Logical transformation of genome-scale metabolic models for gene level applications and analysis

Bioinformatics, 31(14), 2015, 2324–2331
doi: 10.1093/bioinformatics/btv134
Advance Access Publication Date: 2 March 2015
Original paper

Systems biology

Logical transformation of genome-scale metabolic models for gene level applications and analysis

Cheng Zhang[1,2], Boyang Ji[2], Adil Mardinoglu[2], Jens Nielsen[2,*] and Qiang Hua[1,3,*]

[1]State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China, [2]Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Göteborg, Sweden and [3]Shanghai Collaborative Innovation Center for Biomanufacturing Technology (SCICBT), Shanghai 200237, China

*To whom correspondence should be addressed.
Associate editor: Jonathan Wren

Received on December 27, 2014; revised on February 6, 2015; accepted on February 25, 2015

Abstract

Motivation: In recent years, genome-scale metabolic models (GEMs) have played important roles in areas like systems biology and bioinformatics. However, because of the complexity of gene–reaction associations, GEMs often have limitations in gene level analysis and related applications. Hence, the existing methods were mainly focused on applications and analysis of reactions and metabolites.

Results: Here, we propose a framework named logic transformation of model (LTM) that is able to simplify the gene–reaction associations and enables integration with other developed methods for gene level applications. We show that the transformed GEMs have increased reaction and metabolite number as well as degree of freedom in flux balance analysis, but the gene–reaction associations and the main features of flux distributions remain constant. In addition, we develop two methods, OptGeneKnock and FastGeneSL by combining LTM with previously developed reaction-based methods. We show that the FastGeneSL outperforms exhaustive search. Finally, we demonstrate the use of the developed methods in two different case studies. We could design fast genetic intervention strategies for targeted overproduction of biochemicals and identify double and triple synthetic lethal gene sets for inhibition of hepatocellular carcinoma tumor growth through the use of OptGeneKnock and FastGeneSL, respectively.

Availability and implementation: Source code implemented in MATLAB, RAVEN toolbox and COBRA toolbox, is public available at https://sourceforge.net/projects/logictransformationofmodel.

Contact: nielsenj@chalmers.se or qhua@ecust.edu.cn

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

1 Introduction

Genome-scale metabolic models (GEMs), are a useful platform for systems biology, and have been widely used in industrial biotechnology applications. There are many successful stories showing the usefulness of GEMs in rational design of genetic modifications of microorganism for improved production of specific metabolites (Brochado *et al.*, 2010; Choi *et al.*, 2010; Fowler *et al.*, 2009; Kim *et al.*, 2013; Matsuda *et al.*, 2011; Nocon *et al.*, 2014; Ranganathan and Maranas, 2010; Ranganathan *et al.*, 2012; Xu *et al.*, 2011). In addition, GEMs have been used as scaffolds to integrate high

throughput experimental data (Chandrasekaran and Price, 2010; Hamilton *et al.*, 2013; Hoppe *et al.*, 2007) and provide new biological insight into the physiology of different microorganisms (Feist and Palsson, 2010; Harrison *et al.*, 2007; Reznik *et al.*, 2013; Segre *et al.*, 2005; Suthers *et al.*, 2009; Tepper *et al.*, 2013). More recently, human GEMs have been employed to identify novel prognostic biomarkers and potential drug targets for designing efficient treatment strategies (Agren *et al.*, 2014; Barabasi *et al.*, 2011; Jerby-Arnon *et al.*, 2014; Ji and Nielsen, 2015; Mardinoglu *et al.*, 2013a, b; Mardinoglu and Nielsen, 2012, 2015; Nam *et al.*, 2014; Ryan *et al.*, 2014; Yizhak *et al.*, 2014).

Although GEMs are widely used, their uses have some limitations in genetic related applications. A number of bi-level linear programming methods have been developed for *in silico* metabolic engineering, such as OptKnock (Burgard *et al.*, 2003), OptReg (Pharkya and Maranas, 2006), OptForce (Ranganathan *et al.*, 2010), Genetic Design through Local Search (GDLS) (Lun *et al.*, 2009) and Genetic Design through Branch and Bound (GDBB) (Egen and Lun, 2012) and these methods have only identified potential targeted reactions (knockouts or up/down regulations) rather than targeted genes which can be modified *in vivo*. Moreover, these reaction-based methods did not consider the genetic interactions in the context of protein complexes and isoenzymes, which result in strategies genetically complicated or infeasible. Recently, another bilevel-based method, OptORF (Kim and Reed, 2010), was developed for identifying metabolic and regulatory gene that are targets for knockout or upregulation which may lead to overproduction of desired chemicals. OptORF employed a three-dimensional gene–enzyme reaction array to enable gene level prediction. Nevertheless, this array needs to be systematically defined by the user for each model, which restrict the widely use of OptORF in strain design. On the other hand, OptGene and its derived algorithms (Chong *et al.*, 2014; Choon *et al.*, 2014; Patil *et al.*, 2005), could identify gene knockout strategies. However, these methods are based on Genetic Algorithm, which is basically random mutation and would probably miss the optimal solution.

Moreover, synthetic lethality (SL) analysis, which identifies combinations of mutations in two or more genes leading to cell lethality *in silico*, could be applied for identification of small hairpin RNA (shRNA) (Bernards *et al.*, 2006) and CRISPR (Cho *et al.*, 2013) targets for cancer treatment. However, the huge computational cost of triple or higher level SL makes it infeasible for its use in biotechnology applications. There were two methods developed for efficient identification of synthetic lethal reactions (Pratapa *et al.*, 2014; Suthers *et al.*, 2009). But when it comes to synthetic lethal genes analysis, one of them becomes unavailable, while the other one requires manual identification of 'appropriate equations' relating binary variables of genes and reactions, which is time consuming.

