Journal of Bioinformatics and Computational Biology

COFACTOR MODIFICATION ANALYSIS: A COMPUTATIONAL FRAMEWORK TO IDENTIFY COFACTOR SPECIFICITY ENGINEERING TARGETS FOR STRAIN IMPROVEMENT

MEIYAPPAN LAKSHMANAN*, BEVAN KAI-SHENG CHUNG[†], CHENGCHENG LIU[†], SEON-WON KIM[‡], DONG-YUP LEE*[,†,§]

**Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576*

*[†]2Bioprocessing Technology Institute, Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01, Centros, Singapore 138668*

[‡]Division of Applied Life Science, Gyeongsang National University, 900 Gajwa-dong, Jinju 660-701, Republic of Korea

[§]cheld@nus.edu.sg

Cofactors, such as NAD(H) and NADP(H), play important roles in energy transfer within the cells by providing the necessary redox carriers for a myriad of metabolic reactions, both anabolic and catabolic. Thus, it is crucial to establish the overall cellular redox balance for achieving the desired cellular physiology. Of several methods to manipulate the intracellular cofactor regeneration rates, altering the cofactor specificity of a particular enzyme is a promising one. However, the identification of relevant enzyme targets for such cofactor specificity engineering (CSE) is often very difficult and labor intensive. Therefore, it is necessary to develop more systematic approaches to find the cofactor engineering targets for strain improvement. Presented herein is a novel mathematical framework, cofactor modification analysis (CMA), developed based on the well-established constraints-based flux analysis, for the systematic identification of suitable CSE targets while exploring the global metabolic effects. The CMA algorithm was applied to *E. coli* using its genome-scale metabolic model, *i*JO1366, thereby identifying the growth-coupled cofactor engineering targets for overproducing four of its native products: acetate, formate, ethanol and lactate, and three non-native products: 1-butanol, 1,4-butanediol and 1,3-propanediol. Notably, among several target candidates for cofactor engineering, glyceraldehyde-3-phosphate dehydrogenase (GAPD) is the most promising enzyme; its cofactor modification enhanced both the desired product and biomass yields significantly. Finally, given the identified target, we further discussed potential mutational strategies for modifying cofactor specificity of GAPD in *E. coli* as suggested by *in silico* protein docking experiments.

Keywords: Metabolic engineering, cofactor specificity engineering (CSE), genome-scale metabolic model, flux balance analysis (FBA), cofactor modification analysis (CMA).

1. Introduction

Microbial synthesis of several value-added biochemicals such as biofuels, pharmaceuticals and fine chemicals is an attractive alternative to the chemical synthesis.[1,]

[§] Corresponding author

[2] However, low productivity and limited capability to synthesize some of the prominent compounds are the main challenges for its industrial applications. In this regard, recombinant technology now allows us to creatively engineer the innate metabolic capabilities of the organism for improving product yield, product range or substrate utilization by up-, down-regulating, inserting and/or deleting numerous metabolic genes at the same time.[3, 4] As a result, in the past decade, several works have shown to produce a myriad of products from both native and non-native substrates in microbial expression hosts, *Escherichia coli* and *Saccharomyces cerevisiae*.[5-9] However, despite these successes, many of the engineered microbes still do not exhibit the desired cellular physiology mainly due to the imbalances in redox cofactor regeneration rates, which are primarily caused by the insertion and/or deletion of certain gene(s). This can be well exemplified by the case of ethanol production from xylose in *S. cerevisiae*. Even though *S. cerevisiae* has been made to assimilate xylose via insertion of three genes, xylose reductase (*XYL1*), xylitol dehydrogenase (*XYL2*) and xylulose kinase (*XKS1*), from *Pichia stipitis*, the cellular growth and ethanol yields are still very low when compared to glucose due to the imbalances in cofactor regeneration rates across the inserted xylose assimilatory pathway; *XYL1* prefers NADPH whereas *XYL2* requires NAD^+.[10] When the cofactor specificity of any of these enzymes was altered from NADPH to NADH or NAD^+ to $NADP^+$, the resulting strains show an appreciable increase in ethanol productivity.[10-12] Similarly, the production of several important secondary metabolites such as terpenoids, aromatic ketones and flavonoids in *E. coli* have also shown drastic improvement by attenuating the cofactor production rates.[13, 14]

Recognizing the importance of manipulating the redox cofactor regeneration rates, numerous approaches have been proposed for increasing the availability of desired cofactor. Over-expression of *de novo* synthesis pathways[15] and the redirection of carbon-flow via specific pathways[16] are two methods for increasing the overall production rates. On the other hand, techniques such as over-expression of endogenous nucleotide transhydrogenase (*PntA*/*UdhA*) enzyme[17], which converts NAD(H) to NADP(H) and its reverse, or the alteration of cofactor specificity from NAD(H) to NADP(H) or vice-versa of a particular enzyme[10-13] can enhance the desired cofactor production rates at the expense of other. Among these different approaches, altering the cofactor specificity of a particular enzyme is an attractive method as it is quite stable from the evolutionary point of view. However, the identification of relevant enzyme targets for cofactor specificity engineering (CSE) in such a method is often very difficult and labor intensive. Therefore, development of more systematic approaches to find the appropriate CSE targets is highly required to achieve the best possible strain design.

Constraints-based flux analysis is a simple and extensible steady-state modeling approach that can accurately simulate the cellular phenotype in terms of metabolic fluxes with the information of reaction stoichiometry and mass-balance from the underlying metabolic models.[18, 19] Therefore, it has been successfully applied to various organisms for analyzing the cellular phenotypes systematically under different genetic and/or environmental perturbations at genome-scale.[20, 21] Furthermore, this framework has also

been extended into several computational algorithms over the past decade for the rational identification of gene targets that can be manipulated for strain improvement. The OptKnock[22] and OptStrain[23] were the first constraints-based algorithms developed to identify genes for knock-out (KO) and knock-in (KI) using a mixed-integer linear programming (MILP) formulation of the metabolic network. Subsequently, the OptKnock was extended into algorithms such as OptGene[24] and GDLS[25] for identifying the gene knockout targets within reasonable computational time based on genetic algorithm- and heuristic search-based frameworks, respectively. Later, OptKnock was also modified into another algorithm, OptReg[26], which can find even the up- and down-regulation genetic targets in addition to the deletion candidates. RobustKnock[27] has modified the OptKnock for identifying the possible genetic manipulations with a guaranteed minimal product synthesis rate by considering all the possible alternate optima. Most recently, another MILP-based problem, OptORF[28], was developed to find genetic manipulation targets that are consistent with known transcriptional regulatory rules. Moreover, most of these frameworks are now conveniently accessible via constraints-based modeling software tools such as COBRA toolbox and OptFlux.[29] However, despite the availability of several computational algorithms for identifying valid genetic targets to either delete, up- or down-regulate, still no method exists for investigating and improving the overall cellular redox cofactor balance, which may often be limiting case in achieving the desired strain design. Therefore, in this study, we introduce a novel mathematical framework, cofactor modification analysis (CMA), based on the well-established constraints-based flux analysis framework to systematically identify the best possible candidates among several reactions in the entire metabolic network whose CSE can lead to an improvement in both product yield and cellular growth. Subsequently, we also demonstrate the efficacy of this newly developed algorithm by applying it on the genome-scale metabolic network model of *E. coli*, predicting the best CSE targets that can lead to a significant improvement in both cellular growth and production rates of various native and non-native chemicals.

