CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks

Katrin Bohl[1,2,†], Luís F. de Figueiredo[1,†], Oliver Hädicke[3], Steffen Klamt[3], Christian Kost[2], Stefan Schuster[1] and Christoph Kaleta[1,*]
[1] Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany,
[2] Department of Bioorganic Chemistry, Max Planck Institute for Chemical Ecology, D-07745 Jena, Germany,
[3] Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstrasse 1, D-39106 Magdeburg, Germany
*Corresponding author e-mail: christoph.kaleta@uni-jena.de
†Both authors contributed equally

Abstract: Metabolic engineering aims to improve the production of desired biochemicals and proteins in organisms and therefore, plays a central role in Biotechnology. However, the design of overproducing strains is not straightforward due to the complexity of metabolic and regulatory networks. Thus, theoretical tools supporting the design of such strains have been developed. One particular method, CASOP, uses the set of elementary flux modes (EFMs) of a reaction network to propose strategies for the overproduction of a target compound. The advantage of CASOP over other approaches is that it does not consider a single specific flux distribution within the network but the whole set of possible flux distributions represented by the EFMs of the network. Moreover, its application results not only in the identification of candidate loci that can be knocked out, but additionally proposes overexpression candidates. However, the utilization of CASOP was restricted to small and medium scale metabolic networks so far, since the entire set of EFMs cannot be enumerated in such networks. This work presents an approach that allows to use CASOP even in genome-scale networks. This approach is based on an estimation of the score utilized in CASOP through a sample of EFMs within a genome-scale network. Using EFMs from the genome-scale metabolic network gives a more reliable picture of the metabolic capabilities of an organism required for the design of overproducing strains. We applied our new method to identify strategies for the overproduction of succinate and histidine in *Escherichia coli*. The succinate case study, in particular, proposes engineering targets which resemble known strategies already applied in *E. coli*. **Availability:** Source code and an executable are available upon request.

1 Introduction

Using microorganisms to overproduce certain metabolites and proteins is the central objective of metabolic engineering [Lee09]. While many applications consider the improvement of the production of native compounds, there is also an increasing number of attempts

in which entire heterologous pathways have been engineered [NKL10, AS09]. Small molecule compounds whose production have been engineered span from alcohols and lipids, for instance, used in bio-fuel production, to plastics and pharmaceuticals. Usually, the design of strains that overproduce a desired target compound involves a large number of modifications of the metabolic and regulatory network including gene knockouts/knockins and the overproduction of proteins [KJSW10]. Yet the complexity of metabolic and regulatory networks associated with the cost and effort to manipulate organisms is still a major challenge in the development of improved production designs. A large set of theoretical tools has been developed that aim at simplifying this process [BPM03, KR10, TS10]. A common feature of these methods is the prediction of metabolic flux distributions prior and after perturbations. These predictions are then combined with either deterministic or stochastic procedures that try to identify knockout and knockin combinations that improve the production of the target compound with as few genetic modifications as possible.

Recently, a new method called **C**omputational **A**pproach for **S**train **O**pti-mization aiming at high **P**roductivity (CASOP, [HK10]) based on the concept of elementary flux modes (EFMs, [SDF99]) has been proposed. However, this approach has been limited to small and medium-scale metabolic networks, so far, since the enumeration of all EFMs can only be performed in such networks [KS02]. In this work we want to outline an approach that allows us to circumvent this limitation of CASOP. Instead of computing the scores utilized in CASOP from the entire set of EFMs we compute them from a subset of the EFMs in a genome-scale network. This subset of EFMs is obtained from a sampling procedure that is similar to a previously described method to enumerate EFMs in genome-scale metabolic networks [KdFBS09]. Using our approach, CASOP can be applied even to genome-scale networks.

2 Methods

2.1 CASOP

Using the EFMs of a reaction network, CASOP calculates for each reaction a Z_2-score that indicates whether the flux through this reaction needs to be increased or decreased, in order to improve the production of a particular target compound. Similar to other methods, the reasoning behind CASOP is that the organism tries to optimize its growth yield. However, in contrast to most methods, CASOP does not assume that the organism attains the optimal flux, but rather uses a combination of optimal and, to a certain extent, sub-optimal flux distributions.

