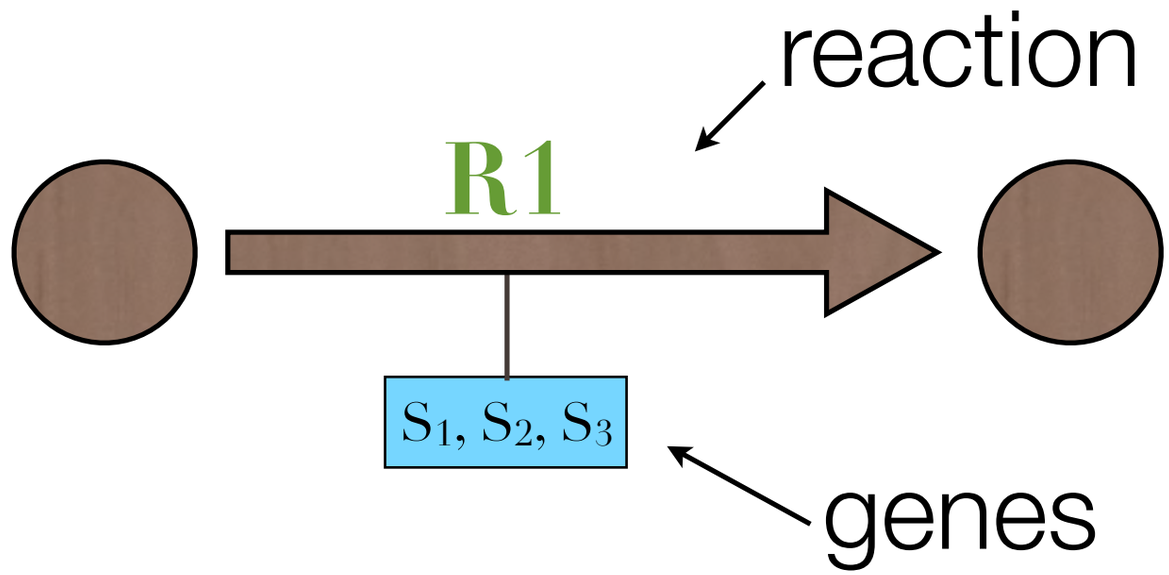
$$S = \begin{array}{c} \\ M_1 \\ M_2 \\ M_3 \\ M_4 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & R_3 & R_4 & R_5 \end{array} \\ \begin{bmatrix} & -2 & & & \\ & -3 & & & \\ & +1 & & & \\ & +4 & & & \end{bmatrix} \end{array}$$

	GLCt1	HEX1	PGI	PFK	FBP	FBA	TPI	EX_glc	
glc-D[e]	−1	0	0	0	0	0	0	−1	
glc-D	1	−1	0	0	0	0	0	0	
atp	0	−1	0	−1	0	0	0	0	
H	0	1	0	1	0	0	0	0	
adp	0	1	0	1	0	0	0	0	
g6p	0	1	−1	0	0	0	0	0	= S
f6p	0	0	1	−1	1	0	0	0	

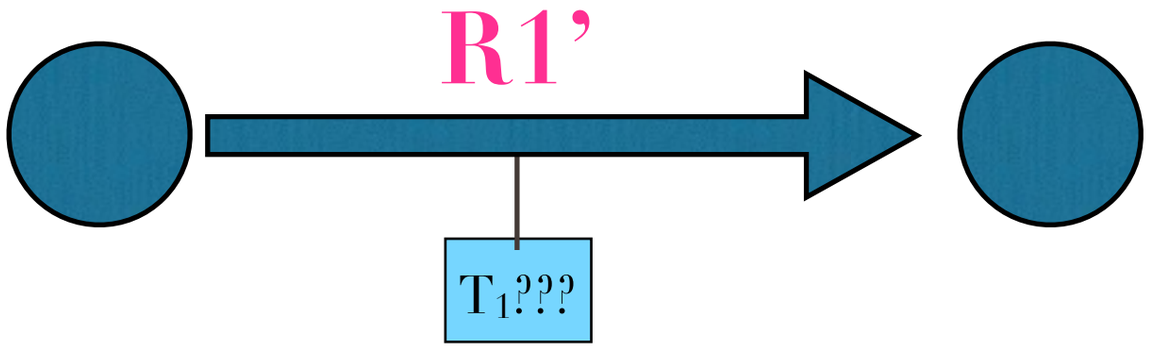
Systems Biology: Properties of Reconstructed Networks
(Palsson, 2006)

Instantiation of a Reaction

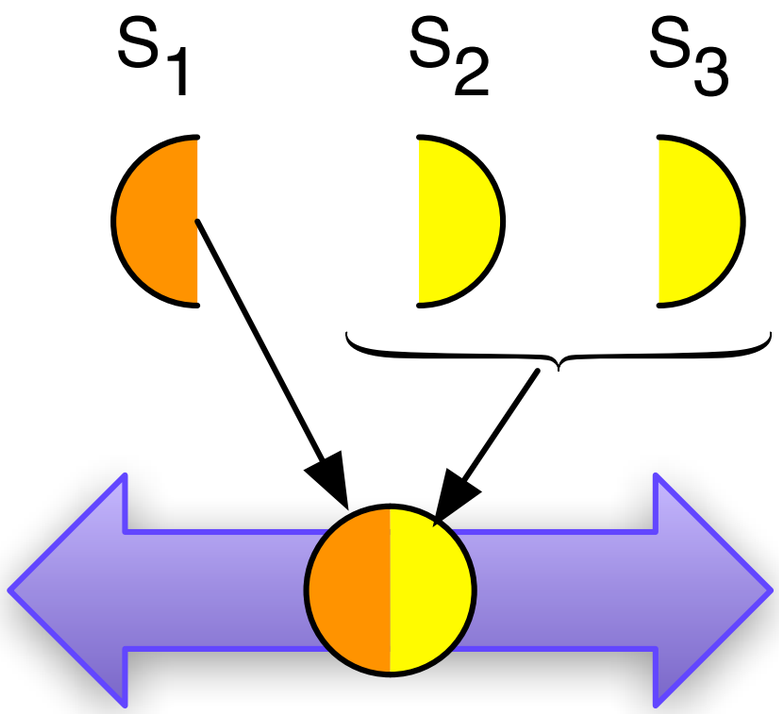
Organism1
(Source)



Organism2
(Target)

