Biological Design Automation for Optimal Cell Factories

Giuseppe Nicosia

Computational Synthetic & Systems Biology Lab

Dept. of Mathematics & Computer Science

University of Catania, Italy

www.dmi.unict.it./nicosia/

BioCAD tools and algorithms

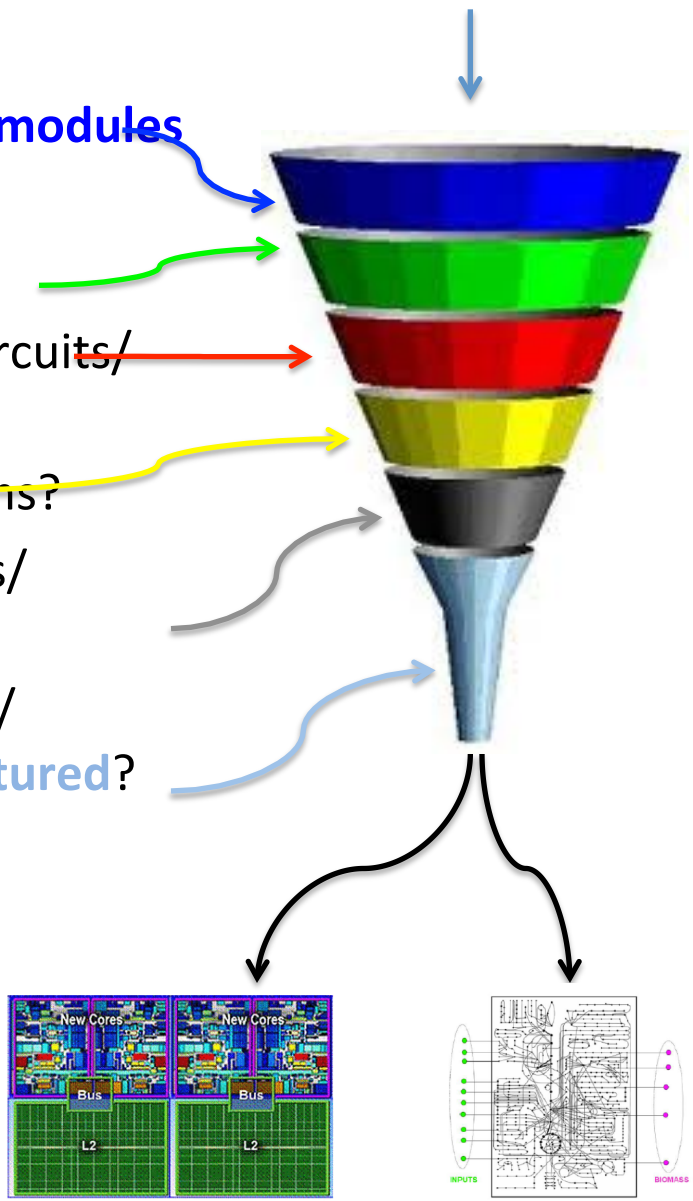
1. Optimisation
   1. Global/Local Optimisation
   2. Single/Multi-objective Optimisation (discrete for the genes and continuous for fluxes)
   3. ε-dominance Analysis
2. Sensitivity Analysis (SA)
   1. Reaction-oriented SA (RoSA)
   2. Species-oriented SA (SoSA)
   3. Pathway-oriented SA, discrete (Gene Sets) or continuous (Fluxes) (PoSA)
3. Robustness Analysis (RA)
   1. Global RA
   2. Local RA
   3. Glocal RA
   4. Pathway-oriented RA
4. Identifiability Analysis (Genotype-Phenotype relationships)

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015

Nicosia et al, "Inferring Pathological States in Cortical Neuron Microcircuits", J. *of Theoretical Biology*, 2015

Requests in Electronic Design Automation vs. Biological Design Automation

- Which are **the most important parameters/parts/modules** of the given Device/Circuit/System* ?
- How many **Feasible** Devices/Circuits/Systems?
- How many **Optimal and/or Suboptimal** Devices/Circuits/Systems?
- How many **Pareto Optimal** Devices/Circuits/Systems?
- How many **Robust Pareto Optimal** Devices/Circuits/Systems?
- Which is the set of **Robust Pareto Optimal** Devices/Circuits/Systems that can be **successfully manufactured**?



* Device = gene/protein/enzyme
Circuit = pathway/organelle
Board/Chip = cell
System (system-of-systems) = tissue/organ/organism

Mixed Integer-Discrete-Continuous Constrained Multi-Objective Optimization Problem

Find	$X=\{x_1, x_2, \ldots, x_n\} = [X^{(i)}, X^{(d)}, X^{(c)}]^T$	
To Minimize/maximize	$f_m(x)$,	$m = 1,2,\ldots,M;$
Subject to	$g_j(x) \geq 0$,	$j = 1,2,\ldots,J;$
	$h_k(x) = 0$,	$k = 1,2,\ldots,K;$
	$x_i^{(L)} \leq x_i \leq x_i^{(U)}$	$i = 1,\ldots,N.$

Where $X^{(i)}$, $X^{(d)}$, $X^{(c)}$ denotes feasible subsets of integer, discrete and continuous variables respectively. While both integer and discrete variables have a discrete nature, only discrete variables can assume floating point values (*they are often unevenly spaced*): [Li, Ui, Si] Integer and discrete variables required different handling.

If a solution x satisfies all of the (J+K) constraints and all of the 2N variable bounds, it is known as a feasible solution.

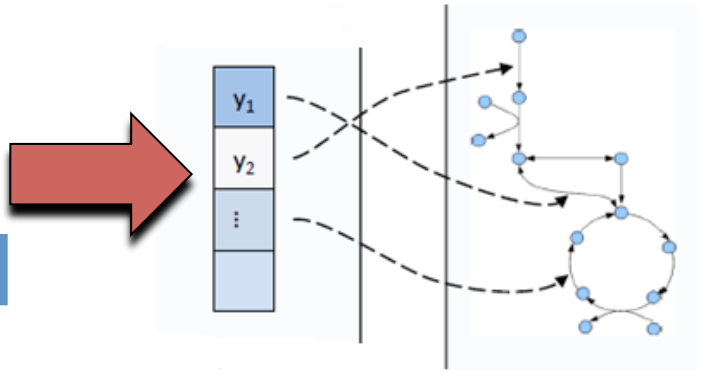
Design variables: *fluxes and/or gene sets*, or *Down- and Up- Regulation of Enzymes*
Objective functions: Biomass vs. ATP, Biomass vs. Succinate (Ethanol, 1,4Butanediol)
Constrains: $O_2=0$, GLC≥10, 5≤Ca≤10, Ph value

