Genome-scale modeling

Isabel Rocha irocha@deb.uminho.pt

September 12, 2017
DSM, Delft, The Netherlands

CENTRE OF BIOLOGICAL ENGINEERING

SILICOLIFE

UNIVERSITY OF MINHO – RESEARCH GROUP

BIOINFORMATICS AND SYSTEMS BIOLOGY TEAM
CENTRE OF BIOLOGICAL ENGINEERING

CENTRE OF BIOLOGICAL ENGINEERING

Collaboration between two departments since 2004:

Computer Science

Biological Engineering

Team:

6 PhD - faculty / post docs

Around 20 PhD students

~ 10 MSc students w/grants

Funding:

Portuguese national agency (FCT)

European Comission

Companies

Main areas:

Constraint based modeling: Metabolic Engineering and health applications

Metabolic/ regulatory network reconstruction

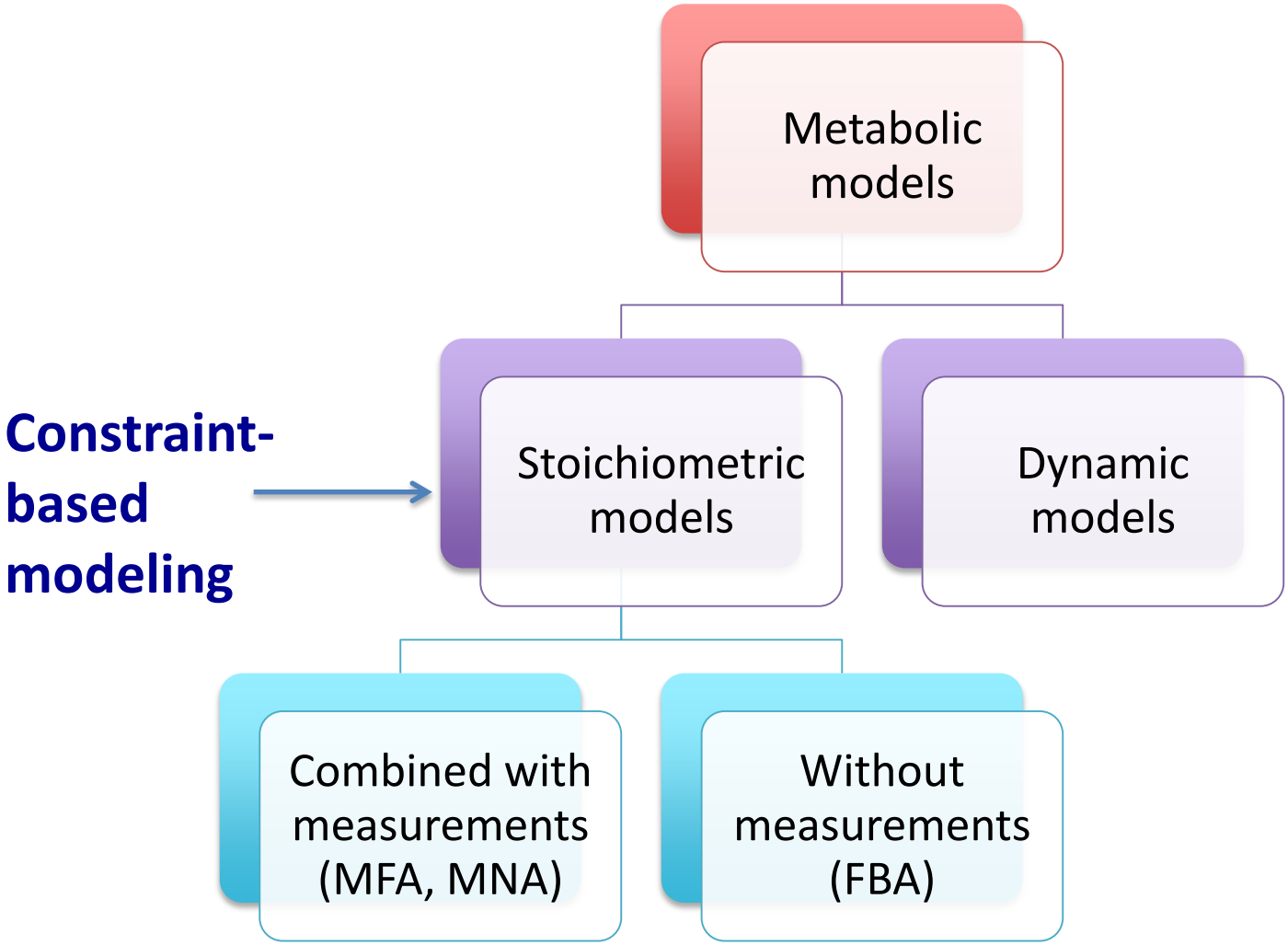
Biomedical Text Mining

http://www.ceb.uminho.pt/biosystems/Labs?lab=1

OUTLINE

- **Metabolic models**
  - Stoichiometric vs dynamic models
  - GSMM reconstruction
- **Simulation Methods**
  - Flux Balance Analysis
  - MOMA
- **Strain design in Metabolic engineering**
  - Metaheuristic Methods (OptGene)
  - MultiObjective Optimization

Metabolic Models



MASS BALANCES

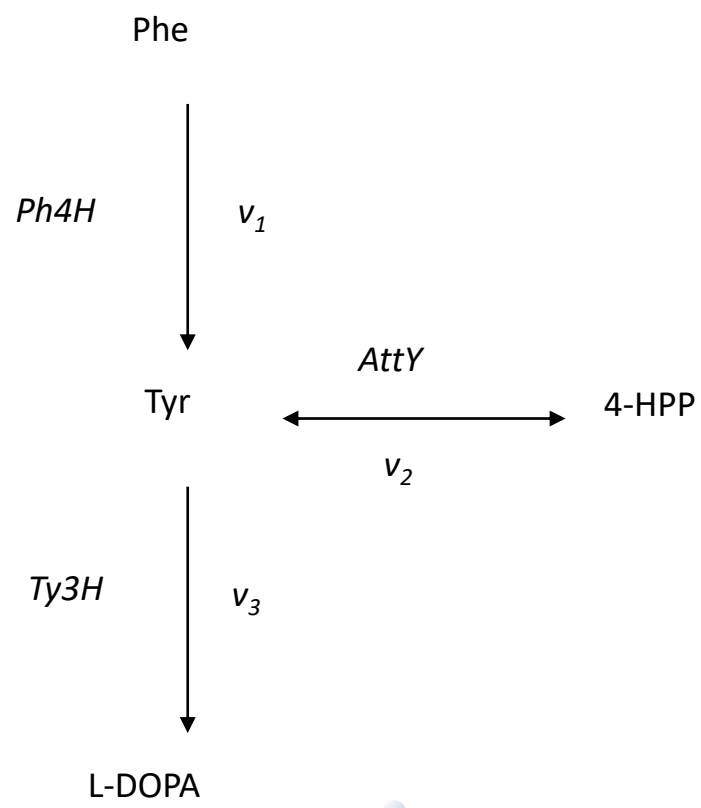
Framework for both dynamic and stoichiometric models:

Mass balance over intra-cellular metabolites

$$\frac{d[Tyr]}{dt} = v_1 - v_2 - v_3 - \mu[Tyr]$$

If, for example, all enzymes can be described by a Michaelis-Menten kinetics

$$\frac{d[Tyr]}{dt} = v_{max1}\frac{[Phe]}{K_{M1}+[Phe]} - v_{max2}\frac{[Tyr]}{K_{M2}+[Tyr]} - v_{max3}\frac{[Tyr]}{K_{M3}+[Tyr]} - \mu[Tyr]$$



If a steady state can be assumed:

$$v_1 - v_2 - v_3 = 0$$

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

Metabolic Models – Dynamic models vs CBM

For *all* considered internal metabolites

1. Mass balance over intracellular metabolites

$$\frac{dx}{dt} = S \cdot v$$

Dynamic or kinetic Models

2. Assumption of (pseudo) steady state

$$S \cdot v = 0 \qquad \beta_j \leq v_j \leq \alpha_j$$

Stoichiometric Models

Result:

<u>Linear equation system</u> described by stoichiometric matrix *S.*

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

Metabolic Models

Stoichiometric models

- Represent only structure: reactions, compounds, stoichiometry, reversibility
- Easier to use in simulation; algebraic methods; constraint-based modeling

Dynamic models

- Represent the concentrations of metabolites and reaction fluxes as a function of time
- Use differential equations
- Harder to simulate
- Require knowledge on enyme kinetics and parameters

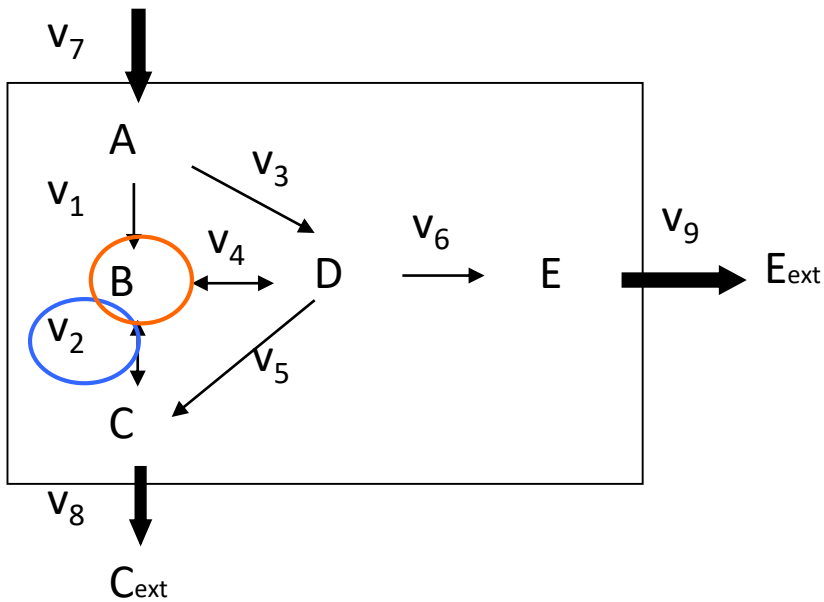
Metabolic Models – assumptions for CBM

Pseudo steady state:

For all intracellular metabolites the fluxes leading to a given metabolite are balanced with the fluxes leading away from the metabolite.

Therefore, <u>there is no net accumulation of metabolites</u>

METABOLIC MODELS – STOICHIOMETRIC MODELS & CBM



Reactions:

$A \xrightarrow{v_1} B$

$B \xleftrightarrow{v_2} 2C$

$A \xrightarrow{v_3} D$

$B \xleftrightarrow{v_4} D$

$D \xrightarrow{v_5} C$

$D \xrightarrow{v_6} E$

$A_{ext} \xrightarrow{v_7} A$

$C \xrightarrow{v_8} C_{ext}$

$E \xrightarrow{v_9} E_{ext}$

Metabolites steady state:

$A: -v_1 - v_3 + v_7 = 0$

$B: v_1 - v_2 - v_4 = 0$

$C: 2v_2 + v_5 - v_8 = 0$

$D: v_3 + v_4 - v_5 - v_6 = 0$

$E: v_6 - v_9 = 0$

Constraints:

$0 \le v_1 \le +\infty$

$-\infty \le v_2 \le +\infty$

$0 \le v_3 \le +\infty$

$-\infty \le v_4 \le +\infty$

$0 \le v_5 \le +\infty$

$0 \le v_6 \le +\infty$

$0 \le v_7 \le a$

$0 \le v_8 \le +\infty$

$0 \le v_9 \le +\infty$

Reactions

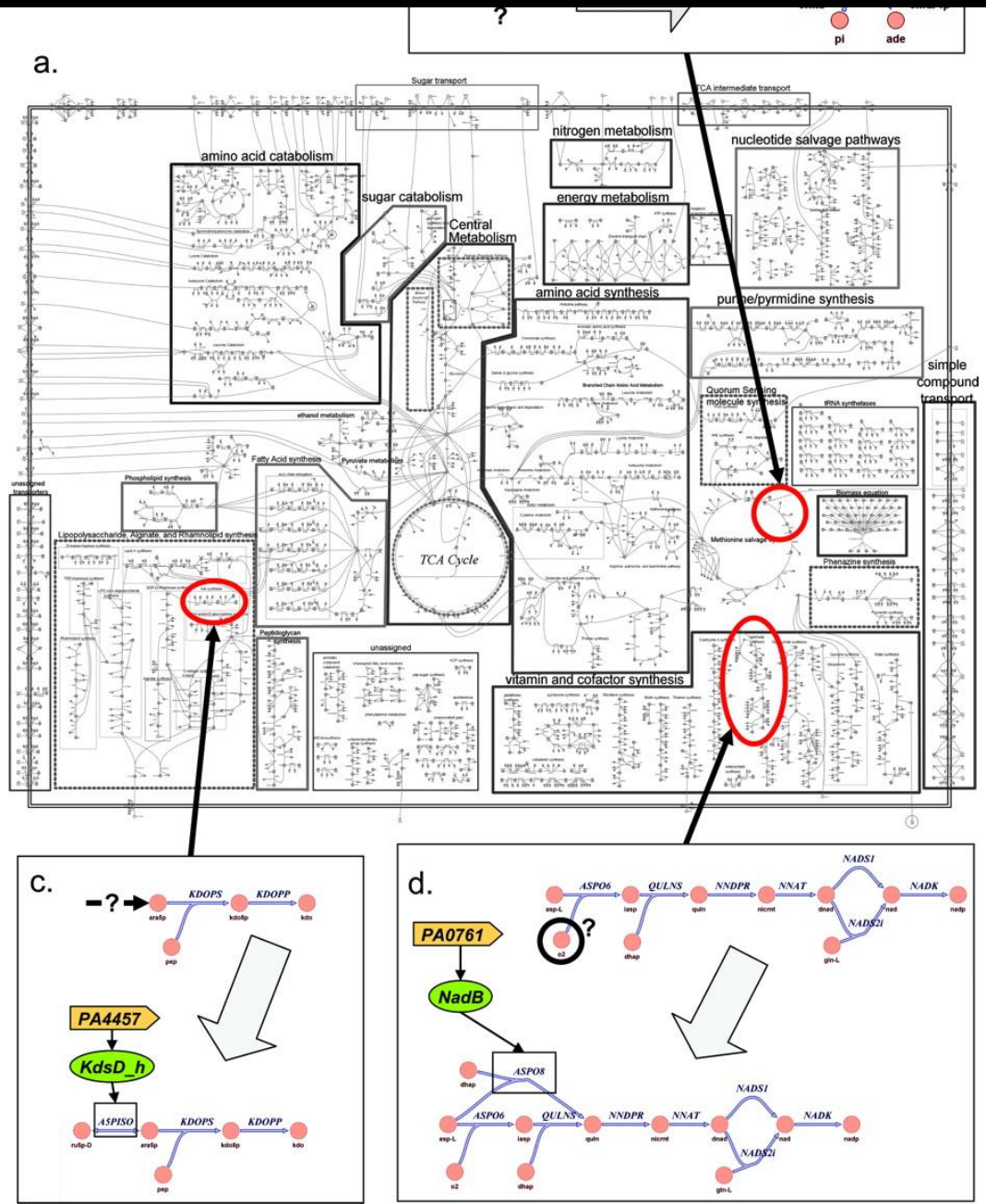
Metabolites

$$\begin{array}{c} A: \\ B: \\ C: \\ D: \\ E: \end{array} \begin{bmatrix} -1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \\ v_8 \\ v_9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

GENOME SCALE METABOLIC MODELS (GSMMs)

This representation is scalable and it is possible to build up these matrices to represent metabolic pathways at a genome-scale level

These GSMMs can account for thousands of genes, reactions and metabolites representing the metabolic capabilities of an organism in a single knowledge-base structure

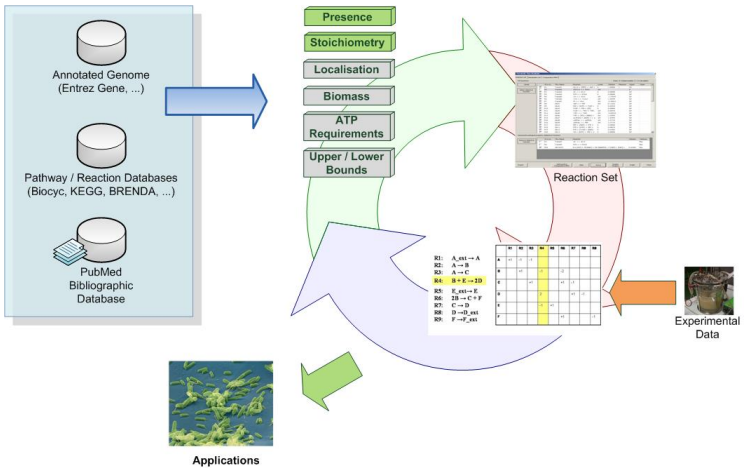


GENOME SCALE METABOLIC MODELS (GSMMS)

HOW CAN WE BUILD THE MODELS IN AN AUTOMATED WAY?

