Computational Approaches for Metabolic and Enzyme Engineering

Isabel Rocha irocha@itqb.unl.pt

Credits for some slides: Paulo Maia (SilicoLife); Cláudio Soares (ITQB)

September 18, 2018
ITQB, Oeiras, Portugal

CENTRE OF BIOLOGICAL ENGINEERING

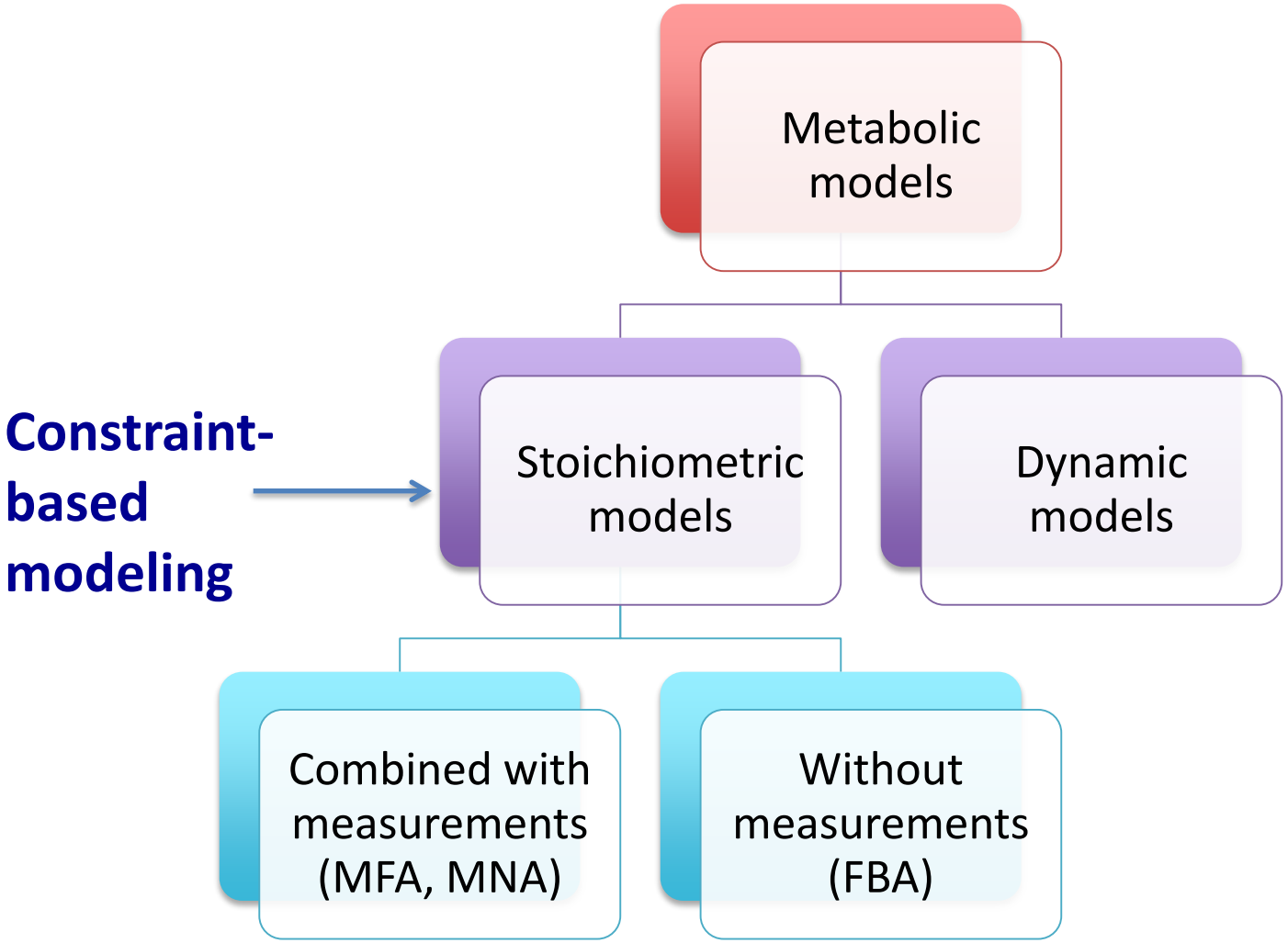
INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA /UNL

SILICOLIFE

OUTLINE

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

METABOLIC MODELS



MASS BALANCES

Framework for both dynamic and stoichiometric models:

Mass balance over intra-cellular metabolites

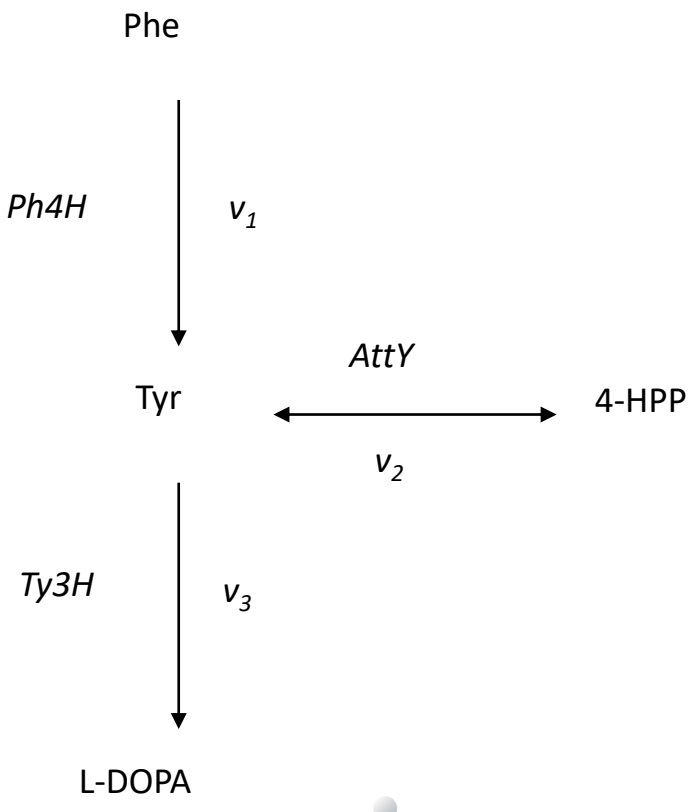
$$\frac{d[Tyr]}{dt} = v_1 - v_2 - v_3 - \mu[Tyr]$$

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:

Linear equation system described by stoichiometric matrix *S.*

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

METABOLIC MODELS

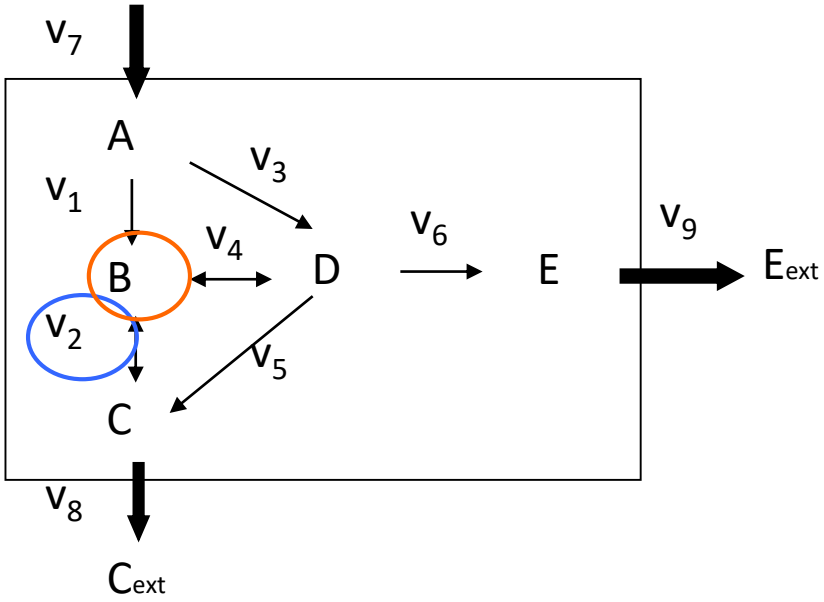
Stoichiometric models

- 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 – 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

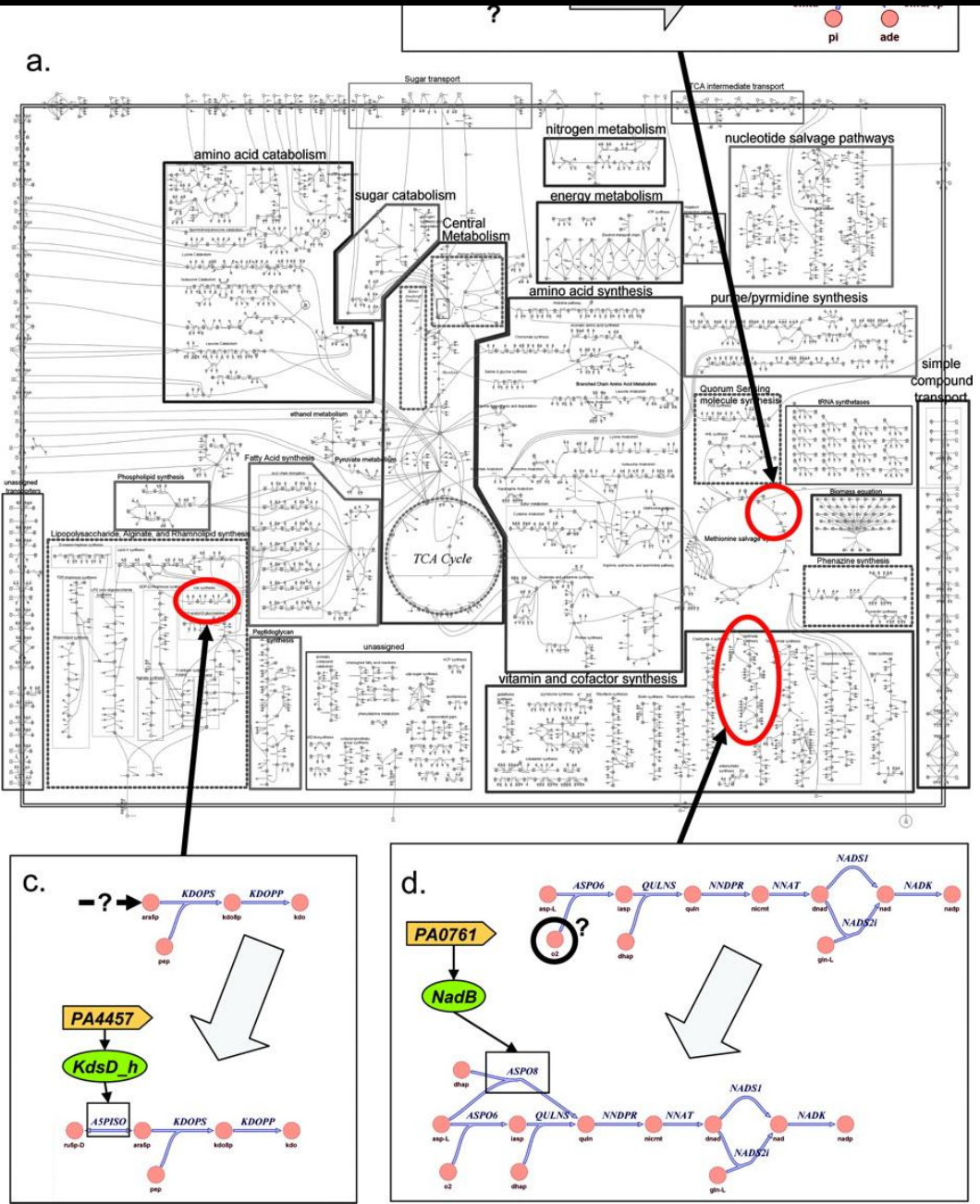
Metabolites

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

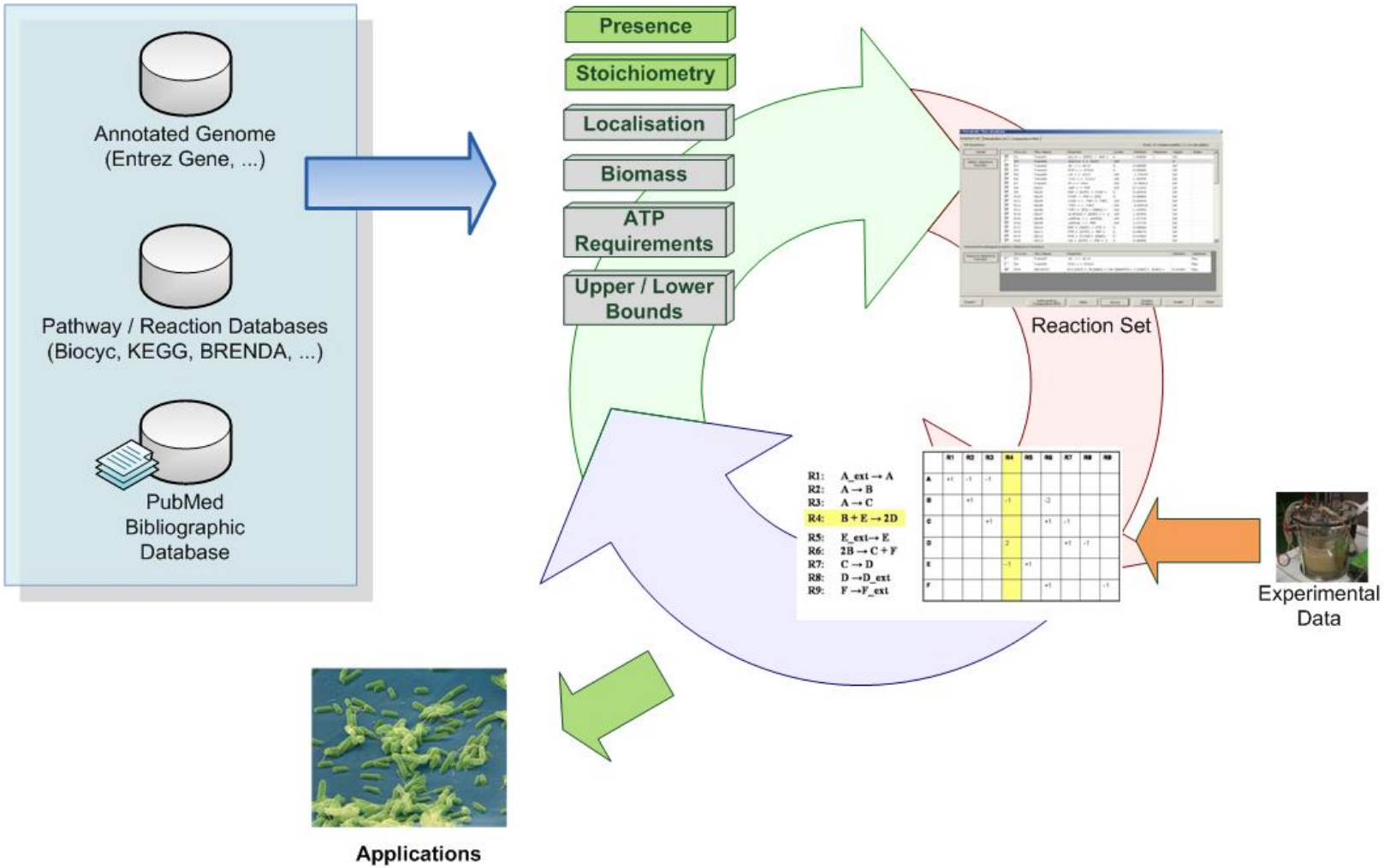
GENOME SCALE METABOLIC MODELS (GSMMS)