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

Genetic Design via MOO

$y_{34} = 1$
$y_{784} = 1$
$y_{432} = 1$

0	1	0	1	0	0	…	0

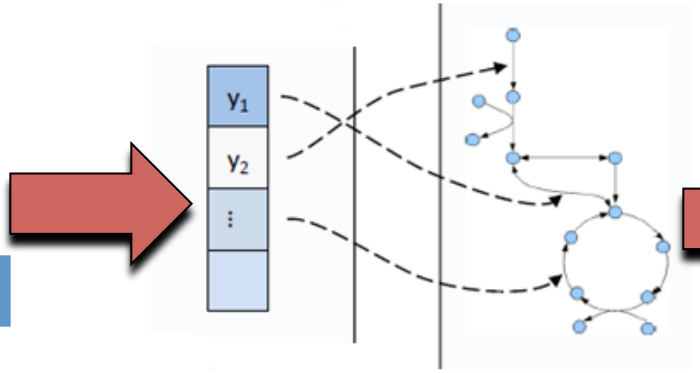


Acetate = + 66% with respect to Wild Type

Knockout = 3

$y_{63} = 1$
$y_{222} = 1$
$y_{562} = 1$
$y_{24} = 1$

0	1	0	1	0	1	…	1



Acetate = + 130% with respect to Wild Type

Knockout = 4

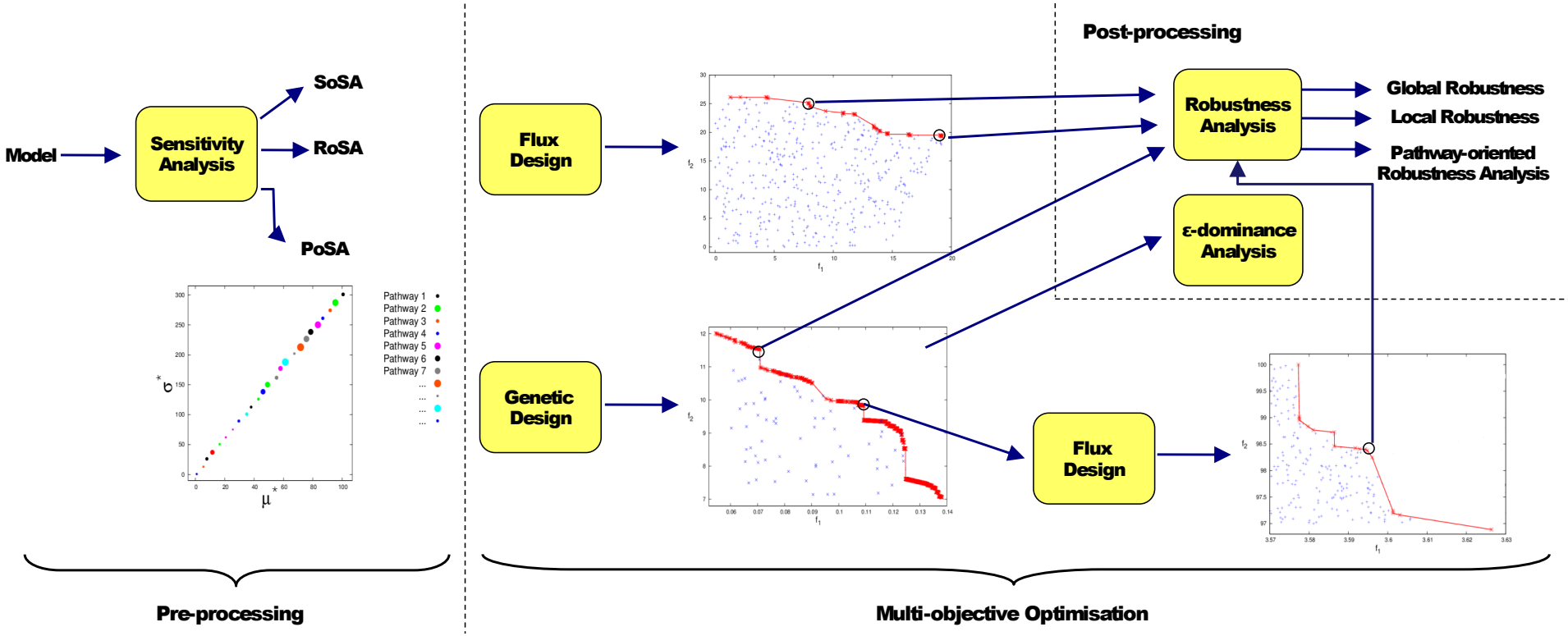
optBioCAD

Algorithm 1 optBioCAD Pseudo-code.

1: **optBioCAD** $(model, d, dup, \tau_B, \rho, \beta, s_a)$
2: $t \leftarrow 0$;
3: $BC_{arch} \leftarrow Create\_Archive(s_a)$;
4: $P^{(t)} \leftarrow Initialise(d)$;
5: $Evaluate(P^{(t)}, model)$;
6: $EvaluateConstraints(P^{(t)}, model)$;
7: **while** $\neg Stop\_Condition(t)$ **do**
8: $P_{cop} \leftarrow Copying(P^{(t)}, dup)$;
9: $P_{LS} \leftarrow Local\_Search\_Operator(P_{cop}, \rho)$;
10: $P_{GS} \leftarrow Global\_Search\_Operator(P_{hyp}, \beta)$;
11: $Evaluate(P_{GS}, model)$;
12: $EvaluateConstraints(P_{GS}, model)$;
13: $Diversity\_Enforcing(P^{(t)}, P_{GS}, \tau_B)$;
14: $BC_{arch} \leftarrow (BC_{arch} \cup P^{(t)} \cup P_{GS})$;
15: $P^{(t+1)} \leftarrow Selection(P^{(t)}, P_{GS}, BC_{arch})$;
16: $t \leftarrow t + 1$;
17: **end while**
18: **return** $(P^{(t)})$; /* output the best d candidate solutions */

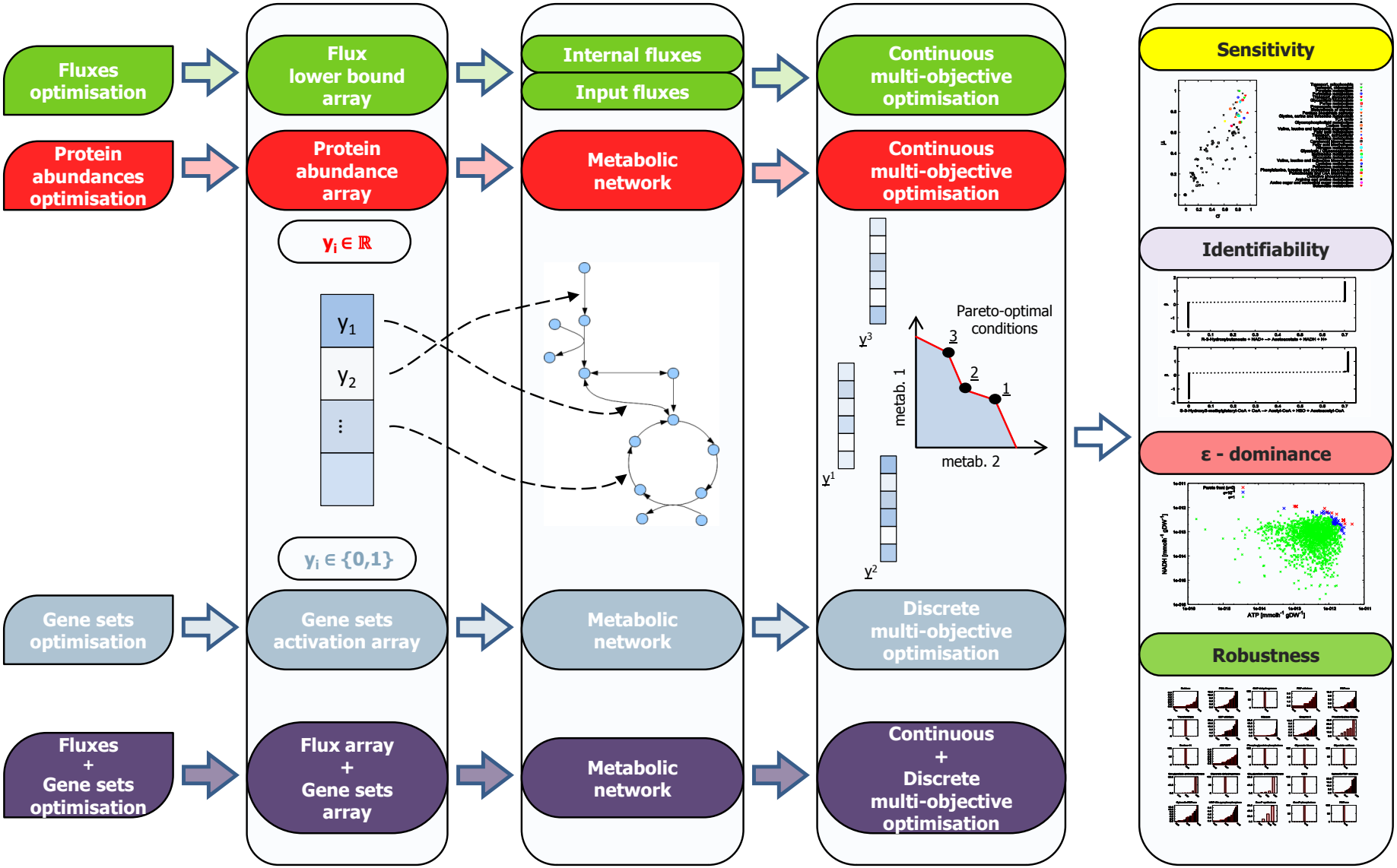
Nicosia et al, *J. Global Opt '12, ACS Synthetic Biology J* '12

Design Flow – 1/2



Nicosia et al, PLOS ONE *'15*.

The Overall Design Flow – part 2/2



Nicosia et al, PLOS ONE *'15*.

