Multi-objective MINLP Optimization Used to Identify Theoretical Gene Knockout Strategies for *E. coli* Cell

G. Maria,[a,*] Z. Xu,[b] and J. Sun[b]

[a]Department of Chemical & Biochemical Engineering,
Polytechnic University of Bucharest, Polizu 1, Bucharest, Romania

[b]Key Laboratory of Systems Microbial Biotechnology,
Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences,
Xiqidao 32, Tianjin Airport Economic Area, Tianjin 300308, China

Original scientific paper
Received: April 14, 2011
Accepted: December 27, 2011

Bioprocess optimization by genetically modifying the microorganism characteristics is an intensively investigated subject due to the immediate economic interest. A large variety of alternatives using elaborated experimental procedures, accompanied by *in-silico* cell design based on topological and dynamic models have emerged. The present study investigates the possibility of using a mixed-integer nonlinear programming (MINLP) approach to determine optimal metabolic fluxes in respect of multi-objective criteria associated to gene knockout strategies. The advantage of the proposed power-law type criterion stems from the possibility of accounting, in a simple way, for the flux nonlinear interactions and complex constraints. The combinatorial rule is included in the iterative MINLP solver, while a large number of constraints could increase the chance of obtaining a reduced set of viable gene-knockout solutions for a given metabolic network. Multiple gene deletion alternatives are thus identified, allowing a high cell growing rate with maximizing externally imposed chemical production targets. Exemplification is made for the case of designing an *E. coli* cell that realizes maximization of succinate production by using a reduced model from literature. Comparatively to the linear procedure that solves a combinatorial problem in a bi-level optimization approach, of dimensionality sharply increasing with the number of removed genes, the MINLP alternative considers an adjustable nonlinear influence of fluxes to the main goal, its performance being less dependent on the number of knockout genes.

Key words:
Flux balance analysis, MINLP, gene knockout, succinate production, *E. coli*

Introduction

Over the last decade, biotechnology and bioengineering have been developing new research directions for improving the metabolic performances of microorganisms used in the processing industry. The new approach, known in literature as "from gene to product concept",[1] is based on the application of fundamental scientific knowledge (biology, biochemistry) and engineering science approaches to understand the cell metabolism, species interactions, and the genetic regulatory circuits responsible for regulation of cell biochemical reactions. The result is the *in-silico* design of new microorganisms, genetically modified, by conferring new properties and functions to the mutant cells (i.e. desired 'motifs'), with applications in various fields, such as improving industrial bioprocesses (biosynthesis, pollutant biotreatment, drug industry), designing new metering devices (biosensors, bioindicators), or in medicine (gene therapy).

Numerical simulation of metabolic cell processes, at a topological or dynamic (kinetic) level, in a holistic, modular, compartmented, lumped or extended approach is necessary for the *in-silico* design of modified microorganisms (based on mathematical models), combining knowledge from various modern fields, such as synthetic biology, systems biology, genetic circuit engineering, and molecular bioengineering.[2–7] Living cells are evolutionary, autocatalytic, self-adjustable structures able to convert nutrients from the environment into additional copies of themselves during the cell cycle. In spite of tremendous progress made in cell process analysis and the development of –omics databanks, various approaches exist in analysing and modelling the genome replication, cell metabolism, and the multiple regulatory functions of the cell syntheses. Different analyses are justified by the very high complexity of metabolic processes, implying thousands of species and tens of thousands of (self)-catalytic reactions, enzymes, co-enzymes, activators and inhibitors, proteic oligomers, intermediates, regulatory and signaling chains, motility, membranar

*Corresponding author: mail address: P.O. 35-107 Bucharest, Romania; gmaria99m@hotmail.com

and internal transport, gene transcription, morphogenesis, and cellular differentiation, all in an inter-connection difficult to decipher. As a consequence, topological cell models have been preferred, as a first step, by applying the 'Metabolic Control Analysis' (MCA)[8,9] to investigate the sensitivity of the stationary cell system vs. external perturbations.

The engineering approach to the analysis and design of new organisms is based on developing dynamic simulators of the cellular processes by applying physico-chemical laws and principles, and including both stationary and kinetic information on metabolic processes. However, the availability of enzymatic reaction kinetic information was fragmented, and consequently such dynamic models rarely include more than hundreds or thousands modelled biochemical reactions. In contrast, due to the very large quantity of qualitative, less structured –omic information, attention has turned to developing methods for analyzing the relative importance of various metabolic events / reactions. In this context, the developed metabolic 'flux balance analysis' (FBA, or 'Metabolic Flux Analysis'),[17] and the 'elementary mode analysis' (EMA)[10] allowed evaluation of cell metabolism efficiency and determination of how to use the resources, as well as the use of the minimum set of enzymes required by the cell growth to preserve the physiological functions and system invariants (so-called 'modes' derived from the null space of the stoichiometric matrix). Similar to EMA, the 'extreme pathways analysis' (ExPA)[11] determines the 'solution space' within which fall all possible steady-state flux distributions of the network, by means of a constraint-based approach (using mass balance and maximum reaction rates). The algebraic calculus is based on Kirchhoff's first law (that the production and utilization rate of a metabolite must balance in steady-state), and second law (that the free-energy change around a biochemical loop must be zero). Rigorous statistical methods can be applied to identify the relations between metabolites in a network,[12,13] by performing a "modal matrix" analysis to assess which metabolites could be grouped ("pooled"), and develop a reduced topological and dynamic model by using those pooled metabolites. In the end, they recommend the type of aggregate variables to be used for kinetic model development when sufficient information is unavailable. Kauffman *et al.*[14] used this "modal matrix" analysis to pool metabolites and extract the dynamic characteristics of a biological network (e.g. the human red blood cell metabolism). They show how "dynamic phase planes, statistical time-lagged correlation analysis, and temporal decomposition" can be used to relate the biochemical mechanistic details and the overall metabolic functions. As such methods are based on analysing the steady-state metabolism of a cell, hybrid stationary-dynamic methods have also been developed. For instance, Mahadevan *et al.*[15] introduced the 'dynamic flux balance analysis' (DFBA), which incorporates rate of change of flux constraints from analysing the evolution of flux distribution over time. Dynamic interpretations of the flux control (sensitivity) coefficients of the MCA have also been studied (e.g. Tušek and Kurtanjek[16]).

FBA is a classical but still very powerful method for determining the stationary distribution of metabolic fluxes (for a known metabolic pathway), and also for relating any change in the environmental conditions or in the cell structure (including genome modification) to the way of using the nutrients inside the cell.[17–19] FBA is based on the stoichiometric mass balance constraints under steady-state conditions, of type $\boldsymbol{S} \cdot \boldsymbol{v} = 0$ (where $\boldsymbol{S}$ is the stoichiometric matrix including stoichiometric coefficients of the metabolites in the reaction pathway, and $\boldsymbol{v}$ is the vector of stationary metabolic fluxes, including internal, transport, and the growth fluxes). In FBA, exchange fluxes are assigned to those metabolites that enter or leave the particular network only, with constraints ranging from negative to positive values. Those metabolites that are consumed within the network are assigned no exchange flux value. As the number of fluxes is much higher than the number of measured fluxes and mass balance constraints, the feasible set of solutions is defined by the intersection of the null space (i.e. vector space for $\boldsymbol{v}$) and max-min type of constraints imposed to the fluxes, i.e. $\boldsymbol{v}_{j,\min} \leq \boldsymbol{v}_j \leq \boldsymbol{v}_{j,\max}$, $j = 1,\ldots, M$ (where M is the number of fluxes in the considered metabolic pathway).[20] In principle, if a sufficient number of constraints (including the kinetic details) are available, a "single point" solution in the flux-space can be obtained. In practice, measurements of some net stationary fluxes allow obtaining the least squares solution of the problem with linear constraints, by means of a simple matrix calculation.[18] The FBA stoichiometric constraints can also be used to correct (reconcile) the measured fluxes affected by gross errors.[21] More elaborated and precise approaches use both stationary and dynamic information on the cellular utilization of a C-labelled substrate (either ^{13}C or ^{14}C) to extract supplementary information on metabolic fluxes in terms of isotopomer distribution.[22–25,52] In fact, FBA and MCA are closely connected because modification of stationary fluxes due to a certain perturbation factor (or cell modification, e.g. by gene knockout) can be transcribed in terms of stationary sensitivities.[26]

FBA can highlight the most effective and efficient pathway through the network in order to

achieve a particular objective function. In fact, multiple stationary flux solutions exist according to the environmental conditions and cell adaptation characteristics. The multiple cellular objectives are associated to the system action to perform regulatory, metabolic, homeostatic, and phenotypic functions that realize the best tradeoff between proliferation and differentiation from one side, and cellular functions and robust growth from other side, with optimal use of available resources. For instance, Selvarasu *et al.*[55] use a weighted multi-objective optimization rule to identify synergistically switching pathways for multi-product strain improvement.[56]

To solve this complex problem, various multi-objective optimization strategies have been proposed, because it is still unclear how to combine and/or prioritize mutually competing objectives to achieve a true optimal solution, or how to select from the very large number of Pareto optimal solutions those that realizes the "best" tradeoff for the cell designers' preferences. A Pareto-solution is one where any improvement in one objective can only take place at the cost of another objective; for continuous variables an infinity of Pareto-optimal solutions might exists.[48] On the other hand, a significant re-routing of flux directions and cycle fluxes are reported when switching from one objective to another within system constraints.

Following the review of Nagrath *et al.*[53] to find an optimal set of stationary net fluxes for a defined number of genes encoding the enzymes that participate in the metabolic reactions, several methods can be followed. One alternative is to use the linear programming (LP) to find the maximum of a weighted linear combination of fluxes of type $Max(\boldsymbol{w}\,\boldsymbol{v})$ (where vector $\boldsymbol{w}$ includes the chosen weights).[27,28] However, in contrast to EMA and ExPA, only a single solution results in the end, even if additional linear constraints (other than the stoichiometric balance $\boldsymbol{S}\,\boldsymbol{v} = 0$) limiting the fluxes are added to the LP formulation. By varying the weights, or by applying an iterative weighting procedure, a still reduced number of Pareto-solutions are usually obtained (especially when Pareto-frontier is non-convex). Besides, the weight selection (usually between 0 and 1 for scaled objectives) in association to physical meanings is difficult. A similar route, the so-called 'Goal Programming', uses sets of upper and lower weights to optimize the composite objective. Even if a larger number of Pareto-optimal solutions are thus obtained, the method suffers from the same disadvantages. Alternatively, the 'Linear Physical Programming' method[27] replaces the a-priori prioritization of cellular objectives (fluxes) by successively relaxing (smoothing) the explicit flux constraints, from very strong ("highly desirable") to very soft ("unacceptable"), and by minimizing the weighted distances from the solution to the boundaries. The number of classes defining the preference degree for each objective is still a subjective decision, somehow equivalent to inspecting various weights in the multi-objective optimization, but realized in a more comprehensive way. The procedure leads to a larger number of Pareto-optimal solutions due to the possibility of gradually relaxing the associated LP problem with an increased physical significance of the imposed constraints.

A similar transformation of the multi-objective optimization in a LP problem in the presence of stoichiometric, enzyme maximum amounts, and solvent capacity constraints, was presented by Vera *et al.*[34] By performing a nonlinear (logarithmic) transformation of the power-law reaction rate expressions (the so-called S-systems), the steady-state metabolic fluxes can be optimized vs. species concentrations and enzyme activities (by means of an evolutionary algorithm).[54] Even if attractive, the procedure requires the knowledge of kinetic expressions of the involved reactions, while the S-type representation of cellular processes, even if computationally convenient, suffer from a number of limitations. [4,57]

Often several Pareto optimal flux solutions are available in cellular systems, representing alternative designs, from which one can subjectively be selected to offer the best trade-off among multiple objectives. An important application is the design of mutant cells by testing the effect of gene knockouts on stationary metabolic fluxes. The enzymatic flux that correlates to the gene that needs to be removed is given a constraint value of zero; then, the reaction that the particular enzyme catalyzes is completely removed from the analysis. Various strategies of gene deletions can be tested, by inspecting the feasibility of the solution in respect to certain constraints or objective, in such a manner. FBA was used in practice, to inspect the successive deletion to only one gene,[20,29] by retaining the optimal solution vs. a certain linear objective function. The simultaneous deletion of more than one gene leads to a very extended combinatorial problem, a total number of C_M^{KG} trials being necessary for KG genes removed from the total of M genes.

It is important to mention that while the gene knockout procedure is trying to drain the cell resources to the over-production of the desired metabolites by cutting alternative metabolic pathways, other proposed cell optimization techniques try to re-design complex regulatory circuits to compensate the removed cell functions.

A completion of the FBA method for solving the gene-knockout problem is the 'minimization of

metabolic adjustment' method (MOMA).[27] MOMA employs quadratic programming to identify the closest point in the flux space to the FBA wild-type point, compatible with the gene deletion constraint. MOMA displays a significantly higher correlation than FBA, being of use for predicting the behavior of the perturbed metabolic networks, whose growth performance is generally sub-optimal. However, as observed by Wunderlich and Mirny,[30] FBA, EMA, and MOMA are unable to separate the role of topology and other parameters in network function, while EMA is computationally very expensive and provides "little insight into why certain mutations are lethal, whereas others are tolerated". They proposed to use the so-called 'total synthetic accessibility' index $S_t = \Sigma_j S_j$, evaluated from summing the synthetic accessibility of the outputs, defined as the minimal number of metabolic reactions needed to produce component j from the network inputs. If an enzyme knockout does not change the index S_t, that is the biomass which can be produced without extra metabolic cost, then the mutant is viable. If S_t becomes infinite, at least one essential component of the biomass cannot be produced from network inputs, and therefore the gene knockout leads to a lethal phenotype.