This gene level application problem originates from the gene–protein relations (GPRs) information loss in the gene–reaction association matrix (GRAM). The GRAM is a predefined binary matrix in GEMs, whose *ij*th elements is 1 if *i*th reaction is associated with *j*th gene and is 0 otherwise. GRAM is the only mathematically friendly part for most of the existing methods to capture the GPRs information. However, GRAM does not include any information of the relationship among gene sets. For example, one cannot know whether a reaction is associated with a protein complex (AND relationship in gene set) or with isoenzymes (OR relationship in gene set) directly from GRAM. Therefore, in order to correctly integrate GPRs information into the simulation algorithms, one needs to interpret the information into mathematical expressions, which is time-consuming and tedious, and sometimes it may even be extremely difficult to reconstruct the correct relationship.

Here, we introduced a GEM modification algorithm, called logic transformation of model (LTM), which extended the original GEM into a logically equivalent with much better genetic applicability. In this algorithm, pseudo reactions and metabolites were added into the GEM based on the GPR relationships described in GEMs to simulate the isoenzyme and protein complex relationships. Hereby reactions are associated with no more than one gene. Therefore, GPRs information was included in the stoichiometric matrix of GEM, and we can directly apply GRAM to all kind of optimization programming algorithms, such as OptKnock, OptReg and OptForce. Furthermore, LTM could extend the use of previously well designed methods that were only developed for reaction simulations in gene level applications. We developed OptGeneKnock, which incorporated LTM with a bilevel mixed integer linear programming (MILP)-based knockout method. The MILP problem was then solved by truncated branch and bound, to obtain near optimal gene knockout strategies with desired phenotype. In addition, by incorporating LTM with a previous method (Pratapa *et al.*, 2014), which was developed for efficient identification of reaction SL, a new method called FastGeneSL was designed for genetic SL analysis. We showed that FastGeneSL greatly outperformed exhaustive analysis in gene SL analysis for GEMs of *Escherichia coli*, *Saccharomyces cerevisiae* and hepatocellular carcinoma (HCC). Finally, we performed two case studies for demonstrating the capability of the here presented methods. First, we identified gene targets that may increase the production of desired biochemicals by designing fast genetic intervention strategies by OptGeneKnock using two different GEMs of *E.coli*. Second, we employed FastGeneSL together with a HCC GEM for gaining biological insight into underlying molecular mechanism of HCC and identifying potential shRNA targets which may inhibit tumor growth.

2 Materials and methods

2.1 Logic transformation of model

Reaction associated with more than two genes were extended into a subnetwork which is a logic equivalent based on its original GPR relationship through the use of LTM. The relationship among genes in gene sets controlling the same reaction were divided into two groups, protein complexes and isoenzymes. For example, if a reaction is associated with gene A and gene B, (A and B) means that both genes are needed to catalyze the reaction, while (A or B) means that either of these genes can catalyze the reaction and the associated reaction can only be blocked when both genes are deleted. Pseudo metabolites and reactions which do not exist in reality were added into the GEM to simulate the logical relationships of reactions and their gene sets. In case of (A and B), the reaction was separated into two pseudo reactions associated with gene A and gene B, respectively (the lower left of Fig. 1). And to keep the original flux bounds, the flux bounds are set the same as the original reaction. While in case of (A or B), the reaction was split into two equivalent pathways which are associated with genes A and B separately, and additional arm reactions were added to control the overall flux (top right of Fig. 1). The arm reactions were set the same upper ($UB_{original}$) and lower ($LB_{original}$) bounds with the original reaction, while the bounds of reactions inside the pseudo network were set differently. If $UB_{original} > 0$ and $LB_{original} > 0$, then for all inner pseudo reactions, $UB = UB_{original}$, $LB = 0$; if $UB_{original} < 0$ and $LB_{original} < 0$, then $UB = 0$, $LB = LB_{original}$; otherwise, $UB = UB_{original}$, $LB = LB_{original}$.

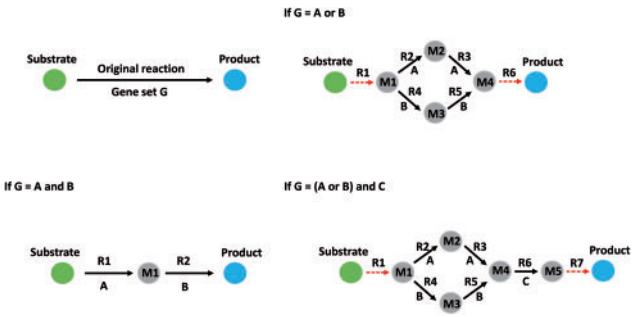

Fig. 1. Simple examples of logical transformation of reactions. Circles represent metabolites, and arrows stand for reactions. Dashed arrows indicate the arm reactions. 'M1' and 'R1' represent 1th pseudo metabolite and reaction, respectively

