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

Cheng Zhang1,2, Boyang Ji2, Adil Mardinoglu2, Jens Nielsen2,* and Qiang Hua1,3,*

1State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China, 2Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Göteborg, Sweden and 3Shanghai 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-Aron 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; Vizhak 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 ith elements is 1 if rth reaction is associated with jth 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 (UBoriginal) and lower (LBoriginal) bounds with the original reaction, while the bounds of reactions inside the pseudo network were set differently. If UBoriginal > 0 and LBoriginal > 0, then for all inner pseudo reactions, UB = UBoriginal, LB = 0; if UBoriginal < 0 and LBoriginal < 0, then UB = 0, LB = LBoriginal; otherwise, UB = UBoriginal, LB = LBoriginal.
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 sides 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. Thus, all reactions that are associated with more than one gene in the GEM were broken down into pseudo reactions.

Three intracellular reactions were associated with more than one gene, and other two were associated with only one 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 \( n \times k \) binary matrix \( G \), which stands for the new GRAM, was obtained. In this matrix, \( G_{ik} = 1 \) if the \( k \)th reaction was associated with \( i \)th gene; otherwise, \( G_{ik} = 0 \) (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):

Maximize \( \mathbf{g}^\top \mathbf{v} \)

Subject to \( \sum_{k=1}^{l} x_k \leq C \)

\( x_k \in \{0, 1\} \quad k = 1, \ldots, l \)

\( S_{ij} = 0 \)

\( G_{ik} = y \)

\( (1 - y_{ij})x \leq v_i - (1 - y_{ij})\beta \quad i = 1, \ldots, m \)

\( f_{ij} = \sum_{i,j=1}^{m} w_{ij}x_i \leq \mu_i x_i \quad i = 1, \ldots, m \)

\( -D_{ij} \leq x_i \leq D_{ij} \quad i = 1, \ldots, m \)

\( \mathbf{w}, \mathbf{\mu} \geq 0 \)

where \( g \) is the target objective vector, whose \( r \)th element is the weight of the \( r \)th reaction that leading to desired overproducing phenotype; \( v \) is the flux distribution vector, whose \( r \)th element stands for the flux of the \( r \)th reaction; \( x \) is the binary gene knockout vector, whose \( k \)th element is 1 if the \( k \)th 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 rth element records the upper and lower bound of flux through the rth reaction, respectively; f is the biological objective vector, whose rth element is the weight of the rth reaction in biological objective; 𝜀 is the dual variable for the equality constraint; μ and ν are the dual variables for the lower and upper bounds, respectively; ξ is the dual variable corresponding to the constraint ν2 = 0 if xi = 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, GF0 if the flux to biological objective is below cutoff (set to 5% of maximum in this study), and GF1, otherwise. Thus, GF0 is the essential gene group, and GF1 is the candidate group for double gene SL analysis. In double gene SL analysis, the flux distributions with single mutation within GF1 were selected one at a time. Similar to the previous process, GF0′ and GF1′ were screened for every mutant. Consequently, all the synthetic lethal gene pairs were identified by combining each GF0′ with the selected gene in GF1. 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 (Aug 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 O2 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 O2 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 the split of isoenzyme associated reactions, which offered alternative

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

‘Original’ represents the original model, and ‘Extended’ represents the logic model after LTM.
enabled identification of gene targets of many previously developed logic models extended by LTM have simplified GRAM, and thus of gene knockouts with truncated branch and bound for fast screening.

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.

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

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

In the ‘Case’ column, ‘Ethanol’ means anaerobic ethanol production in updated jR904, 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).
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 1h. 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 applications.

### 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.
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 multomic data integrating methods like GIMME and PROM (Becker and Palsson, 2008; Chandrasekaran et al., 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

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.
Matsuda, F. et al. (2011) Engineering strategy of yeast metabolism for higher alcohol production, Microbial Cell Factories, 10, 70.
Patil, K.R. et al. (2005) Evolutionary programming as a platform for in silico metabolic engineering, BMC Bioinformatics, 6, 308.
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.