Constraints-Based Modeling to Identify Gene Targets for Overproduction of Ethanol by *Escherichia coli:* The Effect of Glucose Phosphorylation Reaction

Desai T and Srivastava S*

DBT-ICGEB Center for Advanced Bioenergy Research, International Center for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi, India

Abstract

E. coli can metabolize both C5 and C6 sugars, but produces many side-products along with ethanol in the mixed-acid fermentation. Identification of knock-out gene targets that maximize ethanol formation with minimum impact on cell growth can help optimize the fermentation. Constraints-based analysis of genome scale metabolic models (GSMMs) helps to identify gene targets for overproducing desired products and also predict the cellular responses with those predictions. We utilized the latest GSMM of *E. coli* (iJO1366) to identify the gene targets associated with increased ethanol production on glucose and xylose as the C-source. Additionally, we also analyzed *in silico* the predicted intracellular flux distributions in the knock-outs by employing the Relative Change (RELATCH) method. For the KO mutants for which RELATCH gave infeasible solution for flux distribution, we employed "hybrid-MOMA", a method we introduce in this paper. We also studied the effect of choosing the glucose phosphorylation reaction and the glucose intake flux on the targets identified. Our results demonstrate that the targets identified by GDLS vary with the glucose phosphorylation system chosen and not with the glucose intake flux. These results have implications for products beyond ethanol and for species beyond *E. coli*, when employing GDLS or other constraints-based approaches for identifying knock-out targets for metabolite overproduction. Our results demonstrate the importance of knowing correct phosphorylation system for identifying meaningful targets with constraints-based approaches.

Introduction

Lignocellulosic biofuels provide a promising alternative to fossil-fuels currently used. This has led to intense research in this area in order to identify the best conditions for the different unit operations to make the product economically viable. Hydrolysis of lignocellulosic biomass releases a mixture of hexose and pentose sugars (primarily glucose and xylose), with pentoses accounting for 20% to 25% of the total sugars [1-3]. There fore, a major requirement for biofuel production is that the process organism must be able to utilize both C5 and C6 sugars. Yeast (*S. cerevisiae*) normally used for ethanol production using molasses cannot ferment C5 sugars; attempts have been made to engineer *S. cerevisiae* to utilize xylose. Alternatives to *S. cerevisiae* which naturally ferment both C5 and C6 sugars, e.g., *E. coli* and *S. stiptis* are being explored. However, most organisms, including *E. coli*, conduct "mixed acid fermentation", i.e., they produce acetic acid and lactic acid in addition to ethanol in order to maintain the redox balance. To overproduce ethanol, competing pathways are typically deleted. While such rational metabolic engineering approaches have helped produce many compounds, it is not possible to predict the effects of such deletions on growth.

Systems biology approaches can help predict cellular responses under different environmental conditions [4], e.g., on different C sources and the effect of deleting metabolic genes. One of the most popular methods of metabolic systems biology is constraints-based analysis of genome scale metabolic models (GSMMs). This is based on Flux Balance Analysis (FBA) approach, which assumes an internal pseudo steady state of metabolites within the cells. FBA approach requires information only about the reaction stoichiometries. It does not take into account the kinetic parameters of enzymes catalyzing them which are often unknown. Thus it is possible to solve the resultant linear system of equations for a much larger scale of model. GSMMs of *E. coli* have undergone many rounds of iterations to make them more comprehensive and predictive. Most recent GSMM of *E. coli*, iJO1366, comprises of 1366 genes, 2251 metabolic reactions, and 1136 unique metabolites [5,6]. Methods have been developed to analyze these models in order to identify gene targets for metabolite overproduction. Additionally it is also possible to predict the cellular response to gene knock outs.

