<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *IEEE/ACM Transactions on Computational Biology & Bioinformatics*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Gu, D., Jian, X., Zhang, C., Hua, Q. (2016)

Reframed genome-scale metabolic model to facilitate genetic design and integration with expression data

IEEE/ACM Transactions on Computational Biology & Bioinformatics

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-248642>

Reframed Genome-Scale Metabolic Model to Facilitate Genetic Design and Integration with Expression Data

Deqing Gu, Xingxing Jian, Cheng Zhang, Qiang Hua

Abstract—Genome-scale metabolic network models (GEMs) have played important roles in the design of genetically engineered strains and helped biologists to decipher metabolism. However, due to the complex gene-reaction relationships that exist in model systems, most algorithms have limited capabilities with respect to directly predicting accurate genetic design for metabolic engineering. In particular, methods that predict reaction knockout strategies leading to overproduction are often impractical in terms of gene manipulations. Recently, we proposed a method named LTM (logical transformation of model) to simplify the gene-reaction associations by introducing intermediate pseudo reactions, which makes it possible to generate genetic design. Here, we propose an alternative method to relieve researchers from deciphering complex gene-reactions by adding pseudo gene controlling reactions. In comparison to LTM, this new method introduces fewer pseudo reactions and generates a much smaller model system named as gModel. We showed that gModel allows two seldom reported applications: identification of minimal genomes and design of minimal cell factories within a modified OptKnock framework. In addition, gModel could be used to integrate expression data directly and improve the performance of the E-Fmin method for predicting fluxes. In conclusion, the model transformation procedure will facilitate genetic research based on GEMs, extending their applications.

Index Terms—Genome-scale metabolic network models; Genetic design; Expression data

1 INTRODUCTION

Genome-scale metabolic network models (GEMs), which are built on extensive experimental data and literatures of gene annotation and enzyme function, provide a context for analysing attributes of a given metabolic system and have been shown to predict cellular growth phenotypes [1-4]. However, the complete applicability of GEMs relies heavily on development of model-based algorithms. Over the past two decades, advances in the microbial production industry have strongly motivated the development of algorithms for *in silico* metabolic engineering based on GEMs [5-7]. Moreover, much effort has been focused on developing algorithms to incorporate multi-omics data in order to improve the prediction power of GEMs [8, 9].

OptKnock, a programming framework that identifies knockout strategies that lead to the highest potential production by using bi-level optimization, is one of the most prestigious GEMs-based algorithms for redesigning metabolic networks for the overproduction

of target metabolites [10]. Several OptKnock-derived methods [11-14] further extend this framework to address more comprehensive applications and enhanced production of desired biochemicals. These methods are all mainly based on the flux balance analysis (FBA) method [15] and on multi-level optimization (usually consisting of two optimization problems: inner optimization of cellular objective and outer optimization of production). Some algorithms have successfully been applied to metabolic engineering; however, most of them face the dilemma of intractability while scanning for comprehensive strategies. Actually, they all aimed to solve a combinatorial optimization problem. The space of all possible genetic manipulations is quite vast. Thus, computational methods seeking to find an optimal set of knockouts generally result in exponentially increased running time, which thus limits optimal genomic designs found *in silico* to a small number of knockouts. Until recently, methods such as ReacKnock [16] and Genetic Design through Branch and Bound (GDBB) [17] challenged the inner dual transformation system of bi-level optimization in OptKnock and greatly improved search efficiency. However, due to the complex gene-reaction relationships, gene knockout strategies cannot be directly proposed by these reaction-based methods. Even for the GDBB algorithm, only genetic strategies, in the form of gene sets (such as a set of [(gene A and gene B) or gene C]), can be predicted. Methods that identify up- or down-regulation strategies are still very limited, amongst which, the most well-

- Deqing Gu is with the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China. E-mail: hellodannygood@sina.com
- Xingxing Jian is with the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China. E-mail: jianxingxing@foxmail.com
- Cheng Zhang is with the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai, China. E-mail: cheng.zhang@scilifelab.se
- Qiang Hua is with the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, and Collaborative Innovation Center for Biomanufacturing Technology, Shanghai, China. E-mail: qhua@ecust.edu.cn

known are FSEOF/FVSEOF [18, 19], OptForce [14] and Redirector [20]. Unfortunately, these methods also lack the ability to identify genetic strategies directly and are limited to reaction strategies identification. Recently, we reported a framework named LTM [21] that is able to simplify the gene-reaction associations by dividing each multiple gene related reaction into several intermediate reactions with the pseudo metabolites involved and enables accurate gene level analysis based on GEMs. In this study, we further developed an alternative method for reframing GEMs by introducing a pseudo gene controlling reaction for each single gene, which could result in a much smaller reframed model, enabling accurate genetic strategies to be generated by most reaction-based algorithms. In addition, the reframed models obtained in this study allow the identification of minimal genome or even designing a minimal cell factory within a very reasonable calculation time.