Computing CASOP-scores

In the following we give a short overview over CASOP. For a more detailed description see [HK10]. In order to determine scores for the knockout or overexpression of enzymes, CASOP considers two versions of a metabolic network that contain a biomass reaction

defining the proportion of building blocks the organism requires for its reproduction. The first network corresponds to the wild-type model. In the second network, the biomass reaction is coupled with the production of the target metabolite such that, in weights, 10% of biomass and 90% of the target metabolite are consumed. EFMs are computed in both networks. Afterward each EFM i is assigned a weight ν_i that depends on its yield in the biomass reaction, $Y^i_{Biomass/S}$ (ratio between carbon source inflow and flux through the (modified) biomass reaction). The weights of the EFMs are adjusted using a parameter k such that increasing values of k attribute higher weights to EFMs with higher yields. In this work we used a value of $k = 5$.

Afterward, a reaction importance measure is computed for each reaction in both networks as the sum of the weights of the EFMs containing this reaction. As the name suggests, the reaction importance measure allows to assess the impact of a perturbation of an enzyme catalyzing it on the production of biomass and/or the target metabolite. If one reaction has a high importance within the network containing the production of the target metabolite, but a low importance in the other network, this reaction is a candidate for overexpression since increasing the flux through it increases the flux through EFMs producing the target metabolite. In contrast, a reaction that has a low importance for the production of the target metabolite, but a high importance for sole biomass production can be removed, since it favors the flux through EFMs that do not produce the target metabolite. Hence, the Z_2-scores of CASOP, that indicate candidates for knockout and overexpression, are computed as the difference between the reaction importances of each reaction between the two networks. These scores take values between -1 and +1. A positive score indicates a reaction that is candidate for overexpression and a negative score indicates a knockout candidate. Please note that, in contrast to [HK10] we split reversible reactions in irreversible forward and backward directions. Thus, reversible reactions are assigned two CASOP scores allowing us to assess the role of forward and backward direction separately.

Building on the Z_2-scores, the CASOP procedure then knocks out the enzymes in silico that catalyzes the reaction with the most negative score by removing all EFMs containing this reaction (or other reactions catalyzed by this enzyme). Subsequently, the Z_2-scores are recomputed for the reduced set of EFMs and the procedure is iterated.

Assessing the production of the desired product

CASOP allows one to assess the impact of a genetic modification on the production of a specific target metabolite. However, no statement about the change of the production after several consecutive modifications, such as multiple knockouts, is possible. In order to observe the improvement in the production of the target metabolite, we introduce the measure Y_M which allows us to assess the relative change in yield of metabolite M after several knockouts. We make use of the weights ν_i that CASOP assigns to each EFM i (see [HK10]) in the network in which the production of the target metabolite is not associated with biomass production. Given a set of n EFMs in this network with the

individual yields in the target metabolite $Y^i_{M/S}$ of each EFM i, we derive Y_M as

$$Y_M = \sum_{i=1}^{n} \nu_i \cdot Y^i_{M/S}.$$

Since we multiply the weight of each EFM with the production of the target metabolite, Y_M can be considered as a weighted average of the yields of the EFMs in the target metabolite. If Y_M increases after a knockout, we expect this knockout to increase the production of the target metabolite M. Note that Y_M does not correspond to an actual yield, but serves as an indicator of the effect of a knockout strategy.

2.2 Enumeration of EFMs in genome-scale metabolic networks

Until recently, the computation of EFMs has been limited to small and medium-scale metabolic networks. However, fluxes within small-scale networks might be inconsistent with the corresponding fluxes within the underlying genome-scale network [KdFS09]. Several approaches for the computation of EFMs in genome-scale metabolic networks have been developed. One approach, the so-called K-shortest procedure, computes EFMs in increasing number of reactions [dFPR$^+$09]. Another approach, the EFMEvolver [KdFBS09] uses a genetic algorithm to sample large numbers of EFMs in these networks more efficiently.