Finally, it should be mentioned that "topology plays a central role in determining network function and malfunction",[30] and the viability of sets of mutants. However, the gene knockout through FBA is intrinsically incomplete as long as it is difficult to separate the contribution of topology from the contributions of kinetic and equilibrium characteristics of the system. Also, inferences among genes are not accounted for, while validation of FBA conclusions by means of simulating gene regulatory circuits, and dynamic response to perturbations is necessary.[4,6,31,32]

Because the main bioengineering objective through gene knockouts (usually maximization of production of a certain metabolite) is associated with a cellular objective (usually maximization of biomass yield), one worthy alternative is to formulate the problem as a bi-level programming problem.[33] Such an approach is justified by the observation that the yields for some metabolites are far below their theoretical maximum given certain nutrient flux entering into the cell. Linear constraints of the bi-level optimization problem impose fixed substrate uptake, fulfillment of the network stoichiometric balance, upper/lower limits of fluxes, and other balance relationships. Following the mathematical rules, the 'primal' problem, aiming at maximizing the bioengineering objective subjected to maximizing the cellular objective in the presence of linear constraints, is equivalent to solving the associated 'dual' problem, of LP type, aiming at maximizing only one composed objective function in the presence of the original and additional constraints. When a gene knockout strategy is investigated, Boolean variables are added to each flux, leading to a mixed-integer LP (i.e. MILP) problem. By limiting the number of knockouts, the solution consists of a set of retained genes and the associated optimum values of stationary fluxes. Exemplification of this procedure (OptKnock) is made for optimizing production of succinate and lactate in *E. coli* cells by using the Edwards and Palsson[20] central metabolism model of 720 reactions. Even if effective, application of OptKnock requires solving a very large combinatorial problem, and multiple solutions might exist for the same objectives. Besides, the right choice of max/min boundary values of the auxiliary variables increases difficulties in applying the procedure.

The aim of this paper is to investigate the possibility of using a mixed-integer nonlinear programming (MINLP) approach in solving the multi-objective cell metabolism optimization problem, in a nonlinear power-law formulation $Max(\Pi \boldsymbol{v}_j^{\beta_j})$, by including the adjustable influence of fluxes for reaching the composite goal. Even if being similar to the weighted multi-objective and goal optimization (if a logarithmic transformation is applied), the procedure includes the possibility of accounting for nonlinear interactions among fluxes and nonlinear constraints without loosing any property by transformation. The MINLP procedure was then used to identify multiple gene deletion combinations that allow a maximum cell growth rate with maximizing externally imposed chemical (product) production targets, as an alternative to the combinatorial MILP procedure. The algorithm is simple to apply, requires no specification of auxiliary variables, and is easily extendable to solve a variety of nonlinear multi-objective optimization problems in a simple way.[9,34,35]

Exemplification is made for the case of finding sets of knockout genes to ensure maximization of succinate and biomass production in *E. coli* cells, by using the Edwards and Palsson[20] central metabolism model (the reduced variant of 95 reactions).[36] As multiple feasible solutions exist, a step-by-step increase of the number of constraints might lead to a reduction in the gene knockout alternatives when using the MINLP formulation. Even if only linear constraints have been included in the tested case study, the nonlinear multi-objective formulation can be easily extended by accounting for flux interdependencies, the use of energy charge, carbon and nitrogen recoveries at steady state, or cell regulatory / thermodynamic properties, thus allowing reduction of the number of viable solutions.

Tested metabolic model of *Escherichia coli* (central metabolism)

The approached case study is the stoichiometric model of central metabolism of *E. coli* K-12 of Edwards and Palsson[20] and Orth *et al.*[36] The reduced variant includes 72 metabolites (Table 1) participating in 95 reactions (Table 2, Fig. 1), the stationary net fluxes being limited by specified minimum/maximum values (–1000 / +1000 mmol gDW^{-1} h^{-1}). The

Table 1 – *Considered metabolites in the Edwards and Palsson[20] model (central metabolism of Escherichia coli)*

Abbreviation	Metabolite	Formula	Charge	CAS Number
13dpg	3-Phospho-D-glyceroyl phosphate	$C_3H_4O_{10}P_2$	–4	38168-82-0
2pg	D-Glycerate 2-phosphate	$C_3H_4O_7P$	–3	None
3pg	3-Phospho-D-glycerate	$C_3H_4O_7P$	–3	None
6pgc	6-Phospho-D-gluconate	$C_6H_{10}O_{10}P$	–3	None
6pgl	6-phospho-D-glucono-1,5-lactone	$C_6H_9O_9P$	–2	None
ac	Acetate	$C_2H_3O_2$	–1	71-50-1
ac[e]	Acetate (extracellular)	$C_2H_3O_2$	–1	71-50-1
acald	Acetaldehyde	C_2H_4O	0	75-07-0
acald[e]	Acetaldehyde (extracellular)	C_2H_4O	0	75-07-0
accoa	Acetyl-CoA	$C_{23}H_{34}N_7O_{17}P_3S$	–4	72-89-9
acon-C	cis-Aconitate	$C_6H_3O_6$	–3	585-84-2
actp	Acetyl phosphate	$C_2H_3O_5P$	–2	19926-71-7
adp	ADP	$C_{10}H_{12}N_5O_{10}P_2$	–3	58-64-0
akg	2-Oxoglutarate	$C_5H_4O_5$	–2	328-50-7
akg[e]	2-Oxoglutarate (extracellular)	$C_5H_4O_5$	–2	328-50-7
amp	AMP	$C_{10}H_{12}N_5O_7P$	–2	61-19-8
atp	ATP	$C_{10}H_{12}N_5O_{13}P_3$	–4	56-65-5
cit	Citrate	$C_6H_5O_7$	–3	77-92-9
co2	CO_2	CO_2	0	124-38-9
co2[e]	CO_2 (extracellular)	CO_2	0	124-38-9
coa	Coenzyme A	$C_{21}H_{32}N_7O_{16}P_3S$	–4	85-61-0
dhap	Dihydroxyacetone phosphate	$C_3H_5O_6P$	–2	57-04-5
e4p	D-Erythrose 4-phosphate	$C_4H_7O_7P$	–2	585-18-2
etoh	Ethanol	C_2H_6O	0	64-17-5
etoh[e]	Ethanol (extracellular)	C_2H_6O	0	64-17-5
f6p	D-Fructose 6-phosphate	$C_6H_{11}O_9P$	–2	643-13-0
fdp	D-Fructose 1,6-bisphosphate	$C_6H_{10}O_{12}P_2$	–4	488-69-7
for	Formate	CHO_2	–1	64-18-6
for[e]	Formate (extracellular)	CHO_2	–1	64-18-6
fru[e]	D-Fructose (extracellular)	$C_6H_{12}O_6$	0	57-48-7
fum	Fumarate	$C_4H_2O_4$	–2	110-17-8
fum[e]	Fumarate (extracellular)	$C_4H_2O_4$	–2	110-17-8
g3p	Glyceraldehyde 3-phosphate	$C_3H_5O_6P$	–2	142-10-9
g6p	D-Glucose 6-phosphate	$C_6H_{11}O_9P$	–2	56-73-5

Table 1 – *continued*

Abbreviation	Metabolite	Formula	Charge	CAS Number
glc-D[e]	D-Glucose (extracellular)	$C_6H_{12}O_6$	0	50-99-7
gln-L	L-Glutamine	$C_5H_{10}N_2O_3$	0	56-85-9
gln-L[e]	L-Glutamine (extracellular)	$C_5H_{10}N_2O_3$	0	56-85-9
glu-L	L-Glutamate	$C_5H_8NO_4$	–1	56-86-0
glu-L[e]	L-Glutamate (extracellular)	$C_5H_8NO_4$	–1	56-86-0
glx	Glyoxylate	C_2HO_3	–1	298-12-4
h2o	H_2O	H_2O	0	7732-18-5
h2o[e]	H_2O (extracellular)	H_2O	0	7732-18-5
h	H^+	H	1	12408-02-5
h[e]	H^+ (extracellular)	H	1	12408-02-5
icit	Isocitrate	$C_6H_5O_7$	–3	30810-51-6
lac-D	D-Lactate	$C_3H_5O_3$	–1	10326-41-7
lac-D[e]	D-Lactate (extracellular)	$C_3H_5O_3$	–1	10326-41-7
mal-L	L-Malate	$C_4H_4O_5$	–2	97-67-6
mal-L[e]	L-Malate (extracellular)	$C_4H_4O_5$	–2	97-67-6
nad	Nicotinamide adenine dinucleotide	$C_{21}H_{26}N_7O_{14}P_2$	–1	53-84-9
nadh	Nicotinamide adenine dinucleotide – reduced	$C_{21}H_{27}N_7O_{14}P_2$	–2	58-68-4
nadp	Nicotinamide adenine dinucleotide phosphate	$C_{21}H_{25}N_7O_{17}P_3$	–3	53-59-8
nadph	Nicotinamide adenine dinucleotide phosphate – reduced	$C_{21}H_{26}N_7O_{17}P_3$	–4	2646-71-1
nh4	Ammonium	H_4N	1	14798-03-9
nh4[e]	Ammonium (extracellular)	H_4N	1	14798-03-9
o2	O_2	O_2	0	7782-44-7
o2[e]	O_2 (extracellular)	O_2	0	7782-44-7
oaa	Oxaloacetate	$C_4H_2O_5$	–2	328-42-7
pep	Phosphoenolpyruvate	$C_3H_2O_6P$	–3	138-08-9
pi	Phosphate	HO_4P	–2	14265-44-2
pi[e]	Phosphate (extracellular)	HO_4P	–2	14265-44-2
pyr	Pyruvate	$C_3H_3O_3$	–1	127-17-3
pyr[e]	Pyruvate (extracellular)	$C_3H_3O_3$	–1	127-17-3
q8	Ubiquinone-8	$C_{49}H_{74}O_4$	0	1339-63-5
q8h2	Ubiquinol-8	$C_{49}H_{76}O_4$	0	56275-39-9
r5p	alpha-D-Ribose 5-phosphate	$C_5H_9O_8P$	–2	4300-28-1
ru5p-D	D-Ribulose 5-phosphate	$C_5H_9O_8P$	–2	4151-19-3
s7p	Sedoheptulose 7-phosphate	$C_7H_{13}O_{10}P$	–2	None
succ	Succinate	$C_4H_4O_4$	–2	110-15-6
succ[e]	Succinate (extracellular)	$C_4H_4O_4$	–2	110-15-6
succoa	Succinyl-CoA	$C_{25}H_{35}N_7O_{19}P_3S$	–5	604-98-8
xu5p-D	D-Xylulose 5-phosphate	$C_5H_9O_8P$	–2	None

