*Bioinformatics*, 33(1), 2017, 87–94
doi: 10.1093/bioinformatics/btw557
Advance Access Publication Date: 7 September 2016
Original Paper

OXFORD

Systems biology

# BeReTa: a systematic method for identifying target transcriptional regulators to enhance microbial production of chemicals

**Minsuk Kim[1], Gwanggyu Sun[1], Dong-Yup Lee[2,3] and Byung-Gee Kim[1,*]**

[1]School of Chemical and Biological Engineering, Institute of Molecular Biology and Genetics, and Bioengineering Institute, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 151-742, Republic of Korea, [2]Department of Chemical and Biomolecular Engineering, National University of Singapore, 4 Engineering Drive 4, Singapore 117576, Singapore and [3]Bioprocessing Technology Institute; Agency for Science, Technology and Research (A*STAR), 20 Biopolis Way, #06-01, Centros, Singapore 138668, Singapore

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on May 3, 2016; revised on August 4, 2016; accepted on August 21, 2016

## Abstract

**Motivation:** Modulation of regulatory circuits governing the metabolic processes is a crucial step for developing microbial cell factories. Despite the prevalence of *in silico* strain design algorithms, most of them are not capable of predicting required modifications in regulatory networks. Although a few algorithms may predict relevant targets for transcriptional regulator (TR) manipulations, they have limited reliability and applicability due to their high dependency on the availability of integrated metabolic/regulatory models.

**Results:** We present BeReTa (Beneficial Regulator Targeting), a new algorithm for prioritization of TR manipulation targets, which makes use of unintegrated network models. BeReTa identifies TR manipulation targets by evaluating regulatory strengths of interactions and beneficial effects of reactions, and subsequently assigning beneficial scores for the TRs. We demonstrate that BeReTa can predict both known and novel TR manipulation targets for enhanced production of various chemicals in *Escherichia coli*. Furthermore, through a case study of antibiotics production in *Streptomyces coelicolor*, we successfully demonstrate its wide applicability to even less-studied organisms. To the best of our knowledge, BeReTa is the first strain design algorithm exclusively designed for predicting TR manipulation targets.

**Availability and Implementation:** MATLAB code is available at https://github.com/kms1041/BeReTa (github).

**Contact:** byungkim@snu.ac.kr

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

## 1 Introduction

During the past decades, systems metabolic engineering has enabled the enhanced microbial production of various chemicals for the development of sustainable processes (Becker and Wittmann, 2015). Various strategies have been suggested, and serially and/or iteratively applied to improve production yields, titers, and productivities in order to meet industrial constraints (Lee *et al.*, 2012). A modification of regulatory circuits (e.g. removal/suppression of negative feedback regulations and upregulation of biosynthetic pathway activators) is one of the important strategies for developing production strains (Lee and Kim, 2015). Indeed, more than half of the genetic manipulations in engineered strains of *Escherichia coli* and yeast belong to the

modifications of regulatory networks, while most of them are employed by human intuitions (Winkler *et al.*, 2015).

Interestingly, fueled by the development of constraint-based models and phenotype prediction methods such as flux balance analysis, a number of *in silico* strain design algorithms have been developed to provide novel non-intuitive genetic designs (Maia *et al.*, 2016). Since the development of OptKnock (Burgard *et al.*, 2003) which is the first computational strain design algorithm for predicting only gene deletion targets, recent methods have been evolved to perform diverse designing tasks in more efficient manners. For example, OptStrain (Pharkya *et al.*, 2004) was developed to account for heterologous gene insertions; OptForce (Ranganathan *et al.*, 2010) was introduced to consider modulations of gene expression levels in addition to gene deletions. Most recently, such algorithms were further improved to exploit transcriptomic data for identifying genetic targets (Kim *et al.*, 2016). Showing the validity and usefulness of the computational strain design algorithms, several patents referring to the algorithms have now become available (Maia *et al.*, 2016).

However, despite increasing popularities of *in silico* strain design methods, most of the genetic modification targets proposed are confined to metabolic genes. Previously, two groups of researchers have tackled this issue by developing algorithms which can predict transcriptional regulator (TR) manipulation targets by using integrated models of metabolism and regulation (Kim and Reed, 2010; Vilaca *et al.*, 2011). However, the use of integrated metabolic/regulatory models has innate limitations. First, the integrated models are available only for few best-studied organisms, e.g. *E.coli* (Covert *et al.*, 2004) and *Saccharomyces cerevisiae* (Herrgard *et al.*, 2006), so that their applicability is greatly limited to those model microbes. Second, they assume complete on/off behavior of target genes/reactions according to Boolean logic representation of gene regulation in the integrated models, regardless of differential control strengths of TR-target interactions that exist in nature.

To overcome such limitations, in this study, we present an entirely new approach for predicting TR manipulation targets without requiring a well-defined integrated model of metabolism and regulation. The *Be*neficial *Re*gulator *Ta*rgeting (BeReTa) algorithm introduced herein differentiates the variable regulatory strengths of TR-target interactions, and thus is able to rank the effects of TR manipulations on target chemical production. The algorithm was developed to answer the following questions in metabolic engineering: (i) Which TR manipulation is the most effective when several TRs are involved in regulations of product biosynthesis? (ii) Is it helpful or not to manipulate a global/pleiotropic TR that has numerous target genes? Its wide applicability was successfully demonstrated via two case studies for *E.coli* and *Streptomyces coelicolor*.

## 2 Methods

### 2.1 BeReTa algorithm

The BeReTa algorithm calculates beneficial scores of TRs and their significance for the production of target chemicals as illustrated in Figure 1. Beneficial scores are calculated from the regulatory strength matrix and the flux slope vector defined below, using the models for the transcriptional regulatory network (TRN) and the genome-scale metabolic network (GSMN) (Fig. 1A).