E. coli Designing

Designing Molecular Machines

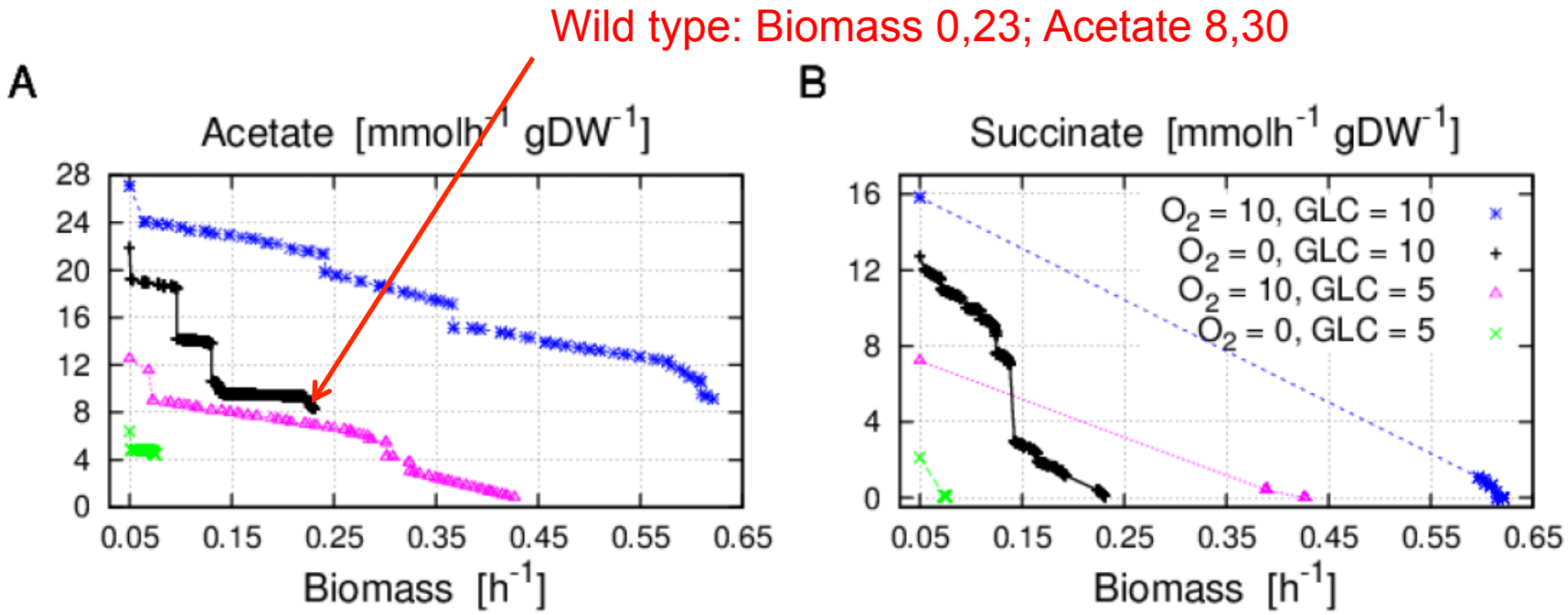
We test the computational framework in the genome-scale metabolic network of *E. coli* *i*AF1260 (Feist et al., 2007)

- 2382 reactions (299 exchange fluxes)
- 1039 metabolites
- 913 gene sets (1040 in E. coli 2011)
- 1260 genes
- 36 pathways
- 3 compartments

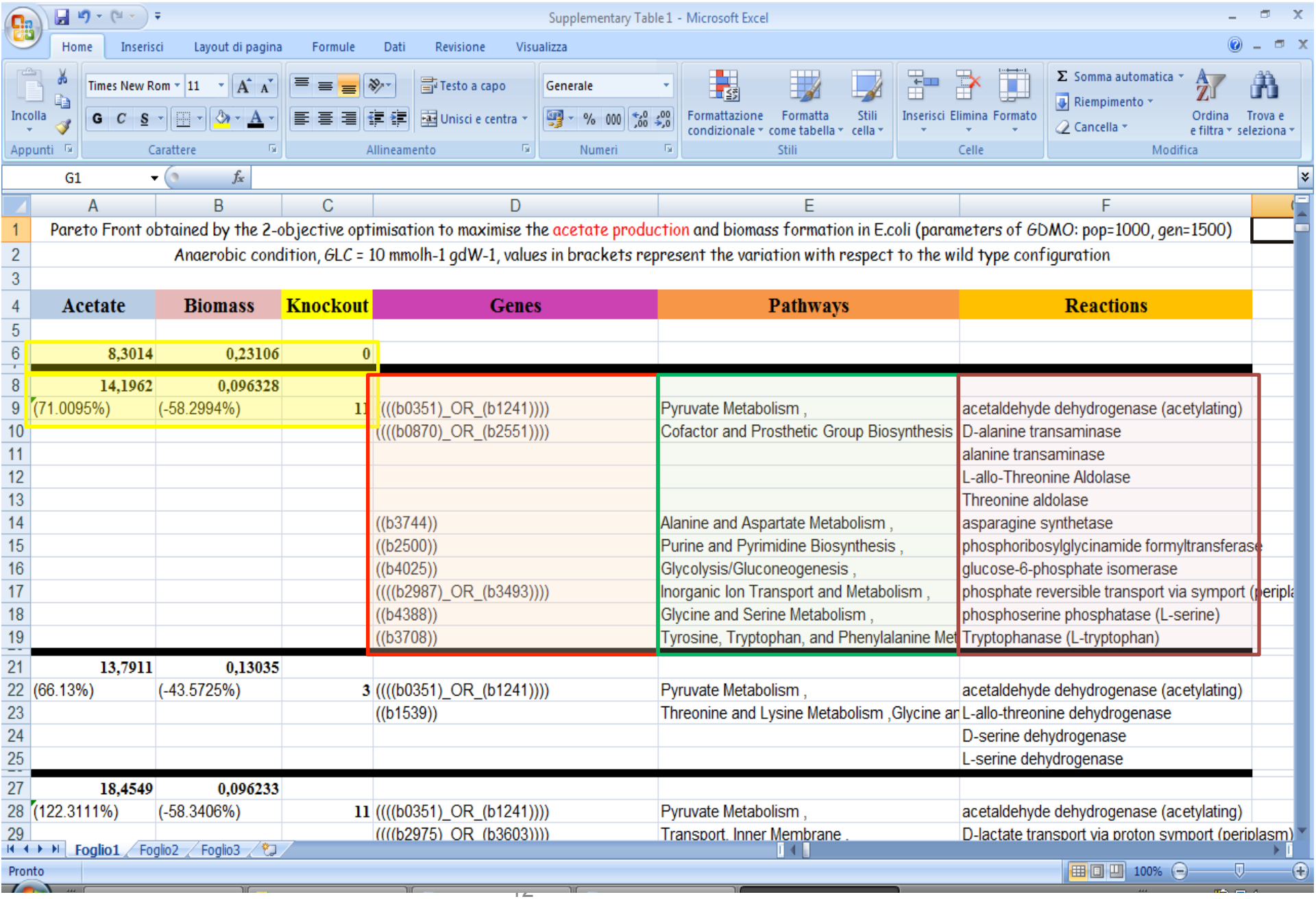
Aim: maximise a metabolite of interest (e.g., *acetate/succinate),* and simultaneously ensure the *biomass formation*, with the *minimum knockout cost*.

Nicosia et al, Theoretical Computer Science, 2015

Acetate vs. Biomass & Succinate vs. Biomass Pareto Fronts



A) *Acetate vs. Biomass maximisation* in ***different environmental conditions***
B) *Succinate vs. Biomass maximisation* in ***different environmental conditions***



Supplementary Table 1 - Microsoft Excel

Home Inserisci Layout di pagina Formule Dati Revisione Visualizza

Incolla Appunti Times New Rom 11 Carattere Testo a capo Unisci e centra Allineamento Generale Numeri Formattazione condizionale Formatta come tabella Stili cella Stili Inserisci Elimina Formato Celle Somma automatica Riempimento Cancella Ordina e filtra Trova e seleziona Modifica