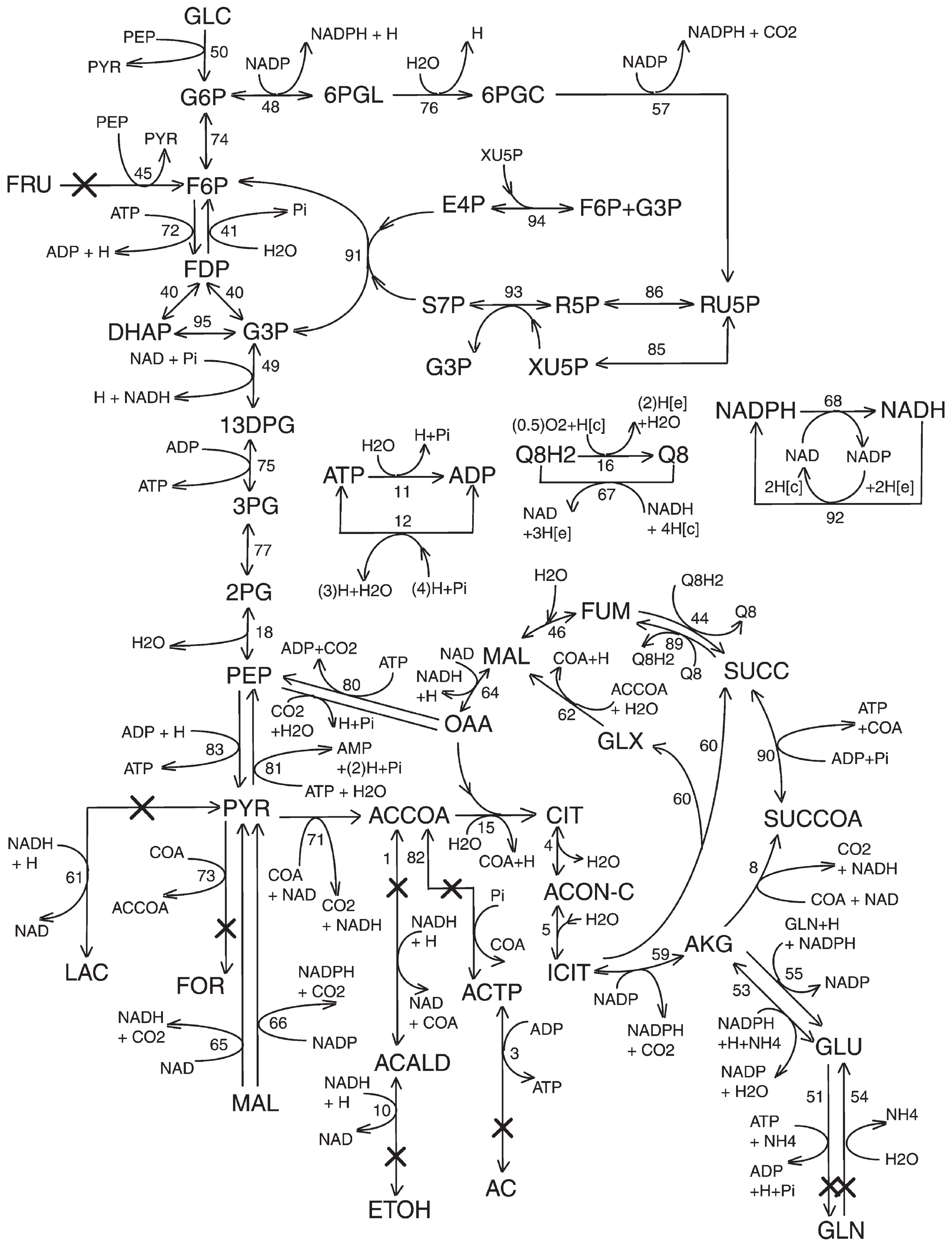


Fig. 1 – *Central metabolic pathway of Escherichia coli (Edwards and Palsson[20] model). Fluxes characterizing the membranar transport (Metabolite[e] ⇔ Metabolite[c]) and the exchange with environment ([e]:Metabolite ⇔ Metabolite[e]) have been omitted from the plot ([e]= environment; [c]= cytosol). This is the case of fluxes no.: 2(ACALD), 6(AC,H), 9(AKG,H), 14(CO2), 17(LAC,H), 19(ETOH,H), 20(AC), 21(ACALD), 22(AKG), 23(CO2), 24(ETOH), 25(FOR), 26(FRU), 27(FUM), 28(GLC), 29(GLN), 30(GLU), 31(H), 32(H2O), 33(LAC), 34(MAL), 35(NH4), 36(O2), 37(Pi), 38(PYR), 39(SUCC), 42(FOR,H), 43(FOR), 47(FUM,H), 52(GLN,ATP), 56(GLU,H), 58(H2O), 63(MAL,H), 69(NH4), 70(O2), 78(Pi,H), 84(PYR,H), 87(SUCC,2H), 88(SUCC,H). Notations: (Met) = diffusional transport of metabolite Met; (Met,H) = transport of metabolite Met via proton symport; (Met,ATP) = transport of metabolite Met via ABC system.*

Table 2 – *Considered reactions in the Edwards and Palsson[20] model (central metabolism of Escherichia coli). Notation: [e]= environment; [c]= cytosol (units correspond to mmol gDW^{-1} h^{-1}; biomass formation is expressed as g-biomass gDW^{-1} h^{-1}, or h^{-1}).*

No.	Reaction name	Equation	Net flux*	Net flux**	Lower limit	Upper limit
1	acetaldehyde dehydrogenase (acetylating)	[c] : acald + coa + nad ⇔ ⇔ accoa + h + nadh	0	$-1.3756 \cdot 10^{-11}$	–1000	1000
2	acetaldehyde reversible transport	acald[e] ⇔ acald[c]	0	$-7.8444 \cdot 10^{-12}$	–1000	1000
3	acetate kinase	[c] : ac + atp ⇔ actp + adp	0	$-6.1391 \cdot 10^{-12}$	–1000	1000
4	aconitase (half-reaction A, Citrate hydro-lyase)	[c] : cit ⇔ acon-C + h2o	6.0072	6.0072	–1000	1000
5	aconitase (half-reaction B, Isocitrate hydro-lyase)	[c] : acon-C + h2o ⇔ icit	6.0072	6.0072	–1000	1000
6	acetate reversible transport via proton symport	ac[e] + h[e] ⇔ ac[c] + h[c]	0	$-6.1391 \cdot 10^{-12}$	–1000	1000
7	adenylate kinase	[c] : amp + atp ⇔ (2) adp	0	$5.2637 \cdot 10^{-11}$	–1000	1000
8	2-Oxogluterate dehydrogenase	[c] : akg + coa + nad → → co2 + nadh + succoa	5.0644	5.0644	0	1000
9	2-oxoglutarate reversible transport via symport	akg[e] + h[e] ⇔ akg[c] + h[c]	0	$-2.8422 \cdot 10^{-12}$	–1000	1000
10	alcohol dehydrogenase (ethanol)	[c] : etoh + nad ⇔ acald + h + nadh	0	$-5.9117 \cdot 10^{-12}$	–1000	1000
11	ATP maintenance requirement	[c] : atp + h2o → adp + h + pi	8.39	8.39	8.39	1000
12	ATP synthase (four protons for one ATP)	adp[c] + (4) h[e] + pi[c] ⇔ ⇔ atp[c] + (3) h[c] + h2o[c]	45.514	45.514	–1000	1000
13	**Biomass Objective Function (with GAMS)**	**[c] : (1.496) 3pg + (3.7478) accoa + (59.8100) atp + (0.3610) e4p + (0.0709) f6p + (0.1290) g3p + (0.2050) g6p + (0.2557) gln-L + (4.9414) glu-L + (59.8100) h2o + (3.5470) nad + (13.0279) nadph + (1.7867) oaa + (0.5191) pep + (2.8328) pyr + (0.8977) r5p → (59.8100) adp + (4.1182) akg + (3.7478) coa + (59.8100) h + (3.5470) nadh + (13.0279) nadp + (59.8100) pi**	**0.8739**	**0.8739**	**0**	**1000**
14	CO2 transporter via diffusion	co2[e] ⇔ co2[c]	–22.8098	–22.8098	–1000	1000
15	citrate synthase	[c] : accoa + h2o + oaa → → cit + coa + h	6.0072	6.0072	0	1000
16	cytochrome oxidase bd (ubiquinol-8: 2 protons)	(2) h[c] + (0.5) o2[c] + q8h2[c] → → (2) h[e] + h2o[c] + q8[c]	43.5989	43.5989	0	1000
17	D-lactate transport via proton symport	h[e] + lac-D[e] ⇔ h[c] + lac-D[c]	0	$-5.0022 \cdot 10^{-12}$	–1000	1000
18	enolase	[c] : 2pg ⇔ h2o + pep	14.7161	14.7161	–1000	1000
19	ethanol reversible transport via proton symport	etoh[e] + h[e] ⇔ etoh[c] + h[c]	0	$-5.9117 \cdot 10^{-12}$	–1000	1000
20	Acetate exchange	[e] : ac ⇔ ac[e]	0	$6.1662 \cdot 10^{-12}$	0	1000
21	Acetaldehyde exchange	[e] : acald ⇔ acald[e]	0	$7.8028 \cdot 10^{-12}$	0	1000
22	2-Oxoglutarate exchange	[e] : akg ⇔ akg[e]	0	$2.9574 \cdot 10^{-12}$	0	1000
23	CO2 exchange	[e] : co2 ⇔ co2[e]	22.8098	22.8098	–1000	1000

Table 2 – *continued*

No.	Reaction name	Equation	Net flux*	Net flux**	Lower limit	Upper limit
24	Ethanol exchange	[e] : etoh ⇔ etoh[e]	0	$5.9292 \cdot 10^{-12}$	0	1000
25	Formate exchange	[e] : for ⇔ for[e]	0	$6.0134 \cdot 10^{-11}$	0	1000
26	D-Fructose exchange	[e] : fru ⇔ fru[e]	0	$3.0363 \cdot 10^{-16}$	0	1000
27	Fumarate exchange	[e] : fum ⇔ fum[e]	0	$3.1576 \cdot 10^{-16}$	0	1000
28	D-Glucose exchange	[e] : glc-D ⇔ glc-D[e]	–10	–10	–10	1000
29	L-Glutamine exchange	[e] : gln-L ⇔ gln-L[e]	0	$3.0640 \cdot 10^{-16}$	0	1000
30	L-Glutamate exchange	[e] : glu-L ⇔ glu-L[e]	0	$2.8755 \cdot 10^{-12}$	0	1000
31	H+ exchange	[e] : h ⇔ h[e]	17.5309	17.5309	–1000	1000
32	H2O exchange	[e] : h2o ⇔ h2o[e]	29.1758	29.1758	–1000	1000
33	D-Lactate exchange	[e] : lac-D ⇔ lac-D[e]	0	$5.0309 \cdot 10^{-12}$	0	1000
34	L-Malate exchange	[e] : mal-L ⇔ mal-L[e]	0	$3.1553 \cdot 10^{-16}$	0	1000
35	Ammonium exchange	[e] : nh4 ⇔ nh4[e]	–4.7653	–4.7653	–1000	1000
36	O2 exchange	[e] : o2 ⇔ o2[e]	–21.7995	–21.7995	–1000	1000
37	Phosphate exchange	[e] : pi ⇔ pi[e]	–3.2149	–3.2149	–1000	1000
38	Pyruvate exchange	[e] : pyr ⇔ pyr[e]	0	$5.0478 \cdot 10^{-12}$	0	1000
39	**Succinate exchange**	**[e] : succ ⇔ succ[e]**	**0**	$\mathbf{4.6482 \cdot 10^{-12}}$	**0**	**1000**
40	fructose-bisphosphate aldolase	[c] : fdp ⇔ dhap + g3p	7.4774	7.4774	–1000	1000
41	fructose-bisphosphatase	[c] : fdp + h2o → f6p + pi	0	$5.6732 \cdot 10^{-11}$	0	1000
42	formate transport via proton symport (uptake only)	for[e] + h[e] → for[c] + h[c]	0	$3.7922 \cdot 10^{-10}$	0	1000
43	formate transport via diffusion	for[c] → for[e]	0	$4.3935 \cdot 10^{-10}$	0	1000
44	fumarate reductase	[c] : fum + q8h2 → q8 + succ	0	497.47	0	1000
45	Fructose transport via PEP:Pyr PTS (f6p generating)	fru[e] + pep[c] → f6p[c] + pyr[c]	0	$3.1681 \cdot 10^{-16}$	0	1000
46	fumarase	[c] : fum + h2o ⇔ mal-L	5.0644	5.0644	–1000	1000
47	Fumarate transport via proton symport (2 H)	fum[e] + (2) h[e] → fum[c] + (2) h[c]	0	$3.0476 \cdot 10^{-16}$	0	1000
48	glucose 6-phosphate dehydrogenase	[c] : g6p + nadp ⇔ 6pgl + h + nadph	4.96	4.96	–1000	1000
49	glyceraldehyde-3-phosphate dehydrogenase	[c] : g3p + nad + pi ⇔ ⇔ 13dpg + h + nadh	16.0235	16.0235	–1000	1000
50	D-glucose transport via PEP:Pyr PTS	glc-D[e] + pep[c] → g6p[c] + pyr[c]	10	10	0	1000
51	glutamine synthetase	[c] : atp + glu-L + nh4 → → adp + gln-L + h + pi	0.2235	0.2235	0	1000
52	L-glutamine transport via ABC system	atp[c] + gln-L[e] + h2o[c] → → adp[c] + gln-L[c] + h[c] + pi[c]	0	$3.1609 \cdot 10^{-16}$	0	1000
53	glutamate dehydrogenase (NADP)	[c] : glu-L + h2o + nadp ⇔ ⇔ akg + h + nadph + nh4	–4.5419	–4.5419	–1000	1000
54	glutaminase	[c] : gln-L + h2o → glu-L + nh4	0	$4.9750 \cdot 10^{-12}$	0	1000
55	glutamate synthase (NADPH)	[c] : akg + gln-L + h + nadph → → (2) glu-L + nadp	0	$5.5288 \cdot 10^{-12}$	0	1000

