Computational Systems Biology Seminar report

System Level Application of Metabolic Engineering

------ *In silico* Microbial Strain Redesign

Hao Wang

Helsinki 9.4.2007

UNIVERSITY OF HELSINKI

Department of Applied Chemistry and Microbiology

1 Introduction

At present, there are about 600 microbial genomes had been sequenced, and 288 undergoing or nearly finished genome projects are processing around the world (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). This escalating condition has turned to be the driving force of new bioinformatics methods and algorisms served for the data-mining requirements in post-genomic era. With more accurate genome annotation and more complete pathway information, promising achievements had been made on *in silico* metabolic network reconstruction (Van Dien and Lidstrom 2002), and noted that one of the most comprehensive stoichiometric models built upon the genome of *E. coli* had been available (Reed and Palsson 2003). Based on these genome-scale metabolic networks, a few bio-engineering applications had be proposed, and one new emerging study is aiming at computationally predict the probability and possibility of overproduction of particular chemical and biochemical compounds, diversely range from industrial interests to environmental usages, through metabolic modification. The outcomes of some positive experimental verification of these microorganism redesigns bring light on their potential engineering application in the future.

2 OptStrain

OptStrain (Pharkya et al. 2003) is newly developed hierarchical computational framework, which could be used as a statistical model for microorganism strain redesign through recombination of available digitalized pathways, its aim is to not only find a proper set of non-native pathways whose combining could lead to desired extrinsic compound production but also figure out possible ways of pathway deletion using OptKnock (Burgard et al. 2003; Pharkya et al. 2003) so as to obtain the maximum output of target compound. The workflow of OptStrain is illustrated in **Figure 1**. Its procedures include four steps described as follow.

2.1 Curation of the database

Firstly, a Universal database was constructed by collecting curated biotransformation data form public domains, such as KEGG (Kanehisa et al. 2004) and *in silico E. coli* model (Reed and Palsson 2003), which laid very important foundation for later studies. Then those downloaded reactions were checked by pre-programmed scripts for automatically parsing and