#### 2.1.1 Regulatory strength matrix

The regulatory strength matrix (RS) is a $m \times n$ matrix of regulatory strength coefficients for a set of $m$ TRs in the TRN and a set of $n$ reactions in the GSMN. To construct the regulatory strength matrix, structures of both TRN and GSMN are required, together with gene expression compendium data. Firstly, the regulatory strength of each TR-gene interaction ($RS_{ij}$ denotes the interaction between TR $i$ and gene $j$) in the TRN is defined as follows:

$$RS_{ij} = \beta_{ij} \cdot |r_{ij}| \quad (1)$$

where $r_{ij}$ is Pearson's correlation coefficient for the gene expression profiles of TR $i$ and gene $j$, and $\beta_{ij}$ is a sign of regulation which is set to $+1$ and $-1$ for activating and repressing interactions, respectively. Then, the TR-gene regulatory strength ($RS_{ij}$) is mapped to generate the TR-reaction regulatory strength ($RS_{ik}$ for TR $i$ and reaction $k$) by using the gene–protein-reaction (GPR) association rules in the GSMN. An example of the GPR mapping process involving an enzyme complex is illustrated in Figure 1B. For instance, if TR $i$ controls multiple genes ($j$ and $j'$) that constitute an enzyme complex which catalyzes reaction $k$, the TR-gene regulatory strengths ($RS_{ij}$ and $RS_{ij'}$) are averaged to yield the regulatory strength coefficient for TR-reaction ($RS_{ik}$). Similarly, for a TR controlling multiple isozymes for a reaction, TR-gene regulatory strengths for isozymes are also averaged to yield a TR-reaction regulatory strength. It should be noted that considering average operator rather than minimum (for enzyme complexes) and maximum (for isozymes) operators, which are often used for mapping transcriptomic or proteomic data, is more appropriate for mapping regulatory strengths. First of all, a gene with minimal or maximal regulatory strengths cannot be regarded as decisive gene for reaction activity. In addition, all the regulatory effects of a TR on metabolic gene expression can be considered simultaneously by using average operator, still reserving any non-minimal and non-maximal values.

Note that the regulatory strength defined herein is an overall regulatory strength of a TR on a gene/reaction across various nutritional and environmental conditions where gene expression profiling experiments were conducted. The gene expression datasets for diverse conditions could provide more relevant estimates of regulatory strengths than using the data only from specific conditions. For example, consider an anaerobic TR and its target genes which are all active under anaerobic conditions. When only examining gene expression data for anaerobic conditions, correlations between the TR and target genes are not evident since they are always active and not differentially expressed for the given datasets. However, the datasets from both anaerobic and aerobic conditions make the correlations evident as the expression patterns of the TR and target genes change together. Therefore, various data in gene expression compendium should be used to build the regulatory strength matrix which in result does not represent condition-specific regulatory strengths of the TRs. It should also be noted that the interactions and hierarchies of TRs have been neglected for simplification. However, the effects of TR–TR interactions and hierarchies might be partially reflected in the regulatory strength matrix through the use of gene expression compendium.

#### 2.1.2 Flux slope vector

The flux slope vector ($q_{slope}$) is a vector of $n$ flux slopes, which quantitatively describes the beneficial effects of the $n$ reactions in the GSMN for target chemical production. The concept of the flux slope and its calculation procedure are adopted from F(V)SEOF which is an algorithm for searching metabolic gene overexpression targets (Choi *et al.*, 2010; Park *et al.*, 2012). Firstly, the following flux balance analysis (Orth *et al.*, 2010) problems, formulated as linear