We identified the gene targets for overproduction of ethanol by *E. coli* on glucose and xylose as the carbon sources (C-sources). These two sugars are the primary sugars in the lignocellulosic hydrolysates. Genetic Design through Local Search (GDLS) method was employed to identify the targets that can significantly improve ethanol production while having less effect on cell growth. GDLS conducts local search of the solution space in order to efficiently find an optimum, though the optimum need not be a global minimum or maximum of the system. We also identified the predicted intracellular fluxes after these knockouts using a method called Relative Change (RELATCH). RELATCH employs 13C metabolic flux analysis (MFA) data from the wild-type strain in order to constrain the flux distribution in the knock-out strain. A new method, "hybrid-MOMA" is introduced to predict flux distribution in the mutants for the cases where RELATCH gives infeasible solution. Last, but not the least, we also evaluated the

*Corresponding author: Srivastava S, DBT-ICGEB Center for Advanced Bioenergy Research, International Center for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhil, India; Tel: +91-11-26741358; Fax: +91-11-26742316; Email: shireesh@icgeb.res.in

effect of glucose phosphorylation reaction on the identified targets and show that the targets identified using GDLS depend on specifying the correct phosphorylation reaction.

Materials and Methods

Materials

The analyses were conducted on a Dell Precision T7600 Tower Workstation with eight cores and 8 GB RAM. These analyses utilized the Constraints-based Reconstruction and Analysis (COBRA) toolbox (version 2.0.5) [7] on MATLAB (2013b) platform. Gurobi 5.6.2 was used as the optimization problem solver. The analyses employed the iJO1366 metabolic model of *E. coli*. The built-in programs for GDLS and RELATCH analyses within the COBRA toolbox were utilized.

Methods

GDLS parameters: GDLS parameters were kept as follows: k, 6; M, 2; maximum number of knockouts, 3 and the minimum flux through biomass reaction was kept to be half of the maximum biomass yield predicted in anaerobic conditions at respective carbon uptake rates. The target reaction search was limited to following metabolic pathways: alternate carbon metabolism, citric acid cycle, glycolysis/gluconeogenesis, glyoxylate metabolism; pentose phosphate pathway and pyruvate metabolism. The ATP maintenance flux was fixed to 8.39 mmol/(gDW·hr).

Testing the effect of glucose intake rates on identified targets: GDLS was run to identify targets for ethanol formation at glucose uptake rates of 10 mmol/gDW/h and 15 mmol/gDW/h. The former flux value is commonly employed in the literature to conduct *in silico* constraints-based analyses [5,7,8].The latter flux value was chosen based on glucose intake flux reported for anaerobic condition [9].The xylose intake rate of 10 mmol/gDW/h was chosen for the analyses.

Testing the effect of glucose transport system: In order to test the effect of glucose transport system on target identification, the reactions corresponding to other transport systems (Glucose isomerase and/or hexokinase) were removed from the GSMM prior to running GDLS simulations.

Predicting intracellular flux distributions using RELATCH and "hybrid-MOMA": For the glucose simulations, possible flux distribution in the knockout model was calculated with respect to the previously published 13C-MFA based flux distribution in wild type (WT) model under anaerobic conditions [9], using RELATCH method [10]. The values of alpha and gamma were one and infinity respectively. These parameter values were chosen to predict the distribution in an "evolved" mutant.

For performing hybrid MOMA, the wild type flux distribution was obtained through RELATCH. The bounds of the reactions for which the 13C-MFA data was available were fixed to values calculated by RELATCH ± error in 13C-MFA data.

Results

GDLS-identified gene targets and their impact on ethanol formation and growth

The reactions pyruvate formate lyase (PFL) and phosphoglucoisomerase (PGI) were identified as knockout targets for ethanol production by GDLS in separate simulations with glucose and xylose as carbon source (Table 1). Indeed, in the KO mutants, the GDLS- predicted maximum flux of ethanol formation was >90%

	Glucose	Xylose
Minimum ethanol production	0.00	0.00
Maximum ethanol production	18.49	15.71
Maximum Predicted yield (% of theoretical maximum)	92.45	94.08
biomass	0.10	0.07
Targets	Pyruvate formate lyase	Pyruvate formate lyase
	G6P isomerase	G6P isomerase

Table 1: Results of GDLS simulations. Uptake rates for both glucose and xylose were chosen as 10 mmol/gDW/h. Product secretion rates are in mmol/gDW·hr. Ethanol yield is shown in percentage of maximum yield.

of the theoretical maximum for both glucose and xylose as the C source. However, the minimum flux through ethanol formation was 0, suggesting that it may be possible that the mutant does not produce ethanol.