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



GSMMs| RECONSTRUCTION - METHODOLOGY



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

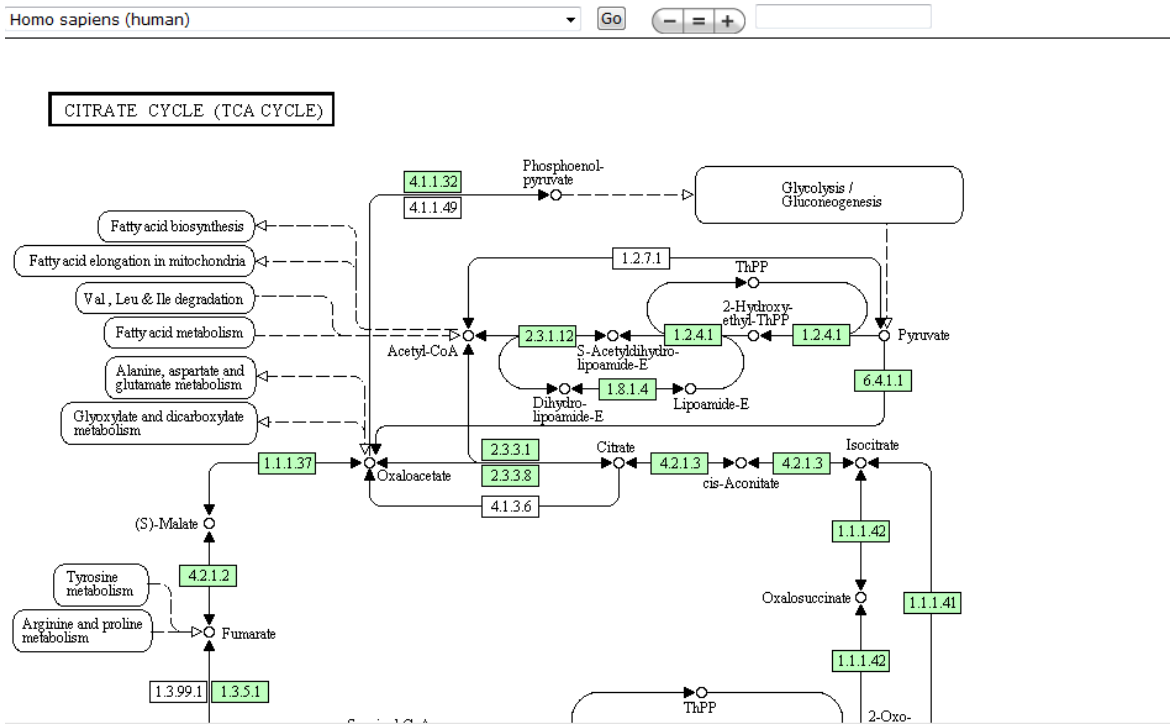
GSMMs| RECONSTRUCTION – DATA SOURCES

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



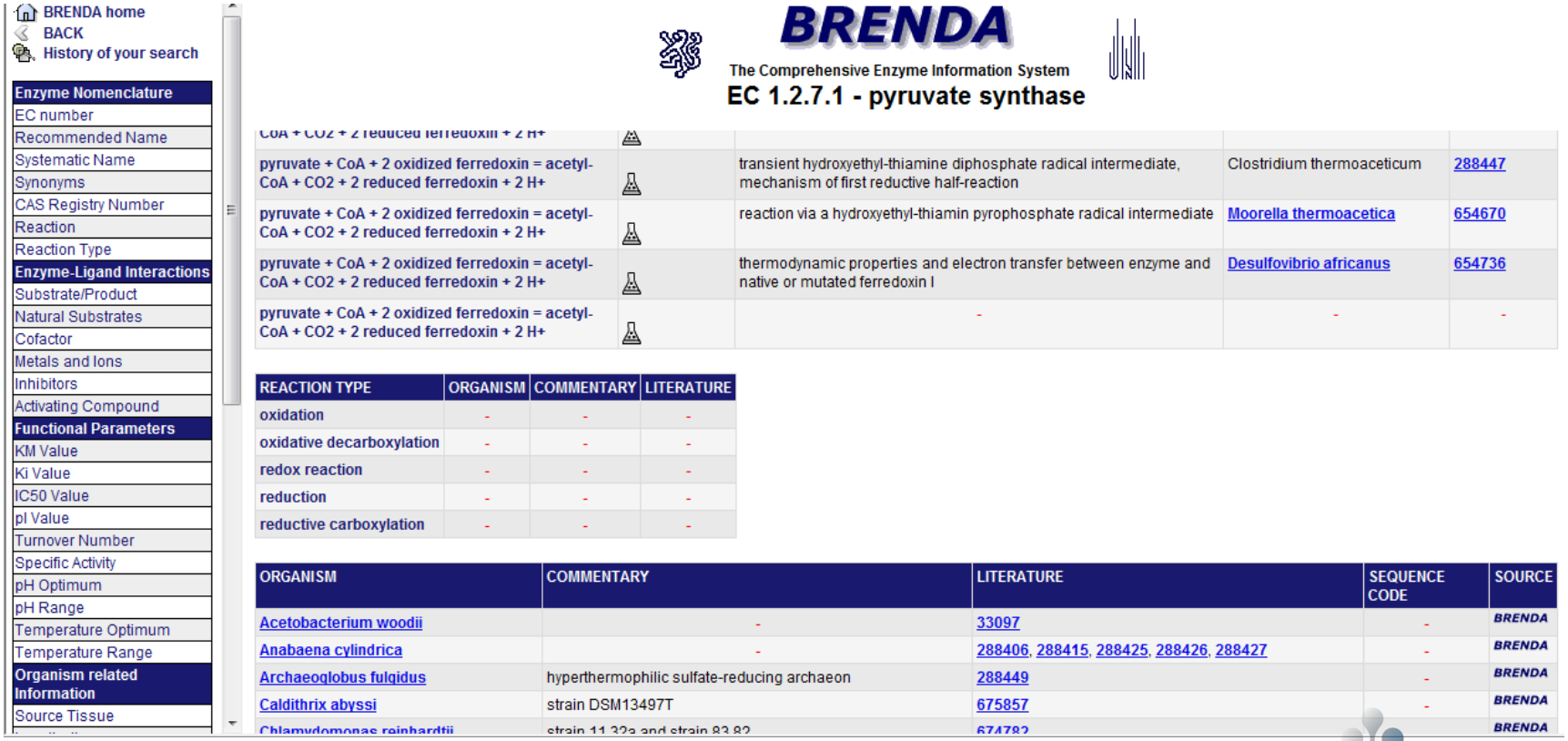
SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

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

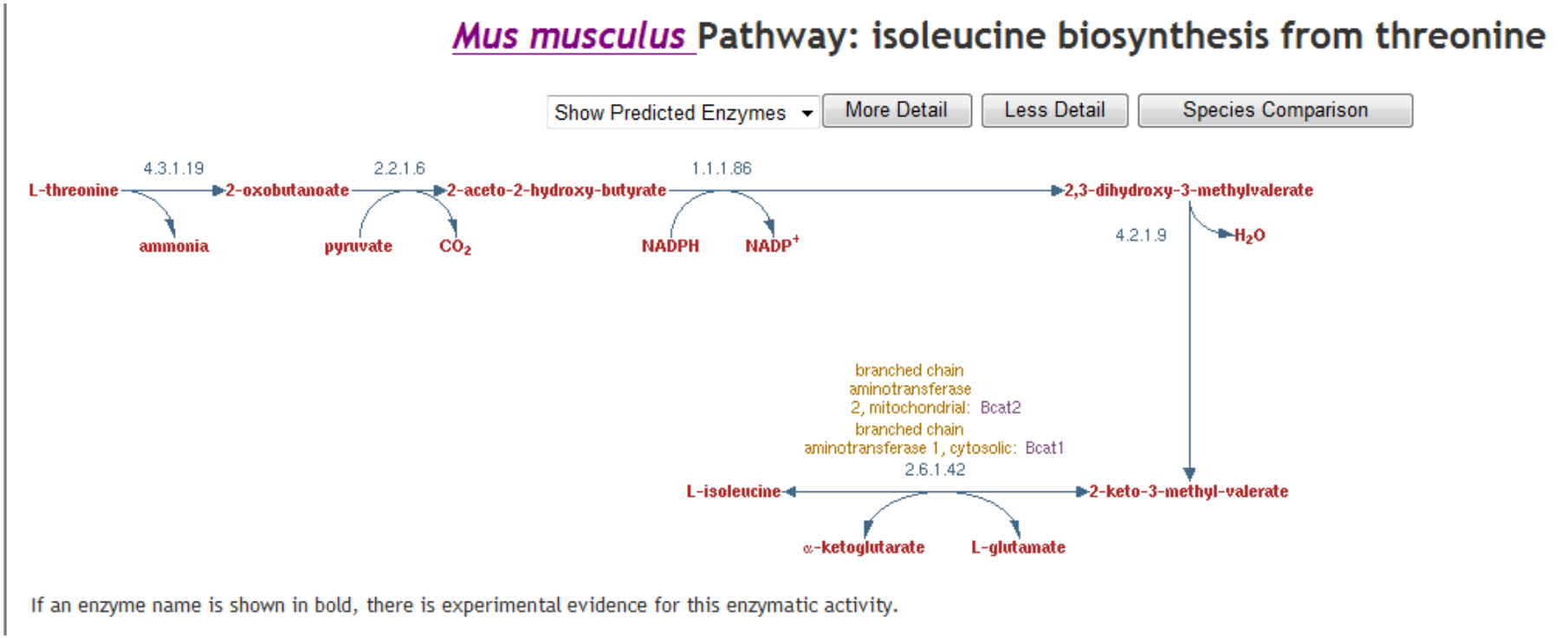
CoA + CO2 + 2 reduced ferredoxin + 2 H+				
pyruvate + CoA + 2 oxidized ferredoxin = acetyl-CoA + CO2 + 2 reduced ferredoxin + 2 H+		transient hydroxyethyl-thiamine diphosphate radical intermediate, mechanism of first reductive half-reaction	Clostridium thermoaceticum	288447
pyruvate + CoA + 2 oxidized 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

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.



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

• 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)

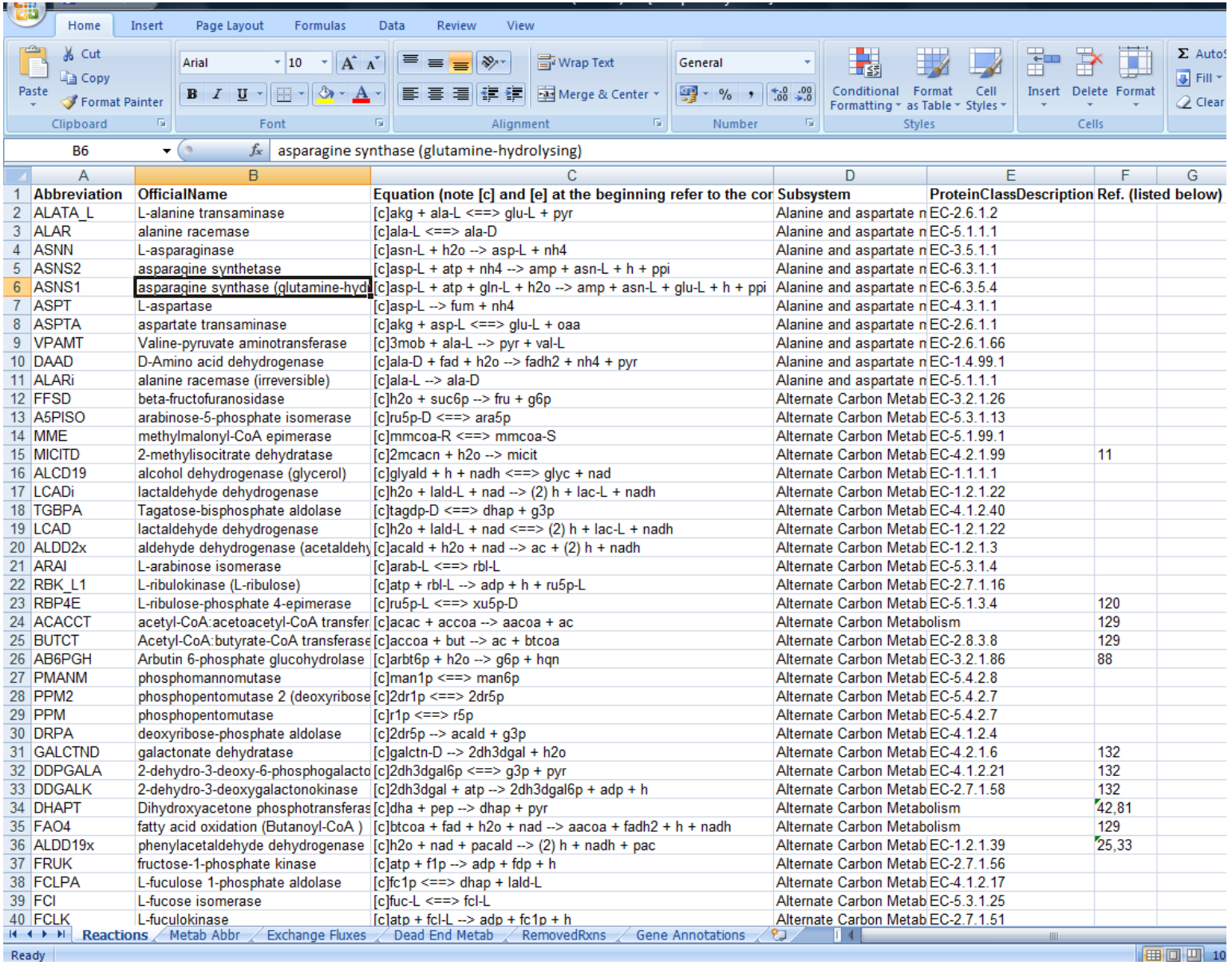
• Growth association requirements related to polymerization of aminoacids, nucleotides, …