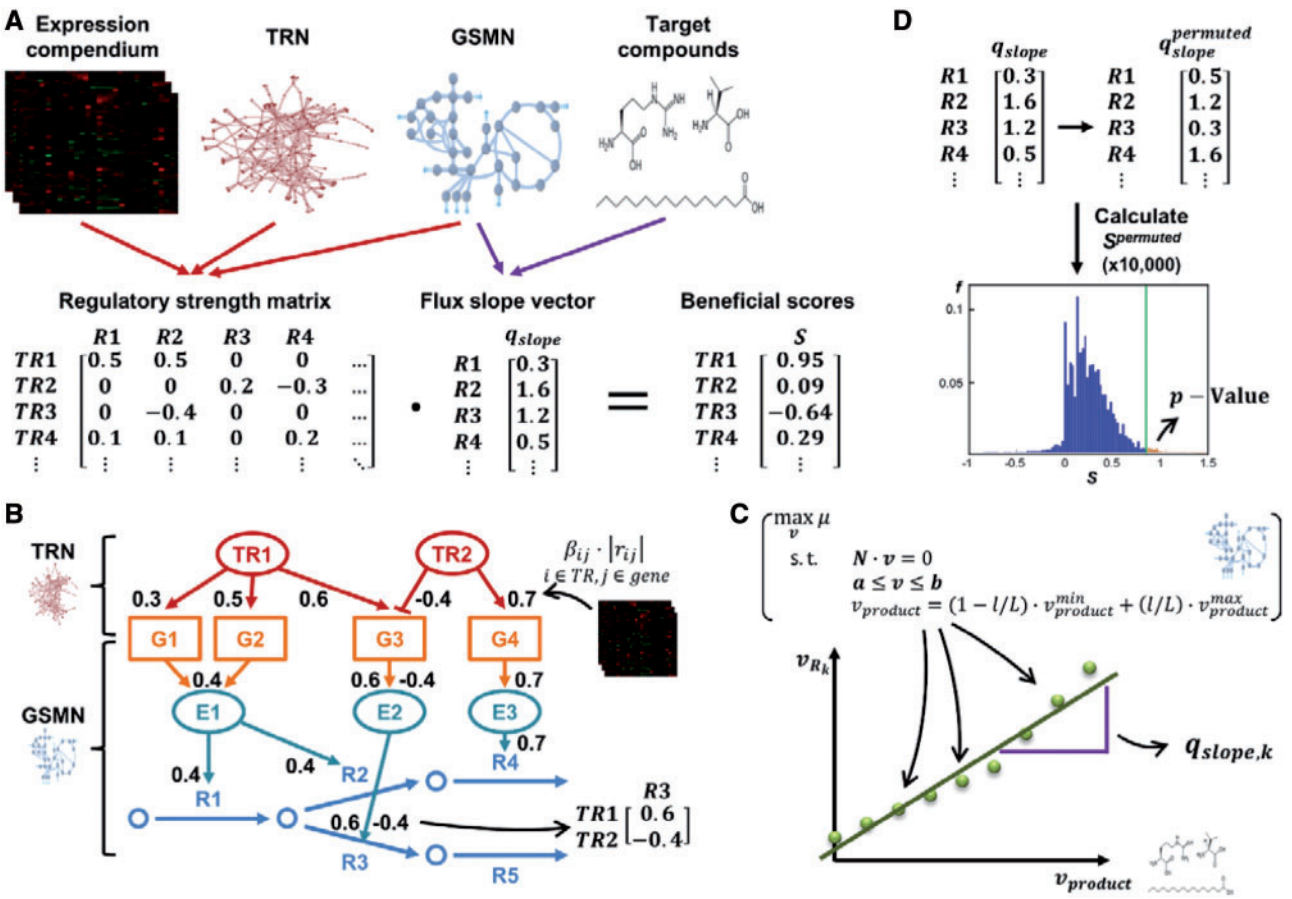


**Fig. 1.** Schematic representation of BeReTa algorithm. **(A)** BeReTa uses models of TRN and GSMN together with expression compendium data to calculate beneficial scores of TRs for the production of target compounds. Beneficial scores can be calculated from multiplication of the regulatory strength matrix and the flux slope vector. **(B)** The regulatory strength matrix is calculated by the mapping of regulatory strength coefficients through the structures of TRN and GSMN (TR, transcriptional regulator; G, gene; E, enzyme; R, reaction). **(C)** The flux slope vector is calculated by using flux balance analysis with constraints for different levels of target chemical production. Linear regression is used to estimate flux slopes. **(D)** A permutation test was introduced to calculate the significance of the beneficial scores

programming (LP) problems, are solved serially for different values of integer $l$ ($l = 0, 1, 2, \ldots, L$):

$$\max_{v} \mu$$

subject to

$$N \cdot v = 0 \tag{2}$$

$$a \leq v \leq b \tag{3}$$

$$v_{\text{product}} = \left(1 - \frac{l}{L}\right) \cdot v^{\min}_{\text{product}} + \left(\frac{l}{L}\right) \cdot v^{\max}_{\text{product}} \tag{4}$$

where $N$ is the stoichiometric matrix derived from the GSMN, $v$ is a flux vector, $a$ is a vector of lower flux bounds, and $b$ is a vector of upper flux bounds. In the standard flux balance analysis, an objective function is maximized or minimized under the mass balance constraints given by Equation (2), together with thermodynamic and mechanistic constraints in Equation (3). For calculating the flux slope, the rate of biomass formation ($\mu$) is maximized as an objective function with the additional constraint in Equation (4) which enforces a fixed flux through the product reaction ($v_{\text{product}}$). The flux through the product reaction is increased from its minimal ($l = 0$) to maximal ($l = L$) values, so that $L + 1$ LP problems should be solved iteratively. For each LP problem, flux vectors, which satisfy the optimality, are further selected for parsimonious enzyme usage (Lewis *et al.*, 2010). Then, flux slopes are obtained by linear regression between absolute reaction fluxes and product fluxes from the $L+1$ flux vectors with varying degrees of product fluxes (Fig. 1C). Finally, the flux slope vector is constructed by replacing the negative values of the flux slopes with zeros to only take into account of beneficial effects of reactions on target chemical production, and disregard negative effects of reactions on cell growth. This procedure is required to avoid the prediction of biased targets toward growth-associated TRs since most of the growth-associated reactions have large negative flux slopes regardless of types of target products. Therefore, the negative values of the flux slopes should be substituted to predict relevant TR targets specific to the target product.

By the definition of flux slope, increasing fluxes through the reactions with high flux slopes is necessary, although often not sufficient, for increasing product flux. Note that the reactions with large flux slopes can be more beneficial for increasing product flux, as they are more sensitive to the enforced product flux than the reactions with smaller flux slopes (Park *et al.*, 2012). In addition, it is important to note that the flux slope vector is condition-specific since some of the nutritional and environmental conditions such as substrate uptake rates and oxygen availability are included in the

lower and upper flux bounds, $\boldsymbol{a}$ and $\boldsymbol{b}$, in Equation (3). Minimal and maximal product fluxes were calculated by flux balance analysis with objective functions that minimize and maximize product formation, and $L = 20$ was the value used for the study. Gurobi (Gurobi Optimization, http://www.gurobi.com) was used as the optimization solver for LP problems.

#### 2.1.3 Beneficial score

Beneficial score ($S$) is defined as follows:

$$S_i = \sum_k \mathrm{RS}_{ik} \cdot q_{\mathrm{slope},k} = \left[\mathrm{RS} \cdot q_{\mathrm{slope}}\right]_i \quad (5)$$

The beneficial score for TR $i$ ($S_i$) is the sum of the products of regulatory strengths ($RS_{ik}$) and flux slopes ($q_{\mathrm{slope},k}$) of its target reactions. Simply, multiplication of the regulatory strength matrix (RS) and the flux slope vector ($q_{\mathrm{slope}}$) yields beneficial scores for all TRs in the TRN for target chemical production (Fig. 1A). Clearly, a TR with high regulatory strengths for its target reactions with high flux slopes will get high beneficial scores in Equation (5).