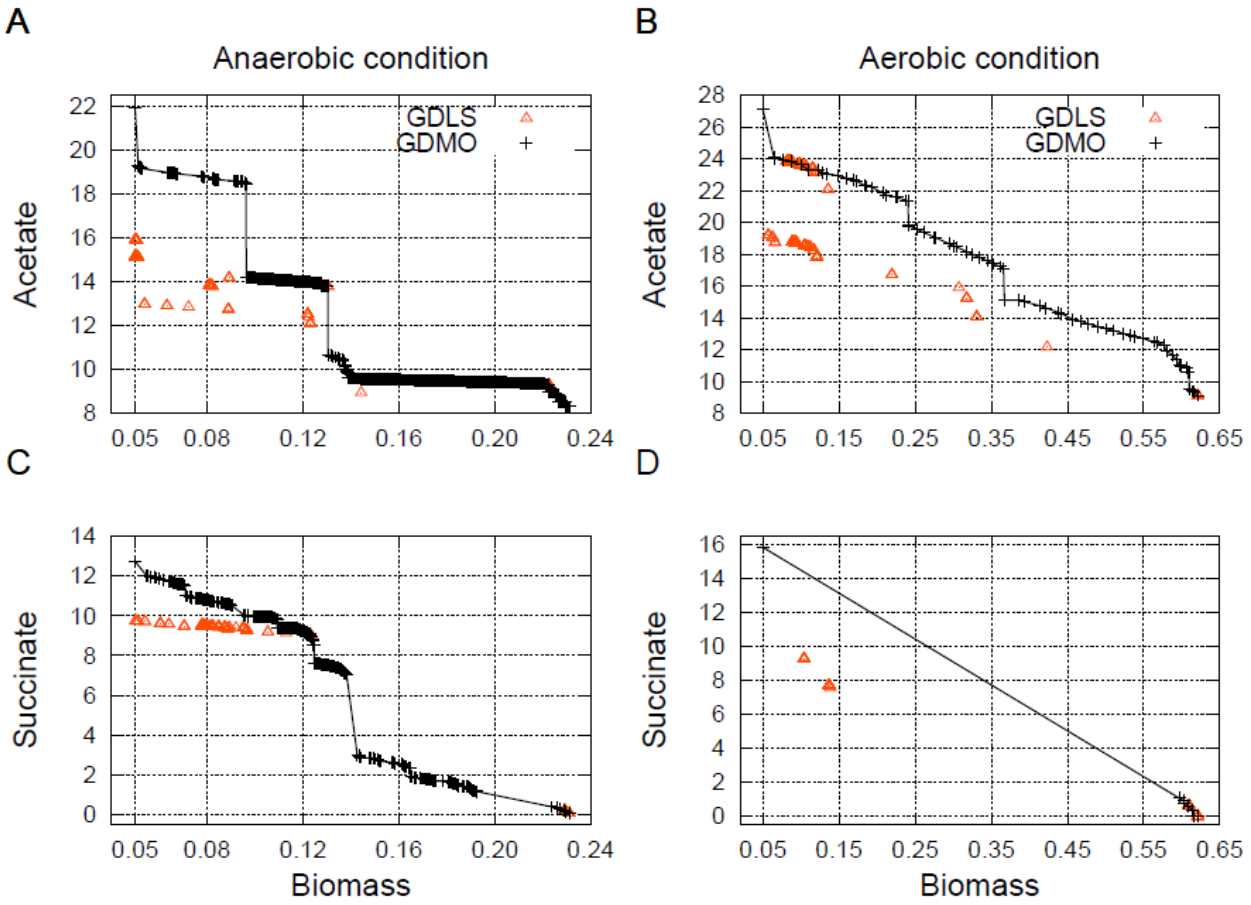
G1

	A	B	C	D	E	F
1	Pareto Front obtained by the 2-objective optimisation to maximise the acetate production and biomass formation in E.coli (parameters of GDMO: pop=1000, gen=1500)					
2	Anaerobic condition, GLC = 10 mmolh-1 gdW-1, values in brackets represent the variation with respect to the wild type configuration					
3						
4	**Acetate**	**Biomass**	**Knockout**	**Genes**	**Pathways**	**Reactions**
5						
6	**8,3014**	**0,23106**	**0**			
8	**14,1962**	**0,096328**				
9	(71.0095%)	(-58.2994%)	11	(((b0351)_OR_(b1241)))	Pyruvate Metabolism ,	acetaldehyde dehydrogenase (acetylating)
10				((((b0870)_OR_(b2551))))	Cofactor and Prosthetic Group Biosynthesis	D-alanine transaminase
11						alanine transaminase
12						L-allo-Threonine Aldolase
13						Threonine aldolase
14				((b3744))	Alanine and Aspartate Metabolism ,	asparagine synthetase
15				((b2500))	Purine and Pyrimidine Biosynthesis ,	phosphoribosylglycinamide formyltransferase
16				((b4025))	Glycolysis/Gluconeogenesis ,	glucose-6-phosphate isomerase
17				((((b2987)_OR_(b3493))))	Inorganic Ion Transport and Metabolism ,	phosphate reversible transport via symport (peripl
18				((b4388))	Glycine and Serine Metabolism ,	phosphoserine phosphatase (L-serine)
19				((b3708))	Tyrosine, Tryptophan, and Phenylalanine Met	Tryptophanase (L-tryptophan)
21	**13,7911**	**0,13035**				
22	(66.13%)	(-43.5725%)	3	((((b0351)_OR_(b1241))))	Pyruvate Metabolism ,	acetaldehyde dehydrogenase (acetylating)
23				((b1539))	Threonine and Lysine Metabolism ,Glycine ar	L-allo-threonine dehydrogenase
24						D-serine dehydrogenase
25						L-serine dehydrogenase
27	**18,4549**	**0,096233**				
28	(122.3111%)	(-58.3406%)	11	((((b0351)_OR_(b1241))))	Pyruvate Metabolism ,	acetaldehyde dehydrogenase (acetylating)
29				((((b2975)_OR_(b3603))))	Transport, Inner Membrane ,	D-lactate transport via proton symport (periplasm)

Foglio1 Foglio2 Foglio3

Pronto 100%

GDMO (us) vs. GDLS (G. Church)



Black: Pareto front obtained by GDMO.

Red: optimal results obtained by GDLS [G. Church, Lun et al.,Molecular Systems Biology, 2009]
[G. Church et al, PLOS Comp. Biol. 2013]

OptFlux (*Rocha et al, BMC Bioinf '08*)
OptGene (*Patil et al, BMC Bioinf '05*)
GDLS (*Lun, G. Church et al, MSB '09*)
OptKnock (*Bugard et al, Biotech & Bioeng'03*)
GDBB (*Lun et al, Bioinformatics'12*)

GDMO vs. OptGene, OptFlux, OptKnock, GDLS, GDBB

	Wild Type	OptFlux [1]	OptGene [2]	GDLS [3]	GDLS [3]	OptKnock [4]	OptKnock [4]	GDMO	GDMO	GDMO
Acetate	8.30	15.129	15.138	15.914	n.a.	n.a.	12.565	**13.797**	**19.150**	**n.a.**
Succinate	0.077	10.007	9.874	n.a.	9.727	9.069	n.a.	**n.a.**	**n.a.**	**10.610**
Biomass [1/h]	0.23	n.a.	n.a.	0.0500	0.0500	0.1181	0.1165	**0.1296**	**0.053**	**0.087**
K cost	n.a.	n.a.	n.a.	14	26	54	53	**3**	**10**	**8**

[1] Rocha M et al. (2008) **Natural computation metaheuristics for the in silico optimization of microbial strains**. MC Bioinformatics 9: 499
[2] Patil K at al. (2005) **Evolutionary programming as a platform for in silico metabolic engineering.** BMC Bioinformatics 6: 308
[3] Lun DS et al. (2009**) Large-scale identication of genetic design trategies using local search**. Molecular Systems Biology 5
[4] Burgard Apat al. (2003) **Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.** Biotechnology and Bioengineering 84: 647657

FBA-GPR Models of *E. Coli* (2007 & 2011)

Robustness Analysis - E. coli iAF1260

- GR: Global Robustness [1]
- LR: Local Robustness [1]
- R: Glocal Robustness[2]
- PoRA: Pathway-oriented Robustness[3]

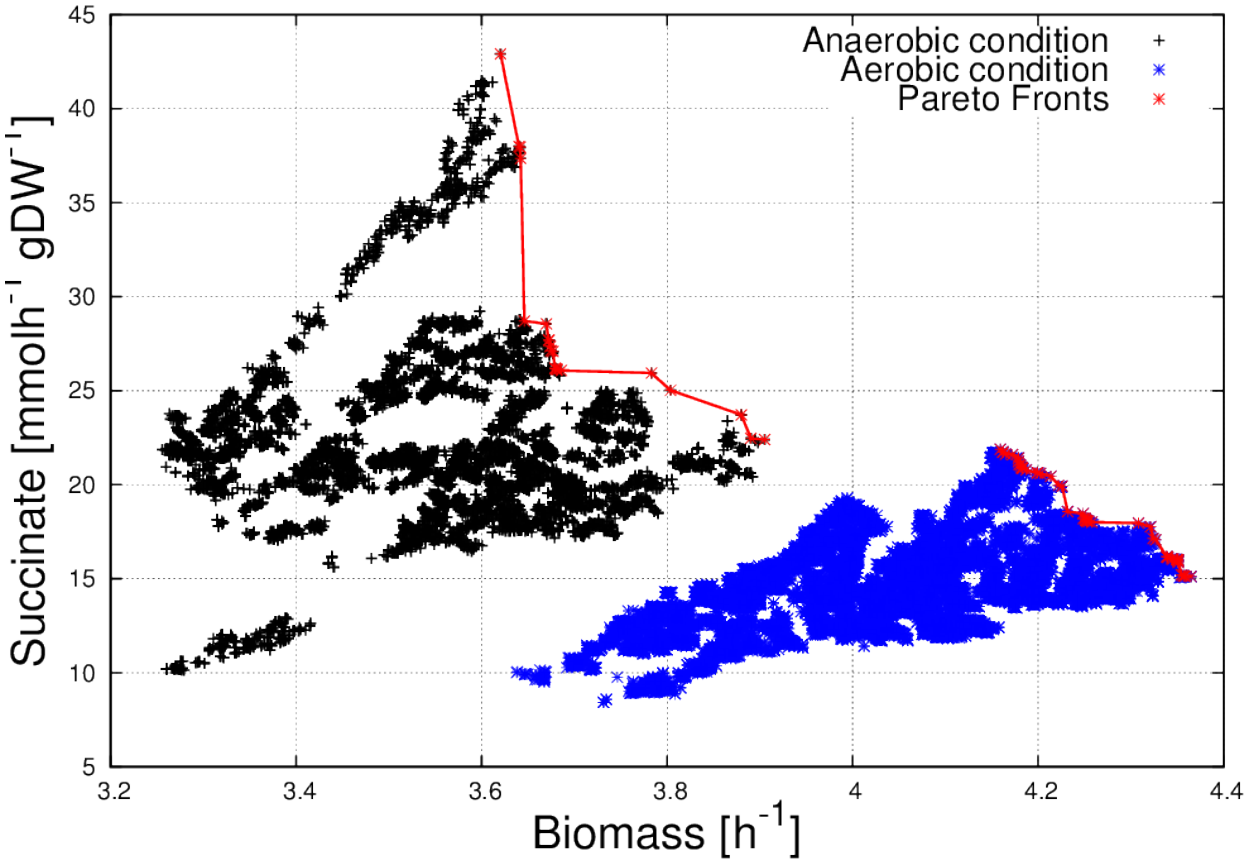
[1] Stracquadanio, G., and Nicosia, G. (2011) **Computational energy-based redesign of robust proteins.** Comput. Chem. Eng.
[2] Hafner,M. et al. (2009) **'Glocal' robustness analysis and model discrimination for circadian oscillators.** PLoS Comput. Biol.
[3] Nicosia et al. (2012) **Robust Design of Microbial Strains**. Bioinformatics J.