• Need to determine energy requirements for maintenance

  • Maintenance of gradients and electrical potential (most important)

  • Turnover of macromolecules

METABOLIC MODELS - FORMATS



Home | Insert | Page Layout | Formulas | Data | Review | View

Cut | Copy | Format Painter | Paste | Clipboard | Arial | 10 | Font | Wrap Text | Merge & Center | Alignment | General | Number | Conditional Formatting | Format as Table | Cell Styles | Styles | Insert | Delete | Format | Cells | AutoSum | Fill | Clear

B6 | fx | 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>
```

GSMMs| RECONSTRUCTION - SOFTWARE

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

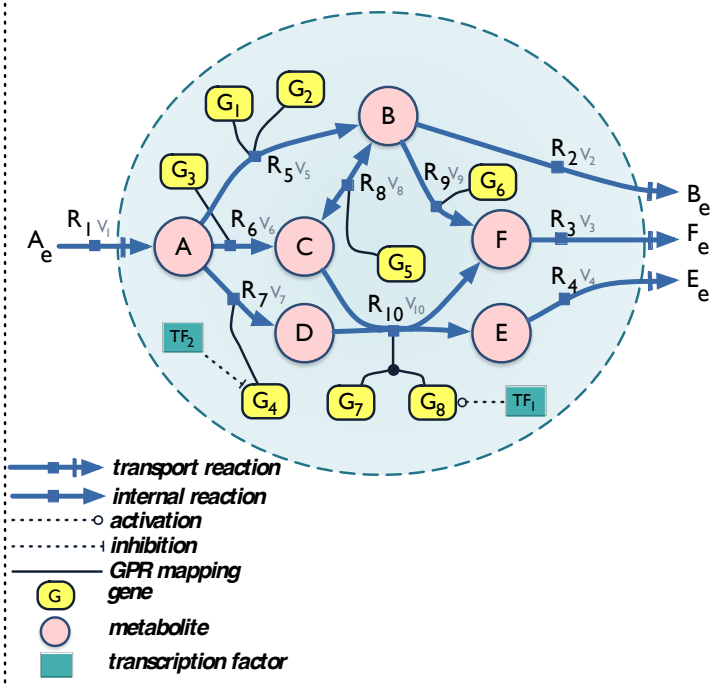
- Supports the main tasks in metabolic genome **re-annotation** and **reaction assignment**
- Supports compartmentalization and transporters identification
- Allows manual curation of the model through a user friendly environment

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)



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

$0 \le V_1 \le a$

$0 \le V_2 \le \infty$

$0 \le V_3 \le \infty$

$0 \le V_4 \le \infty$

$0 \le V_5 \le \infty$

$0 \le V_6 \le \infty$

$0 \le V_7 \le \infty$

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

$0 \le V_9 \le \infty$

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

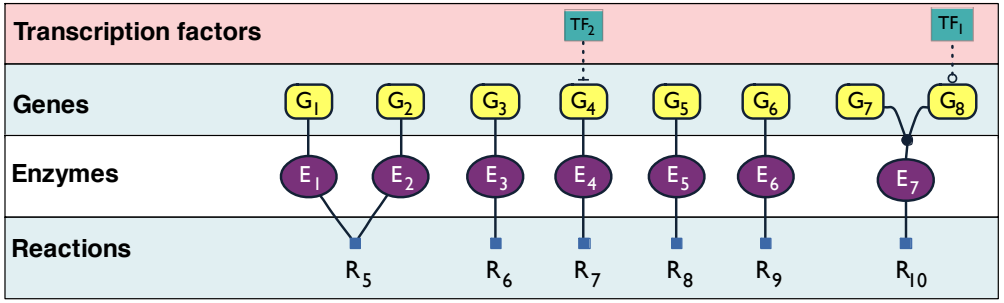
b)

Stoichiometric matrix **Fluxes**

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
A	1	0	0	0	-1	-1	-1	0	0	0
B	0	-1	0	0	1	0	0	-1	-1	0
C	0	0	0	0	0	1	0	1	0	-1
D	0	0	0	0	0	0	1	0	0	-1
E	0	0	0	-1	0	0	0	0	0	1
F	0	0	-1	0	0	0	0	0	1	1

$S(m,n) * v(n) = 0(m)$

c)



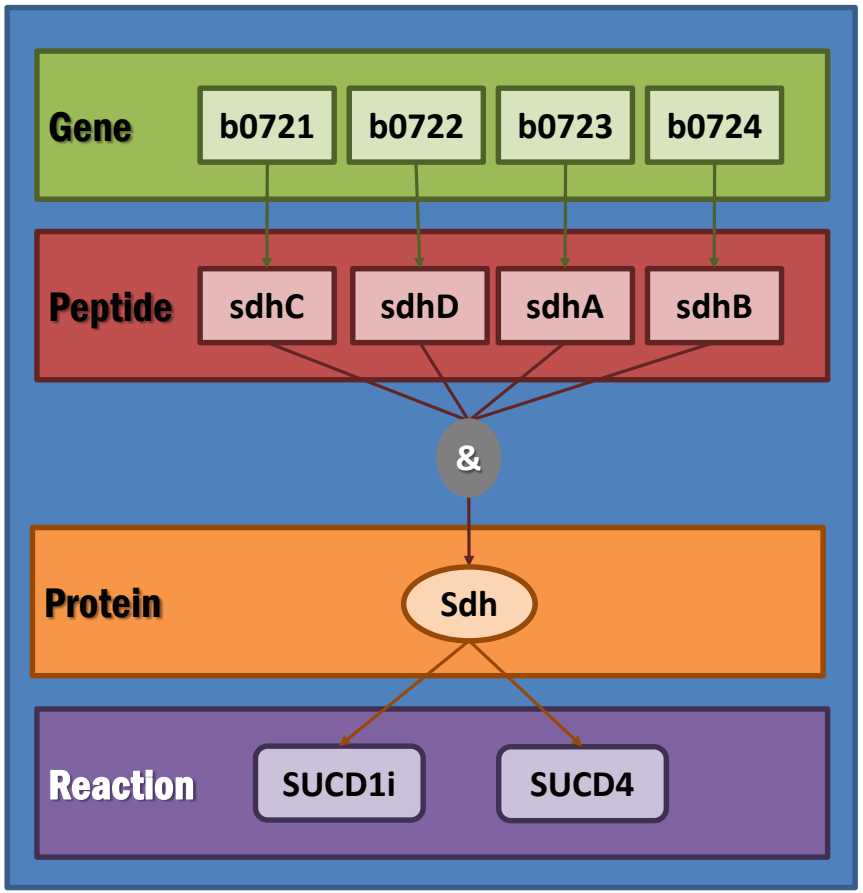
Reaction	GPR
R_5	G_1 or G_2
R_6	G_3
R_7	G_4
R_8	G_5
R_9	G_6
R_{10}	G_7 and G_8

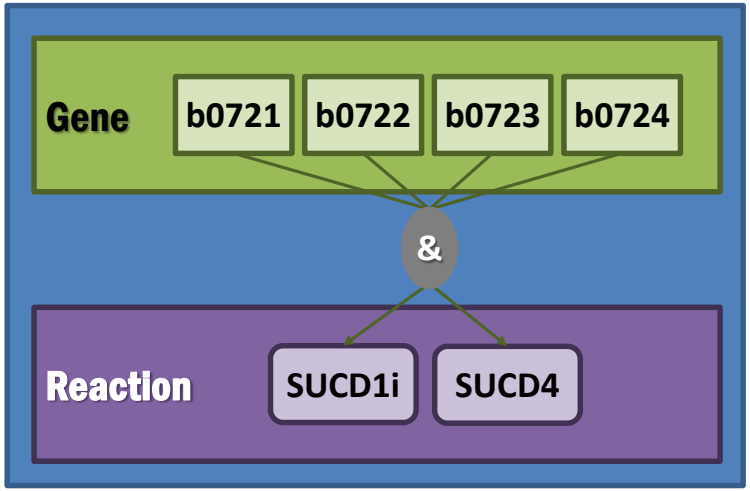
Gene	Regulatory rule
G_4	not TF_2
G_8	TF_1