Flux distribution in WT and the knockout suggested by GDLS was calculated using RELATCH method. However, lactate was predicted as a major byproduct in this KO. This was in conjunction with previous report that knocking out pyruvate formate lyase (pfl) gene increased lactate production in *E. coli* [11]. For further simulations, we tested the effect of lactate dehydrogenase (LDH) reaction KO by constraining zero flux through the reaction. Indeed, ethanol secretion was higher in PFL, PGI and LDH knockout model. PFL and LDH reactions have been used as metabolic engineering targets previously in *E. coli* and other organisms [11-14]. In order to evaluate the role of PGI as a KO target, the flux distribution in PFL and LDH mutant was calculated. Pyruvate was predicted to be a major product with no flux through ethanol secretion.

GDLS simulations up to maximum of ten reaction targets were also performed. However, the increase in target metabolite production was marginal for reaction targets greater than three (not shown).

RELATCH analysis of flux distribution in the GDLS-predicted knock-out mutants

In order to understand the mechanism of improved ethanol formation in the GDLS-predicted KO, the intracellular flux distribution in the PFL, PGI and LDH reaction knockout strain was predicted using the RELATCH method [10]. The predicted flux distributions of WT and knockout models are shown in Figure 1a and 1b, respectively.

A major determinant of ethanol formation is the redox state of the cell. In the WT cells, the excess NADH is utilized to reduce Acetyl-CoA to ethanol. The metabolism of pyruvate to acetyl-CoA through PFL in anaerobic conditions generates formate as a product. In the KO mutant, glucose metabolism was primarily through the xylose (glucose) isomerase instead of the EMP pathway. Additionally, it was observed that in the "evolved" mutant, the predicted flux from pyruvate to acetyl-CoA was through Pyruvate Dehydrogenase (PDH) system. This flux through PDH generated additional NADH than in the WT strain. It must be noted that PDH is inactive under anaerobic conditions in WT *E. coli* due to repression of its genes under anaerobic conditions. The predicted flux in the KO mutant suggests that the flux from pyruvate to Acetyl-CoA must go through PDH to improve ethanol formation. This would require PDH to be expressed under promoters that are active in anaerobic conditions. Munjal et al. [15] had used the same
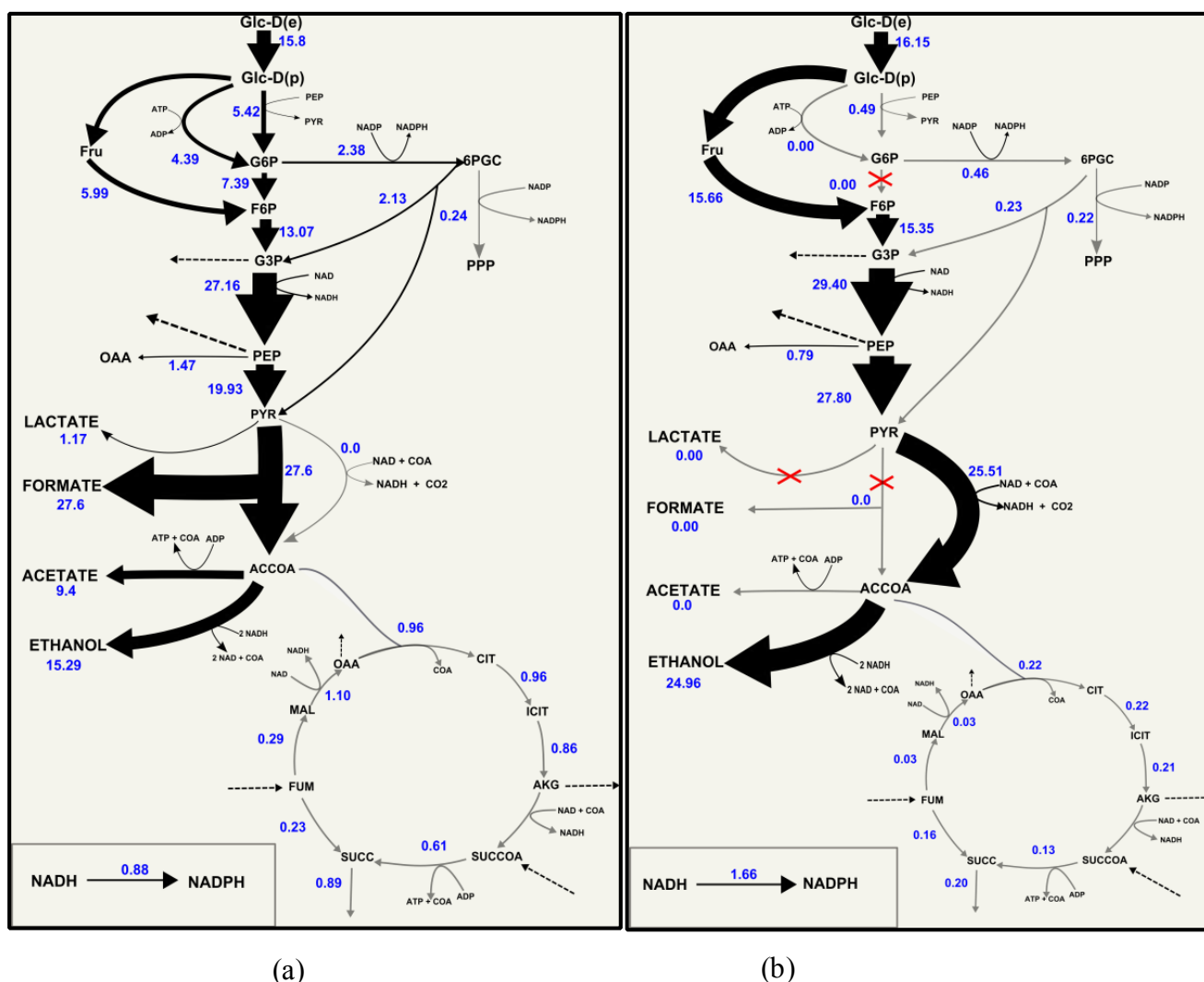
Figure 1: Simulated flux distribution in WT (a) and KO (b) model using RELATCH. Biomass_WT = 0.23, Biomass_KO = 0.19.