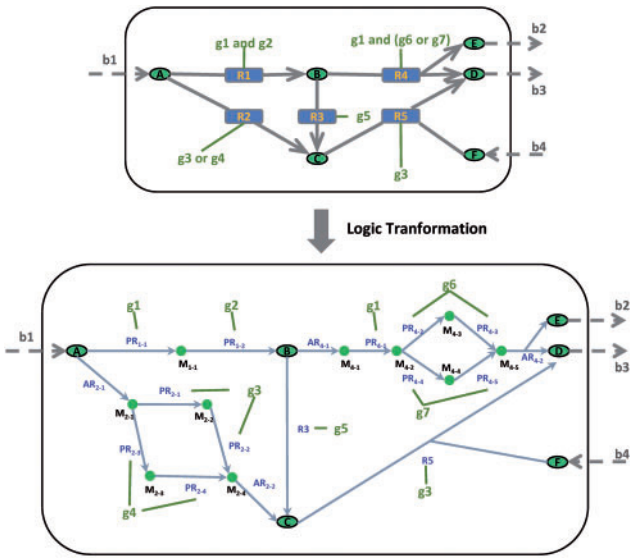

Fig. 2. Illustration of LTM using a toy model. Dashed arrows present exchange reactions and are marked as 'b1' to 'b4'. Intracellular reactions were marked as 'R1' to 'R5', and each has an arrow indicating the direction. 'g1' to 'g7' were genes. 'ARx-y', 'PRx-y' and 'Mx-y' denote the *y*th arm reaction, pseudo reaction and pseudo metabolite of reaction 'RX', respectively

By setting reaction bounds as above, all kinds of *in silico* genetic manipulation (up/down regulation and knockout) would have the same effect as in the original GEM. For more complex gene sets, this principle can also be applied step by step (Supplementary Fig. S1). However, in order to reduce the number of redundant pseudo reactions, arm reactions were only added to both side of the overall network (the lower right of Fig. 1). This kind of transformation decomposed reactions with comprehensive GPRs into several pseudo reactions. Thus, all reactions that are associated with more than one gene in the GEM were broken down into pseudo reactions associated with at most one gene.

In order to further elucidate how LTM simplifies the GRAM, we illustrated the principles of LTM using a toy model (Fig. 2). This toy model included four exchange reactions and five intracellular reactions. Three intracellular reactions were associated with more than one gene, and other two were associated with only one gene. After applying LTM, reactions without gene association (exchange reactions) and associated with only one gene were kept the same and other reactions were replaced with an extended subnetwork according to their GPR associations.

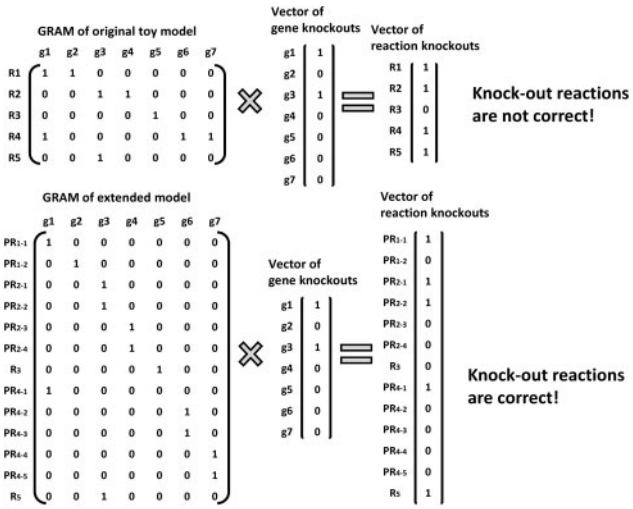

Fig. 3. Comparison between GRAMs of toy model before and after LTM. In vectors of gene or reaction knockouts, if 'g1' or 'R1' is knocked out, the corresponding binary variable is 1; otherwise, it is 0

In the extended model, all reactions were associated with no more than 1 gene. Therefore, the GRAM of the extended model became simple and easy to use for gene target identification. For example, if 'g1' and 'g3' were knocked out, the binary reaction knockout vector could be easily obtained by multiplying GRAM with a binary gene knockout vector. It should be noticed that, incorrect results could be obtained if the same principle is applied to the original GRAM (Fig. 3).

2.2 OptGeneKnock

We developed OptGeneKnock by applying LTM to previously developed OptKnock (Burgard *et al.*, 2003). Before the implementation of OptGeneKnock, GEMs were extended by LTM to obtain the simplified GRAM. After that, an m^*k binary matrix G, which stands for the new GRAM, was obtained. In this matrix, G_{ik} is 1 if the ith reaction was associated with kth gene; otherwise, G_{ik} is 0 (Fig. 3). Therefore, the problem of searching the best gene knockout strategy was converted into a single-level MILP problem that similar to a previously published study (Egen and Lun, 2012):

$$
\begin{aligned}
&\text{Maximize} \quad g'v \\
&\text{Subject to} \quad \sum_{k=1}^{l} x_k \le C \\
&\quad x_k \in \{0,1\} \qquad k = 1, \ldots, l \\
&\quad Sv = 0 \\
&\quad Gx = y \\
&\quad (1-y_i)\alpha \le v_i \le (1-y_i)\beta \qquad i = 1, \ldots, m \\
&\quad f'v = \sum_{i=1}^{m} w_i\beta_i - \mu_i\alpha_i \qquad i = 1, \ldots, m \\
&\quad f_i - \sum_{j=1}^{n} \lambda_j S_{ij} - w_i + \mu_i - \xi_i = 0 \qquad i = 1, \ldots, m \\
&\quad -Dy_i \le \xi_i \le Dy_i \qquad i = 1, \ldots, m \\
&\quad w, \mu \ge 0
\end{aligned}
$$

where g is the target objective vector, whose ith element is the weight of the ith reaction that leading to desired overproducing phenotype; v is the flux distribution vector, whose ith element stands for the flux of the ith reaction; x is the binary gene knockout vector, whose kth element is 1 if the kth is mutated and is 0 otherwise; C is the maximum number of genes that are allowed to be knocked out;

S is the $m*n$ stoichiometric matrix; α, β are the upper and lower bound vectors, whose ith element records the upper and lower bound of flux through the ith reaction, respectively; f is the biological objective vector, whose ith element is the weight of the ith reaction in biological objective; λ is the dual variable for the equality constraints; μ and ν are the dual variables for the lower and upper bounds, respectively; ξ is the dual variable corresponding to the constraint $v_i = 0$ *if* $x_i = 1$ *and* D is a scalar which is set to 100. In this MILP problem, the gene knockout vector, x, was easily converted to a reaction blocking vector, y, by simply multiplying with the G matrix. The problem was then solved by the truncated branch and bound algorithm from a previous study (Egen and Lun, 2012).