	Wild Type	OptFlux	OptGene	GDLS	GDLS	OptKnock	OptKnock	*GDMO*	*GDMO*	*GDMO*
Acetate	8.30	15.129	15.138	15.914	n.a.	n.a.	12.565	***13.797***	***19.150***	***n.a.***
Succinate	0.077	10.007	9.874	n.a.	9.727	9.069	n.a.	***n.a.***	***n.a.***	***10.610***
Biomass [1/h]	0.23	n.a.	n.a.	0.0500	0.0500	0.1181	0.1165	***0.1296***	***0.053***	***0.087***
~~K cost~~	n.a.	n.a.	n.a.	14	26	54	53	***3***	***10***	***8***
GR (%)	54.76/53.68	n.a.	n.a.	13.76	16.6	43.24	43.08	***45.32***	***27.6***	***40.40***
LR(%)	54.0/54.67	n.a.	n.a.	16.0	21.33	40.0	40.60	***39.33***	***24.0***	***46.0***
R	1.30/1.34	n.a.	n.a.	1.45	1.45	1.18	1.02	***0.78***	***0.44***	***1.32***
PoRA(%)	100.0/99.33	n.a.	n.a.	19.33	28.67	87.33	76.67	***81.33***	***43.33***	***83.33***

The **Robustness estimates how robust is a strain when it undergoes small perturbations**

Flux Design in E. coli iAF1260

Power law & specific operational regions



297 fluxes

Nicosia et al. (2012) **Robust Design of Microbial Strains**. Bioinformatics J.

Overproduction of BDO in *E. coli*
optBioCAD/GDMO vs. Genomatica Inc. (and BASF)

- 2 Genome-scale metabolic networks of *E. coli* (iJR904 and iJO1366) *(Palsson et al. Gen. Biol. 2003, Mol. Syst. Biol. 2011)*

  - 931/2251 reactions
  - 625/1136 metabolites
  - 904/1366 genes
  - 729/1041 enzymes

- Synthetic pathway of **1,4 butanediol** (BDO)

- Genetic and Flux Design to overproduce BDO in *E. coli*

- BioCAD software

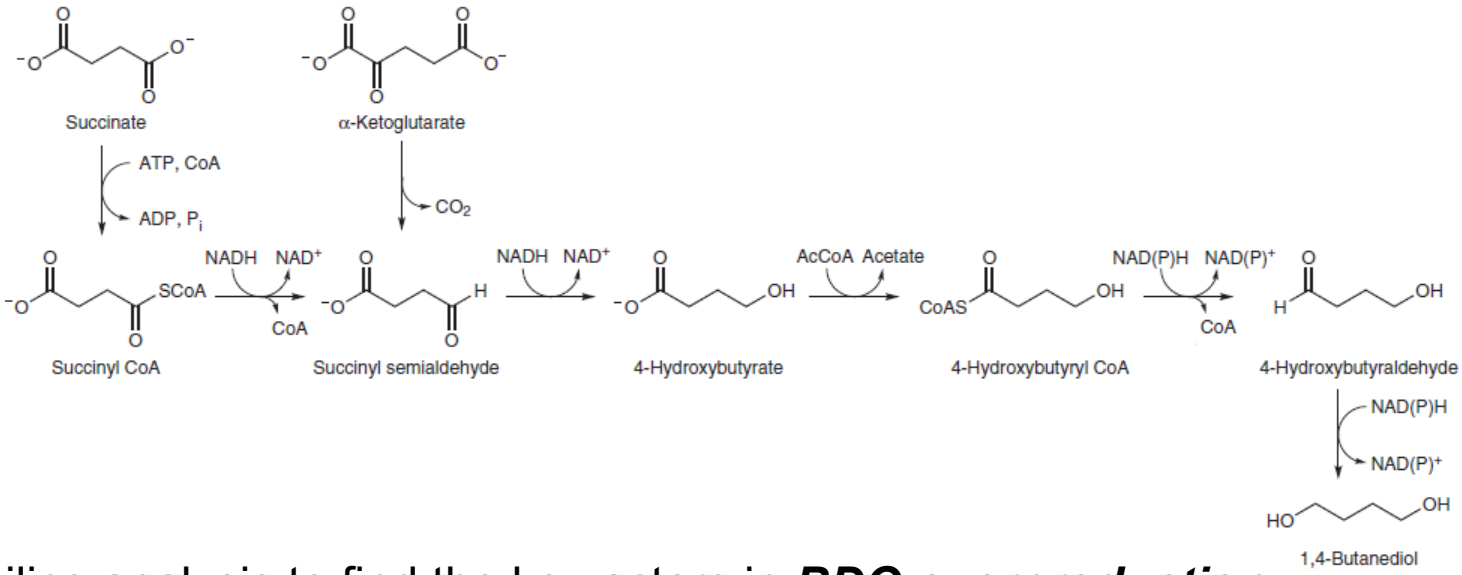
Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

1,4-Butanediol

- BDO is an inorganic compound; it is used industrially as a solvent and in the manufacture of some types of plastics, elastic fibers and polyurethanes

- BDO currently is manufactured entirely from petroleum-based feedstocks

- Inclusion of BDO synthetic pathway in *E. coli* model – ***BDO production***



- In silico analysis to find the key actors in ***BDO overproduction***

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

BioCAD results

Deletions				Max BIO_ECO	EX_14btd(e) at Max BIO_ECO
ADHEr	LDH_D			0,378	5,72
ADHEr	LDH_D	MDH		0,296	7,72
ADHEr	LDH_D	CO2t		0,363	6,88
ADHEr	LDH_D	PTAr		0,219	6,60
ADHEr	LDH_D	ACKr		0,219	6,60
ADHEr	LDH_D	THD2		0,367	6,56
ADHEr	LDH_D	PGI		0,195	6,21
ADHEr	LDH_D	TPI		0,199	6,13
ADHEr	LDH_D	FUM		0,250	6,04
ADHEr	LDH_D	C140SN		0,307	5,94
ADHEr	LDH_D	TKT2		0,375	5,86
ADHEr	LDH_D	GLCpts		0,333	5,83
ADHEr	LDH_D	GLUDy		0,352	5,78
ADHEr	LDH_D	RPE		0,376	5,78
ADHEr	LDH_D	PFK		0,360	5,76
ADHEr	LDH_D	FBA		0,360	5,76
ADHEr	LDH_D	FRD3		0,368	5,74
ADHEr	LDH_D	NADH8		0,368	5,74
ADHEr	LDH_D	CBMK2		0,374	5,73
ADHEr	LDH_D	MDH	FORt	0,140	15,17
ADHEr	LDH_D	PFLi	MDH	0,140	15,17
ADHEr	LDH_D	MDH	ATPS4r	0,203	12,17
ADHEr	LDH_D	PGDHY	PGI	0,131	11,86
ADHEr	LDH_D	EDA	PGI	0,131	11,86
ADHEr	LDH_D	FUM	ACKr	0,127	10,92

203 solutions/strains

ADHEr	[c] : accoa + (2) h + (2) nadh <==> coa + etoh + (2) nad
LDH_D	[c] : lac-D + nad <==> h + nadh + pyr
MDH	[c] : mal-L + nad <==> h + nadh + oaa
PFLi	[c] : coa + pyr --> accoa + for

ADHEr	b1241	1
LDH_D	b2133 OR b1380	2
MDH	b3236	1
PFLi	(b0902 AND b0903 AND b2579) OR (b3114) OR (b3951ANDb3952)	3