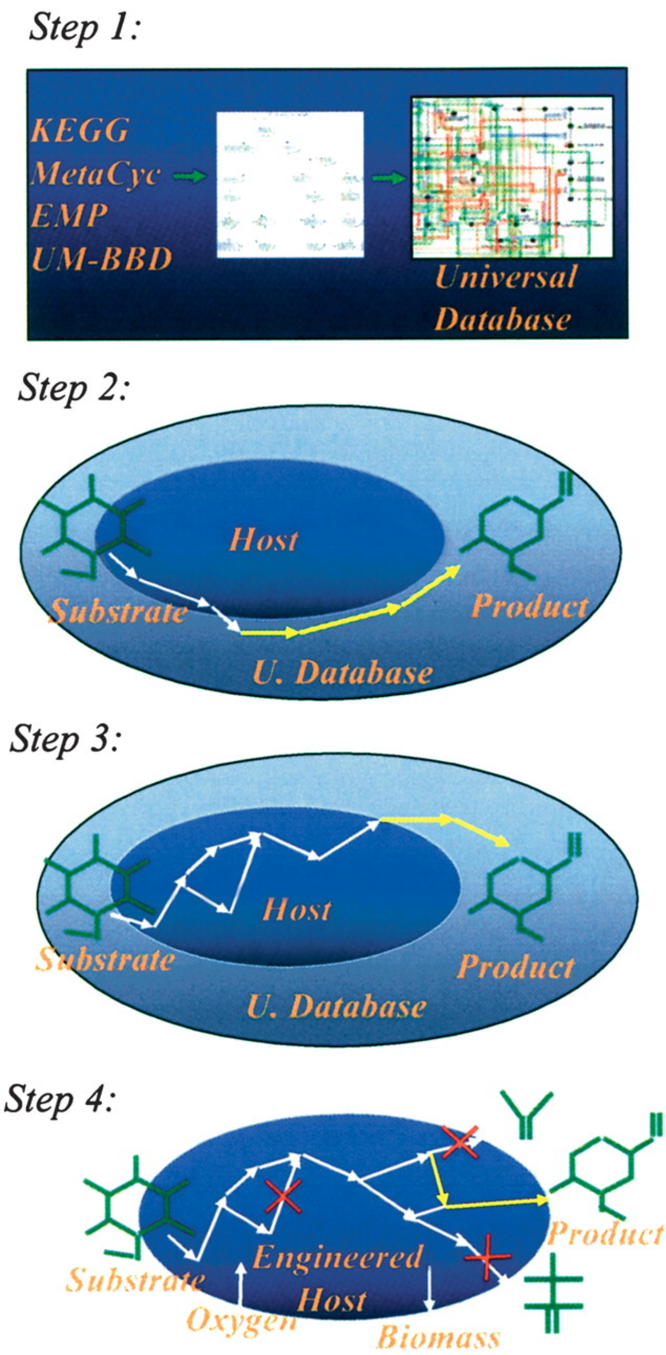


Figure 1. Flowchart of OptStrain procedure. Step 1 Curation of database(s) of reactions to compile the Universal database, only elementally balanced reactions included. Step 2 identifies a maximum-yield path enabling the desired biotransformations from a substrate to product, the white arrows represent native reactions of the host and the yellow arrows denote non-native reactions. Step 3 minimizes the reliance on non-native reactions, and Step 4 incorporates the non-native functionalities into the microbial host's stoichiometric model and applies the OptKnock framework to identify and eliminate reactions competing with the targeted product. (Pharkya et al., 2003).

elementally balancing check before entering the database. The metabolite set *N* was comprised of ～4800 metabolites, and the reaction set *M* consisted of >5700 reactions. Moreover, compounds with an unspecified number of repeat units [e.g., trans-2-Enoyl-CoA represented by C25H39N7O17P3S(CH2)n] or unspecified alkyl groups R in their chemical formulas are removed. A few Perl (Brown 1999) scripts were developed to automatically process this task routinely.

2.2 Determination of the maximum yield

With the elementally balanced datasets of functionalities obtained from previous step, theoretical maximum yields of the target product **P** are calculated by maximizing the sum of all reaction fluxes producing minus those consuming the target metabolite, weighted by the stoichiometric coefficient of the target metabolite in all reactions within the database, which comprised of a set *N* = {1,..., N} of metabolites (~4800) and a set *M* = {1,..., M} of reactions (>5700). The uptake rate of substrate is set to pre-defined value (default is 1 unit of substrate). This calculation could be simulated as a linear programming (LP) problem as following equations, which now could be analyzed in a large scale and solved very well.

$$\begin{aligned} &\underset{v_j}{\mathbf{Max}} && MW_i \cdot \sum_{j=1}^{M} S_{ij} v_j, \quad i = \mathbf{P} && \\ &\textbf{subject to} && \sum_{j=1}^{M} S_{ij} v_j \geq 0, \qquad \forall\, i \in \mathrm{N},\, i \notin \Re && (1) \\ & && \sum_{i \in \Re} \left(MW_i \cdot \sum_{j=1}^{M} S_{ij} v_j \right) = -1 && (2) \end{aligned}$$

where *MWi* is the molecular weight of metabolite *i*, v_j is the molar flux of reaction *j* (can either be irreversible or reversible), and *Sij* is the stoichiometric coefficient of metabolite i in reaction j. The inequality in constraint (1) allows only for secretion and prevents the uptake of all metabolites in the network other than the substrates in *R*. Constraint (2) scales the results for a total substrate uptake flux of one unit of mass. Solutions of these equations could be the initial values used in later calculations.

2.3 Identification of the minimum number of non-native reactions for a host organism

One key purpose of strain engineering is to endow microbial hosts with function of producing additional compounds with economical or environmental interests. In this step, OptStrain is trying to determine the minimum number of non-native functionalities or reactions, that are absent in the examined microbial host's metabolic model and needed to be combined into the original network. The simulation of this step could be achieved through adding heterologous fluxes as following equations:

$$\mathbf{Min}_{v_j, y_j} \qquad \sum_{j \in M_{non\text{-}native}} y_j$$

$$\mathbf{subject\ to} \quad \sum_{j=1}^{M} S_{ij} v_j \geq 0, \qquad \forall\, i \in \mathrm{N},\ i \notin \Re \qquad (1)$$

$$\sum_{i \in \Re} \left(MW_i \cdot \sum_{j=1}^{M} S_{ij} v_j \right) = -1, \qquad (2)$$

$$MW_i \cdot \sum_{j=1}^{M} S_{ij} v_j \geq Yield^{t\arg et}, \quad i = \mathbf{P} \qquad (3)$$

$$v_j \leq v_j^{\max} \cdot y_j, \qquad \forall\, j \in \mathrm{M}_{non\text{-}native} \qquad (4)$$

$$v_j \geq v_j^{\min} \cdot y_j, \qquad \forall\, j \in \mathrm{M}_{non\text{-}native} \qquad (5)$$

$$y_j \in \{0, 1\}, \qquad \forall\, j \in \mathrm{M}_{non\text{-}native} \qquad (6)$$