A transcriptional activator (repressor) which regulates at least one beneficial reaction, i.e. the reaction with positive flux slope, should have positive (negative) beneficial score, since the sign of beneficial score is only dependent on the signs of regulatory strengths which are all positive (negative) for transcriptional activator (repressor). For a dual regulator which has both activating and repressing interactions, beneficial score will be positive (negative) if its activating (repressing) effects are stronger than repressing (activating) effects. Therefore, a TR with positive (negative) beneficial score can be considered as an overexpression (knockout or downregulation) target.

Note that manipulating TRs with large absolute beneficial scores would be more effective for increasing product flux than manipulating TRs with smaller scores for the same target product. However, comparison of beneficial scores of TRs for different target products is often meaningless since theoretical maximum beneficial scores for each chemical, a case where regulatory strengths are equal to one for all reactions meaning that a TR manipulation is maximally effective for increasing product flux, are largely different. It is also noteworthy to mention that the beneficial score defined herein does not have any further biologically meaningful interpretation as we employed correlation coefficients to define the regulatory strengths.

#### 2.1.4 Permutation test

The beneficial score defined in Equation (5) is not normalized by the number of targets of the TR, and thus TRs with many targets, i.e. global regulators, usually receive higher scores than the TRs with a small number of targets. Consider two TRs, that one is a global TR controlling 100 target reactions among which only two reactions are beneficial while the other is a local TR controlling only one reaction which is beneficial. If the values of regulatory strengths and flux slopes are identical for both cases, the beneficial score of the global TR will be twice as large as that of the local TR. However, it is evident that the local TR is more likely to be related to target chemical production than the global TR. Furthermore, perturbations in the 98 non-beneficial reactions by manipulating the global TR might induce great detrimental effects on cell growth and target chemical production.

To deal with this problem, a permutation test was introduced to calculate the significance of the beneficial scores (Fig. 1D). The flux slope vector ($q_{\mathrm{slope}}$) was permuted 10 000 times to yield 10 000 permuted flux slope vectors ($q_{\mathrm{slope}}^{\mathrm{permuted}}$), and the corresponding 10 000 permuted beneficial scores ($S^{\mathrm{permuted}}$) were obtained. The $P$-value was defined as the number of permuted beneficial scores greater (smaller) than the non-permuted beneficial score divided by the number of permutations for TRs with positive (negative) beneficial scores. Note that only gene-associated reactions were considered for permutation, i.e. exchange reactions and orphan reactions were excluded from the permutation.

#### 2.1.5 Target criteria

We set up the following criteria for the selection of TR manipulation targets for target chemical production.

1. The TR should have a non-zero beneficial score.
2. The $P$-value of the beneficial score should be $<0.05$.
3. The TR should have two or more effective (beneficial) gene/reaction targets.
4. At least 10% of the target metabolic genes of the TR should be beneficial, i.e. have positive flux slopes.

The third criterion was introduced to select meaningful TR manipulations which can simultaneously change the expression of multiple genes and reactions, and thus could perform better than single enzyme manipulations. The fourth criterion was applied to rule out TRs which regulates mostly non-beneficial reactions to prevent unexpected detrimental effects on cell growth and target chemical production. Finally, TRs with positive (negative) beneficial scores were designated as overexpression (knockout/downregulation) targets if the TRs had passed all the criteria.

### 2.2 Network models and datasets

Using the BeReTa algorithm, we predicted TR manipulation targets for the production of various chemicals in *E.coli*, and antibiotics in *S.coelicolor*. The network models and datasets used for the predictions are summarized in Table 1.

#### 2.2.1 Genome-scale metabolic network

The most recent version of the *E.coli* GSMN, *i*JO1366 (Orth *et al.*, 2011), was used for making BeReTa predictions. To simulate the production of non-native chemicals in *E.coli*, experimentally validated heterologous pathway reactions were obtained from literature, and implemented into *i*JO1366 (Supplementary Table S1). For simulating antibiotics production in *S.coelicolor*, a recently published high-quality GSMN for *S.coelicolor*, *i*MK1208 (Kim *et al.*, 2014), was employed. The glucose uptake rate was constrained to

**Table 1.** A summary of network models and expression datasets used for this study

| | *E.coli* | *S.coelicolor* |
|---|---|---|
| GSMN | *i*JO1366 | *i*MK1208 |
| Included genes (ORF coverage) | 1366 (32%) | 1208 (15%) |
| Reactions | 2251 | 1643 |
| Metabolic/transport | 1473/778 | 1443/200 |
| Gene associated/no gene associated/spontaneous | 2088/128/35 | 1376/238/29 |
| Exchange reactions | 332 | 216 |
| Metabolites | 1136 | 1246 |
| TRN | RegulonDB | Correlation network |
| Included TRs | 180 | 389 |
| Regulatory interactions | 3672 | 10 000 |
| Expression compendium | COLOMBOS | COLOMBOS |
| Genes | 4294 | 7767 |
| Conditions | 2470 | 371 |

10 mmol gDCW$^{-1}$ h$^{-1}$ to simulate chemical production from glucose. Additional details on the model simulations are available in Supplementary Table S1.

#### 2.2.2 Transcriptional regulatory network

The *E.coli* TRN was obtained from RegulonDB (Gama-Castro *et al.*, 2016), version 9.0 in October 2015 (http://regulondb.ccg.unam.mx). A total of 3672 interactions with both strong and weak confidence levels involving 180 single TRs were used for the predictions. RegulonDB also provides information on the regulatory modes of the interactions, such as activator, repressor, dual or unknown. For dual and unknown modes of interactions, regulatory modes were reassigned as activator or repressor according to the sign of Pearson's correlation coefficients between the gene expression profiles of the TR and the target gene.

Although RegulonDB is a curated primary reference database for the regulatory network of the best-studied organism, such a database does not exist for *S.coelicolor*. Therefore, we generated a relevance network for *S.coelicolor* using the expression compendium. The inferred TRN consists of 10 000 edges that are ranked based on the absolute values of Pearson's correlation coefficients, which is a method shown to perform fairly well (Marbach *et al.*, 2012). 389 TRs in total are included in the inferred *S.coelicolor* TRN.

