*Bioinformatics*, 37(12), 2021, 1717–1723
doi: 10.1093/bioinformatics/btaa996
Advance Access Publication Date: 27 November 2020
Original Paper

OXFORD

Systems biology

# GeneReg: a constraint-based approach for design of feasible metabolic engineering strategies at the gene level

**Zahra Razaghi-Moghadam[1,2] and Zoran Nikoloski[1,2,*]**

[1]Department of Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam and [2]Department of Systems Biology and Mathematical Modeling, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on July 2, 2020; revised on October 24, 2020; editorial decision on November 9, 2020; accepted on November 17, 2020

## Abstract

**Motivation:** Large-scale metabolic models are widely used to design metabolic engineering strategies for diverse biotechnological applications. However, the existing computational approaches focus on alteration of reaction fluxes and often neglect the manipulations of gene expression to implement these strategies.

**Results:** Here, we find that the association of genes with multiple reactions leads to infeasibility of engineering strategies at the flux level, since they require contradicting manipulations of gene expression. Moreover, we identify that all of the existing approaches to design gene knockout strategies do not ensure that the resulting design may also require other gene alterations, such as up- or downregulations, to match the desired flux distribution. To address these issues, we propose a constraint-based approach, termed GeneReg, that facilitates the design of feasible metabolic engineering strategies at the gene level and that is readily applicable to large-scale metabolic networks. We show that GeneReg can identify feasible strategies to overproduce ethanol in *Escherichia coli* and lactate in *Saccharomyces cerevisiae*, but overproduction of the TCA cycle intermediates is not feasible in five organisms used as cell factories under default growth conditions. Therefore, GeneReg points at the need to couple gene regulation and metabolism to design rational metabolic engineering strategies.

**Availability and implementation:** https://github.com/MonaRazaghi/GeneReg

**Contact:** zniko@uni-potsdam.de

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Computational modeling of metabolic networks has become a prominent approach in the design of cell factories that accomplish a desired goal, e.g. improving yield of a chemical compound (Burgard *et al.*, 2003; Kamp and Klamt, 2017; Patil *et al.*, 2005; Pharkya and Maranas, 2006; Pharkya *et al.*, 2004; Ranganathan *et al.*, 2010; Rocha *et al.*, 2010). The design of cell factories is achieved by metabolic engineering strategies which involve a combination of knockouts, over- or under-expression of native genes of a host organism along with insertion of genes from other organisms and even synthetic enzymes which do not occur in nature (Erb *et al.*, 2017). The creation of metabolic manipulation strategies has largely been propelled by the developments in the constraint-based modeling framework (Burgard *et al.*, 2003; Kamp and Klamt, 2017; Patil *et al.*, 2005; Pharkya and Maranas, 2006; Pharkya *et al.*, 2004; Ranganathan *et al.*, 2010; Rocha *et al.*, 2010). The existing computational approaches for prediction of metabolic engineering strategies can be roughly grouped into optimization-based or enumeration-based. The former use optimization-based formulation of the engineering task and can accommodate the gene manipulations, while the latter enumerate set of reactions whose knockouts achieve a desired goal (Maia *et al.*, 2015).

Common to both types of approaches is that they result in engineering strategies that specify how the fluxes through selected reactions are to be manipulated to achieve the desired goal. For instance, OptKnock uses bilevel programming to identify reaction knockouts which improve a desired goal while ensuring growth (Burgard *et al.*, 2003). In addition, OptReg approach also uses bilevel programming to identify up- or downregulation of reaction fluxes and reach a metabolic engineering goal (Pharkya and Maranas, 2006). In contrast, the concept of minimal cut sets can be used to enumerate reaction knockouts of specific size that disrupt a function in the network. The latter has been recently employed to demonstrate that growth-coupled overproduction is feasible for almost all metabolites in five organisms used in metabolic engineering (Kamp and Klamt,

2017). Yet, what a majority of the existing approaches neglects to determine is whether or not the predicted strategies can, in fact, be realized by appropriate manipulation of the underlying set of genes, included in the models via the gene–protein–reaction (GPR) rules.

GPR rules characterize the genes and respective proteins that catalyze reactions. They can either be included as additional annotation of reactions (Reed *et al.*, 2003) or as part of the stoichiometric matrix that represents the stoichiometry of the reactions included in the model (Machado *et al.*, 2016). GPR rules specify whether a protein catalyzes a reaction on its own or if there are isoenzymes, enzyme complexes or their combinations (Reed *et al.*, 2003). For instance, a reaction catalyzed by two isoenzymes, encoded by the genes $g_1$ and $g_2$, is denoted by the GPR rule ($g_1$ OR $g_2$); further, a reaction catalyzed by an enzyme complex composed of two units, encoded by $g_1$ and $g_2$, is denoted by ($g_1$ AND $g_2$). We refer to GPR rules that include more than one gene as complex rules. For instance, in the metabolic network in Figure 1A, the metabolite $M_1$ is taken up as a substrate from the environment to produce metabolite $M_5$, the undesired by-product $M_4$, and the metabolite $M_6$, which denotes the engineering objective. The GPR rules in this example are such that $R_4$ is a reaction catalyzed by an enzyme complex composed of two proteins encoded by the genes of $g_1$ and $g_3$. Further, to carry flux, reaction $R_2$ requires any of the two isoenzymes encoded by the genes $g_1$ and $g_2$.

The participation of genes in multiple GPR rules and the inclusion of complex GPR rules may lead to conflicts on the level of gene manipulations. For instance, in the network presented in Figure 1A and B, with the objective of maximizing the production of metabolite $M_5$, the pathway through $R_2$ and $R_3$ yield the optimal solution (Fig. 1C). On the other hand, maximizing the production of $M_6$ requires flux via the reaction $R_4$ and blocking any flux through reactions $R_2$ and $R_3$ (Fig. 1D). On the level of gene manipulations, $R_4$ can carry flux if the proteins encoded by the genes $g_1$ and $g_3$ are present. However, to remove the reaction $R_2$, it is necessary to knockout genes $g_1$ and $g_2$, which encode the associated isoenzymes. Therefore, this leads to an infeasible design for the objective of maximizing the production of $M_6$ due to contradicting gene manipulation—the need for simultaneous presence and absence of the product encoded by the gene $g_1$.