The set $M_{non\text{-}native}$ comprises the non-native reactions for the examined host, and (1) and (2) are identical to those in step 2. (3) ensures that the product yield meets the maximum theoretical yield calculated in Step 2. The binary variable y_j is a set of binary values (1 or 0) for turning the responding reactions on or off, and this constraint is imposed only on reactions associated with genes heterologous to the specified production host. The parameters v_j^{min} and v_j^{max} can either be assumed values or calculated by minimizing and maximizing every reaction flux v_j subject to stoichiometric constraints. The addition of constraints derived from binary controls turned such a case into solving a simulation problem of Mixed Integer Linear Programming (MILP) model; its solution is a set of non-native pathways, which are obtained through balancing between minimizing numbers of heterologous genes and maximizing theoretical product yield.

2.4 OptKnock: pruning of the host organism's stoichiometric model in order to achieve highest production of compounds (bi-level computational framework)

Even addition of optimal set(s) of non-native biotransformations would lead to production of desired compounds; it would not guarantee an overproduction of this compound. In order to maximize this output, another computational framework, OptKnock (Burgard et al. 2003;

Pharkya et al. 2003) was used to modulate the flux distribution toward specific orientation by containing other competing reactions and byproducts.

This framework is aim to optimize a bilevel problem, that is balancing between two competing optimal strategists (cellular objective and chemical production). As one promising achievement from the same group of OptStrain, OptKnock also utilize genome-scale metabolic models as stoichiometric basis and flux balance analysis (FBA), then maximization of biomass formation (cellular objective), optimization of target chemical or biochemical compounds (chemical production), and candidate gene deletions are set as additional constraints in looking for a likely flux distribution. Optimal solutions after overall considering all alternative solutions would give out suggested genes for knocking out. It should be noted that OptKnock's results from testing a wide range of products (succinate, lactate, 1,3-propanediol, glutamate, alanine, hydrogen, vanillin) showed good congruence with the experimental data published in literatures.

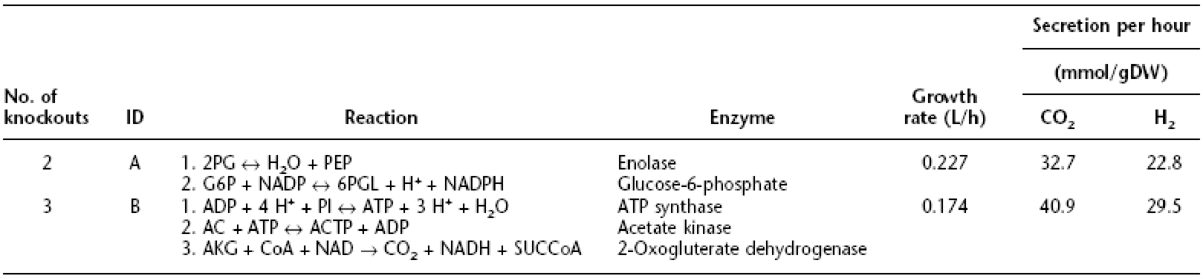
3. Case studies results from OptStrain

To verify the effectiveness of OptStrain, two diversified compounds, hydrogen and vanillin, are selected as test examples of *in silico* strain design. In the case of hydrogen production, three different hosts are used, and a few substrates are screened. As for the case of vanillin, minimum set of non-native reactions required for compound production was identified, and this highlights the outcome of OptStrain. Nevertheless, the suggestions of reshaping strain made by this framework look like only are *in silico* predictions, and have little support form existing experimental data. Hence, the results of implementation of those suggestions are really expected, not only for model validation but also for preliminary engineering trial.

3.1 Hydrogen production

The result of LP formulation (step 2) among many substrates shows that methanol turned to be the most efficient 'raw material' for hydrogen production, and glucose also selected for further test owning to its economic availability. Three strains are used for analysis: 1). *E. coli*, 2). *C. acetobutylicum*, 3). *M. extorquens*.