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 under-determined equation system 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**

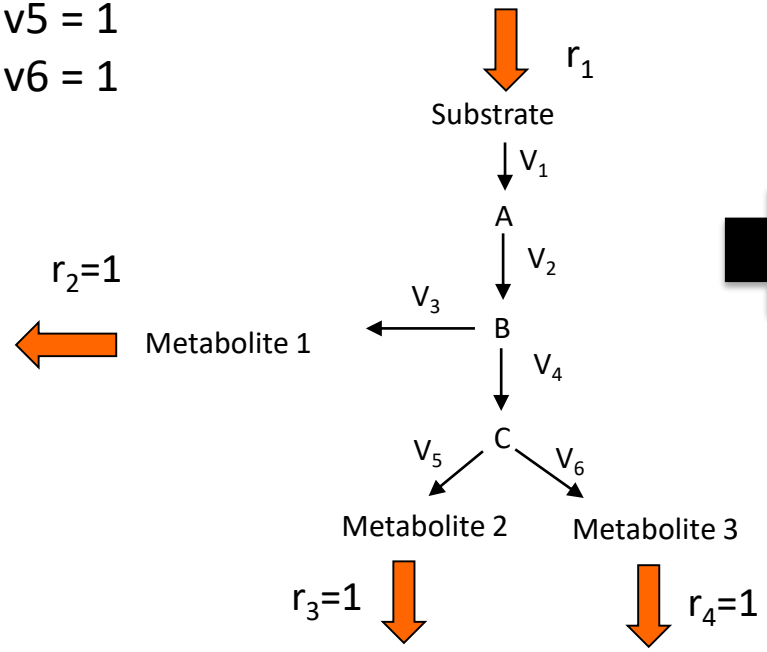
Under-determined!

If we measure 3 exchange fluxes

v3 = 1

v5 = 1

v6 = 1



A: v1 - v2 = 0

B: v2 – v3 – v4 = 0 ⇔ v2 – 1 – v4 = 0 ⇔

C: v4 – v5 –v6 =0 v4 – 1 – 1 =0

v1 - v2 = 0

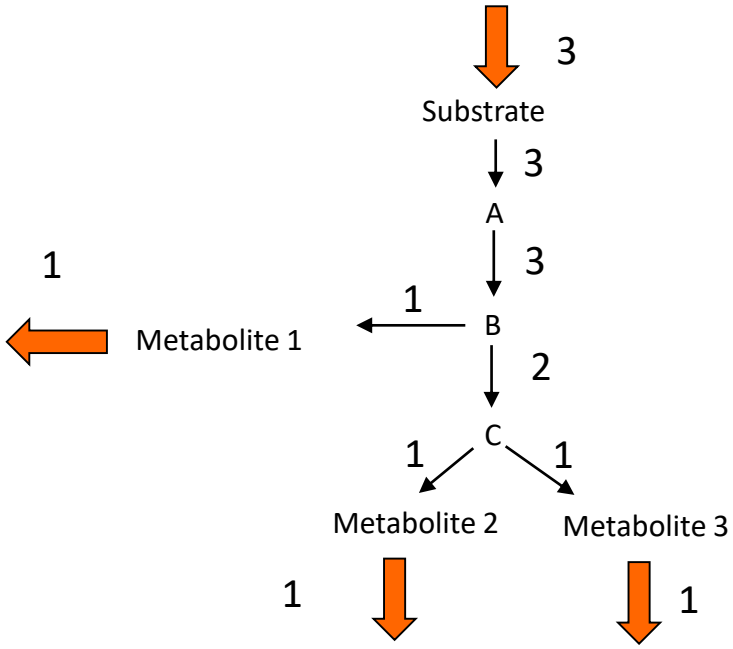
v1 = v2

v2 = v4 +1 ⇔ v2 = 3

v4 = 2 v4 = 2

v1 = 3

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:

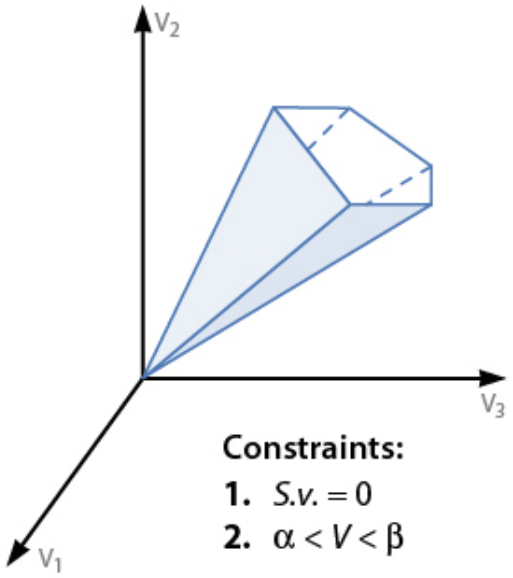
$$S\,v = 0$$

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

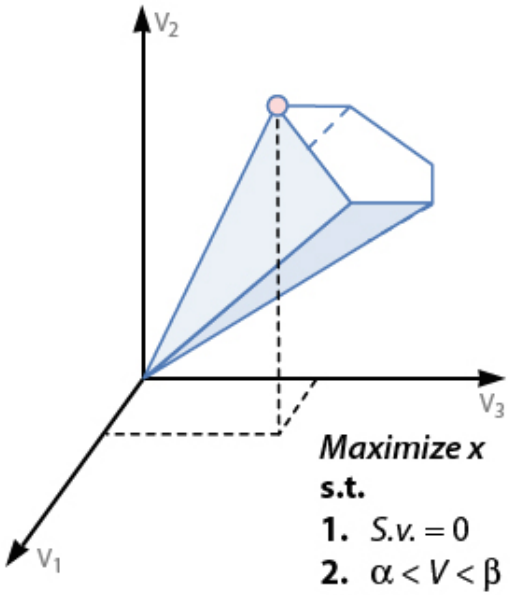
α, β = lower and upper limits for fluxes. Use

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

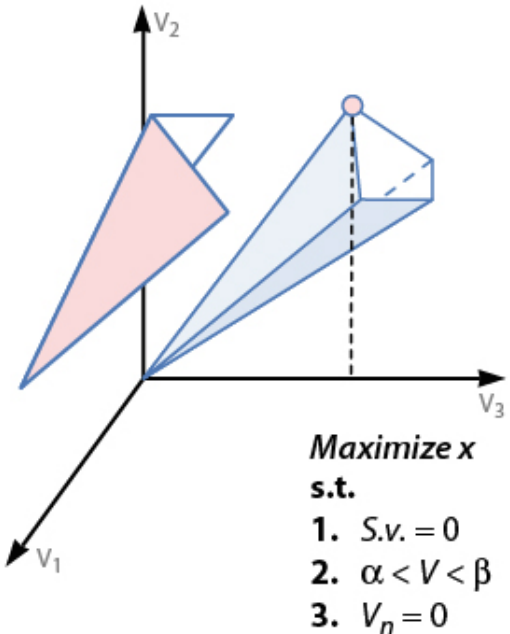
A. Admissible flux space



B. Optimal flux distribution



C. Further constraints to redirect the flux



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.

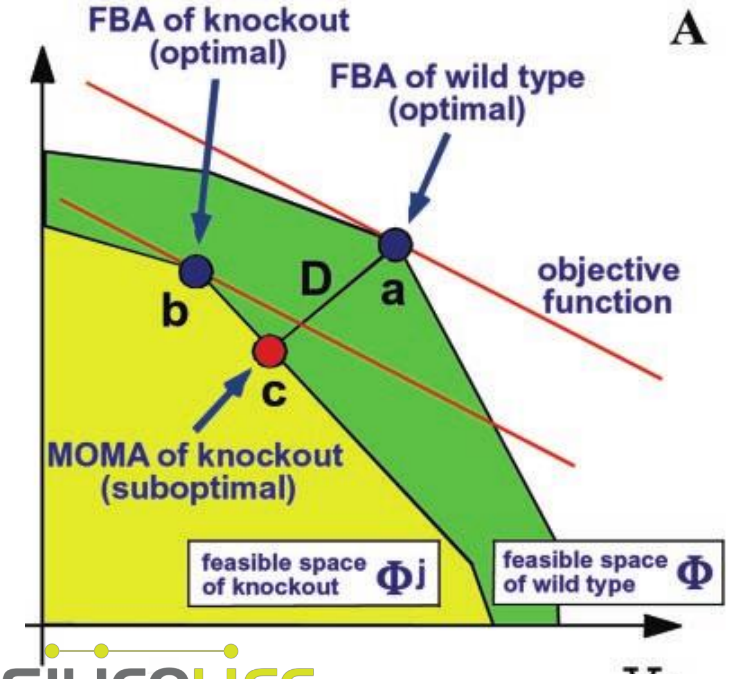
- However, it has been shown that for mutants and wild-type organisms grown on some unusual carbon sources the hypothesis of optimal growth is not always real.

- 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, but the optimal growth flux for mutants is relaxed.

- 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:

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

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

PHENOTYPE PREDICTION – DYNAMIC MODELS

	Number pathways	Number metabolites	Number parameters	Number reactions	Number comparments	Publication year
Chassagnole	2	18	127	48	2	2002
Kadir	6	23	179	51	2	2010
Peskov	7	49	662	73	2	2012
Jahan	8	52	365	129	1	2016
Millard	10	68	441	62	3	2016
GSMM iAF1260		1668	NA	2382	3	2007

10 metabolic pathways

- Glucose Phosphotransferase System (PTS)
- Glycolysis and Gluconeogenesis (EMP)
- Pentose Phosphate Pathway (PPP)