2.3 FastGeneSL

We developed FastGeneSL which accelerates the identification of synthetic lethal genes by prescreening the gene candidates by applying LTM to FastSL (Pratapa *et al.*, 2014; Suthers *et al.*, 2009). First, we extended the GEM by LTM to simplify the GPR relationships. As a result, we could easily find genes from reaction sets. Then, the principle of FastSL is applied. The biological objective was optimized, and the flux distribution was obtained. Here, reactions carrying non-zero fluxes, *Rf*, were selected, and then genes associated to these reactions, *Gf*, were also screened. It should be noted that, all essential genes were included in *Gf*. Therefore, it was possible to perform exhaustive gene single mutations within *Gf* and calculate the flux distributions for each mutant. As a result, *Gf* was divided into two parts, Gf_0, if the flux to biological objective is below cutoff (set to 5% of maximum in this study), and Gf_1, otherwise. Thus, Gf_0 is the essential gene group, and Gf_1 is the candidate group for double gene SL analysis. In double gene SL analysis, the flux distributions with single mutation within Gf_1 were selected one at a time. Similar to the previous process, Gf_0' and Gf_1' were screened for every mutant. Consequently, all the synthetic lethal gene pairs were identified by combining each Gf_0' with the selected gene in Gf_1. As described, triple, tetra and high-order synthetic lethal gene sets can be obtained in the same way.

2.4 Models and simulations

An *E.coli* core metabolic model (Orth *et al.*, 2010) was used as a platform for demonstrating the basic feature of the LTM. In addition, GEMs of *E.coli*, iAF1260 (Feist *et al.*, 2007), and *S.cerevisiae*, Yeast 7.11 (Aung *et al.*, 2013), which are manually reconstructed and commonly used models, were directly employed for knockout simulation and genetic SL analysis, respectively. In the case of microaerobic succinate production in *E.coli* GEM iAF1260, the glucose uptake was set to 10 mmol/(gDW*h), and the O_2 intake was set to 5 mmol/(gDW*h). For comparison with OptORF, we also employed a updated *E.coli* GEM from iJR904 (Reed *et al.*, 2003) to simulate ethanol production under anaerobic condition, and the uptake of glucose and O_2 were set to 18.5 mmol/(gDW*h) and 0 mmol/(gDW*h), respectively. In addition, genes encoding ATP synthase were excluded from knockout so that all conditions were consistent with the original study (Kim and Reed, 2010). For the shRNA targets identification, a generic human HCC, liver cancer GEM was used (Agren *et al.*, 2014). However, since there are a lot of gene sets with unknown gene–reaction relationships (isoenzymes or protein complex), in order to get more reliable results, all these relationships were treated as 'or' in HCC GEM to make it a stronger case against shRNA. We also performed toxicity test for the identified targets through the use of GEM for healthy hepatocytes within liver tissue, *iHepatocytes2322* (Mardinoglu *et al.*, 2014) and tested if the identified targets disrupt the biological tasks known to occur in normal human liver tissue. These 256 metabolic tasks were defined based on the known biological functions of the healthy hepatocytes (Mardinoglu *et al.*, 2014).

In this study, Matlab (version 8.0.0.783 (R2012b)) incorporated with COBRA Toolbox 2.0 (Schellenberger *et al.*, 2011) as well as RAVEN Toolbox (Agren *et al.*, 2013) was employed for GEM operation and analysis. Flux balance analysis (FBA) which assumes that the metabolic network is a pseudo steady-state system was used for flux calculation. The Gurobi solver (version 5.6.3, academic) and Mosek solver (version 6) were used to solve all the optimizing problems. All procedures were implemented on a personal computer with 3.00 GHz Intel(R) Core(TM) i7-3540M CPU and 8.00 GB RAM.

3 Results

3.1 Comparison between original model and the extended logic model

To illustrate the effect of LTM in the model content, we made direct comparison between the original model and the extended logic model in several aspects.

First, we compared the number of reactions, metabolites and genes in five models used in this study before and after applying LTM. As shown in Table 1, the numbers of reactions and metabolites were increased around 2- to 5-fold, while the number of genes was kept constant as expected. The explanation to this difference in fold changes of numbers of reactions and metabolites among different GEMs was that the models have different prevalence of protein complexes and isoenzymes reactions (Supplementary Table S1). LTM only works on reactions controlled by genes, thus reactions with more complex gene associations will be extended more. It should be noticed that, HCC GEM and *iHepatocytes2322* had more significant size increase compared with models of microorganisms, which could be explained by the superior complexity of human GPRs. However, the size of HCC GEM was still larger than the *iHepatocytes2322* despite the small gene and original reaction numbers which is caused by the different way of GPRs interpretation.