Here we used a more direct approach than EFMEvolver to compute EFMs. The similarity between both methods concerns the linear programming formulation to compute a single EFM given a metabolic network (for more details see [KdFBS09]). However, instead of using a genetic algorithm, we used an iterative procedure to enumerate EFMs. Starting from an initial EFM using the target reaction, one of its reactions is selected randomly. Subsequently this reaction is blocked by setting its flux to zero and therefore, a new EFM is computed by solving the linear programming formulation. Iterating this procedure, several EFMs are obtained while the number of blocked reactions increases. If no EFM is found given a particular set of blocked reactions, the last reaction is removed from this set. Additionally, with a small probability, all reactions are removed from the set of blocked reactions. This procedure allows one to increase the diversity of the EFMs that are detected since resetting the set of blocked reactions corresponds to initializing a new independent sampling procedure. More details on the sampling procedure will be given elsewhere.

3 Results and Discussion

We applied our method to two cases: the production of succinate from glucose (studied in [HK10]) and the production of histidine from fructose in *Escherichi coli*. In each case, we started with an initial sample of 10^6 EFMs for the two networks that are required in our procedure. As a genome-scale metabolic model of *E. coli*, we used *i*AF1260 [FHR$^+$07]. Besides the carbon source, we supplied the network with the following compounds: NH_4^+,

NO_3^-, SO_4^{2-}, Fe^{2+}, Fe^{3+}, CO_2, H^+, K^+, Ca^{2+}, cobalt, molybdate, Na^+, Pi, O_2, H_2O, Cl^-, Cu^{2+}, Mg^{2+}, Mn^{2+} and Zn^{2+} that are required for the survival of the cell.

3.1 Case study I: Succinate production

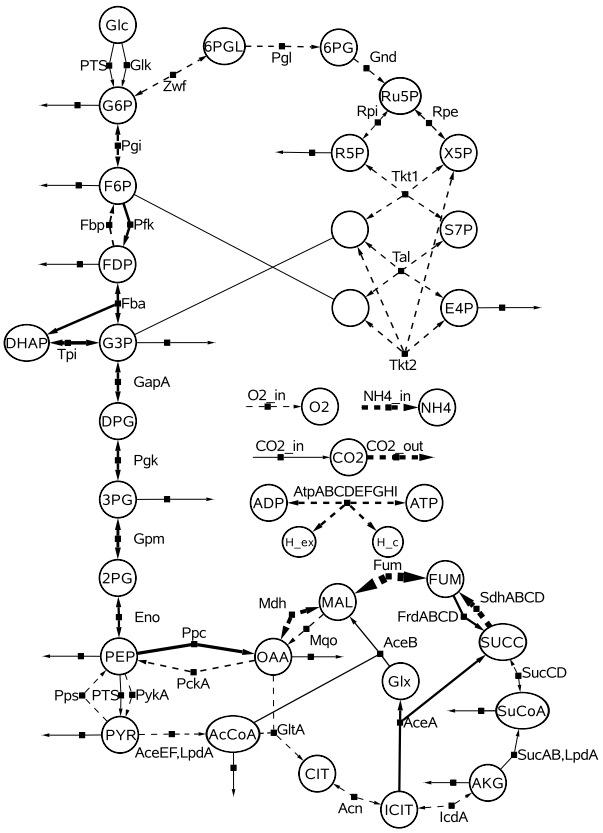

Figure 1: Z_2 scores of central metabolism of the wild-type network for succinate overproduction. The width of the arrows corresponds to the values of the scores. Dashed lines indicate negative scores, bold lines positive scores. Metabolite nodes connected by straight lines are identical. A list of abbreviations can be found in the supplementary material of [FHR+07].

The Z_2-scores for reactions within central metabolism are displayed in Fig. 1. While the relative scores for many reactions matched those discussed in [HK10] there were some differences. For instance, reactions of the glyoxylate shunt have high overexpression ratings, while this was not the case in [HK10]. The importance of such a modification to increase succinate production has been demonstrated by [LBS05]. Additionally, the overexpression of Ppc, as indicated by our analysis, is also known to improve succinate production [LBS05]. Most interestingly, fumarase (Fum) that reversibly converts fumarate into

malate received the highest knockout rating. This case exemplifies the advantage of computing the Z_2-scores of both directions of reversible reactions independently. In [HK10] both directions of reversible reactions were not considered independently and, in consequence, the score of Fum was relatively low. However, knocking out fumarase increases Y_{SUCC} almost ten-fold (Fig. 2A). This strong increase in production is probably due to the fact that this deletion interrupts the TCA cycle. In consequence, the concentration of fumarate increases which entails an increase in the concentration of the desired target metabolite succinate. Furthermore, fumarate, which is a side-product of several biosynthetic pathways, can only be disposed through conversion into succinate after this knockout if fumarate is not excreted.