- Anaplerotic Reactions (AR)
- Tricarboxylic Acid Cycle (TCA)
- Entner-Doudoroff Pathway (EDP)
- Acetate Metabolism (AC)
- Oxidative Phosphorylation (OP)
- Nucleotide Interconversion Reactions (NC)
- Glyoxylate Shunt (GS)

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

SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

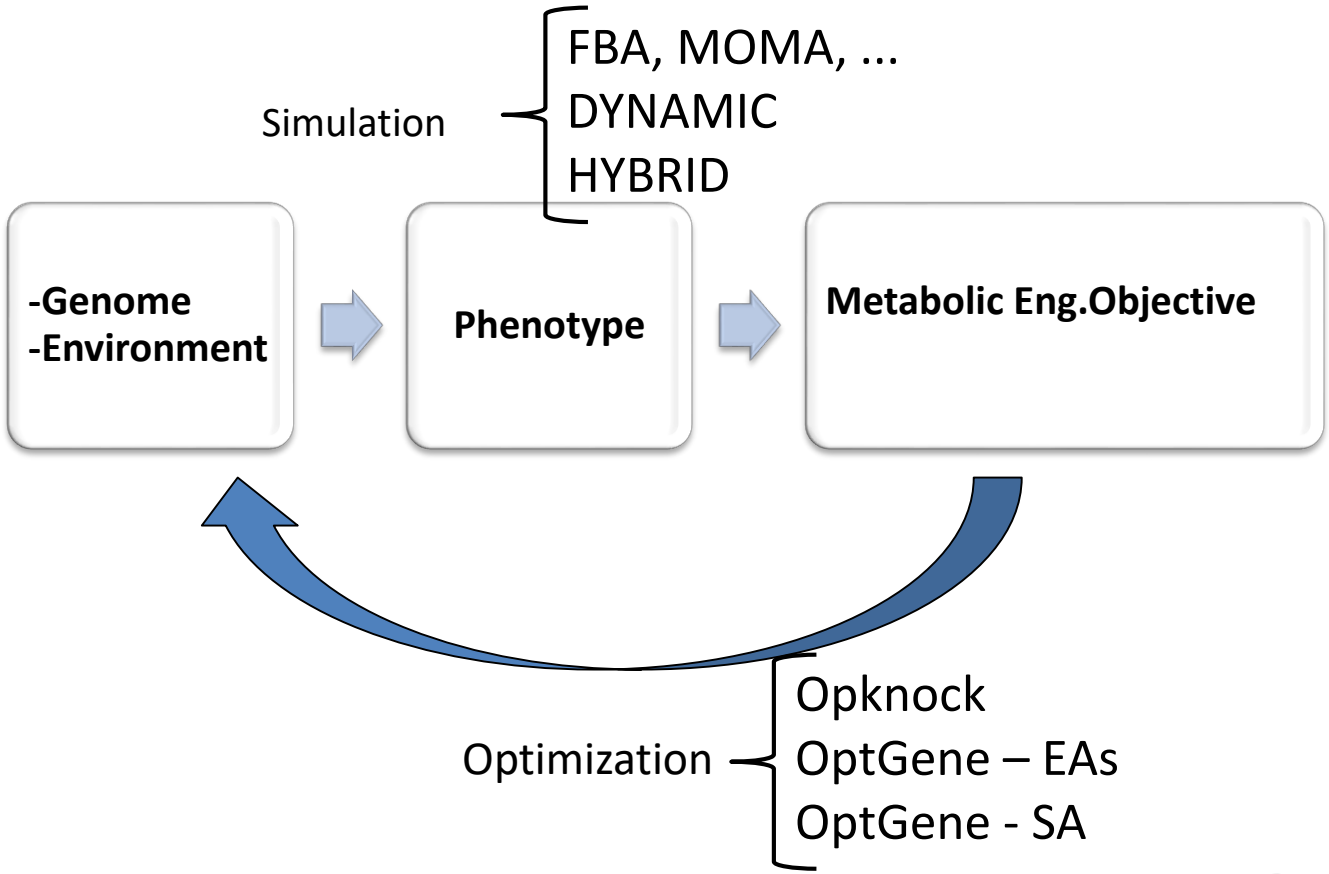
METABOLIC ENGINEERING

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

A view of the Strain Design Problem



Strain Optimization

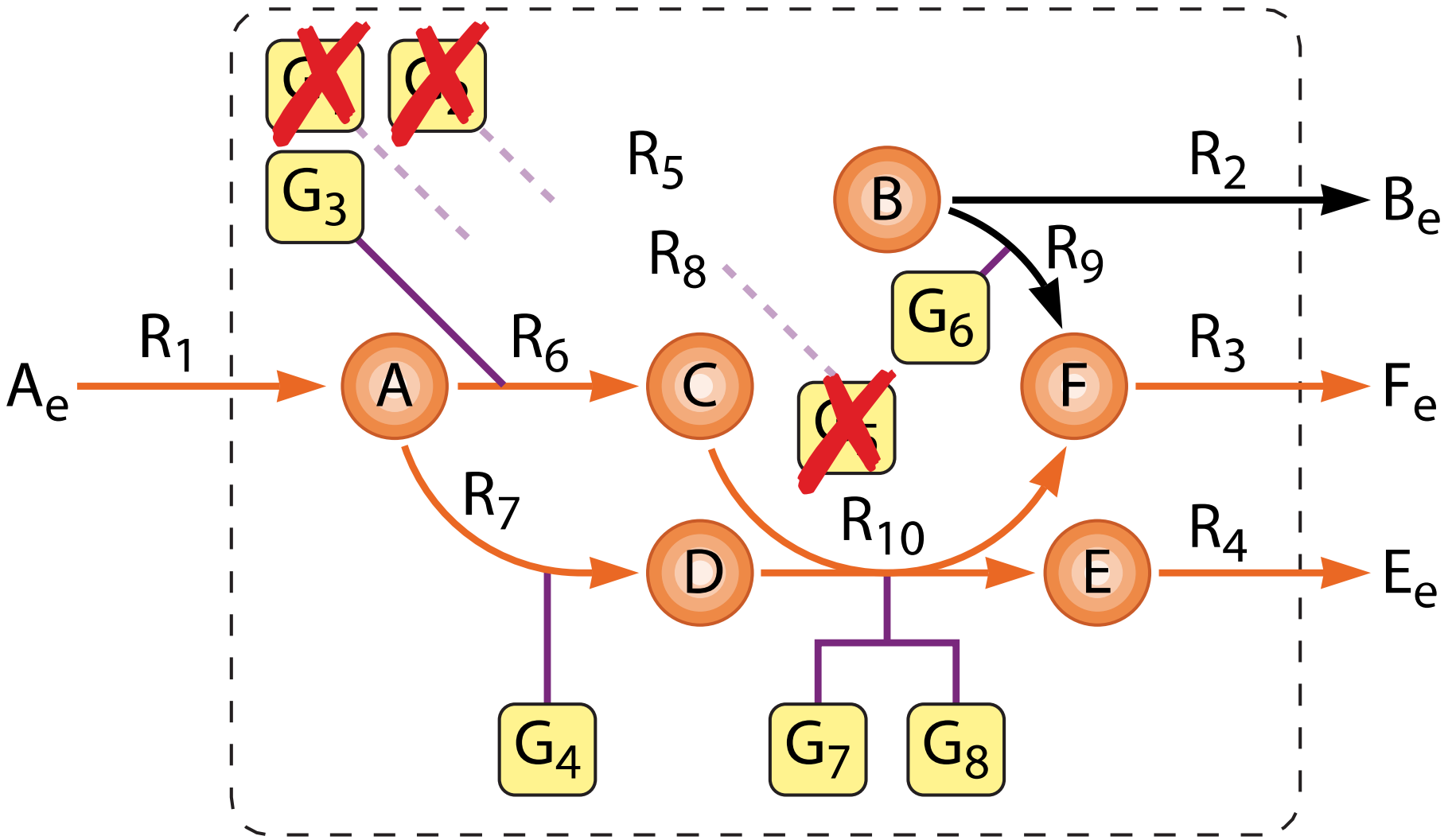
Possible aims

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

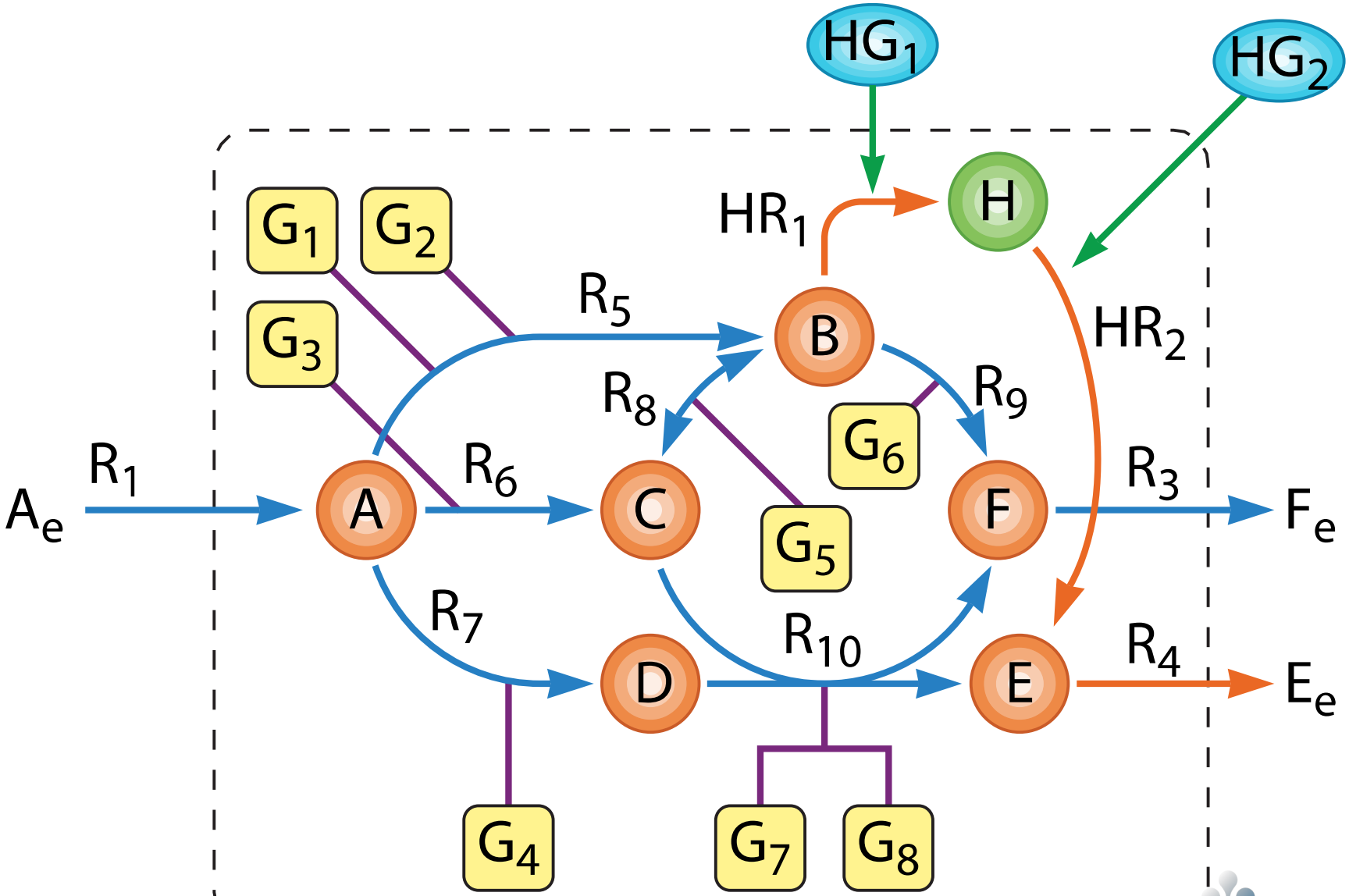
Objective function

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

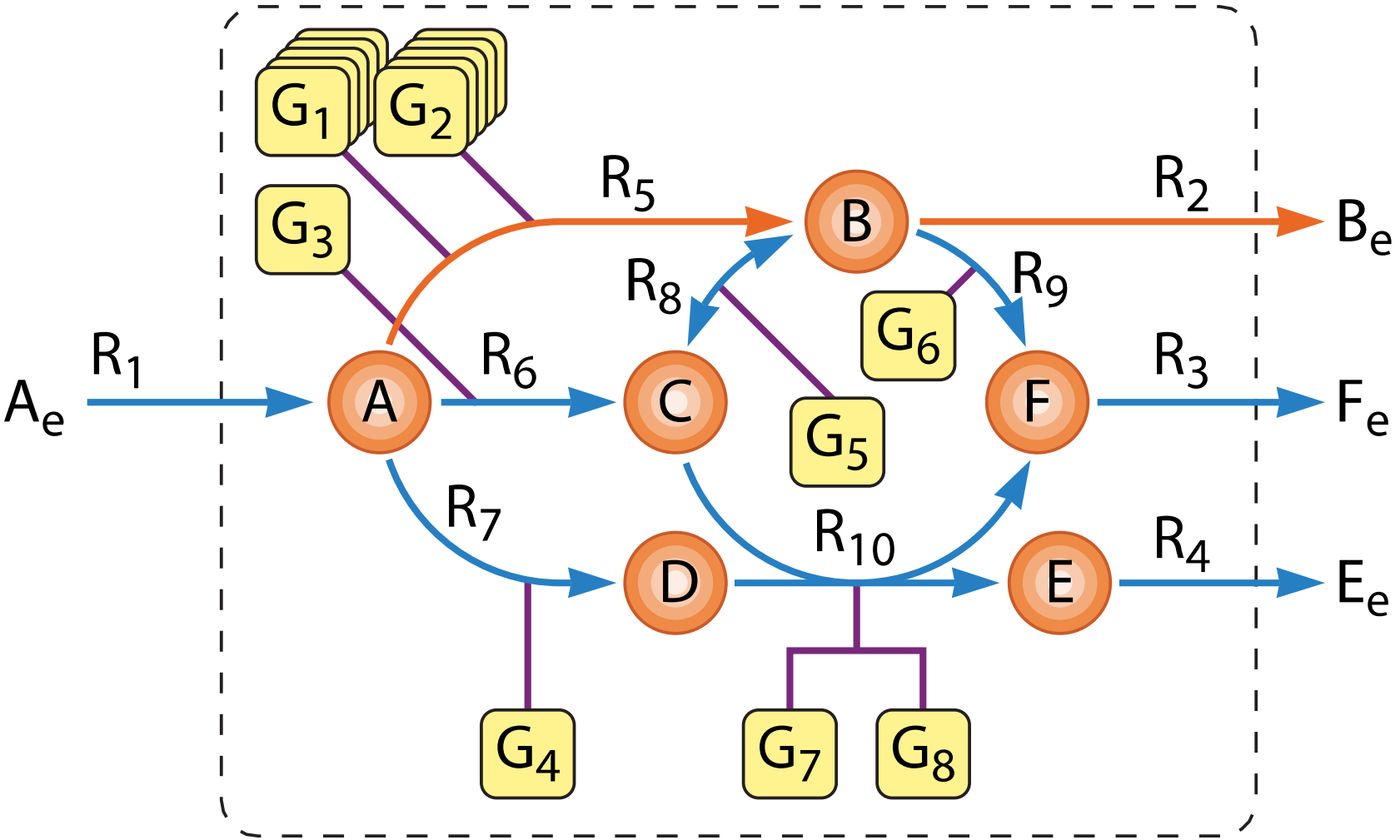
Strain Optimization – Gene deletion



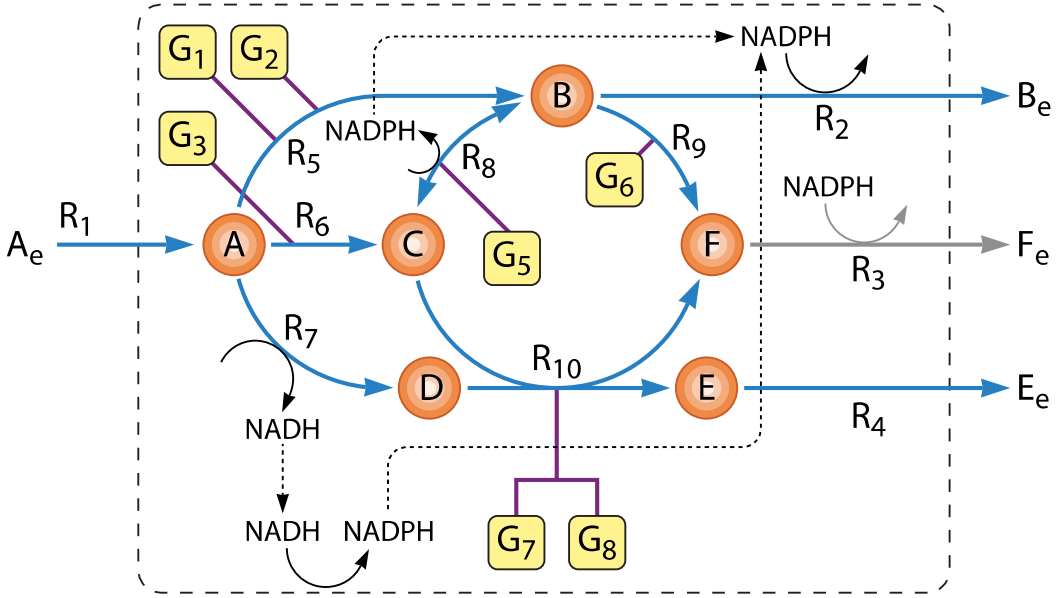
Strain Optimization – Heterologous insertion

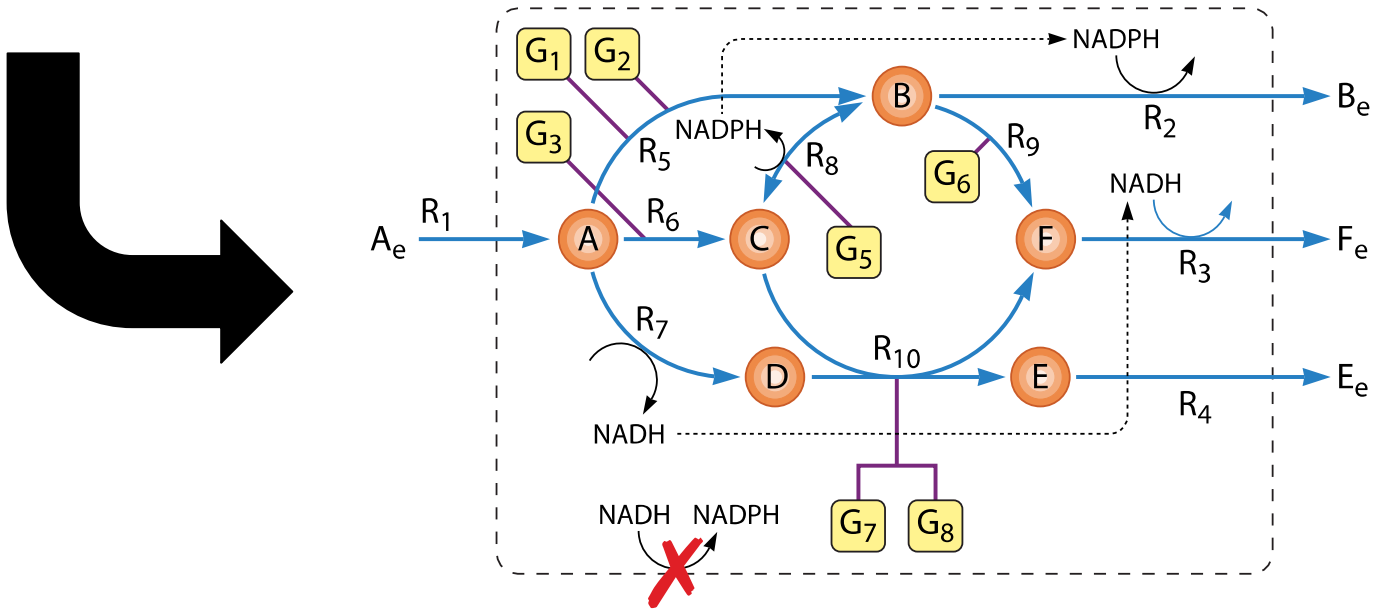


Strain optimization – gene over/under expression



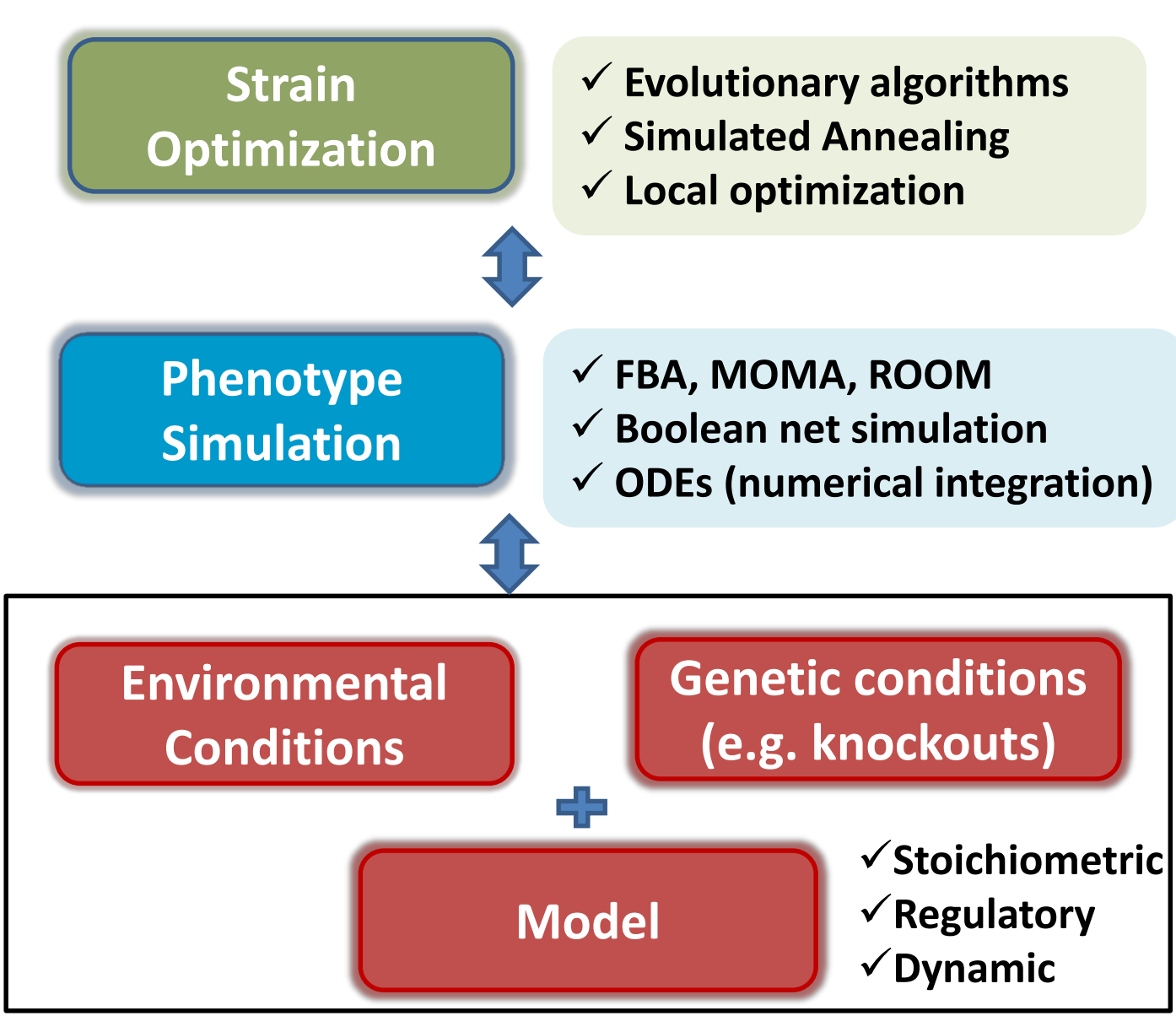
Strain Optimization: co-factor swapping



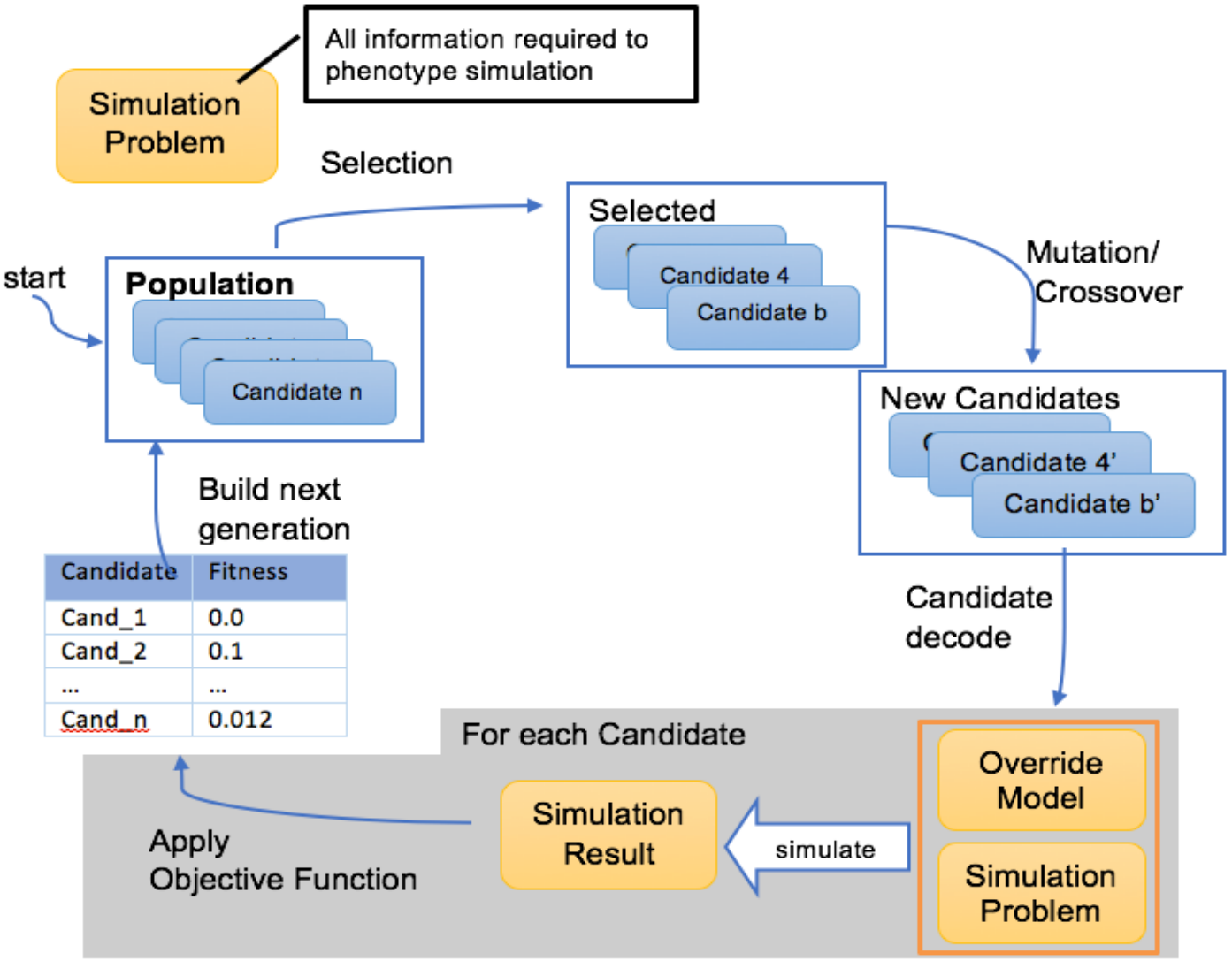


SILICOLIFE

Strain Optimization: Decoupled optimization



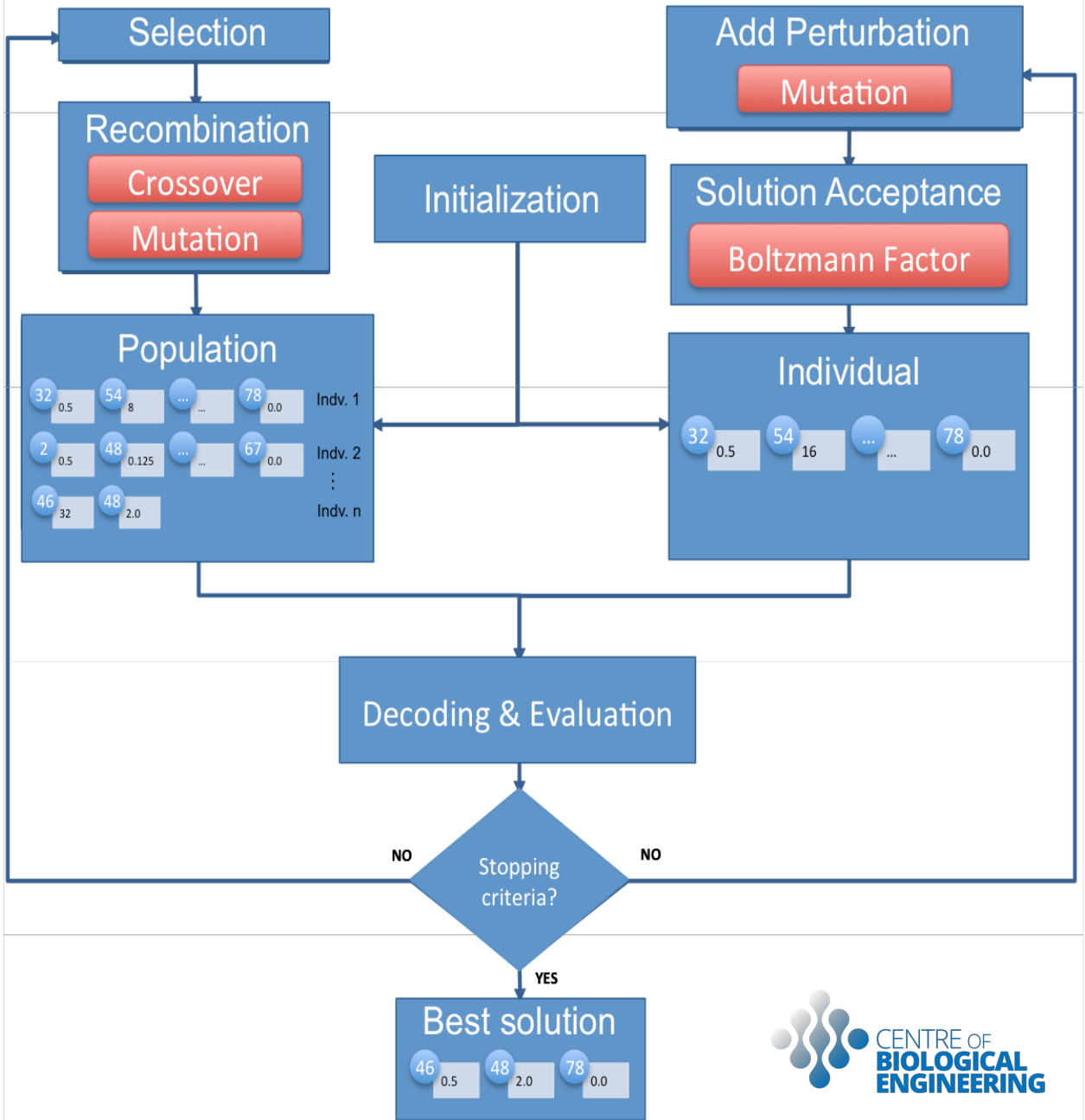
STRAIN OPTIMIZATION: DECOUPLED OPTIMIZATION



STRAIN OPTIMIZATION – ALGORITHMS

Evolutionary Algorithms (EA)

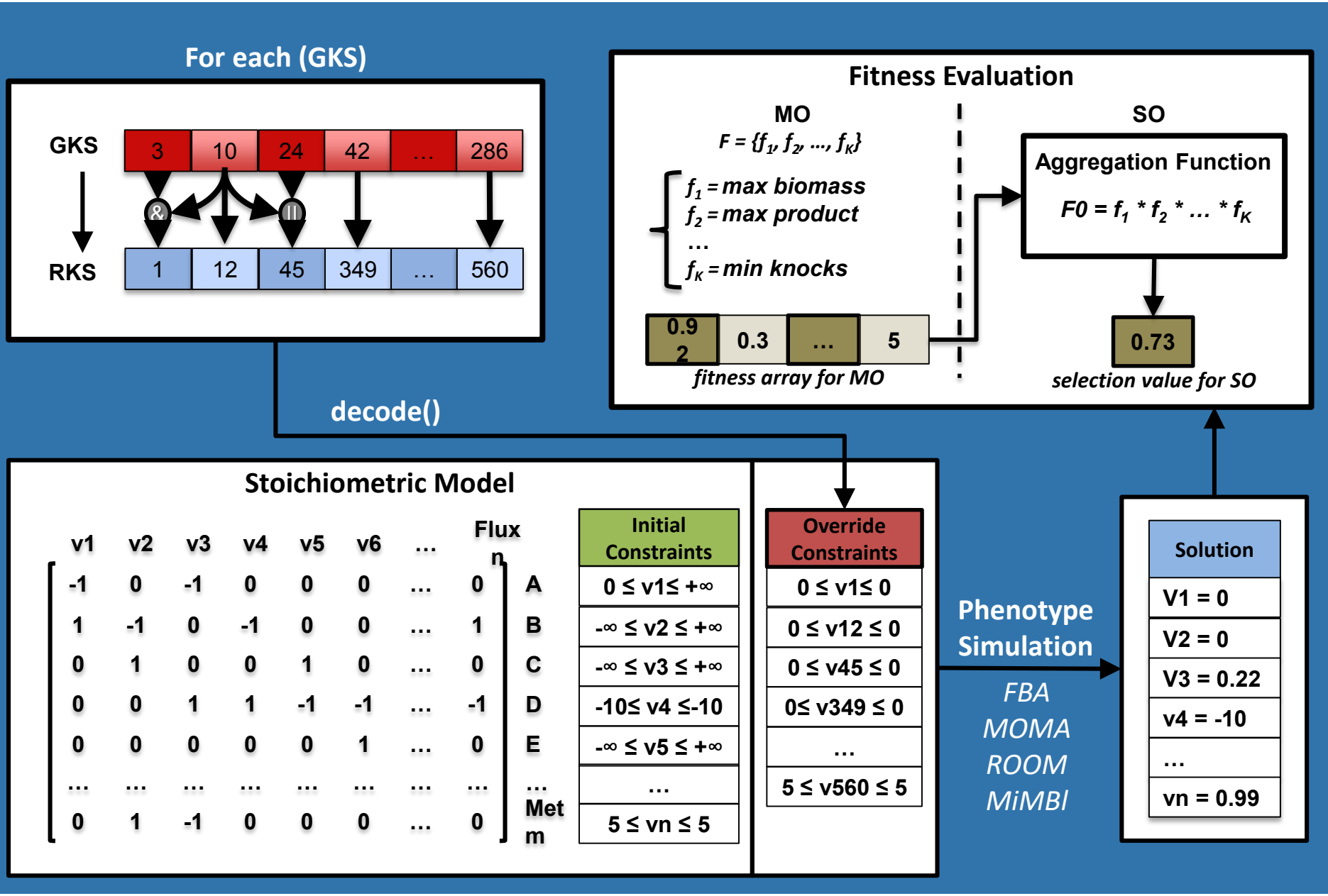
Simulated Annealing (SA)



SILICOLIFE

CENTRE OF BIOLOGICAL ENGINEERING

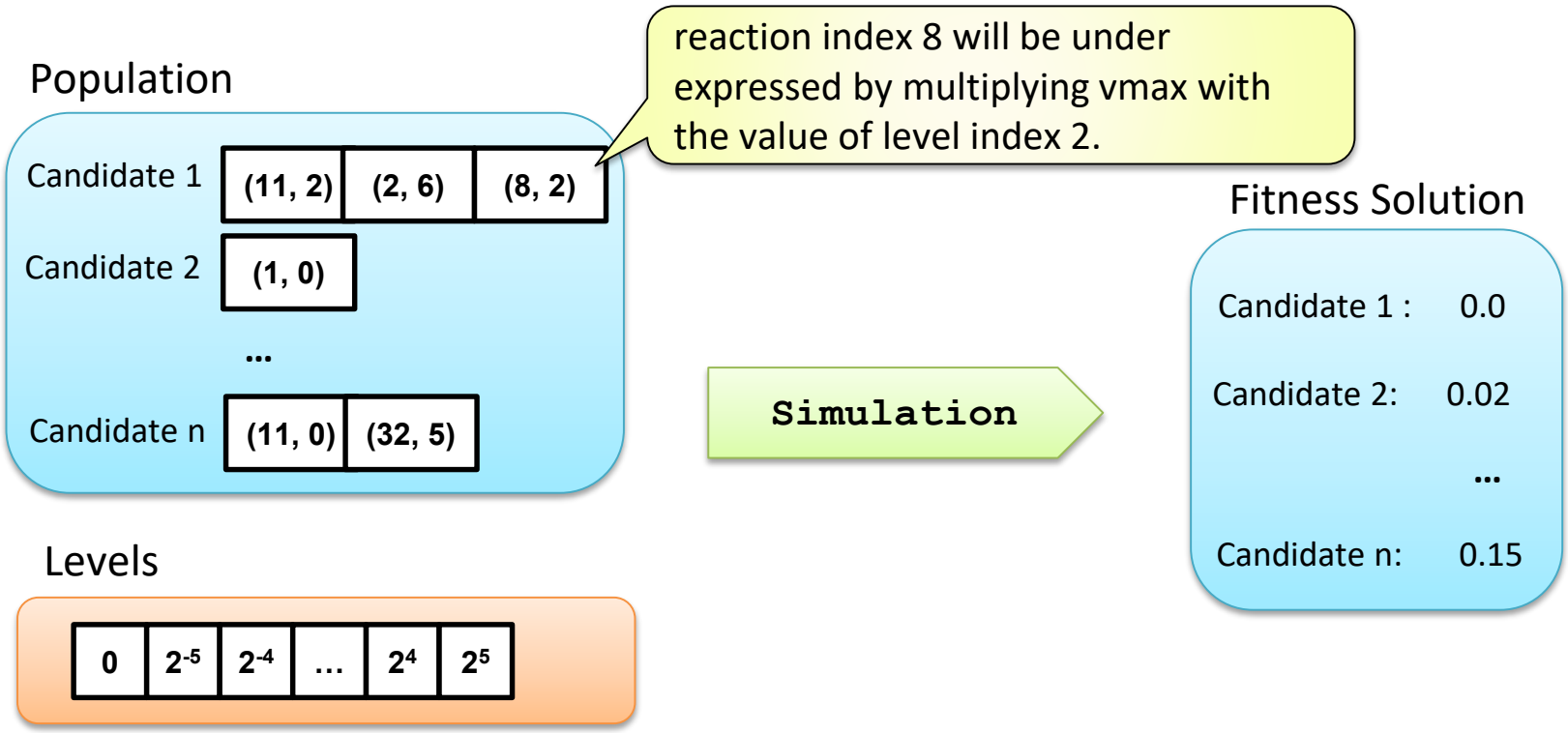
STRAIN OPTIMIZATION – EXAMPLE USING STOICH MODELS



STRAIN OPTIMIZATION – OVER/UNDER EXPRESSION

Enzyme Over/Under Expression

each element of candidate solution is a tuple of 2 integers, the first identifies the reaction to manipulate and the second the level of expression.



STRAIN OPTIMIZATION – SUCCINATE WITH STOICH MODELS

- Production of Succinate with *S. cerevisiae*

In vivo validation:

30-fold improvement in succinate titer

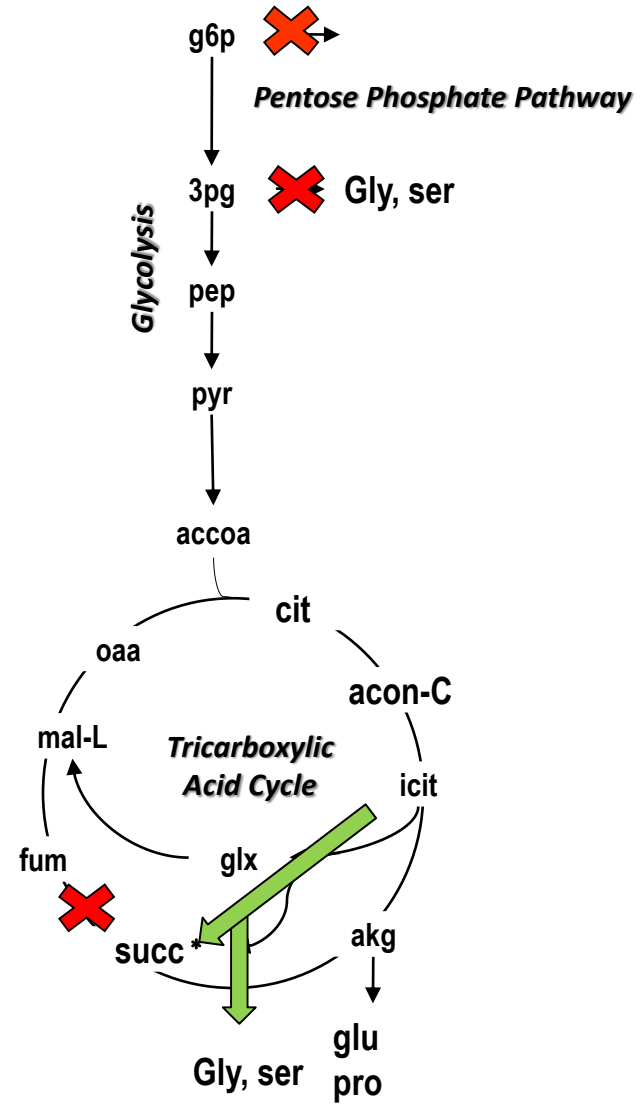
43-fold improvement in succinate

yield on biomass

2.8-fold decrease in the specific growth rate

Otero et al, PlosOne 2013

Lopes et al (under preparation)



STRAIN OPTIMIZATION – SUCCINATE W/ DYNAMIC MODELS

Best solution:

vPTS4_max', 32

v6PGDH_max : 0.03125

vAck_max: 0

SDH: 0.0625

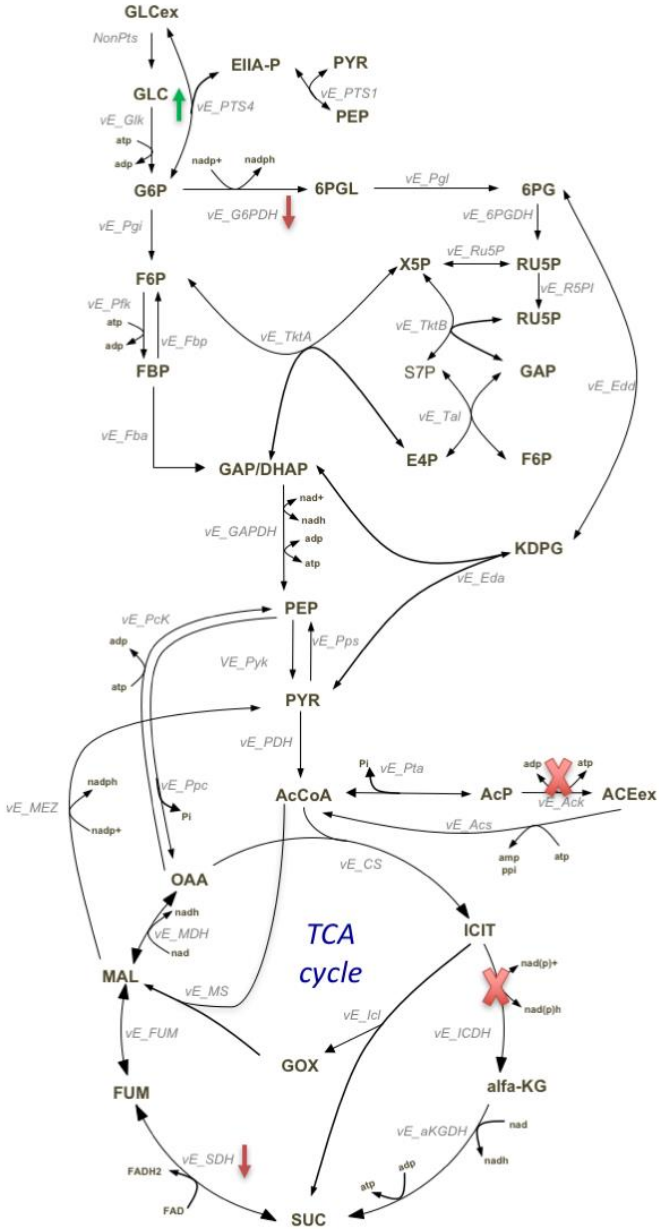
ICDH: 0

At least 3 of the targets have been experimentally validated

Advantages of using Dynamic Models:

- No need for steady-state considerations
- No need for further assumptions for steady state simulation
- Under/overexpression is directly accounted for

Correia et al (under preparation)



RESULTS FROM IN SILICO MET ENG

- **KNOCKOUTS – often too stringent for biological systems**
- **Cofactor swapping – difficult to implement**
- **Underexpression – often a good susbtitute for Kos**
- **Overexpression – works only when the catalytic properties of the enzyme are adequate**
- **Otherwise-> need for Enzyme Engineering (same targets as Overexpression or cofactor swapping)**

Basis for Rational Enzyme Engineering

Comparative Modelling Methods:

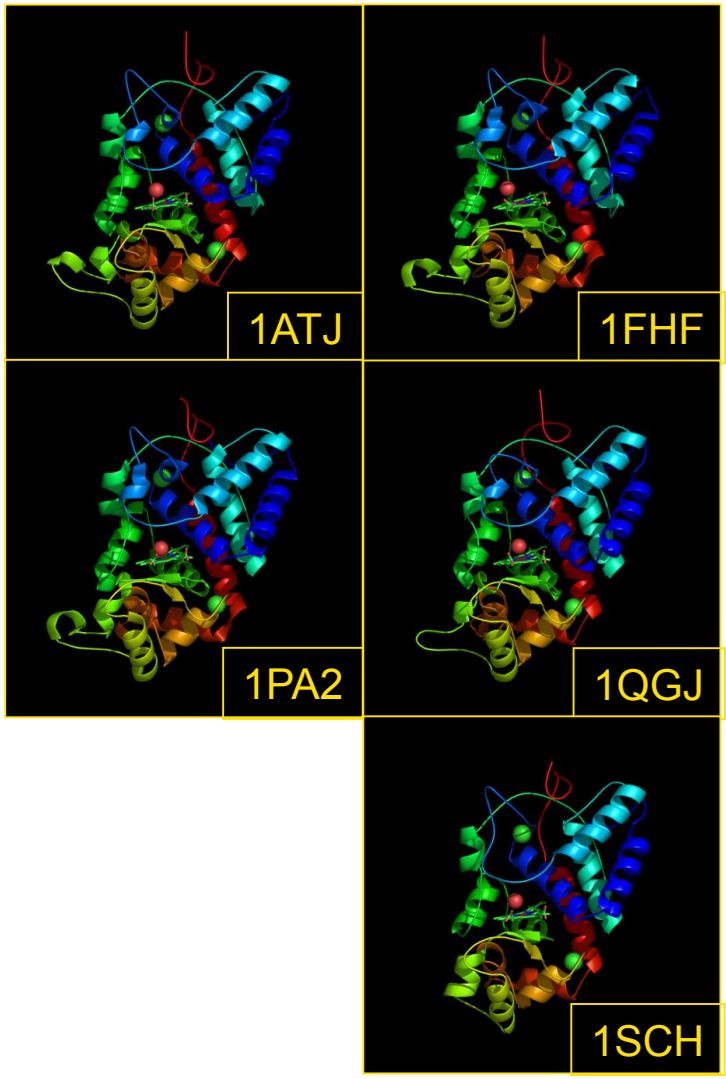
Homologous amino acid residue sequences $\Rightarrow$ homologous protein tertiary structures

Ex: trypsin, elastase, tonin

Homologous sequence - Homologous structure

Peroxidases

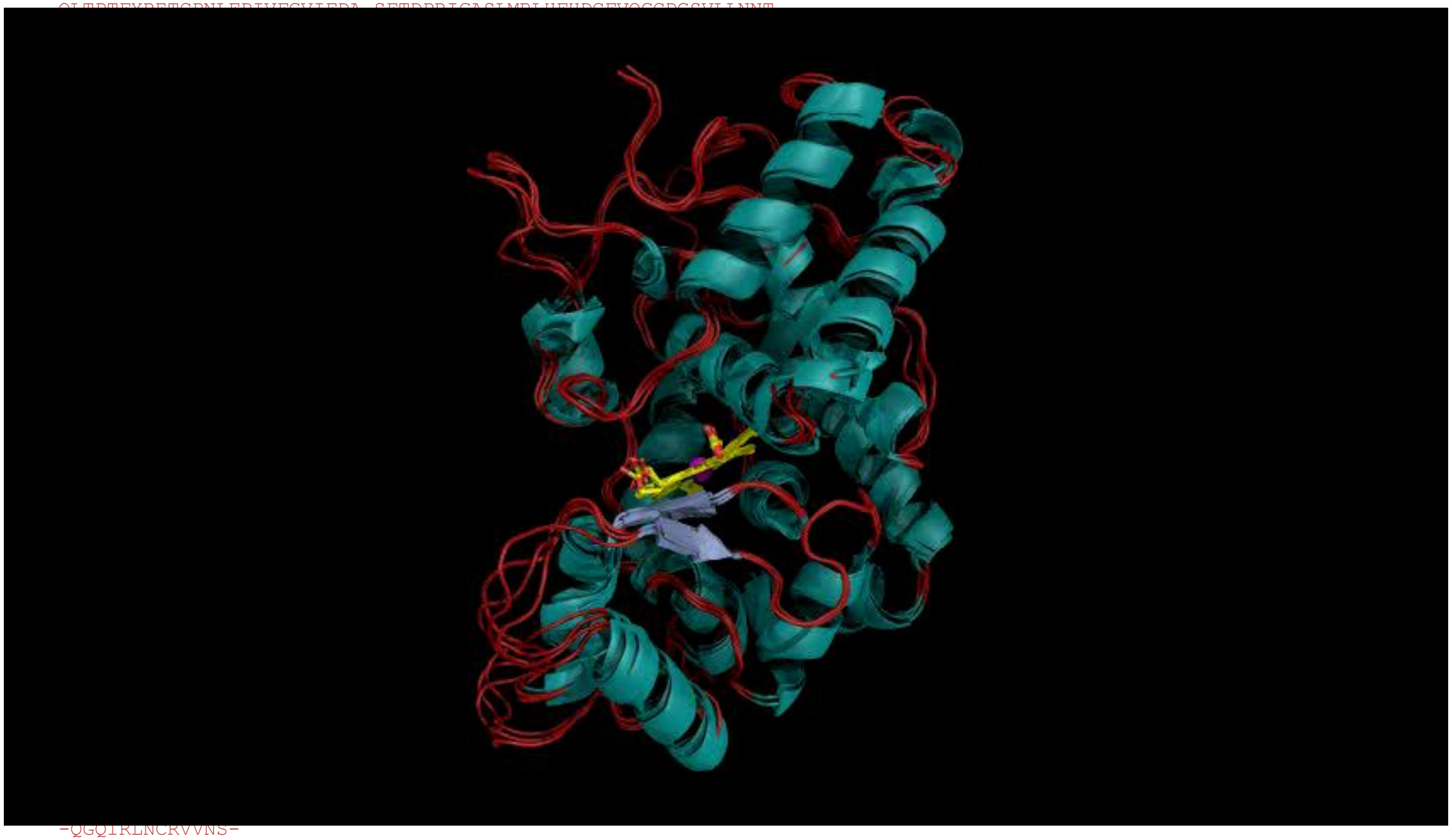
```
1ATJ            QLTPTFYDNSCPNVSNIVRDTIVNE-LRSDPRIAASILRLHFHDCFVNGCDASILLDNT-
1FHF            QLTPTFYRETCPNLFPIVFGVIFDA-SFTDPRIGASLMRLHFHDCFVQGCDGSVLLNNT-
1PA2            QLNATFYSGTCPNASAIVRSTIQQA-LQSDTRIGASLIRLHFHDCFVNGCDASILLDDT-
1QGJ            QLSPDIYAKSCPNLVQIVRKQVAIAL-KAEIRMAASLIRLHFHDCFVNGCDASLLLDG--
1SCH            -LSSNFYATKCPNALSTIKSAVNSAVAK-EARMGASLLRLHFHDCFVQGCDASVLLDDTS
                 *.. :*  .***    :   :       : *:.**::*********:***.*:**:.