- Ideally, it should be possible to extract most knowledge necessary to construct cellular models from the information obtained during genome sequencing

- However, the knowledge extracted is still limited mainly regarding data needed for dynamic models

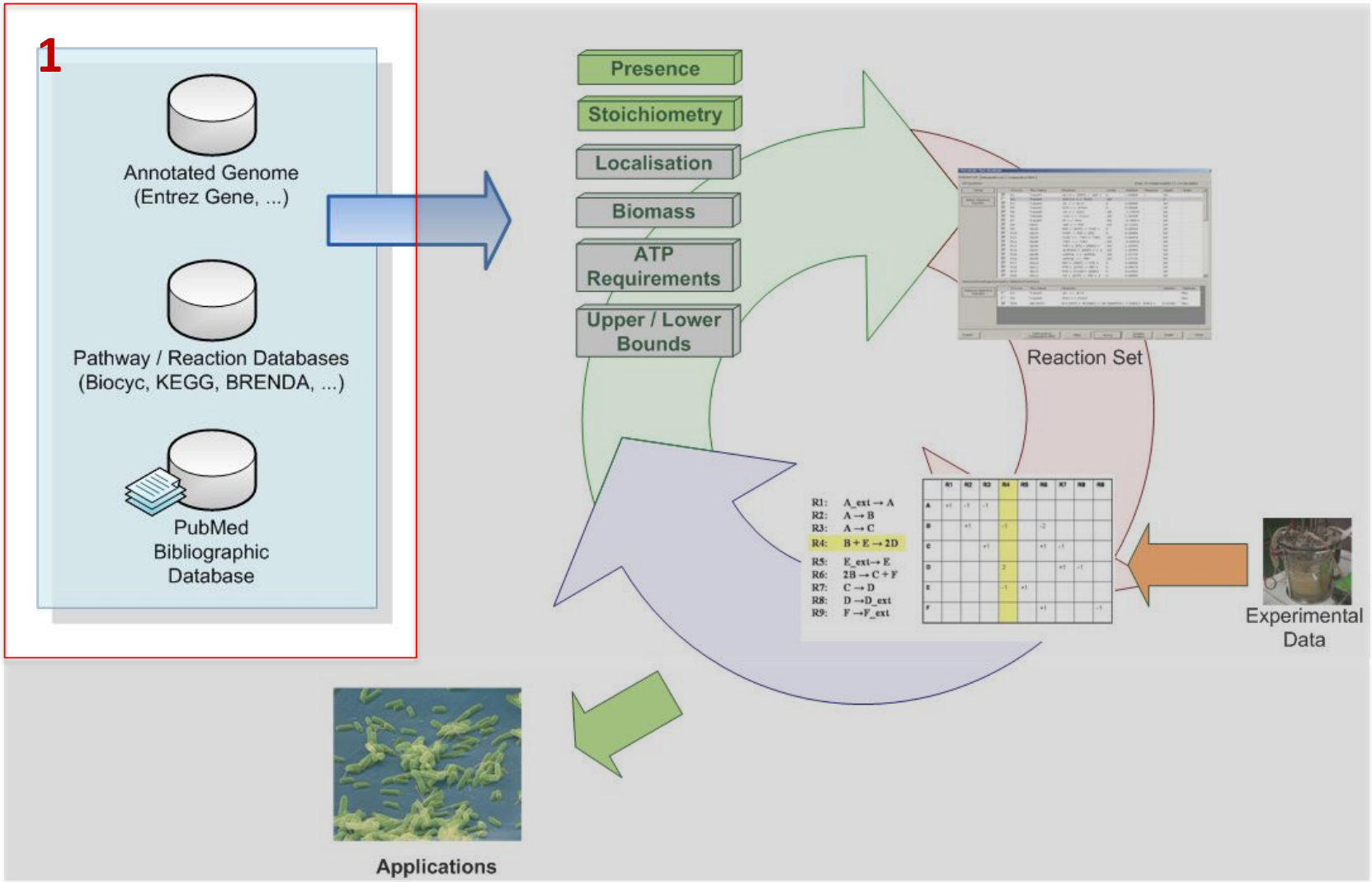


➢ The methodology for semi-automatically obtaining stoichiometric models from genome annotation is quite developed.

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

GSMMs| RECONSTRUCTION - METHODOLOGY



Rocha et al (2008) Methods in Molecular Biology, Vol. 416, Ch. 29, 409

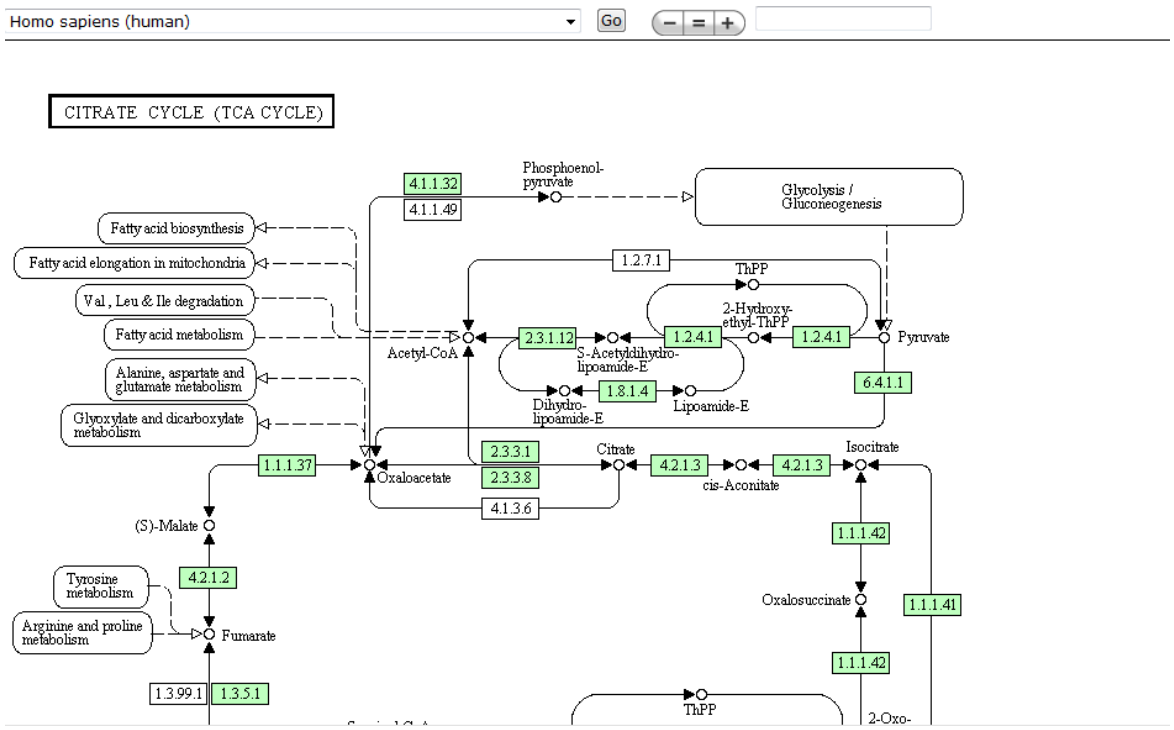
GSMMs| RECONSTRUCTION – DATA SOURCES

Database	Web address	Description
GOLD – Genomes Online Database	http://www.genomesonline.org/	Monitoring of genome sequencing projects, including complete and ongoing projects around the world
NCBI – National Centre for Biotechnology Information – databases	http://www.ncbi.nlm.nih.gov/Genomes/index.html	Contains diverse information related with both microbial and higher organisms genomes, like sequence data, and homology information
KEGG – Kyoto Encyclopedia of Genes and Genomes	http://www.genome.ad.jp/kegg/	Database that includes all microorganisms with publicly available genome sequence. Stores both genomic and metabolic information
BioCyc Database Collection	http://biocyc.org/	Contains several databases (like EcoCyc) that comprise genome and metabolic pathways of single organisms, and also a reference database (MetaCyc) on metabolic pathways from many organisms
ExPASy - Expert Protein Analysis System - Molecular Biology Server	http://www.expasy.org/	The Swiss-Prot and TrEMBL available though ExPASy are protein sequence databases that provide organism specific annotation information. ENZYME is another functionality where enzyme-specific information can be found
BRENDA enzyme database	http://www.brenda-enzymes.org/	Contains information about enzymes. It covers organism related information for most sequenced organisms
TCDB – Transport Classification database	http://tcdb.ucsd.edu/	Classification system for membrane transport proteins known as the Transporter Classification (TC) system (analogous to the Enzyme Commission system for classification of enzymes). Allows similarity searches

GSMMs| RECONSTRUCTION – DATA SOURCES

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

- Several multi-organism databases
- PATHWAY database - knowledge on molecular interaction networks
- GENES database - genes and proteins generated by genome sequencing projects
- LIGAND database - information about chemical compounds and chemical reactions relevant to cellular processes



BRaunschweig ENzyme DAtabase (BRENDA)

(http://www.brenda-enzymes.info/)

- Manually curated and literature-based resource for organism-specific enzymatic data such as kinetics, substrates/products, inhibitors/activators and cofactors.
- Reactions are classified according to the EC system

BRENDA home
BACK
History of your search

Enzyme Nomenclature
EC number
Recommended Name
Systematic Name
Synonyms
CAS Registry Number
Reaction
Reaction Type
Enzyme-Ligand Interactions
Substrate/Product
Natural Substrates
Cofactor
Metals and Ions
Inhibitors
Activating Compound
Functional Parameters
KM Value
Ki Value
IC50 Value
pI Value
Turnover Number
Specific Activity
pH Optimum
pH Range
Temperature Optimum
Temperature Range
Organism related Information
Source Tissue

BRENDA
The Comprehensive Enzyme Information System
EC 1.2.7.1 - pyruvate synthase

CoA + CO2 + 2 reduced ferredoxin + 2 H+				
pyruvate + CoA + 2 oxidized ferredoxin = acetyl-CoA + CO2 + 2 reduced ferredoxin + 2 H+		transient hydroxyethyl-thiamine diphosphate radical intermediate, mechanism of first reductive half-reaction	Clostridium thermoaceticum	288447
pyruvate + CoA + 2 oxidized ferredoxin = acetyl-CoA + CO2 + 2 reduced ferredoxin + 2 H+		reaction via a hydroxyethyl-thiamin pyrophosphate radical intermediate	Moorella thermoacetica	654670
pyruvate + CoA + 2 oxidized ferredoxin = acetyl-CoA + CO2 + 2 reduced ferredoxin + 2 H+		thermodynamic properties and electron transfer between enzyme and native or mutated ferredoxin I	Desulfovibrio africanus	654736
pyruvate + CoA + 2 oxidized ferredoxin = acetyl-CoA + CO2 + 2 reduced ferredoxin + 2 H+		-	-	-

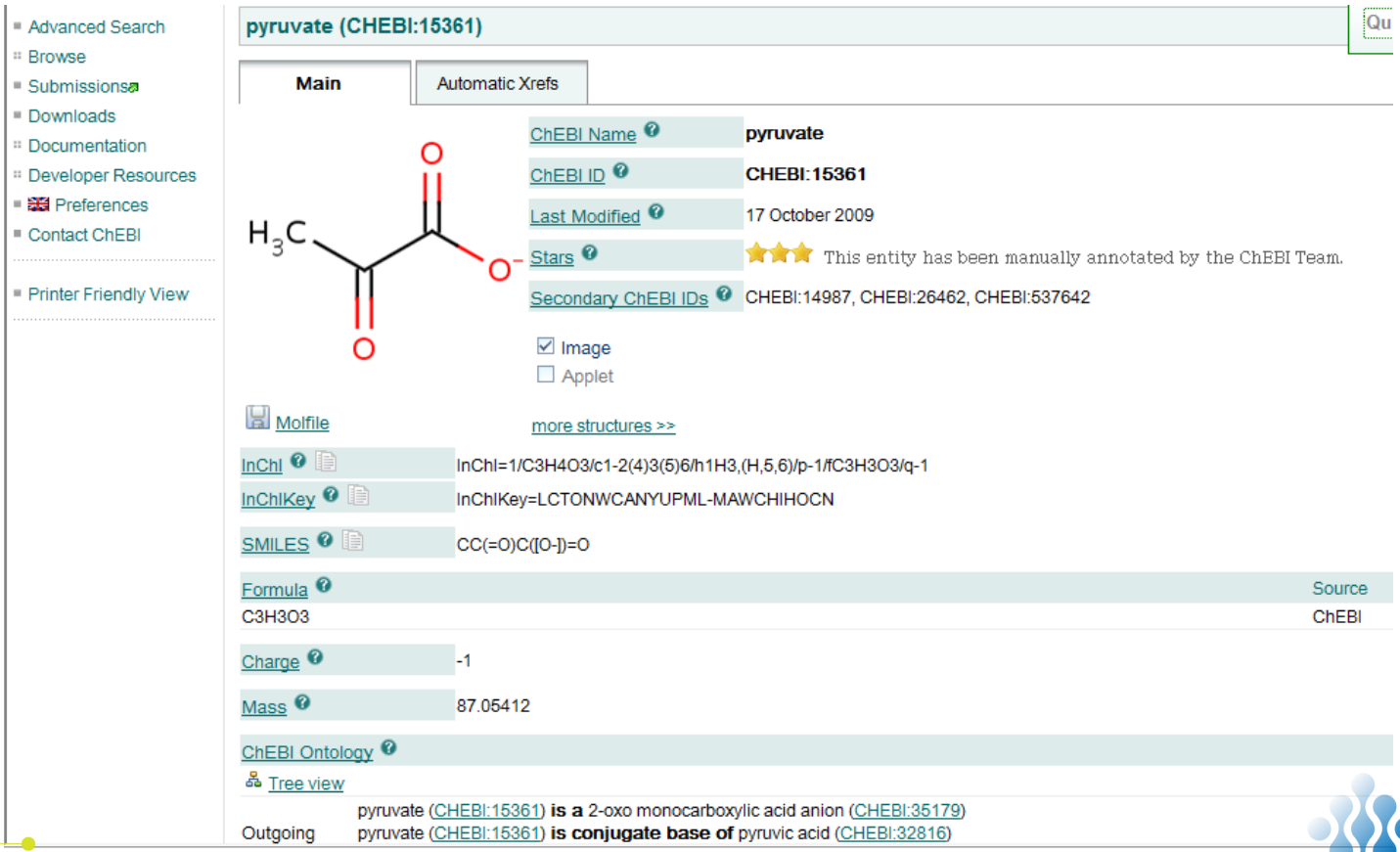
REACTION TYPE	ORGANISM	COMMENTARY	LITERATURE
oxidation	-	-	-
oxidative decarboxylation	-	-	-
redox reaction	-	-	-
reduction	-	-	-
reductive carboxylation	-	-	-