3.1.1 glucose substrate in *E. coli*



No. of knockouts	ID	Reaction	Enzyme	Growth rate (L/h)	Secretion per hour (mmol/gDW)	
					CO_2	H_2
2	A	1. 2PG ↔ H_2O + PEP 2. G6P + NADP ↔ 6PGL + H^+ + NADPH	Enolase Glucose-6-phosphate	0.227	32.7	22.8
3	B	1. ADP + 4 H^+ + PI ↔ ATP + 3 H^+ + H_2O 2. AC + ATP ↔ ACTP + ADP 3. AKG + CoA + NAD → CO_2 + NADH + SUCCoA	ATP synthase Acetate kinase 2-Oxogluterate dehydrogenase	0.174	40.9	29.5

Table 1. Deletion mutants for enhanced hydrogen production in E. coli

For this case, OptStrain doesn't find any non-native reaction needed for hydrogen producing. But OptKnock (step 4) identified two sets of reactions (shown in Table 1), whose deletions would decrease competition with hydrogen production at most degrees.

3.1.2 glucose substrate in *C. acetobutylicum*

C. acetobutylicum had been intensively studied as a model organism of hydrogen production (Chin et al. 2003), and OptStrain also does not find any non-native reactions are needed. Importantly, the output of OptKnock's two knockout candidates is highly congruent with experimental data.

3.1.2 methanol substrate in *M. extorquens AM1*

M. extorquens AM1 could live on methanol as solely carbon source, therefore, also been thoroughly studied (Van Dien et al. 2003; Chistoserdova et al. 2004), even a stoichiometric model of central metabolism had been established (Van Dien and Lidstrom 2002). Single non-native reaction was identified by OptStrain to enable hydrogen ability on this host; this might be the reason why *M. extorquens* AM1 can not produce hydrogen originally.

3.2 Vanillin production

There is constant demand for overproduction of vanillin in case of its limited natural yields and economical value. So initially OptStrain was used to figure out the maximum theoretical yields and minimum set of non-native reactions required for strain redesign of *E. coli*. Based on a maximum theoretical yield of glucose at 0.63 g/g, OptKnock was used again for overproduction optimization in such an augmented genome-scale model with three non-native bioconversions, and many alternative pathways were found that their deletions could meet the criteria of maximum vanillin production. Finally the author look like choose the three cases which have experimental data support to be further discussed in the paper. They are one deletion (removal of acetaldehyde dehydrogenase, EC 1.2.1.10), double deletion (with