We hypothesize that such infeasibilities are likely to occur in metabolic engineering strategies obtained with the existing computational approaches in genome-scale metabolic models. For instance, by using a metabolic model of *Escherichia coli*, OptKnock predicts that 63% increase in production of succinate in comparison to the wild-type is possible by knocking out pyruvate formate lyase and lactate dehydrogenase (Burgard *et al.*, 2003). Looking closer at the GPR rules of pyruvate formate lyase (EC2.3.1.54) in the *E.coli* iJR904 model (Reed *et al.*, 2003), we observe that it involves five genes (three enzymes) in a complex GPR rule: (b3951 AND b3952) OR (b0902 AND b0903) OR b3114. Knocking out this reaction requires that the three isoenzymes are knockedout. However, the last two terms form the GPR rule of the 2-oxobutanoate formate lyase, which will be also knockedout although the metabolic engineering strategy must not alter the flux through the respective reaction. As another example, OptReg proposes a metabolic manipulation strategy that leads to 94.3% of the maximum theoretical yield of ethanol production of 19.87 mmol/(gDW h). In this design, the overproduction is possible by knocking out pyruvate formate lyase, which is likewise also infeasible (Pharkya and Maranas, 2006). Moreover, Machado *et al.* (2016) tested the feasibility of strain designs proposed by the enumeration-based method (Kamp and Klamt, 2017) and the results show that fewer than 10% of the proposed solutions are actually feasible, when GPR rules are considered. All these examples demonstrate that complex GPR rules often lead to contradicting gene manipulations, referred to as gene conflicts, and that there is a need for overcoming this important problem in the design of metabolic engineering strategies.

Two constraint-based approaches, OptGene (Patil *et al.*, 2005) and OptORF (Kim and Reed, 2010), facilitate manipulation of genes. However, OptGene suffers from non-linear constraints, which render it impractical for usage with large-scale models. This issue is partly addressed by using an evolutionary programming approach, which, however, allows only to approximate the solution. In contrast, OptORF allows only gene knockouts and off–on overexpression, while finer manipulations are not considered. Although OptORF includes linear constraints that capture the environmental regulation of metabolism, it is challenging to implement this approach since it requires additional information about condition-specific activation of enzyme-coding genes that is not readily available even for model organisms that experience multi-factorial environmental cues (see Section 3.2, for further details). There is a modified version of OptORF which is based on bilevel programming that only considered knockout manipulations while claiming to couple reaction and gene manipulations (Kim *et al.*, 2011). This variant can be seen as analogous to OptKnock (Burgard *et al.*, 2003), with a modified objective that prefers fewer gene knockout. While this variant of OptORF is feasible to implement (see https://github.com/MonaRazaghi/GeneReg), the considered constraints that couple reaction flux and its activity [on–off, see constraint A.3 in Supplementary Text of Kim *et al.* (2011)] leads to the case in which a reaction does not carry flux (since the lower flux bound is often set to zero), although it is considered active. Therefore, we recognize that there is no working solution that resolves gene conflicts while allowing fine tuning of gene manipulations, which is the problem we address in this study.

Another elegant way to approach this problem is to rely on logical transformation of model (LTM) that enforces gene–reaction associations. These associations can then be combined with constraint-based approaches for metabolic engineering (Zhang *et al.*, 2015). LTM is based on inclusion of pseudo-reactions and metabolites that facilitate the simulation of complex GPR rules (i.e. isoenzymes and protein complexes). As a result, the metabolic models extended based on LTM have usually 2- to 5-fold more reactions and metabolites in comparison to the original models, while leaving the number of gene intact (Zhang *et al.*, 2015). However, LTM only allows the simulation and design of knockout strategies, while fine-tuned up- and downregulations of genes are not considered. The question that remains is whether the metabolic engineering strategies feasible at the gene level can be achieved with a considerably smaller number of variables, rendering the models applicable with more involved constrained-based approaches (e.g. considering integer-value variables).

Moreover, and most importantly, both approaches with and without LTM when applied to design knockout strategies do not ensure that the engineering goal is achieved only by gene knockout (i.e. without any other transformations, such as up- or downregulation). For instance, in the metabolic network in Figure 1A and B, with the objective of maximizing the production of metabolite $M_5$, knockout strategies only propose blocking flux through reactions $R_2$ and $R_3$ (Fig. 1D) and ignore the fact that the engineering goal is impossible without upregulation of $R_4$. This possibility is also not precluded by the elegant LTM approach.

Here, we first investigate how complex the GPR rules are across organisms, including those relevant for biotechnological applications, for which there are high-quality metabolic models available. We then propose an optimization-based approach, called GeneReg, which predicts gene manipulation strategies while avoiding gene conflicts. Finally, we show that the approach is applicable to genome-scale metabolic networks of *E.coli* and *Saccharomyces cerevisiae*, and pinpoint the differences in comparison to the contenders. Applications of GeneReg brings into question recent claims about feasibility of growth-coupled overproduction of key metabolites in major production organisms, namely *E.coli*, the Gram-positive bacterium Corynebacterium *glutamicum*, the filamentous fungus Aspergillus *niger* and the cyanobacterium *S.* sp. PCC6803 (Kamp and Klamt, 2017). Altogether, GeneReg provides the means for rational and feasible design of cell factories.