2. Methods

2.1. *Constraints-based flux analysis*

The constraints-based flux analysis is the most commonly used mathematical framework for analyzing cellular metabolisms.[19] It involves the constraining of cellular metabolic network by considering the balance of the metabolic fluxes (reactions) around each metabolite. Under steady state, the mass balance of intracellular metabolites in a network can be represented as:

$$\sum_j S_{ij} v_j = 0 \tag{1}$$

where S_{ij} refers to the stoichiometric coefficient of metabolite *i* involved in reaction *j*, v_j denotes to the flux or specific rate of metabolic reaction *j*. Since eq. 1 is under-determined most of the times in many metabolic networks, linear programming (LP) can be used to solve it with an objective function such as maximization of biomass or minimization of substrate uptake. Additionally, constraints on flux-capacity can be incorporated using the following constraint:

$$v_j^{\min} \le v_j \le v_j^{\max} \tag{2}$$

where $v_j^{\min}$ and $v_j^{\max}$ represent the lower and upper limits on the flux of reaction *j*.

Mathematically, the constraints-based flux analysis problem for biomass flux maximization can be represented as:

$$\begin{aligned} &\max v_{biomass} \qquad\qquad (P1)\\ &\text{s.t.} \quad \sum_j S_{ij} v_j = 0 \quad \forall \text{ metabolite } i \\ &\qquad v_j^{\min} \le v_j \le v_j^{\max} \quad \forall \text{ reaction } j \end{aligned}$$

2.2. *Cofactor modification analysis (CMA)*

To identify the best possible candidate for CSE in the entire metabolic network that can lead to a growth coupled product synthesis, first, we need to examine the consequences of replacing the cofactor specificity from NAD(H) to NADP(H) or vice-versa for all the reactions that involve them. However, it is quite tedious to iteratively examine the effect of CSE on each of the reaction in a metabolic network as there may be several reactions involving redox cofactors. Therefore, in order to remedy this issue, we formulate a MILP problem with two stoichiometric matrices where one representing the original metabolic state and the other signifying reactions with switched cofactors such that the formulation can select the best possible reaction with original or altered cofactor specificity for the stated objective. In this regard, we first modified the mass-balance constraints of the constraints- based framework (P1) by duplicating the stoichiometric matrix to account for reactions with engineered cofactor specificity:

$$\sum_j (S_{ij} v_j + S_{ij}^{cMod} v_j^{cMod}) = 0 \tag{3}$$

In eq. 3, S_{ij}^{cMod} refers to the modified stoichiometric coefficient of metabolite involved in reaction *j* and v_j^{cMod} denotes to the flux or specific rate of cofactor modified reaction *j*. Here, it should be noted that S_{ij}^{cMod} is as same as S_{ij}, except the reactions which involve NAD(H) or NADP(H) where it will be simply swapped such that $S_{(NAD)j}^{cMod} = S_{(NADP)j}$, $S_{(NADP)j}^{cMod} = S_{(NAD)j}$, $S_{(NADH)j}^{cMod} = S_{(NADPH)j}$ and $S_{(NADPH)j}^{cMod} = S_{(NADH)j}$. Additionally, a binary variable y_j^{cMod} is introduced in the flux capacity constraints (eq. 2) of v_j and v_j^{cMod} to ensure that there is no duplication of same reaction with both the cofactors. Eq. 2 can be appropriately modified as:

$$(1-y_j^{cMod}).v_j^{\min} \le v_j \le (1-y_j^{cMod}).v_j^{\max} \quad (4)$$

$$y_j^{cMod} v_j^{\min} \le v_j \le y_j^{cMod} v_j^{\max} \quad (5)$$

$$y_j^{cMod} = \{0,1\} \quad \forall \text{ reaction } j \quad (6)$$

where the binary variable y_j^{cMod} assume a value one if the corresponding reaction has its cofactor specificity modified and a value of zero if the reaction has its original cofactor specificity. Here, it should be noted that since all the reactions in a metabolic network does not involve cofactor, the selection of reaction *j* by the variable y_j^{cMod} in eq. 5 shall be restricted to a small subset of reactions, *C* via elimination of unwanted reactions (e.g. reactions which do not involve any cofactors, transport and exchange reactions) as follows:

$$y_j^{cMod} v_j^{\min} \le v_j \le y_j^{cMod} v_j^{\max} \quad \forall \text{ reaction } j \in C \quad (5a)$$

With all the above-mentioned changes to the constraints of original framework, CMA can be formulated as a bi-level MILP problem in which the formulation identifies the reactions that need to be modified for cofactor specificity so that the optimization of cellular objective indirectly leads to the overproduction of required biochemical. Mathematically, it can be represented as:

$$\max Z = v_{product}$$

$$\text{s.t.} \begin{bmatrix} \max v_{biomass} \\ \text{s.t.} \sum_j \left(S_{ij} v_j + S_{ij}^{cMod} v_j^{cMod} \right) = 0 \quad \forall \text{ metabolite } i \\ v_{biomass} \ge v_{biomass}^{\min} \\ \left(1 - y_j^{cMod}\right) \cdot v_j^{\min} \le v_j \le \left(1 - y_j^{cMod}\right) \cdot v_j^{\max} \\ y_j^{cMod} \cdot v_j^{\min} \le v_j^{cMod} \le y_j^{cMod} \cdot v_j^{\max} \quad \forall \text{ reaction } j \in C \\ y_j^{cMod} = \{0,1\} \quad \forall \text{ reaction } j \end{bmatrix}$$

$$\sum_j y_j^{cMod} \le k$$

where $v_{biomass}^{\min}$ is the minimum amount of biomass that needs to be produced and k is the number of allowed switches.

In general, bi-level optimization problems involving thousands of variables are computationally quite intensive to solve with their nested structure. Therefore, in order to solve the above mentioned bi-level optimization in an easier manner, it has to be re-formulated into a single-level MILP using the duality theory. For this purpose, first, the

inner problem is re-formulated in its dual form. The dual problem equivalent to the inner problem is as follows:

$$\begin{aligned}
\min \quad & -v_{biomass}^{\min}.\tau + \sum_j \left(\alpha_j v_j^{\max} \left(1 - y_j^{cMod}\right) - \beta_j v_j^{\min} \left(1 - y_j^{cMod}\right) \right) \\
& + \sum_j \left(\chi_j v_j^{\max} y_j^{cMod} - \delta_j v_j^{\min} y_j^{cMod} \right) \\
\text{s.t.} \quad & \sum_i \left(\lambda_i S_{i,biomass} \right) - \tau + \alpha_j - \beta_j = 1 \\
& \sum_i \left(\lambda_i S_{ij} \right) + \alpha_j - \beta_j = 0 \quad \forall \text{ reaction } j \\
& \sum_i \left(\lambda_i S_{ij}^{cMod} \right) + \chi_j - \delta_j = 0 \quad \forall \text{ reaction } j \in C \\
& y_j^{cMod} = \{0,1\} \quad \forall \text{ reaction } j
\end{aligned}$$

where λ_i is the dual variable associated with stoichiometric constraints in eq. 3, α_j and β_j are the dual variables associated with flux in eq.4, and are the dual variables associated with in eq. 5.