ORGANISM	COMMENTARY	LITERATURE	SEQUENCE CODE	SOURCE
Acetobacterium woodii	-	33097	-	BRENDA
Anabaena cylindrica	-	288406, 288415, 288425, 288426, 288427	-	BRENDA
Archaeoglobus fulgidus	hyperthermophilic sulfate-reducing archaeon	288449	-	BRENDA
Caldithrix abyssi	strain DSM13497T	675857	-	BRENDA
Chlamydomonas reinhardtii	strain 11-32a and strain 83-82	674782		BRENDA

Chemical Entities of Biological Interest (ChEBI)

http://www.ebi.ac.uk/chebi/

- EBI's freely available dictionary of molecular entities focused on chemical compounds
- Includes an ontological classification and employs nomenclature and terminology recommended by the IUPAC and NC-IUBMB



Universal Protein Resource (UniProt) (http://www.uniprot.org/)

- UniProt Knowledgebase (UniProtKB/Swiss-Prot) - fully classified, richly and accurately annotated protein sequence knowledgebase and fully curated entries
- UniProt Reference Clusters (UniRef)
- UniProt Archive (UniParc)

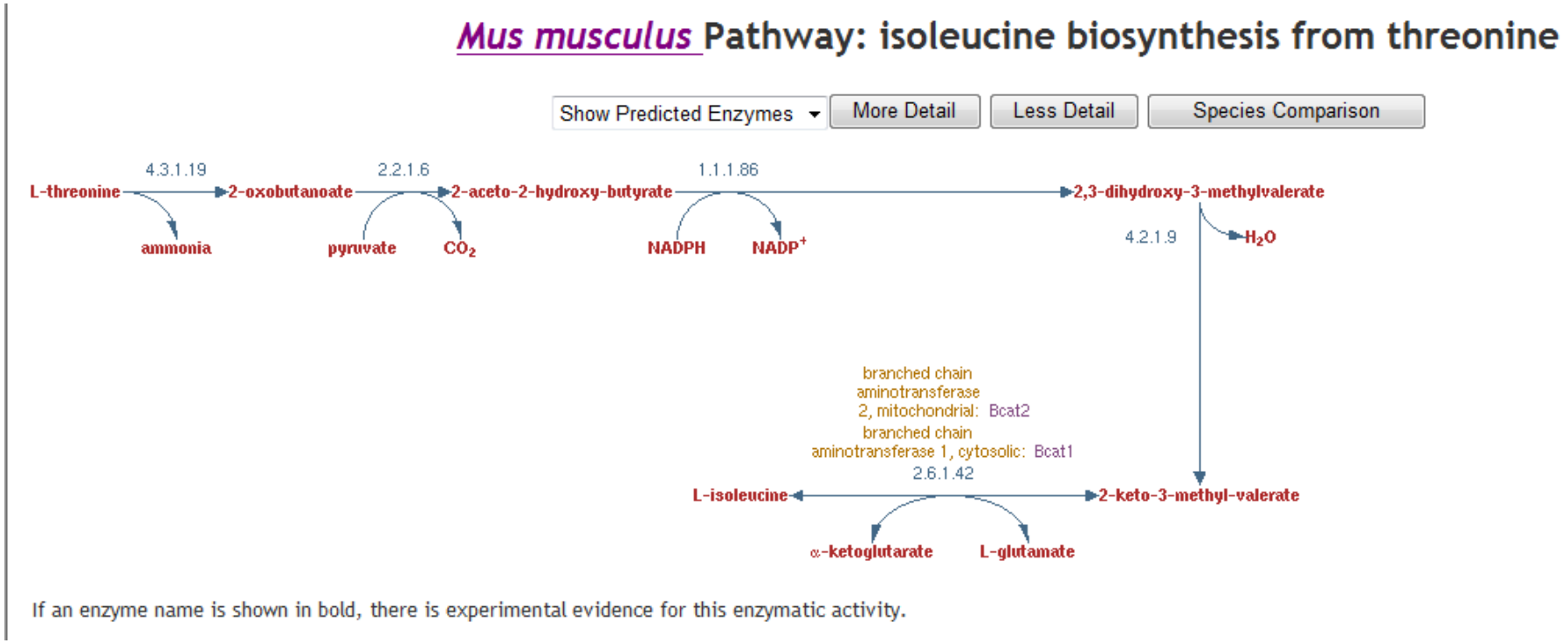
Names and origin	
Protein names	*Recommended name:* **Phosphoenolpyruvate synthase** Short name=PEP synthase EC=2.7.9.2 *Alternative name(s):* Pyruvate, water dikinase
Gene names	Name: **ppsA** Synonyms: pps Ordered Locus Names: b1702, JW1692
Organism	**Escherichia coli (strain K12)** [Complete proteome] [HAMAP]
Taxonomic identifier	83333 [NCBI]
Taxonomic lineage	Bacteria › Proteobacteria › Gammaproteobacteria › Enterobacteriales › Enterobacteriaceae › Escherichia

Protein attributes	
Sequence length	792 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

General annotation (Comments)

GSMMs| RECONSTRUCTION – DATA SOURCES

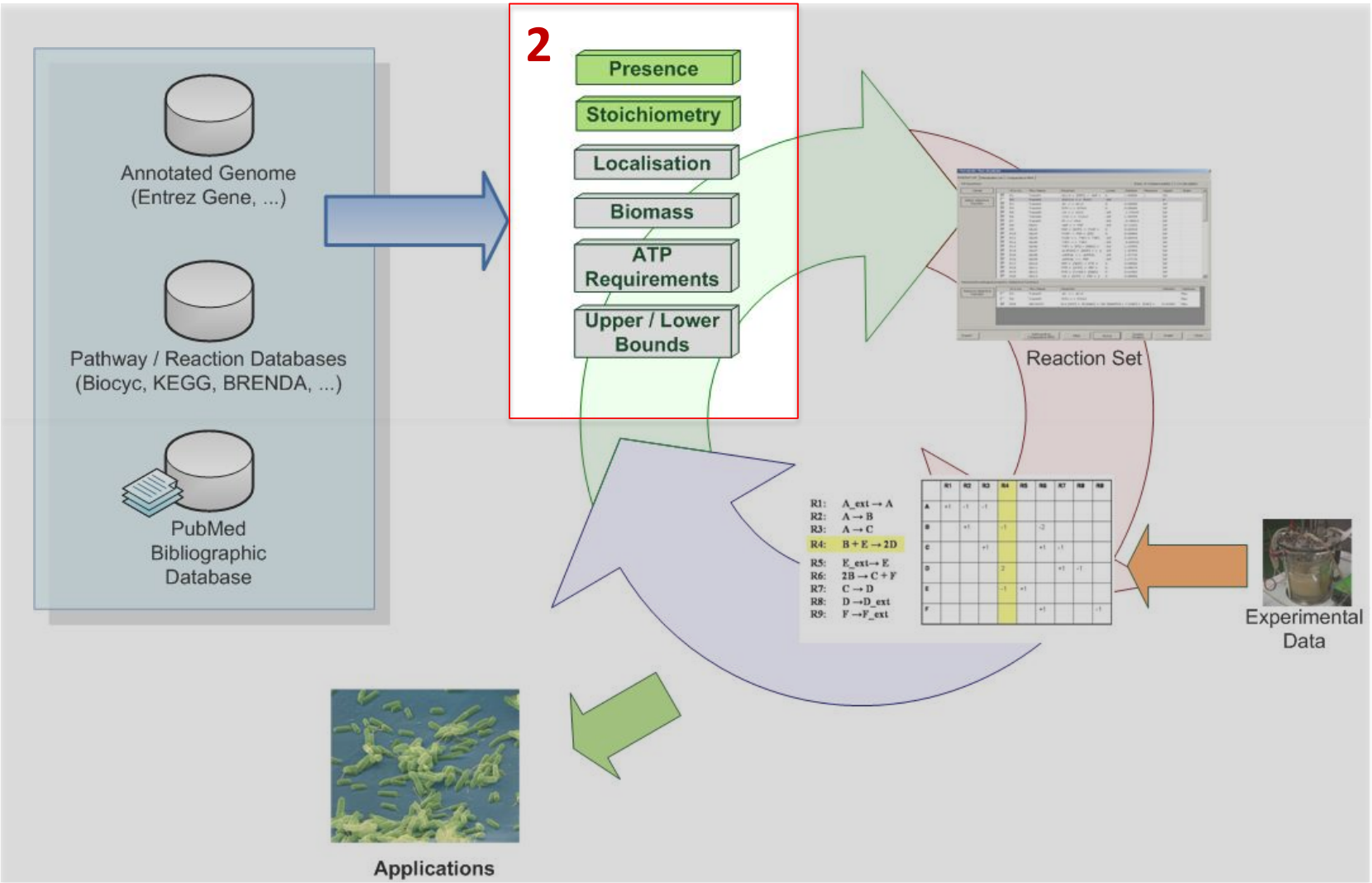
BioCyc (http://biocyc.org/) is a collection of 505 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism. Has 3 tiers, depending on the level of curation.



SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

GSMMs| RECONSTRUCTION - METHODOLOGY



Rocha et al (2008) Methods in Molecular Biology, Vol. 416, Ch. 29, 409

GSMMs| RECONSTRUCTION – BIOINFORMATICS TOOLS

- A number of Bioinformatics tools are used towards the identification of the set of reactions that make the portfolio of a given organism's metabolism
- Homology searching tools, such as BLAST or HMMER can be used to provide a metabolic (re)-annotation comprising sets of homologous genes for each gene (or CDS) of the target organism
- These results, together with databases as UniProt or KEGG are used to identify putative enzymatic functions for those genes/ CDSs in a semi-automatic way (manual curation is still needed for some cases)
- The annotation of transporters is typically challenging requiring other tools and data sources (as TCDB)

GSMMs| RECONSTRUCTION - COMPARTMENTS

- Compartmentalization is important, particularly for metabolites for which there are no specific transporters and diffusion is unlikely to occur
- For prokaryotic organisms:
  - Cytosol
  - Intermembrane compartment (in some cases)
- For eukaryotic microorganisms:
  - Mitochondrion, endoplasmic reticulum, lysosome, glyoxisome, Golgi apparatus, etc.
- For complex organisms, it is also necessary to differentiate between different tissues
- This task might be aided by Bioinformatics tools for protein localization

GSMMs| RECONSTRUCTION – BIOMASS FORMATION

- For *p* biomass constituents, this reaction can be represented as:

$$\sum_{k=1}^{p} c_k X_k \longrightarrow \text{Biomass}$$

- The values of c_k are given by the biomass composition on each metabolite, building block or macromolecule X_k.
- The ATP, NADH and NADPH requirements have to be determined/ known (found in the literature or estimated by fitting the model results to experimental data)
- Growth association requirements related to polymerization of aminoacids, nucleotides, …
- Need to determine energy requirements for maintenance
  - Maintenance of gradients and electrical potential (most important)
  - Turnover of macromolecules

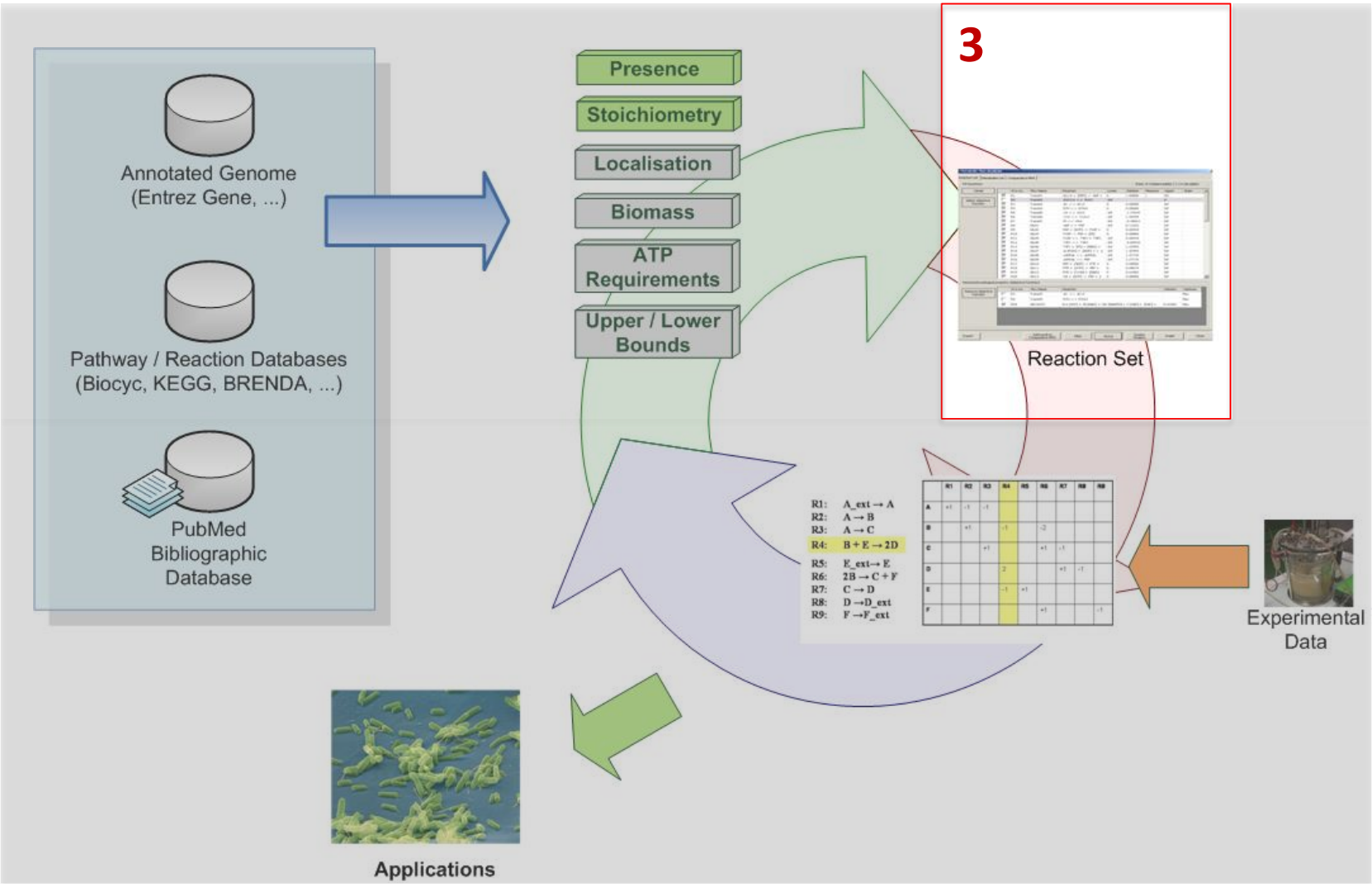
GSMMs| RECONSTRUCTION – OTHER CONSTRAINTS

- Reversibility/irreversibility of the reactions: determined from thermodynamics information regarding the reaction:
  - Setting the minimum of a given reaction flux to zero for irreversible reactions
  - Setting to minus infinity for reversible reactions
  - Information can be collected from databases or inferred with Bioinformatics tools, but not always possible to do it accurately

- When maximal fluxes through a given reaction are known, this can also be added to the model as a constraint.