strategy to improve ethanol formation from glucose and xylose. They had expressed PDH genes under different constitutive promoters so that this reaction is active under anaerobic conditions.

The effect of glucose intake rates and phosphorylation reaction on ethanol targets identified by GDLS

RELATCH predicted a significant flux through the xylose (glucose) isomerase (XI) reaction to generate fructose from glucose in both WT and KO (Figure 1a and 1b). The fructose was then phosphorylated by the fructokinase to F6P. While in the WT strain, the glucose phosphorylation occurred through all the three reactions (XI, hexokinase and PTS) to almost equal extent, in the KO mutant XI was predicted to be the primary phosphorylation reaction. However, it is well known that in *E. coli*, PTS is the primary glucose transport system and there are no reports on significant flux through the XI reaction. Therefore, we investigated whether the glucose phosphorylation reaction affects the KO targets predicted by GDLS. For this analysis, all the glucose intake flux was restricted through (PTS + hexokinase) or PTS only (Table 2). The targets identified by the GDLS for the

conditions where the flux through XI was not allowed were different than those when the flux through XI was allowed (compare targets in Table 1 vs. Table 2), though PFL was identified as the target in all the cases. The minimum ethanol production predicted by GDLS was still zero, though the maximum was greater than ninety percent of the theoretical maximum. It must be mentioned that the targets identified for xylose as C-source did not change with changing the glucose phosphorylation reaction.

Increasing the glucose uptake rates from a traditional value of 10 mmol/gDW·hr to near-13C-MFA value of 15 mmol/gDW·hr did not affect the targets identified. Lactate was again predicted to be a major byproduct by RELATCH. Hence we included LDH as a target for further analysis. However, RELATCH was unable to identify the flux distribution in these KOs. Therefore, we derived the flux distribution in these mutants by hybrid-MOMA (see methods) which predicted a succinate as a major byproduct in the KOs. *E. coli* AFP111 and *E. coli* NZN111, which lacks PFL and LDH genes, have been reported to produce higher amount of succinate under anaerobic condition [16]. Deleting fumarate reductase (FRD) reactions prevented the