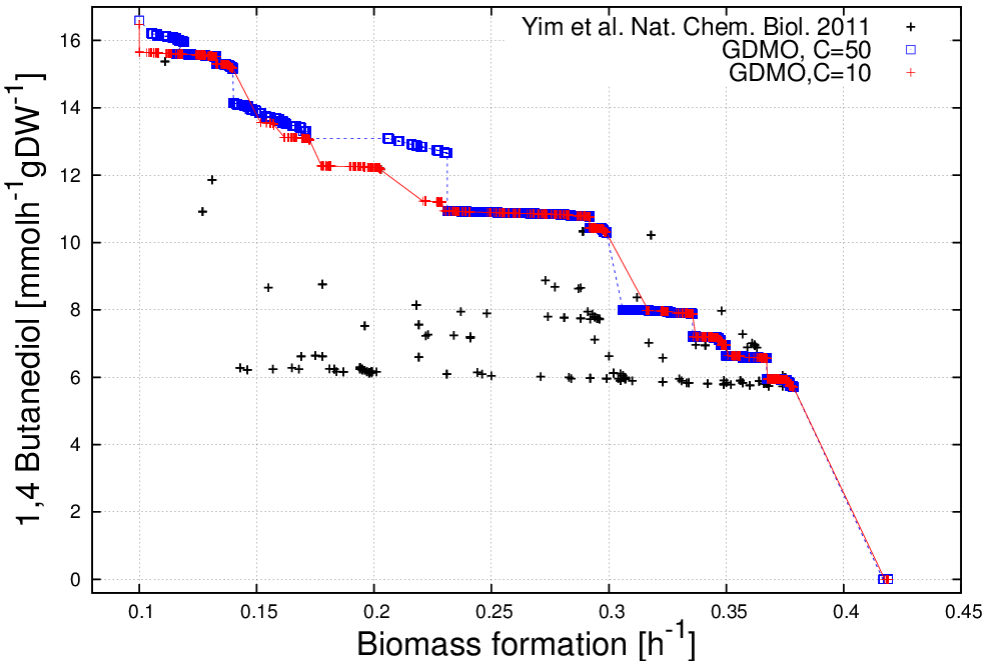
Knockout cost = 7

From Yim H. et al. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat Chem Biol. 2011

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

BioCAD results

➢ MOO to maximise BDO production and biomass formation in synthetic *E.coli* model *i*JR904[1]



Pareto fronts obtained by **GDMO** algorithm[2]

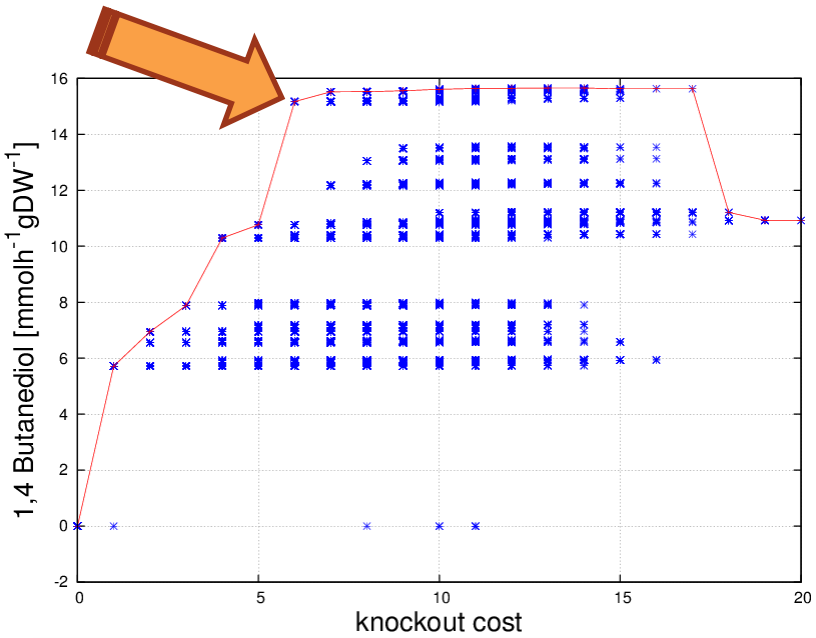
- C=10 (12836 Pareto strains)
- C=50 (49876 Pareto strains)

versus 203 solutions

C is the maximum knockout number allowable

[1] Reed et al. **An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR)** Gen. Biol. 2003
[2] Nicosia et al. (2012**) Robust Design of Microbial Strains**. Bioinformatics

BioCAD results



12836 Pareto optimal strains (**C=10)**

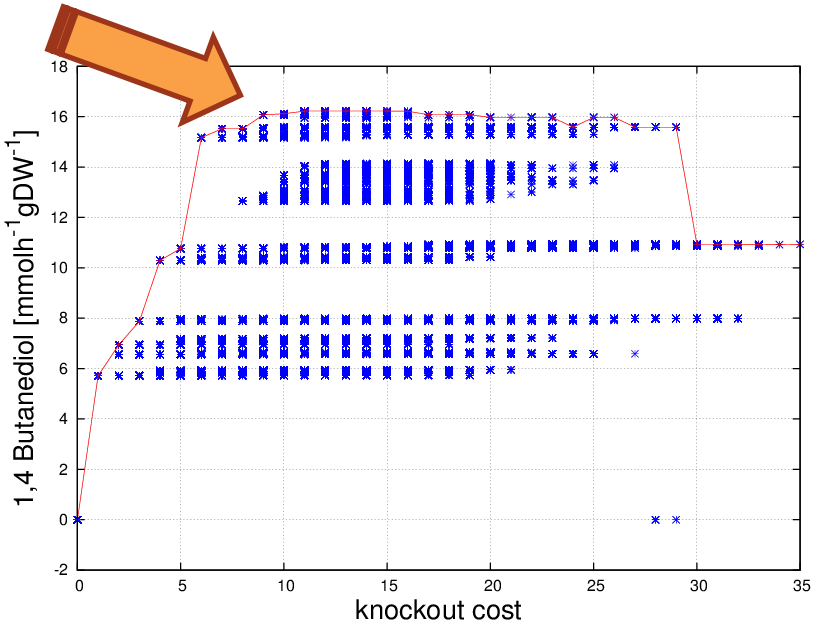
vs BDO=15,1660 & Biomass=0,1398

ADHEr	b1241
LDH_D	b2133 OR b1380
MDH	b3236
PFLi	(b0902 AND b0903 AND b2579) OR (b3114) OR (b3951ANDb3952)

BDO	Biomass	kcost	Gene knockout	Pathway	Reactions
15.1683	0.13977 (-66.6334%)	6	b1241	Alternate Carbon Metabolism, Pyruvate Metabolism	LCADi , ADHEr
			b3236	Citrate Cycle (TCA)	MDH
			b2975, b3603	Transport (Extracellular)	D-LACt2, GLYCLTt2r L-LACt2
			b2492, b0904	Transport (Extracellular)	FORt

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

BioCAD results



49876 Pareto optimal strains (**C=50**)

vs BDO=15,1660 & Biomass=0,1398

ADHEr	b1241
LDH_D	b2133 OR b1380
MDH	b3236
PFLi	(b0902 AND b0903 AND b2579) OR (b3114) OR (b3951ANDb3952)

BDO	Biomass	kcost	Gene knockout	Pathway	Reactions
16.0736	0.11551 (-72.4245%)	9	b1241	Alternate Carbon Metabolism , Pyruvate Metabolism	LCADi, ADHEr
			b2661	Arginine and Proline Metabolism	SSALy
			b3236	Citrate Cycle	MDH
			b1602+b1603	Oxidative Phosphorylation	THD2
			b0767	Pentose Phosphate Pathway	PGL
			b2975, b3603	Transport (Extracellular)	D-LACt2 , GLYCLTt2r, L-LACt2
			b2492, b0904	Transport (Extracellular)	FORt

Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

BioCAD results

- **GDMO** considers the gene-protein-reaction (GPR) mapping and each bit of y is related to a gene set

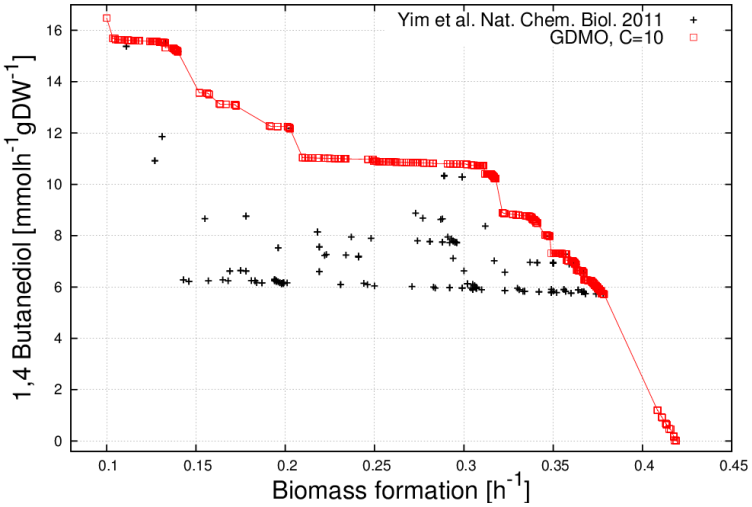
- A gene set can be formed by
  - a single gene
  - more genes linked by a Boolean relationship