- The transport flux for nutrients present in the medium can also be constrained.

GSMMs| RECONSTRUCTION - METHODOLOGY



Rocha et al (2008) Methods in Molecular Biology, Vol. 416, Ch. 29, 409

METABOLIC MODELS - FORMATS

B6 | asparagine synthase (glutamine-hydrolysing)

	A	B	C	D	E	F	G
1	Abbreviation	OfficialName	Equation (note [c] and [e] at the beginning refer to the cor	Subsystem	ProteinClassDescription	Ref. (listed below)	
2	ALATA_L	L-alanine transaminase	[c]akg + ala-L <==> glu-L + pyr	Alanine and aspartate n	EC-2.6.1.2		
3	ALAR	alanine racemase	[c]ala-L <==> ala-D	Alanine and aspartate n	EC-5.1.1.1		
4	ASNN	L-asparaginase	[c]asn-L + h2o --> asp-L + nh4	Alanine and aspartate n	EC-3.5.1.1		
5	ASNS2	asparagine synthetase	[c]asp-L + atp + nh4 --> amp + asn-L + h + ppi	Alanine and aspartate n	EC-6.3.1.1		
6	ASNS1	asparagine synthase (glutamine-hyd	[c]asp-L + atp + gln-L + h2o --> amp + asn-L + glu-L + h + ppi	Alanine and aspartate n	EC-6.3.5.4		
7	ASPT	L-aspartase	[c]asp-L --> fum + nh4	Alanine and aspartate n	EC-4.3.1.1		
8	ASPTA	aspartate transaminase	[c]akg + asp-L <==> glu-L + oaa	Alanine and aspartate n	EC-2.6.1.1		
9	VPAMT	Valine-pyruvate aminotransferase	[c]3mob + ala-L --> pyr + val-L	Alanine and aspartate n	EC-2.6.1.66		
10	DAAD	D-Amino acid dehydrogenase	[c]ala-D + fad + h2o --> fadh2 + nh4 + pyr	Alanine and aspartate n	EC-1.4.99.1		
11	ALARi	alanine racemase (irreversible)	[c]ala-L --> ala-D	Alanine and aspartate n	EC-5.1.1.1		
12	FFSD	beta-fructofuranosidase	[c]h2o + suc6p --> fru + g6p	Alternate Carbon Metab	EC-3.2.1.26		
13	A5PISO	arabinose-5-phosphate isomerase	[c]ru5p-D <==> ara5p	Alternate Carbon Metab	EC-5.3.1.13		
14	MME	methylmalonyl-CoA epimerase	[c]mmcoa-R <==> mmcoa-S	Alternate Carbon Metab	EC-5.1.99.1		
15	MICITD	2-methylisocitrate dehydratase	[c]2mcacn + h2o --> micit	Alternate Carbon Metab	EC-4.2.1.99	11	
16	ALCD19	alcohol dehydrogenase (glycerol)	[c]glyald + h + nadh <==> glyc + nad	Alternate Carbon Metab	EC-1.1.1.1		
17	LCADi	lactaldehyde dehydrogenase	[c]h2o + lald-L + nad --> (2) h + lac-L + nadh	Alternate Carbon Metab	EC-1.2.1.22		
18	TGBPA	Tagatose-bisphosphate aldolase	[c]tagdp-D <==> dhap + g3p	Alternate Carbon Metab	EC-4.1.2.40		
19	LCAD	lactaldehyde dehydrogenase	[c]h2o + lald-L + nad <==> (2) h + lac-L + nadh	Alternate Carbon Metab	EC-1.2.1.22		
20	ALDD2x	aldehyde dehydrogenase (acetaldehy	[c]acald + h2o + nad --> ac + (2) h + nadh	Alternate Carbon Metab	EC-1.2.1.3		
21	ARAI	L-arabinose isomerase	[c]arab-L <==> rbl-L	Alternate Carbon Metab	EC-5.3.1.4		
22	RBK_L1	L-ribulokinase (L-ribulose)	[c]atp + rbl-L --> adp + h + ru5p-L	Alternate Carbon Metab	EC-2.7.1.16		
23	RBP4E	L-ribulose-phosphate 4-epimerase	[c]ru5p-L <==> xu5p-D	Alternate Carbon Metab	EC-5.1.3.4	120	
24	ACACCT	acetyl-CoA:acetoacetyl-CoA transfer	[c]acac + accoa --> aacoa + ac	Alternate Carbon Metabolism		129	
25	BUTCT	Acetyl-CoA:butyrate-CoA transferase	[c]accoa + but --> ac + btcoa	Alternate Carbon Metab	EC-2.8.3.8	129	
26	AB6PGH	Arbutin 6-phosphate glucohydrolase	[c]arbt6p + h2o --> g6p + hqn	Alternate Carbon Metab	EC-3.2.1.86	88	
27	PMANM	phosphomannomutase	[c]man1p <==> man6p	Alternate Carbon Metab	EC-5.4.2.8		
28	PPM2	phosphopentomutase 2 (deoxyribose	[c]2dr1p <==> 2dr5p	Alternate Carbon Metab	EC-5.4.2.7		
29	PPM	phosphopentomutase	[c]r1p <==> r5p	Alternate Carbon Metab	EC-5.4.2.7		
30	DRPA	deoxyribose-phosphate aldolase	[c]2dr5p --> acald + g3p	Alternate Carbon Metab	EC-4.1.2.4		
31	GALCTND	galactonate dehydratase	[c]galctn-D --> 2dh3dgal + h2o	Alternate Carbon Metab	EC-4.2.1.6	132	
32	DDPGALA	2-dehydro-3-deoxy-6-phosphogalacto	[c]2dh3dgal6p <==> g3p + pyr	Alternate Carbon Metab	EC-4.1.2.21	132	
33	DDGALK	2-dehydro-3-deoxygalactonokinase	[c]2dh3dgal + atp --> 2dh3dgal6p + adp + h	Alternate Carbon Metab	EC-2.7.1.58	132	
34	DHAPT	Dihydroxyacetone phosphotransferas	[c]dha + pep --> dhap + pyr	Alternate Carbon Metabolism		42,81	
35	FAO4	fatty acid oxidation (Butanoyl-CoA)	[c]btcoa + fad + h2o + nad --> aacoa + fadh2 + h + nadh	Alternate Carbon Metabolism		129	
36	ALDD19x	phenylacetaldehyde dehydrogenase	[c]h2o + nad + pacald --> (2) h + nadh + pac	Alternate Carbon Metab	EC-1.2.1.39	25,33	
37	FRUK	fructose-1-phosphate kinase	[c]atp + f1p --> adp + fdp + h	Alternate Carbon Metab	EC-2.7.1.56		
38	FCLPA	L-fuculose 1-phosphate aldolase	[c]fc1p <==> dhap + lald-L	Alternate Carbon Metab	EC-4.1.2.17		
39	FCI	L-fucose isomerase	[c]fuc-L <==> fcl-L	Alternate Carbon Metab	EC-5.3.1.25		
40	FCLK	L-fuculokinase	[c]atp + fcl-L --> adp + fc1p + h	Alternate Carbon Metab	EC-2.7.1.51		

Reactions | Metab Abbr | Exchange Fluxes | Dead End Metab | RemovedRxns | Gene Annotations

Ready

Reed,J.L et al. (2003) Genome Biology, **4,** R54.

METABOLIC MODELS – SBML FORMAT

Notes

Reactants

Products

Kinetic Law

```
<reaction id="R_PYK" name="R_pyruvate_kinase" reversible="false">
    <notes>
        <html:p>GENE_ASSOCIATION: ( b1854  or  b1676 )</html:p>
        <html:p>PROTEIN_ASSOCIATION: ( Pyka ) or ( Pykf )</html:p>
        <html:p>SUBSYSTEM: S_GlycolysisGluconeogenesis</html:p>
        <html:p>PROTEIN_CLASS: 2.7.1.40</html:p>
    </notes>
    <listOfReactants>
        <speciesReference species="M_adp_c" stoichiometry="1.000000"/>
        <speciesReference species="M_h_c" stoichiometry="1.000000"/>
        <speciesReference species="M_pep_c" stoichiometry="1.000000"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="M_atp_c" stoichiometry="1.000000"/>
        <speciesReference species="M_pyr_c" stoichiometry="1.000000"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply>
                <ci> LOWER_BOUND </ci>
                <ci> UPPER_BOUND </ci>
                <ci> OBJECTIVE_COEFFICIENT </ci>
                <ci> FLUX_VALUE </ci>
                <ci> REDUCED_COST </ci>
            </apply>
        </math>
        <listOfParameters>
            <parameter id="LOWER_BOUND" value="0.000000" units="mmol_per_gDW_per_hr"/>
            <parameter id="UPPER_BOUND" value="999999.000000" units="mmol_per_gDW_per_hr"/>
            <parameter id="OBJECTIVE_COEFFICIENT" value="0.000000"/>
            <parameter id="FLUX_VALUE" value="0.000000" units="mmol_per_gDW_per_hr"/>
            <parameter id="REDUCED_COST" value="0.000000"/>
        </listOfParameters>
    </kineticLaw>
</reaction>
```

Merlin

metabolic models reconstruction using genome-scale information

Developed by

University of Minho
School of Engineering

CENTRE OF BIOLOGICAL ENGINEERING

www.merlin-sysbio.org

Main features

- Supports the main tasks in metabolic genome **re-annotation** and **reaction assignment**
- Supports compartmentalization and transporters identification
- Allows manual curation of the model through a user friendly environment

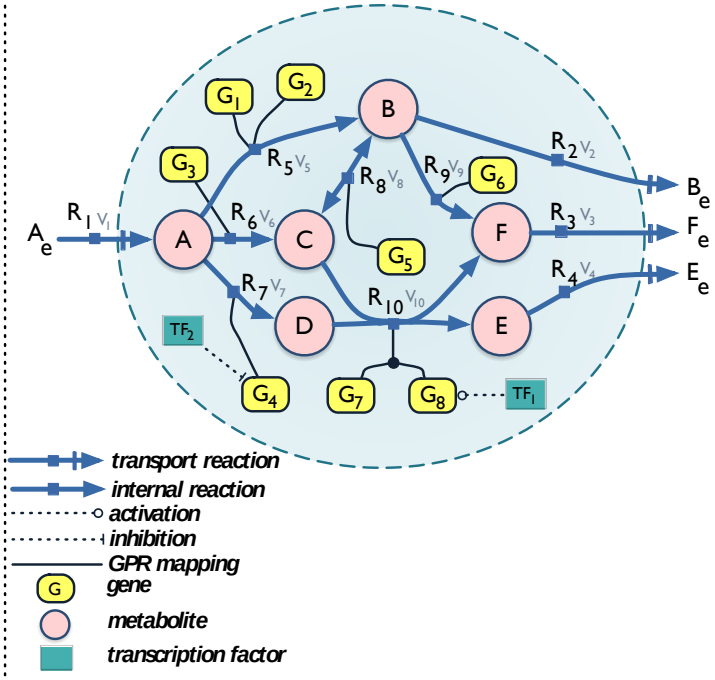
Dias, O., Rocha, M., Ferreira, E.C. and Rocha, I. (2015)
Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Res.

http://nar.oxfordjournals.org/content/early/2015/04/06/nar.gkv294

GSMMs| Model information

a)

Metabolic network (w/ transcriptional regulation)



Reactions

R_1 : $A_e \xrightarrow{V_1} A$

R_2 : $B \xrightarrow{V_2} B_e$

R_3 : $F \xrightarrow{V_3} F_e$

R_4 : $E \xrightarrow{V_4} E_e$

R_5 : $A \xrightarrow{V_5} B$

R_6 : $A \xrightarrow{V_6} C$

R_7 : $A \xrightarrow{V_7} D$

R_8 : $B \xleftrightarrow{V_8} C$

R_9 : $B \xrightarrow{V_9} F$

R_{10} : $C+D \xrightarrow{V_{10}} E+F$

Flux limits

$0 \le V_1 \le a$

$0 \le V_2 \le \infty$

$0 \le V_3 \le \infty$

$0 \le V_4 \le \infty$

$0 \le V_5 \le \infty$

$0 \le V_6 \le \infty$

$0 \le V_7 \le \infty$

$-\infty \le V_8 \le \infty$

$0 \le V_9 \le \infty$

$0 \le V_{10} \le \infty$

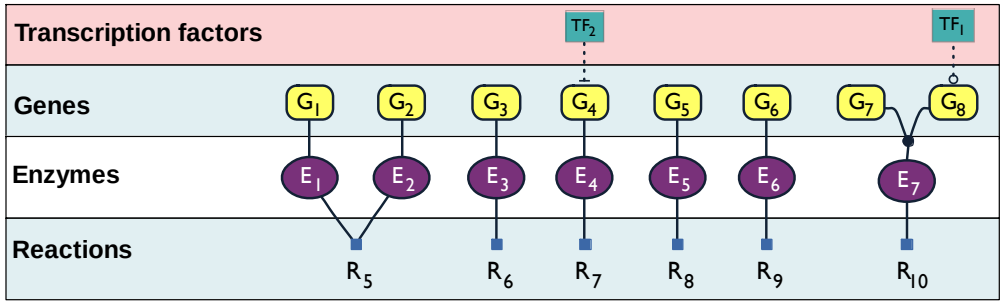
b)