high flux towards succinate (Table 3). *In silico* prediction of flux through product is biased by method of calculating flux distributions. However, minimizing the product flux may give an idea about how much minimum flux can be expected from the KO strain. We found that the minimum flux towards ethanol is similar irrespective of FRD being included as target. When only PFL, LDH, and FRD were used as KO targets, the minimum ethanol flux was lower than when PFK, SGL, PFL and LDH were used as KO targets. While the addition of FRD as the target did not impact the minimum ethanol formation predicted by FBA, the ethanol formation predicted by hybrid-MOMA was significantly increased. Flux distribution predictions by hybrid MOMA for the WT and in PFK, SGL, PFL, LDH and FRD knockout mutant is shown in Figure 2 a and b. The deletion of PFK significantly increased the flux through pentose phosphate pathway and the deletion of SGL prevented the loss of C to E4P. Also, in the mutant, there was increased flux through PTS system and DHAP-phosphotransferase which led to increased pyruvate synthesis. The synthesis of acetyl-CoA from pyruvate in the mutant was increased by distributing the flux through PDH and pyruvate synthase. The increased acetyl-CoA led to increased ethanol formation.

Discussion

Constraints-based analyses of GSMM are very useful to predict cellular behavior under different environmental and genetic conditions. Associated methods have been developed that can identify KO and overexpression candidates to overproduce a desired metabolite. An important factor investigated in this study was the effect of glucose transport system on the targets identified. The PTS system has been reported to be the primary transport and phosphorylation systems for

E. coli. Glucose intake and phosphorylation through PTS consumes PEP during the import which is regenerated during the glycolytic metabolism. The ABC transporter requires ATP which is generated through a variety of substrate-level phosphorylation under anaerobic conditions. The observation that different phosphorylation reactions lead to different targets raises the point that in order to make meaningful predictions, the correct phosphorylation reaction must be known. While such an information is known for many of the well-characterized organisms, for the newly-sequenced and less characterized organisms, it is not sufficient to just know all the transport systems, but to know which is the primary pathway through which glucose or the primary C-source is phosphorylated.

Typically, most simulations in literature have used 10 mmol/gDW/h glucose intake rates. This value is closer to glucose intake rates in aerobic conditions [9]. However, it has been reported that glucose intake flux is higher under anaerobic conditions, perhaps to compensate for reduced ATP generation per mol of glucose under anaerobic conditions. Here we investigated the effect of glucose intake rate on targets identified by GDLS. Our results show that beyond 10 mmol/gDW/h, glucose intake rate did not affect the targets identified.

GDLS also provides the minimum and maximum flux through the product. Both these values are useful when deciding the KO targets. However, these extreme values, while providing a useful range, may over- or under- predict the actual product formation rates. For example, GDLS predicted >90% flux through ethanol formation with just PFL and PGI KOs, even though it is known that such a mutant will produce lactate and pyruvate as products. RELATCH, which utilizes 13C-MFA data for the WT cells may provide more realistic

Carbon Uptake Mode	Only PTS		PTS and Hexokinase	
Targets	PFK	PFK	PFK	FBA
	S7PK	SGL	S7PK	SGL
	PFL	PFL	PFL	PFL
Minimum ethanol production	0.00	0.00	0.00	0.00
Maximum ethanol production	18.23	18.23	18.12	18.12
Biomass	0.12	0.12	0.13	0.13