Once the dual form of the inner problem is formulated, it is further equated with the original inner primal problem according to duality theory; if the primal and dual problems are appropriately bounded then their objective values must be equal at one another at optimality. Using this condition, the bi-level formulation shown previously is easily transformed into a single-level optimization problem as:

$$\begin{aligned}
& \max Z = v_{product} \\
& v_{biomass} = -v_{biomass}^{\min}.\tau + \sum_j \left(\alpha_j v_j^{\max} \left(1 - y_j^{\text{cMod}}\right) - \beta_j v_j^{\min} \left(1 - y_j^{\text{cMod}}\right) \right) \\
& \qquad + \sum_j \left(\chi_j v_j^{\max} y_j^{\text{cMod}} - \delta_j v_j^{\min} y_j^{\text{cMod}} \right) \\
& \text{s.t.} \quad \sum_j \left(S_{ij} v_j + S_{ij}^{\text{cMod}} v_j^{\text{cMod}} \right) = 0 \quad \forall \text{ metabolite } i \\
& \qquad v_{biomass} \geq v_{biomass}^{\min} \\
& \qquad \left(1 - y_j^{\text{cMod}}\right) \cdot v_j^{\min} \leq v_j \leq \left(1 - y_j^{\text{cMod}}\right) \cdot v_j^{\max} \quad \forall \text{ reaction } j \\
& \qquad y_j^{\text{cMod}} \cdot v_j^{\min} \leq v_j^{\text{cMod}} \leq y_j^{\text{cMod}} \cdot v_j^{\max} \quad \forall \text{ reaction } j \in C \\
& \qquad \sum_i \left(\lambda_i S_{i,biomass} \right) - \tau + \alpha_j - \beta_j = 1 \\
& \qquad \sum_i \left(\lambda_i S_{ij} \right) + \alpha_j - \beta_j = 0 \quad \forall \text{ reaction } j \\
& \qquad \sum_i \left(\lambda_i S_{ij}^{\text{cMod}} \right) + \chi_j - \delta_j = 0 \quad \forall \text{ reaction } j \in C \\
& \qquad y_j^{\text{cMod}} = \{0,1\} \quad \forall \text{ reaction } j \\
& \qquad \sum_j y_j^{\text{cMod}} \leq k
\end{aligned}$$

For all simulations in this work, we fixed k=1 to avoid unrealistic simulations as cofactor specificity engineering requires substantial efforts.

2.3. *Cofactor flux-sum*

The redox cofactors of interest, NAD(H) and NADP(H), are involved in many reactions of *E. coli* metabolism. The resultant fluxes v_j from flux analysis will just indicate the rates of consumption/generation of these cofactors for each of the respective metabolic reactions but not the overall turnover rates. Therefore, we utilized the previously developed concept of ''flux-sum'' to quantify the metabolite turnover rates.[30, 31] Since the overall consumption and generation rates are equal under the steady-state assumption, the flux-sum of metabolite i can be formulated as $\Phi_i = 0.5\sum_j \left|S_{ij}v_j\right|$. Each $\left|S_{ij}v_j\right|$ term in this summation series gives us the absolute rate of consumption/generation of metabolite i due to reaction j and thus by halving the sum of these terms, we can obtain the overall turnover rate for metabolite i.

2.4. *In silico model settings*

The latest genome-scale metabolic model of *E. coli*, *i*JO1366, accounting for 1366 genes, 2583 reactions and 1805 metabolites in three cellular compartments, cytosol, periplasm and extracellular matrix was used throughout the work described herein.[32] 23 new reactions were added into this model to facilitate the synthesis of non-natural products in *E. coli*: 6 for 1-butanol, 11 for 1,4-butanediol and 5 for 1,3-propanediol (see additional file 1 for the list of reactions added). Furthermore, the *i*JO1366 model was also pre-processed to identify the possible candidates for CSE (See additional file 1 for list of candidates considered for CSE). For all simulations, a glucose uptake rate of 20 mmol g^{-1} DCW hr^{-1} was used and the non-growth associated maintenance energy was maintained at 3.15 mmol g^{-1} DCW hr^{-1}. The oxygen uptake rate was set to zero in all simulations corresponding to anaerobic conditions. The minimum level of target biomass in CMA simulations was set at an arbitrary value, i.e. 0.05 hr^{-1}. Additionally, the CAT, DHPTDNR, DHPTDNRN, FHL, SPODM, SPODMpp, SUCASPtpp, SUCFUMtpp, SUCMALtpp and SUCTARTtpp reactions were also constrained to zero for the reasons previously established.[32] All simulations were performed in GAMS IDE software using CPLEX solver.

3. Results

3.1. *Improving native products synthesis ability in E. coli*

We first conducted CMA on the *E. coli i*JO1366 model to analyze whether CSE of any particular enzyme can improve the cofactor (re-)generation rates, and thus, enhance the synthesis capability of four of its natural products: acetate, formate, ethanol and lactate. Table 1 summarizes the CMA identified CSE targets for each of them. Note that the abbreviations of enzymes/reactions presented in table 1 and elsewhere in the paper adhere

to the *i*JO1366 metabolic model definitions. From the analysis results, it could be observed that the production ability of all the four compounds analyzed can be improved significantly upon modifying the cofactor specificity of a certain enzyme (Table 1). Most importantly, the CSE design for acetate, ethanol and formate overproducing strain also indicated a significant increase in the biomass yield.

Table 1. CSE strategies predicted by CMA for overproducing the natural products in wild-type *E. coli*.