1ATJ            TSF-RTEKDAFGNANSARGFPVIDRMKAAVESACPRTVSCADLLTIAAQQSVTLAGGPSW
1FHF            DTI-ESEQDALPNINSIRGLDVVNDIKTAVENSCPDTVSCADILAIAAEIASVLGGGPGW
1PA2            GSI-QSEKNAGPNVNSARGFNVVDNIKTALENACPGVVSCSDVLALASEASVSLAGGPSW
1QGJ            ---ADSEKLAIPNINSARGFEVIDTIKAAVENACPGVVSCADILTLAARDSVVLSGGPGW
1SCH            N-F-TGEKTAGPNANSIRGFEVIDTIKSQVESLCPGVVSCADILAVAARDSVVALGGASW
                      *: *  * ** **: *:: :*: :*. ** .***:*:*::*:. :    **..*

1ATJ            RVPLGRRDSLQAFLDLANANLPAPFFTLPQLKDSFRNVGLNRSSDLVALSGGHTFGKNQC
1FHF            PVPLGRRDSLTANRTLANQNLPAPFFNLTQLKASFAVQGL-NTLDLVTLSGGHTFGRARC
1PA2            TVLLGRRDSLTANLAGANSSIPSPIESLSNITFKFSAVGL-NTNDLVALSGAHTFGRARC
1QGJ            RVALGRKDGLVANQNSANN-LPSPFEPLDAIIAKFVAVNL-NITDVVALSGAHTFGQAKC
1SCH            NVLLGRRDSTTASLSSANSDLPAPFFNLSGLISAFSNKGF-TTKELVTLSGAHTIGQAQC
                 * ***:*.  *    **  :*:*:  *  :   *   .:    ::*:***.**:*: :*