## 2 Materials and methods

Inspired by the OptReg approach, we devise an optimization-based approach, termed GeneReg, to predict metabolic engineering

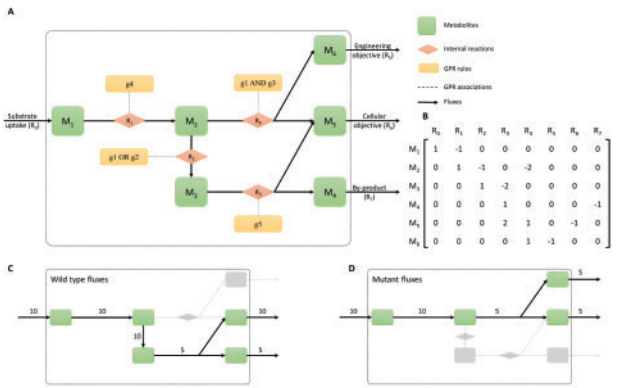

**Fig. 1.** Illustration of gene conflicts. (**A**) Metabolic network with eight reactions, four internal and four exchange, transforming six metabolites together with (**B**) the stoichiometric matrix representation. (**C**) Maximizing the production of metabolite $M_5$ is realized by the pathway through $R_2$ and $R_3$, and allowing only knockouts, (**D**) maximizing the production of metabolite $M_6$ requires flux via the reaction $R_4$ and blocking reactions $R_2$ and $R_3$. The latter leads to a gene conflict since the product of gene $g_1$ has to be both present and absent from the network to facilitate flux through $R_4$ and blocking of $R_2$. Note that the mutant flux distribution also requires upregulation of flux through reaction $R_4$

strategies without gene conflicts. A reaction flux is called up-/downregulated if the flux value is considerably higher/lower with respect to its steady-state value at the optimal growth (made precise by introducing a parameter $C$, below). To this end, we first determine the flux range of every reaction subject to the steady-state constraints. The steady-state assumption can be described by a system of linear equations as:

$$\frac{dx}{dt} = S \cdot v = 0, \tag{1}$$

where $S_{m\times r}$ is the stoichiometry matrix, in which $m$ is the number of metabolites and $r$ is the number of reactions. Here, $x$ is the vector of size $m$, representing metabolite concentration, and $v$ is the vector of size $r$ of flux values. We apply flux variability analysis (FVA) to calculate the feasible minimum and maximum flux values, $v_i^{\min}$ and $v_i^{\max}$, for each reaction $i$ ($1 \leq i \leq r$). We then calculate the minimum and maximum fluxes, $v_i^L$ and $v_i^U$, that each reaction $i$ supports at the optimal growth (i.e. biomass production).

The parameter $C \in [0, 1]$ is defined to be tunable and quantifies the deviation from the maximum and the minimum flux values at the optimal growth: reaction $i$ is referred to as upregulated if its flux value is in the range

$$v_i^U(1 - C) + v_i^{\max} C \leq v_i \leq v_i^{\max}. \tag{2}$$

In contrast, reaction $i$ is considered downregulated if

$$v_i^{\min} \leq v_i \leq v_i^L(1 - C) + v_i^{\min} C. \tag{3}$$

For every reaction $i$ ($1 \leq i \leq r$), we introduce two binary variables $y_i^u$ and $y_i^d$ which take, respectively, a value of 0, if the reaction is up- or downregulated, and 1, otherwise. These binary variables allow to bound the range of flux values according to the corresponding gene manipulations. Based on the following constraints, the binary variables $y_i^u$ and $y_i^d$ enable us to switch between the flux ranges and determine whether or not the reaction $i$ is up- and downregulated, respectively:

$$v_i^L y_i^u + (v_i^U(1 - C) + v_i^{\max} C)(1 - y_i^u) \leq v_i, \tag{4}$$

$$v_i \leq (v_i^{\max})(1 - y_i^u) + v_i^U y_i^u, \tag{5}$$

$$v_i^{\min}\left(1 - y_i^d\right) + v_i^L y_i^d \leq v_i, \tag{6}$$

$$v_i \leq \left(v_i^L(1 - C) + v_i^{\min} C\right)\left(1 - y_i^d\right) + v_i^U y_i^d. \tag{7}$$

Equations (4) and (5) show that in case of upregulation, i.e. when $y_i^u = 0$, the flux range is $[v_i^U(1 - C) + v_i^{\max} C, v_i^{\max}]$ and $[v_i^L, v_i^U]$, otherwise. Similarly, Eqs. (6) and (7) restrict the flux range in response to downregulation. The constraints in Eqs. (2)–(7) are motivated by OptReg (Pharkya and Maranas, 2006), and resolve the issue of this approach so that down- and upregulated reactions in an engineering strategy have fluxes which do not fall in the FVA ranges at optimal growth (i.e. those of the wild-type) if they are declared as altered.

Another constraint is imposed by the fact that a reaction can be the target of at most a single type of manipulation; thus, the following constraint is added:

$$(1 - y_i^u) + \left(1 - y_i^d\right) \leq 1. \tag{8}$$

In addition and in analogy to OptReg, a limit on the total number of reaction manipulations ($L$) can be captured by:

$$\sum_{i=1}^{r} (1 - y_i^u) + \left(1 - y_i^d\right) \leq L. \tag{9}$$

Besides, for reversible reactions either the forward or the backward reactions can be the target of a manipulation. To impose this constraint, we first split all reversible reactions into forward and backward reactions, and then set the following constraints:

$$y_{i_f}^d + y_{i_b}^d \geq 1, \tag{10}$$

$$y_{i_f}^u + y_{i_b}^u \geq 1, \tag{11}$$

where $i_f$ and $i_b$ are, respectively, the indices of the forward and backward reactions of the reversible reaction $i$. Constraints (8)–(11) also appear in OptReg (Pharkya and Maranas, 2006).

To overcome the shortcomings of reaction-based strategies, we opt to design a manipulation strategy that is feasible at the gene level. Therefore, we introduce novel constraints that capture the coupling between the modulation of the reactions and the respective genes. Our main contribution is the encoding of constraints for the different GPR rules in a form of a mixed-integer linear program (MILP).