Although GEMs, also known as constraint-based models, have proven to be useful for helping understanding cellular metabolism, these models have quite a limited predicting capacity owing to fewer constraints on flux space. Thus, several methods have been proposed by integrating available expression into GEM data in different ways [22-24]. Due to the complexity of Gene-Protein Relations (GPRs) and uncertain relationships between gene expression profiles and reaction fluxes, it remains difficult to obtain a method that is reliable enough [8]. Specifically, numerous FBA-driven algorithms have been introduced, which use experimentally derived mRNA transcript levels to constrain the network's reactions, either by inactivating them entirely, or by constraining their activity levels. Nonetheless, despite the advantages of these algorithms, a key issue remains to be addressed. If a multi-functional gene was highly expressed, how could the related functions that were activated be determined? In most existing methods, all related functions are considered as activated, resulting in equally imposed constraints on the related reactions. However, this solution is not always reasonable because high expression levels may suggest that only one function needs to be fully activated. Though we are not sure such a case really exists or not in vivo, it might be better to take this special case into account. In our case, we may simply assume that multi-function enzyme are functioning just like other common enzymes to support a fast growing purpose that lies behind the FBA modeling method. In addition, for a gene involved in multiple reactions, its expression data may be reused several times by using the existing data mapping method. In the current study, we present a framework for translating a GEM into a special model with unique gene-controlling reactions. Thus, the fluxes through the pseudo controlling reactions indirectly impose constraints on the determination of whole flux distributions. Therefore, gene expression data can be mapped onto GEMs by adding constraints of gene

controlling reactions. Using this proposed reframed model with an appropriate objective, it is possible for the model to automatically decide which functions are activated when a multi-functional gene is highly expressed. For example, if the objective is cell growth, then only functions that support the fast growing purpose will be favored.

2 MATERIALS AND METHODS

2.1 Modelling platform

FBA is the most commonly used and validated method to determine flux distribution [4, 7, 25], which assumes that organisms use substrates to maximize their growth rate and, thus, eventually to adapt to changes in the environment. In this study, we mainly used FBA to calculate flux distributions. The *Escherichia coli* model *iAF1260* [26] used in this study is a large and high-quality genome-scale metabolic network reconstruction that has been proved to be able to predict a wild-type flux distribution that is highly consistent with experimental data. In addition, a *E. coli* core model was also used. Both models were downloaded from website (<http://systemsbiology.ucsd.edu/Downloads>).

All optimization problems were solved in Matlab (version 7.12.0.635 (R2011a)) with the COBRA Toolbox 2.0 incorporated. The solvers 'glpk' (version 4.35) and Tomlab/CPLEX (version 7.9) were used for FBA and OptKnock, respectively. All the procedures were implemented on a personal computer with 3.40 GHz Intel(R) Core(TM) i7-2600k and 16.0 GB RAM.

2.2 Transformation Procedures

GEMs, also known as the Cobra model in SBML format [27], comprise thousands of reactions associated with hundreds of enzyme coding genes. The GPRs in GEMs are quite complex for the presence of isoenzymes and protein complex. That is to say, one gene might be responsible for a single or multiple enzymes, while multiple genes might code for a single enzyme. Therefore, most GEMs-based redesigning methods for metabolic engineering can only suggest reaction regulation strategies or gene set (here a gene set refers to a set of genes that code one enzyme) manipulation strategies. Strictly speaking, this type of strategy might be not practical in terms of gene manipulation. For example, a knockout strategy might contain one reaction associated with an essential gene. In addition, for a reaction strategy, which indicates up-regulation of one reaction and down-regulation of another reaction, contradiction may exist if the two reactions are related to a common gene. Both examples involve paradox regulation strategies, which have been efficiently addressed in LTM. Nevertheless, we propose an alternative, easily implementable method (illustrated in Fig. 1) to reframe GEMs.

Fig. 1. Overview of transformation rules. 'Rxn', 'S' and 'P' short for reaction, substrates and products, respectively. Boxes left beside arrows represents the reaction status in original GEMs and boxes right beside arrows represent the corresponding reaction status in reframed GEMs..

Fig. 2. Toy models. The left part denotes an original toy model and the right part denotes a reframed one (gModel). Rectangles (within inner box), ellipses and small rectangles (outside inner box) represent enzyme reactions, real metabolites and exchange reactions, respectively. Particularly, in the right figure, hexagons and diamonds represent pseudo gene metabolites and pseudo reaction metabolites, respectively.

As illustrated in **Fig. 1**, for each gene in the model, a corresponding pseudo exchange reaction (referred to as the gene controlling reaction in this study) was introduced. In the exchange reaction, only one pseudo metabolite representing a gene was involved. In general, a reaction is associated with multiple gene products and their relationships are described using Boolean operators such as "AND" and "OR." The AND operation represents the involvement of multiple gene products in catalysing a reaction, whereas the OR operation signifies

that one or both of the gene products are involved in the reaction. The following are gModel transformation principles: (1) For reactions associated with single genes (such as Rxn1 in Fig.1 and R3 in **Fig.2**), the pseudo metabolites representing the genes were added into them as substrates. (2) For reactions related to multiple genes (such as Rxn2 in Fig.1 and R1 in Fig.2), the number of different gene sets that could independently code for an enzyme involved in the reaction was first determined. Thereafter, the reactions were divided into several reactions according to the gene sets with Boolean