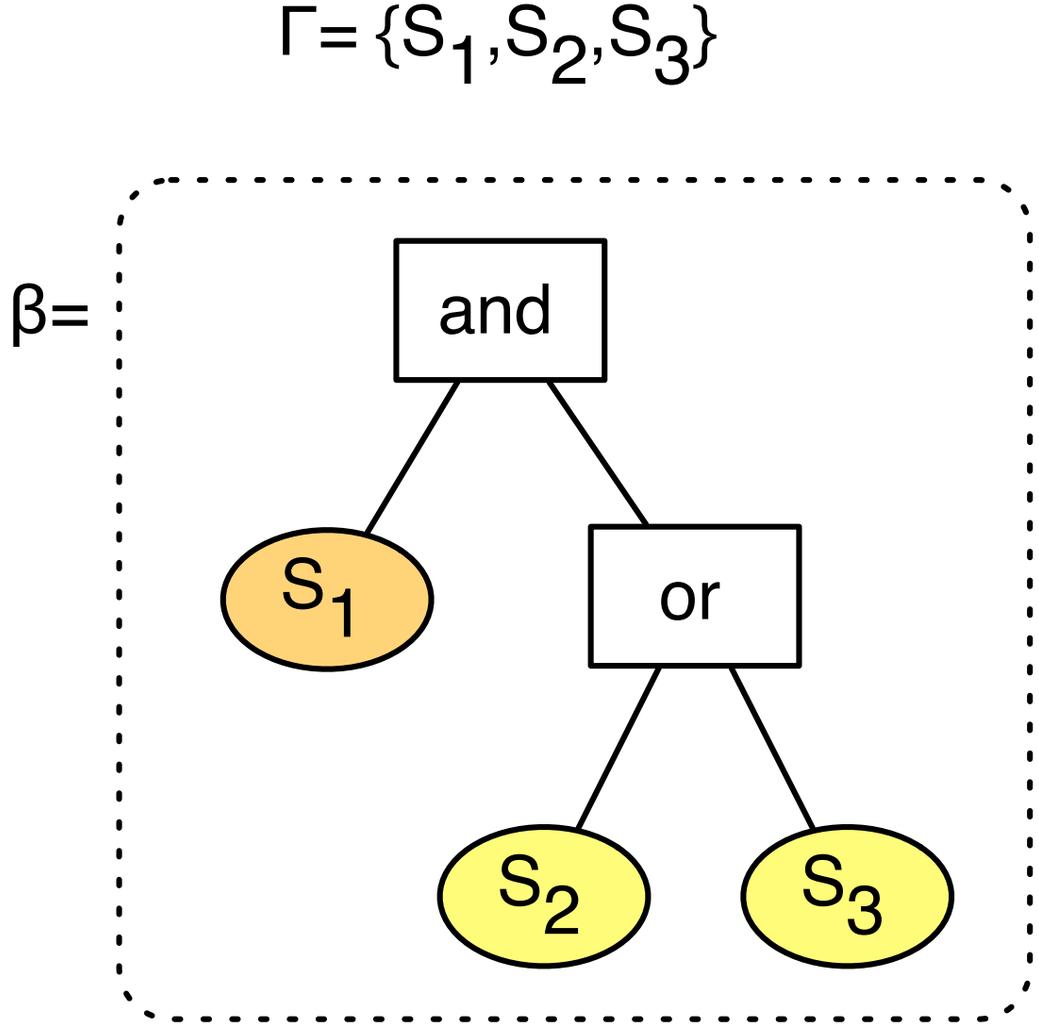


Gene association

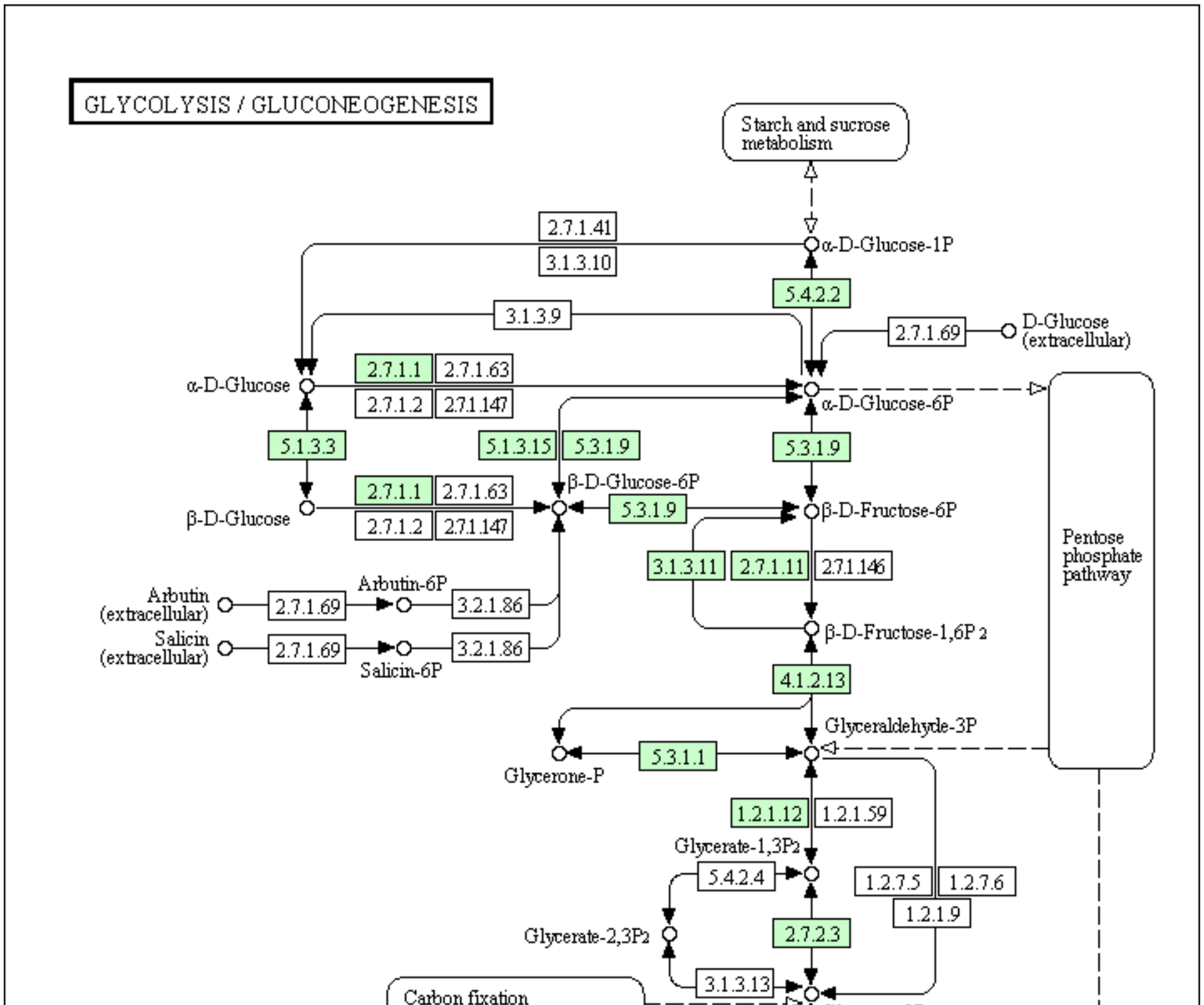


Reaction R

Gene association: S_1 and (S_2 or S_3)

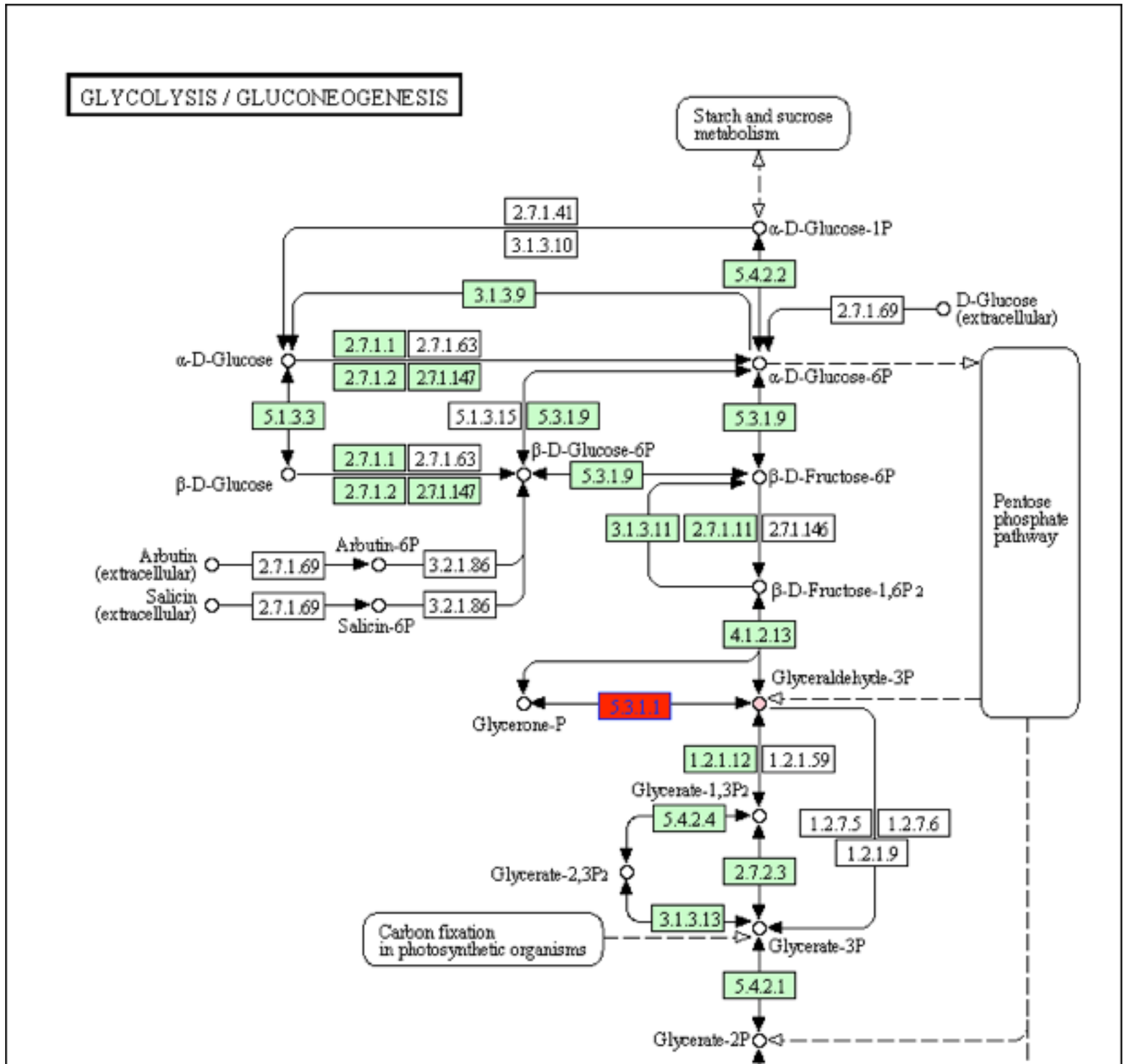


KEGG http://www.genome.jp/kegg/



KEGG Color Mapper

http://www.genome.jp/kegg/tool/map_pathway2.html



Current metabolic models of *S.cerevisiae*

Model ID	**Publication**	**Genes**	**Reactions**	**Metabo-lites**	**Compart-ments**
iFF708	(Förster, 03)	708	1,175	825	4
iND750	(Duarte, 04)	750	1,489	972	8
iLL672	(Kuepfer, 05)	672	1,038	636	3
iIN800	(Nookaew, 08)	800	1,446	1,118	4
iJM832	(Herrgård, 08)	832	1,857	2,152	15
iJM832 no compartments	(Herrgård, 08)	832	1,573	1,748	2
iMM904	(Mo, 09)	904	1,577	1,392	8

SBML

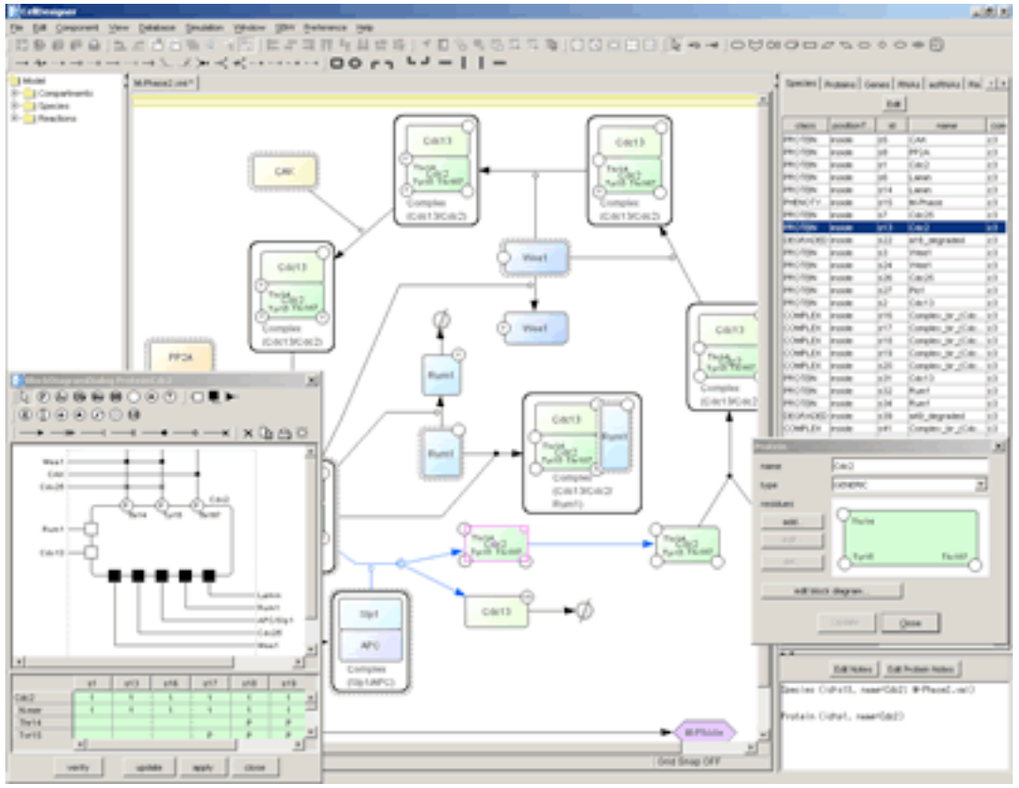
```
    <species compartment="c_02" id="s_5012" name="butyrate [cytoplasm]"/>
    <species compartment="c_14" id="s_5013" name="butyrate [peroxisome]"/>
    <species compartment="c_14" id="s_5014" name="butyryl-CoA [peroxisome]"/>
  </listOfSpecies>
  <listOfParameters>
    <parameter id="dummy_flux" units="flux_unit" value="0"/>
  </listOfParameters>
  <listOfReactions>
    <reaction id="r_0001" name="(R)-lactate:ferricytochrome-c 2-oxidoreductase"
      <notes>
        <html:body>
          <html:p>GENE_ASSOCIATION: (YALI0D09273g and (YALI0E03212g or YALI0C06
        </html:body>
      </notes>
      <listOfReactants>
        <speciesReference species="s_0028"/>
        <speciesReference species="s_0679" stoichiometry="2"/>
      </listOfReactants>
      <listOfProducts>
        <speciesReference species="s_0680" stoichiometry="2"/>
        <speciesReference species="s_1277"/>
      </listOfProducts>
      <kineticLaw>
        <ns6:math>
```

beginning of model definition
- list of function definitions (optional)
- list of unit definitions (optional)
- list of compartment types (optional)
- list of species types (optional)
- list of compartments (optional)
- list of species (optional)
- list of parameters (optional)
- list of initial assignments (optional)
- list of rules (optional)
- list of constraints (optional)
- list of reactions (optional)
- list of events (optional)