In the simplest case, when the reaction $i$ is catalyzed by the product of a single gene $g_{i_1}$, can be captured by two equality constraints:

$$g_{i_1}^u + y_i^u = 1, \tag{12}$$

$$g_{i_1}^d + y_i^d = 1. \tag{13}$$

Equations (12) and (13) model the coupling between gene manipulation and the up-/downregulation of reaction $i$, i.e. $y_i^{u/d} = 0$ implies that $g_{i_1}$ must be manipulated, and $y_i^{u/d} = 1$ implies that $g_{i_1}$ remains unchanged. To prevent conflicts, for every gene $j$, we add the constraint that $g_j^d + g_j^u \leq 1$.

Further, if the reaction $i$ is catalyzed by $n$ isoenzymes each encoded by an independent gene, the GPR rule ($g_{i_1}$ OR $g_{i_2}$ OR ... OR $g_{i_n}$), is captured by two sets of inequality constraints:

$$1 \leq \left(\sum_{j=1}^{n} {g_{i_j}}^u\right) + n \cdot y_i^u \leq n, \tag{14}$$

$$n \leq \left(\sum_{j=1}^{n} {g_{i_j}}^d\right) + n \cdot y_i^d \leq 2n - 1. \tag{15}$$

The left-hand side of Eq. (14) models the upregulation of reaction $i$, i.e. $y_i^u = 0$, which implies that $\sum_{j=1}^{n} {g_{i_j}}^u \geq 1$, i.e. at least one of the genes must be upregulated. The right-hand side of Eq. (14)

models the case when the reaction $i$ is not upregulated, i.e. $y_i^u = 1$, whereby ${g_{i_j}}^u = 0$, which implies that no gene encoding an isoenzyme is upregulated. In addition, Eq. (15) models the presence or absence of downregulation of reaction $i$. Similarly, for the case that the reaction $i$ is catalyzed by an enzyme complex composed to $n$ units, the GPR rule ($g_{i_1}$ AND $g_{i_2}$ AND ... AND $g_{i_n}$) is captured by two sets of inequality constraints:

$$n \leq \left(\sum_{j=1}^{n} {g_{i_j}}^u\right) + n y_i^u \leq 2n - 1, \tag{16}$$

$$1 \leq \left(\sum_{j=1}^{n} {g_{i_j}}^d\right) + n y_i^d \leq n. \tag{17}$$

In the case of a more complicated reaction $i$, where the reaction is catalyzed by $n$ isoenzymes each encoded by either a single gene or an enzyme complex, which we refer to as *complex isoenzymes*, the GPR rule ($g_{1,1}$ AND $g_{1,2}$ AND ... AND $g_{1,i_1}$) OR ($g_{2,1}$ AND $g_{2,2}$ AND ... AND $g_{2,i_2}$) OR ... ($g_{n,1}$ AND $g_{n,2}$ AND ... AND $g_{n,i_n}$) is captured by four sets of inequality constraints:

$$\forall 1 \leq j \leq n: \; 0 \leq \left(\sum_{k=1}^{i_j} g_{j,k}^u\right) - i_j G_j^u \leq i_j - 1, \tag{18}$$

$$1 \leq \left(\sum_{j=1}^{\mathrm{n}} G_j^u\right) + n \cdot y_i^u \leq n, \tag{19}$$

$$\forall 1 \leq j \leq n: \; 1 - i_j \leq \left(\sum_{k=1}^{i_j} g_{j,k}^d\right) - i_j G_j^d \leq 0, \tag{20}$$

$$n \leq \left(\sum_{j=1}^{n} G_j^d\right) + n y_i^d \leq 2n - 1, \tag{21}$$

where $G_j^u$ and $G_j^d$ $(j = 1, \ldots, n)$ are variables defined to represent the status (up-/downregulation) of the complex of ($g_{j,1}$ AND $g_{j,2}$ AND ... AND $g_{j,i_j}$).

Moreover, for each gene, there is either a possibility for upregulation or downregulation, and to avoid conflict in manipulation, the following constraint is added for each gene $g$ in the model:

$$g^u + g^d \leq 1. \tag{22}$$

Instead of the bilevel optimization formulation of OptReg, which maximizes the flux toward the synthesis of the desired biochemical while simultaneously maximizing growth, here, we minimize the number of gene manipulations while guaranteeing that: (i) a certain factor, $\alpha_{\mathrm{growth}}$, of optimal growth is achieved and (ii) the flux toward the objective is at least a given factor of the maximum product yield. The objective of the optimization problem in GeneReg is to minimize the number of gene manipulations while guaranteeing that a certain factor of optimal growth is achieved, and the flux toward the synthesis of the engineering objective is at least a given factor, $f_{\mathrm{Obj}}$, of its maximum possible value (when optimized as a sole objective). This can be guaranteed by having the flowing constraints:

$$v_{\mathrm{growth}} \geq \alpha_{\mathrm{biomass}} \cdot v_{\mathrm{growth}}^{\max}, \tag{23}$$

$$v_{\mathrm{engineering\ obj}} \geq f_{\mathrm{Obj}} \cdot v_{\mathrm{engineering\ obj}}^{\max}, \tag{24}$$

where $\alpha_{\mathrm{biomass}} < 1$ and $f_{\mathrm{Obj}} \leq 1$ are given constants. The objective function can be written as $\sum_{g_i \in \mathrm{Genes}} (g_i^d + g_i^u)$. Combining the linear objective function with the linear constraints presented in Eqs. (1)–(24) leads to our proposed MILP formulation of GeneReg, which is implemented in Matlab2017b and is available at https://github.com/MonaRazaghi/GeneReg.