operators, and the corresponding gene representing pseudo metabolites were added into the divided reactions. (3) Reversible reactions (such as Rxn4 in Fig.1 and R2 in Fig.2) were first divided into two irreversible ones (a forward and a reverse reaction) and then rewritten with pseudo metabolites added following the principles described above. (4) For each reaction that would have been divided, an additional pseudo metabolite representing the reaction would also have been added to the system if there were real constraints on the bounds of the reaction (such as Rxn3 and ExRxnR3 in Fig.1, R5 and Er5 in Fig.2). This newly added metabolite would participate in every divided reaction. Thus, the exchange reaction of the new metabolite would have played the role of controlling the bounds of the special reaction fluxes. For example, a 10 upper bound for Rxn3 will result in a 10 upper bound for ExRxnR3, which will ensure that the total flux of Rxn3-1 and Rxn3-2 being smaller than 10. (5) To avoid the addition of unnecessary reactions, irreversible reactions that associated with one gene, which related only to that particular reaction were not rewritten (Rxn1 in Fig.1 would be such a special case if geneA only presents in this reaction), and the reaction itself could serve as a special gene controlling reaction. After applying the above principles, a special reframed model with gene switches was obtained. The fluxes through gene controlling reactions determined the flux distribution of the whole reframed system, and the deletion of one gene controlling reaction equalled to the deletion of one gene. In the following sections, reframed models will be referred to as “gModels.” The transformation principles are also illustrated with before and after toy models in Fig. 2. The major transforming difference between gModel and LTM model is that gModel does not divide a reaction which is associated with a gene set that codes only an enzyme. For example, when a reaction is associated with a gene set (gene A, gene B, and gene C), the LTM model will divide this reaction into three reactions with three unique pseudo metabolites being used, while the gModel will not divide the reaction but introduce three gene representing metabolites instead.

2.3 Integration with expression data

The initial step in predicting fluxes using E-Fmin [28] is to map gene expression data onto reactions based on GPR (gene-protein-reaction associations). The weights of reactions are calculated with the following mathematical representation. The next step for E-Fmin is to minimize the sum of absolutely weighted flux through reactions. For reactions whose associated genes have no available expression data, their weights are assigned as 1.0. In this study, we adjusted the E-Fmin code and made it suitable for gModel.

$$\text{weight}_{\text{rxn}} = 1 - \min(g_x \text{ and } g_y) \div \max(g_1, g_2, \dots, g_n)$$

$$\text{weight}_{\text{rxn}} = 1 - \max(g_x \text{ or } g_y) \div \max(g_1, g_2, \dots, g_n)$$

2.4 Bi-level optimization framework

OptKnock [10], which was first proposed in 2003 and has been incorporated into Cobra toolbox, presents the earliest method employing bi-level optimization framework for identifying reaction knockout candidates for overproduction. However, as pointed out by Xu et al. [16], there are some problems with the current low-efficiency OptKnock in the Cobra toolbox. In this study, we reframed the OptKnock, as detailed in the Fig. 3.

bi-level_optimization_objective: product

bi-level_optimization_s.t._followed:

$$\begin{aligned} &\text{primal_problem_objective: } C_{\text{growth}}[C_1, C_2, \dots, C_j] \\ &\text{primal_problem_s.t._followed:} \\ &\sum S_{i,j} \times V_j = 0, \dots, i = 1, 2, \dots, M; j = 1, 2, \dots, N; \dots (1) \\ &Y_k \times V_{j,k} = 0, \dots, V_{j,k} = V_j; k = 1, 2, \dots, K; K \leq N; \dots (2) \\ &V_{j,l} \leq ub_{j,l}, \dots, V_{j,l} = V_j; l = 1, 2, \dots, L; j \in (1, 2, \dots, N); L \leq N; \dots (3) \\ &-V_{j,l} \leq -lb_{j,l}, \dots, V_{j,l} = V_j; l = 1, 2, \dots, L; j \in (1, 2, \dots, N); L \leq N; \dots (4) \\ &V_p \leq 0, \dots, p = 1, 2, \dots, P; P \leq N; \dots (5) \\ &V_q \geq 0, \dots, q = P, P+1, \dots, Q; Q \leq N; \dots (6) \\ &V_e \text{ is free}, \dots, e = Q, Q+1, \dots, N; \dots (7) \\ &\text{dual_problem_objective: } C_{\text{dual}}, [0, 0, ub, lb] \\ &\text{dual_problem_s.t._followed:} \\ &\sum S_{j,i} \times W_i + W_{M+k} + W_{M+K+1} - W_{M+K+L+1} \geq C_j, \dots, j = 1, 2, \dots, P; \dots (8) \\ &\sum S_{j,i} \times W_i + W_{M+k} + W_{M+K+1} - W_{M+K+L+1} \leq C_j, \dots, j = P, P+1, \dots, Q; \dots (9) \\ &\sum S_{j,i} \times W_i + W_{M+k} + W_{M+K+1} - W_{M+K+L+1} = C_j, \dots, j = Q, Q+1, \dots, N; \dots (10) \\ &W_i \text{ is free}, \dots, i = 1, 2, \dots, M; \dots (11) \\ &W_{\min} \leq W_{M+k} \times Y_k \leq W_{\max}, \dots, k = 1, 2, \dots, K; \dots (12) \\ &W_{M+1} \geq 0, \dots, L = 1, 2, \dots, L; \dots (13) \\ &W_{M+L+1} \geq 0, \dots, L = 1, 2, \dots, L; \dots (14) \\ &\text{additional_s.t._for_bi-level_optimization} \\ &C_{\text{growth}} \times V - C_{\text{dual}} \times W = 0, \dots (15) \\ &\sum Y_k \leq \text{maxKnockNumber}, \dots (16) \end{aligned}$$

Fig. 3 A mathematical framework of OptKnock.