Stoichiometric matrix — Fluxes

$$
\begin{array}{c}
 \\ A \\ B \\ C \\ D \\ E \\ F
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 & R_8 & R_9 & R_{10} \end{array} \\
\left[\begin{array}{cccc|cccccc}
1 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}\right]
\end{array}
* \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \\ V_5 \\ V_6 \\ V_7 \\ V_8 \\ V_9 \\ V_{10} \end{bmatrix}
= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
$$

S (m,n) v(n) 0(m)

c)



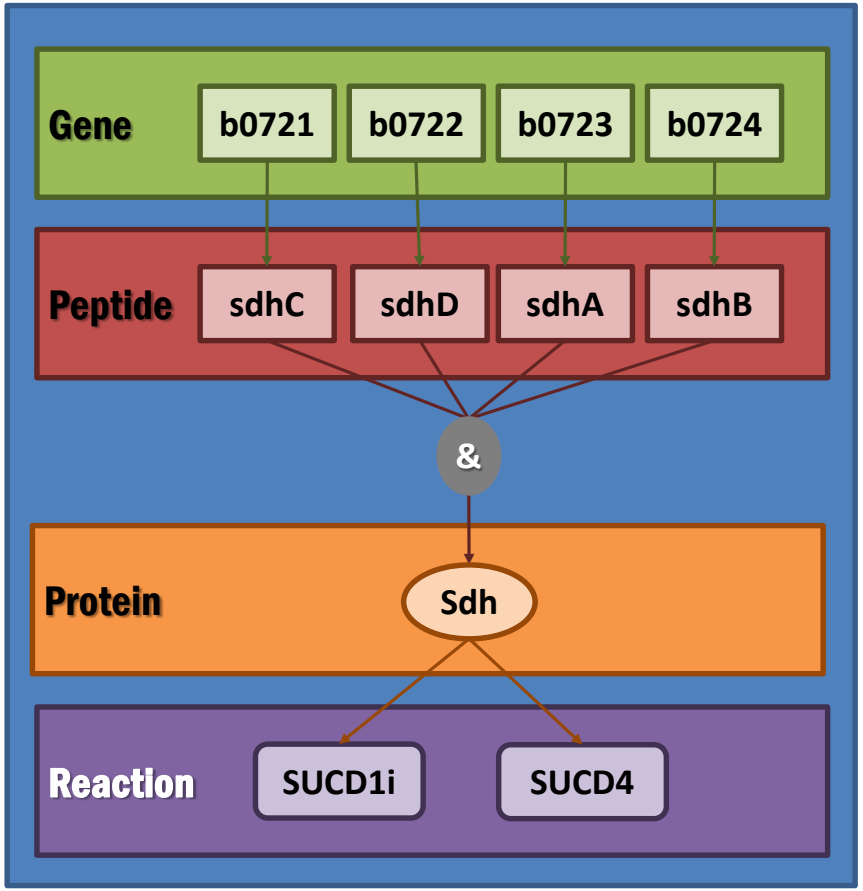
Reaction	GPR
R_5	G_1 or G_2
R_6	G_3
R_7	G_4
R_8	G_5
R_9	G_6
R_{10}	G_7 and G_8

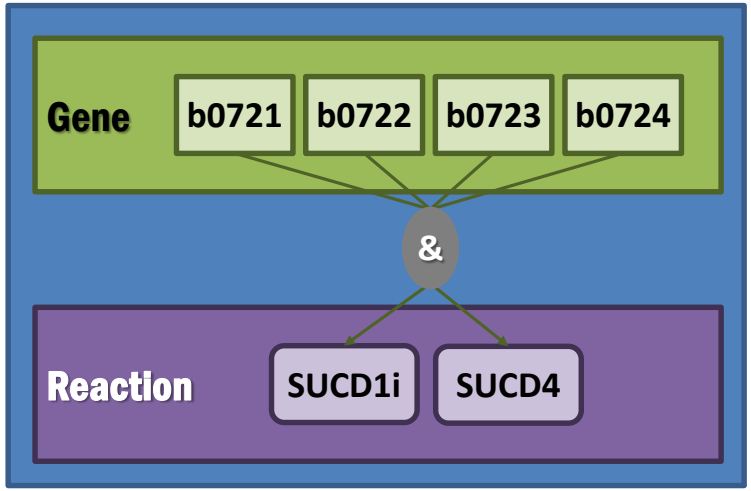
Gene	Regulatory rule
G_4	not TF_2
G_8	TF_1

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

GSMMs| Gene Protein Reaction rules





Gene-reaction rules:

SUCD1i = b0721 AND b0722 AND b0723 AND b0724

SUCD4 = b0721 AND b0722 AND b0723 AND b0724

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

Metabolic Models – stoichiometric models & CBM

- Stoichiometric models typically have more fluxes than balanced metabolites.
- The equation system S • v = 0 thus has more variables than equations. This is a so-called <u>under-determined equation system</u> with infinitely many solutions:

Under-determined system	$a_{11}x_1+a_{12}x_2=b_1$
Determined system	$a_{21}x_1+a_{22}x_2=b_2$
Over-determined system	$a_{31}x_1+a_{32}x_2=b_3$

How do we deal with underdetermination?

Experimental approaches

Generation of additional constraints from:

- measurement of exchange fluxes (MFA)
- experiments with labeled substrates (MNA)

Computational or *in silico* approaches

- adding assumptions, e.g. objective function (FBA)
- Enumeration of all possible solutions (Elementary modes)

METABOLIC MODELS – MFA EXAMPLE

Metabolic Flux Analysis

- Example

Before:

F= #fluxes – #metabolites ⇔ F = 6 – 3 **= 3**

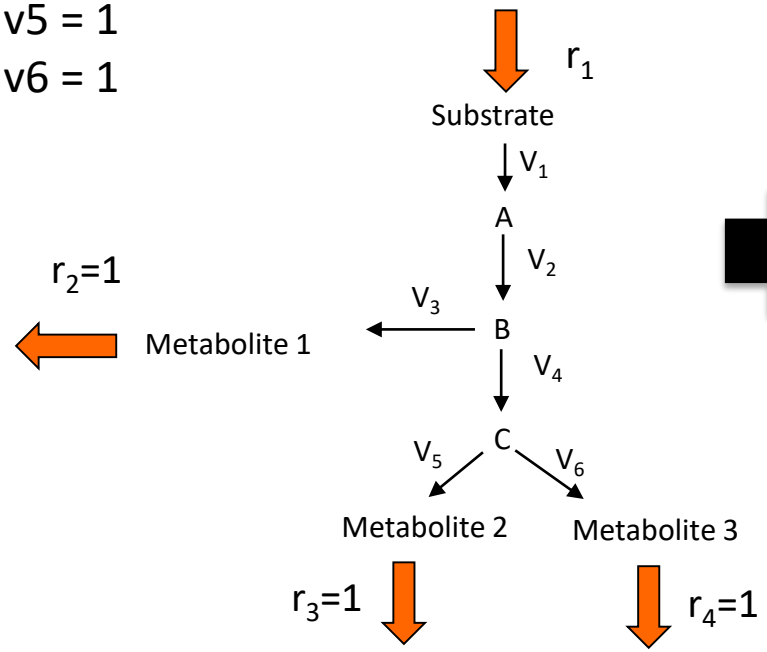
Under-determined!

If we measure 3 exchange fluxes

v3 = 1

v5 = 1

v6 = 1



A: v1 - v2 = 0

B: v2 – v3 – v4 = 0 ⇔

C: v4 – v5 –v6 =0

v1 - v2 = 0

v2 – 1 – v4 = 0 ⇔

v4 – 1 – 1 =0

v1 = v2

v2 = v4 +1 ⇔

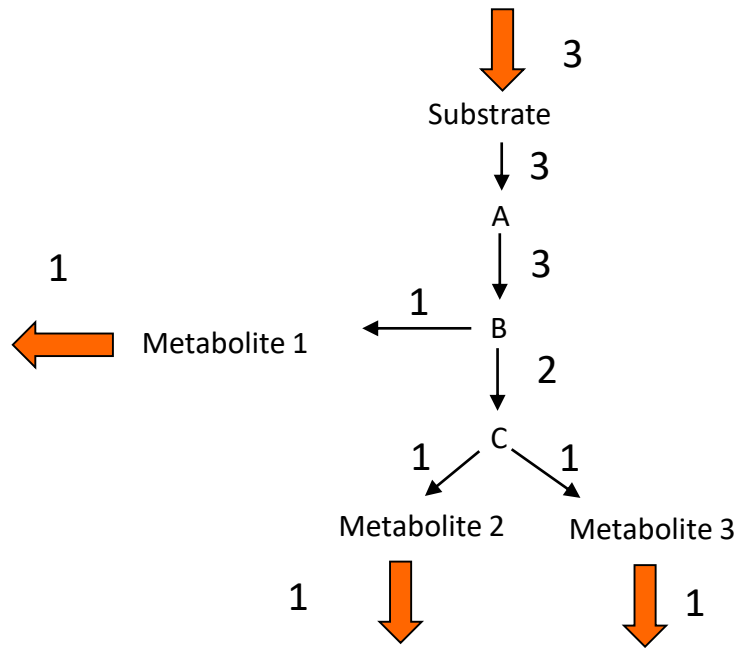
v4 = 2

v1 = 3

v2 = 3

v4 = 2

Solution:



Optimization problem

- The system is undetermined – one solution is to transform into an **optimization problem**
- **Flux Balance Analysis**: assumes organisms have evolved "perfectly" to maximize a given objective function with a biological rationale
- **Objective function**: most common – to maximize biomass flux – artificial flux determined experimentally including all biomass precursors
- Linear OF; linear constraints – **Linear Programming** problem
- Easy to solve (e.g. simplex algorithm)
- Methods for mutant simulation adopt different objective function: MOMA, ROOM

PHENOTYPE PREDICTION – FBA PROBLEM

Maximize:

$$z = c^T v = v_{prod}$$

c = row vector containing weights specifying what combination of fluxes to optimize

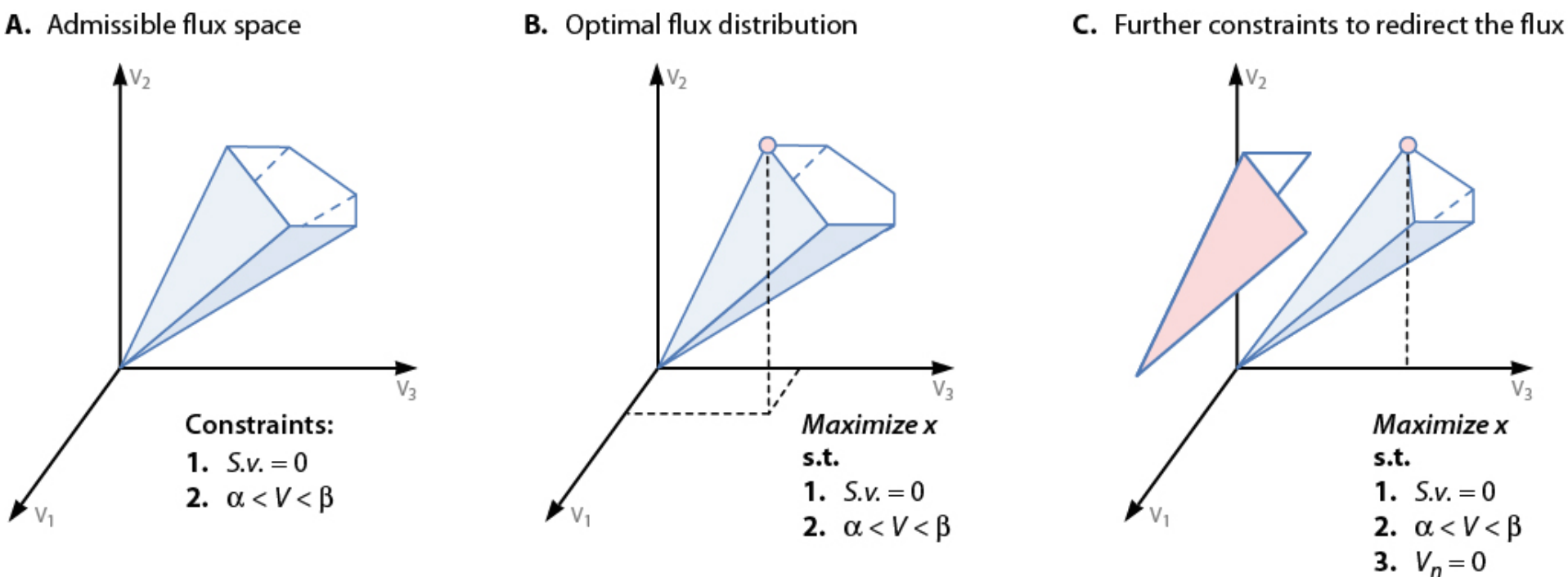
Subject to:

$$S\,v = 0$$

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

α, β = lower and upper limits for fluxes. Use

- to model irreversible reactions
- to limit uptake and secretion rates
- to specify measured fluxes



PHENOTYPE PREDICTION – PARSIMONIOUS FBA

- FBA has an important limitation, since it provides a solution with a unique optimal value for the objective function, while a large number of flux distributions may exist that lead to this value, i.e. **multiple optima may exist.**

- One way to address this issue was proposed by the **Parsimonious enzyme usage FBA (pFBA)** method that chooses a particular flux distribution (or a smaller set of flux distributions) from these multiple optima, by performing a second LP optimization that minimizes the sum of the flux values, while keeping biomass flux at an optimum level.

PHENOTYPE PREDICTION – OBJECTIVE FUNCTIONS

- Studies in several organisms demonstrated that their metabolic network has evolved for optimization of the specific growth rate under several carbon source limiting conditions.

- Thus, for simulating cellular behavior the most common objective function is the maximization of biomass production.

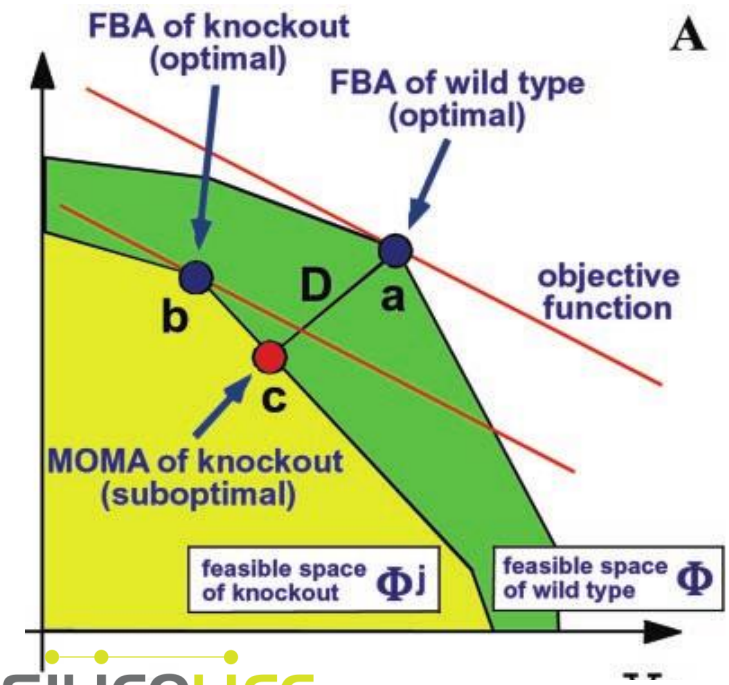
- However, it has been shown that for mutants and wild-type organisms grown on some unusual carbon sources the hypothesis of optimal growth is not always real.