end of model definition

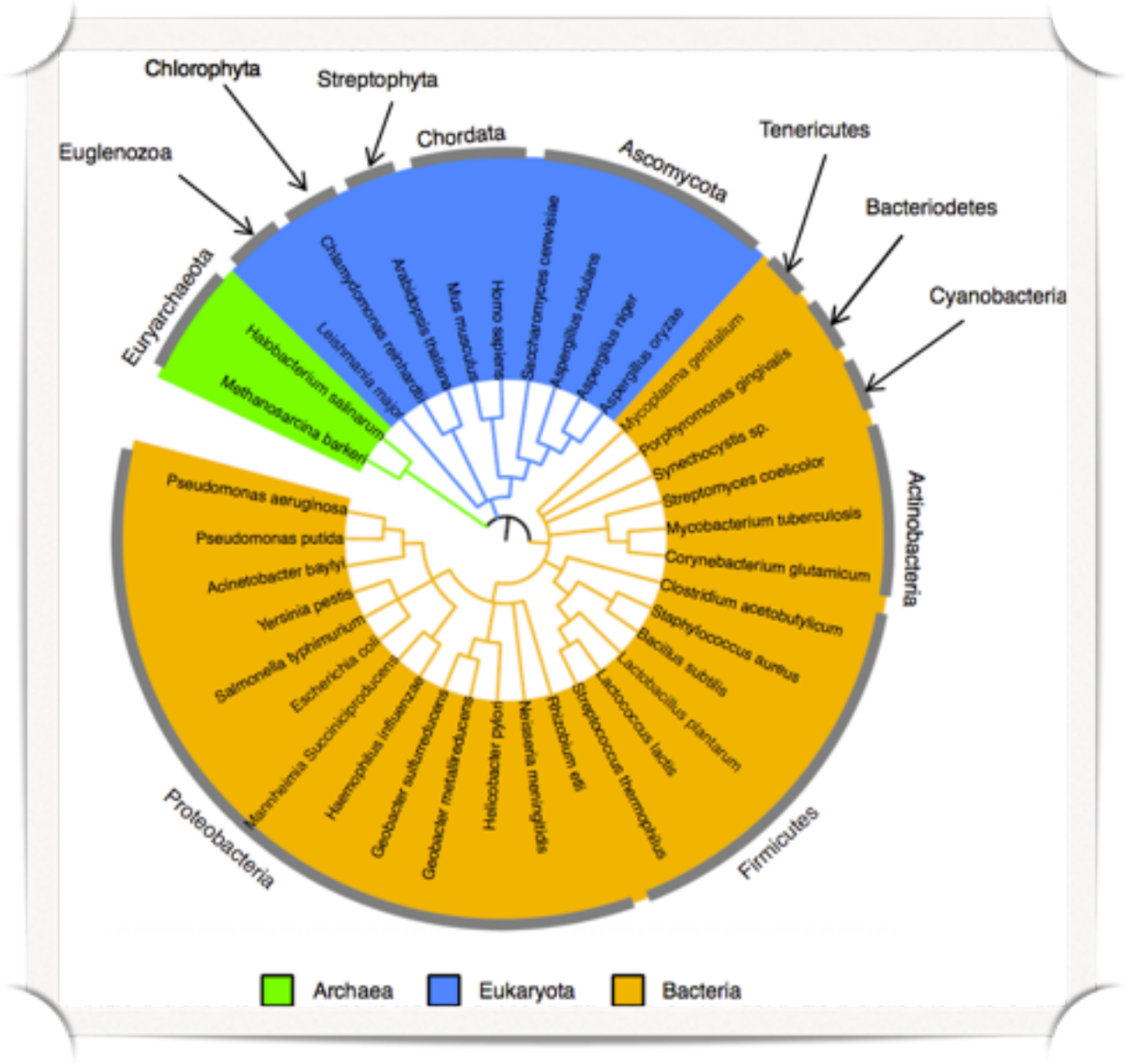
Visualization tools

Cytoscape



CellDesigner

Current reconstructions

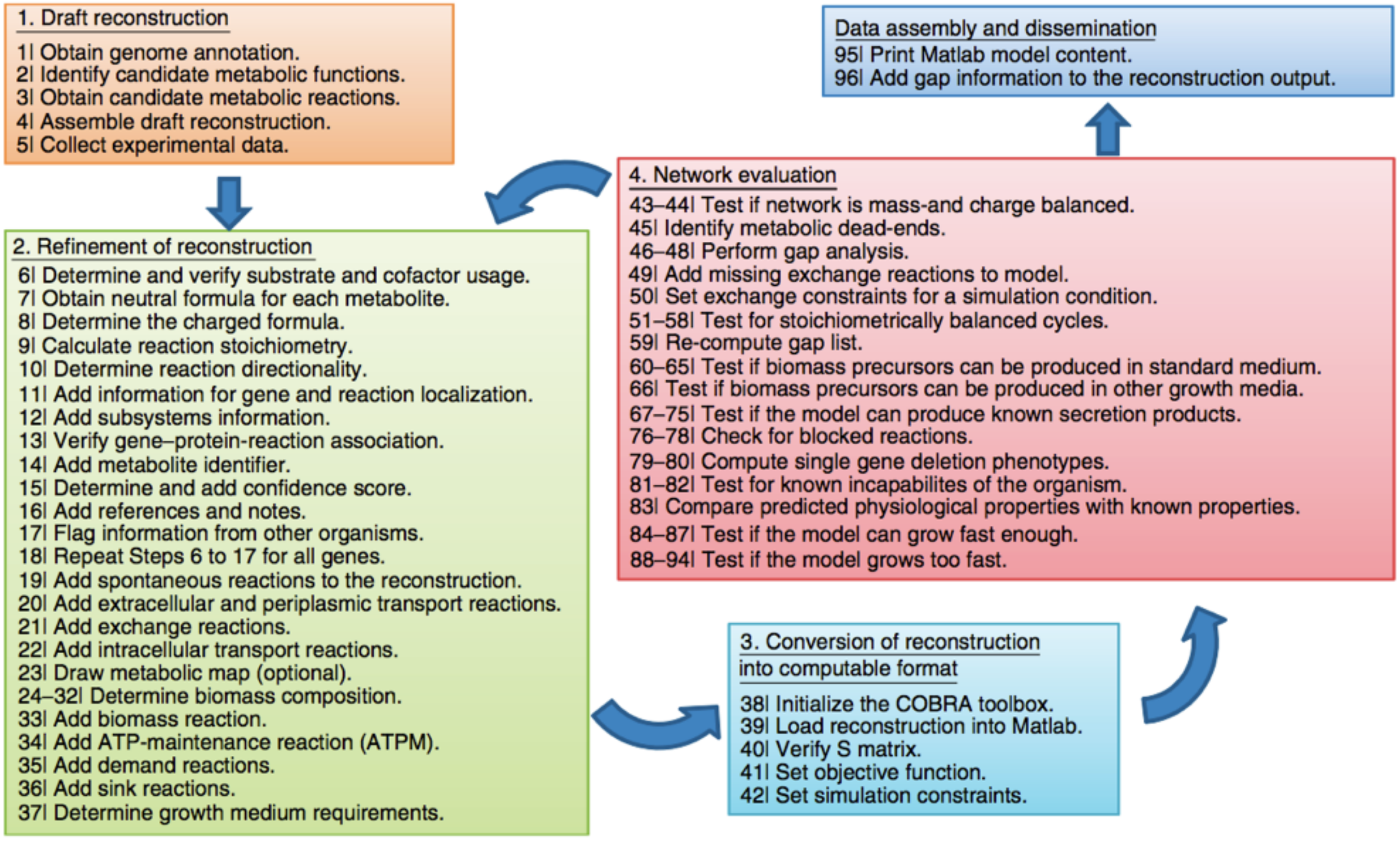


[Oberhardt, 2009]

- Genome-scale models are hard and expensive to build
- Most reconstructions are for bacteria
- There are no tools to correctly reconstruct models for eukaryotes

de novo reconstruction

[Thiele, Nature Protocols, 2010]



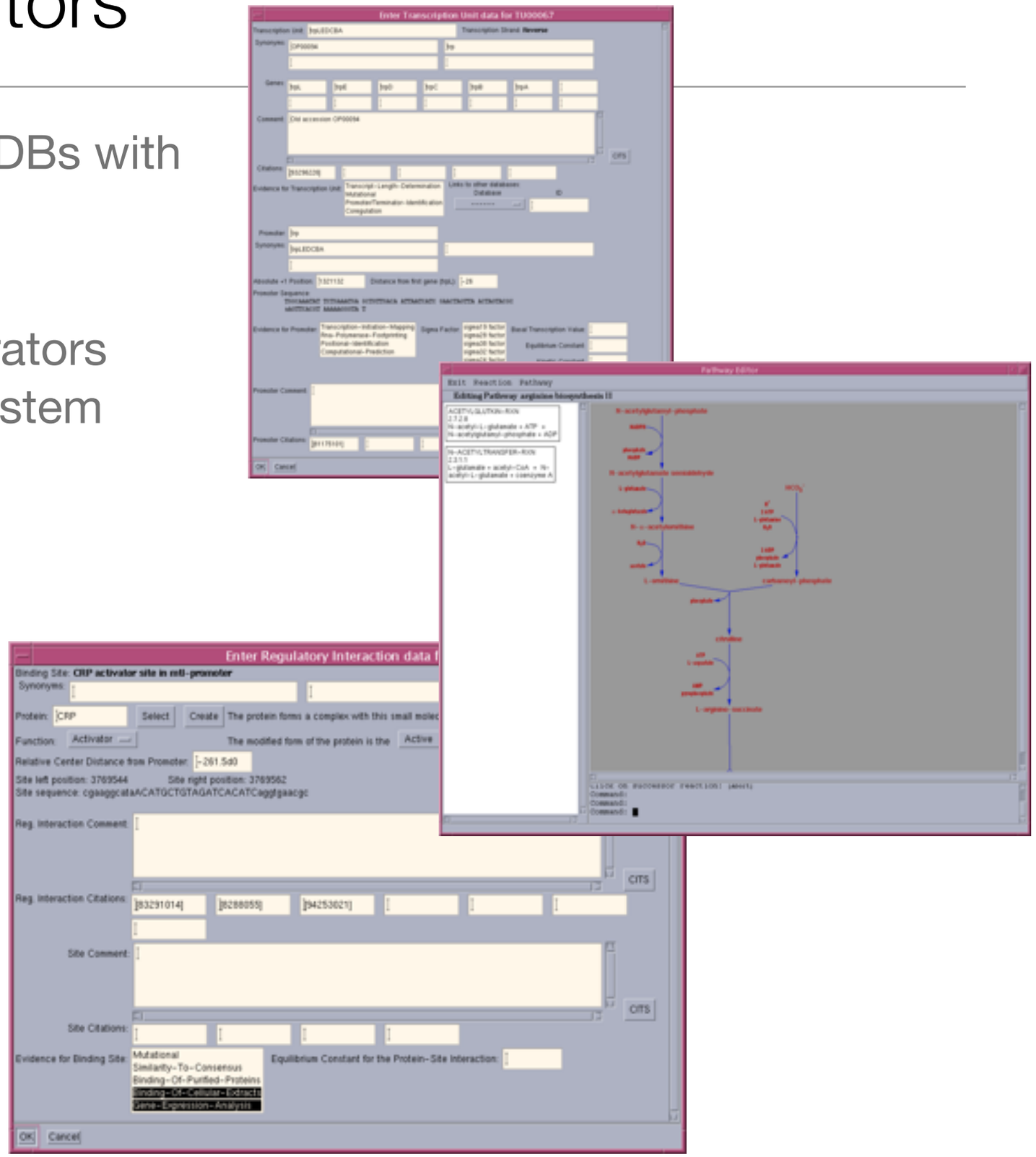
Auto: Pathway Tools Software: PathoLogic