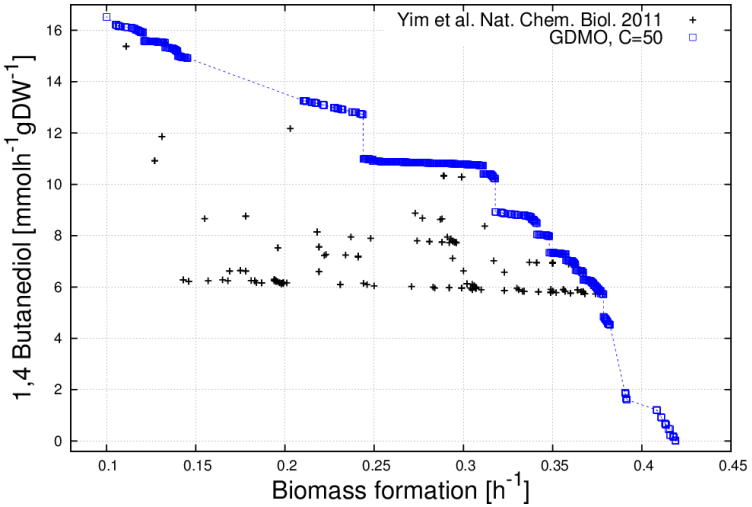
- The **method in Yim** et al. **2011 does not consider the GPR map, and turns off/on the flux of the reactions**

- In order to compare our method with Yim et al. 2011 results, **we also perform knockout research in the reaction space.**

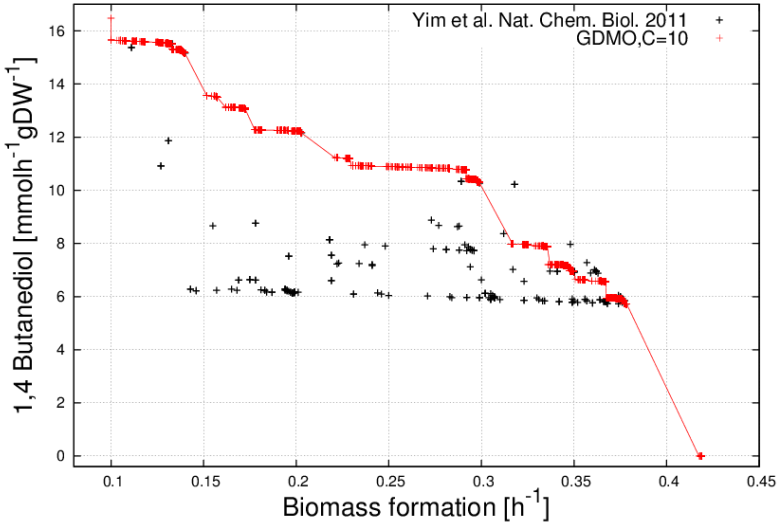
BioCAD results

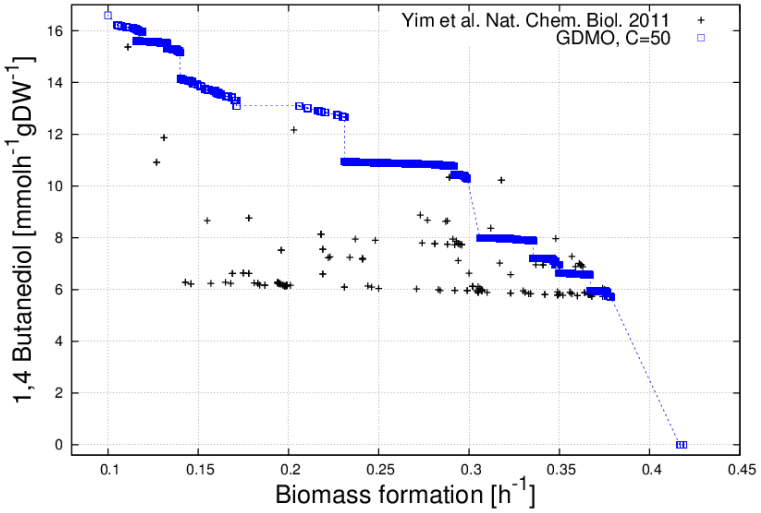
Reaction space





Gene set space





Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

Down- and Up- Regulation of Enzymes

Myristoyl-CoA Optimization for the iAF1260 E. coli: Fatty acids production

	BioCAD[**]	Redirector [*]
Biomass of the best strain	0.17 (**+21.43%**)	0.14
Myristoyl-CoA of the best strain	1.62 (**+5.19%**)	1.54
CPU time [s]	**2400s**	15200s

Anaerobic condition. Glucose uptake rate 8mmol-1 gDW-1.

Notable strain we obtain **dominates** the one obtained by *G. Church* et al [*].

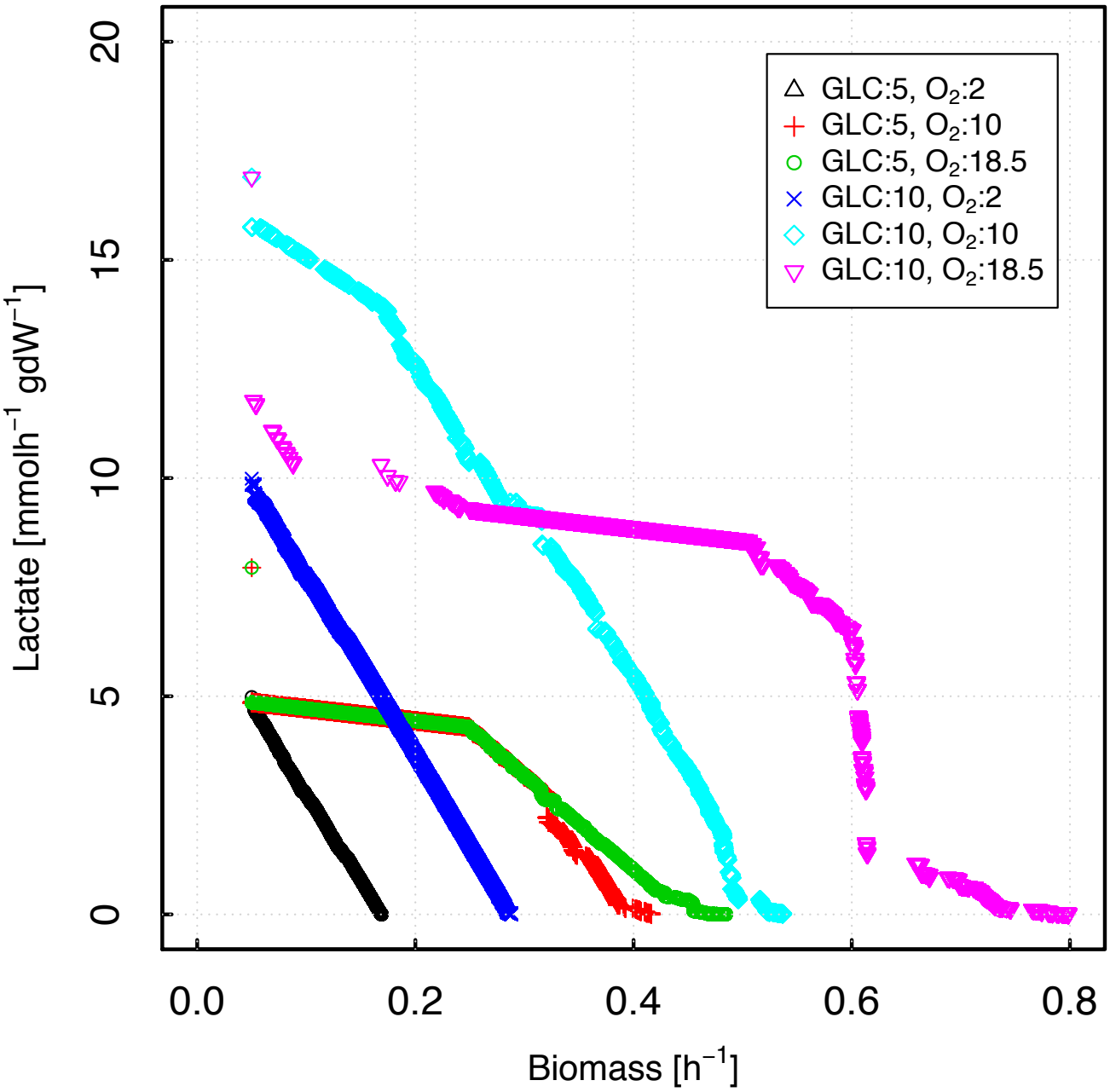
[*] G. Rockwell, N. J. Guido and G. M. Church. *Redirector: designing cell factories by reconstructing the metabolic objective.* PLoS Computational Biology 9, 1, 2013.

[**] Nicosia et al, IEEE Transactions on Biomedical Circuits & Systems, 2015.