## 3 Results

### 3.1 Complex GPR rules across organisms

To determine whether gene conflicts may arise in the design of metabolic engineering strategies, we first investigate the occurrence of complex GPR rules in 21 metabolic models across different species from the BiGG database (King *et al.*, 2016) (Supplementary Table S1). We find that between 48.5% and 93.1% of reactions in the models are associated to at least one gene. Of these reactions, between 12.4% and 54.8% have complex rules, with an average number of rules per gene ranging from 1.4 to 13.7 (Fig. 2 and Supplementary Table S1). In addition, we observe that the maximum number of GPR rules in which a gene participates varies from 8 to 1040. Since a gene can be associated with multiple reactions, it can be included in different rules. Calculating the maximum percentage gene occurrence in rules over all genes in a model, we find that it varies between 1.5% and 28.5% over the considered models (Fig. 2 and Supplementary Table S1). We also observe that some of the GPR rules for one reaction are embedded as part of other, more complex rules, and refer to these as *rule repetitions*. We find that the average rule repetition ranges from 0.02% to 0.4%, while the maximum rule repetition ranges from 1.4% to 25.1% across the models (Fig. 2 and Supplementary Table S1). Therefore, we conclude that complex rules are present to a considerable degree in the analyzed metabolic models and, therefore, gene conflicts are expected to be present in the design of metabolic engineering strategies. As a result, to arrive at metabolic engineering strategies that are feasible at the gene level, we rely on the GeneReg approach explained in detail in Section 2.

### 3.2 Feasible strategy for overproduction of ethanol in *E.coli*

To assess the performance of the proposed optimization approach, GeneReg, we employed it on the iJR904 metabolic model for *E.coli*, to determine a modification strategy for overproducing ethanol feasible at the gene level. This model includes 761 metabolites participating in 1075 reactions (821 irreversible and 254 reversible reactions), along with 904 genes associated to 873 GPR rules. The maximum possible flux for ethanol synthesis ($v_{\mathrm{ethanol}}^{\max}$) is 20 mmol/(gDW h), while the maximum flux of ethanol synthesis ($v_{\mathrm{ethanol}}^{U}$) at the optimal growth (of 0.922 per hour) is 0.046 mmol/(gDW h). We are interested in designing a metabolic engineering strategy that results in 30% of the maximum production of ethanol at reduction in growth not larger than 50% of the optimal. To ensure the existence of a feasible solution, the tunable parameter $C$ is set to 0.01, and no limit is placed on the total number of reaction

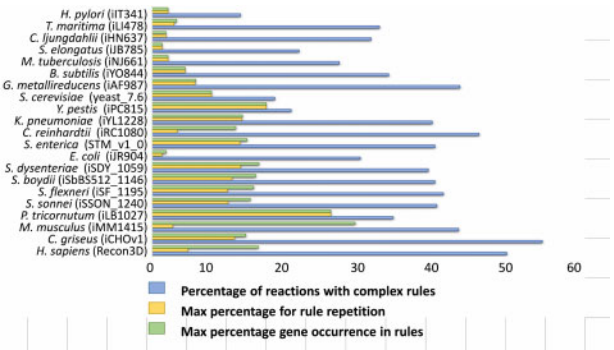


**Fig. 2.** GPR rules in metabolic models across species. GPR rules are collected from metabolic models of 21 organisms across kingdoms of life. The bar-plot illustrates the maximum percentage of rules in which any gene occurs (green, max gene rule occurrence), the maximum percentage of rule repetition in more complex rules (yellow, max GPR repetition), and the percentage of reactions with complex GPR rules (blue, percentage of reactions with complex rules). Complex rules are prevalent in large-scale metabolic networks, increasing the likelihood for gene conflicts in metabolic engineering strategies

manipulations. With this setting, GeneReg identifies a modification strategy which involves 85 genes and 300 reactions, of which 114 include GPR rules (see Supplementary Tables S2 and S3).

The modified reactions take part in 14 metabolic subsystems. The eight metabolic subsystems with the largest number of modified reactions include the cell envelope biosynthesis pathway (28), cofactor and prosthetic groups (24), nucleotide salvage pathway (12), histidine metabolism (9), membrane lipid metabolism (7), tyrosine, tryptophan and phenylalanine metabolism (7), valine, leucine and isoleucine metabolism (6), methionine metabolism (6) (Supplementary Table S3).

It can be expected that avoiding conflicts in gene-based strategies may lead to a larger number of manipulations in comparison to the number of manipulations in reaction-based designs. To compare the strategy proposed by GeneReg with that of reaction-based designs with respect to the number of reactions manipulated, we applied a modified version of OptReg. The modification to optReg include: (i) flux ranges in response to up-/downregulations (see Section 2) and (ii) the minimization of the number of reaction manipulations, while guaranteeing that a certain factor of optimal growth is achieved. The strategy predicted by the modified OptReg implies the manipulation of 99 reactions, of which 74 are associated with GPR rules (see Supplementary Table S4). This implies that, with the comparable number of manipulations based on the reaction associated with genes, GeneReg proposes a feasible strain design of similar size while addressing the shortcomings of the reaction-based design.

To assess GeneReg and the modified version of OptReg, we examine the reactions associated with GPR rules proposed for manipulation by both approaches. To increase the production of ethanol, the solution of GeneReg includes 114 reactions with GPR rules, 110 of which need to be downregulated and the other four are upregulated. Respectively, out of 74 reactions considered by the modified OptReg, 73 have to be downregulated and the only reaction to be upregulated is fatty acid oxidation (*n*-C16:0). The complex GPR rule corresponding to fatty acid oxidation (*n*-C16:0) comprises four genes that also participate in GPR rules of three other reactions, whose flux must remain unchanged. This again indicates at the infeasibility of the strategy proposed by modified OptReg. In comparison, the reactions proposed by GeneReg to be upregulated involve three genes b1262, b1263 and b1264, which participate in no other reactions (see Supplementary Table S5).