- Computational creation of new Pathway/Genome Databases
- Transforms genome into Pathway Tools schema and layers inferred information above the genome
- Predicts operons
- Predicts metabolic network
- Predicts pathway hole fillers
- Infers transport reactions

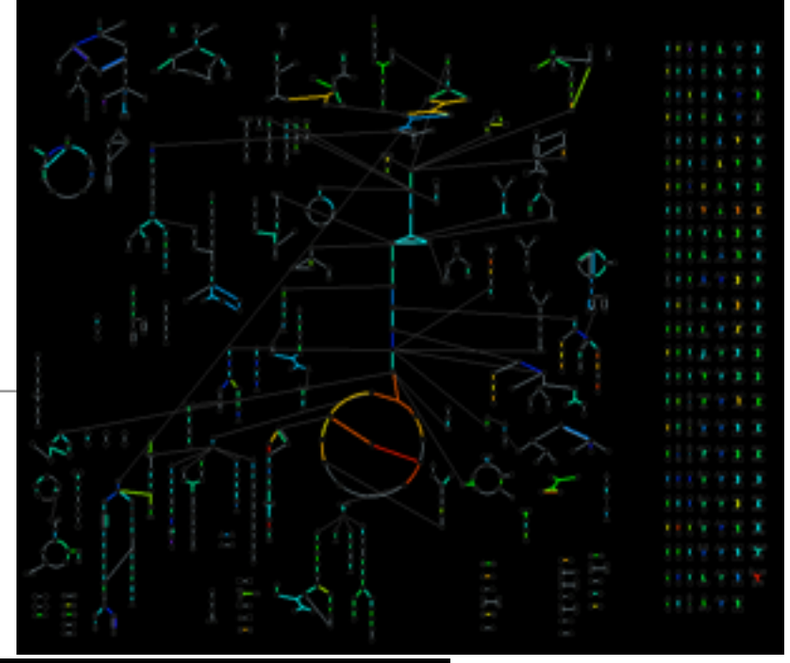
[Slides by Peter Karp]

Pathway Tools Software: Pathway/Genome Editors

- Interactively update PGDBs with graphical editors
- Support geographically distributed teams of curators with object database system
- Gene editor
- Protein editor
- Reaction editor
- Compound editor
- Pathway editor
- Operon editor
- Publication editor

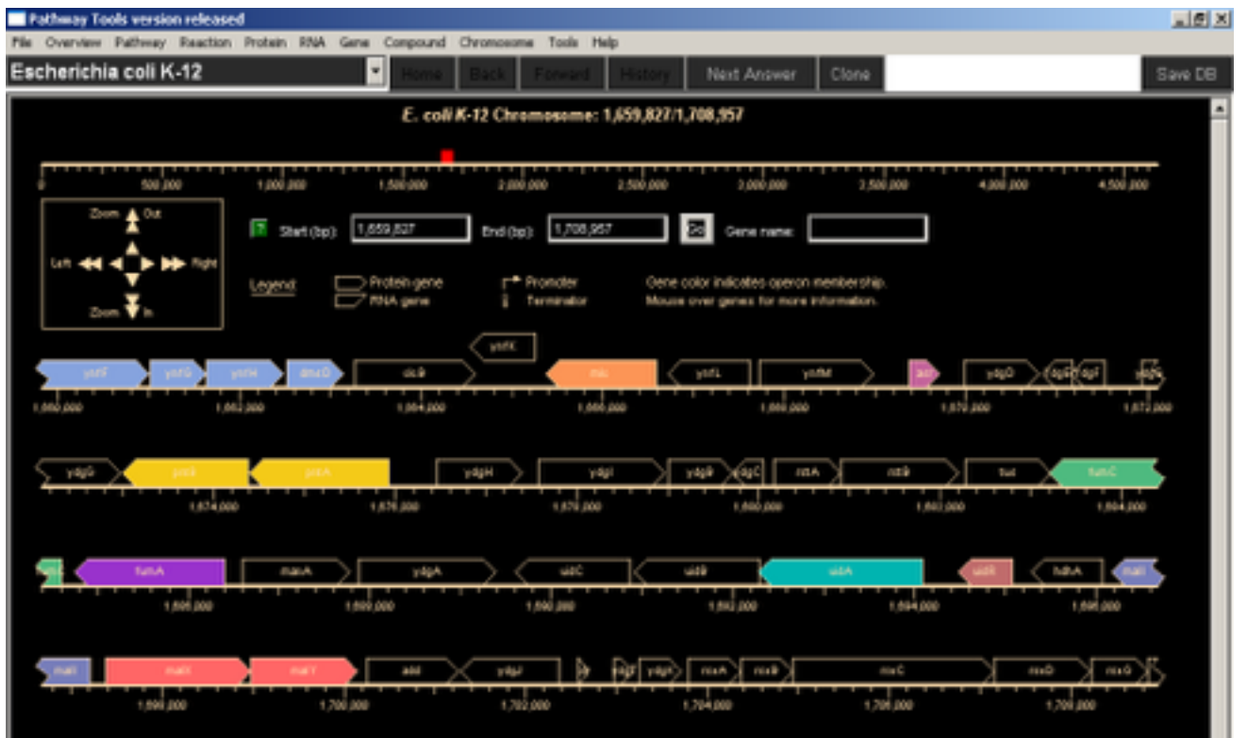


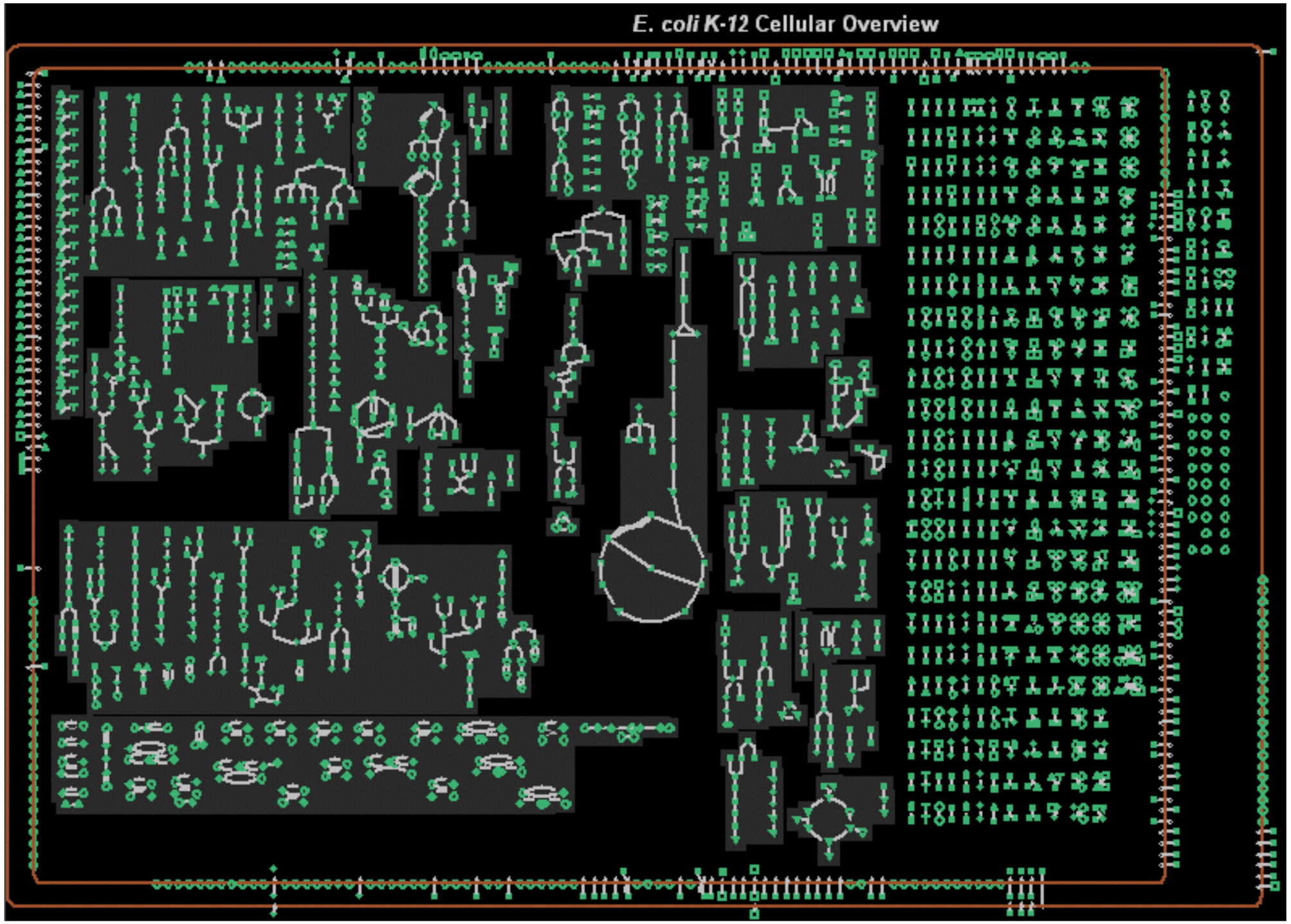
Pathway Tools Software: Pathway/Genome Navigator