Designing BioPlastic: Polylactic acid from Lactate acid production in S. cerevisiae

Lactate Production in S. cerevisiae by Redesigning the Metabolic Networks

- Organism: Saccharomyces cerevisiae S288c
- Model: iMM904
- Genome: PRJNA128
- Metabolites: 1226
- Reactions: 1577
- Genes: 905
- Database: http://bigg.ucsd.edu/models/iMM904/
- Publication PMID: 19321003
- Mo ML, Palsson BO, Herrgård MJ., Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol. 2009 Mar 25;3:37. doi: 10.1186/1752-0509-3-37.



Lactate Production and Biomass Optimization in S. cerevisiae

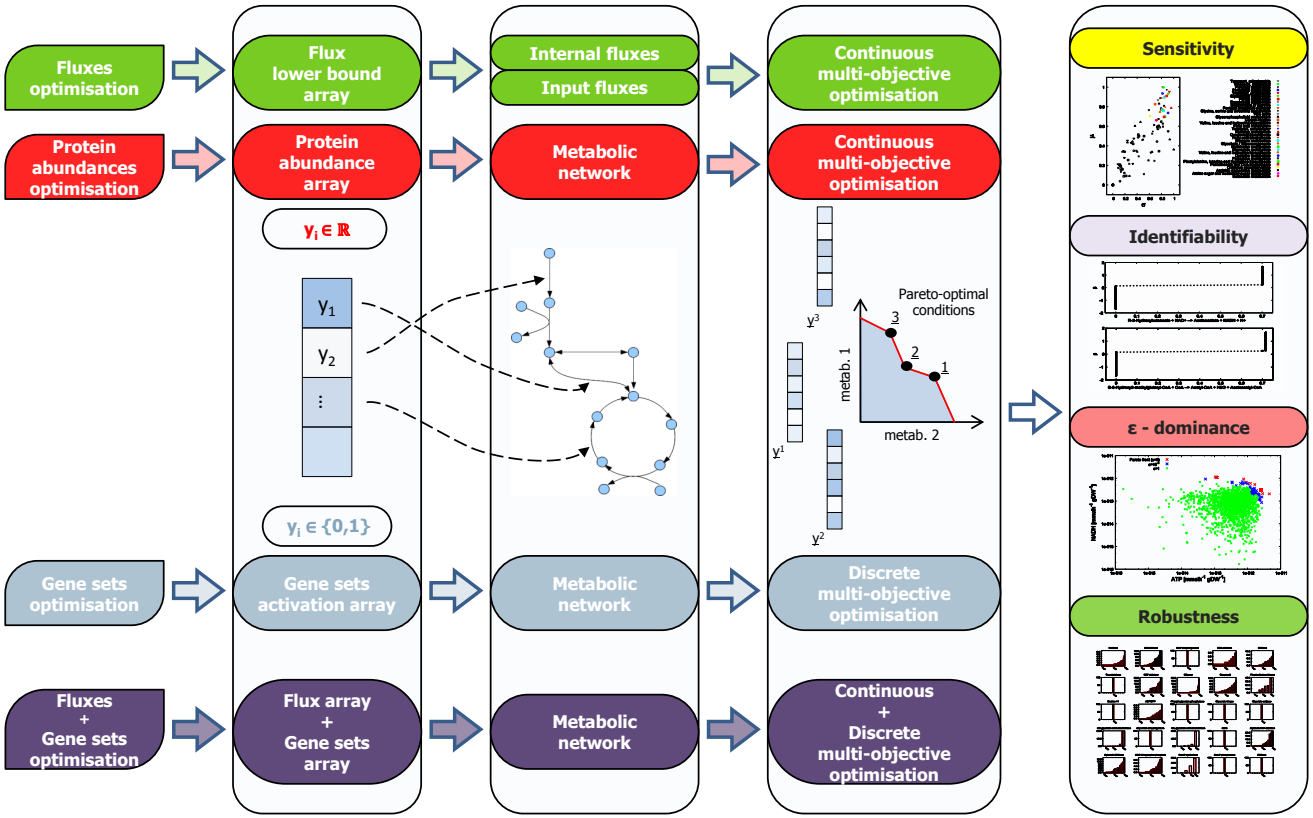
Conclusions

A complete Computational Flow for Biological Design Automation

SA: SoSA, RoSA, PoSA
IA: Genotype-Phenotype relationships
RA: LR, GR, GlocalR
Opt: SOO, MOO

Models: ODEs, DAEs, FBA, FBA-GPR.

Systems: Pathways, Organelles & Organisms



Results

1. 1,4 Butanediol Production in E. coli
2. ATP maximixation in the Mitochondrion
3. Down- and Up- regulation of Enzymes for Fatty acids production
4. BioPlastic production by Engineered Yeast

Acknowledgments

Jole Costanza, Center for Genomic Science@IIT, Milano, Italy

Giovanni Carapezza, Nerviano Medical Sciences, Catania, Italy

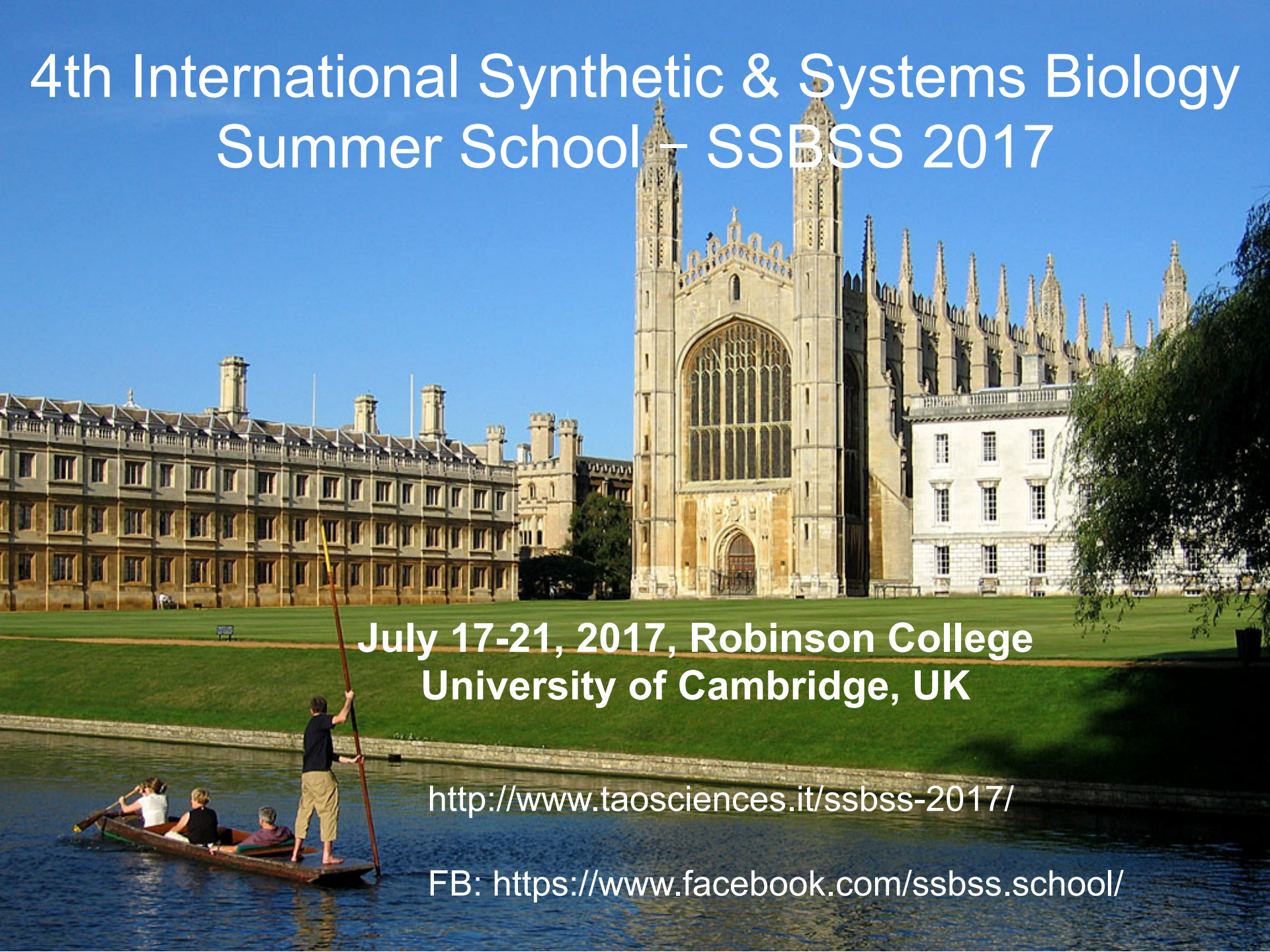
Piero Conca, CNR − Catania, Italy

Giorgio Jansen, University of Catania, Italy

Andrea Patanè, University of Catania, Italy

Andrea Santoro, University of Catania, Italy

4th International Synthetic & Systems Biology Summer School – SSBSS 2017



July 17-21, 2017, Robinson College
University of Cambridge, UK

http://www.taosciences.it/ssbss-2017/

FB: https://www.facebook.com/ssbss.school/