Product	CSE targets	Biomass Flux (hr^{-1})	Product Flux (mmol g^{-1} DCW hr^{-1})
Acetate	-	0.518	16.157
	GAPD	0.644	30.101
Formate	-	0.518	34.175
	GAPD	0.644	62.516
Ethanol	-	0.518	15.878
	PDH	0.536	19.014
Lactate	-	0.518	0
	ALCD2x/ACALD	0.426	33.218

The CMA framework identified glyceraldehyde-3-phosphate dehydrogenase (GAPD) as the best enzyme target for CSE to overproduce acetate and formate. In *E. coli*, all the fermentation products including acetate, formate and ethanol are produced from pyruvate through pyruvate formate lyase (PFL). The acetyl-coA resulting from PFL is further metabolized into ethanol and acetate via acetaldehyde dehydrogenase (ACALD) and phosphotransacetylase (PTAr), respectively (Fig. 1A). Therefore, in order to increase the production of acetate, the flux through PFL must be increased whereas the flux through ACALD must be reduced. In this regard, the GAPD-CSE mutant strain eliminates the carbon flow via ACALD by eradicating the need to regenerate the NADH lost in the GAPD as it is replaced with NADPH (Fig. 1B). Furthermore, the replacement of GAPD's cofactor also has a positive effect on biomass synthesis due to better NADP(H) regeneration rates. Here, it should be noted that although both NAD(H) and NADP(H) can serve as redox carriers, NAD(H) is primarily involved in the catabolic reactions whereas NADP(H) is required to synthesize most of the building blocks of the cell. Therefore, the replacement of NAD(H) with NADP(H) in GAPD enables *E. coli* to meet the biosynthetic requirements of NADP(H) more efficiently using glycolysis rather than the pentose phosphate pathway or the *UdhA*/*PntAB* enzymes (Fig. 1B). In general, these observations are in very good agreement with the experimental work by Martinez, et al. who have previously shown that altering GAPD's cofactor specificity from NAD(H) to NADP(H) can improve both biomass and acetate yields significantly.[13]

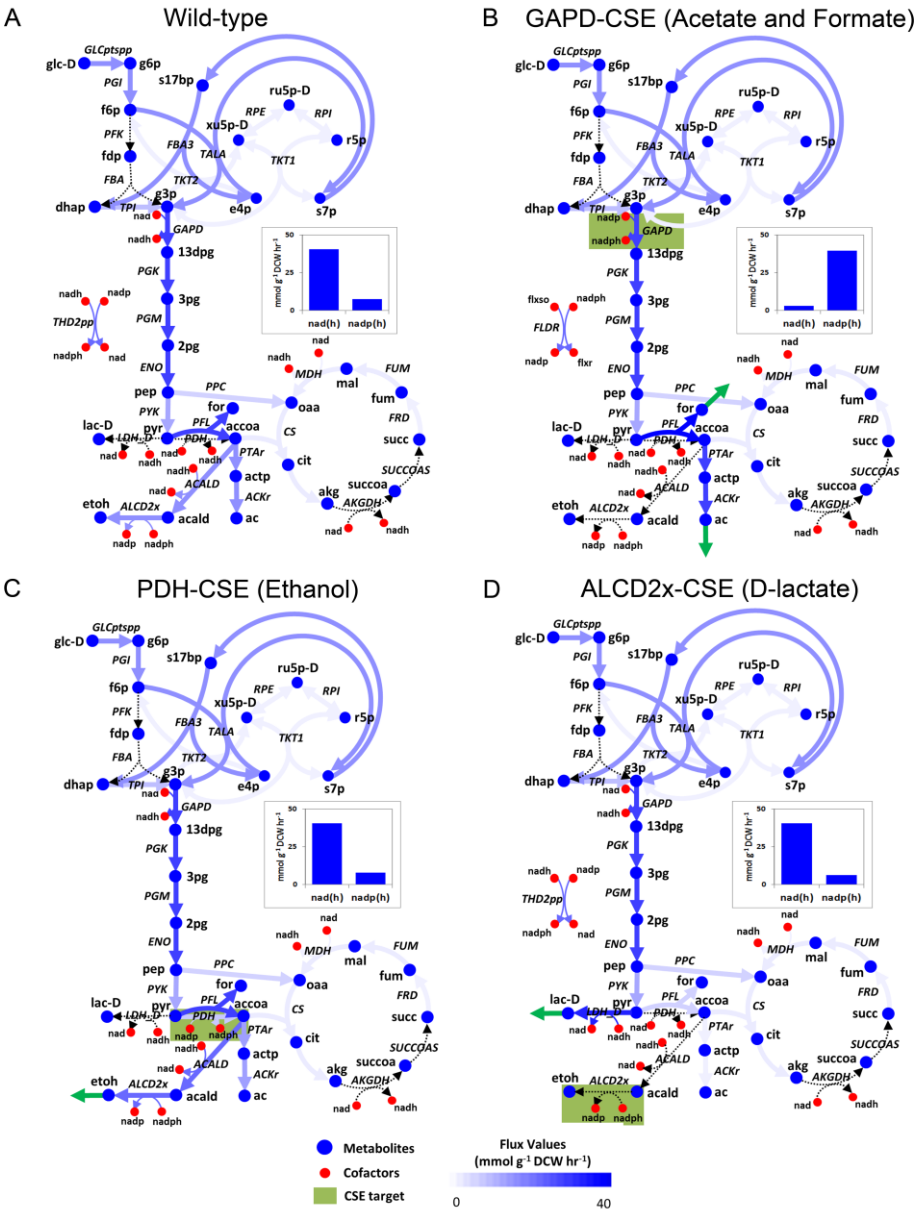

Fig. 1. *In silico* simulated flux maps of central metabolism and cofactor flux-sums in (A) wild-type, (B) GAPD-CSE, (C) PDH-CSE and (D) ALCD2x-CSE *E. coli*. The flux maps are generated using the results from constraints-based flux analysis. The color intensity of the edges between two nodes in the flux maps correspond to the amount of flux flow through them whereas the dotted lines represent the reactions with no active fluxes. The insert bar-graph in each figure represent the flux-sums of NAD(H) and NADP(H).

In case of ethanol overproducing design, pyruvate dehydrogenase (PDH) was identified as the best candidate for CSE by CMA. Unlike the wild-type *E. coli*, the PDH-CSE strain enables the ethanol overproduction by allowing an increased carbon flow from pyruvate to acetyl-coA by utilizing the PDH in addition to PFL (Fig. 1C). Interestingly, this observation is in very good agreement with the earlier report by Pharkya et al., who suggested an up-regulation of PDH is mandatory to increase ethanol production.[26] When PDH is up-regulated, it minimizes the flux through PFL, thus, reducing the amount of carbon flux wasted in formate synthesis. However, despite the up-regulation of PDH gene, wild-type *E. coli* may not produce acetyl-coA efficiently via PDH due to the NAD^+ requirements. In this regard, the PDH-CSE strain can successfully utilize the PDH as it can regenerate $NADP^+$ via several other NADP-dependent pathways. Furthermore, apart from improving the ethanol flux, the PDH-CSE strain also improve the biomass yield by allowing more flux towards the biomass synthesis pathways (Table 1).