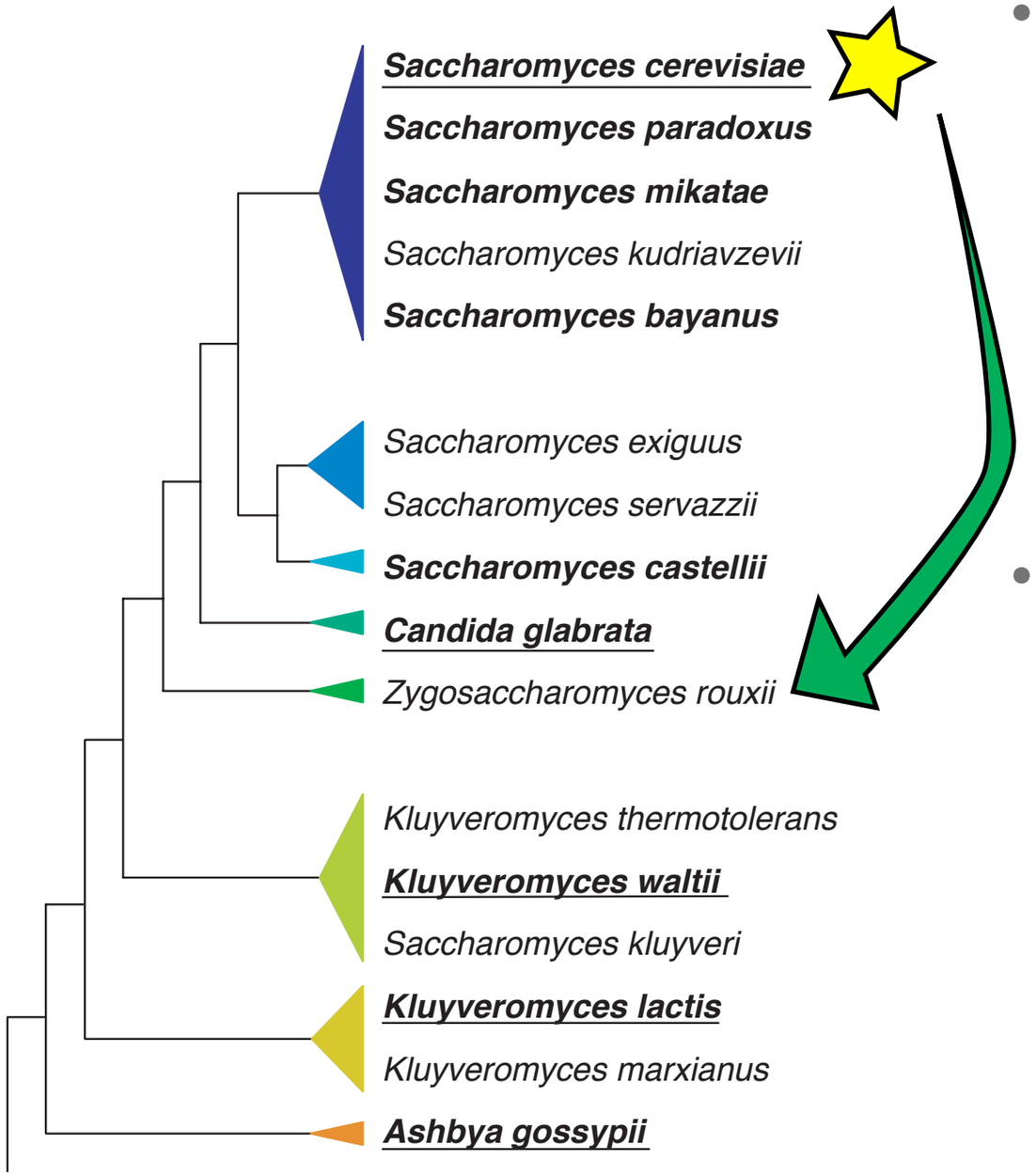
- Querying, visualization of pathways, chromosomes, operons
- Analysis operations
  - Pathway visualization of gene-expression data
  - Global comparisons of metabolic networks
  - Comparative genomics

- WWW publishing of PGDBs
- Desktop operation




E. coli K-12 Cellular Overview

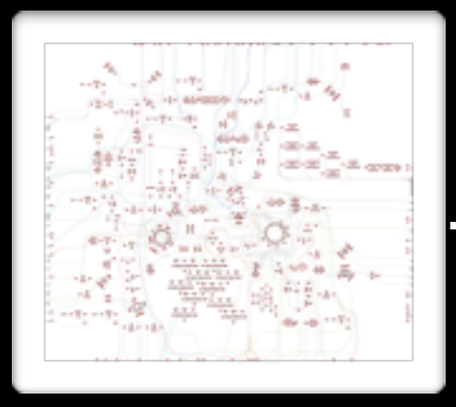
Auto: Pantograph
Re-using existing models

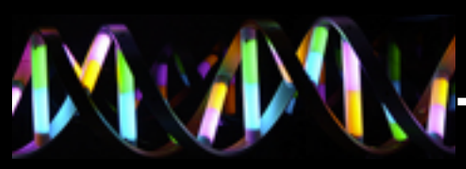


- We need methods that can use an existing model to provide a base to build models for other organisms
- This should decrease the amount of work needed to build a metabolic model

Auto: Pantograph

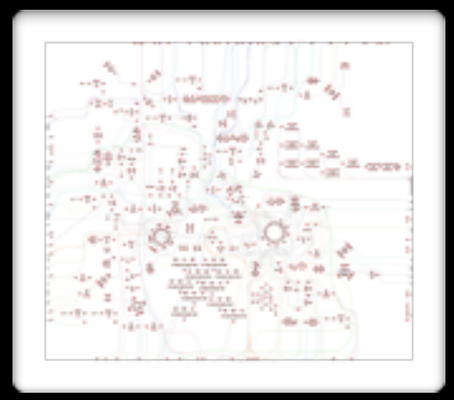
Model Organism 1



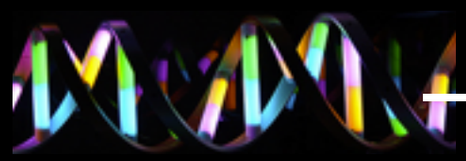


Genes Organism 1

PANTOGRAPH METHOD

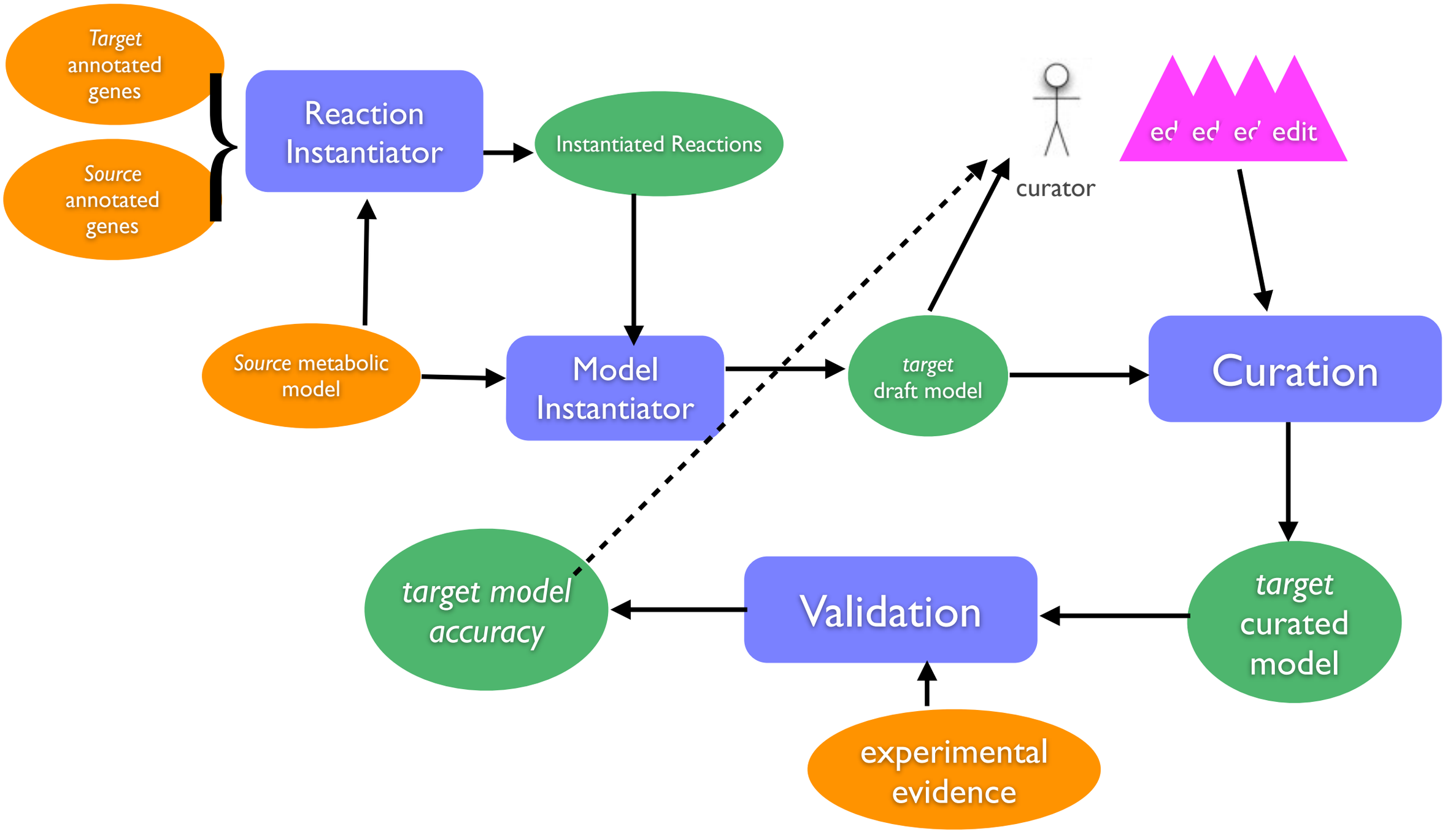


Model Organism 2



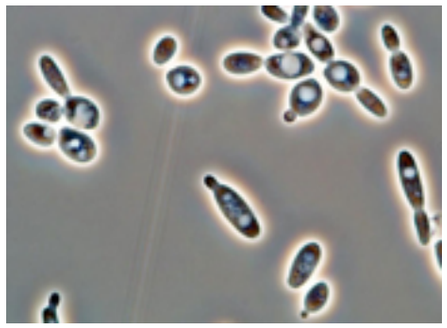
Genes organism 2

Pantograph Workflow



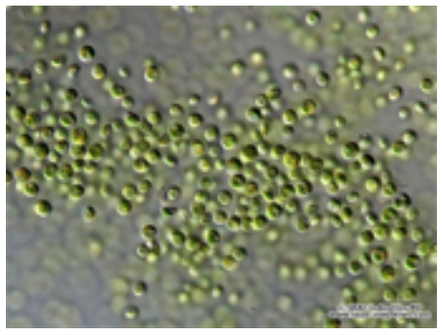
Reconstructions with **Pantograph**

[Loira et al., 2012]



Yarrowia lipolytica

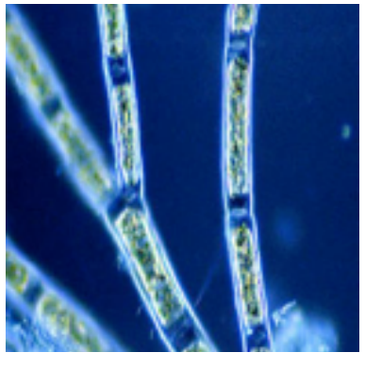
(INRA Micalis, France)



Nannochloropsis salina

(AUSTRAL Biotech, Chile)

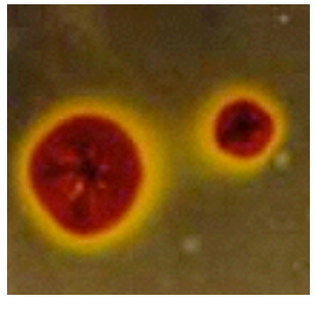
Biofuels



Ectocarpus siliculosus

(INRIA Dylis, France)

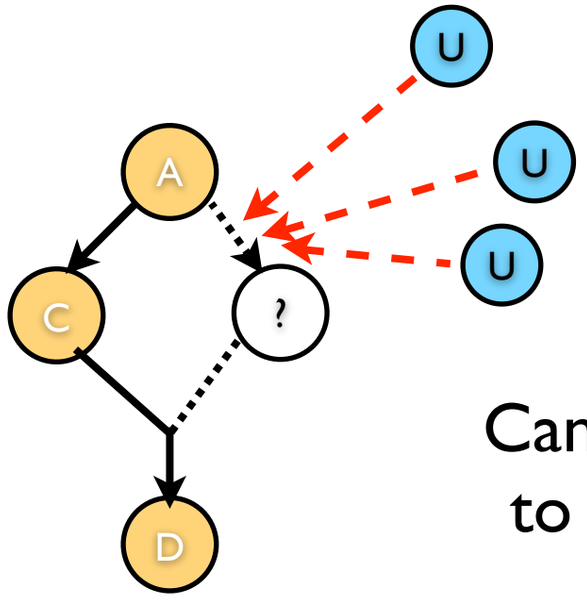
Biomining



Acidithiobacillus ferrooxidans

(MATHomics, CMM, U.Chile)