1ATJ            RFIMDRLYNFSNTGLPDPTLNTTYLQTLRGLCPLNG--NLSALVDFDLRTPTIFDNKYYV
1FHF            STFINRLYNFSNTGNPDPTLNTTYLEVLRARCPQNAT-G-DNLTNLDLSTPDQFDNRYYS
1PA2            GVFNNRLFNFSGTGNPDPTLNSTLLSTLQQLCPQNG--SASTITNLDLSTPDAFDNNYFA
1QGJ            AVFSNRLFNFTGAGNPDATLETSLLSNLQTVCPLGG--NSNITAPLDRSTTDTFDNNYFK
1SCH            TAFRTRIY-N------ESNIDPTYAKSLQANCPSV--GGDTNLSPFDVTTPNKFDNAYYI
                  :  *::        :..::.:  . *:  **     .      :*  *.  *** *:

1ATJ            NLEEQKGLIQSDQELFSSPN-A-TDTIPLVRSFANSTQTFFNAFVEAMDRMGNITPLTGT
1FHF            NLLQLNGLLQSDQELFSTP-GA--DTIPIVNSFSSNQNTFFSNFRVSMIKMGNIGVLTGD
1PA2            NLQSNDGLLQSDQELFSTT-GS--STIAIVTSFASNQTLFFQAFAQSMINMGNISPLTGS
1QGJ            NLLEGKGLLSSDQILFSSDL-AVNTTKKLVEAYSRSQSLFFRDFTCAMIRMGNIS--NGA
1SCH            NLRNKKGLLHSDQQLFN-G--V-S-TDSQVTAYSNNAATFNTDFGNAMIKMGNLSPLTGT
                ** . .**: *** **.        *   * ::: .   *   *  :* .***:   .*

1ATJ            -QGQIRLNCRVVNS-
1FHF            -EGEIRLQCNFVNG-
1PA2            -NGEIRLDCKKVNGS
1QGJ            S-GEVRTNCRVINN-
1SCH            -SGQIRTNCRKTN--
                  *::* :*.  *
```



Homologous sequence - Homologous structure

Peroxidases

```
1ATJ          QLTPTFYDNSCPNVSNIVRDTIVNE-LRSDPRIAASILRLHFHDCFVNGCDASILLDNT-
1FHF
1PA2
1QGJ
1SCH

1ATJ
1FHF
1PA2
1QGJ
1SCH

1ATJ
1FHF
1PA2
1QGJ
1SCH

1ATJ
1FHF
1PA2
1QGJ
1SCH

1ATJ
1FHF
1PA2
1QGJ
1SCH
```



```
1ATJ          -QGQIRLNCRVVNS-
1FHF          -EGEIRLQCNFVNG-
1PA2          -NGEIRLDCKKVNGS
1QGJ          S-GEVRTNCRVINN-
1SCH          -SGQIRTNCRKTN--
                *::* :*.  *
```

Homologous sequence - Homologous structure

This opens the way for modelling proteins with unknown structure on basis of protein structures of homologous proteins.

- This is called Comparative Modelling
  - Some people call it Homology Modelling
- These proteins will be quite similar
  - But not exactly the same
- We will amplify the current structural knowledge

Comparative modelling can be applied:

- Identities above 30-40%
  For some families it can be lower

- The existence of at least one homologous structure

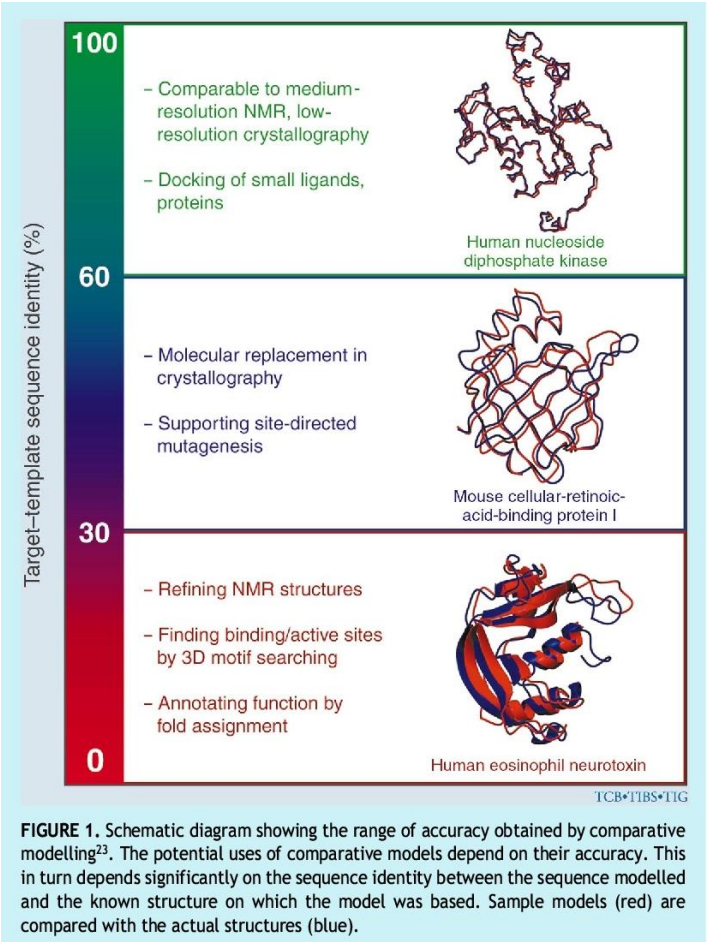


FIGURE 1. Schematic diagram showing the range of accuracy obtained by comparative modelling[23]. The potential uses of comparative models depend on their accuracy. This in turn depends significantly on the sequence identity between the sequence modelled and the known structure on which the model was based. Sample models (red) are compared with the actual structures (blue).

Recipe:

- For the amino acid residue sequence of the protein of unknown tertiary structure *(target)*, find proteins of known structure *(templates)* whose sequences are homologous to that of the target protein
  - Do a BLAST (or FASTA) against the PDB (Protein Data Bank)
- Align (in 3D) the structure of the known proteins and derive a 3D based sequence alignment.

- Align the amino acid residue sequences of the unknown protein to the *alignment template* of the known structures

- Build a model of the target protein from the information contained in:
  - the alignment of the sequences
  - the tertiary structures of the template proteins

Internet based comparative modelling servers

- Swiss Model: http://swissmodel.expasy.org/
- 3D-JIGSAW: http://bmm.cancerresearchuk.org/~3djigsaw/
- ModWeb: https://modbase.compbio.ucsf.edu/scgi/modweb.cgi

Modelling by satisfaction of spatial restraints

Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spacial restraints. *J.Mol.Biol.* **234,** 779-815.
Sánchez, R. & Sali, A. (1997). Advances in comparative protein-structure modelling. *Curr.Opin.Struct.Biol.* **7,** 206-214.

Figure 1: Comparative protein modeling by satisfaction of spatial restraints. First, the sequence to be modeled (target) is aligned with the known 3D structures (templates). Second, a large number of restraints (dashed lines) on distances and dihedral angles in the target sequence are extracted from the alignment. Third, the 3D model is obtained by satisfying all the restraints as well as possible. Reprinted with permission from [79].

Modelling whole genomes: fully automated protein modelling

Sánchez, R., Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *PNAS,* **95**, 13597-13602.
Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R., Melo, F., Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annu.Rev.Biophys.Biomol.Struct.,* **29**, 291-325.

Comparative modelling approaches have been used to do massive structure prediction of whole genomes

Limited to ORFs (or domains) that show significant homology with the experimental database (PDB)

Ex: MODBASE http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

Sanchez et al (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res* **28**:250-3; Pieper *et al.* (2014) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336-46,.

77% of the transcripts of the genomes in the database have models present (in 2016).

Genome Datasets: (at 2016-01-24)

Genome	# of Transcripts	# of modeled Transcripts	# of Models	Status/Year	%Modelled transcripts
Aeropyrum pernix	1,700	1,183	2,698	2007	70
Arabidopsis thaliana	47,160	36,520	116,372	2013	77
Archaeoglobus fulgidus	2,409	1,794	3,980	2007	74
Aspergillus oryzae	12,051	9,546	27,539	2013	79
Bacillus subtilis	4,105	3,373	9,244	2007	82
Brugia Malayi	11,397	7,850	23,216	2008	69
Burkholderia mallei	4,798	3,908	11,032	2007	81
Caenorhabditis elegans	22,698	18,996	52,232	2007	84
Canis familiaris	30,264	22,614	65,595	2007	75
Clostridium tetani	2,412	2,158	5,864	2007	89
Coxsackievirus B3 (strain Woodruff)	15	13	24	2012	87
Cryptosporidium hominis	3,886	1,614	3,287	2006	42
Cryptosporidium parvum	3,806	1,918	3,969	2006	50
Dengue virus 2 16681-PDK53	12	12	32	2012	100
Drosophila melanogaster	17,104	9,381	24,683	2006	55
Edwardsiella tarda EIB202	3,625	3,073	6,806	2014	85
Escherichia coli	4,206	3,150	5,994	2006	75
Escherichia coli K12	4,069	3,952	11,132	2011	97
Giardia Lamblia P15	5,054	5,004	21,917	2011	99
Helicobacter pylori 26695	1,552	1,528	4,744	2010	98
Homo sapiens	133,658	111,905	317,653	2013	84
Homo sapiens	63,075	58,631	219,440	2010	93
Homo sapiens	30,552	26,034	94,276	2008	85
Homo sapiens	32,010	21,270	51,076	2007	66
Human enterovirus 71	25	23	49	2012	92
Human poliovirus 1 Mahoney	11	10	29	2012	91
Lactobacillus reuteri DSM 20016	1,865	1,611	4,126	2012	86
Lactobacillus reuteri MM4-1A	2,023	1,683	4,268	2013	83
Lactobacillus reuteri SD2112	2,201	1,784	4,529	2013	81
Leishmania major	8,070	6,796	22,157	2011	84
Leishmania major	8,009	3,975	8,285	2006	50
Methanobrevibacter ruminantium M1	2,209	1,745	3,986	2012	79
Methanococcus jannaschii	1,785	1,480	1,707	2007	83
Methanopyrus kandleri	1,687	1,111	2,466	2007	66
Mus musculus	30,133	25,337	70,765	2008	84
Mycobacterium bovis BCG str. Pasteur 1173P2	3,887	3,845	11,363	2012	99
Mycobacterium leprae	1,609	1,606	4,989	2011	100
Mycobacterium leprae	1,601	1,178	2,493	2006	74
Mycobacterium smegmatis str. MC2 155	6,598	3,820	11,781	2012	58
Mycobacterium tuberculosis	3,956	3,856	12,638	2011	97
Mycobacterium tuberculosis	3,972	3,373	10,617	2011	85
Mycobacterium tuberculosis	3,954	2,808	5,913	2006	71
Mycobacterium tuberculosis H37Rv	3,966	3,936	11,746	2012	99
Mycoplasma pneumoniae	687	426	857	2006	62
Nanoarchaeum equitans	536	447	496	2007	83
Picrophilus torridus	1,535	1,260	2,902	2007	82
Plasmodium falciparum	8,547	8,064	28,260	2015	94
Plasmodium falciparum	5,342	2,599	5,053	2006	49
Plasmodium vivax	5,334	2,359	4,670	2006	44
Pseudomonas aeruginosa	5,559	3,806	9,222	2006	68
Pyrobaculum aerophilum	2,600	1,566	3,497	2007	60
Pyrococcus furiosus	2,113	1,524	3,373	2007	72
Rickettsia prowazekii	835	754	2,135	2007	90
Saccharomyces cerevisiae	6,600	3,440	9,463	2009	52
Schistosoma mansoni	12,720	8,616	26,193	2010	68
Staphylococcus aureus subsp. aureus MRSA252	2,635	1,184	3,161	2008	45
Streptococcus mutans	2,385	1,403	2,818	2012	59
Streptococcus pyogenes	1,691	1,440	3,984	2007	85
Sulfolobus solfataricus	2,922	2,006	4,451	2007	69
Thermoplasma acidophilum	1,480	1,220	2,801	2007	82
Thermoplasma volcanium	1,497	1,204	2,806	2007	80
Toxoplasma gondii	7,787	1,530	3,064	2008	20
Treponema pallidum	1,030	1,027	2,764	2009	100
Trypanosoma brucei	8,965	3,900	8,054	2006	44
Trypanosoma cruzi	19,245	7,390	14,858	2006	38
Wolbachia	805	621	1,873	2008	77
Xenopus	27,952	25,453	69,168	2007	91
Yersinia pestis	3,882	3,214	8,370	2007	83
Total	663863	511857			77