- Growth of these microorganisms is better explained through the hypothesis that such strains undergo minimal redistribution of fluxes with respect to the wild-type strains.

PHENOTYPE PREDICTION – MUTANTS: MOMA

- Minimization of Metabolic Adjustment (MOMA) is a flux-based analysis technique similar to FBA and based on the same stoichiometric constraints, <u>but the optimal growth flux for mutants is relaxed</u>.

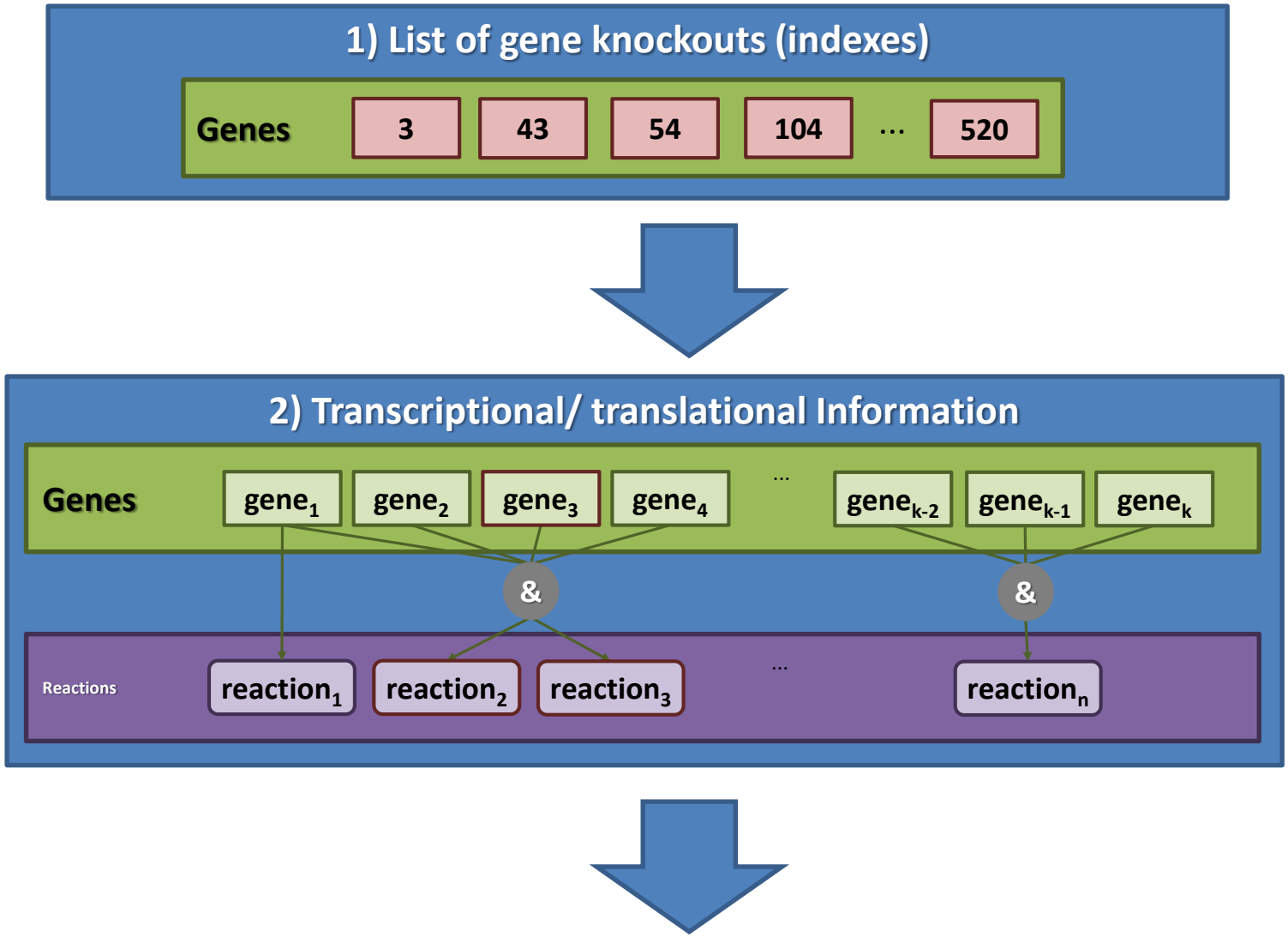
- Instead, MOMA provides an approximate solution for a sub-optimal growth flux state, which is nearest in flux distribution to the unperturbed state.



Formulated as a Quadratic Programming problem:

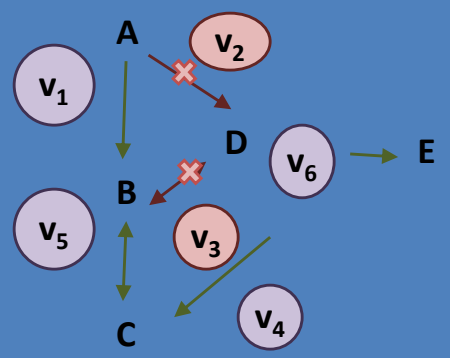
$$\min \ ||\mathbf{v_w} - \mathbf{v_d}||^2 \qquad s.t. \quad \mathbf{S} \cdot \mathbf{v_d} = 0$$

Mutant simulation with Gene reaction rules



Mutant simulation with Gene reaction rules

3) Metabolic Model Modification



	v_1	v_2	v_3	v_4	v_6	...	v_n
A	-1	-1	0	0	0	..	?
B	1	0	-1	0	0	..	?
C	0	0	0	1	0	..	?
D	0	1	1	-1	-1	..	?
...	...	...	...	...	...	..	...
Met_m	?	?	?	?	?	...	?

0	$\leq v_1 \leq$	∞
0	$\leq v_2 \leq$	0
0	$\leq v_3 \leq$	0
0	$\leq v_4 \leq$	∞
$-\infty$	$\leq v_5 \leq$	∞
0	$\leq v_6 \leq$	∞
	...	

4) Mutant Phenotype Simulation

FBA
MOMA
ROOM

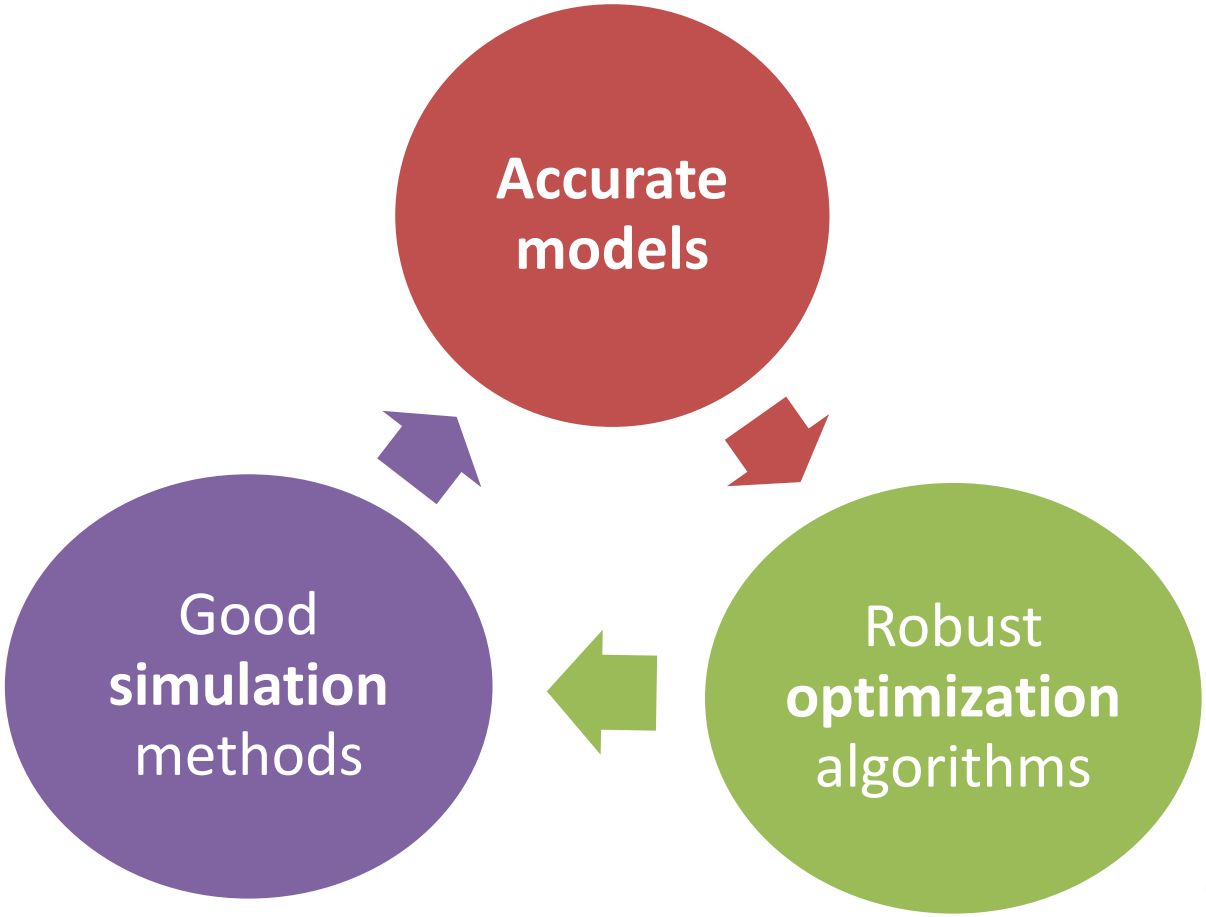
METABOLIC ENGINEERING

To produce **desired compounds** (e.g. antibiotics, fuels, vitamins) from **microbial cell factories** it is generally necessary to **retrofit the metabolism**

Metabolic Engineering envisages the introduction of **directed genetic modifications** leading to desirable phenotypes, as opposed to traditional methods

METABOLIC ENGINEERING

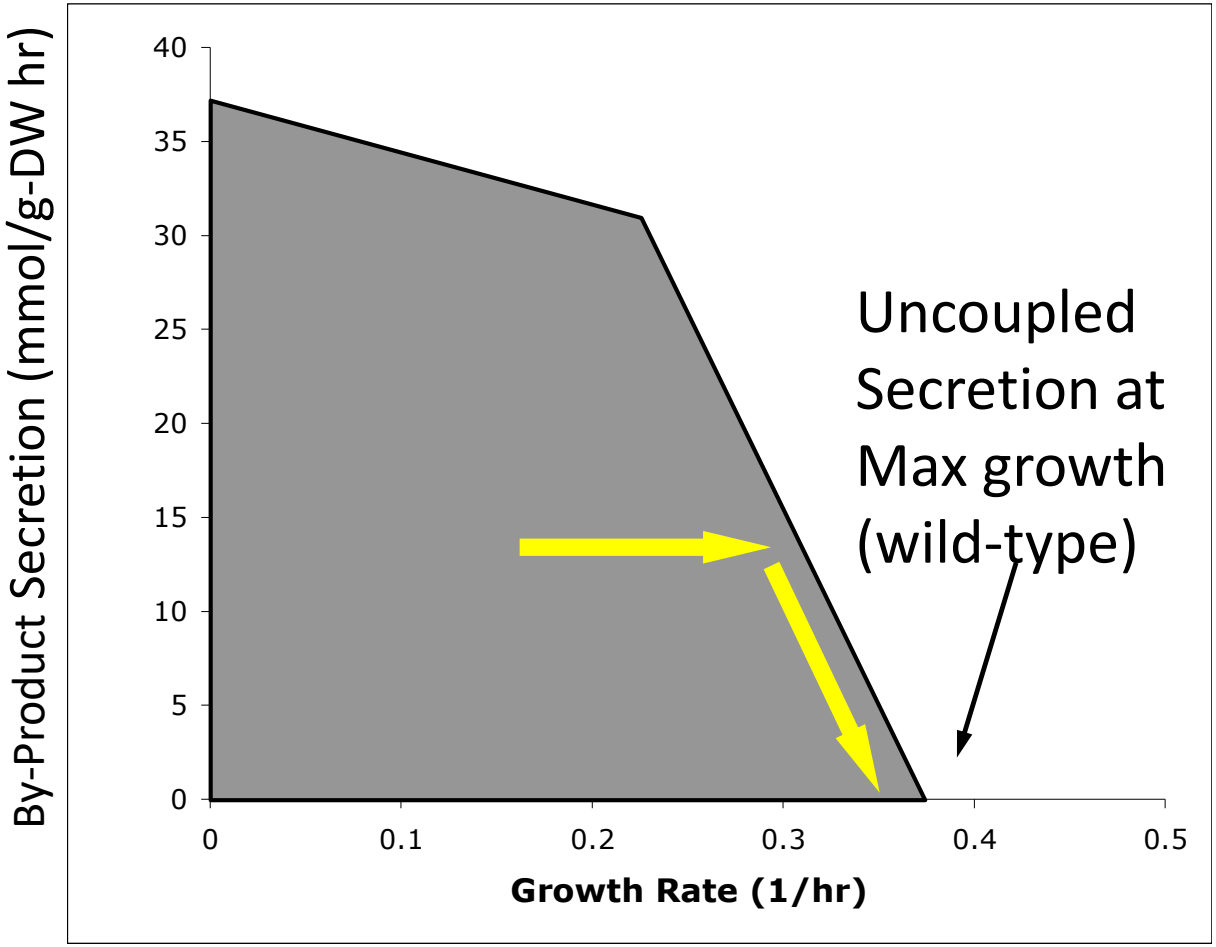
It is often difficult to identify which genetic manipulations will originate a given desired phenotype



SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

STRAIN OPTIMIZATION: WHY ?



No production of desired compound in wild type strains !