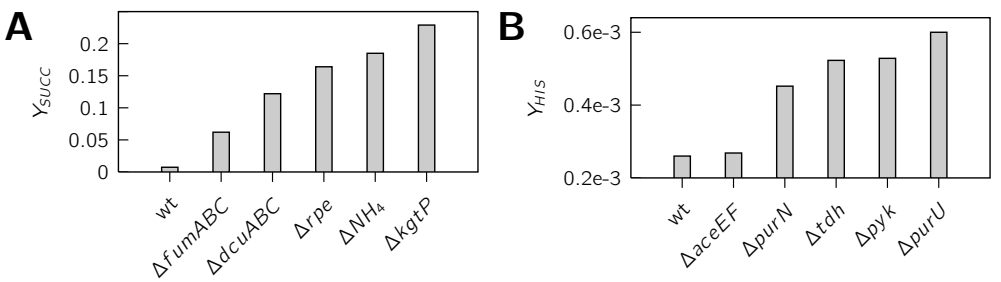

Figure 2: Y_M in the two case-studies. Knockouts are cumulative from left to right. **A** Succinate production. **B** Histidine production.

After knocking out fumarase, the fumarate transporter (Dcu) that exists in 3 isoforms, received the lowest Z_2-score. Knocking out the corresponding genes yielded a strain in which succinate production is coupled to growth. That is, biomass can only be produced when co-producing succinate. Interestingly, after this knockout, the Z_2-score of the succinate dehydrogenase SdhABCD, which is known to improve succinate production [LBS05] and had the second lowest score in the wild-type, indicates that there is no influence of a SdhABCD knockout on the production of succinate anymore. Thus, the knockout of fumarase and the fumarate transporter appears to represent an alternative knockout strategy to the knockout of succinate dehydrogenase.

In the next step, the ribulose-5-phosphate-3-epimerase (Rpe) was suggested for knockout. The forth proposed knockout involves the inflow reaction of ammonium. Knocking out this reaction corresponds to removing ammonium from the growth medium. This modification is not lethal, since we provided nitrate as alternative nitrogen source, but at the expense of reducing the growth rate [BZ90]. Moreover, nitrate can only serve as nitrogen source in the absence of oxygen [LK87]. Indeed, oxygen inflow is also assigned a relatively low Z_2-score. This indicates that the utilization of nitrate as electron acceptor and ammonium source under anaerobic conditions can improve succinate production. The fifth proposed knockout removed the export of α-ketoglutarate further reducing the number of possible pathways to excrete TCA cycle intermediates besides succinate.

3.2 Case study II: Histidine production

As a second case study we examined the production of histidine from fructose (Fig. 2B). In the first step, pyruvate dehydrogenase was suggested as knockout (Fig. 3). In the second step, the phosphoribosylglycinamide formyltransferase (PurN) was suggested as knockout. Removing this reaction drastically increased Y_{HIS} (Fig. 2B). This knockout illustrates the need of considering all reactions within a genome-scale metabolic network as knockout candidates. The reaction catalyzed by PurN consumes 10-Formyltetrahydrofolate (10-FTHF) as a co-factor. However, 10-FTHF is also required for histidine biosynthesis. In purine biosynthesis, the reaction catalyzed by PurN can also be catalyzed by the transformylase PurT that uses formate rather than 10-FTHF. Thus, knocking out PurN increases the 10-FTHF pool available for histidine biosynthesis. Furthermore, a strain with a PurN knockout grows slower than the wild-type [BAH$^+$06], indicating that the capacity of purine production might be reduced. This is of additional advantage for histidine production, since 5-Phospho-α-D-ribose-1-diphosphate (PRPP) is a common precursor of histidine and purine biosynthesis. In the following two steps, threonine dehydrogenase and pyruvate kinase were knocked out. Especially, the knockout of the threonine dehydrogenase is of interest, since it removes one of the two pathways of glycine biosynthesis from threonine. Thus, glycine biosynthesis via serine might be increased which in turn increases the cellular 10-FTHF pool whose major source is glycine biosynthesis via serine. In the fifth step, the formyltetrahydrofolate deformylase PurU that converts 10-FTHF to formate and tetrahydrofolate was knocked out.