#### 2.2.3 Expression dataset

The gene expression compendia for *E.coli* and *S.coelicolor* were obtained from the COLOMBOS database (Meysman *et al.*, 2014), version 2.0 in October 2015 (http://www.colombos.net). The COLOMBOS database provides re-normalized cross-platform gene expression compendia which are derived from public resources including Gene Expression Omnibus and ArrayExpress. The sizes of the expression compendia are summarized in Table 1.

## 3 Results

### 3.1 Production of chemicals in *E.coli*

*E.coli* is one of the most frequently employed organisms for the industrial production of various chemicals owing to its well-characterized metabolism and regulation, and ease of genetic modification and maintenance. Accordingly, *E.coli* is the best option for demonstrating and validating the predictive power of the BeReTa algorithm, as there exist high-quality models of metabolism and regulation which enable the reliable prediction of TR targets, and substantial literature describing the metabolic engineering strategies for the validation. Using the network models and datasets for *E.coli* listed in Table 1, TR manipulation targets for 22 native and 16 non-native compounds were identified by BeReTa. Four top-scoring TR manipulation targets for each chemical ranked by the absolute values of beneficial scores are shown in Figure 2. Detailed results including the values of beneficial scores and *P*-values are available in Supplementary Table S2.

Validating the predictions of the BeReTa algorithm, metabolic engineering strategies, which include the predicted TR manipulations, could be found for 12 of 38 simulated chemicals covering both native (7 of 22) and non-native (5 of 16) compounds. Although the majority of validated TR targets are involved in the biosynthesis

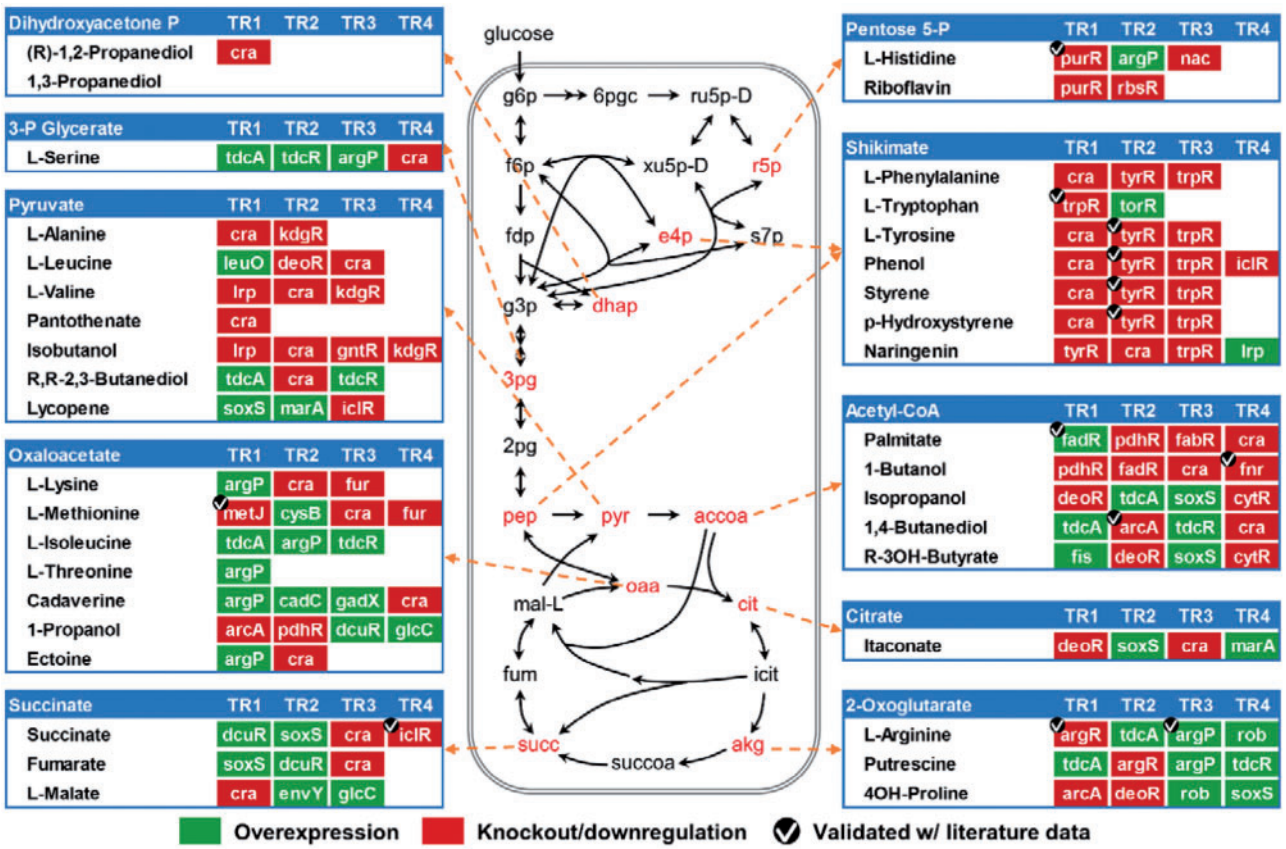


**Fig. 2.** BeReTa predictions for the production of various chemicals in *E.coli*. Four top-scoring TR manipulation targets for each chemical ranked by absolute values of beneficial scores are shown. TRs in green boxes are overexpression targets, whereas TRs in red boxes are knockout or downregulation targets. Manipulations validated with literature data are indicated with check marks

of amino acids and their simple derivatives, previous studies supporting the predictions for other classes of chemicals are also available. For example, *fadR* overexpression was correctly predicted as the top-scoring strategy for fatty acid overproduction (Zhang *et al.*, 2012). Furthermore, *fnr* deletion and *arcA* deletion were also properly suggested for the overproduction of two non-native compounds from acetyl-CoA, 1-butanol and 1,4-butanediol, respectively (Atsumi *et al.*, 2008; Yim *et al.*, 2011). These examples also clearly demonstrated that BeReTa algorithm can appropriately evaluate the effects of manipulating global/pleiotropic TRs.