Taking a closer look at the list of reactions to downregulate shows that the strategy devised by GeneReg specifies 37 more reactions in addition to the 73 proposed by the modified OptReg. To illustrate the issue, we consider a reaction that is predicted to be downregulated by both approaches, namely UDP-sugar hydrolase. This reaction is associated with gene b0480, which itself is present in the GPR rules of 11 other reactions. While all of these reactions must have unaltered flux, according to OptReg, they are all considered as downregulated in the strategy proposed by GeneReg (see Supplementary Tables S3 and S4).

Comparison with OptGene is not possible due to the mentioned non-linear constraints involved, rendering the application to genome-scale models impractical. Comparison to OptORF is not feasible due to the lack of public implementation and availability of fully parameterized model, detailing the active reactions as input to this approach. For a comparison on a small example, akin to that shown in Figure 1, please, see Supplementary Material.

### 3.3 General observations for the tunable parameter

As previously explained, the tunable parameter *C* determines the degree of deviation from the wild-type flux values at the optimal growth. To investigate the effect of this parameter on the proposed strategy for overproducing ethanol, we used five different values of *C* in the optimization formulation, and determine the effect of tuning this parameter on the number of reactions and genes required to be manipulated (Fig. 3). As shown in Figure 3, the higher value of the tunable parameter results in a larger number of genes and reactions (with or without GPR rules) to be manipulated. For values of the parameter *C* of at least 0.05, no feasible solution can be identified. The question is then how general are these observations?

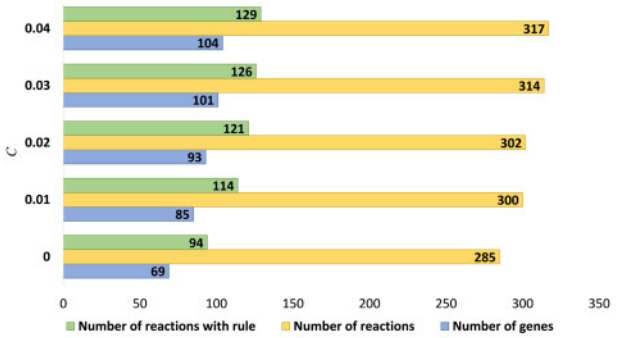


**Fig. 3.** The effect of the tunable parameter in GeneReg. The figure summarizes the effect of five values for the tunable parameter *C* on the characteristics of the engineering strategy for overproducing ethanol in iJR904 metabolic model of *E.coli* (Reed *et al.*, 2003) in terms of the number of reactions (with or without GPR associations) and genes. Increasing the value of the tunable parameter leads to an increase in the number of genes and reactions to be manipulated, up to the value when infeasibility is reached

The parameter *C* determines the range in which a particular reaction is considered up- or downregulated. By increasing the value of parameter *C* the ranges for up-/downregulations, determined by FVA, become smaller. Therefore, the number of reactions and genes that need to be manipulated increases. Importantly, in the case of knockouts, the sets of genes and reactions that need to be knocked-out for a given value of *C* are always subsets of the sets proposed for a higher value of the *C*. However, this is not guaranteed for up-/downregulation, since the changes in the ranges of up-/downregulation have effect on the definition of correspondingly regulated reactions. Therefore, GeneReg looks for new sets of upregulated reactions and genes (see Supplementary Table S6).

To determine which value of *C* to use, we recommend that a search in the interval [0, 1], with increments of 0.01, is conducted up to reaching infeasibility for the MILP formulation of GeneReg and based on the strength of the modifications (e.g. promoters, silencing) used in the engineering strategy. Following this suggestion, we also conducted empirical tests for the time required to design an engineering strategy. Our results indicate that the running time on the different values for *C* ranges from 18.8 to 992.8 min, for overproducing ethanol in *E.coli* (see Supplementary Table S6). The timing in other models is expected to depend on the size of the model and the engineering strategy to be reached.

### 3.4 Feasible strategy for overproduction of lactate in *S.cerevisiae*

The findings from the *E.coli* model have motivated us to examine the performance of GeneReg on a larger metabolic model. To this end, we employed GeneReg on the metabolic model of *S.cerevisiae*, yeast 7.6 (Aung *et al.*, 2013), to identify a metabolic engineering strategy that increases the production of lactate—a compound that has diverse applications in food-processing and industry (Chang *et al.*, 1999). We used the yeast 7.6 model that contains 2220 metabolites participating in 3493 reactions, of which 2302 are associated with GPR rules accounting for 909 genes. The optimal growth is 23.73 per hour associated with the maximum flux toward lactate synthesis is 56.66 mmol/(gDW h), while the maximum possible flux toward lactate synthesis is 1000 mmol/(gDW h). Here, we are interested in achieving 20% of the maximum production of lactate, while guaranteeing that 80% of optimal growth is achieved. The tunable parameter *C* is set to 0.001 and no limit is placed on the total number of reaction manipulations. GeneReg predicts a modification strategy in which 134 reactions needed to be manipulated, out of which 41 are associated with GPR rules. The minimum number of genes required to implement these manipulations is 36 (see Supplementary Table S7).

**Table 1.** Comparison of the number of variables and constraints in GeneReg and LTM

| | Number | *E.coli* (iJR904) | *S.cerevisiae* (yeast 7.6) | *C.glutamicum* (iJM658) | *A.niger* (iMA871) | *S.* sp. PCC6803 (iJN678) |
|---|---|---|---|---|---|---|
| Original | Metabolites | 761 | 2220 | 982 | 1049 | 795 |
| | Reactions | 1329 | 4821 | 1452 | 1988 | 1087 |
| | Genes | 904 | 909 | 658 | 731 | 622 |
| LTM | Metabolites | 2103 | 5112 | 2461 | 4862 | 1828 |
| | Reactions | 2930 | 8630 | 3408 | 7225 | 2313 |
| | Variables | 6764 | 18 169 | 7474 | 15 181 | 5305 |
| | Eq. constraints | 5033 | 13 742 | 5869 | 12 087 | 4141 |
| | In. constraints | 5860 | 17 260 | 6818 | 14 450 | 4626 |
| GeneReg | Variables | 3614 | 10 622 | 3593 | 4707 | 2886 |
| | Eq. constraints | 1539 | 4426 | 1838 | 1513 | 1404 |
| | In. constraints | 3324 | 10 986 | 3484 | 5026 | 2650 |

Note: The comparison is performed for five models of organisms used as cell factories. Eq. and In. constraints stand for Equality and Inequality constraints, respectively. The number of reactions considers the splitting of reversible reactions.