Automatic procedures can generate models with gaps



Candidates to fill gap

Metabolic network

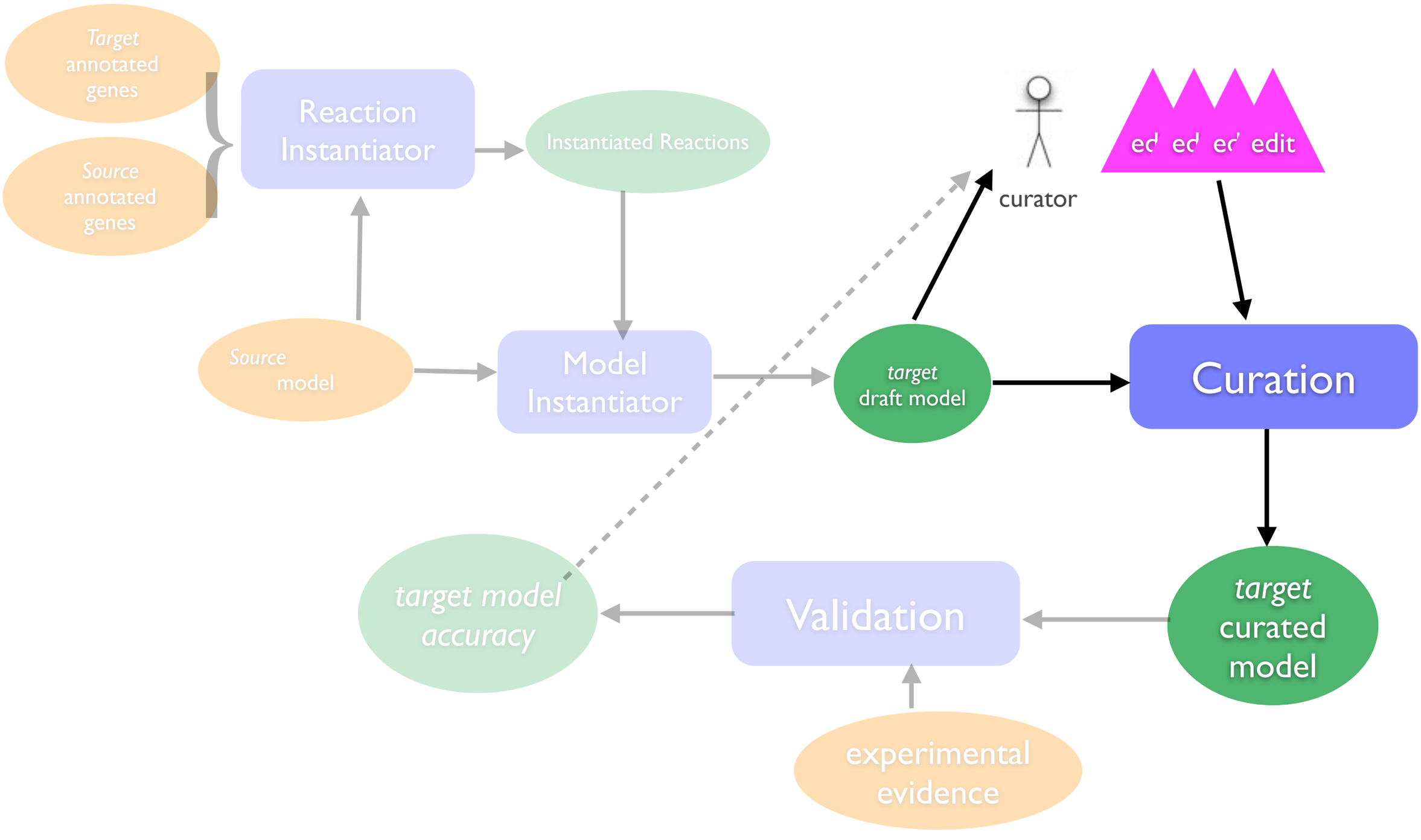
Gaps can exist because:

- The organism doesn't have that reaction
- The tools could not identify an homolog for the genes
- The organism have a different way to produce this enzymatic reaction

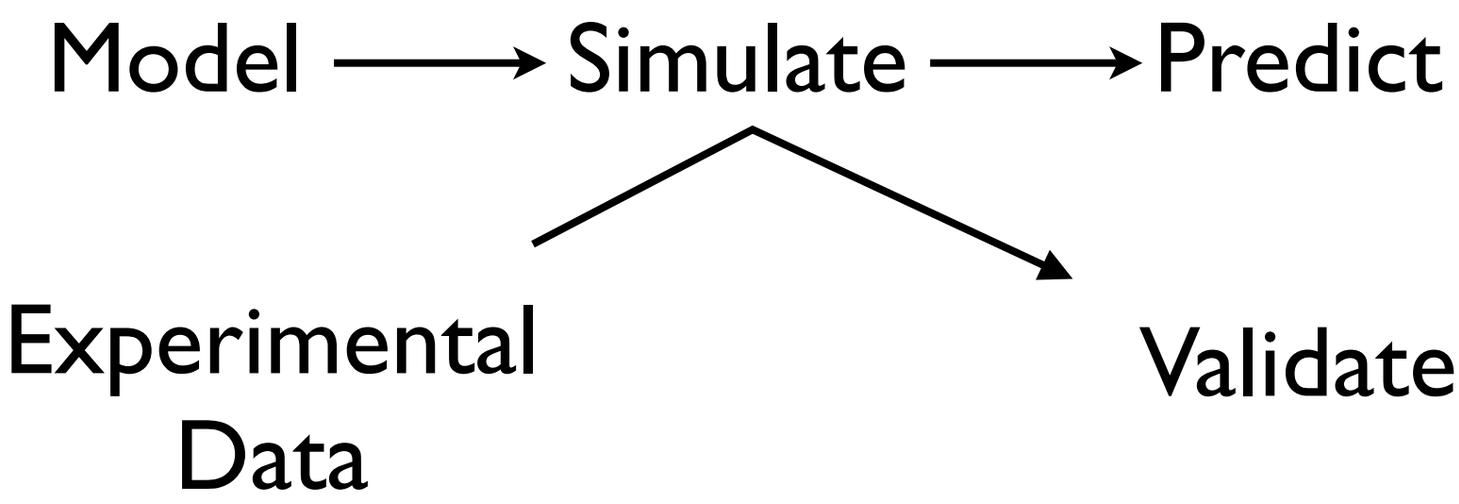
Models reconstructed by automatic methods still require manual curation

- To Fix non-obvious gaps (but obvious for an expert in the organism)
- To Add relevant knowledge from the literature

Manual curation



Models can be used to predict behavior



Fluxes

Fluxes account for the number of times a reaction happens

- Representation of the variation of Metabolites in a time unit, given some flux:

$$S * v = dx/dt$$

S: Stoichiometric Matrix

v: Flux Vector

x: Metabolite vector

$$x=(M_1\ M_2\ M_3\ M_4 \ldots)$$

Steady State

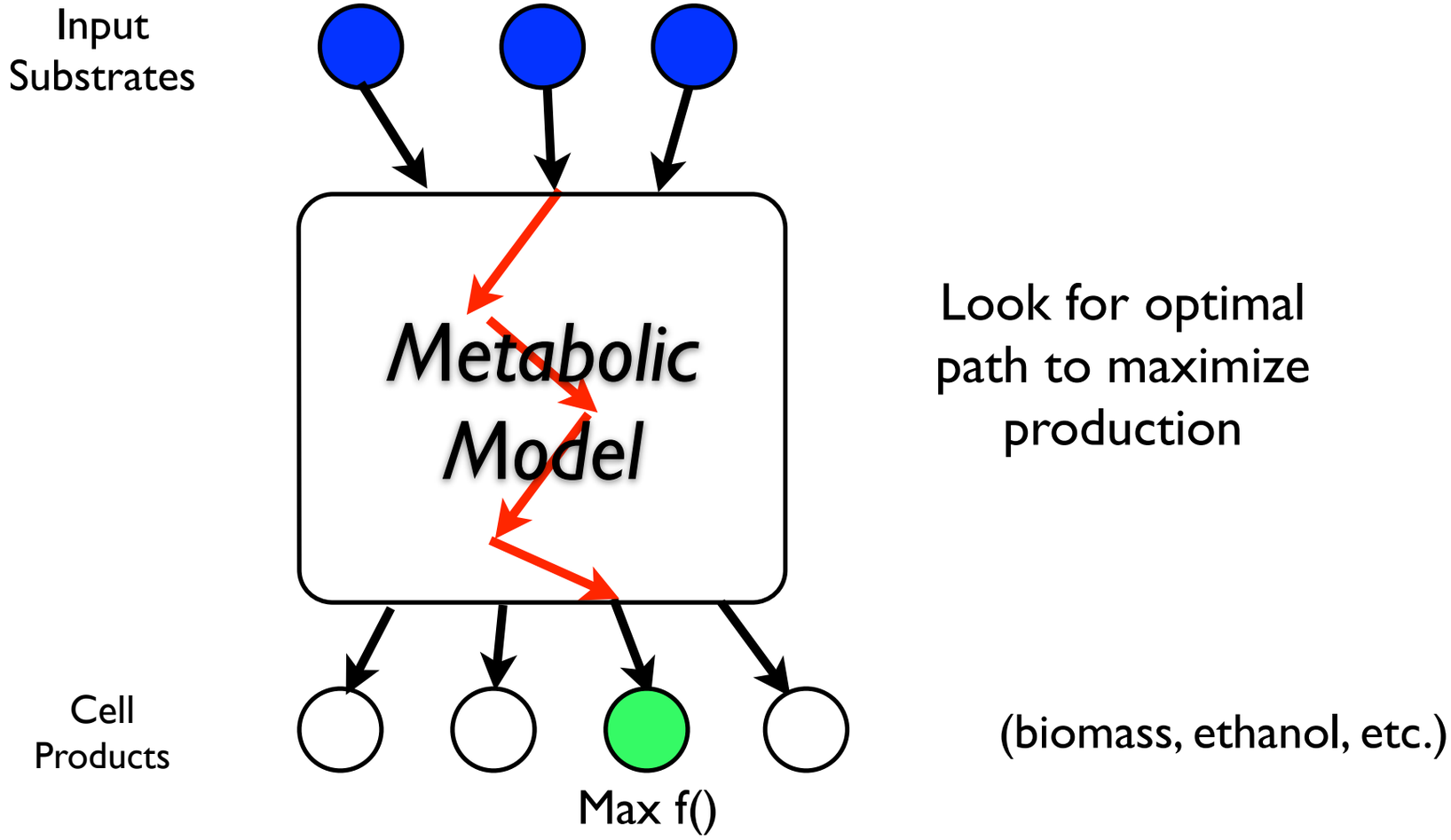
Steady-state mean number of internal metabolites is constant in time