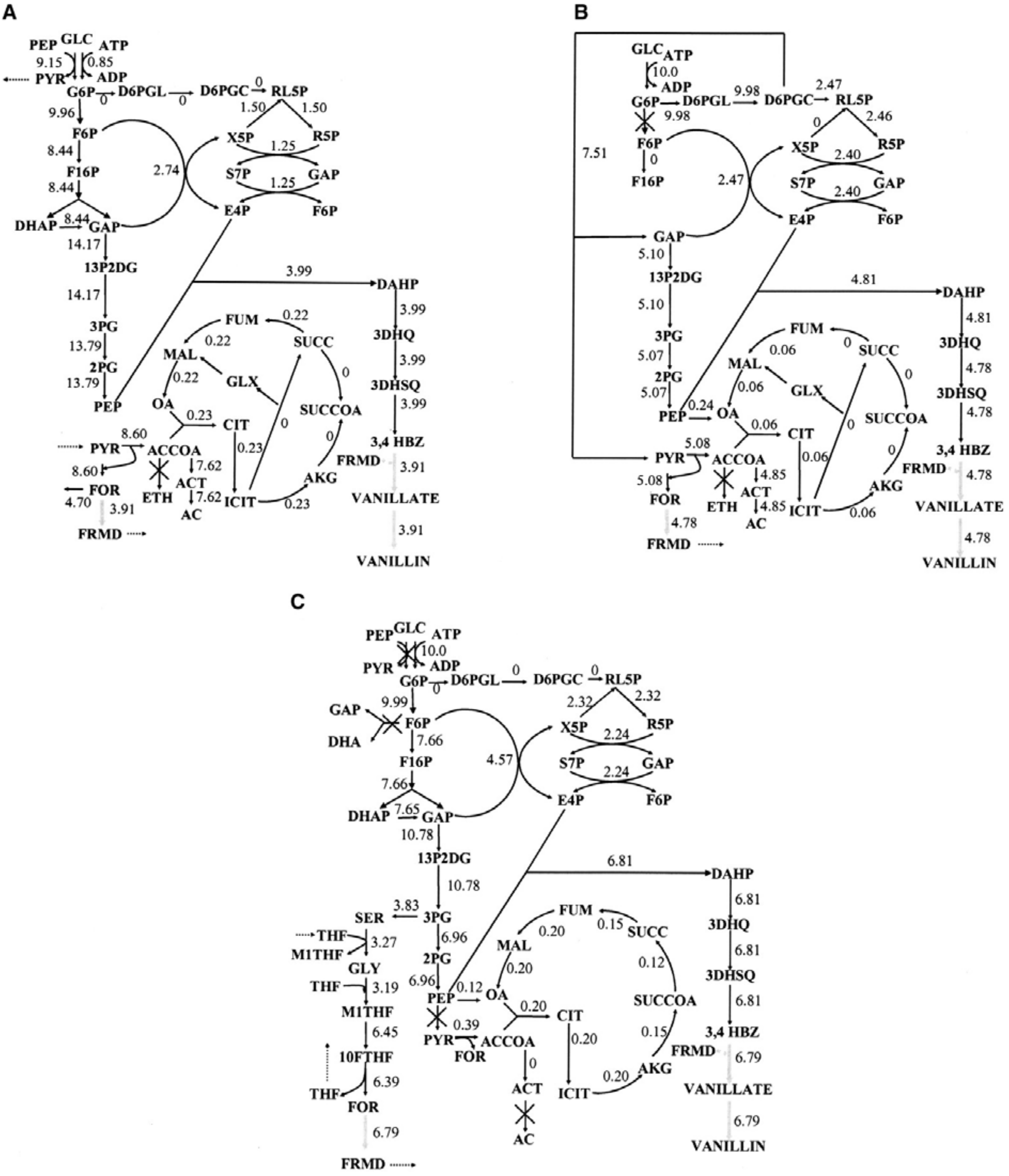


Figure 2. Calculated flux distributions at the maximum growth rates in the (A) one, (B) two, and (C) four deletion *E. coli* mutant networks for overproducing vanillin. Non-native reactions are denoted by the thicker gray arrows. A basic glucose uptake rate of 10 mmol/gDW per hour was assumed.

additional removal of glucose -6-phophate isomerase EC 5.3.1.9), and four-reaction deletion (with deletion of acetate kinase EC 2.7.2.1, pyruvate kinase EC 2.7.1.40, the PTS transport mechanism, and fructose 6-phosphate aldolase). These modifications on flux distribution network (shown in Figure 2) presume higher vanillin production level under all culture conditions.

4 OptReg

As an upgraded version of OptKnock, OptReg (Pharkya and Maranas, 2006) was developed to maximize the production of desired compound through modulation on pathways by up- or down-regulating reactions besides knocking them out. However, since this extended computational framework take into account much complex conditions than the former one did, its computational complexity is magnified.

OptReg still applied Mixed Integer Linear Programming (MILP) simulation to solve the optimization problem in this framework. In this case, regulation strength parameter *C* (Figure 3) was introduced into the framework to simulate the conditions of up- and down-regulation, so there are actually three states (up/down-regulating and knocking out) for any reaction in the model for optimal secreening. In present paper (Pharkya and Maranas, 2006) all results were calculated while fixing the value of *C* as 0.5, however they might be completely different if *C* were assigned to another value; and this might be a potential drawback of OptReg.

It is still unrealistic to check those predictions by present experimental techniques, therefore, another computational criterion for anticipating microbial systems, MOMA (Segre et al., 2002), was applied to evaluate OptReg’s performance. Anyway, even the authors agree that a similar conclusion from both statistical methods could not make confirmative remarks on OptReg.