Table 2 – *continued*

No.	Reaction name	Equation	Net flux*	Net flux**	Lower limit	Upper limit
56	L-glutamate transport via proton symport, reversible (periplasm)	glu-L[e] + h[e] ⇔ glu-L[c] + h[c]	0	$-2.9559 \cdot 10^{-12}$	–1000	1000
57	phosphogluconate dehydrogenase	[c] : 6pgc + nadp → → co2 + nadph + ru5p-D	4.96	4.96	0	1000
58	H2O transport via diffusion	h2o[e] ⇔ h2o[c]	–29.1758	–29.1758	–1000	1000
59	isocitrate dehydrogenase (NADP)	[c] : icit + nadp ⇔ ⇔ akg + co2 + nadph	6.0072	6.0072	–1000	1000
60	Isocitrate lyase	[c] : icit → glx + succ	0	$1.8071 \cdot 10^{-10}$	0	1000
61	D-lactate dehydrogenase	[c] : lac-D + nad ⇔ h + nadh + pyr	0	$-5.1159 \cdot 10^{-12}$	–1000	1000
62	malate synthase	[c] : accoa + glx + h2o → → coa + h + mal-L	0	$1.8071 \cdot 10^{-10}$	0	1000
63	Malate transport via proton symport (2 H)	(2) h[e] + mal-L[e] → → (2) h[c] + mal-L[c]	0	$3.0499 \cdot 10^{-16}$	0	1000
64	malate dehydrogenase	[c] : mal-L + nad ⇔ h + nadh + oaa	5.0644	5.0644	–1000	1000
65	malic enzyme (NAD)	[c] : mal-L + nad → co2 + nadh + pyr	0	$4.0902 \cdot 10^{-11}$	0	1000
66	malic enzyme (NADP)	[c] : mal-L + nadp → → co2 + nadph + pyr	0	$5.8181 \cdot 10^{-11}$	0	1000
67	NADH dehydrogenase (ubiquinone-8 & 3 protons)	(4) h[c] + nadh[c] + q8[c] → → (3) h[e] + nad[c] + q8h2[c]	38.5346	38.5346	0	1000
68	NAD transhydrogenase	[c] : nad + nadph → nadh + nadp	0	$6.2334 \cdot 10^{-11}$	0	1000
69	ammonia reversible transport	nh4[e] ⇔ nh4[c]	4.7653	4.7653	–1000	1000
70	O2 transport via diffusion	o2[e] ⇔ o2[c]	21.7995	21.7995	–1000	1000
71	pyruvate dehydrogenase	[c] : coa + nad + pyr → → accoa + co2 + nadh	9.2825	9.2825	0	1000
72	phosphofructokinase	[c] : atp + f6p → adp + fdp + h	7.4774	7.4774	0	1000
73	pyruvate formate lyase	[c] : coa + pyr → accoa + for	0	$6.0134 \cdot 10^{-11}$	0	1000
74	glucose-6-phosphate isomerase	[c] : g6p ⇔ f6p	4.8609	4.8609	–1000	1000
75	phosphoglycerate kinase	[c] : 3pg + atp ⇔ 13dpg + adp	–16.0235	–16.0235	–1000	1000
76	6-phosphogluconolactonase	[c] : 6pgl + h2o → 6pgc + h	4.96	4.96	0	1000
77	phosphoglycerate mutase	[c] : 2pg ⇔ 3pg	–14.7161	–14.7161	–1000	1000
78	phosphate reversible transport via proton symport	h[e] + pi[e] ⇔ h[c] + pi[c]	3.2149	3.2149	–1000	1000
79	phosphoenolpyruvate carboxylase	[c] : co2 + h2o + pep → h + oaa + pi	2.5043	2.5043	0	1000
80	phosphoenolpyruvate carboxykinase	[c] : atp + oaa → adp + co2 + pep	0	$2.8585 \cdot 10^{-11}$	0	1000
81	phosphoenolpyruvate synthase	[c] : atp + h2o + pyr → → amp + (2) h + pep + pi	0	$5.2692 \cdot 10^{-11}$	0	1000
82	phosphotransacetylase	[c] : accoa + pi ⇔ actp + coa	0	$6.2528 \cdot 10^{-12}$	–1000	1000
83	pyruvate kinase	[c] : adp + h + pep → atp + pyr	1.7582	1.7582	0	1000
84	pyruvate reversible transport via proton symport	h[e] + pyr[e] ⇔ h[c] + pyr[c]	0	$-5.0022 \cdot 10^{-12}$	–1000	1000
85	ribulose 5-phosphate 3-epimerase	[c] : ru5p-D ⇔ xu5p-D	2.6785	2.6785	–1000	1000
86	ribose-5-phosphate isomerase	[c] : r5p ⇔ ru5p-D	–2.2815	–2.2815	–1000	1000

No.	Reaction name	Equation	Net flux*	Net flux**	Lower limit	Upper limit
87	succinate transport via proton symport (2 H)	(2) h[e] + succ[e] → (2) h[c] + succ[c]	0	$1.1830 \cdot 10^{-10}$	0	1000
88	succinate transport out via proton antiport	h[e] + succ[c] → h[c] + succ[e]	0	$1.2295 \cdot 10^{-10}$	0	1000
89	succinate dehydrogenase (irreversible)	[c] : q8 + succ → fum + q8h2	5.0644	5.0644	0	1000
90	succinyl-CoA synthetase (ADP-forming)	[c] : atp + coa + succ ⇔ ⇔ adp + pi + succoa	–5.0644	-5.0644	–1000	1000
91	transaldolase	[c] : g3p + s7p ⇔ e4p + f6p	1.497	1.497	–1000	1000
92	NAD(P) transhydrogenase	(2) h[e] + nadh[c] + nadp[c] → → (2) h[c] + nad[c] + nadph[c]	0	$5.7674 \cdot 10^{-10}$	0	1000
93	transketolase	[c] : r5p + xu5p-D ⇔ g3p + s7p	1.497	1.497	–1000	1000
94	transketolase	[c] : e4p + xu5p-D ⇔ f6p + g3p	1.1815	1.1815	–1000	1000
95	triose-phosphate isomerase	[c] : dhap ⇔ g3p	7.4774	7.4774	–1000	1000

*Fluxes correspond to equilibrated stationary growth with a glucose uptake rate of –10 mmol gDW^{-1} h^{-1}, and an oxygen uptake rate of –1000 mmol gDW^{-1} h^{-1}.

**Flux values are obtained by means of LP procedure from solving the single level optimization problem of biomass production maximization in *Escherichia coli* cells with basic CONSTR (not the global optimum).

fluxes correspond to an equilibrated growth of cell, with a glucose uptake rate of –10 mmol gDW^{-1} h^{-1}, and an oxygen uptake rate of –1000 mmol gDW^{-1} h^{-1}. The model was obtained by lumping the extended model of Edwards and Palsson[20] that includes 720 reactions and 436 metabolites. In the extended variant, unconstrained uptake routes for inorganic phosphate, carbon dioxide, oxygen, sulphate, potassium, sodium, and ammonia are provided, and the capacity constraints were used to define the reaction reversibility. Lower limits for the internal fluxes were set to zero for all irreversible fluxes, and all reversible fluxes were upper bounded at a large value. Transport fluxes for metabolites not available in the media were always restricted to zero, while forward and backward reactions result in positive and negative fluxes respectively.

Due to the applied lumping procedure, the reduced model contains many overall reactions that sum ‘elementary’ metabolic steps. For instance, the rate of biomass production, $v_{biomass} = v_{13}$, results as a sum of many contributory steps leading to the overall stoichiometry of Table 2.

Burgard *et al.*[33] solved the dual-optimization problem of succinate production maximization ($v_{succinate}$), subjected to biomass production maximization ($v_{biomass}$) in the presence of linear constraints. They obtained a large number of gene knockout solutions, such as: removed genes no. {61,73}, or no. {1,10,61,73} in the reduced model, with $v_{succinate}$ = 11 mmol gDW^{-1} h^{-1}, $v_{biomass}$ = 0.3 h^{-1}; removed genes {3,50,82,83} in the reduced model with $v_{succinate}$ = 15 mmol gDW^{-1} h^{-1}, $v_{biomass}$ = 0.16 h^{-1}, etc. (see representation of some solutions in Fig. 2). On the other hand, the size of the MILP combinatorial problem increases largely with the number of removed genes. Simulations and experiments also revealed existence of a non-linear relationship between $v_{succinate}$ and $v_{biomass}$, that is large $v_{succinate}$ (of maximum 16.4 mmol gDW^{-1} h^{-1} in the studied cell growth conditions) corresponds to $v_{biomass}$ close to zero, and vice-versa. Consequently, it was concluded that several sub-optimal solutions can exist when designing a mutant cell, according to the considered sets of removed genes.

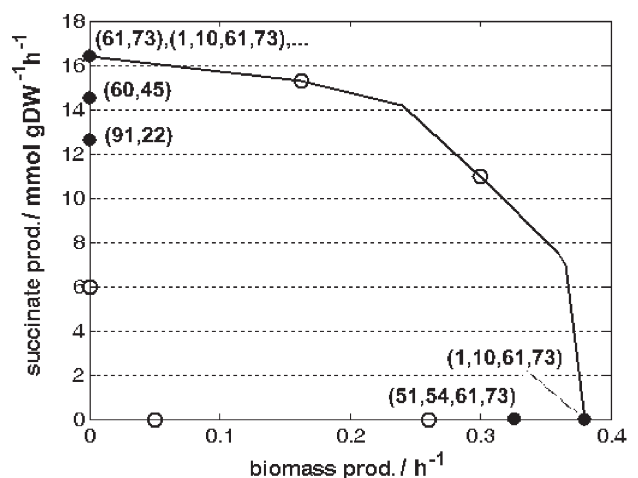


Fig. 2 – *Succinate and biomass production in wild-type E. coli cells (——), and location of some restricted gene-knockout mutants predicted by Burgard et al.*[33] *(○) (anaerobic conditions), and by the present study for $v_{biomass_target}$ (●) (aerobic conditions). Numbers in parenthesis denote removed genes from wild-type E. coli cell.*

Formulating the succinate production optimization problem

When designing an optimal phenotype, a bi-/multi-objective cell flux optimization problem should be formulated, by accounting for several goals:[37]

– maximize ATP production to determine conditions of optimal metabolic energy efficiency.

– minimize nutrient uptake, by determining the conditions under which the cell will perform its metabolic functions while consuming the minimum amount of available nutrients.

– minimize redox production by finding conditions where the cells operate to generate the minimum amount of redox potential, and minimum adenylate energy charge (ATP, ADP, AMP) necessary to draw the inner cell syntheses.

– minimize the Euclidean norm, i.e. the sum of the fluxes allowing to channel the metabolites as efficiently as possible through the metabolic pathways.

– maximize target metabolite production, by optimizing the cell capabilities to produce a certain compound of practical (industrial) interest.

– maximize the biomass production ensuring the cellular network to evolve and proliferate.

It should be mentioned that the last two goals, i.e. *maximize the biomass and metabolite production,* are competing objectives in a cell due to the requirement of using the resources for a maximum responsiveness to the environmental changes rather than for the overproduction of a specific compound. In a design strain, a certain tradeoff should be realized between cell growth and forced metabolite production to preserve the cell growth and proliferation objectives.

The single level optimization problem can be formulated in terms of biomass production maximization ("cellular objective"), in the form:

$$[v_1,\dots,v_M] = arg\ Max\ \Phi = v_{biomass} (= v_{13})$$
$$\text{s.t.}\ \sum_{j=1}^{M} S_{ij} v_j = 0,\ \ i = 1,\dots,N \qquad (1)$$
$$v_{j,\min} \le v_j \le v_{j,\max},\ \ j = 1,\dots,M$$

where M = 95 fluxes and N = 72 metabolites are considered in the metabolic system (flux notations correspond to those of Table 2). The way to formulate the stoichiometric matrix and mass balance for the considered metabolic reactions in the $\boldsymbol{S}\,\boldsymbol{v} = 0$ format is exemplified in Annex 1.

Other objectives can be considered with a similar formulation, for instance a linear combination of target metabolic fluxes.[20,27] Successive solutions derived under various environmental conditions allow for instance determination of the correct sequence of byproduct secretion under increasingly anaerobic conditions.[38] When the same objective is associated with gene knockout alternatives, estimated fluxes can give information on the essential genes in the cell (lethality of gene knockout).[20,27] Additional linear constraints to the optimization problem usually account for environmental requirements (e.g. nutrient limitation).[28]

When not only the "cellular objective" is optimized, but also the production of a certain metabolite (the so-called "bioengineering / chemical objective"), a bi-level optimization problem results (e.g. by using succinate as target metabolite):

Primal:

$$[v_1,\dots,v_M] = arg\ Max\ \Phi = v_{succinate}$$
$$\text{s.t.}\ Max\ v_{biomass}$$
$$\text{s.t.}\ \sum_{j=1}^{M} S_{ij} v_j = 0,\ \ i = 1,\dots,N \qquad (2)$$
$$v_{j,\min} \le v_j \le v_{j,\max},\ \ j = 1,\dots,M;$$
$$v_{biomass} \ge v_{biomass_target}$$

where $v_{biomass_target}$ is the minimum level of biomass production which has to be realized by the 'optimized' cell. One alternative to solve the 'primal' problem is to transform it into an equivalent LP problem (called 'dual' problem):

Dual:

$$[v_1,\dots,v_M] = arg\ Max\ \Phi = v_{succinate}$$
$$\text{s.t.}\ v_{biomass} = \mu_{biomass} v_{biomass_target}$$
$$\sum_{j=1}^{M} S_{ij} v_j = 0,\ \ i = 1,\dots,N;$$
$$\sum_{i=1}^{N} \lambda_i^{stoich} S_{i,biomass} + \mu_{biomass} = 1$$
$$\sum_{i=1}^{N} \lambda_i^{stoich} S_{i,j} + \mu_j = 0,\ j \ne biomass;\ \lambda_i^{stoich} \in R$$
$$v_{j,\min} \le v_j \le v_{j,\max},\ \ j = 1,\dots,M; \qquad (3)$$
$$v_{biomass} \ge v_{biomass_target}$$
$$\mu_{j,\min} \le \mu_j \le \mu_{j,\max},$$
(for reversible reactions, but not secr_only)
$$\mu_j \ge \mu_{j,\min},$$
(for reversible reactions, and secr_only)
$$\mu_j \le \mu_{j,\max},$$
(for irreversible reactions, and not secr_only)
$$\mu_j \in R,$$
(for irreversible reactions, and secr_only)

where 'secr_only' denotes transport fluxes for metabolites that can only be secreted from the network; μ_j = dual variable associated with any restriction of the corresponding flux v_j in the primal problem; λ_i^{stoich} = dual variable associated with the stoichiometric constraints). Such a transformation of (2) to (3) is possible due to the observation that if the optimal solutions of primal and dual problems are bounded, their objective functions must be equal at optimality.[39–42]

However, the previous LP problem raises several complications when additional constraints are formulated and/or the extended cellular model is approached, making the problem preparation very laborious and computation-intensive to be solved. For instance, in the Burgard *et al.*[33] formulation, two additional constraints have to be added to (2), while finding the optimal gene knockout strategy requires searching for optimal additional Boolean variables $y_j \in \{0,1\}$ that multiply the fluxes (i.e. $y_j v_j$ instead of v_j), complicating the formulation (3). Besides, other algorithm disadvantages have to be mentioned, as follows: I) the LP transformation of primal problem is valid only for the bi-level optimization and not for several objective functions; ii) exclusion of nonlinear constraints; iii) effective solution of the dual LP problem requires the correct setting of upper/lower bounds of the dual variables, i.e. $\mu_{j,\min}$, $\mu_{j,\max}$; iv) when more than 2–3 genes are simultaneously removed from the network, the resulting LP combinatorial problem becomes extremely computation-intensive, and practically ineffective for cells including a large number of genes.