For the design of lactate overproducing strain, CMA identified alcohol dehydrogenase (ALCD2x) or aldehyde dehydrogenase (ACALD) as the top candidates for CSE. Replacing the cofactor specificity of ALCD2x or ACALD allows *E. coli* to produce lactate as a homo-fermentation product with a theoretical maximum yield of 33 mmol g^{-1} DCW h^{-1}. Most importantly, unlike the ethanol or acetate over-producing designs where all the CSE targets offered a direct solution to improve the cofactor balancing in product producing pathways, this strategy perturbs the cofactor regeneration profile in the ethanol producing pathway, and thus, indirectly redirecting the flux towards lactate (Fig. 1D). Again, it should be noted that this strategy identified by CMA is comparable to the KO designed mutants for lactate overproducing *E. coli* where the ethanol production pathways are blocked by knocking out ALCD2x and/or ACALD.[22]

We further explored the product flux envelopes of ethanol, acetate, formate and lactate, in the wild-type and the CMA identified CSE mutants to analyze how the biomass flux and product fluxes are modified upon cofactor specificity alternation in *E. coli*. All the four flux envelopes in Figure 2 show that CSE mutants have an enlarged production envelope, suggesting that the CSE mutants possess better product synthesis capability than wild-type. Notably, the engineered strains also showed a significant improvement in the biomass fluxes except lactate case, highlighting that CSE benefits both product and biomass yield such that the specific productivity of these strains are superior than those of wild-type. Furthermore, the flux envelopes of acetate and formate overproducing mutants showed interesting characteristics where the maximum possible product fluxes of CSE-mutants where significantly higher than wild-type *E. coli* at zero biomass flux (Fig. 2B and 2C). To better understand this phenomenon, we performed constraints-based flux simulations by maximizing acetate production in the wild-type and GAPD-CSE *E. coli*. The wild-type strain produced mixed products consisting of 2 moles acetate, 0.5 moles of ethanol and 1 mole of formate per mole of glucose. On the other hand, the GAPD-CSE mutant strain with altered cofactor specificity from NAD(H) to NADP(H) synthesized significantly high amount of acetate (2.82 moles instead 2) along

with negligible quantities of formate and ethanol. Such substantial change in metabolic behavior is most likely because of the cofactor balancing in acetate pathways; the NADH produced by enzyme GAPD in wild-type can be regenerated only via the fermentative pathways, i.e. ACALD and ALCD2x, and thus producing significant amounts of ethanol. Conversely, in the GAPD-CSE mutant strain, the NADPH produced by GAPD can be regenerated via several other NADP-dependent reactions which are involved in biomass synthesis. In summary, these results indicate that CSE offers multiple advantages over the knockout design for strain design by improving both product yield and biomass yield even beyond its theoretical maximum via better cofactor regeneration.

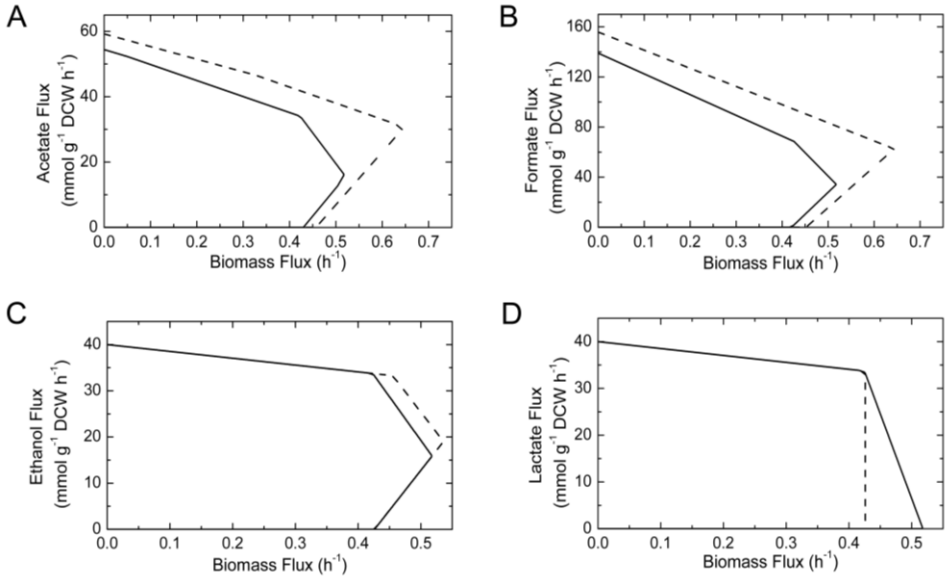

Fig. 2. Biomass vs. product flux envelopes for acetate (A), formate (B), ethanol (C) and lactate (D) in wild-type (solid line) and CSE-mutant (dashed line) *E. coli*. The flux envelopes were drawn by maximizing and minimizing the product synthesis fluxes while constraining the biomass flux at several pre-defined values as shown in the x-axis.

3.2. *Improving non-native products synthesis ability in E. coli*

Following our success in improving the growth coupled product synthesis for ethanol, acetate and formate via CSE in *E. coli*, we applied it to the synthetic pathway designed mutant strains that are specifically designed for producing non-native products such as 1-butanol, 1,4-butanediol and 1,3-propanediol. 1-butanol is an important biofuel, naturally produced from *Clostridum* species has significant potential in replacing gasoline as it possess a very similar energy content (27 MJ L^{-1}).[33] 1,3-propanediol and 1,4-butanediol are important building blocks for several polymers.[34] It has been earlier shown that *E. coli* can be successfully utilized to produce these compounds by manifesting their synthetic pathways into the host chassis where some of them have further optimized the engineered strain to produce these synthetic compounds by knocking out several genes.[8,

[35, 36] However, as explained earlier, the insertion and/or deletion of several reactions that involve a redox cofactor as it is in the 1,4-butanediol case where among the total 10 gene modifications (6 inserted and 4 deleted), 6 of them involves NAD(P), and thus could significantly alter the overall cofactor regeneration rates. Therefore, in this study we also analyzed the best possible CSE targets in the engineered *E. coli* using our newly developed CMA. To do so, first, the KO design for all the products where identified using the OptKnock algorithm as these non-native compounds cannot be synthesized just by the insertion of relevant synthetic reactions (See Table 2 for the KO design of each product used in this study). Secondly, we applied the CMA on these mutants as same as the wild-type after removing the reactions that are suggested to be knocked out from *i*JO1366 model.

Table 2. CSE strategies predicted by CMA for overproducing the non-native products in *E. coli*.

Product	KO targets	CSE targets	Biomass Flux (hr^{-1})	Product Flux (mmol g^{-1} DCW hr^{-1})
1-butanol	GLCptspp, ALCD2x,	-	0.461	8.166
	LDH_D, MDH	PDH	0.476	9.392
1,4-butanediol	ADHEr, LDH_D,	-	0.432	8.242
	FUM, ATPS4r	G6PDH2r	0.386	10.273
1,3-propanediol	TPI, MGSA, F6PA,	-	0.062	13.996
	G6PDH2r, ALCD2x	G3P2	0.076	15.045

Table 2 summarizes the CMA identified enzyme targets for CSE in the KO-designed non-native product synthesizing *E. coli* strains. PDH is identified as the best candidate for CSE to overproduce 1-butanol. Similar to ethanol, 1-butanol is also synthesized through alcohol and aldehyde dehydrogenases from acetyl-coA. Therefore, analogous to the ethanol overproducing strain design, the replacement of PDH's cofactor from NAD(H) to NADP(H) enables the conversion of more pyruvate into acetyl-coA, and thus facilitates the 1-butanol synthesis.