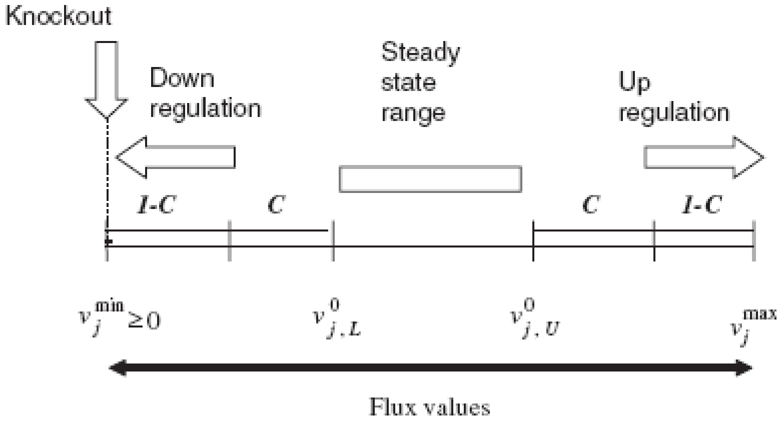


Figure 3: A pictorial overview of the definitions of up/down regulations and deletions. 0⩽*C* ⩽1, The lower bound v_j min for a flux j may be greater than zero if it is required for biomass formation, and the reaction cannot be knocked out.

5 Calculation and Software

All the optimization problems above were solved using CPLEX 7.0 accessed via the GAMS (Brooke et al. 1998) modeling environment on an IBM RS6000-270 workstation.

6 Conclusions

Such a series of frameworks (OptKnock, OptStrain and OptReg) highlighted continuous efforts laid on the advancement of metabolic engineering modification from this group in Pennsylvania State University (PSU). Their studies would be essentially helpful in improving the application of biotechnology in industry innovations, and also paved solid basis for more comprehensive understanding on this leading edge field as well. Admittedly, some of the examples discussed in their papers are in good agreement with experimental data in published literatures. But this could not provide us enough confidence on their performance due to only a limited number of experimental test results could be available until now. As we known, The stoichiometry models used in metabolic engineering study could have alternative solutions because only extracellular parameters could be accurately measured, thus intracellular information is required in order to obtain better predictions from present highly redundant models. Consequently, more accurate data from intracellular conditions would be tremendous helpful in polishing these frameworks, and new techniques (e.g. NMR, mass spectrometry, and isotopic atoms labeling *etc.*) that could intensively measure those intracellular parameters would provide more precise and stringent constraints over those optimization models, therefore, could make significant advancement for the frameworks.

7 References

Brooke, A., Kendrick, D., Meeraus, A., and Raman, R. 1998. *GAMS: A user's guide*. GAMS Development Corp., Washington, D.C.

Brown, M. 1999. *Perl programmer's reference.* Osborne/McGraw-Hill, Berkeley, CA.

Burgard, A., Pharkya, P. and Maranas, C. 2003. OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization, *Biotech. Bioeng.*, **84**: 647-657.

Chin, H.L., Chen, Z.S., and Chou, C.P. 2003. Fedbatch operation using *Clostridium acetobutylicum* suspension culture as biocatalyst for enhancing hydrogen production. *Biotechnol. Prog.* **19:** 383–388.

Chistoserdova, L., Laukel, M., Portais, J.C., Vorholt, J.A., and Lidstrom, M.E. 2004. Multiple formate dehydrogenase enzymes in the facultative methylotroph *Methylobacterium extorquens* AM1 are dispensable for growth on methanol. *J. Bacteriol.* **186:** 22–28.

Fischer, E., Zamboni, N., Sauer, U., 2004. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Anal. Biochem.* **325**: 308–316.

Ignizio, J.P., Cavalier, T.M., 1994. *Linear programming*. Prentice Hall, Englewood Cliffs, N.J.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. 2004. The KEGG resource for depichering the genome, *Nucleic Acid res.* **32**: D277-D280.

Pharkya, P., Burgard, A.P., Maranas, C.D. 2003. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol. Bioeng.* **84**: 887–899.

Pharkya, P., Burgard, A.P., Maranas, C.D., 2004. OptSrain: a computational framework for redesign of microbial production networks. *Genome Res.* **14**: 2367–2376.

Pharkya, P., Maranas, C.D., 2006. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic Engineering* **8**: 1–13.

Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4:** R54.

Segre, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**: 15112–15117.

Van Dien, S.J. and Lidstrom, M.E. 2002. Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C(3) and C(4) metabolism. *Biotechnol. Bioeng.* **78:** 296–312.

Van Dien, S.J., Strovas, T., and Lidstrom, M.E. 2003. Quantification of central metabolic fluxes in the facultative methylotroph *Methylobacterium extorquens* AM1 using 13C-label tracing and mass spectrometry. *Biotechnol. Bioeng.* **84:** 45–55.