Consequently, to solve the multi-objective LP or NLP problems, various alternatives can also be approached: min-max formulation, single-composite function, or Pareto front method.[43–47] There is no general approach for such a choice because the decision is case-dependent.[34] For instance, if the individual objective functions f_j are scaled in the same range, a composite objective function can be defined as a linear combination, of the form $\Phi_{LP} = \Sigma_j w_j f_j$, where the adopted weights usually satisfy the conditions $\Sigma_j w_j = 1$.[20,48] However, nonlinear combinations of individual f_j are also possible, depending on their physical significance.[34]

For instance, one possibility retained for comparison is to transform the two-objective problem (2) into a single level LP optimization by using the composite function:

$$[v_1,\dots,v_M] = arg\ Max\ \Phi_{LP} =$$
$$= w_{succinate} v_{succinate} + w_{biomass} v_{biomass} (= w_{39} v_{39} + w_{13} v_{13})$$
$$\text{s.t.} \sum_{j=1}^{M} S_{ij} v_j = 0, \quad i = 1,\dots,N \qquad (4)$$
$$v_{j,\min} \le v_j \le v_{j,\max}, \quad j = 1,\dots,M$$

In the present case study, we want to investigate another route to achieve the multi-objective optimization of cell fluxes, by formulating a nonlinear programming (NLP) problem using a power-law type composite objective function, of the form:

$$[v_1,\dots,v_M] = arg\ Max\ \Phi_{NLP};$$
$$\Phi_{NLP} = \prod_{j=1}^{M} v_j^{\beta_j}, \quad \beta_j \in R$$

s.t. CONSTR:

$$\sum_{j=1}^{M} S_{ij} v_j = 0, \quad i = 1,\dots,N, \ (\text{or } \boldsymbol{Sv} = 0) \qquad (5)$$
$$v_{j,\min} \le v_j \le v_{j,\max}, \quad j = 1,\dots,M$$

The individual fluxes can be included ($\beta_j \ne 0$) or not ($\beta_j = 0$) in the optimization, in an (un)scaled form, with an exponent sign and magnitude depending on the maximization / minimization goal and its relative importance in the metabolism. As another observation, when some fluxes (index 'exp') are measured, a supplementary equality constraint should be added to (5), of the form $\boldsymbol{S}_{unk}\boldsymbol{v}_{unk} = -\boldsymbol{S}_{exp}\boldsymbol{v}_{exp}$, where 'unk' index denotes the unknown vector of fluxes.

In fact, if a logarithmic transformation is applied to the goal function Φ_{NLP} in (5), an equivalent weighted multi-objective LP optimization problem of type $Max(\boldsymbol{wv})$ is obtained. However, the NLP formulation is not fully equivalent with the LP formulation, as long as linear transformation distorts the flux contribution to the main goal and can not represent the nonlinear inter-dependencies among fluxes. For instance, various nonlinear objectives of power-law type can be formulated according to the desired modification of phenotype, e.g.: I) maximum of v_s with minimum of v_b, $\text{Max}\Phi_{NLP} = v_s / v_b$; ii) maximum of a series of scaled fluxes v_k, v_l, v_m, …, $\text{Max}\Phi_{NLP} = (v_k/\|\boldsymbol{v}\|_2)(v_l/\|\boldsymbol{v}\|_2)(v_m/\|\boldsymbol{v}\|_2)$; iii) maximum of a flux v_k related to the corresponding overall production of entropy into the cell, $\text{Min}\Phi_{NLP} = (-v_k)/\Sigma_j (v_j A_j / T)$, etc. (where: A_j = reaction affinity, T = temperature; see Heinrich and Schuster[9] for other nonlinear objectives and constraints). Besides, NLP formulation can include also nonlinear constraints derived from imposed properties of the metabolic pathway, accounting for gene inferences[7] and regulatory network properties,[4] or other thermodynamic properties.[56]

In the present study, a particularization of (5) is used for maximizing the succinate and biomass positive fluxes, of the form:

$$[v_1, \ldots, v_M] = arg\ Max\ \Phi_{NLP} = arg\ Min\,(-\Phi_{NLP});$$

$$\Phi_{NLP} = v_{succinate}^{\beta_s} v_{biomass}, \quad \beta_s > 0. \tag{6}$$

The degree of freedom (*DF*) of the NLP problem depends on the number $NEQ<N$ of equality constraints of $\boldsymbol{Sv} = 0$ balance set. The constraints are accounted during the solution search in a simple way, for instance by evaluating the constraint violation degree by means of two indices:

$$CV_{eq} = \left(\sum_{i=1}^{N} \left| \sum_{j=1}^{M} S_{ij} v_j \right| \right); \quad \overline{CV}_{eq} = CV_{eq}/N;$$

$$CV_{eq}^2 = (CV_{eq})^2 \quad \text{(equality constraints)},$$

$$CV_{ineq}^2 = \sum_{j=1}^{M} C_{j,up}^2 + \sum_{j=1}^{M} C_{j,low}^2,$$

$$\overline{CV}_{ineq}^2 = (CV_{ineq}^2)/M \quad \text{(inequality constraints)},$$

where:

$$C_{j,up}^2 = \begin{cases} (v_j - v_{j,\max})^2, & \text{if } v_j > v_{j,\max} \\ 0, & \text{else} \end{cases},$$

$$C_{j,low}^2 = \begin{cases} (v_{j,\min} - v_j)^2, & \text{if } v_j < v_{j,\min} \\ 0, & \text{else} \end{cases} \tag{7}$$

When solving the NLP problem, constraint violations "penalise" the objective function by means of an extended Lagrange function, of the form:

$$[v_1, \ldots, v_M] = arg\ Min\ L = $$
$$= -\Phi_{NLP} + \lambda_{eq} CV_{eq}^2 + \lambda_{ineq} CV_{ineq}^2, \tag{8}$$

where the Lagrange multipliers λ_{eq}, λ_{ineq} are chosen to be zero if the constraints are not violated, and receive positive values (constant, or increasing / decreasing numbers according to the search failure / success).[43,49] The constraints are scaled according to the objective function range. In the present study, as Φ_{NLP} is a product of two fluxes, squared indices of constraint violation are included, with a uniform weight of $\lambda_{eq} = \lambda_{ineq} = 1$.

To make our results comparable to similar case studies, the considered constraints in the optimization problem of biomass and target metabolite production maximization are those indicated by Burgard *et al.*[33] for the extended model, but adapted to the reduced model structure, that is, glucose uptake balance, maintenance requirements, and minimum level of biomass production:

$$\begin{bmatrix} v_{pts} + v_{glk} = v_{glc_uptake} \\ v_{biomass} \geq v_{biomass_target} \\ v_{atp} \geq v_{atp_maintenance} \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} v_{pts} = v_{glc_uptake} \\ v_{biomass} \geq v_{biomass_target} \\ v_{atp} \geq v_{atp_maintenance} \end{bmatrix} \Rightarrow \begin{bmatrix} |v_{28}| = |v_{50}| \\ v_{13} \geq v_{13}^{\min} \\ v_{12} \geq v_{11} \end{bmatrix} \tag{9}$$

where: M = number of considered fluxes / reactions in the metabolic network (720 in the extended model, and 95 in the reduced model); N = number of metabolites (436 in the extended model, and 72 in the reduced model); v_j = generic metabolic flux; S_{ij} = stoichiometric coefficient of the metabolite i in the reaction j. In the previous formulations, v_{glc_uptake} is the basic glucose uptake scenario, $v_{atp_maintenance}$ is the non-growth associated ATP maintenance requirement, while $v_{biomass_target}$ is the minimum level of biomass production imposed by the designer (theoretically being higher than zero, or even zero).[33] Adaptation of metabolic constraints reported some modified relationships due to lumped species and reactions. Thus, the equality constraint "$v_{pts} + v_{glk} = v_{glc_uptake}$" refers to phosphotransferase (glc-D[e] + pep[c] → g6p[c] + pyr[c]) and glucokinase ([c] : atp + glc-D → adp + g6p + h) fluxes of the extended model. In the reduced model, v_{pts} refers to GLCpts reaction in Table 2 (i.e. v_{50} flux), but the *E. coli* core reduced model does not include the glk/glucokinase flux. Consequently, this constraint reduces to "$v_{pts} = v_{glc_uptake}$" constraint as long as v_{glk} is not accounted for. The constraint "$v_{biomass} \geq v_{biomass_target}$" was implicitly considered during the MINLP optimization problem of maximum biomass production, so $v_{biomass_target}$ is usually set to zero.

Solving the FBA problem to determine various gene knockout strategies

To determine possible optimal phenotypes of the analysed microorganism, coupled FBA with the multi-objective optimization can be applied, by estimating the stationary fluxes associated with a proposed sub-set of genes encoding the enzymes participating in the metabolic pathway. To point-out the importance of the number and structure of the problem constraints, a step-by-step strategy to identify the optimal *E. coli* mutant for succinate production was developed.

By setting the number *KG* of genes which have to be removed from the cell, the basic MINLP rule consists of simultaneously finding the removed genes and the optimal fluxes, in a problem formulation similar to (6–8), that is:

$$[v_1,\dots,v_M,y_1,\dots,y_M] = arg\ Min\ \Phi_{NLP} = -\Phi_{NLP} + \lambda_{eq} CV_{eq}^2 + \lambda_{ineq} CV_{ineq}^2$$

s.t. CONSTR:

$$\sum_{j=1}^{M} S_{ij} v_j = 0, \quad i = 1,\dots,N; \qquad (10)$$

$$v_{j,\min} y_j \le v_j \le v_{j,\max} y_j, \; j = 1,\dots,M; \; y_j = \{0,1\};$$

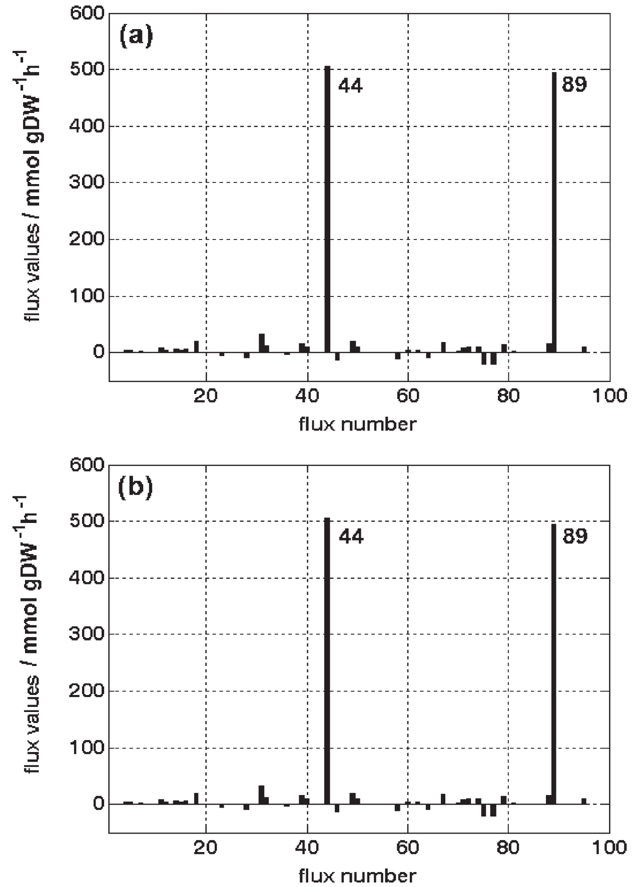
$$\sum_{j=1}^{M} (1 - y_j) = KG; \; y_{13} = y_{biomass} = 1; \; y_{39} = y_{succinate} = 1$$

When one gene (and its encoded enzyme) is removed from the cell, the associated flux is also omitted from the reaction pathway, by setting $y_j = 0$. The obtained mutant presents optimized fluxes, but it is not necessarily able to preserve the main cell functions to ensure cell survival. Consequently, every mathematical solution has to be metabolically viable for being interpreted as a physical meaning before validation.