Second, differences in FBA performance between the original and extended models were observed. We did singular value decomposition (SVD) to the *E.coli* core model and its extended version and the singular values were ranked (Fig. 4). It has been shown that the extended model had bigger singular values, which implies a higher degree of freedom in the solution space. Explanation to this is that the split of isoenzyme associated reactions, which offered alternative

Table 1. Model sizes before and after LTM

Models	Type	Reaction number	Metabolite number	Gene number
E.coli core model	Original	95	72	137
	Extended	337	273	137
iAF1260	Original	2382	1668	1261
	Extended	7287	5372	1261
Yeast 7.11	Original	3490	2220	910
	Extended	6780	4742	910
iHepatocytes2322	Original	7930	5686	2322
	Extended	20 787	18 097	2322
HCC GEM	Original	4820	4099	1779
	Extended	26 282	18 030	1779

'Original' represents the original model, and 'Extended' represents the logic model after LTM.

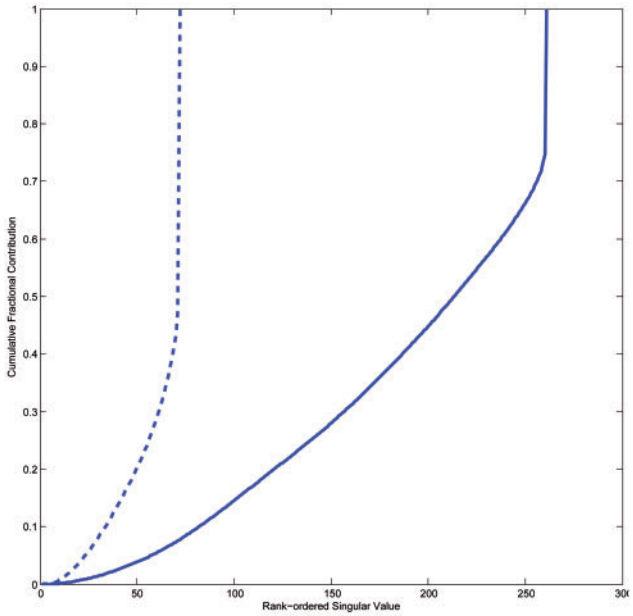

Fig. 4. Cumulative fractional contribution versus rank-ordered singular value. Dashed line represents the behavior of original model, whereas solid line represents the extend model's

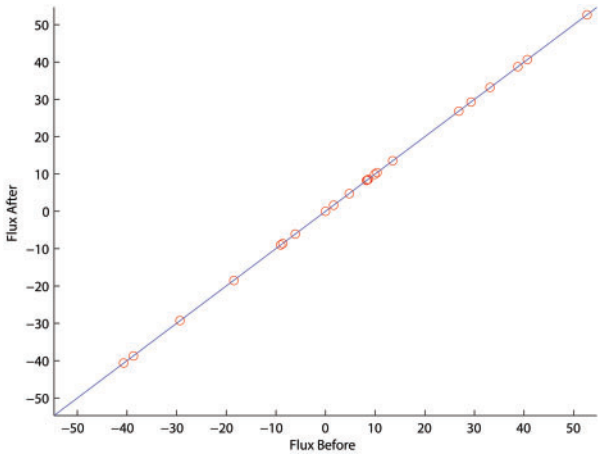

Fig. 5. Flux comparison of unchanged reactions between *E.coli* core models before and after LTM

solutions. However, this would not affect the fluxes of the unchanged reactions (as shown in Fig. 5).

At last, the GPRs of the extended model were validated by double gene deletion analysis. Growth rate of all combinatory knockouts of two genes were calculated for both models, and the result of extended model was exactly the same as that of the original model (Supplementary Fig. S2), which clearly showed that GPRs between both models are equivalents.

Therefore, as described above, it was concluded that LTM increases the number of reactions and metabolites, but it kept the major characteristics of flux distribution and GPRs of the original model.

3.2 Case study 1: incorporating OptGeneKnock with truncated branch and bound for fast screening of gene knockouts

Logic models extended by LTM have simplified GRAM, and thus enabled identification of gene targets of many previously developed bilevel MILP-based methods. In this study, we developed a new method, named OptGeneKnock, by straightforwardly integrating LTM with a previous method (Egen and Lun, 2012). The previous one could only deal with predictions of gene set knockouts, which is actually discovered by reaction knockouts. Searching for the optimal knockout strategies of reactions (gene sets) and individual genes are two independent questions (Supplementary Fig. S3). In addition, only reactions with exactly the same gene sets were simultaneously blocked, which means that the comprehensive GPRs were not correctly captured. However, when using OptGeneKnock, these problems are solved since reactions are only associated with no more than one gene. Therefore, the *in silico* gene knockouts can be performed by solving the MILP problem with truncated branch and bound.

Table 2. Predicted gene knockout strategies for ethanol and succinate production in *E.coli*

Case	C	Time	Knockout genes	N	G	P
Ethanol	1	20	*pntB*	1	0.46	17.98
	2	20	*focA, focB*	1	0.42	30.74
	3	100	*focA, focB, pgi*	2	0.24	33.47
	4	200	*focA, focB, pgi, ptsH*	16	0.19	34.22
	5	200	*pflA, focB, pgi, ptsH, gdhA*	17	0.17	34.47
Succinate	1	100	*sdhD*	1	0.40	0.00
	2	100	*sdhD*	1	0.40	0.00
	3	800	*pykA, pykF, ptsI*	18	0.21	0.41
	4	3100	*pta, eutD, sdhD, atpF*	4	0.14	5.55
	5	3400	*ptsI, focA, focB, pykA, pykF*	20	0.19	8.32

In the 'Case' column, 'Ethanol' means anaerobic ethanol production in updated iJR904, while 'Succinate' stands for microaerobic succinate production in iAF1260. 'C' represents the maximum number for knockout. 'Time' means computational time for obtaining the strategy, and the unit is second. 'N' represents the number of reactions blocked by the knockout strategies. 'G' and 'P' stands for the growth and succinate production rates of the knockout mutants, respectively, and the units are mmol/(gDW*h).