$$S * \vec{v} = d\vec{x}/dt = 0$$

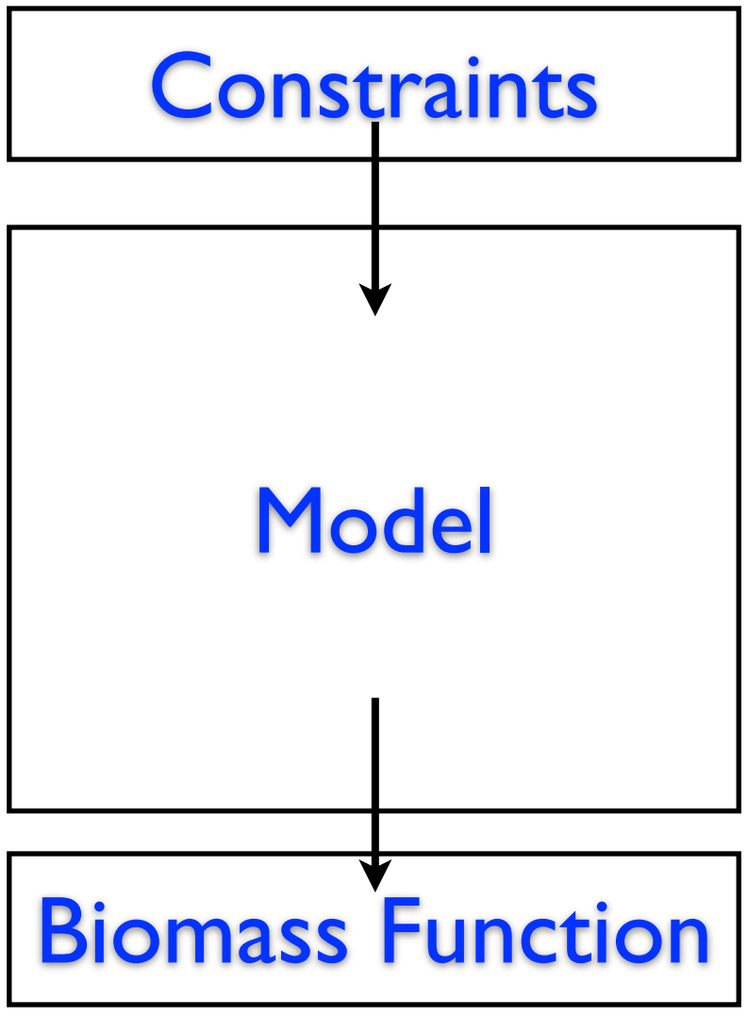
- Biological Systems tend to stabilize with time in a Steady State

Flux Balance Analysis (FBA)

Flux Balance Analysis (FBA) is useful to search fluxes that maximize a function



FBA solutions are restricted by constrains



FBA can be used to maximize biomass production

Biomass = 1.134800 1,3betaDglcn + 0.458800 ala + 0.046000 amp + 0.160700 arg + 0.101700 asn + 0.297500 asp + 59.276000 atp + 0.044700 cmp + 0.006600 cys + 0.003600 damp + 0.002400 dcmp + 0.002400 dgmp + 0.003600 dtmp + 0.000700 ergst + 0.105400 gln + 0.301800 glu + 0.290400 gly + 0.518500 glycogen + 0.046000 gmp + 59.276000 h2o + 0.066300 his + 0.192700 ile + 0.296400 leu + 0.286200 lys + 0.807900 mannan + 0.050700 met + 0.000006 pa + 0.000060 pc + 0.000045 pe + 0.133900 phe + 0.164700 pro + 0.000017 ps + 0.000053 ptdlino + 0.185400 ser + 0.020000 so4 + 0.191400 thr + 0.023400 tre + 0.000066 triglyc + 0.028400 trp + 0.102000 tyr + 0.059900 ump + 0.264600 val + 0.001500 zymst

From a constrained flux problem we build an LP problem

Maximize: $Z = \omega \cdot v$

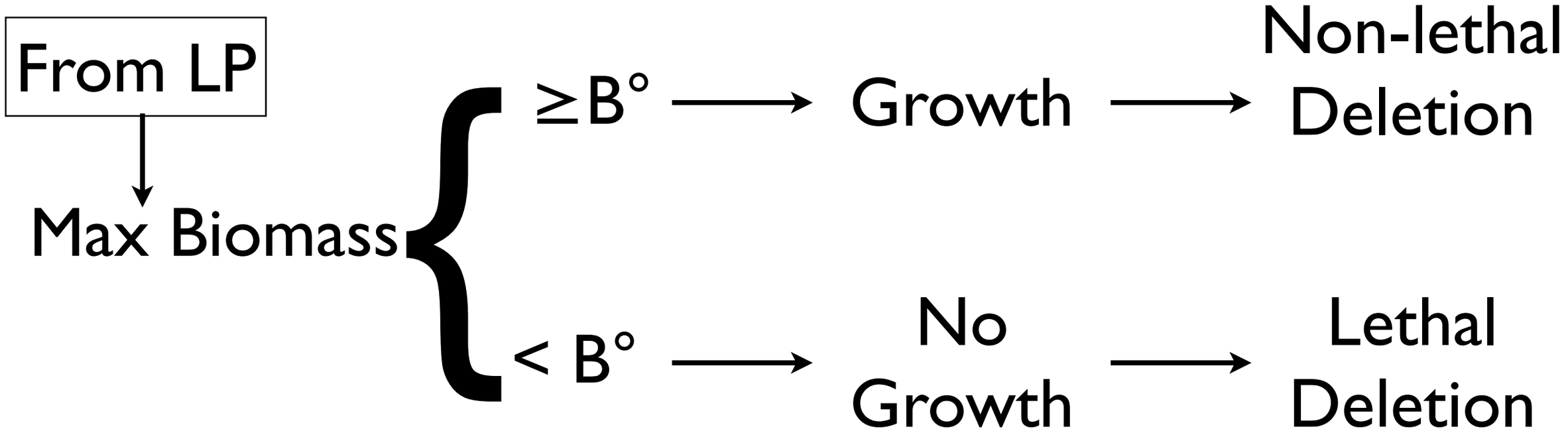
ω define metabolites in biomass function

Subject to: $S \cdot v = 0$

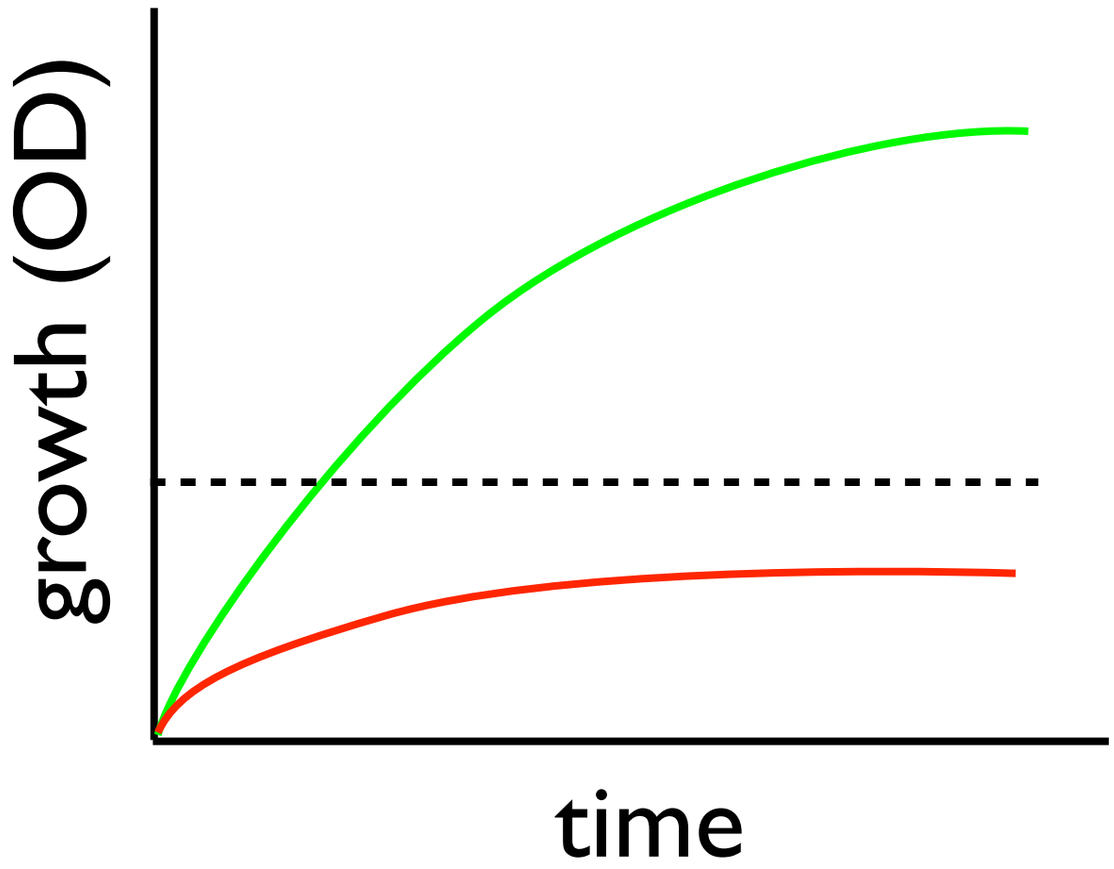
$\alpha \leq v \leq \beta$

α and β are bounds to fluxes

We use FBA to measure effects of reaction deletion in biomass



Knocking out a gene and measuring growth



Viable

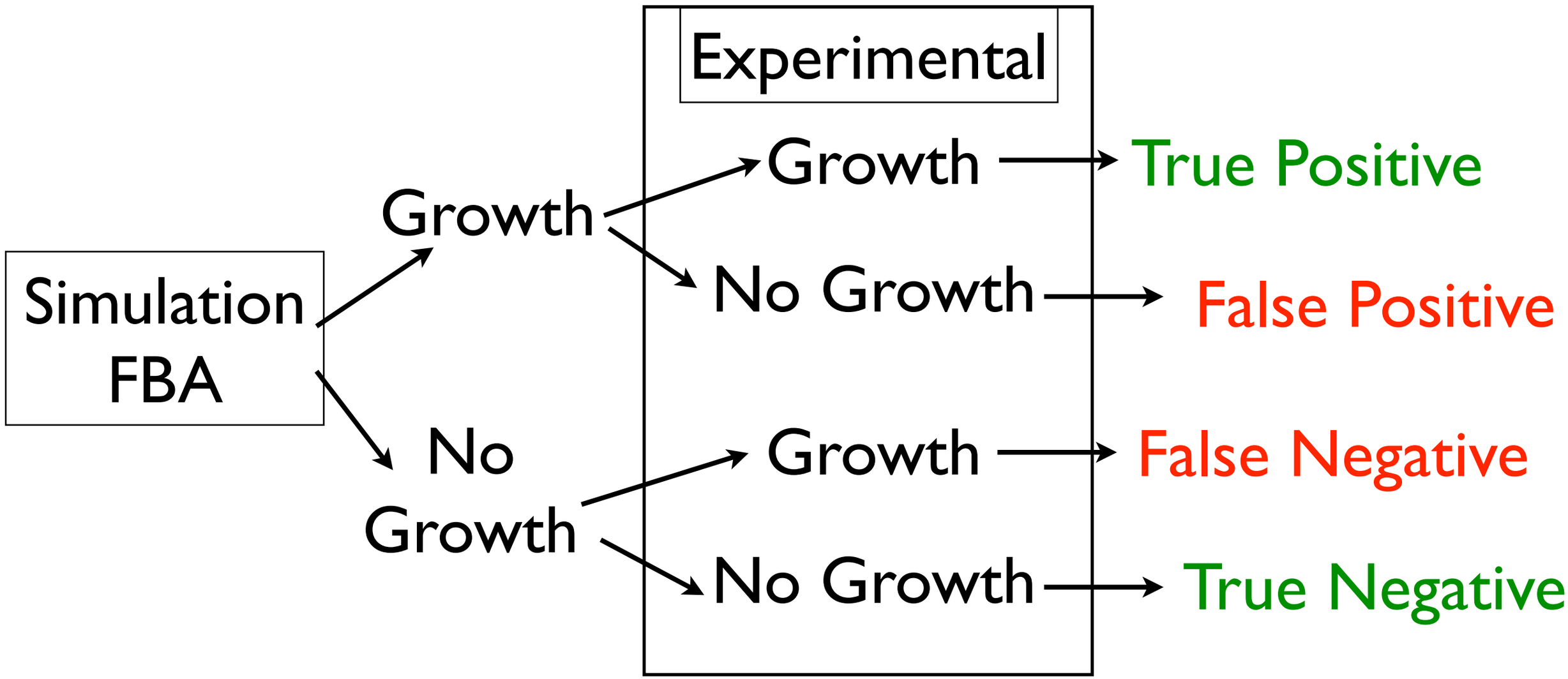
[Joice, 2006]