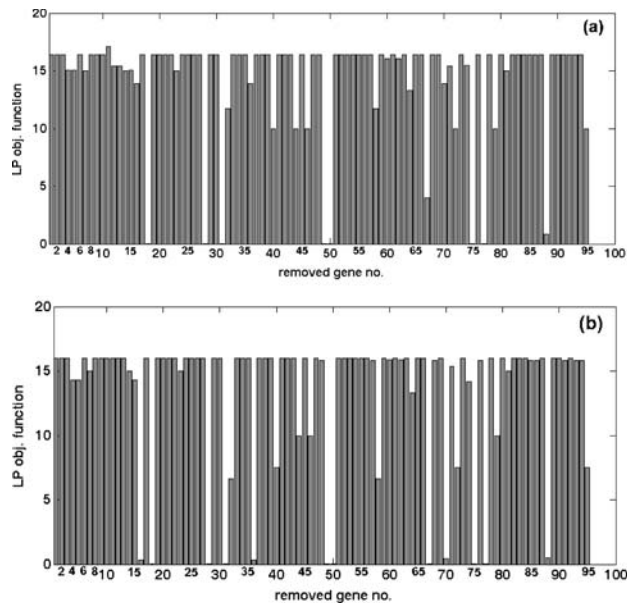
Solving gene knockout LP and MINLP problem by accounting for only stoichiometric constraints. One starts by finding optimal fluxes in the *E. coli* cell in the presence of only basic stoichiometric constraints of the metabolic pathway (Table 2). The problem is solved by using both the LP formulation (4) (with $w_{succinate} = w_{biomass} = 1$), and the NLP formulation (5) with the CONSTR set of constraints and $\beta_s = 1$. The results, presented in Fig. 3, indicate practically the same (unique) solution, irrespective of the used method. As expected, the resulting very large values for the reversible succinate-to-fumarate transformation (fluxes #44 and #89) is the main reaction responsible for succinate production maximization. But this theoretical solution does not necessarily ensure cell viability, and other dynamic / thermodynamic constraints should be further considered.

When genes are *in-silico* knockout from the cell, the problem solution ceases to be unique, and several alternatives might exist. Theoretically, a number of C_M^{KG} mutant cells results from removing *KG* genes. For *KG = 1,* the number of knockout trials equals the number of genes (*M*). This number increases sharply with *KG*, following an approximate power law given by the Stirling formula: $N! = (N/e)^N \sqrt{2\pi N}$.[50]

By successively removing one gene after one (from gene #1 to gene #95, *KG = 1*), every time evaluating the optimal fluxes vs. criterion (4), optimal LP solutions are thus obtained, being presented in Fig. 4-a (the Matlab™ LP solver has been used).[44] By analyzing the results, it should be mentioned that a large number of alternatives exhibit the same performance index $\Phi_{NLP} = v_{succinate} + v_{biomass} = 16.384$ mmol gDW^{-1} h^{-1}, even if two additional constraints ($|v_{28}| = |v_{50}|$ and $v_{12} \ge v_{11}$) have been added (Fig.



F i g . 3 – *Flux distribution (absolute values in mmol gDW^{-1} h^{-1}) in wild-type E. coli for succinate and biomass production maximization (with only basic CONSTR constraints): (a) LP solution (v_{13} = 2.90 · 10^{-12}, v_{39} = 16.3840, L = 1.36 · 10^{-3}); (b) NLP solution (v_{13} = 2.97 · 10^{-12}, v_{39} = 16.3842, L = 1.07 · 10^{-5}, DF=94).*

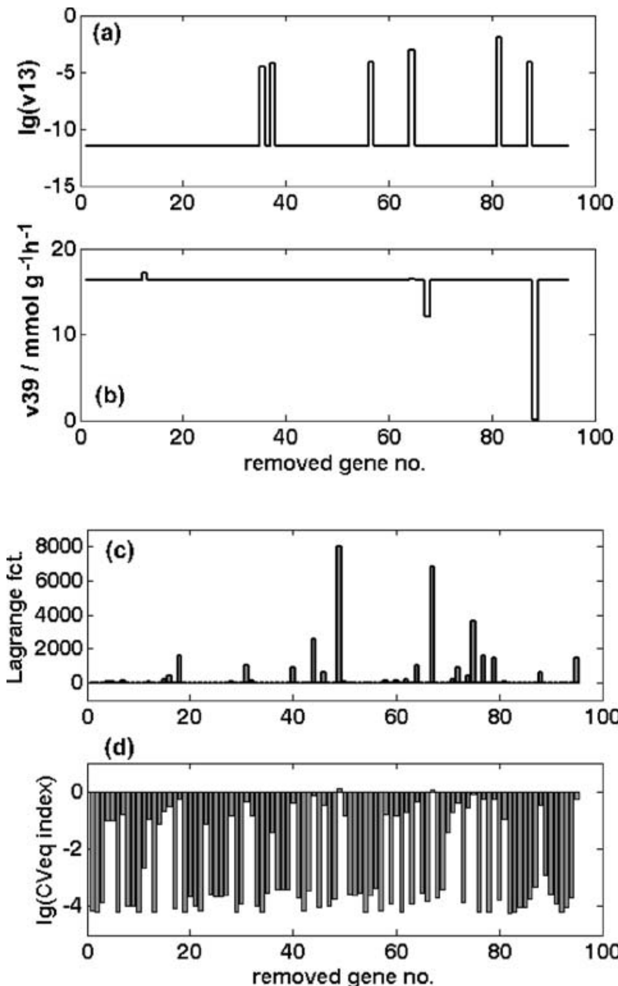


F i g 4 – *LP objective function ($\Phi_{LP} = v_{39} + v_{13}$) for succinate and biomass production maximization of various mutants of E. coli when successively removing one single gene (from gene #1 to gene #95). (a) Imposed basic constraints CONSTR; (b) Imposed basic constraints CONSTR, and the supplementary constraints $|v_{28}| = |v_{50}|$ and $v_{12} \ge v_{11}$.*

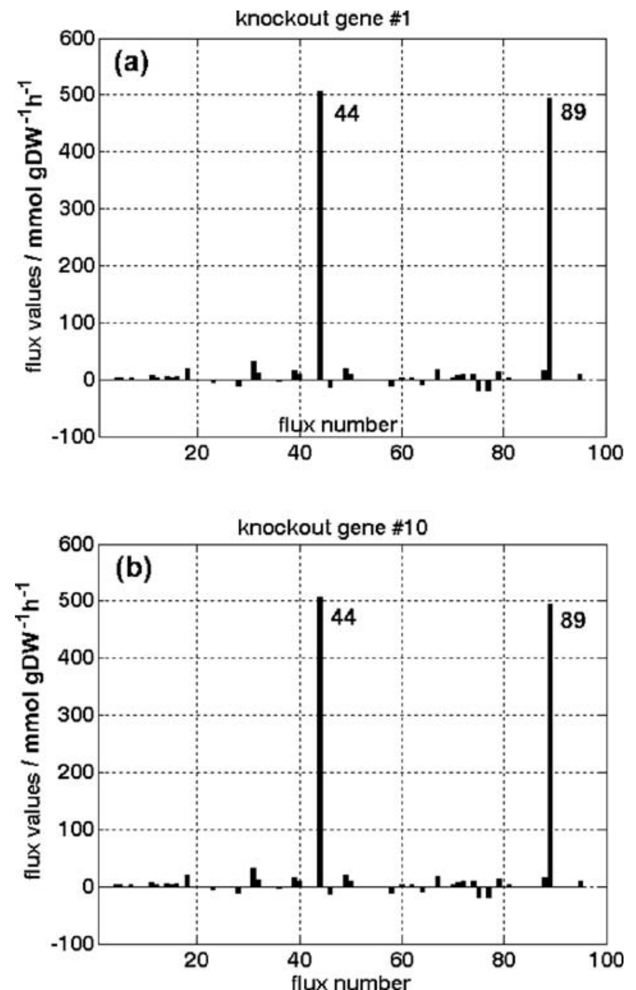
4-b). This result is similar to those of Edwards and Palsson,[20] suggesting that "a large number of the central metabolic genes can be removed without eliminating the capability of the metabolic network to support growth under the conditions considered", due to the interconnectivity of the metabolic reactions.

The same single-gene knockout rule was repeated by using the MINLP criterion (10) with the basic stoichiometric constraints included in the Lagrange function *L*. The results, presented in Fig. 5, indicate the same conclusion as those obtained from using the LP criterion, that is, a large number of genes can be removed by keeping the succinate production at the highest level of $v_{succinate}$ = 16.384 mmol g^{-1} h^{-1} (Fig. 5-b). The inequality constraints are all time fulfilled, while the violation index $\overline{CV}_{eq}$ of equality constraints is roughly negligible (more precise solutions are possible but with the expense of a significant supplementary computational effort). Slight violation of equality constraints ($\overline{CV}_{eq}$ index) is sharply penalized by large Lagrange functions *L* in Fig. 5c. It should be mentioned that many removed genes lead practically to the same optimal fluxes into the cell, as for instance the removed genes 1 and 10 in Fig. 6 (reactions #1 and #10). These removed reactions block the EtOH production, being also removed in one of the mutant cells obtained by Burgard *et al.*[33]

To investigate multiple gene knockout solutions, two or four genes have been concomitantly removed from the pathway. It should be mentioned that the number of possible solutions of the same quality increases very much (i.e. of approximately the same objective function *L*). Some of the *in silico* mutant *E. coli* cells obtained by means of MINLP criterion (10) are presented in Table 3. As



F i g . 5 – *Example of optimal solutions with the MINLP for succinate (v_{39})(b) and biomass (v_{13})(a) production maximization of various mutants of E. coli when successively removing one single gene (from gene #1 to gene #95; DF=95, basic CONSTR). The MINLP Lagrange objective function (L) (c) is displayed together with the logarithm to the base ten of $\overline{CV}_{eq}$ equality constraint index (d). All inequality constraints are met ($\overline{CV}^2_{ineq}$ = 0).*



F i g . 6 – *Example of local solution – flux distribution (absolute values in mmol gDW^{-1} h^{-1}) of two E. coli mutants for succinate and biomass production maximization (MINLP solution with basic CONSTR, DF = 95): (a) gene #1 knockout (v_{13} = –2.90 · 10^{-12}, v_{39} = –16.384, L = 2.41 · 10^{-5}); (b) gene #10 knockout (v_{13} = –2.90 · 10^{-12}, v_{39} = –16.384, L = 2.21 · 10^{-5}) (units are in mmol gDW^{-1} h^{-1}, and in g-biomass gDW^{-1} h^{-1} for biomass formation).*

Table 3 – *Various local solutions of the MINLP problem for succinate and biomass production maximization in mutant E. coli by removing two or four genes from the central metabolism (Edwards and Palsson[20] model). The $\overline{CV}^2_{ineq}$ inequality constraint violation index (7) is zero for all below solutions (all inequality constraints fulfilled). The $\overline{CV}_{eq}$ index is defined by eq. (7).*