The most promising example that shows the predictive power and utility of BeReTa is the case study of L-arginine production. As shown in Figure 2, BeReTa identified ArgR, TdcA, ArgP, and Rob as TR manipulation targets for increasing arginine production. The compositions of beneficial scores for the regulators, i.e. associated flux slopes and regulatory strengths, along with arginine pathway structure are given in Supplementary Figure S1. ArgR and ArgP are directly related to the arginine biosynthesis as ArgR controls entire arginine biosynthetic pathway while ArgP controls arginine secretion and glutamate formation. Meanwhile, TdcA and Rob are involved in precursor supply as TdcA controls generation of acetyl-CoA, and Rob controls TCA cycle for providing glutamate and pentose phosphate pathway for NADPH supply. Interestingly, among these manipulations, *argR* deletion and *argP* upregulation can be found in an arginine-producing *E.coli* strain which was recently developed by Ginesy and coworkers (Ginesy *et al.*, 2015). As wild-type *E.coli* strains do not excrete any arginine, they first constructed a base strain for arginine production by deleting *argR* and three other genes (*adiA*, *speC* and *speF*) which are responsible for degrading arginine and ornithine. To further increase arginine production, they overexpressed feedback-resistant, thus constitutively active, *argP* and *argA* (N-acetylglutamate synthase) alleles, thereby engineering the final arginine-overproducing strain with high productivity. The comparison of these experimental results and our predictions shows that the BeReTa algorithm could not only identify multiple valid TR manipulation targets, but also assign higher scores (ranks) to more significant TR manipulations. Moreover, BeReTa provides an easier way to discover novel metabolic engineering strategies such as *argP* overexpression, which was originally identified through expensive and labor-intensive classical random mutagenesis study.

Finally, it is noteworthy to mention that the predicted TR manipulations for the products synthesized from the same precursors are similar. For example, TyrR deletion was consistently predicted for the production of shikimate-derived chemicals except tryptophan. Accordingly, actual metabolic designs including TyrR deletion or inactivation can be found for four of the chemicals, tyrosine, phenol, styrene, and *p*-hydroxystyrene (Supplementary Table S1). Likewise, ArgP upregulation for oxaloacetate-based chemicals and PurR deletion/inactivation for ribulose 5-phosphate-based chemicals were predicted. Interestingly, no TR manipulation targets were identified for dihydroxyacetone phosphate-derived chemicals, except Cra downregulation for 1,2-propanediol production.

### 3.2 Production of antibiotics in *S.coelicolor*

*Streptomyces* are biotechnologically important bacteria, as they produce various secondary metabolites including many of the currently used antibiotic drugs (Bentley *et al.*, 2002). To examine whether BeReTa can be applied for the identification of TR manipulation targets in such less-studied organisms, we applied BeReTa for the production of antibiotics in *S.coelicolor*, a model species of *Streptomyces*. Here, it should be noted that the correlation network derived from the expression compendium data was used for making the predictions, since a high-quality TRN model does not exist for this organism (Table 1). The predicted TR targets for three antibiotics produced by *S.coelicolor*, actinorhodin (ACT), undecylprodiginine (RED), and calcium-dependent antibiotic (CDA), are shown in Figure 3. Details of the predictions, including beneficial scores, *P*-values and lists of target reactions are available in Supplementary Table S3.

BeReTa identified both known and novel TR manipulation targets for increasing the production of antibiotics in *S.coelicolor*. First, cluster-situated regulators (CSRs) were correctly suggested as top-scoring overexpression targets. Overexpression of ActII-ORF (SCO5085) for ACT production and RedD (SCO5877), RedZ (SCO5881) for RED production were predicted. However, CdaR, a CSR for the CDA biosynthetic gene cluster, was not identified as a TR manipulation target for CDA production. Interestingly, RedD was also predicted as an overexpression target for increasing ACT production, which might indicate cross-regulation of CSRs across different secondary metabolite biosynthetic gene clusters (Huang *et al.*, 2005).

AbsR1 (SCO6992) is a novel overexpression target predicted by BeReTa for increasing both ACT and RED productions. AbsR1 is a fascinating target since it can control both primary and secondary metabolisms (Supplementary Table S3). Furthermore, there is also a report describing the effects of AbsR1 deficiency on the reduced production of ACT and RED (Park *et al.*, 2000). Therefore, it will be interesting to experimentally investigate the effects of AbsR1 overexpression as well.

### 3.3 Comparison of BeReTa with OptORF

We compared BeReTa with OptORF (Kim and Reed, 2010), an existing method which can predict TR manipulation targets using integrated metabolic/regulatory models. A key difference between BeReTa and OptORF is that OptORF relies on actual flux prediction for TR mutants by using the integrated models to find target genes, whereas BeReTa focuses on using statistical correlations and flux slopes as proxy for effects of TR manipulations instead of predicting flux distribution in TR mutants. As a result, BeReTa has

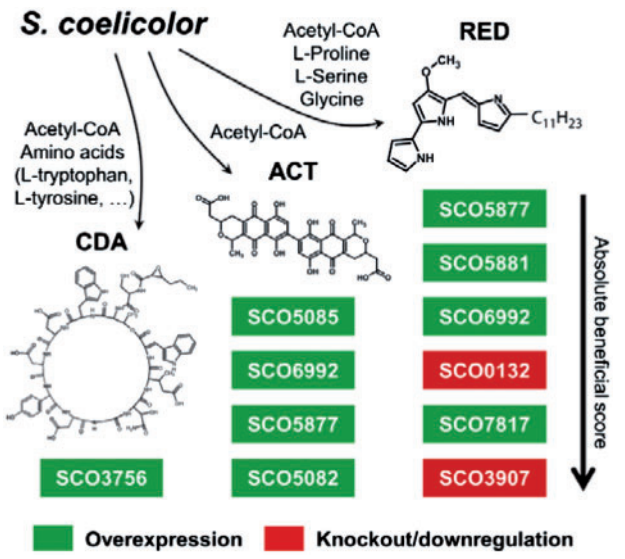


**Fig. 3.** BeReTa predictions for the production of antibiotics in *S.coelicolor*. *S.coelicolor* produces actinorhodin (polyketide, ACT), undecylprodiginine (prodiginine, RED), and calcium-dependent antibiotic (nonribosomal peptide, CDA). Predicted TR targets were sorted by absolute values of beneficial scores. TRs in green boxes are overexpression targets, whereas TRs in red boxes are knockout or downregulation targets