For the over-production of 1,4-butanediol, CMA identified the oxidative pentose phosphate pathway enzyme, glucose 6-phosphate dehydrogenase (G6PDH2r) as the best candidate. The replacement of G6PDH2r's cofactor from NADP(H) to NAD(H) allows the mutant strain to have more NADH pool that can be regenerated via the alcohol and/or dehydrogenases present in the downstream steps of 1,4-butanediol production.

In case of 1,3-propanediol overproducing strain, CMA identified glycerol-3-phosphate dehydrogenase (G3P2) as the best candidate for CSE where the product yield improved by 36% upon cofactor modification. G3P2 is an important enzyme present upstream in the 1,3-propanediol synthetic pathway, producing glycerol-3-phospahte from dihydroxyacetone phosphate (DHAP). In wild-type *E. coli*, this enzyme prefers NADP(H) as a cofactor, whereas in the CSE strain, it is replaced with NAD(H) to better regenerate the NADH required in the downstream step: 1,3-propanediol dehydrogenase.

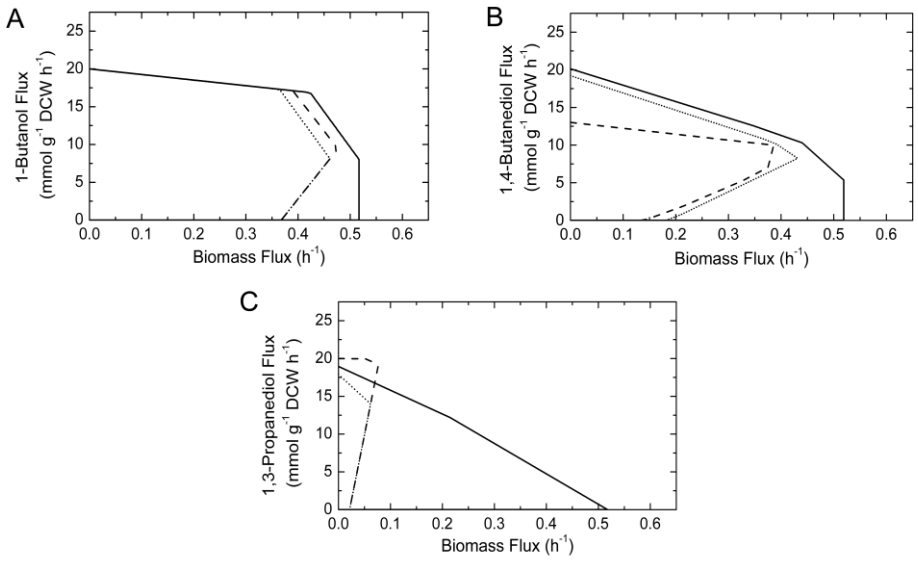

Fig. 3. Biomass vs. product flux envelopes for 1-butanol (A), 1,4-butanediol (B) and 1,3-propanediol (C) in wild-type (solid line), KO-designed mutant (dotted line), and KO+ CSE-mutant (dashed line) *E. coli*. The flux envelopes were drawn by maximizing and minimizing the product synthesis fluxes while constraining the biomass flux at several pre-defined values as shown in the x-axis.

In order to better understand the influence of CSE on product production and biomass growth in the KO designed mutants, we also examined the product production and biomass flux envelopes of all the three strains before and after CSE. Overall, the KO designed mutants show a significant improvement in the synthesis ability of all products analyzed (Fig. 3). Notably, as in the acetate and ethanol overproducing strains, two out of the three KO designed mutants also showed a significant increase in the biomass yield along with the increase in product yield after CSE (Table 2). Furthermore, similar to the acetate and formate flux envelopes, the 1,3-propanediol overproducing strain also showed an improvement in the maximum product flux at zero biomass fluxes (Fig. 3C). This indicates that this mutant strain possess better redox balance after CSE, and thus will be able to produce the desired compound even beyond the theoretical maximum of wild-type *E. coli*.

4. Discussion

Manipulation of cofactor regeneration rates via cofactor specificity alternation is an important technique to alleviate the existing redox imbalances across metabolic pathways. Systems biology now has the impetus to assist such methodology through the model-driven analyses as they offer a holistic approach to investigate and engineer cellular metabolism.[21] Toward this goal, we herein presented the algorithm, CMA, for identifying the CSE targets that can simultaneously improve both cellular growth and

product synthesis ability of microbial organisms. Furthermore, we also demonstrated its usefulness using *E. coli in silico* model, *i*JO1366, for producing several native and non-native products. Therefore, with the increasing recognition of constraints-based model reconstruction and analysis, this procedure can be simply adapted to any other organism of interest whose genome-scale metabolic model is readily available.

In the current work, we analyzed the cofactor modification targets for four native (acetate, ethanol, formate and lactate) and three non-native products (1-butanol, 1,4-butanediol and 1,3-propanediol). From the analysis, we observed some important attributes in cofactor balancing that can be generalized for *E. coli*. Firstly, in order to produce any alcoholic compounds via acetyl-coA, it is important to utilize PDH instead of PFL for the conversion of pyruvate to acetyl-coA; PFL wastes significant amounts of carbon flux in the form of formate. However, as mentioned earlier, PDH cannot be utilized in wild-type *E. coli* due to the cofactor requirements. Therefore, in order to alleviate this problem, replacing PDH's cofactor from NAD(H) to NADP(H) is an attractive solution. Secondly, by replacing the cofactor of GAPD from NAD(H) to NADP(H), glycolysis can facilitate the NADP(H) pathways without the involvement of pentose phosphate pathway or the transdehydrogenases, i.e. *PntAB* or *UdhA*. This observation is remarkable as improvement of NADP(H) regeneration rates has multiple advantages such as enhanced biomass synthesis and increase in the production capacity of several value-added products.

During cofactor engineering, it is important to note that following CMA, there are two different approaches available to achieve cofactor specificity alternation. In the first method, the enzyme target identified has to be engineered by mutating the protein sequences such that its cofactor specificity gets altered. The second method involves knockout of the original enzyme, i.e. enzyme identified from CMA, and the insertion of an equivalent enzyme from another organism with altered cofactor specificity using genetic engineering techniques. While comparing both the methods, the former method is preferable against the latter, because the expression levels of a foreign gene can be completely different when compared to the host gene expression levels.[37] Further, in order to modify the cofactor specificity of a particular enzyme via mutagenesis, the data available in literature for similar studies can be utilized for developing the mutation strategies and can be subsequently tested using the *in silico* protein docking experiments before testing it *in vivo*; it can predict the structure or structural ensemble of the intermolecular complex between NAD(H) or NADP(H) and the binding site of enzyme together with their binding energies. To better illustrate the utility of such protein docking experiments, we have successfully applied it to GAPD for altering its cofactor specificity from NAD(H) to NADP(H) using the mutation strategies suggested for *Bacillus stearothermophilus*[38] (see additional file 2 for detailed workflow and results).