Knockout genes	v_{13} (biomass rate, h^{-1})	v_{39} (succinate rate, mmol gDW^{-1} h^{-1})	v_{39}/v_{50} (Note b)	Lagrange function (L)	$\overline{CV}_{eq}$ index	$v_{39} > v_{39}^{min}$
two genes removed under basic constraints CONSTR						
61, 73	$2.8477 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$4.88 \cdot 10^{-7}$	$9.71 \cdot 10^{-6}$	
51, 82	$2.8566 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$1.71 \cdot 10^{-7}$	$5.75 \cdot 10^{-6}$	
45, 51	$2.6734 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$2.39 \cdot 10^{-7}$	$6.79 \cdot 10^{-6}$	
45, 55	$2.7374 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$5.80 \cdot 10^{-7}$	$1.05 \cdot 10^{-5}$	
10, 94	$2.7726 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$3.92 \cdot 10^{-7}$	$8.70 \cdot 10^{-6}$	
9, 61	$3.0328 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$6.01 \cdot 10^{-7}$	$1.07 \cdot 10^{-5}$	
1, 84	$2.7496 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$2.94 \cdot 10^{-7}$	$7.53 \cdot 10^{-6}$	
51, 82	$2.9069 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$1.19 \cdot 10^{-5}$	$4.79 \cdot 10^{-5}$	
51, 54	$2.9085 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$1.55 \cdot 10^{-7}$	$5.47 \cdot 10^{-6}$	
1, 78	$2.9555 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$3.16 \cdot 10^{-6}$	$2.47 \cdot 10^{-5}$	
1, 3	$3.2408 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$1.68 \cdot 10^{-7}$	$5.70 \cdot 10^{-6}$	
(82, 83), Etc.	$2.7578 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$2.60 \cdot 10^{-7}$	$7.08 \cdot 10^{-6}$	
four genes removed under basic constraints CONSTR						
1, 10, 61, 73	$2.9187 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$5.69 \cdot 10^{-7}$	$1.04 \cdot 10^{-5}$	
61, 73, 51, 54	$2.8809 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$7.35 \cdot 10^{-7}$	$1.19 \cdot 10^{-5}$	
25, 45, 55, 78	$2.9001 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$2.10 \cdot 10^{-5}$	$6.37 \cdot 10^{-5}$	
1, 30, 51, 94	$2.8650 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$1.46 \cdot 10^{-5}$	$5.31 \cdot 10^{-5}$	
43, 1, 78, 41	$2.9872 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$2.12 \cdot 10^{-7}$	$6.40 \cdot 10^{-6}$	
85, 73, 41, 1	$3.1386 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$3.51 \cdot 10^{-7}$	$8.23 \cdot 10^{-6}$	
(78, 42, 59, 76), Etc.	$3.0323 \cdot 10^{-12}$	$1.6384 \cdot 10^{1}$	1.6384	$5.03 \cdot 10^{-7}$	$9.85 \cdot 10^{-6}$	
two genes removed by three more additional constraints [(a)]						
61, 73	$2.3229 \cdot 10^{-12}$	$1.6514 \cdot 10^{1}$	1.6514	$4.84 \cdot 10^{2}$	$3.05 \cdot 10^{-1}$	Yes
51, 82	$2.4282 \cdot 10^{-12}$	$1.6498 \cdot 10^{1}$	1.6498	$6.28 \cdot 10^{2}$	$3.48 \cdot 10^{-1}$	Yes
1, 10	$2.4588 \cdot 10^{-12}$	$1.6497 \cdot 10^{1}$	1.6497	$4.77 \cdot 10^{2}$	$3.03 \cdot 10^{-1}$	Yes
60, 45	$4.7780 \cdot 10^{-12}$	$1.4549 \cdot 10^{1}$	1.4549	$8.40 \cdot 10^{2}$	$4.02 \cdot 10^{-1}$	Yes
91, 22	$2.9067 \cdot 10^{-12}$	$1.2613 \cdot 10^{1}$	1.2613	$4.16 \cdot 10^{2}$	$2.83 \cdot 10^{-1}$	Yes
61, 73	$8.7392 \cdot 10^{-1}$	$4.6207 \cdot 10^{-12}$	~ 0	$1.85 \cdot 10^{-6}$	$1.89 \cdot 10^{-5}$	No
(60, 45), Etc.	$8.7393 \cdot 10^{-1}$	$4.6496 \cdot 10^{-12}$	~ 0	$2.29 \cdot 10^{-5}$	$6.64 \cdot 10^{-5}$	No
four genes removed by three more additional constraints [(a)]						
1, 10, 61, 73	$9.7773 \cdot 10^{-13}$	$1.7213 \cdot 10^{1}$	1.7213	$5.42 \cdot 10^{2}$	$3.23 \cdot 10^{-1}$	Yes
61, 73, 51, 54	$3.4073 \cdot 10^{-12}$	$1.6721 \cdot 10^{1}$	1.6721	$6.08 \cdot 10^{2}$	$3.42 \cdot 10^{-1}$	Yes
90, 21, 55, 48	$3.2508 \cdot 10^{-12}$	$1.6377 \cdot 10^{1}$	1.6377	$8.72 \cdot 10^{2}$	$4.10 \cdot 10^{-1}$	Yes
84, 73, 41, 1	$3.3764 \cdot 10^{-12}$	$1.6186 \cdot 10^{1}$	1.6186	$6.08 \cdot 10^{2}$	$3.42 \cdot 10^{-1}$	Yes
78, 42, 59, 76	$3.0796 \cdot 10^{-12}$	$1.6510 \cdot 10^{1}$	1.6510	$5.71 \cdot 10^{2}$	$3.32 \cdot 10^{-1}$	Yes
94, 84, 39, 85	$2.9541 \cdot 10^{-12}$	$1.6436 \cdot 10^{1}$	1.6436	$1.24 \cdot 10^{3}$	$4.89 \cdot 10^{-1}$	Yes
1, 10, 61, 73	$3.8510 \cdot 10^{-1}$	$1.6744 \cdot 10^{-12}$	~ 0	$8.44 \cdot 10^{3}$	$1.27 \cdot 10^{0}$	No
(61, 73, 51, 54), Etc.	$3.2685 \cdot 10^{-1}$	$8.0169 \cdot 10^{-12}$	~ 0	$1.32 \cdot 10^{4}$	$1.59 \cdot 10^{0}$	No
two genes removed by three more additional constraints [(a)] and $\Phi_{NLP} = v_s^2 v_b$						
61, 73	$2.3229 \cdot 10^{-12}$	$1.6434 \cdot 10^{1}$	1.6434	$1.03 \cdot 10^{2}$	$1.41 \cdot 10^{-1}$	No
(1, 10), Etc.	$5.10 \cdot 10^{-7}$	$1.6314 \cdot 10^{1}$	1.6314	$9.52 \cdot 10^{1}$	$1.35 \cdot 10^{-1}$	No

[(a)] Solutions of MINLP problem with the additional constraints of Burgardt *et al.*[33]

[(b)] succinate production rate/glucose consumption rate.

marked in Fig. 1, some of the solutions are expected, for instance the use of cell resources for succinate production maximization by blocking formation of ethanol (removed fluxes #1 and #10), lactate (removed flux #61), formate (removed flux #73), glutamine (removed fluxes #51 and #54), etc. (see some of these solutions in Fig. 2). Also, from the mathematical point of view it appears that F6P production by two alternative routes (#50 and #45) is redundant, and one of them should be removed. Such multiple solutions require a physical meaning evaluation to check viability of each *in-silico* resulting mutant cell. For instance, the removed gene set (82,83) is not viable as long as the flux #82 is responsible for PYR production and its elimination will not ensure the essential energy pathway of the cell. As the visual inspection of a larger number of solutions is difficult for complex cell system cases, an automatic rule is preferable, by using, for instance, the Wunderlich and Mirny[30] synthetic accessibility concept (described in the introductory part) to identify unfeasible cases of non-viable cells pathways when products cannot be synthesized from the network inputs. As another observation, Burgard *et al.*[33] have found several optimal-Pareto solutions by eliminating the oxygen uptake reactions (#36, #70) that maximize the succinate production. This alternative was not identified by our procedure after a significantly large number of trials, probably due to the reduced form of the model requiring the use of oxygen in Q8 production (#16), which is essential for the metabolism.

From the numerical point of view, solving the associated MILP/LP multi-objective problem leads to an extended combinatorial calculus when optimizing the sum of fluxes and also removing certain genes. The advantage of using the MINLP formulation comes from the concomitant random search for optimal fluxes and gene knockout alternatives during the same iterative rule, with the risk of missing gene knockout alternatives of similar quality in terms of objective function. This risk can be reduced when additional constraints are added to the optimization problem, or when a suitable flux prioritisation (by means of exponents β_j) is formulated. However, the proposed MINLP formulation presents some limitations. Similar to the weighted multi-objective and goal optimization cases, the proposed criterion leads to a reduced number of Pareto-optimal solutions (see the extended discussion of Nagrath *et al.*[53] in this respect). Two such Pareto-optimal solutions, corresponding to the removed gene set (1,10,61,73) are displayed in Fig. 2. An increased number of solutions can be obtained if a repeated application of the procedure is performed by using different sets of weights/exponents (varied within certain limits), with the expense of a considerable computational effort.

Solving gene knockout LP and MINLP problem with additional constraints. To check the effect of introducing new constraints to the optimization problem for reducing the set of gene knockout solutions, three additional constraints (9) suggested by Burgard *et al.*[33] are added to the multi-objective optimization problem (10). Some of the MINLP solutions, for the case of two- or four-genes simultaneously removed from the wild-type *E. coli* cell, are presented in Table 3. The results reveal several conclusions.

i) Multiple solutions are obtained as possible knockout alternatives to optimize the biomass and succinate production, valid for removing one, two, four or more possible genes from the metabolic network. Beside Pareto-optimal solutions (two of them are represented in Fig. 2 for removed genes #1, #10, #61, #72), the slow convergence of the used MINLP algorithm may lead to approximate (sub-optimal) solutions also, as indicated by the approximate fulfillment of the equality constraints, with an average error from 10^{-6} to up to 1.5 flux units compared to the ±1000 range. However, the sub-optimal solutions can be easily identified from the constraint fulfillment analysis and eventually removed. The used MINLP algorithms were the adaptive random search of Maria,[51] and a modification of the Nelder-Mead algorithm.[44] The use of other solvers (e.g. evolutionary algorithms) might improve the solution quality, but a trade-off between an acceptable precision level and the computational effort remains an open question.

ii) By inspecting the solutions (Fig. 2, Table 3) it is to observe that high biomass production rates correspond to low production in succinate, and vice-versa. This result is in perfect agreement with the experimental data and findings of Burgard *et al.*[33] from Fig. 2, which display a nonlinear dependence between the two mentioned stationary fluxes. For instance, negligible values of $v_{biomass}$ corresponds to ca. $v_{succinate}$ = 17 mmol gDW^{-1} h^{-1}.

iii) The analysis of the physical meaning of fluxes in the designed *E. coli* can indicate the viability of the solution. As previously mentioned and marked in Fig. 1, succinate production maximization can be achieved by removing some metabolic unessential steps, e.g. synthesis of ethanol, lactate, formate, glutamine, etc., which correspond to the findings of Burgard *et al*.[33] with using an extended cell model (i.e. removed genes no. 1, 3, 10, 50, 61, 73, 82, 83). However, such a check of cell viability for each of the multiple solutions is very laborious, and an automatic rule might be preferred (e.g. the Wunderlich and Mirny[30] synthetic accessibility rule). Other imposed system constraints can also be used in this respect.

Solving gene knockout MINLP problem by inducing high succinate production. The conflicting succinate and biomass production in *E. coli* cells indicates that several mutant cell solutions can be obtained, some with large $v_{succinate}$ and low $v_{biomass}$ levels, and vice-versa. To design cells with a high succinate production rate, one alternative is to "artificially" introduce a constraint to the MINLP problem defining the lower limit, e.g. $v_{succinate} = v_{39} > v_{39}^{\min} = 10$ mmol gDW^{-1} h^{-1}. The results indicate still a large number of possible reduced networks, of close efficiency. Some of the obtained *E. coli* cell solutions when removing two or four genes are presented in Table 3. For instance, the solution of removing genes {91,22} and $[v_{39}, v_{13}, L] = [12.6, 2.90 \cdot 10^{-12}, 416]$ is close to the solution of removing genes {61,73} and $[v_{39}, v_{13}, L] = [16.5, 2.32 \cdot 10^{-12}, 484]$. However, the plausible solution is that corresponding to the missed genes {61,73}, thus blocking the production of LAC and FOR without disturbing the main metabolic pathway (as is the case when removing the flux #91; Fig. 1).

An alternative to imposing "artificial" thresholds for some fluxes is to increase the importance of some fluxes in the composite MINLP objective function. For instance, by adopting $\beta_s = 2$ in (6), other *E. coli* cell solutions are obtained (two of them are displayed in Table 3). Not only derivation of mutant cells of high succinate productivity is thus favoured, but also the precision of the Pareto-optimal solution is roughly doubled (i.e. smaller $\overline{CV}_{eq}$ index) for the same computational effort. The benefit of such an adjustable flux weight in achieving the composite goal is obvious, by the expense of requiring an extended computational effort when investigating adjustable relative prioritizations of fluxes in the MINLP formulation. Comparatively to the LP approach, the advantage of including nonlinear correlations among fluxes and constraints recommends the MINLP approach as a worthy alternative for designing optimal cells.

Conclusions

Application of a MINLP procedure to find theoretical gene knockout alternatives that optimize several formulated objectives (e.g. maximize target metabolite production) has been proved to be very promising to *in-silico* design mutant cells. This computational strategy can partly overcome the complex combinatorial MILP problem that corresponds to a multi-layer LP formulation, and can save considerable computing time by superposing the knockout rule to the basic NLP optimization approach. However, the identified multiple solutions of the MINLP problem, explained by the cell metabolism complexity, must be further 'filtered' by adding supplementary (non)linear constraints (other than the stoichiometric ones), leading to a considerable reduction in the number of gene-knockout alternatives.

The use of the LP formulation $Max(\boldsymbol{wv})$, with subjective weights allocated to the target fluxes, or transformation of the primal problem into a dual LP problem (for only bi-level optimization) is laborious, requiring to formulate and prepare the derived dual problem. Moreover, the solution is dependent on the adopted upper/lower bounds of the dual variables, while the resulting LP combinatorial problem when removing several genes from a large number becomes computationally intensive. Recent improvements of the LP method lead to a better description of the optimal solution set, but do not overcome the combinatorial problem.

The advantage of the proposed power-law type MINLP multi-objective function comes from the possibility of accounting, in a simple way, for the flux nonlinear interactions and complex constraints as mentioned in the literature.[28,34,35] The combinatorial rule is included in the iterative MINLP solver, while the larger number of considered (nonlinear) constraints can increase the chance to obtain a reduced set of feasible gene-knockout solutions for a given metabolic network. The preferred random search can offer a higher reliability in finding a global solution (if any) of the optimal-flux-gene- knockout problem, with also providing the opportunity for the integer variables to span their range of possible values during the flux optimization. In such a manner, a continuous evaluation of the effects of removing various genes during the MINLP solver iterations is realized. Because the random searches are usually slowly convergent near the problem solution, approximate solutions are usually retained, with an acceptable precision of fulfilling the problem constraints. However, derivation of a larger set of optimal solutions by using an adjustable relative prioritization of fluxes in the MINLP formulation will lead to an extended computational effort.

In any variant, the resulting multiple gene knockout solutions have to be validated from several points of view, both theoretically (physical meaning) and experimentally. To reduce the number of solutions, formulation of problem constraints is crucial. Gene inference,[6] feasible reaction paths,[30] or any information on protein-gene interactions, and on the regulatory circuits can be used for such purposes. On the other hand, several criteria to check the design optimal cells for viability can be used, for instance the Wunderlich and Mirny[30] synthetic accessibility index, or additional information on the gene inferences.