Using the extended logical model of *E.coli*, we have identified gene knock-out strategies (one to five knockouts) for anaerobic ethanol and microaerobic succinate overproduction with OptGeneKnock (Table 2). Unlike most of the previous methods (OptKnock *et al.*), the strategies obtained by OptGeneKnock were all actual gene targets, which are user friendly for *in vivo* validation. Interestingly, the single deletion of *pntB* (encoding pyridine nucleotide transhydrogenase beta subunit) will significantly improve the production of ethanol. In *E.coli*, *pntA* and *pntB* encode the alpha and beta subunit of membrane-bound pyridine nucleotide transhydrogenase, which couples the reversible reduction of NADP by NADH. Thus, *pntA* and *pntB* involve in two reactions: NADTRHD and THD2. While *sthA* encodes a soluble pyridine nucleotide transhydrogenase by the oxidation of NADPH. Therefore, the deletion of *pntB* will only block the reaction THD2, and lead to the overproduction of ethanol (Table 2). In the case of double deletion, both *focA* and *focB* encode the formate transporters. The deletion of *focA* and *focB* will block reaction FORt, and thus block the transport of formate into cytosol. The best single and double gene knock-out strategies for ethanol production both resulted in blocking of one reactions, and this clearly showed the difference between reaction and gene level mutation. In addition, the optimal 5 gene knockout strategy for succinate overproduction obtained resulted in simultaneous blocking of 20 reactions, which is quite difficult to be predicted by other reaction-based methods. Mutations of *ptsI*, *pykA* and *pykF* cut off all direct transformation from PEP to pyruvate, which shifted the flux from pyruvate to the tricarboxylic

acid (TCA) cycle and hereby enhanced the production of succinate. On the other hand, knockouts of *focA* and *focB* blocked the flux to formate as byproduct. Since formate is not a common byproduct of *E.coli*, the knockout of these two genes may be less important than the others. This complex gene knock strategies obtained by OptGeneKnock further proved that the optimal gene-level knockouts are not limited by the number of reactions affected.

In addition, the results highlighted the great efficiency of OptGeneKnock. Most of the strategies were obtained within minutes, and none of them exceeded 1 h. In addition, our ethanol-producing strategies could result in better solutions (even global optimal solutions for levels 1–level 3) in all conditions in comparison to those obtained by OptORF without regulation (Supplementary Table S2). Therefore, we concluded that OptGeneKnock had less computational cost and found better solutions. Thus, the development of OptGeneKnock gave a successful example showing that LTM could extend previously developed reaction-based bilevel MILP method to gene level applications.

3.3 Comparison between FastGeneSL and exhaustive search

We also developed a method called FastGeneSL for efficient gene SL analysis by combining LTM with FastSL which is a previous reaction-based method (Pratapa *et al.*, 2014). GEMs pretreated by LTM which simplifies the GPRs is a prerequisite for FastGeneSL. Firstly, we performed both FastGeneSL and an exhaustive search for double gene SL analysis of iAF1260 and Yeast 7.11. Compared to the exhaustive algorithm, FastGeneSL was nearly 20 times faster while obtaining the same results (Supplementary Table S3). In addition, triple gene SL analysis was implemented by FastGeneSL for both GEMs as well as GEM for HCC, and the computational costs of exhaustive search were calculated (Table 3). FastGeneSL was much faster than exhaustive search (up to more than 100 times) and thus enabled triple gene SL analysis for large GEMs. Furthermore, FastGeneSL showed an increased improvement in both double and triple SL analysis for Yeast 7.11 (910 genes), iAF1260 (1261 genes) and HCC GEM (1779 genes), which suggested it can be applied to more comprehensive models. Therefore, FastGeneSL is another successful example showing that LTM could extend the use of existing reaction-based methods to gene level analysis.

3.4 Case study 2: identification of shRNA/CRISPR targets for cancer therapy with FastGeneSL

We exhibited the applicability of FastGeneSL in novel shRNA and CRISPR targets identification. These both techniques are used in suppression of the expression of genes in biotechnology applications. Gene SL analysis could help us find potential shRNA/CRISPR targets that can inhibit or kill the growth of the tumor while keep normal cells alive. In this study, we used HCC GEM for identifying the shRNA/CRISPR targets (Agren *et al.*, 2014) and performed a toxicity test using healthy hepatocytes model *iHepatocytes2322* (Mardinoglu *et al.*, 2014).

Essential genes as well as double and triple synthetic lethal genes sets were identified by applying LTM and FastGeneSL to HCC GEM. 67 essential genes as well as 85 double and 175 triple synthetic lethal gene sets were found to inhibit the growth of HCC tumors. We also performed *in silico* toxicity test to reveal the effect of these identified targets in healthy hepatocytes. We found that 34 of the essential genes as well as 44 of double and 95 of triple synthetic lethal gene sets did not disrupt any of the 256 biological functions that are known to occur in healthy hepatocytes

Table 3. Computational cost of FastGeneSL and exhaustive algorithm

Model	SL level	CPU time of exhaustive search	CPU time of FastGeneSL
Yeast 7.11	Double	~8.0 hours	~0.5 hours
	Triple	~101.4 days*	~3.7 days
iAF1260	Double	~15.4 hours	~0.8 hours
	Triple	~270.1 days*	~2.0 days
HCC model	Double	~30.8 hours*	~2.0 hours
	Triple	~759.0 days*	~6.1 days

Data marked with '*' means time is estimated by the number of linear programs (LP) required for exhaustive search. Each LP takes an average of 0.07 s in this estimation. All results were validated with original GEMs, and the results were consistent before and after LTM.