(threshold=⅓ of average OD)

Lethal

Simulation vs experimental evidence

A simulation is accurate if predicts (lack of) growth correctly



Metabolic model accuracy

- Example: Model iND750, Biomass Function iND750
- Against experimental results from (Winzeler,1999)

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

	Rich Media	Minimal Media
TP	49	48
TN	1	6
FP	3	4
FN	9	4
Accuracy	0.806	0.871

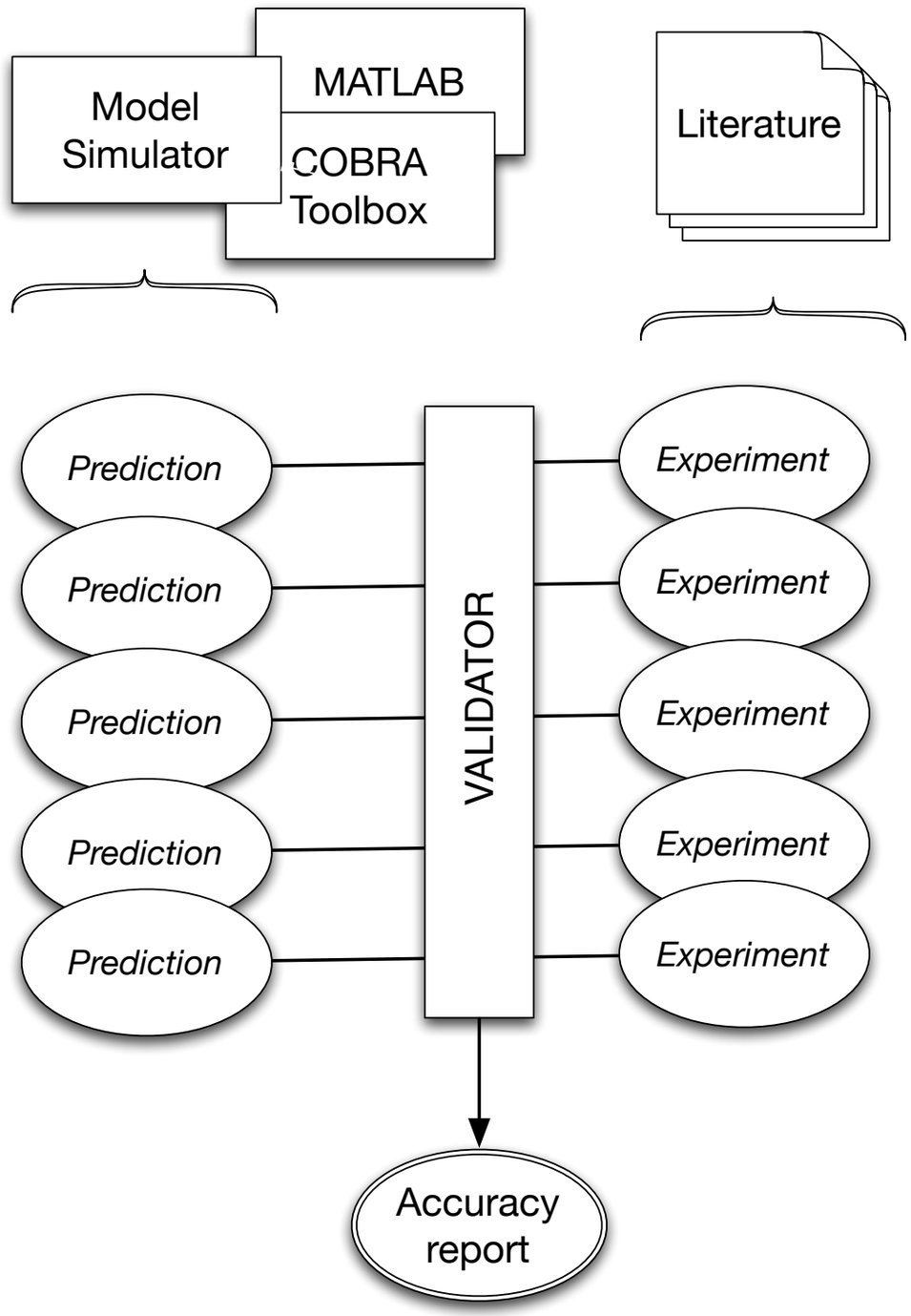
Accuracy=
(Tp+Tn)/(tp+tn+fp+fn)

sensitivity=
tp/(tp+fn)

specificity=
tn/(tn+fp)

geom.mean=
sqr(sens*specif)

Accuracy report



Detailed validation

Ref	Media	*Y. lipolytica* knocked locus	*S. cerevisiae* ortholog	Gene name	Exp. Growth	Simul. Growth	Result
[36]	Aspartate	YALI0C24101g	YGL062W	*PYC1*	+	+	TP
[36]	Glutamate	YALI0C24101g	YGL062W	*PYC1*	+	-	FN
[36]	YNBD	YALI0C16885g	YER065C	*ICL1*	+	+	TP
[36]	Ethanol	YALI0C16885g	YER065C	*ICL1*	-	-	TN
[36]	Aspartate	YALI0C16885g	YER065C	*ICL1*	+	+	TP
[36]	Glutamate	YALI0C16885g	YER065C	*ICL1*	+	-	FN
[36]	YNBD	YALI0C24101g YALI0C16885g	YGL062W YER065C	*ICL1* *PYC1*	-	-	TN
[36]	Ethanol	YALI0C24101g YALI0C16885g	YGL062W YER065C	*ICL1* *PYC1*	-	-	TN
[36]	Aspartate	YALI0C24101g YALI0C16885g	YGL062W YER065C	*ICL1* *PYC1*	+	+	TP

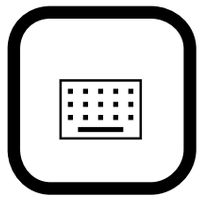
Y. lipolytica Model validation

152 experiments

98 simulations

39 True Positives 16 False Positives

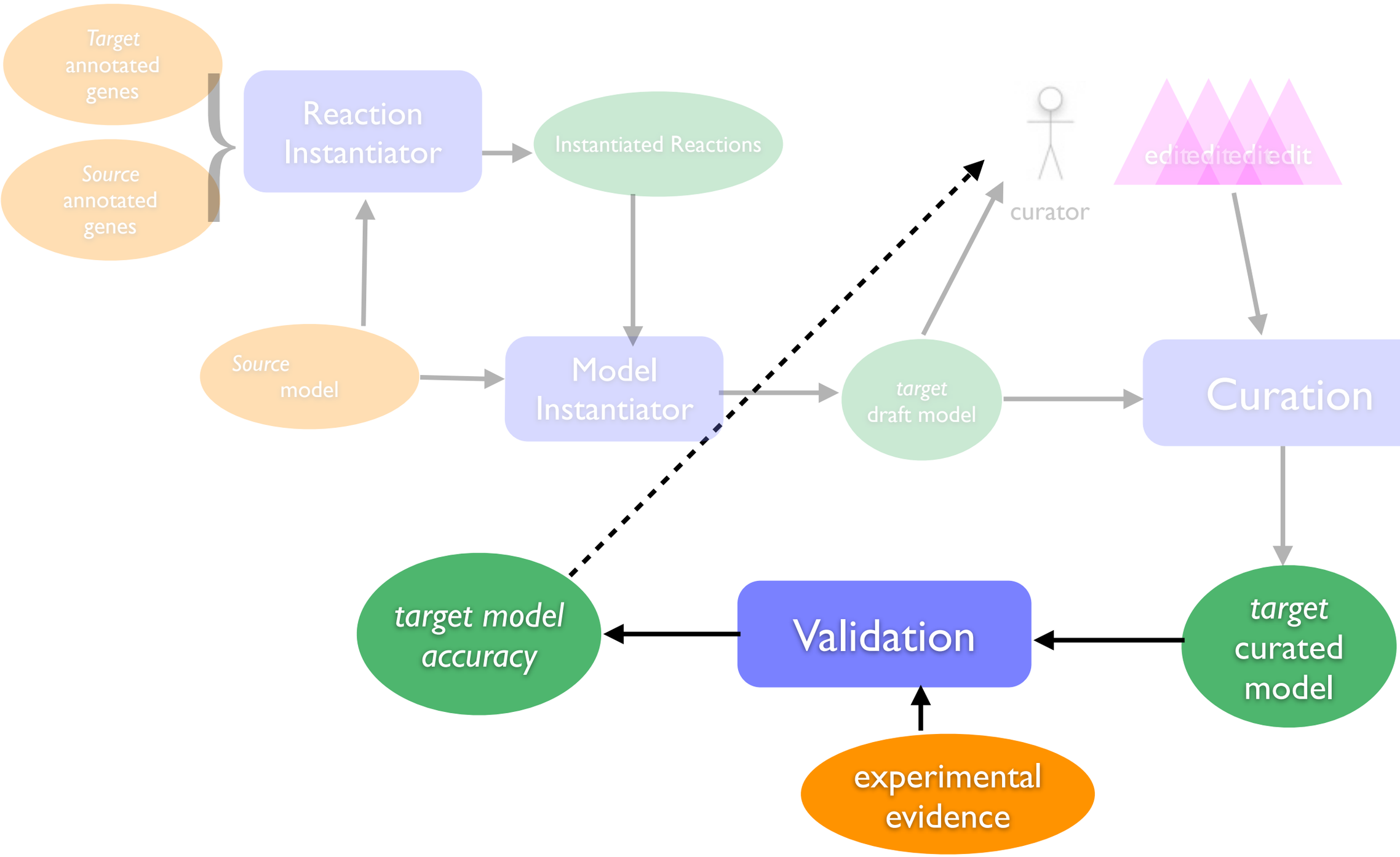
25 True Negatives18 False Negatives



Tools used for this analysis

- COBRA toolbox: for FluxBalanceAnalysis
- Matlab
- glpk/glpkmex: for solving LP
- libsbml/SBMLTools: for SBML handling
- Pantograph

Validation and iterative improvement



FIN

Nicolás Loira
Center for Mathematical Modeling
November 2014
nloira@gmail.com