A more systematic rule for designing mutant cells should be based on using hybrid stationary-dynamic models to incorporate stationary and kinetic information on flux distribution over time.

ACKNOWLEDGMENT

All the three authors thank the Chinese Academy of Sciences and the Chinese Ministry of Science and Technology for their financial support (with the project KIP KSCX2-YW-G-030 and the 973 program sub-project 2007CB707804, respectively).

Annex 1 – Metabolic stoichiometric balance

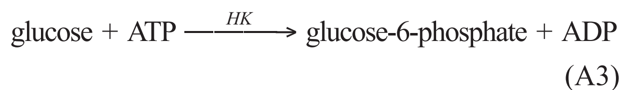
Chemical and biochemical kinetics are based on the postulate that a reaction rate, v_j, can be expressed as a unique (usually nonlinear) function of the concentrations, c_i, of all participating chemical species at a certain time t. When (bio)chemical reactions are the only cause of concentration changes, that is the transport processes are negligible, the concentration dynamics is given by the balance equation:

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = \sum_{j=1}^{M} S_{ij} v_j , \quad i = 1,\dots, N \qquad \text{(A1)}$$

When the biochemical system subsists in a steady state, the balance equation (A1) in a matrix formulation becomes:

$$\boldsymbol{S}\, \boldsymbol{v}(\boldsymbol{c}^*) = 0 \qquad \text{(A2)}$$

where '*' superscript denominates the steady-state values of concentrations. To exemplify the way to relate the stoichiometric matrix to the steady state reaction rates (denominates as fluxes), the following metabolic reactions are considered:[9]



glucose + ATP $\xrightarrow{HK}$ glucose-6-phosphate + ADP

(A3)

glucose-6-phosphate $\xrightarrow{PGM}$ glucose-1-phosphate

The reactions are catalysed by the enzymes hexokinase (HK, EC 2.7.1.1) and phosphoglucomutase (PGM, EC 5.4.2.2), respectively. By attaching the stoichiometric matrix of the two reactions (indexed with 1 and 2), the steady-state mass balance (A2) can be written as:

$$\boldsymbol{S}\,\boldsymbol{v} = \begin{matrix} \\ gluc \\ G6P \\ G1P \\ ATP \\ ADP \end{matrix} \begin{pmatrix} HK & PGM \\ -1 & 0 \\ 1 & -1 \\ 0 & 1 \\ -1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{(A4)}$$

In the mass balance, all reacting species involved have been included. To be feasible, the mass balance has to be completed with input/output fluxes into/from the system. For larger biochemical systems, this procedure is neither necessary nor useful, allowing checking the conservation relationships to derive large models, and to study some properties of the system, such as conservation relations and reaction invariants, as long as the stationary flux vector is the null space of the matrix $\boldsymbol{S}$.[9]

Notations

A_j – reaction affinity[9]
C_a^b – $a!/b!\,(a-b)!$ = number of ways of choosing b objects from a collection of a objects regardless of order
CV_{eq} – equality constraint violation index
CV_{ineq} – inequality constraint violation index
c_i – species i concentration
f_j – individual objective functions
KG – number of removed genes
L – Lagrange function of the optimization problem
$\lambda_{eq}, \lambda_{ineq}$ – Lagrange multipliers of the equality and inequality constraints
M – number of fluxes in the considered metabolic pathway
N – number of metabolites
$\boldsymbol{S}$ – stoichiometric matrix (of elements S_{ij}, i.e. the stoichiometric coefficient of the metabolite i in the reaction j).
S_j – synthetic accessibility of an output j, i.e. the minimal number of metabolic reactions needed to produce component j from the network inputs
S_t – total synthetic accessibility index, $S_t = \Sigma_i S_i$.
T – temperature
t – time
v, v_j – vector of stationary metabolic fluxes, or reaction rate
y_j – Boolean variables
w_j – weights of individual objective functions

Greeks

β_j – exponent of v_j flux in the NLP objective function
Φ – objective function
λ_i^{stoich} – dual variable associated with the stoichiometric constraint involving reaction i
μ_j – dual variable associated with any restriction of the flux v_j

Index

low – lower limit
max – maximum
min – minimum
up – upper limit

Abbreviations

ADP – adenosine diphosphate
AMP – adenosine monophosphate
arg – argument of
ATP – adenosine-5'-triphosphate coenzyme
DF – degree of freedom
DFBA – dynamic flux balance analysis
DW – dry weight
EMA – elementary mode analysis
ExPA – extreme pathways analysis
FBA – flux balance analysis
LP – linear programming
LSQ – least squares
MCA – metabolic control analysis
MILP – mixed integer LP
MINLP – mixed integer NLP
MOMA – minimization of metabolic adjustment' method
NEQ – number of equality constraints
NLP – nonlinear programming
rank – rank of a matrix
sign – sign of
s.t. – subject to
$\|\bullet\|_2$ – Euclidean norm

References

1. *Buchholz, K., Hempel, D. C.,* Eng. Life Sci. **6** (2006) 437.
2. *Ma, H. W., Zhao, X. M., Yuan, Y. J., Zeng, A .P.,* Bioinformatics **20** (2004) 1870.
3. *Maria, G.,* Chem. Biochem. Eng. Q. **19** (2005) 213.
4. *Maria, G.,* Chem. Biochem. Eng. Q. **21** (2007) 417.
5. *Maria, G.,* Rev. Chim. **61** (2010) 172.
6. *He, F., Balling, R., Zeng, A. P.,* Jl. Biotechnol. **144** (2009) 190.
7. *Jandt, U., Shao, S., Wirth, M., Zeng, A. P.,* Biotechnol. Bioeng. **108** (2011) 2205.
8. *Kacser, H., Burns, J. A.*, Proc. Symp. Soc. Exper. Biol. Rate Control Biol. Proc., Cambridge Univ. Press, Cambridge, 1973, pp. 65–104.
9. *Heinrich, R., Schuster, S.,* The regulation of cellular systems, Chapman & Hall, New York, 1996.
10. *Stelling J, Klamt, S., Bettenbrock, K., Schuster, S., Gilles, E. D.,* Nature **420** (2002) 190.
11. *Papin, J. A., Price, N. D., Palsson, B. O.*, Genome Res. **12** (2002) 1889.
12. *Arkin, A., Ross, J.,* J. Phys. Chem. **99** (1995) 970.
13. *Arkin, A., Shen, P. D., Ross, J.*, Science **277** (1997) 1275.
14. *Kauffman, K. J., Pajerowski, J. D., Jamshidi, N., Palsson, B. O., Edwards, J. S.,* Biophys. J. **83** (2002) 646.
15. *Mahadevan, R., Edwards, J. S., Doyle, F. J.,* Biophys. J. **83** (2002) 1331.
16. *Tušek, A., Kurtanjek, Z.,* Chem. Biochem. Eng. Q. **23** (2009) 527.
17. *Stephanopoulos, G. N., Aristidou, A. A., Nielsen, J.,* Metabolic engineering. Principles and methodologies, Academic Press, New York, 1998.
18. *Vallino, J. J., Stephanopoulos, G.,* Flux determination in cellular bioreaction networks: Applications to lysine fermentations, In: Frontiers in bioprocessing (Sikdar, S. K., Bier, M., Todd, P., Eds.), CRC Press, Boca Raton, 1990.
19. *Vallino, J. J., Stephanopoulos, G.,* Biotechnol. Bioeng. **41** (1993) 633.
20. *Edwards, J. S., Palsson, B. O.,* Proc. Natl. Acad. Sci. USA **97** (2000) 5528.
21. *Raghunathan, A. U., Perez-Correa, J. R., Biegler, L. T.,* Biotechnol. Bioeng. **84** (2003) 700.
22. *Schmidt, K., Carlsen, M., Nielsen, J., Villadsen, J.,* Biotechnol. Bioeng. **55** (1997) 831.
23. *Nöh, K., Grönke, K., Luoa, B., Takors, R., Oldiges, M., Wiechert, W.,* J. Biotechnol. **129** (2007) 249.
24. *Deshpande, R. R., Yang, T. H, Heinzle, E.,* J. Biotechnol. **4** (2009) 247.
25. *Yuan, Y., Yang, T. H., Heinzle, E.,* Metab. Eng. **12** (2010) 392.
26. *Modak, J., Deckwer, W. D., Zeng, A. P.,* Biotechnol. Prog. **18** (2002) 1157.
27. *Segre, D., Vitkup, D., Church, G. M.,* ONAS **99** (2002) 15112.
28. *Sainz, J., Pizarro, F., Perez-Correa, J. R., Agosin, E.,* Biotechnol. Bioeng. **81** (2003) 818.
29. *Edwards, J. S., Ibarra, R. U., Palsson, B. O.,* Nature Biotechnol. **19** (2001) 125.
30. *Wunderlich, Z., Mirny, L. A.,* Biophys. J. **91 (**2006) 2304.
31. *Styczynski, M. P., Stephanopoulos, G.,* Comput. Chem. Eng. **29** (2005) 519.
32. *Zak, D. E., Vadigepalli, R., Gonye, G. E., Doyle III, F. J., Schwaber, J. S., Ogunnaike, B. A.*, Comput. Chem. Eng. **29** (2005) 547.
33. *Burgard, A. P., Pharkya, P., Maranas, C. D.,* Biotechnol. Bioeng. **84** (2003) 647.
34. *Vera, J., Atauri, P., Cascante, M., Torres, N. V.,* Biotechnol. Bioeng. **83** (2003) 335.
35. *Kayser, A., Weber, J., Hecht, V., Rinas, U.,* Microbiology **151** (2005) 693.
36. *Orth, J. D., Fleming, R. M. T., Palsson, B. O.,* Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide, EcoSal Web: Escherichia coli and Salmonella: Cellular and molecular biology, ASM Press, 2008, Module 10.2.1. http://www.ecosal.org/. doi:10.1128/ecosal.10.2.1
37. *Palsson, B.,* Flux-balance analysis: Basic concepts, Lecture #6, Systems Biology Research Group, University of California, San Diego, 2003. http:// systemsbiology.ucsd.edu.
38. *Varma, A., Palsson, B. O.,* J. Theor. Biol. **165** (1993) 503.
39. *Bard, J. F.,* Practical bilevel optimization, Kluwer Academic Publ., Boston, 2005.
40. *Gumus Z. H., Floudas, C. A.,* J. Global Opt. **20** (2001) 1.
41. *Gumus Z. H., Floudas, C. A.,* Comput. Manag. Sci. **2** (2005) 163.
42. *Mitsos, A., Lemonidis, P., Barton, P. I.,* J. Global Opt. **42** (2008) 475.
43. *Floudas, C. A.,* Nonlinear and mixed-integer optimization, Oxford University Press, Oxford, 1995.
44. *Coleman, T., Branch, M. A., Grace, A.,* Optimization toolbox for use with MATLAB. User's Guide (ver. 2), MathWorks Inc., Natick (MA), 1999.
45. *Sbalzarini, I. F., Mueller, S., Koumoutsakos, P.,* Multi-objective optimization using evolutionary algorithms, Proc. Summer Program, Center for Turbulence Research, NASA Ames, Stanford Univ., 2000, pp. 63–74.

46. *Sunar, M., Kahraman, R.,* Turk. J. Engin. Environ. Sci. **25** (2001) 69.
47. *Tang, K. Z., Sun, T. K., Yang, J. Y.,* Comput. Chem. Eng. **35** (2011) 615.
48. *Rao, S. S.,* Engineering optimization – Theory and practice, Wiley, New York, 2009. pp. 761–767.
49. *Fletcher, R.,* Practical methods of optimization: Constrained optimization, Wiley, New York, 1980, vol. 2.
50. *Zwillinger, D., Krantz, S. G., Rosen, K. H.,* Standard Mathematical Tables and Formulae, CRC Press, Boca Raton, 1996.
51. *Maria, G.,* ARS combination with an evolutionary algorithm for solving MINLP optimization problems, In: Modelling, Identification and Control, M.H. Hamza (Ed.), IASTED/ACTA Press (Canada), 2003, pp. 112–118.
52. *Maier, K., Hofmann, U., Reuss, M., Mauch, K.,* Biotechnol. Bioeng. **100** (2008) 355.
53. *Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., Yarmush, M. L.,* Metab. Eng. **12** (2010) 429.
54. *Link, H., Vera, J., Weuster-Botz, D., Darias, N. T., Franco-Lara, E.,* Comput. Chem. Eng. **32** (2008) 1707.
55. *Selvarasu, S., Lee, D. Y., Karimi, I. A.,* Comput. Aided Chem. Eng. **24** (2007) 1007.
56. *Saha, R., Suresh, S., Park, W., Lee, D. Y., Karimi, I. A.,* In: Proc. 17th European Symposium on Computer Aided Process Engineering (V. Plesu, P.S. Agachi, Eds.), 2007, Elsevier, Amsterdam. Paper #T4–379.
57. *Voit, E. O.,* IEE Proc. Syst. Biol. **152** (2005) 207.
58. KEGG – Kyoto encyclopedia of genes and genomes, Kanehisa Laboratories, Bioinformatics Center of Kyoto University, 2011. http://www.genome.jp/kegg/pathway.html