3.3 Influence of sample sizes on Z_2-scores

In order to test the reliability of the Z_2-scores we obtained using a sample of 10^6 EFMs (Sample A), we recomputed the scores for independent samples with a higher number of EFMs: $2 \cdot 10^6$ EFMs (Sample B) and $3.7 \cdot 10^6$ EFMs (Sample C). The maximum deviations over the five knockouts between sample A and B increased over the knockout depth from 0.05 to 0.09 after the forth knockout. In all cases this maximum deviation was smaller between sample B and sample C. Here, the maximum deviation was 0.07. Slight deviations occurred in the order by which the reactions were knocked out in the three samples. After the forth knockout, the export of pyruvate rather than α-ketoglutarate received the lowest Z_2-score in the larger samples. Thus, the Z_2-scores are relatively robust if sample sizes are sufficiently large. However, for greater knockout depths, larger samples of EFMs might be required.

4 Conclusions

In this work we have presented CASOP GS as an approach that allows one to apply CASOP to genome-scale metabolic networks. Furthermore, we have introduced a mea-

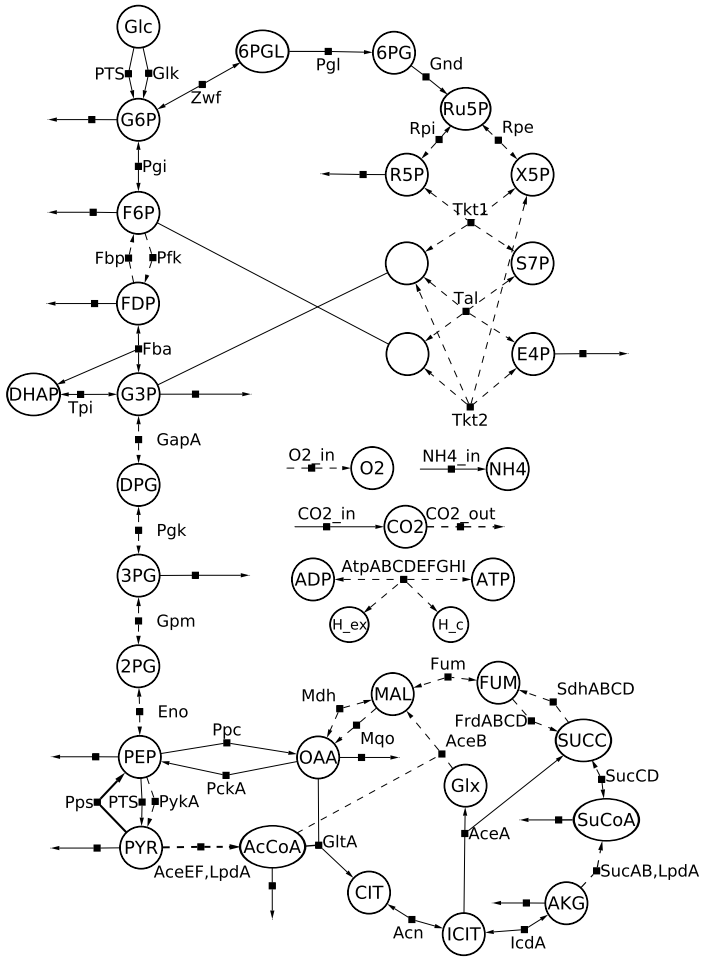

Figure 3: Z_2 scores for histidine production. For details see Fig. 1.

sure that allows one to assess the changes in the production of a target metabolite after multiple genetic modifications. Besides these improvements of CASOP, our approach offers several important advantages over other theoretical methods for strain improvement such as OptKnock [BPM03], OptGene [PRFN05] and other recently proposed approaches [KR10, TS10].