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

Strain Optimization

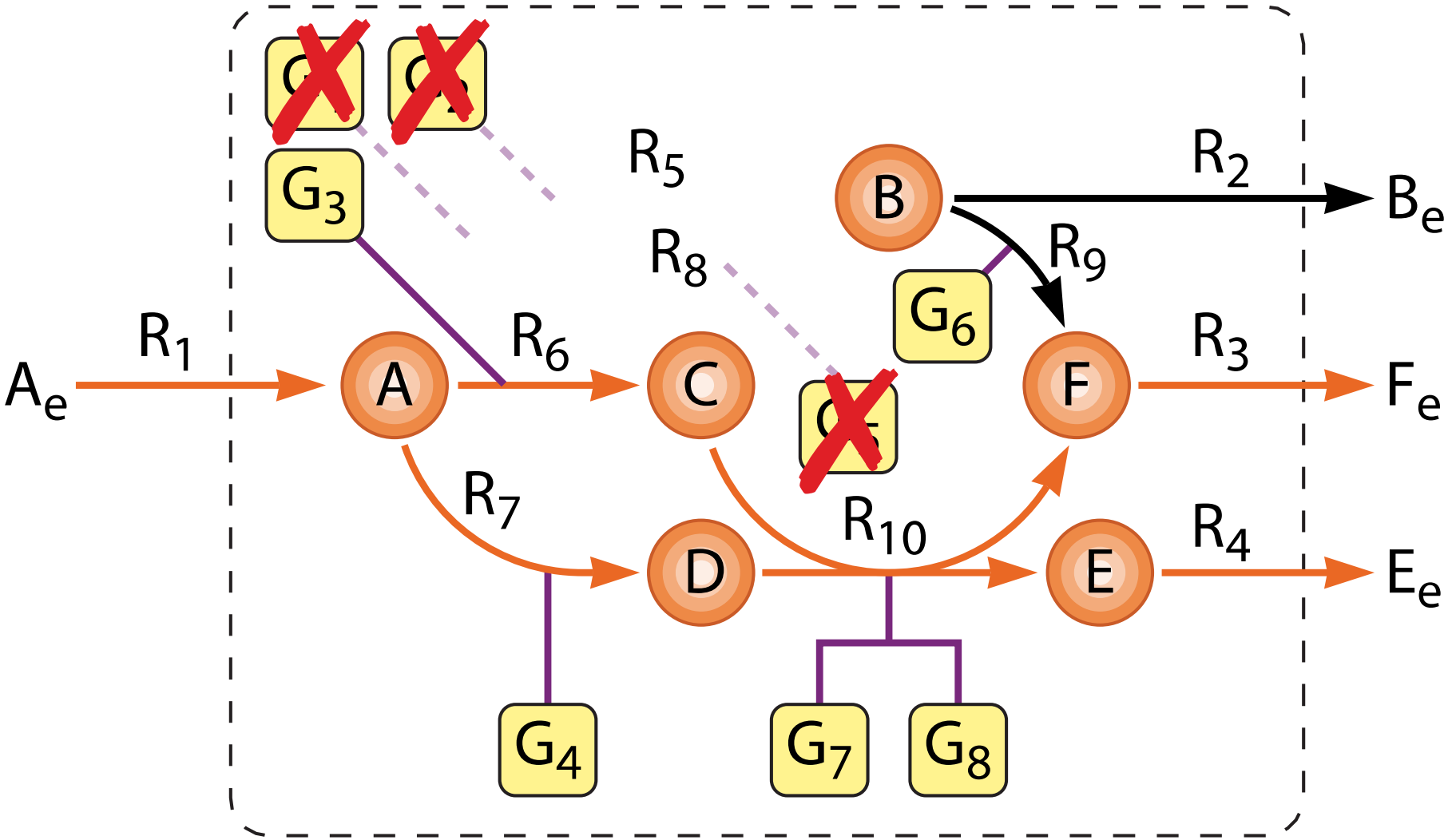
Possible aims

- Select appropriate gene/ reaction deletions
- Select genes to over/under express
- Select set of reactions to add to a metabolic model

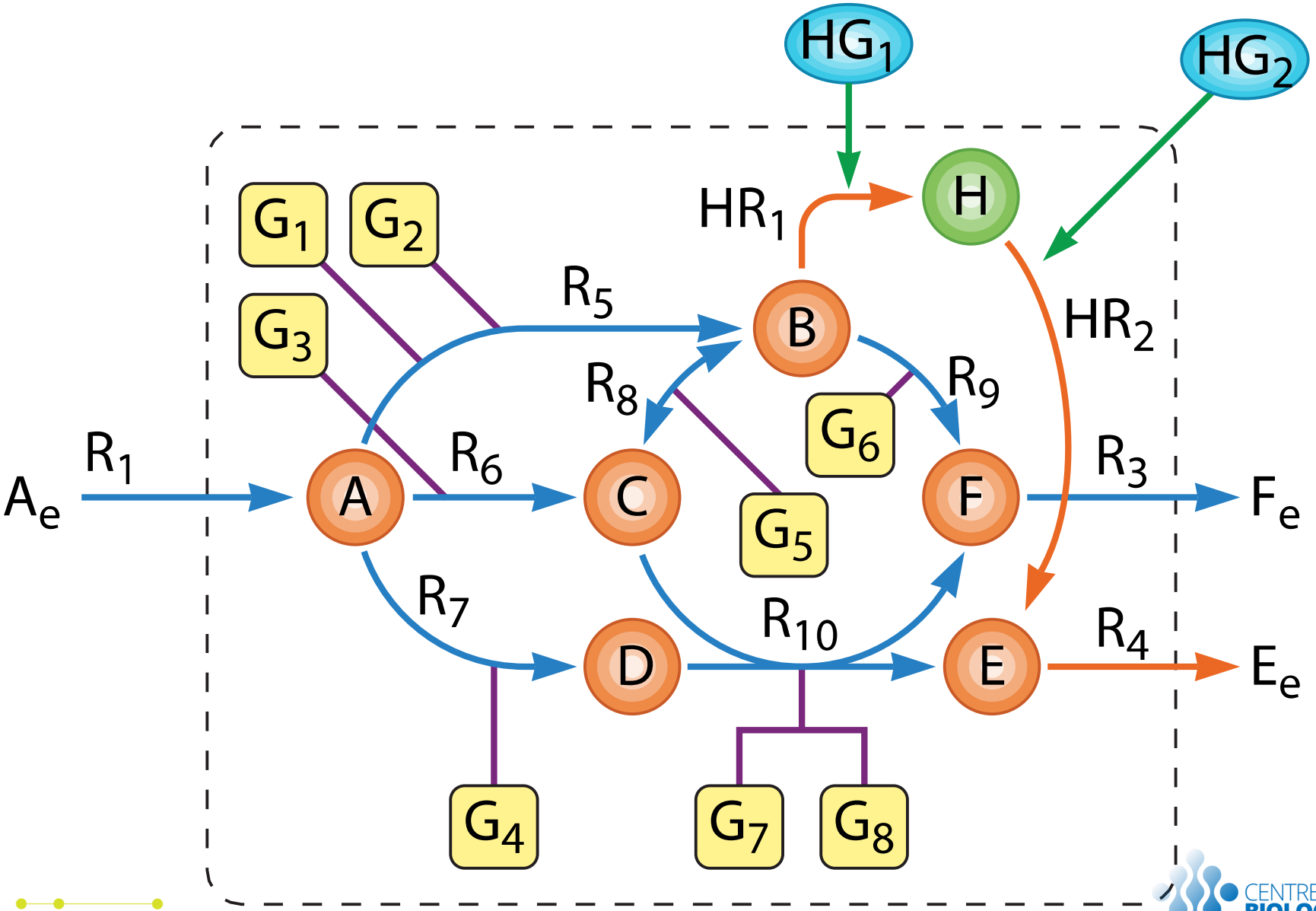
Objective function

- Maximizing the production of a compound
- Keeping the organism viable
- One alternative - **Biomass product coupled yield (BPCY)**: multiplies biomass and compound production fluxes and divides by substrate intake flux

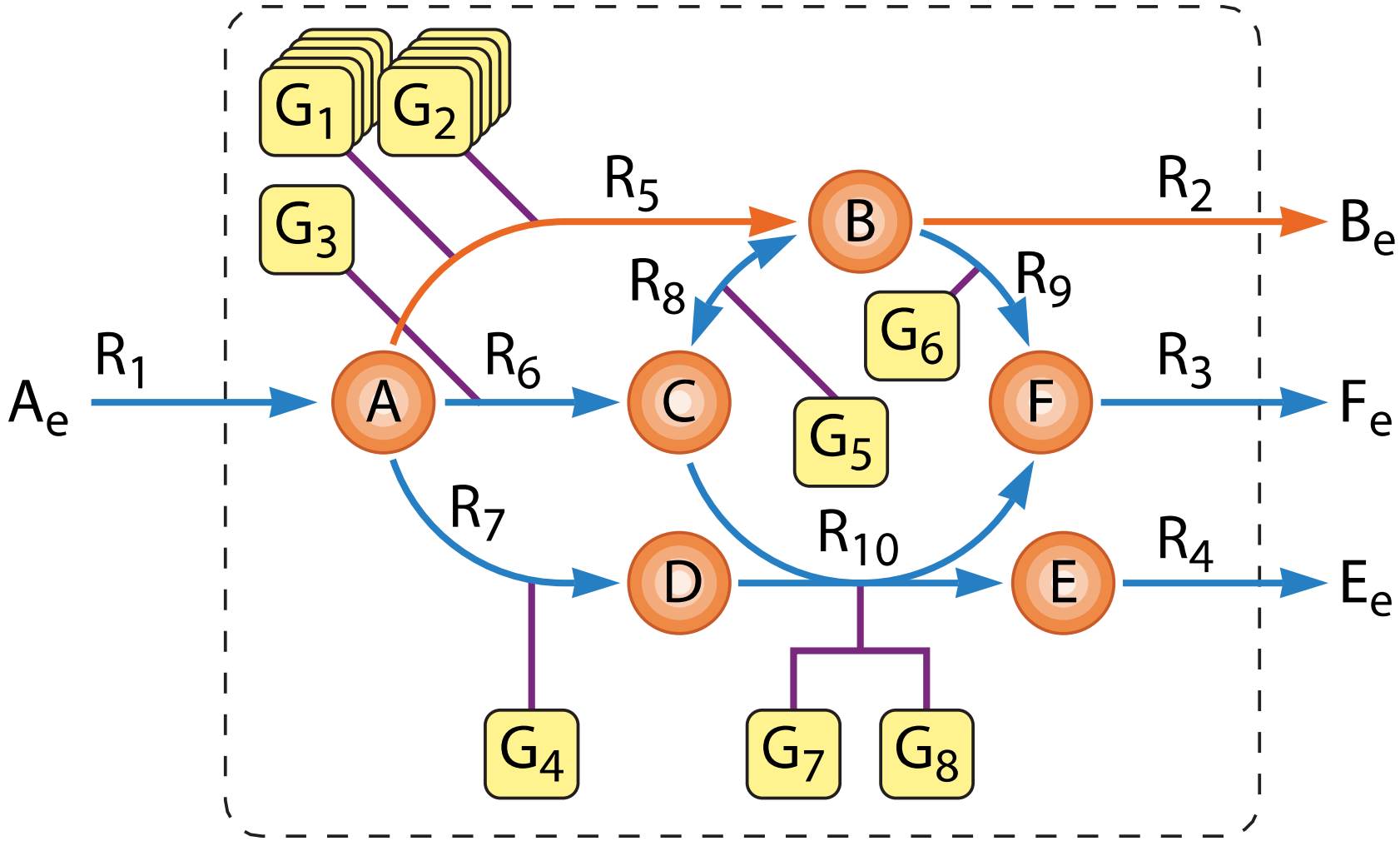
Strain Optimization – Gene deletion



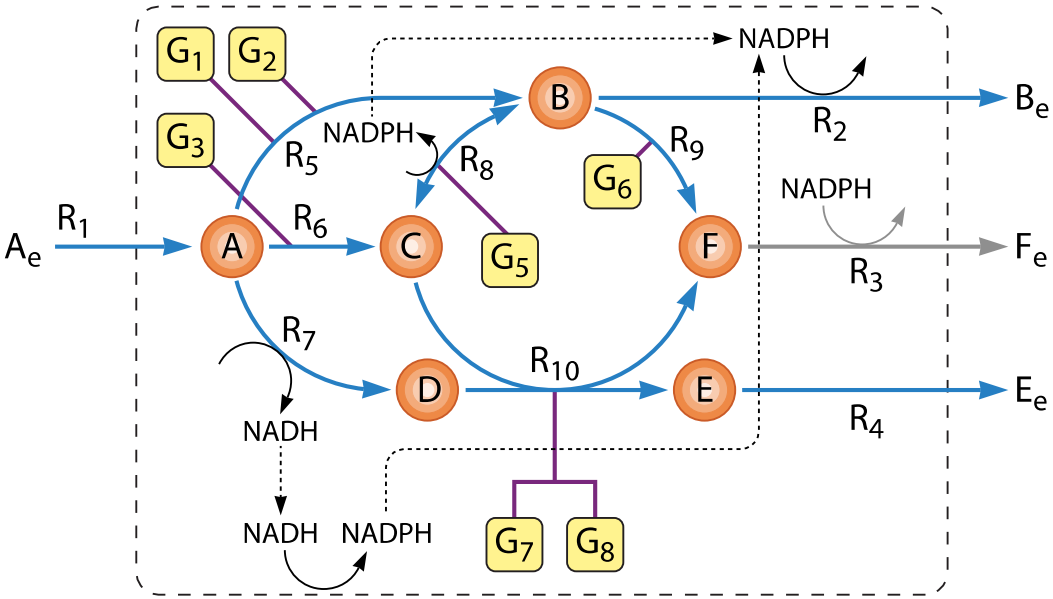
Strain Optimization – Heterologous insertion

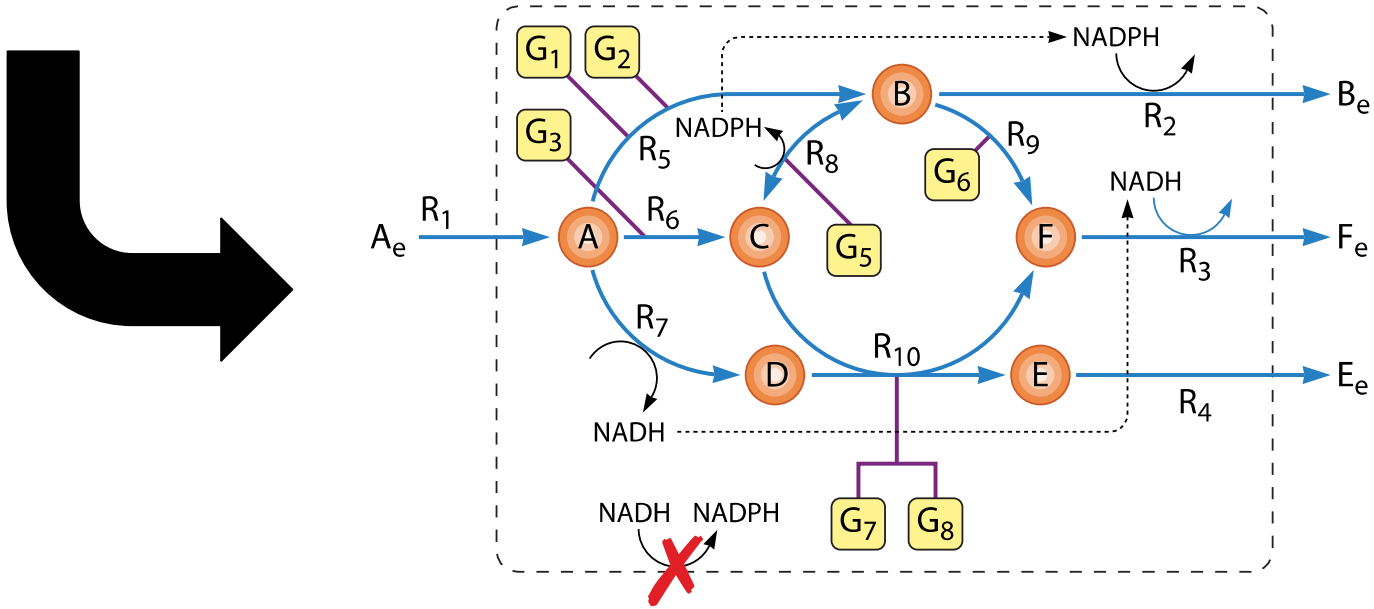


Strain optimization – gene over/under expression



Strain Optimization: co-factor swapping





STRAIN OPTIMIZATION: BILEVEL OPTIMIZATION APPROACH

Aim: select the appropriate genetic modifications (e.g. gene knockout sets) to enable the production of a desired product

Bi-level optimization problem:

Maximize (compound) – bioengineering objective

Candidate solution evaluation

Maximize (biomass) – cellular objective

constraints:
- steady state
- reversibility
(...)