**Table 2.** A comparison of BeReTa with OptORF

| | OptORF (Kim and Reed, 2010) | | BeReTa (This study) | |
|---|---|---|---|---|
| Requirements | Integrated metabolic and regulatory model | | **Unintegrated metabolic and regulatory model** Multiple gene expression data | |
| Type of prediction | Knockout of TRs **together with knockout and upregulation of metabolic genes** | | Knockout/downregulation and **upregulation of TRs** | |
| Type of optimization problem | MILP | | **LP** | |
| Example of prediction | Strategy | Yield | Strategy | Beneficial score |
| Anaerobic ethanol production in *E.coli* | **Δ*arcA*** Δ*pgi* | 83.5% | **Δ*pdhR*** | −0.668 |
| | **Δ*fnr* Δ*gntR*** Δ*pflB* Δ*tdcE* Δ*tpiA* | 86.2% | (↑)***fur*** | 0.355 |
| | **Δ*fnr*** Δ*pflB* Δ*tdcE* Δ*pgi* (↑)*edd* | 86.2% | **Δ*cra*** | −0.186 |
| | **Δ*fnr*** Δ*pflB* Δ*tdcE* Δ*pgi* Δ*ptsH* (↑)*edd* (↑)*fbp* | 90.4% | **Δ*gntR*** | −0.114 |
| | **Δ*arcA*** Δ*pta* Δ*eutD* Δ*tpiA* Δ*ptsH* (↑)*edd* | 91.6% | **Δ*kdgR*** | −0.111 |
| Anaerobic isobutanol production in *E.coli* | Δ*adhE* **Δ*gntR*** Δ*pgi* | 93.8% | **Δ*gntR*** | −0.054 |
| | Δ*adhE* Δ*pntA* Δ*gdhA* (↑)*edd* (↑)*fbp* | 95.5% | **Δ*kdgR*** | −0.052 |

The advantageous features and predicted TR targets are highlighted in bold. The gene targets for upregulation are denoted with (↑) symbol.

several advantages over OptORF (Table 2). First, BeReTa has wider applicability than OptORF as it can use unintegrated models of metabolism and regulation while OptORF requires the integrated one. Reconstruction of integrated metabolic/regulatory model demands extensive manual work and additional information for generating a well-defined set of Boolean logical rules for genes in TRN (Vivek-Ananth and Samal, 2016). Defining the Boolean logical rules is extremely difficult and time-consuming process which requires biochemical and genetic knowledge such as TR hierarchies, effector molecules for TRs, and phenotype of TR mutants. Thus, integrated models are only available for very few best-studied model organisms. Obviously, thus, OptORF is not applicable for *S.coelicolor* at this moment. Second, BeReTa has unique ability to predict TR targets for upregulation. Meanwhile, OptORF is not able to predict TR overexpression targets since it uses Boolean approximation which assumes only two states of the genes, presence and absence. Lastly, BeReTa needs far less computational power than OptORF as BeReTa only solves LP problems while OptORF solves mixed-integer linear programming (MILP) problems.

However, BeReTa also has some drawbacks compared with OptORF. First, BeReTa requires multiple gene expression data as an additional input. Therefore, the applicability of BeReTa relies on sufficient amount and good quality of data for inferring regulatory strengths. In addition, BeReTa is not able to provide product flux or yield expected for the mutant, as opposed to OptORF. Finally, BeReTa cannot predict combined effects of manipulating multiple TRs and metabolic genes, while OptORF can as shown in the examples of the predictions in Table 2. Nevertheless, BeReTa provides more options for TR manipulations including TR targets for upregulation, whereas OptORF predicts TR targets among the set of frequently found TR deletion strategies (Kim and Reed, 2010).

The other method (Vilaca *et al.*, 2011), which is also able to predict TR manipulation targets using integrated models, is similar to OptORF while it uses meta-heuristics to perform the task, and has similar features to OptORF.

## 4 Discussion

To shift the microbial metabolism towards productions of a desired chemicals, a system-wide engineering of regulatory and metabolic networks is required. Advances in modern systems biology tools, e.g. high-throughput sequencing techniques and automated genome annotation tools, have enabled the comprehensive reconstruction of biological networks for diverse microbes. Computational strain design algorithms were then developed to make use of such reconstructed networks for the system-wide identification of engineering targets. However, until now, only metabolic networks are commonly used by the algorithms except for a few cases (King *et al.*, 2015). In this study, we devised the BeReTa algorithm to fully exploit both metabolic and regulatory networks for identifying important TR manipulation targets. Through the *E.coli* and *S.coelicolor* case studies, we demonstrated that BeReTa can identify both known and novel TR manipulation targets for the production of various chemicals. It can not only predict and rank plausible TR manipulation targets but also properly evaluate the effect of global/pleiotropic TR manipulations on the production of desired chemical. To the best of our knowledge, BeReTa is the first strain design algorithm exclusively designed for predicting TR manipulation targets. As we have identified various TR manipulation targets for a variety of chemicals, it will be interesting to experimentally investigate the effects of such manipulations.