(Supplementary Table S4). Hence, we suggested that these identified targets can be used for designing novel effective cancer treatment strategies.

The result of gene SL analysis also provided biological insight into the metabolism of HCC. In cancer cells, glutamine can be converted to citrate via conversions to glutamate, α-ketoglutarate and isocitrate which partially reverses the TCA cycle in the process. This glutamine-derived citrate could then be incorporated in the fatty acid biosynthesis (FAB) which is an essential pathway for tumor growth. Through our analysis, we found that the targeting genes involved in glutaminolysis including glutamate dehydrogenases (GLUD1 and GLUD2), glutamic-oxaloacetic transaminase (GOT2), glutamic-pyruvate transaminase (GPT1) as well as glutaminase (GLS) in synthetic lethal gene sets can be used to inhibit or kill the growth of HCC tumors.

We identified the citrate synthase (CS), pyruvate carboxylase (PC) and pyruvate dehydrogenase (PDHA2) in synthetic lethal gene sets as potential targets for inhibition of the HCC tumor growth. Pyruvate transported into the mitochondria is converted to acetyl-CoA by PC and PDHA2. Next acetyl-CoA is converted to citrate by CS and exported to the cytosol where citrate is reconverted to acetyl-CoA. This cytosolic acetyl-CoA can then enter the FAB and thus serve as a substrate for the production of lipids in HCC. Previously, it has been shown that fatty acid synthase which overexpressed in tumor cells is a general feature for tumor cells (Flavin *et al.*, 2010). There are also mounting evidence to suggest that altered lipid metabolism is a common feature of cancer cells (Zhang and Du, 2012).

FAB requires large amounts of NADPH and the increased activity of the pentose phosphate (PP) pathway would indeed satisfy the increased demand for reducing power resulting from induction of FAB. We identified genes involved in PP pathway including glucose-6-phosphate dehydrogenase (G6PD), transaldolase 1 (TALDO1) and transketolase (TKT) as targets for inhibiting/killing the growth of HCC tumors.

We also identified the genes involved in fatty acid beta-oxidation, cholesterol biosynthesis, amino acid synthesis and amino acid transportation that are known to have major roles in the formation of biomass components as anticancer drug targets. Previously, anticancer drug targets involved in the similar pathways were identified in HCC tumors (Agren *et al.*, 2014) and our analysis confirmed the potential use of previously identified targets. In addition, SLC38A4 which controls the transportation of amino acids, especially cationic amino acid was identified as anticancer drug target. This implicated that HCC tumors had a specific amino acid metabolism that is different from other cancer cells. We observed that the predicted anticancer drug targets for inhibiting/killing the growth of HCC tumors

correctly captured the general feature of HCC through the use of our method.

4 Discussion

In this article, we developed a method, called LTM, which transforms a GEM to a logically equivalent extended model to enable gene level applications. The transformation enlarges the size of GEMs while keeps the correct GPRs, and the FBA results of extended models are equivalent to the original models despite the increased degree of freedom.

We showed the applicability of our method in two different biotechnology applications. We first showed the utility of LTM in optimization frameworks. The optimization frameworks could be easily applied to genetic solution prediction by incorporating with LTM. This implicated the wide application of LTM with many well designed methods such as OptKnock and OptForce (Burgard *et al.*, 2003; Ranganathan *et al.*, 2010). We developed OptGeneKnock by combining OptKnock and LTM as well as truncated branch and bound algorithm, and showed its capability in designing gene knockout strategies leading to overproduction of desired biochemical.

In addition, we developed a gene SL analysis method called FastGeneSL. This method combined LTM and a previous method FastSL, and enabled efficient genet SL analysis in GEMs for different organisms. We demonstrated that the efficiency was greatly improved by FastGeneSL and this enable identification of triple SL genes in GEMs which was previously infeasible. Next, by employing FastGeneSL, we identified novel anticancer drug targets that can be used for inhibition of the HCC tumors and performed *in silico* toxicity test for these identified targets. We found the most prominent distinguishing metabolic features of HCC and observed that here identified targets can be used in the drug development process for efficient treatment of HCC tumors.

By the two biotechnology applications tested in this study, we could conclude that although the principle of LTM is simple, it greatly benefits gene level analysis and applications. In addition, many previous methods (e.g. OptKnock and FastSL) could be extended to gene level applications (e.g. OptGeneKnock and FastGeneSL) by incorporating with LTM in a straightforward way.

Furthermore, as GPRs are simplified after LTM, complex gene set controlled reactions are avoided and it becomes very convenient to integrate different omics data into extended models. Thus, LTM could also be used together with multiomic data integrating methods like GIMME and PROM (Becker and Palsson, 2008; Chandrasekaran and Price, 2010). In total, we could conclude that LTM is a useful tool which facilitates the application of GEMs and we are therefore confident that it will find wide use in the community.

Acknowledgements

We thank the GUROBI and MOSEK group for providing the Gurobi and Mosek solver.

Funding

Knut and Alice Wallenberg Foundation, the Bill and Melinda Gates Foundation and the European Research Council (247013); National Basic Research Program of China (973 Program) (2012CB721101); China Scholarship Council.

Conflict of Interest: none declared.