STRAIN OPTIMIZATION – HEURISTIC ALGORITHMS

Combinatorial optimization

- Solutions can be represented as sets of genes or reactions

Complexity

- Solution space is very large; problems are NP – hard
- Approaches based on MILP (e.g. OptKnock) cannot handle non-linear objective functions or multiple objectives
- Solution: use of metaheuristic optimization methods such as **Evolutionary Algorithms** and **Simulated Annealing**

Strain Optimization: Decoupled optimization

Strain Optimization

- ✓ Evolutionary algorithms
- ✓ Simulated Annealing
- ✓ Local optimization

Objective function: Maximize the production of a given metabolite

Phenotype Simulation

- ✓ FBA, MOMA, ROOM
- ✓ Boolean net simulation
- ✓ ODEs (numerical integration)

Determine the phenotype (e.g. set of fluxes in steady state)

Environmental Conditions

Genetic conditions (e.g. knockouts)

Override Model

Model

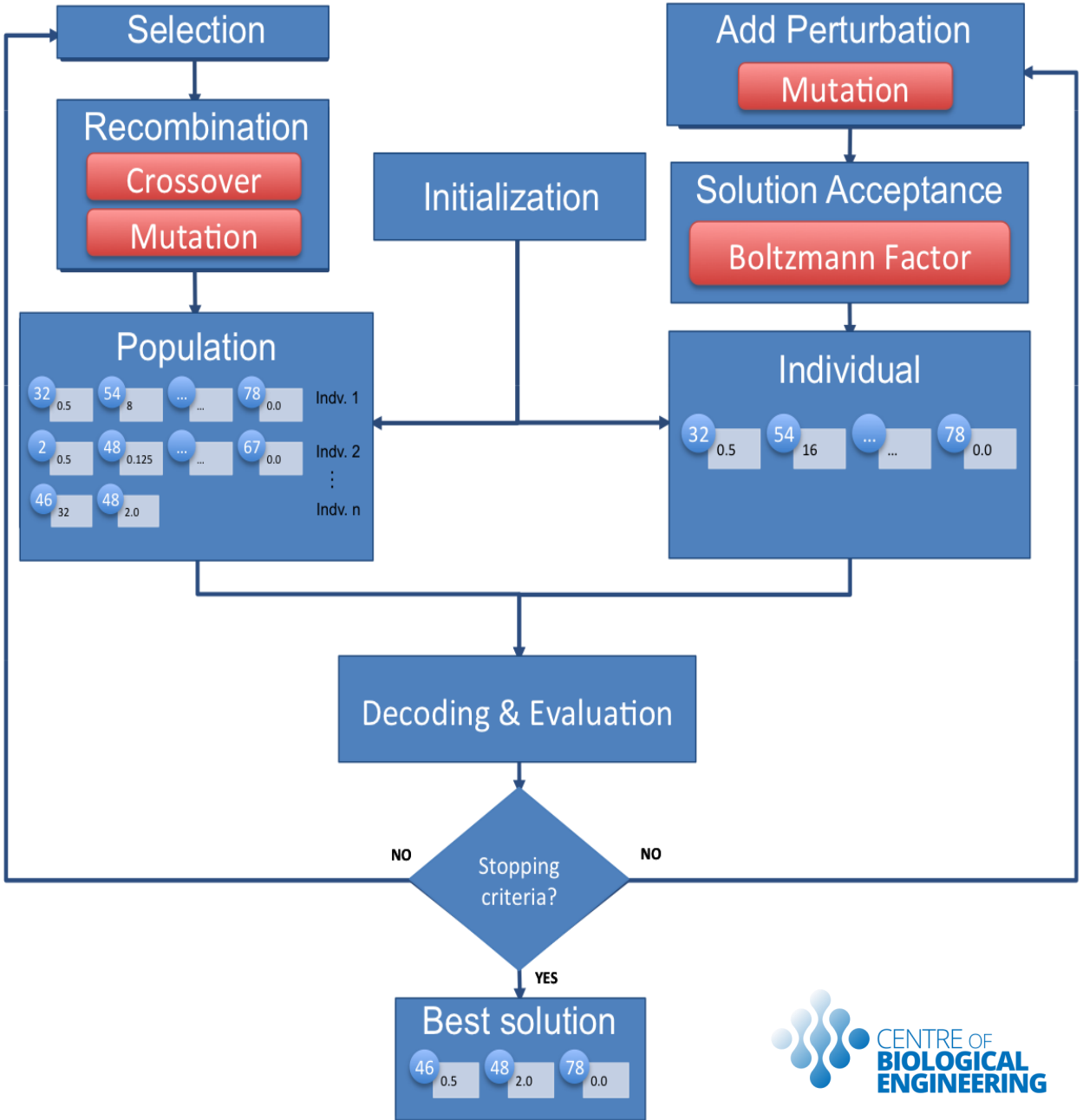
- ✓Stoichiometric
- ✓Regulatory
- ✓Dynamic

Original Model

STRAIN OPTIMIZATION – ALGORITHMS

Evolutionary Algorithms (EA)

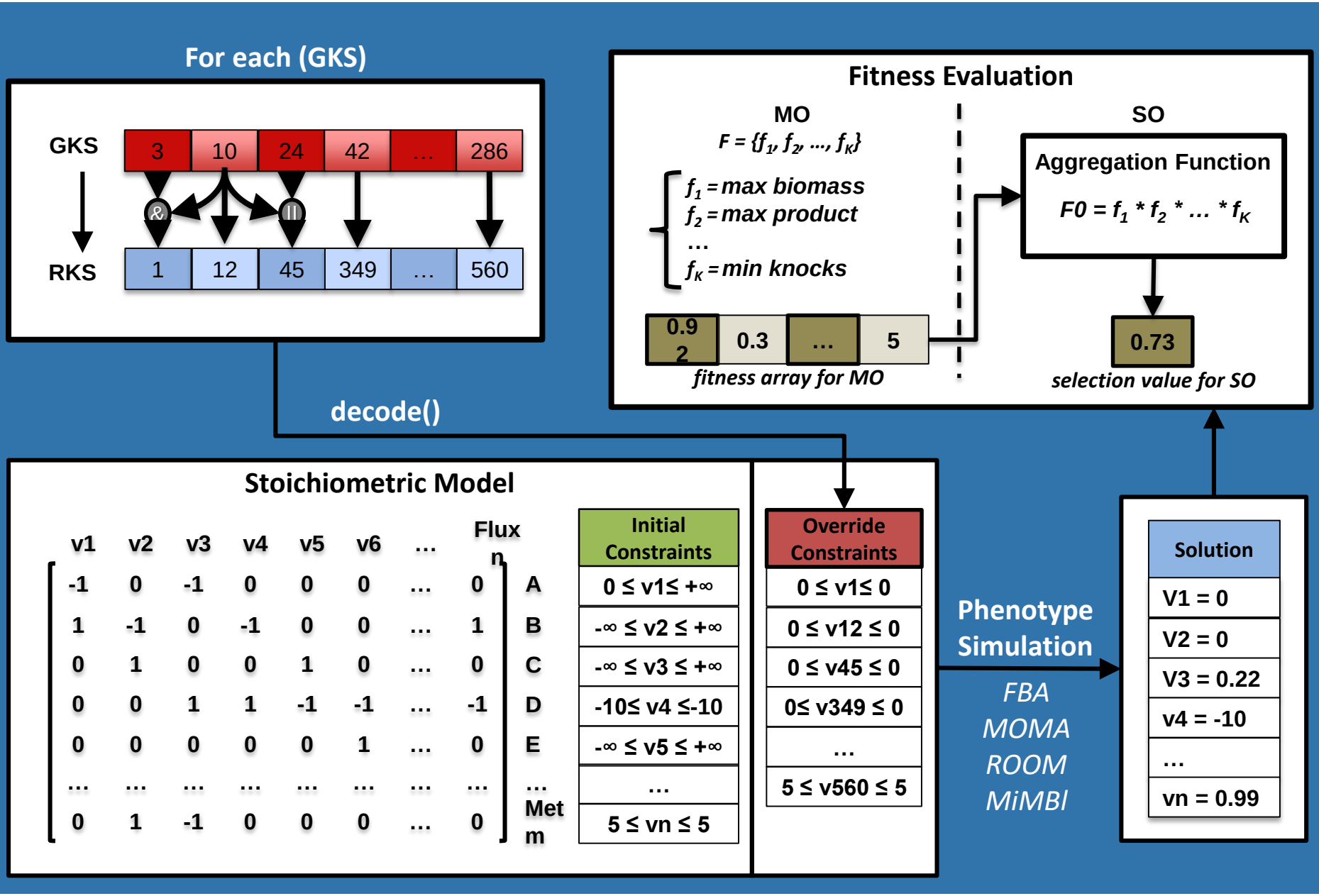
Simulated Annealing (SA)



SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

STRAIN OPTIMIZATION – ALGORITHMS



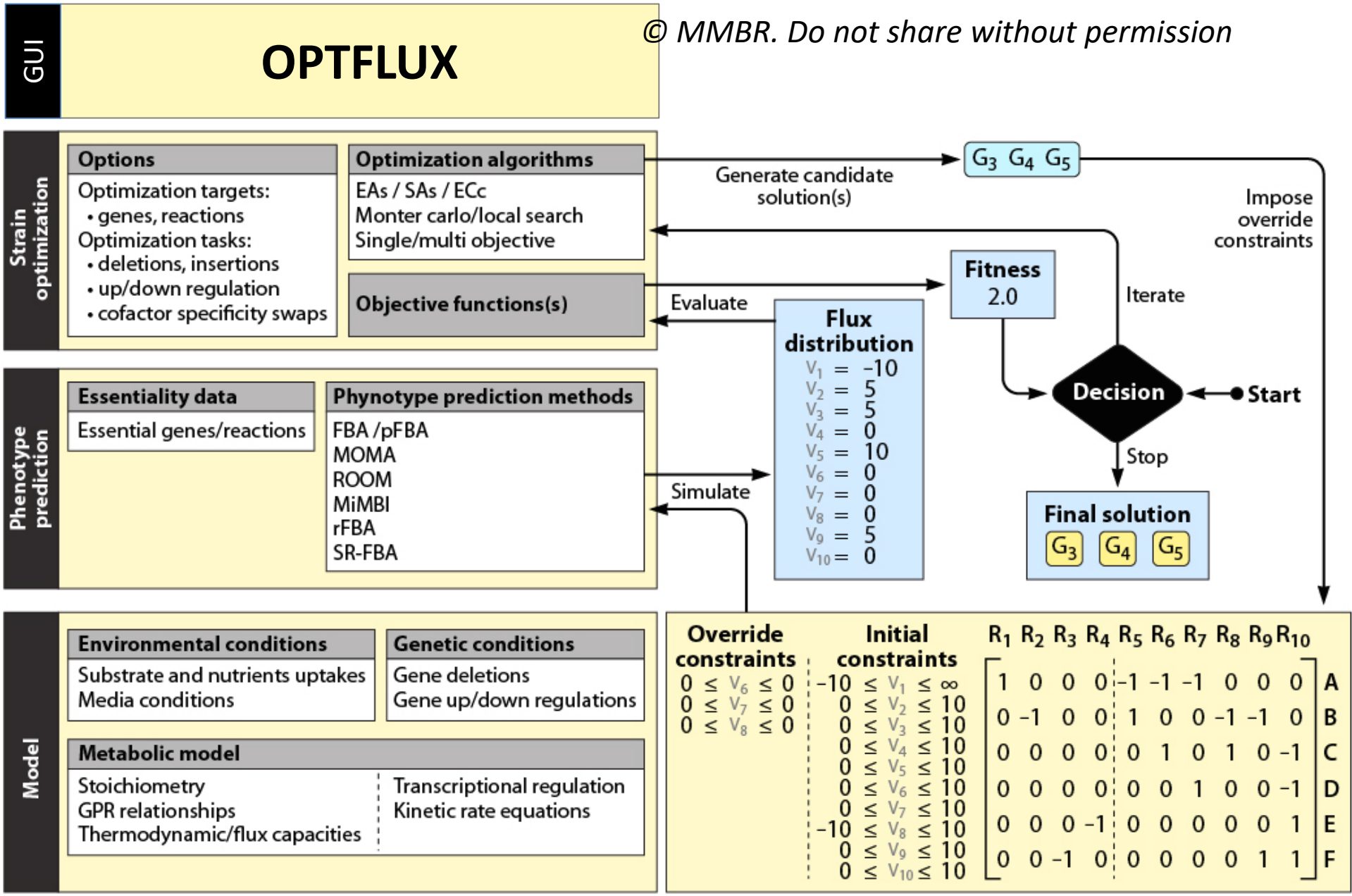
Strain optimization – multiobjective

Some of the aims of **strain optimization algorithms** are to find strains that, e.g. (i) produce the compound of interest; (ii) are biologically viable; (iii) have minimal changes from wild type

In **optimization** these are distinct **aims**: (i) max **compound** producing flux; (ii) max **biomass** flux; (iiI) min number of deletions

It makes sense to have **multiobjective algorithms** that compute solutions that are different **trade-offs** of (all or part of) these objectives: result is a set of solutions.

STRAIN OPTIMIZATION – HEURISTIC APPROACH



Points For Discussion

- CBM approaches have provided a robust framework to perform large-scale simulation and optimization of microbes

- However, CBM approaches have limitations mainly in simulating non-steady state conditions, metabolite concentrations and absolute flux values

- Large-scale dynamic models are still only available for very few organisms and require further validation

- Novel methodologies for both modeling and data collection for model inference are required

Genome-scale modeling

Isabel Rocha irocha@deb.uminho.pt

September 12, 2017
DSM, Delft, The Netherlands

Credits for some slides: Paulo Maia (SilicoLife)

CENTRE OF BIOLOGICAL ENGINEERING

SILICOLIFE