In Fig. 3, M and N denote the number of metabolites and reactions respectively. V denotes the reaction fluxes and S denotes the stoichiometry. C represents the vector of the objective of the optimization problem. K represents the number of knockout candidates. L represents the number of reaction with exact flux bounds. lb and ub denote lower bound and upper bound of a reaction respectively. P denotes the index of reaction whose flux is constrained to be smaller than zero; and Q denotes the index of reaction whose flux is constrained to be bigger than zero. W represents the dual variables of the dual problem. $W_{\max}$ represents the upper bound of W , which was specified to be 100 according to the study of Kim et al. [29]. Y denotes a binary variable. **maxKnockNumber** denotes the allowable knockout number set by users. The **upper box** depicts the primal problem of optimizing biomass, while the **lower box** depicts the dual problem of optimizing biomass and the two boxes together depict the bi-level optimization problem. The goal of outer optimization is

TABLE 1
MODEL BEFORE AND AFTER REFRAMING

Models	Reactions	Metabolites	Genes	Essential Genes	Flux Consistency
model	2382	1668	1261	187	-
gModel	5907	2597	1261	187	98.15%
LTM model	7287	5372	1261	187	100%

Note: model denotes the original model, whereas gModel and LTM model denote the reframed models.

the maximum production of target metabolite, while the goal of inner optimization is the maximum growth of biomass that will be guaranteed by the constraint in the bi-level optimization problem. The constraint (2) is originally a non-linear constraint, but can be easily translated into a linear constraint as follows: $(1-Y)*lb < V < (1-Y)*ub$. To better understand the calculation framework, readers are strongly suggested to read the original paper of OptKnock and ReacKnock. The scale of the OptKnock problem is determined by the numbers of variables including reaction number, knockout candidate number and metabolite number. Thus, it is recommended to first reduce the number of knockout candidates as much as possible. In the original OptKnock framework, reversible reactions will be divided into two irreversible reactions. However, this process was avoided in our framework according to mathematical dual transformation rules and thus introduced less variables. This newly formulated OptKnock is slightly more efficient and easier to use compared with the one in Cobra toolbox. For example, there is no longer need to pass bound constraints already included in the model as parameters into the OptKnock problem again. In addition, the OptKnock problem could also be efficiently solved with a truncated branch and bound algorithm like GDBB.

3 RESULTS

To validate the proposed transformation procedure, we made a direct comparison between the original model and the reframed *iAF1260* model. Algorithms such as OptKnock were shown to be able to generate genetic designs directly based on the gModel. Furthermore, we showed that OptKnock could be employed to suggest a minimal genome based on gModel. Finally, we adapted the method E-Fmin for integrating expression data into the gModel to improve phenotype prediction.

3.1 Validation of gModel and comparison with LTM model

The reframing rules were successfully applied on the large-scale *iAF1260* model. To facilitate a direct comparison, detailed information of the original model and reframed model was listed in **Table 1**. The LTM was also applied to the same model and LTM model information was listed in the same table. According to the data we obtained, the gModel has a much smaller

Fig. 4 Comparison of flux distribution and single gene deletion. The left two graphs are simulation results of original model, while the right two graphs of the figure are simulation results of the gModel.

size than that of the LTM model, thus may resulting in reduced computational burden.

To validate the reliability of the reframed model, we compared the before and after flux distributions obtained by implementing standard FBA (**Fig. 4**). As expected, the growth rates were consistent with each other. The flux distributions were slightly different between the gModel and the original model, while those of the LTM model were identical to those of the original model. The small discrepancy between the gModel and the original model does not necessarily mean that the gModel is incorrectly transformed. By constraining the reaction flux with the solution of gModel in the original model, we had validated that the solution of gModel is also a solution of the original model. Therefore, the small discrepancy should not indicate any loss of information. In fact, before reframing the model, the fluxes of some reactions (usually more than 500 mmol/gDW/h) were not reasonable. As discussed by Verwoerd and Mao [30], unreasonable flux is often caused by futile cycles, which may be overcome by minimizing total flux. As depicted in **Fig. 4**, the flux distribution obtained from the gModel was more reasonable (detailed flux distributions are provided in S1 Table of Appendix A). Actually, our many simulation results showed that the gModel tends to avoid most futile cycles (data not shown). Moreover, we compared the results of performing single gene deletion analysis in the original model and single gene controlling reaction deletion analysis in the gModel. Both analyses indicated a same set of 187 genes were essential under the condition of minimal glucose medium (see **Fig. 4**).

Furthermore, we made a double gene deletion analysis in model *iAF1260*, and the result of gModel is consistent with that of the original model (results displayed in **Appendix A**). Therefore, based on the flux analysis and gene deletion analysis, we conclude that the gene controlling reactions work well in the gModel.