Table 2: GDLS solution when glucose uptake rates are restricted through only PTS system or PTS and hexokinase. Uptake rates for both glucose and xylose were chosen as 10 mmol/gDW/h. Product secretion rates are in mmol/gDW·hr. Ethanol yield is shown in percentage of maximum yield.
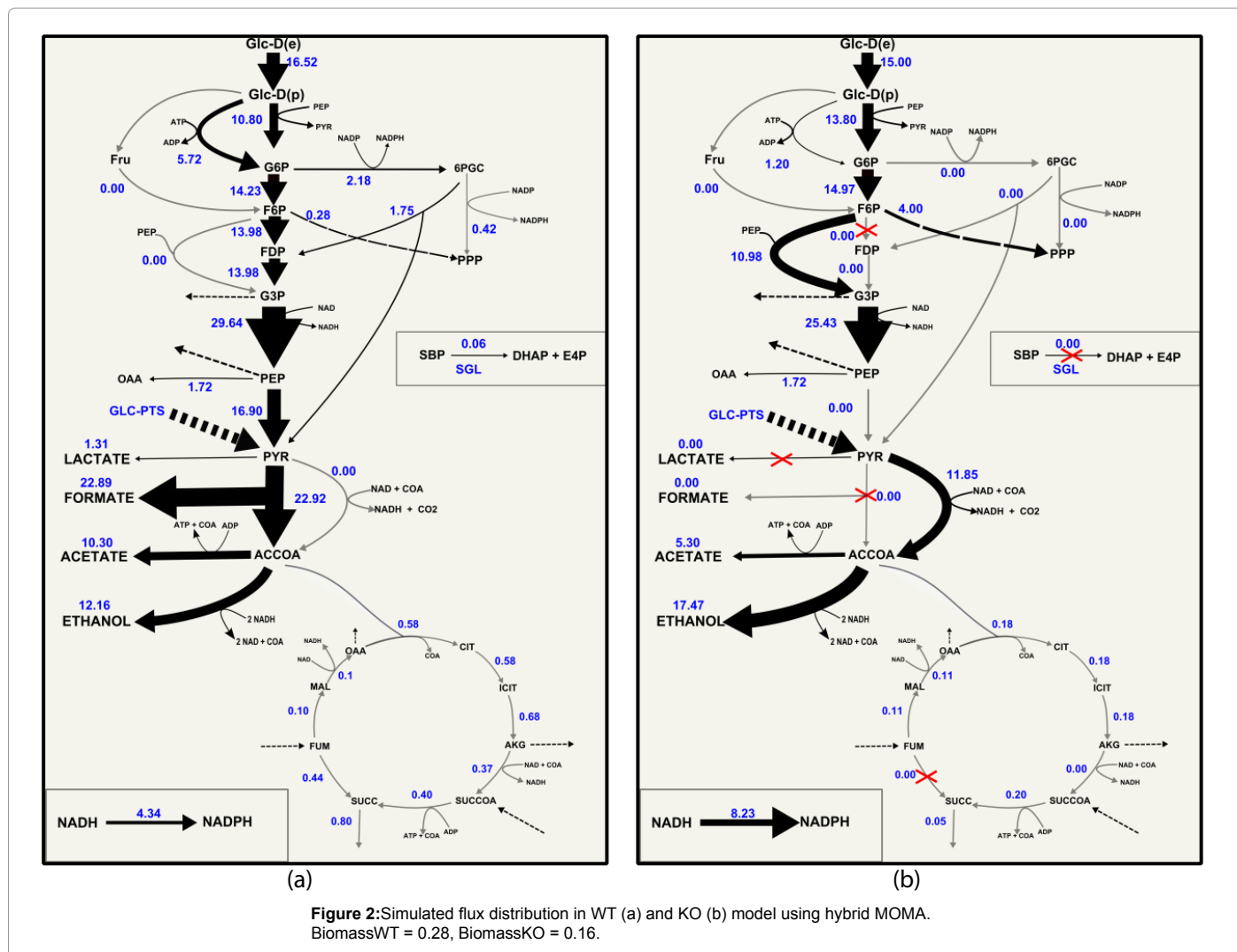
Abbreviations: PFK: Phosphofructokinase; S7PK: Sedoheptulose 7-Phosphate Kinase; SGL: Sedoheptulose1,7-Bisphosphate D-Glyceraldehyde-3-Phosphate-Lyase; FBA: Fructose-Bisphosphate Aldolase; PFL: Pyruvate Formate Lyase.

	WT	PFK, SGL, PFL	PFK, SGL, PFL, LDH	PFK, SGL, PFL, LDH, FRD	PFL, LDH, FRD
Biomass	0.28	0.16	0.16	0.16	0.16
Glucose uptake rate	16.52	15.00	15.00	15.00	15.00
Ethanol	12.38	8.47	14.63	17.47	15.89
Lactate	1.31	9.23	0.00	0.00	0.00
Formate	22.44	0.00	0.00	0.00	0.00
Acetate	10.30	5.25	5.52	5.30	4.55
Succinate	0.80	5.55	8.05	0.05	0.05
minimum ethanol *	0	0	13.42	13.56	12.82

*FBA prediction

Table 3: Hybrid MOMA-predicted fluxes through major products. All flux values except shown in mmol/gDW·hr. Biomass flux shown in hr-1. The minimum ethanol flux was calculated using FBA by minimizing the flux through this reaction and not through hybrid MOMA.