References

Agren,R. *et al.* (2013) The RAVEN toolbox and its use for generating a genome-scale model for *Penicillium chrysogenum*. *PLoS Comput. Biol.*, **9**, e1002980.

Agren,R. *et al.* (2014) Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.*, **10**, 721.

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

Barabasi,A.L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Becker,S.A. and Palsson,B.O. (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.*, **4**, e1000082.

Bernards,R. *et al.* (2006) shRNA libraries and their use in cancer genetics. *Nat. Methods*, **3**, 701–706.

Brochado,A.R. *et al.* (2010) Improved vanillin production in baker's yeast through in silico design. *Microb. Cell Fact.*, **9**, 84.

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.

Cho,S.W. *et al.* (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.

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

Chong,S.K. *et al.* (2014) A hybrid of ant colony optimization and minimization of metabolic adjustment to improve the production of succinic acid in *Escherichia coli*. *Comput. Biol. Med.*, **49**, 74–82.

Choon,Y.W. *et al.* (2014) A hybrid of bees algorithm and flux balance analysis with OptKnock as a platform for in silico optimization of microbial strains. *Bioprocess Biosyst. Eng.*, **37**, 521–532.

Egen,D. and Lun,D.S. (2012) Truncated branch and bound achieves efficient constraint-based genetic design. *Bioinformatics*, **28**, 1619–1623.

Feist,A.M. and Palsson,B.O. (2010) The biomass objective function. *Curr. Opin. Microbiol.*, **13**, 344–349.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Flavin,R. *et al.* (2010) Fatty acid synthase as a potential therapeutic target in cancer. *Future Oncol.*, **6**, 551–562.

Fowler,Z.L. *et al.* (2009) Increased malonyl coenzyme A biosynthesis by tuning the Escherichia coli metabolic network and its application to flavanone production. *Appl. Environ. Microbiol.*, **75**, 5831–5839.

Hamilton,J.J. *et al.* (2013) Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. *Biophys. J.*, **105**, 512–522.

Harrison,R. *et al.* (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc. National Acad. Sci. USA*, **104**, 2307–2312.

Hoppe,A. *et al.* (2007) Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst. Biol.*, **1**, 23.

Jerby-Arnon,L. *et al.* (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, **158**, 1199–1209.

Ji,B. and Nielsen,J. (2015) New insight into the gut microbiome through metagenomics. *Adv. Genomics Genet.*, **5**, 77–91.

Kim,H.J. *et al.* (2013) Genome-wide analysis of redox reactions reveals metabolic engineering targets for D-lactate overproduction in *Escherichia coli*. *Metab. Eng.*, **18**, 44–52.

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

Lun,D.S. *et al.* (2009) Large-scale identification of genetic design strategies using local search. *Mol. Syst. Biol.*, **5**, 296.

Mardinoglu,A. *et al.* (2013a) Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol. Syst. Biol.*, **9**, 649.

Mardinoglu,A. *et al.* (2013b) Genome-scale modeling of human metabolism—a systems biology approach. *Biotechnol. J.*, **8**, 985–996.

Mardinoglu,A. and Nielsen,J. (2012) Systems medicine and metabolic modelling. *J. Intern. Med.*, **271**, 142–154.

Mardinoglu,A. and Nielsen,J. (2015) New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.*, **34C**, 91–97.

Mardinoglu,A. *et al.* (2014) Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.*, **5**, 3083.

Matsuda,F. *et al.* (2011) Engineering strategy of yeast metabolism for higher alcohol production, *Microbial Cell Factories*, **10**, 70.

Nam,H. *et al.* (2014) A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks. *PLoS Comput. Biol.*, **10**, e1003837.

Nocon,J. *et al.* (2014) Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production. *Metab. Eng.*, **24**, 129–138.

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

Patil,K.R. *et al.* (2005) Evolutionary programming as a platform for in silico metabolic engineering, *BMC Bioinform.*, **6**, 308.

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

Pratapa,A. *et al.* (2014) Fast-SL an efficient algorithm to identify synthetic lethal reaction sets in metabolic networks. arXiv: 1406.6557v2 [q-bio.MN]. Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, INDIA; Department of Computer Science and Engineering, Indian Institute of Technology Madras.

Ranganathan,S. and Maranas,C.D. (2010) Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions. *Biotechnol. J.*, **5**, 716–725.

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

Ranganathan,S. *et al.* (2012) An integrated computational and experimental study for overproducing fatty acids in Escherichia coli. *Metab. Eng.*, **14**, 687–704.

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

Reznik,E. *et al.* (2013) Flux imbalance analysis and the sensitivity of cellular growth to changes in metabolite pools. *PLoS Comput. Biol.*, **9**, e1003195.

Ryan,C.J. *et al.* (2014) DAISY: picking synthetic lethals from cancer genomes. *Cancer Cell*, **26**, 306–308.

Schellenberger,J. *et al.* (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.*, **6**, 1290–1307.

Segre,D. *et al.* (2005) Modular epistasis in yeast metabolism. *Nat. Genet.*, **37**, 77–83.

Suthers,P.F. *et al.* (2009) Genome-scale gene/reaction essentiality and synthetic lethality analysis, *Mol. Syst. Biol.*, **5**, 301.

Tepper,N. *et al.* (2013) Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load. *PloS One*, **8**, e75370.

Xu,P. *et al.* (2011) Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.*, **13**, 578–587.

Yizhak,K. *et al.* (2014) A computational study of the Warburg effect identifies metabolic targets inhibiting cancer migration. *Mol. Syst. Biol.*, **10**, 744.

Zhang,F. and Du,G. (2012) Dysregulated lipid metabolism in cancer. *World J. Biol. Chem.*, **3**, 167–174.