First, and most importantly, our approach provides the user with a ranking of reactions whose removal/overexpression improves the production of a target metabolite. Thus, rather than presenting a complete knockout strategy, the user has the possibility to choose, which reaction is most suitable for knockout or overexpression. This is of particular importance for the incorporation of prior knowledge about difficulties and side-effects of certain gene-manipulations. For example the principal knockout candidate might require removing a gene whose deletion is known to cause pleiotropic effects (e.g. a slow growth

rate), while the second rated knockout might yield a strain with a only slightly reduced growth rate.

Second, some approaches only consider a specific part of the metabolic network due to computational limitations. In contrast, our approach takes all reactions within an organism into account. In consequence, we do not only identify candidates for knockouts in the primary metabolism, but also in other parts of the metabolism. This is of particular importance for the overproduction of histidine, since reactions from nucleotide and amino acid metabolism appear to be suitable knockout targets.

Third, most approaches concentrate only on knockouts, while our approach, since it is an extension of CASOP, also proposes overexpression candidates to increase the production of the target metabolite. This is important since the overexpression of genes is frequently used for strain improvement.

CASOP GS offers many advantages over other approaches for the design of production strains. However, a shortcoming is that the regulatory network is not considered. In order to circumvent this problem, we are currently working on an improved version that takes into account regulatory rules by only allowing for EFMs that are consistent with the regulation of metabolism. This, regulation will be implemented in the form of Boolean logic. Moreover, the proposed knockouts of the histidine case study are currently being implemented in *E. coli* in order to validate our results.

References

[AS09] H. Alper and G. Stephanopoulos. Engineering for biofuels: Exploiting innate microbial capacity or importing biosynthetic potential? *Nat Rev Microbiol*, 7(10):715–723, Oct 2009.

[BAH+06] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2:2006.0008, 2006.

[BPM03] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–657, Dec 2003.

[BZ90] H. J. Brons and A. J. Zehnder. Aerobic nitrate and nitrite reduction in continuous cultures of *Escherichia coli* E4. *Arch Microbiol*, 153(6):531–536, 1990.

[dFPR+09] L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, and F. J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165, Dec 2009.

[FHR+07] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[HK10] O. Hädicke and S. Klamt. CASOP: A computational approach for strain optimization aiming at high productivity. *J Biotechnol*, 147(2):88–101, May 2010.

[KdFBS09] C. Kaleta, L. F. de Figueiredo, J. Behre, and S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P. Stadler, editors, *Lecture Notes in Informatics - Proceedings*, volume P-157, pages 179–189, Bonn, 2009. Gesellschaft für Informatik.

[KdFS09] C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res*, 19(10):1872–1883, Oct 2009.

[KJSW10] S. Kind, W. K. Jeong, H. Schröder, and C. Wittmann. Systems-wide metabolic pathway engineering in *Corynebacterium glutamicum* for bio-based production of diaminopentane. *Metab Eng*, Apr 2010. In print.

[KR10] J. Kim and J. L. Reed. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol*, 4(1):53, Apr 2010.

[KS02] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236, 2002.

[LBS05] H. Lin, G. N. Bennett, and K.-Y. San. Metabolic engineering of aerobic succinate production systems in *Escherichia coli* to improve process productivity and achieve the maximum theoretical succinate yield. *Metab Eng*, 7(2):116–127, Mar 2005.

[Lee09] S. Y. Lee. Systems biotechnology. *Genome Inform*, 23(1):214–216, Oct 2009.

[LK87] E. C. C. Lin and D. R. Kuritzkes. *Escherichia coli and Salmonella typhimurium - Cellular and Molecular Biology*, volume I, chapter 16 - Pathways for anaerobic electron transport, pages 201–221. ASM, Washington, 1987.

[NKL10] D. Na, T. Y. Kim, and S. Y. Lee. Construction and optimization of synthetic pathways in metabolic engineering. *Curr Opin Microbiol*, Mar 2010. In print.

[PRFN05] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308, 2005.

[SDF99] S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60, Feb 1999.

[TS10] N. Tepper and T. Shlomi. Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways. *Bioinformatics*, 26(4):536–543, Feb 2010.