Although CSE could be an attractive technique in metabolic engineering, one particular concern that could lie with it is the possibility of reduction in reaction rates. For example, if the enzyme possesses a particular reaction rate with NAD(H) as cofactor, the engineered protein with NADP(H) may have a reduced reaction rate owing to the

differences in enzyme-substrate binding timings. Since these issues cannot be assessed computationally with the current algorithm, further extensions to the presented framework can be expected by including the kinetic data of enzyme-substrate binding where a priority to the lower value conformations can be designed. Nevertheless, despite these limitations, CMA can still offer useful suggestions for strain improvement in a systematic manner as shown in this study.

Finally, the workflow and procedure presented here can be complemented with the existing strain design algorithms such as OptKnock or OptStrain since the KO designed mutant designs showed more improvements in product yield after CSE. Moreover, many of the results presented in this work suggested that altering the cofactor specificity will not only improve the product yield but also the biomass yield. Therefore, it would be a good practice to use the presented algorithm as the next immediate step to the existing strain improvement algorithms that can identify the genetic targets for deletion/insertion so that the newly designed strains will not possess any limitations in terms of cofactor imbalance that are actually caused by genetic deletions/insertions.

5. Conclusions

In the current work, we presented a novel mathematical framework, CMA, for systematically identifying the cofactor specificity engineering targets that can improve both cellular growth and product synthesis. The developed framework has been demonstrated in *E. coli* via its genome-scale model *i*JO1366, thereby identifying the growth-coupled CE targets for overproducing four native and three non-native products. In general, the results presented here show that CSE can improve not only the product yield but also the biomass yield significantly. Among the several candidates identified for CSE, GAPD is the most promising one as it increases the biomass yield by improving the overall NADP(H) regeneration rates. Therefore, in this work, we have also suggested a mutation strategy for altering cofactor specificity of GAPD in *E. coli* from NAD(H) to NADP(H) using the data obtained from literature, and subsequently tested it successfully with *in silico* protein docking experiments. The procedure presented here can be complemented with the existing strain design algorithms where the CMA could be the next immediate step after identifying the gene deletion/insertion targets for obtaining an overall improved strain design.

Acknowledgments

This work was supported by the National University of Singapore, Biomedical Research Council of A*STAR (Agency for Science, Technology and Research), Singapore, and a grant from the Next-Generation BioGreen 21 Program (SSAC, No. PJ009520), Rural Development Administration, Republic of Korea. The authors also thank Dr. Ahn Jungoh for helpful discussions.

Additional Files

Additional File 1 – List of reactions added for non-native product synthesis and candidate reactions for cofactor engineering in the *i*JO1366 model.

Additional File 2 – Detailed information on the computationally-driven enzyme redesign workflow for altering cofactor specificity from NADPH to NADH or vice-versa.

Additional files can be downloaded from: http://bioinfo.nus.edu.sg/paper/cma.

References

1. Alper, H, Stephanopoulos, G, Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential?, *Nature Reviews Microbiology*, **7**(10):715-723, 2009.
2. Chartrain, M, Salmon, PM, Robinson, DK, Buckland, BC, Metabolic engineering and directed evolution for the production of pharmaceuticals, *Curr Opin Biotechnol*, **11**(2):209-214, 2000.
3. Nielsen, LK, Metabolic Engineering: From retrofitting to green field, *Nat Chem Biol*, **7**(7):407-408, 2011.
4. Jarboe, LR, Zhang, XL, Wang, X, Moore, JC, Shanmugam, KT, Ingram, LO, Metabolic Engineering for Production of Biorenewable Fuels and Chemicals: Contributions of Synthetic Biology, *J Biomed Biotechnol*, 2010.
5. de Jong, B, Siewers, V, Nielsen, J, Systems biology of yeast: enabling technology for development of cell factories for production of advanced biofuels, *Curr Opin Biotechnol*, **23**(4):624-630, 2012.
6. Scalcinati, G, Partow, S, Siewers, V, Schalk, M, Daviet, L, Nielsen, J, Combined metabolic engineering of precursor and co-factor supply to increase alpha-santalene production by Saccharomyces cerevisiae, *Microb Cell Fact*, **11**:117, 2012.
7. Yang, JM, Nie, QJ, Ren, M, Feng, HR, Jiang, XL, Zheng, YN, Liu, M, Zhang, HB, Xian, M, Metabolic engineering of Escherichia coli for the biosynthesis of alpha-pinene, *Biotechnol Biofuels*, **6**, 2013.
8. Yim, H, Haselbeck, R, Niu, W, Pujol-Baxley, C, Burgard, A, Boldt, J, Khandurina, J, Trawick, JD, Osterhout, RE, Stephen, R, Estadilla, J, Teisan, S, Schreyer, HB, Andrae, S, Yang, TH, Lee, SY, Burk, MJ, Van Dien, S, Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol, *Nat Chem Biol*, **7**(7):445-452, 2011.
9. Jeffries, TW, Jin, YS, Metabolic engineering for improved fermentation of pentoses by yeasts, *Applied Microbiology and Biotechnology*, **63**(5):495-509, 2004.
10. Krahulec, S, Klimacek, M, Nidetzky, B, Engineering of a matched pair of xylose reductase and xylitol dehydrogenase for xylose fermentation by Saccharomyces cerevisiae, *Biotechnol J*, **4**(5):684-694, 2009.
11. Petschacher, B, Nidetzky, B, Altering the coenzyme preference of xylose reductase to favor utilization of NADH enhances ethanol yield from xylose in a

metabolically engineered strain of Saccharomyces cerevisiae, *Microb Cell Fact*, **7**:9, 2008.