Abbreviations: PFK: Phosphofructokinase; SGL: Sedoheptulose 1,7-Bisphosphate D-Glyceraldehyde-3-Phosphate-Lyase; FBA: Fructose-Bisphosphate Aldolase; PFL: Pyruvate Formate Lyase; FRD: Fumarate Reductase.

Figure 2: Simulated flux distribution in WT (a) and KO (b) model using hybrid MOMA. BiomassWT = 0.28, BiomassKO = 0.16.

product formation rates in mutants. The RELATCH-predicted ethanol formation flux in the best mutant is ~58% of the maximum. This value is very close to a yield of ~54% of the maximum reported for LDH-FRD mutant expressing PDH under GAPDH promoter growing on defined medium [15].

Some of the targets identified by us, e.g., PFL, LDH and FRD, are logical targets for ethanol production and have been reported previously to improve ethanol formation. A consistent observation in all the "evolved" KOs is the flux through PDH which increases the NADH yield and improves ethanol formation. This approach has been successfully employed in a previous study [15]. Our analyses suggest that addition of PFK and SGL may further improve the ethanol formation.

Therefore, our results show that, while FBA-based *in silico* methods are very useful to identify suitable targets with minimal experimental information input; additional information on cellular systems such as correct phosphorylation reaction is needed to identify meaningful targets. Also, utilizing the 13C-MFA data may provide more realistic predictions of performance of the mutants than those given by purely FBA-based methods.

References

1. Jørgensen H, Kristensen JB, Felby C (2007) Enzymatic conversion of lignocellulose into fermentable sugars: Challenges and opportunities. Biofuel Bioprod Bior 1: 119-134.

2. Lawther JMS R, Banks WB (1995) Extraction, fractionation, and characterization of structural polysaccharides from wheat straw. J Agric Food Chem 43: 667-675.

3. Templeton DW, Scarlata CJ, Sluiter JB, Wolfrum EJ (2010) Compositional analysis of lignocellulosic feedstocks. 2. Method uncertainties. J Agric Food Chem 58: 9054-9062.

4. Sun L, Li J, Zhou K, Zhang M, Yang J, et al. (2013) Metabolomic analysis reveals metabolic disturbance in the cortex and hippocampus of subchronic MK-801 treated rats. PLoS one 8: e60598.

5. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84: 647-657.

6. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, et al. (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. Mol Syst Biol 7: 535.

7. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, et al. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6: 1290-1307.

8. Lun DS, Rockwell G, Guido NJ, Baym M, Kelner JA, et al. (2009) Large-scale identification of genetic design strategies using local search. Mol Syst Biol 5: 296.

9. Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y (2011) Synergy between (13) C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in *E. coli*. Metab Eng 13: 38-48.

10. Kim J, Reed JL (2012) RELATCH: Relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. Genome Biol 13: R78.

11. Zhu J, Shimizu K (2004) The effect of pfl gene knockout on the metabolism for optically pure D-lactate production by *Escherichia coli*. Appl Microbiol Biotechnol 64: 367-375.

12. Cai Y, Lai C, Li S, Liang Z, Zhu M, et al. (2011) Disruption of lactate dehydrogenase through homologous recombination to improve bioethanol production in *Thermoanaerobacterium aotearoense*. Enzyme Microb Technol 48: 155-161.

13. Chen J, Zhang W, Tan L, Wang Y, He G (2009) Optimization of metabolic pathways for bioconversion of lignocellulose to ethanol through genetic engineering. Biotechnol Adv 27: 593-598.

14. Liu S, Nichols NN, Dien BS, Cotta MA (2005) Metabolic engineering of a *Lactobacillus plantarum* double ldh knockout strain for enhanced ethanol production. Journal of industrial microbiology and biotechnology 33: 1-7.

15. Munjal N, Mattam AJ, Pramanik D, Srivastava PS, Yazdani SS (2012) Modulation of endogenous pathways enhances bioethanol yield and productivity in Escherichia coli. Microb Cell Fact 11: 145.

16. Vemuri GN, Eiteman MA, Altman E (2002) Effects of growth mode and pyruvate carboxylase on succinic acid production by metabolically engineered strains of *Escherichia coli*. Appl Environ Microbiol 68: 1715-1727.