### 3.5 Comparison of model size between GeneReg and LTM

LTM provides an elegant way to fully capture the gene–reaction association while considering the specific types of GPR rules. While LTM leads to the increase in the number of metabolites and reactions (Zhang *et al.*, 2015), it is also important to determine the number of variables and constraints used in the design of a metabolic engineering strategy. These numbers determine the size of the MILPs used. Therefore, we examined the number of variables as well as equality and inequality constraints resulting from the application of GeneReg and LTM to models of five organisms: *E.coli* [iJR904 (Reed *et al.*, 2003)], *S.cerevisiae* [yeast 7.6 (Aung *et al.*, 2013)], the Gram-positive bacterium *C.glutamicum* [iJM658 (Mei *et al.*, 2016)], the filamentous fungus *A.niger* [iMA871 (Andersen *et al.*, 2008)] and the cyanobacterium *S.* sp. PCC6803 [iJN678 (Nogales *et al.*, 2012)]. For the case when only knockouts are considered, as shown in Table 1, the number of variables in LTM in comparison to GeneReg is 1.7-fold larger for the yeast 7.6 model and as much as 3.2-fold larger for the model of *A.niger*. Similarly, the number of constraints is at least 2-fold larger in LTM in comparison to GeneReg across all five models. Therefore, we concluded that GeneReg consistently results in smaller models. Moreover, for the case of GeneReg with up- and downregulation of reaction fluxes, the number of variables in LTM is 1.1-fold larger for the yeast 7.6 model and as much as 2-fold larger for the

model of *A.niger*. In this case, the number of constraints in GeneReg and LTM is similar (see Supplementary Table S8). Most importantly, LTM allows the design of engineering strategies that involve knockouts, but does not facilitate the development of strategies that include up- and downregulation of genes, which is the principal novelty of GeneReg.

### 3.6 Infeasibility of overproduction for key metabolites in five major production organisms

Kamp and Klamt (2017) used the concept of minimum cut set to design optimal reaction-based strategies, and showed that it is feasible to overproduce almost all metabolites in genome-scale metabolic models of the five organisms, enumerated above, even at no growth. To investigate the feasibility of gene-based strategies, we applied GeneReg to increase the production of metabolites participating in the tricarboxylic acid (TCA) cycle as well as amino acids present in the five models analyzed above (see Supplementary Tables S9 and S10). Looking at the structure of GPR associations in each organism, we find that more than a fifth of the rules involve at least two genes (Fig. 4). Therefore, this result suggests infeasibility of optimizing the production of metabolites at the core of central metabolism, like the intermediates of the TCA cycle.

To increase the likelihood of finding a solution, the tunable parameter $C$ is set to 0.005 and 0, allowing small flux changes to be

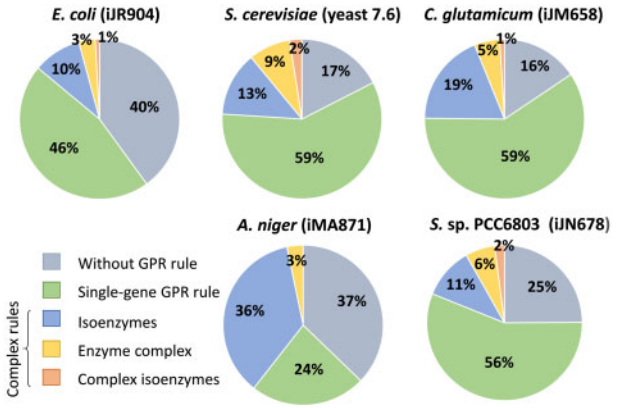


**Fig. 4.** The structure of GPR associations in five organisms used as cell factories. This figure summarizes the structure of GPR associations in the models of *E.coli* [iJR904 (Reed *et al.*, 2003)], *S.cerevisiae* [yeast 7.6 (Aung *et al.*, 2013)], the Gram-positive bacterium *C.glutamicum* [iJM658 (Mei *et al.*, 2016)], the filamentous fungus *A.niger* [iMA871 (Andersen *et al.*, 2008)) and the cyanobacterium *S.* sp. PCC6803 [iJN678 (Nogales *et al.*, 2012)]. More than a fifth of the rules involve at least two genes, which suggests infeasibility of optimizing the production of metabolites at the core of central metabolism

considered, and we place no limit on the total number of reaction manipulations. A metabolite is considered as an optimization objective, if there is room for improvement of the metabolite synthesis, i.e. the ratio of $v^{max}$ to $v^{U}$ is bigger than 1 for its sink (or exchange) reaction. The results show that feasible metabolic strategies at the gene level exist only for the increased production of malate, citrate, aspartate, cysteine, glutamine and methionine in the *E.coli* model and the increased production of pyruvate in *S.cerevisiae*, when the parameter $C$ is set to 0.005 and at least 20% of optimal growth is achieved (see Supplementary Tables S9 and S10).