12. Bengtsson, O, Hahn-Hagerdal, B, Gorwa-Grauslund, MF, Xylose reductase from Pichia stipitis with altered coenzyme preference improves ethanolic xylose fermentation by recombinant Saccharomyces cerevisiae, *Biotechnol Biofuels*, **2**:9, 2009.
13. Martinez, I, Zhu, J, Lin, H, Bennett, GN, San, KY, Replacing Escherichia coli NAD-dependent glyceraldehyde 3-phosphate dehydrogenase (GAPDH) with a NADP-dependent enzyme from Clostridium acetobutylicum facilitates NADPH dependent pathways, *Metabolic Engineering*, **10**(6):352-359, 2008.
14. Chemler, JA, Fowler, ZL, McHugh, KP, Koffas, MA, Improving NADPH availability for natural product biosynthesis in Escherichia coli by metabolic engineering, *Metabolic Engineering*, **12**(2):96-104, 2010.
15. San, KY, Bennett, GN, Berrios-Rivera, SJ, Vadali, RV, Yang, YT, Horton, E, Rudolph, FB, Sariyar, B, Blackwood, K, Metabolic engineering through cofactor manipulation and its effects on metabolic flux redistribution in Escherichia coli, *Metabolic Engineering*, **4**(2):182-192, 2002.
16. Kabir, MM, Shimizu, K, Fermentation characteristics and protein expression patterns in a recombinant Escherichia coli mutant lacking phosphoglucose isomerase for poly(3-hydroxybutyrate) production, *Applied Microbiology and Biotechnology*, **62**(2-3):244-255, 2003.
17. Sanchez, A, Andrews, J, Hussein, I, Bennett, G, San, K, Effect of Overexpression of a Soluble Pyridine Nucleotide Transhydrogenase (UdhA) on the Production of Poly(3-hydroxybutyrate) in *Escherichia coli*, *Biotechnology Progress*, **22**(2):420-425, 2006.
18. Oberhardt, MA, Chavali, AK, Papin, JA, Flux balance analysis: interrogating genome-scale metabolic networks, *Methods Mol Biol*, **500**:61-80, 2009.
19. Orth, JD, Thiele, I, Palsson, BO, What is flux balance analysis?, *Nat Biotechnol*, **28**(3):245-248, 2010.
20. Edwards, JS, Covert, M, Palsson, B, Metabolic modelling of microbes: the flux-balance approach, *Environ Microbiol*, **4**(3):133-140, 2002.
21. Lewis, NE, Nagarajan, H, Palsson, BO, Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods, *Nat Rev Microbiol*, **10**(4):291-305, 2012.
22. Burgard, AP, Pharkya, P, Maranas, CD, Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnol Bioeng*, **84**(6):647-657, 2003.
23. Pharkya, P, Burgard, AP, Maranas, CD, OptStrain: a computational framework for redesign of microbial production systems, *Genome Res*, **14**:2367-2376, 2004.
24. Patil, KR, Rocha, I, Forster, J, Nielsen, J, Evolutionary programming as a platform for in silico metabolic engineering, *BMC Bioinformatics*, **6**, 2005.
25. Lun, DS, Rockwell, G, Guido, NJ, Baym, M, Kelner, JA, Berger, B, Galagan, JE, Church, GM, Large-scale identification of genetic design strategies using local search, *Mol Syst Biol*, **5**:296, 2009.

26. Pharkya, P, Maranas, CD, An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems, *Metabolic Engineering*, **8**(1), 2006.
27. Tepper, N, Shlomi, T, Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways, *Bioinformatics*, **26**(4):536-543, 2010.
28. Kim, J, Reed, JL, Maravelias, CT, Large-scale bi-level strain design approaches and mixed-integer programming solution techniques, *PLoS One*, **6**(9):e24162, 2011.
29. Lakshmanan, M, Koh, G, Chung, BK, Lee, DY, Software applications for flux balance analysis, *Brief Bioinform*, Epub ahead of print, 2012.
30. Kim, PJ, Lee, DY, Kim, TY, Lee, KH, Jeong, H, Lee, SY, Park, S, Metabolite essentiality elucidates robustness of Escherichia coli metabolism, *Proceedings of the National Academy of Sciences of the United States of America*, **104**(34):13638-13642, 2007.
31. Chung, BK, Lee, DY, Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network, *BMC Syst Biol*, **3**:117, 2009.
32. Orth, JD, Conrad, TM, Na, J, Lerman, JA, Nam, H, Feist, AM, Palsson, BO, A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011, *Mol Syst Biol*, **7**:535, 2011.
33. Durre, P, Biobutanol: an attractive biofuel, *Biotechnol J*, **2**(12):1525-1534, 2007.
34. Lee, JW, Na, D, Park, JM, Lee, J, Choi, S, Lee, SY, Systems metabolic engineering of microorganisms for natural and non-natural chemicals, *Nat Chem Biol*, **8**(6):536-546, 2012.
35. Atsumi, S, Cann, AF, Connor, MR, Shen, CR, Smith, KM, Brynildsen, MP, Chou, KJY, Hanai, T, Liao, JC, Metabolic engineering of Escherichia coli for 1-butanol production, *Metabolic Engineering*, **10**(6):305-311, 2008.
36. Wang, FH, Qu, HJ, Zhang, DW, Tian, PF, Tan, TW, Production of 1,3-propanediol from glycerol by recombinant E-coli using incompatible plasmids system, *Mol Biotechnol*, **37**(2):112-119, 2007.
37. Chung, BK, Lee, DY, Computational codon optimization of synthetic gene for protein expression, *BMC Syst Biol*, **6**:134, 2012.
38. Didierjean, C, Rahuel-Clermont, S, Vitoux, B, Dideberg, O, Branlant, G, Aubry, AA, A crystallographic comparison between mutated glyceraldehyde-3-phosphate dehydrogenases from *Bacillus stearothermophilus* complexed with either NAD+ or NADP+, *J Mol Biol*, **268**(4):739-759, 1997.

Meiyappan Lakshmanan is a Ph.D. student in Systems Biology and Bioinformatics Group, Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore. He obtained his B.Tech. degree in Chemical Engineering from Anna University, India in 2007. His research interests include systems biological analysis of plant metabolism and development of computational frameworks for strain improvement.

Bevan K.S. Chung is a Research Scientist in the Bioinformatics group of Bioprocessing Technology Institute, Agency for Science Technology and Research of Singapore, Singapore. He received his B.Eng. in Chemical Engineering from National University of Singapore (NUS) in 2008 and obtained his Ph.D. in Integrative Sciences and Engineering from the NUS Graduate School in December 2012. His main research interests are in systems biology and synthetic biology, especially in the areas of metabolic network modelling and synthetic gene design.

Chengcheng Liu graduated with a B.Sc. in Computational Biology from National University of Singapore in 2008. She obtained her Ph.D. in Computation and Systems Biology with concentration in protein structure prediction from Singapore-MIT Alliance in February 2013. She is currently a research scientist in bioinformatics group of Bioprocessing Technology Institute, Agency for Science Technology and Research of Singapore. Her research interests include conformational sampling of intrinsically disordered proteins, molecular modeling, simulation and docking as well as statistical analysis and data mining.

Seon-Won Kim is a professor in the Division of Applied Life Science, Gyeongsang National University, Republic of Korea. He received his B.S. in Food Engineering degree from Yonsei University, Republic of Korea. He obtained his Ph.D. in Biological Science from Korea Advanced Institute of Science and Technology, Republic of Korea in 1996. His research interests include pathway engineering and optimization of terpenoids, genetic tool development for terpenoids and high throughput screening for novel antibiotics, anti-malaria, and hypercholesterolemia drugs based on microbial cells.

Dong-Yup Lee is an assistant professor in the Department of Chemical and Biomolecular Engineering, National University of Singapore, Singapore. He is also a leader of Bioinformatics group in Bioprocessing Technology Institute, Agency for Science, Technology and Research, Singapore. He received his B.S. (Chemical Engineering) from Yonsei University, Republic of Korea in 1998, and obtained his Ph.D. in Systems Biology from Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, Republic of Korea in 2004. His research interests include computational analysis and engineering of biological systems, bioinformatics analysis of multi-omics profiles and bioinformatics platform and tool development.