Integrated metabolic/regulatory models together with appropriate simulation methods can be used for predicting beneficial genetic modifications at the regulatory level for metabolic engineering (Imam *et al.*, 2015). However, integrated metabolic/regulatory models are only available for very few organisms owing to the extensive manual work and information required to define Boolean regulatory logic for each regulatory interaction. Although probabilistic regulation of metabolism (PROM) (Chandrasekaran and Price, 2010) had been developed for the automatic generation of integrated metabolic/regulatory models, PROM models are still not appropriate for simulating TR overexpression and cannot handle activating and repressing interactions simultaneously. Considering these limitations, BeReTa was designed to avoid the use of integrated metabolic/regulatory models. As a result, BeReTa can be widely applicable for not only the best-studied but also less-studied organisms as long as enough transcriptomic data do exist, and it accounts for both activating and repressing interactions for predicting overexpression and knockout/downregulation targets.

Despite the success of our new algorithm, performance of BeReTa can be further improved in some aspects. Firstly, regulatory strength can be defined in different ways rather than using Pearson's correlation coefficients. For example, the connectivity strength calculated from network component analysis (Liao *et al.*, 2003) can be an

alternative option to test. Secondly, the coupling of BeReTa and the state-of-the-art methods for inferring TRNs (Marbach *et al.*, 2012) would provide somewhat different, but improved predictions. For the best-studied organisms, use of the inferred TRN might provide different hypotheses that are based on novel inferred interactions which are not documented in primary reference databases of regulatory networks. On the other hand, for the less-studied organisms, using TRNs inferred by better inference methods would result in better predictions than the simple and well-performing correlation-based method used in this study. Finally, the current version of BeReTa is not able to predict combinatorial effects of simultaneous manipulations of TR and metabolic genes, or multiple TR targets. Future work may enable the prediction of a set of TR and metabolic gene targets with a maximal combinatorial effect on target chemical production.

## Funding

This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A2A2A01069197, NRF-2014K1A3A1A20034749). M.K. was also supported by the project of Global PhD Fellowship which NRF conducts from 2012 (NRF-2012H1A2A1001956). D.Y.L. was supported by the Synthetic Biology Initiative (DPRT/943/09/14) and the Academic Research Fund (R-279-000-476-112) of the National University of Singapore.

*Conflict of Interest*: none declared.

## References

Atsumi,S. *et al.* (2008) Metabolic engineering of *Escherichia coli* for 1-butanol production. *Metab. Eng.*, **10**, 305–311.

Becker,J. and Wittmann,C. (2015) Advanced biotechnology: metabolically engineered cells for the bio-based production of chemicals and fuels, materials, and health-care products. *Angew. Chem. Int. Ed. Engl.*, **54**, 3328–3350.

Bentley,S.D. *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.

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

Chandrasekaran,S. and Price,N.D. (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA*, **107**, 17845–17850.

Choi,H.S. *et al.* (2010) In silico identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microb.*, **76**, 3097–3105.

Covert,M.W. *et al.* (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.

Gama-Castro,S. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, **44**, D133–D143.

Ginesy,M. *et al.* (2015) Metabolic engineering of *Escherichia coli* for enhanced arginine biosynthesis. *Microb. Cell Fact.*, **14**, 29.

Herrgard,M.J. *et al.* (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.*, **16**, 627–635.

Huang,J. *et al.* (2005) Cross-regulation among disparate antibiotic biosynthetic pathways of *Streptomyces coelicolor*. *Mol. Microbiol.*, **58**, 1276–1287.

Imam,S. *et al.* (2015) Data-driven integration of genome-scale regulatory and metabolic network models. *Front. Microbiol.*, **6**, 409.

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

Kim,M. *et al.* (2014) Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in *Streptomyces coelicolor* A3(2). *Biotechnol. J.*, **9**, 1185–1194.

Kim,M. *et al.* (2016) Transcriptomics-based strain optimization tool for designing secondary metabolite overproducing strains of *Streptomyces coelicolor*. *Biotechnol. Bioeng.*, **113**, 651–660.

King,Z.A. *et al.* (2015) Next-generation genome-scale models for metabolic engineering. *Curr. Opin. Biotechnol.*, **35C**, 23–29.

Lee,J.W. *et al.* (2012) Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.*, **8**, 536–546.

Lee,S.Y. and Kim,H.U. (2015) Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.*, **33**, 1061–1072.

Lewis,N.E. *et al.* (2010) Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, **6**, 390.

Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA*, **100**, 15522–15527.

Maia,P. *et al.* (2016) In silico constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiol. Mol. Biol. Rev.*, **80**, 45–67.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Meysman,P. *et al.* (2014) COLOMBOS v2.0: an ever expanding collection of bacterial expression compendia. *Nucleic Acids Res.*, **42**, D649–D653.

Orth,J.D. *et al.* (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol.*, **7**, 535.

Orth,J.D. *et al.* (2010) What is flux balance analysis? *Nat. Biotechnol.*, **28**, 245–248.

Park,J.M. *et al.* (2012) Flux variability scanning based on enforced objective flux for identifying gene amplification targets. *BMC Syst. Biol.*, **6**, 106.

Park,U. *et al.* (2000) Genetic analysis of absR, a new abs locus of *Streptomyces coelicolor*. *J. Microbiol. Biotechnol.*, **10**, 169–175.

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.

Vilaca,P. *et al.* (2011) A computational tool for the simulation and optimization of microbial strains accounting integrated metabolic/regulatory information. *Biosystems*, **103**, 435–441.

Vivek-Ananth,R.P. and Samal,A. (2016) Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems*, **147**, 1–10.

Winkler,J.D. *et al.* (2015) The LASER database: Formalizing design rules for metabolic engineering. *Metab. Eng. Commun.*, **2**, 30–38.

Yim,H. *et al.* (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol.*, **7**, 445–452.

Zhang,F. *et al.* (2012) Enhancing fatty acid production by the expression of the regulatory transcription factor FadR. *Metab. Eng.*, **14**, 653–660.