We would like to emphasize that the metabolic engineering strategies from constraint-based modeling approaches depend strongly on the constraints and the model used for their design. For instance, the overproduction of pyruvate, succinate, serine and threonine is not feasible when the upper-bound of 20 mmol/(gDW h) for oxygen uptake is imposed and glucose (the default carbon source) is the sole carbon source used in the iJR904 *E.coli* model (see Supplementary Table S11). While GeneReg predicts that an increase in production of fumarate is not feasible under the previously mentioned constraints, the usage of glycerol as a carbon source allows the possibility for overproduction of this compound, in line with experimental evidence (Li *et al.*, 2014). By considering three different environments (see Supplementary Table S11), we found that only for three amino acids, i.e. tryptophan, tyrosine and valine, GeneReg reports infeasibility for increase of production by 10% in these

environments in contrast to feasible engineering strategies documented in the literature (Chávez-Béjar *et al.*, 2008; Panichkin *et al.*, 2016; Park *et al.*, 2007). However, closer look at this strategy point out they consider alteration (e.g. removal) or allosteric regulation effects of metabolites on enzymes, for which the constraint-modeling framework remains underdeveloped. In addition, the parameter $C$ also will have effect on the reported infeasibilities. For instance, when $C = 0$, GeneReg predicts that fumarate production can be increased by 20% even under default constraints of iJR904 *E.coli* model. Therefore, GeneReg demonstrates the need to couple gene regulation and metabolism to understand the limits of metabolic engineering under specific assumptions of the constraint-based modeling framework.

## 4 Conclusion

Here, we proposed GeneReg, a constraint-based approach based on MILP formulation to design gene manipulation strategies that facilitate under- and overexpression of genes at a genome-scale level. GeneReg formulation of up- and downregulation of reactions is motivated by OptReg, but is modified to account for differences between wild-type and mutant ranges. The novelty of GeneReg consists of constraints that ensures coupling of the reaction and gene manipulations as well as constraints that avoid conflict at the gene level. As a result, GeneReg facilitates the generation of manipulation strategies feasible at the gene level.

Our findings demonstrate that GeneReg is applicable with large-scale metabolic networks in model organisms as well as organisms used as cell factories. In contrast to the results about growth-coupling of the overproduction of almost every metabolite, we found that there is no strategy feasible at the gene level that leads to overproduction of any intermediates in the TCA cycle as well as amino acids present in the analyzed models (with few exceptions in *E.coli* and *S.cerevisiae*) under the default growth conditions. However, we also identify that changes in environmental cues (e.g. carbon used) has an effect on the proposed engineering strategies, in line with experimental evidence. Therefore, coupling of gene and reaction levels, along with context-dependent activity and allosteric regulation of enzymes, should be given further emphasis in future studies aimed at design of feasible and more accurate metabolic engineering strategies, and motivates the integration of gene regulatory and metabolic networks for improved prediction of complex traits.

## Funding

This work was supported by the MELICOMO project 031B0358B of the German Federal Ministry of Science and Education to Z.N.

*Conflict of Interest*: none declared.

## References

Andersen,M. *et al.* (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol. Syst. Biol.*, **4**, 178.

Aung,H. *et al.* (2013) Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol.*, **9**, 215–228.

Burgard,A. *et al.* (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.*, **84**, 647–657.

Chang,D. *et al.* (1999) Homofermentative production of D- or L-lactate in metabolically engineered *Escherichia coli* RR1. *Appl. Environ. Microbiol.*, **65**, 1384–1389.

Chávez-Béjar,M.I. *et al.* (2008) Metabolic engineering of *Escherichia coli* for L-tyrosine production by expression of genes coding for the chorismate mutase domain of the native chorismate mutase-prephenate dehydratase and a cyclohexadienyl dehydrogenase from *Zymomonas mobilis*. *Appl. Environ. Microbiol.*, **74**, 3284–3290.

Erb,T. *et al.* (2017) Synthetic metabolism: metabolic engineering meets enzyme design. *Curr. Opin. Chem. Biol.*, **37**, 56–62.

Kamp,A. and Klamt,S. (2017) Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms. *Nat. Commun.*, **8**, 15956.

Kim,J. and Reed,J. (2010) OptORF: optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst. Biol.*, **4**, 53.

Kim,J. *et al.* (2011) Large-scale bi-level strain design approaches and mixed-integer programming solution techniques. *PLoS One*, **6**, e24162.

King,Z. *et al.* (2016) BiGG models: a platform for integrating, standardizing, and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.

Li,N. *et al.* (2014) Engineering *Escherichia coli* for fumaric acid production from glycerol. *Bioresour. Technol.*, **174**, 81–87.

Machado,D. *et al.* (2016) Stoichiometric representation of gene-protein-reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS Comput. Biol.*, **12**, e1005140.

Maia,P. *et al.* (2015) In silico constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiol. Mol. Biol. Rev.*, **80**, 45–67.

Mei,J. *et al.* (2016) Reconstruction and analysis of a genome-scale metabolic network of *Corynebacterium glutamicum* S9114. *Gene*, **575**, 615–622.

Nogales,J. *et al.* (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc. Natl. Acad. Sci. USA*, **109**, 2678–2683.

Panichkin,V.B. *et al.* (2016) Metabolic engineering of *Escherichia coli* for L-tryptophan production. *Appl. Biochem. Microbiol.*, **52**, 783–809.

Park,J.H. *et al.* (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proc. Natl. Acad. Sci. USA*, **104**, 7797–7802.

Patil,K. *et al.* (2005) Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinform.*, **6**, 1–12.

Pharkya,P. and Maranas,C. (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.*, **8**, 1–13.

Pharkya,P. *et al.* (2004) OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.*, **14**, 2367–2376.

Ranganathan,S. *et al.* (2010) OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.*, **6**, e1000744.

Reed,J. *et al.* (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, **4**, R54.

Rocha,I. *et al.* (2010) OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.*, **4**, 1–12.

Zhang,C. *et al.* (2015) Logical transformation of genome-scale metabolic models for gene level applications and analysis. *Bioinformatics*, **31**, 2324–2331.