3.2 Minimal genome prediction and minimal cell factory design

The OptKnock framework was first extended to generate gene knockout strategies based on the developed gModel. The *E. coli* core model was used as a testing model platform to reduce computational time. In this study, the overproduction of succinate was used as an example, and the maximal oxygen and glucose uptake rates were set to 18.5 mmol/gDW/h and 10 mmol/gDW/h respectively. To identify genetic strategies, the gModel was applied and only the gene controlling reactions were taken as knockout candidates. To identify reaction knockouts, the original model was used and all reactions that were associated with at least one gene were considered to be knockout candidates. The detailed results are listed in S2 Table of Appendix A, and partially illustrated in **Fig. 5**. It is worth mentioning that the same knockout strategies were obtained based on the LTM model and the gModel

Fig. 5. Comparison between two types of knockout strategies. The circle above the bar indicates that the corresponding knockout strategy is not practical in terms of gene manipulation. Gene knockouts were obtained based on gModel, while reaction knockouts were obtained based on the original model.

(results not shown).

As illustrated in **Fig. 5**, the best global reaction knockout strategy was found to be the one with eight knockouts, while the best gene knockout strategy consisted of 12 knockouts. The obtained optimal production rates generated by both strategies were quite similar, 14.016 mmol/gDW/h for reaction knockouts and 13.959 mmol/gDW/h for gene knockouts; even though, reaction knockout strategies generally lead to higher production rates in comparison to gene knockout strategies. However, note that only the octuplet reaction knockout strategy was deemed feasible. Based on the data shown in **S2 Table**, we concluded that the complex gene-reaction relationships account for the impractical solution of reaction knockouts. It should be noted that

other algorithms like ReacKnock, OptForce, and FSEOF could also utilize gModel to directly generate reasonable genetic design strategies. FSEOF, in particular, is usually employed to identify large sets of reaction amplification targets; however, it is quite inconvenient for users to extract gene amplification targets when considering complex gene-reaction relationship. Fortunately, gModel could be helpful in solving this problem by directly suggesting gene regulation targets.

Determination of the minimal set of protein-coding genes necessary to maintain a living cell has been an appealing issue for years [31]. Such a minimal gene set should include “the smallest possible group of genes that would be sufficient to sustain a functioning cellular life under the most favourable conditions imaginable, that is, in the presence of a full complement of essential nutrients and in the absence of environmental stress” [32]. The increasing knowledge of comparative genomics should facilitate the reconstruction of the minimal gene set, also referred to as minimal genome [33]. However, few computational methods based on genome-scale metabolic model were reported to find minimal genome. In this study, we proposed a straightforward framework to predict the minimal genome with gModel. The minimal genome problem can be rephrased as “Which genes could be removed simultaneously while maintaining a minimal growth rate? If this problem had multiple solutions, then which solution would result in the best growth rate?” Thus, this is a bi-level optimization problem: the outer optimization goal is gene knockout number, and the inner optimization target is growth rate. As such, based on the gModel, this problem can be solved directly using OptKnock with the gene controlling reaction knockout number as the outer optimization goal. Consequently, we modified the OptKnock procedure slightly to allow the optimization of knockout number, so that the OptKnock procedure provided here is capable of finding a minimal genome. In fact, it took less than 16 hours to find a minimal genome for the large-scale GEM *iAF1260*. It should be noted that OptKnock provides only one solution of minimal genome while multiple solutions may exist. Therefore, ReacKnock may serve as a good substitute for OptKnock as it could give a set of equivalent optimal

Fig. 6. Pathway distribution of genes in minimal genome identified for *E. coli* cells grown in rich medium.

solutions.

To obey the definition of a minimal gene set, we set the minimal growth rate to 0.005 mmol/gDW/h and simulated the cell growth in rich medium with all exchange reactions opened with lower bounds of 100 mmol/gDW/h. As a result, a minimal genome with only 145 genes was obtained. According to the PEC database (<http://www.shigen.nig.ac.jp/ecoli/pec/>), 62 of the 145 genes are essential (essential genes are listed in S3 Table of Appendix A). The rest of the genes are mostly involved in cofactor and prosthetic group synthesis, cell membrane synthesis, inorganic ion transport and nucleotide synthesis (Fig. 6). According to minimal genome research, all of these pathways are important for cell survival [34]. Hence, the predicted minimal genome could provide useful physiological and metabolic information for systems biologists.

It is believed that a minimal cell could have beneficial properties for industrial applications. *E. coli* strains without redundant metabolic pathways produced much higher quantities of ethanol from hexose and pentose sugars [35, 36]. *E. coli* lacking 22% of the genomic DNA produced 2.4-fold more threonine than the wild-type strain did [37]. To the best of our knowledge, currently few methods were reported to predict a minimal cell factory based on GEM. Here, we used gModel to predict the minimal genome of a minimal cell factory having the previously specified

minimum growth rate (0.05 h⁻¹) and production rate (13.378 mmol/gDW/h, 80% of theoretical maximum production), under aerobic conditions with a glucose uptake rate of 10 mmol/gDW/h. As a result, we obtained a mini-cell factory with only 282 genes, which could maintain a growth rate of 0.259 h⁻¹ and a minimal succinate production rate of 7.297 mmol/gDW/h. The minimal genome is provided in S4 Table of Appendix A. This functional mini-cell factory could be further improved by tuning parameters such as carbon source and oxygen uptake rates, thus enabling the design of excellent engineering strategies for strain improvement.

3.3 Integration of Expression Data

“Expression data-guided flux minimization” (E-Fmin) is a newly reported method that uses transcriptional data taken from a single experimental condition and has been proved more effective on predicting context-specific flux distributions than many existing methods [28]. This algorithm minimizes a weighted sum of flux magnitudes, where the weights are calculated based on expression data (described in the methods section). E-Fmin maps gene expression data onto reactions based on GPR in a similar way to many other methods, which arbitrarily impose the same constraints on reactions catalysed by a common active enzyme. In addition, for a gene involved in multiple reactions, its expression data may be reused several times by using the above data

Fig. 7 Comparison of fluxes predicted by E-Fmin and experimental fluxes under six different conditions. Lines with “x” markers denote flux calculated by minimizing total weighted fluxes of gene controlling reactions. Lines with “o” markers denote experiment determined fluxes and Lines with “.” markers denote fluxes calculated by minimizing total weighted fluxes of enzyme-catalysed reactions. corr(gene,C13) represent correlation coefficients between fluxes predicted by minimizing total weighted fluxes of gene controlling reactions and fluxes determined experimentally. corr(rxn,C13) represent correlation coefficients between fluxes predicted by minimizing total weighted fluxes of enzyme-catalysed reactions and fluxes determined experimentally. All p-values were less than 0.001. The conditions denoted with four genes *pgm*, *pgi*, *zwf* and *rpe* means that four corresponding mutants with single gene deleted were examined.

mapping method. However, the traditional way of mapping expression data onto reactions is flawed. For example, in case that a multi-functional enzyme encoded by a gene with two units of expression (assuming that one unit of gene expression is corresponding to one unit of flux), the total flux of the reactions catalyzed by the enzyme should be limited to two units according to the assumption. However, according to the traditional way, four units of flux are actually allowed for same constraints are imposed on reactions catalysed by the multi-functional gene. Fortunately, in case of gModel, *in silico* cells may mirror the expression data by imposing constraints on gene controlling reactions rather than exact functional reactions. It therefore could be more appropriate to directly minimize the weighted flux of gene controlling reactions rather than minimizing the weighted flux of functional reactions.

Based on six sets of whole genome expression data and the corresponding six sets of ^{13}C flux data reported by Ishii et al. [38], we performed E-Fmin by minimizing the flux of gene controlling reactions as well as minimizing enzyme-catalysed reactions. Fluxes predicted by E-Fmin were then compared with experimental flux data, and the correlation coefficients and p-values for testing the hypothesis of no correlation were calculated with the function 'corrcoef' incorporated in Matlab. The model *iAF1260* with 80% optimal growth rate maintained was used. Significant improvement on predictions was observed in the results generated for three among the six conditions tested, while the predictions for the other three conditions remained almost unchanged (Fig. 7). Moreover, the small p-values (< 0.001) support a reliable correlation between prediction data and experimental data. Similarly, expression data of 85 genes and ^{13}C flux data of the central metabolism under 28 conditions were used as input data and validation data respectively [38]. The test results were recorded in S5 Table of Appendix A, where it can be seen that correlation coefficients calculated based on gModel commonly have higher values through all conditions. Similar trends were also observed when 70% or 90% of the optimal growth rate was maintained (Appendix B including codes and raw data was provided in the supplementary material for testing purposes). Thus, we concluded that gModel tends to improve the prediction of fluxes in case of integration with expression data by directly mapping single gene expression data onto single gene controlling reaction.

4 DISCUSSION

A special framework was proposed in this study to reframe genome-scale metabolic model and facilitate genetic research as well as omics data integration. In addition, OptKnock was rejuvenated and rendered able to predict exact genetic strategies for metabolic engineering with gModel. The format of gModel is consistent with any other GEMs, thus it should also be compatible with existing algorithms such as ReactKnock

and OptForce. It could achieve the same efficiency of LTM model in the application of genetic knockout design. In addition, compared with LTM, gModel provides a more straightforward and elegant way in other applications such as integration with expression data and over-expression target identification. In this study, more interestingly, we shown that gModel could be used as a novel tool for designing minimal genomes and minimal cell factories for the production of target chemicals. Though the current genome-scale models are not yet perfect and could generate unreliable predictions, we believe that the gModel constitutes a very good starting point that will assist systems biologists to advance a step forward towards their goal of building a minimal cell factory.

While gModel may release researchers from complex gene-reaction relationships, the relatively increased scale of model structure leads to increased computational burden; it is encouraging, though, that efficient algorithms are being developed quickly. GDBB and Reacknock are the two most efficient methods that suggest near-optimal solutions within a few minutes. In this study, the OptKnock framework was modified to adapt to the truncated branch and bound algorithm employed in GDBB. In this way, the modified OptKnock achieved the same efficiency as that of OptGeneKnock designed for LTM model. For example, a near-optimal solution for succinate production under fully aerobic condition was obtained within one hour. The results of this example suggested that the knockout of 14 genes (*aceE*, *dut*, *gapA*, *gcd*, *poxB*, *sdhC*, *serB*, *udp*, *focA*, *pyrH*, *pgl*, *nudI*, *ylil*, and *focB*) could result in a minimal yield of 11.609 mmol/gDW/h with a maintained growth rate of 0.050 h^{-1} .

A direct and efficient way to constrain a GEM is the incorporation of fluxomic data; however, this data is not readily available. In contrast, loads of expression data are available nowadays. In this study, we proved that the performance of the newly reported E-Fmin method integrating expression data could be improved within the gModel framework. There is an assumption underlying the construction of gModel: for each reaction, one unit flux of a functional reaction requires one unit flux of gene controlling reaction, implying that the stoichiometric coefficient of each gene-representing metabolite is one. This assumption may be vulnerable if the elaborate connection between gene expression and enzyme function is considered. However, considering that only soft-constraints were imposed on reaction fluxes, the above assumption might be acceptable. In conclusion, gModel promises important contributions to the usefulness of genome-scale models.

ACKNOWLEDGMENT

Qiang Hua is the corresponding author. The work was financially supported by the National Basic Research Program of China (973 Program) (2012CB721101), and the National Special Fund for State Key Laboratory of Bioreactor Engineering (2060204).

REFERENCES

- [1] M. Akesson, J. Forster, and J. Nielsen, "Integration of gene expression data into genome-scale metabolic models," *Metab Eng*, vol. 6, no. 4, pp. 285-93, Oct, 2004.
- [2] I. Borodina, and J. Nielsen, "From genomes to in silico cells via metabolic networks," *Curr Opin Biotechnol*, vol. 16, no. 3, pp. 350-5, Jun, 2005.
- [3] J. M. Schwartz, and C. Gaugain, "Genome-scale integrative data analysis and modeling of dynamic processes in yeast," *Methods Mol Biol*, vol. 759, pp. 427-43, 2011.
- [4] N. E. Lewis, K. K. Hixson, T. M. Conrad et al., "Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models," *Mol Syst Biol*, vol. 6, pp. 390, Jul, 2010.
- [5] A. M. Feist, and B. O. Palsson, "The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*," *Nature Biotechnology*, vol. 26, no. 6, pp. 659-667, Jun, 2008.
- [6] T. Osterlund, I. Nookaew, and J. Nielsen, "Fifteen years of large scale metabolic modeling of yeast: developments and impacts," *Biotechnol Adv*, vol. 30, no. 5, pp. 979-88, Sep-Oct, 2012.
- [7] K. A. Curran, N. C. Crook, and H. S. Alper, "Using flux balance analysis to guide microbial metabolic engineering," *Methods Mol Biol*, vol. 834, pp. 197-216, 2012.
- [8] A. S. Blazier, and J. A. Papin, "Integration of expression data in genome-scale metabolic network reconstructions," *Front Physiol*, vol. 3, pp. 299, 2012.
- [9] Y. Matsuoka, and K. Shimizu, "Current status and future perspectives of kinetic modeling for the cell metabolism with incorporation of the metabolic regulation mechanism," *Bioresources and Bioprocessing*, vol. 2, no. 1, pp. 4, 2015.
- [10] A. P. Burgard, P. Pharkya, and C. D. Maranas, "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization," *Biotechnol Bioeng*, vol. 84, no. 6, pp. 647-57, Dec 20, 2003.
- [11] N. Tepper, and T. Shlomi, "Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways," *Bioinformatics*, vol. 26, no. 4, pp. 536-43, Feb 15, 2010.
- [12] P. Pharkya, and C. D. Maranas, "An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems," *Metab Eng*, vol. 8, no. 1, pp. 1-13, Jan, 2006.
- [13] P. Pharkya, A. P. Burgard, and C. D. Maranas, "OptStrain: a computational framework for redesign of microbial production systems," *Genome Res*, vol. 14, no. 11, pp. 2367-76, Nov, 2004.
- [14] S. Ranganathan, P. F. Suthers, and C. D. Maranas, "OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions," *PLoS Comput Biol*, vol. 6, no. 4, pp. e1000744, Apr, 2010.
- [15] J. D. Orth, I. Thiele, and B. O. Palsson, "What is flux balance analysis?," *Nat Biotechnol*, vol. 28, no. 3, pp. 245-8, Mar, 2010.
- [16] Z. Xu, P. Zheng, J. Sun et al., "ReacKnock: identifying reaction deletion strategies for microbial strain optimization based on genome-scale metabolic network," *PLoS One*, vol. 8, no. 12, pp. e72150, 2013.
- [17] D. Egen, and D. S. Lun, "Truncated branch and bound achieves efficient constraint-based genetic design," *Bioinformatics*, vol. 28, no. 12, pp. 1619-23, Jun 15, 2012.
- [18] H. S. Choi, S. Y. Lee, T. Y. Kim et al., "In Silico Identification of Gene Amplification Targets for Improvement of Lycopene Production," *Applied and Environmental Microbiology*, vol. 76, no. 10, pp. 3097-3105, May, 2010.
- [19] J. M. Park, H. M. Park, W. J. Kim et al., "Flux variability scanning based on enforced objective flux for identifying gene amplification targets," *Bmc Systems Biology*, vol. 6, Aug 21, 2012.
- [20] G. Rockwell, N. J. Guido, and G. M. Church, "Redirector: designing cell factories by reconstructing the metabolic objective," *PLoS Comput Biol*, vol. 9, no. 1, pp. e1002882, 2013.
- [21] C. Zhang, B. Ji, A. Mardinoglu et al., "Logical transformation of genome scale metabolic models for gene level applications and analysis," *Bioinformatics*, Mar 2, 2015.
- [22] T. Shlomi, Y. Eisenberg, R. Sharan et al., "A genome-scale computational study of the interplay between transcriptional regulation and metabolism," *Mol Syst Biol*, vol. 3, pp. 101, 2007.
- [23] P. A. Jensen, and J. A. Papin, "Functional integration of a metabolic network model and expression data without arbitrary thresholding," *Bioinformatics*, vol. 27, no. 4, pp. 541-547, February 15, 2011, 2011.
- [24] P. A. Jensen, K. A. Lutz, and J. A. Papin, "TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks," *BMC Syst Biol*, vol. 5, pp. 147, 2011.
- [25] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data," *Nature Biotechnology*, vol. 19, no. 2, pp. 125-130, Feb, 2001.
- [26] A. M. Feist, C. S. Henry, J. L. Reed et al., "A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information," *Molecular Systems Biology*, vol. 3, Jun, 2007.
- [27] M. Hucka, A. Finney, H. M. Sauro et al., "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524-31, Mar 1, 2003.
- [28] H. S. Song, J. Reifman, and A. Wallqvist, "Prediction of metabolic flux distribution from gene expression data based on the flux minimization principle," *PLoS One*, vol. 9, no. 11, pp. e112524, 2014.
- [29] J. Kim, J. L. Reed, and C. T. Maravelias, "Large-scale bi-level strain design approaches and mixed-integer programming solution techniques," *PLoS One*, vol. 6, no. 9, pp. e24162, 2011.
- [30] W. S. Verwoerd, and L. Mao, "The Problem of Futile Cycles in Metabolic Flux Modeling: Flux Space Characterization and Practical Approaches to Its Solution," *Modelling and Simulation of Diffusive Processes, Simulation Foundations, Methods and Applications* S. K. Basu and N. Kumar, eds., pp. 233-256: Springer International Publishing, 2014.
- [31] M. Lluch-Senar, J. Delgado, W. H. Chen et al., "Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium," *Mol Syst Biol*, vol. 11, pp. 780, 2015.
- [32] E. V. Koonin, "How many genes can make a cell: the minimal-gene-set concept," *Annu Rev Genomics Hum Genet*, vol. 1, pp. 99-116, 2000.
- [33] W. Davids, and Z. Zhang, "The impact of horizontal gene

transfer in shaping operons and protein interaction networks--direct evidence of preferential attachment," *BMC Evol Biol*, vol. 8, pp. 23, 2008.

- [34] R. Gil, F. J. Silva, J. Pereto et al., "Determination of the core of a minimal bacterial gene set," *Microbiol Mol Biol Rev*, vol. 68, no. 3, pp. 518-37, table of contents, Sep, 2004.
- [35] C. T. Trinh, P. Unrean, and F. Sreenc, "Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses," *Applied and environmental microbiology*, vol. 74, no. 12, pp. 3634-3643, 2008.
- [36] D. Sarkar, and K. Shimizu, "An overview on biofuel and biochemical production by photosynthetic microorganisms with understanding of the metabolism and by metabolic engineering together with efficient cultivation and downstream processing," *Bioresources and Bioprocessing*, vol. 2, no. 1, pp. 17, 2015.
- [37] H. Mizoguchi, Y. Sawano, J. Kato et al., "Superpositioning of deletions promotes growth of Escherichia coli with a reduced genome," *DNA Res*, vol. 15, no. 5, pp. 277-84, Oct, 2008.
- [38] N. Ishii, K. Nakahigashi, T. Baba et al., "Multiple high-throughput analyses monitor the response of *E. coli* to perturbations," *Science*, vol. 316, no. 5824, pp. 593-7, Apr 27, 2007.

Deqing Gu has an MSc degree (2014) in Biochemical Engineering from East China University of Science and Technology (ECUST). His main research interest is in silico metabolic engineering based on genome-scale model metabolic network and is now working as a bioinformatician in a private enterprise providing high-throughput technology services.

Xingxing Jian received the BSc degree in Bioengineering from Inner Mongolia University of Science and Technology in 2013. He is now a PhD student in the School of Biotechnology, East China University of Science and Technology (ECUST). His research interests focus on the reconstruction of genome-scale metabolic model and its application in metabolic engineering.

engineering and systems medicine.

Cheng Zhang received the BSc degree in Bioengineering at Department of Biotechnology, East China University of Science and Technology (ECUST). He also studied in Systems and Synthetic Biology group in Chalmers University as a visiting student for two years during his PhD period. His main research interests include developing and expanding genome-scale metabolic models to other omics level, and applying them into research fields such as metabolic

engineering of yeasts for the production of useful compounds. He is the executive editor of a Springer journal of *Bioresources and Bioprocessing*.

Qiang Hua received the PhD degree in Bioengineering from Kyushu Institute of Technology, Japan, in 2000. He is a professor in the School of Biotechnology and the State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, China. His research interests include isotope-based quantification and analysis of cellular metabolism, simulation and design of metabolic networks, and metabolic