Predicting the carbon source for *Bacillus subtilis* by integrating gene expression profiles into a constraint-based metabolic model

Kulwadee Thanamit[1], Franziska Hoerhold[1], Marcus Oswald[1], Rainer Koenig[1*]

[1]Systems Biology Research Group, Center for Sepsis Control and Care, Jena University Hospital, Kollegiengasse 10, 07743, Jena, Germany

* Corresponding author

E-mail: Rainer.Koenig@uni-jena.de (RK)

Abstract

Finding drug targets for antimicrobial treatment is a central focus in biomedical research. To discover new drug targets, we are interested in finding out which nutrients are essential for pathogenic microorganisms in the host or under specific circumstances. Besides, metabolic fluxes have been successfully constructed and predicted by employing flux balance analysis (FBA) technique. While ^{13}C metabolic data is the most informative way to explore metabolism, the data can be difficult to acquire in complicated environments, for example, osteomyelitis from *S. aureus*. On the other hand, although gene expression data is less informative in this case as compared to ^{13}C metabolic data, it is easier to generate, and it still provides us informative insights. We develop FBA models using the stoichiometric knowledge of the metabolic reactions of a cell and combine them with gene expression profiles. We aim to identify essential drug targets for specific nutritional uptakes of pathogenic microorganisms. As a case study, we implemented our method by applying data from *B. subtilis* to predict carbon sources based on given gene expression profiles. We validated our flux prediction results by comparing with ^{13}C metabolic flux data. With our method, we efficiently utilized gene expression profiles to predict carbon sources and investigate the metabolic network of *B. subtilis*. We show that our method is promising, generalizable, and versatile. We present that using FBA model with gene expression data is a good starting point to support subsequent hypotheses to conduct further studies; especially, in the environment that ^{13}C metabolic flux data is hard to achieve. Besides, from a technical aspect, our method performed faster in order to remove thermodynamically infeasible loops as compared to loopless COBRA (ll-COBRA), which is the well-established method in the community.

Keywords: flux balance analysis, mixed integer linear programming, bacillus subtilis, carbon source

Introduction

Understanding mechanisms behind metabolism systemically is one of the significant challenges in systems biology. With aid from high-throughput technology, it evolutionarily changed the way we investigated metabolism. Different types of data, such as transcriptomics, proteomics, metabolomics, and fluxomics, have been generated in a large number in such a short time as compared to the past. It is undeniable that the growth of data greatly contributes to elucidate cellular metabolism. Besides filling gaps of knowledge, it is our mission as computational biologists to systemically explore and connect missing links between different types of data to gain more insights to unravel key players behind metabolism.

Over the past years, constraint-based modeling has been successfully constructed and employed by flux balance analysis (FBA) technique [1-3]. FBA has been served as a tool to provide *in silico* simulations which allows researchers to discover new targets and support further studies. It has been applied widely to model metabolic fluxes in a scalable way; especially, in the field of metabolic engineering or agriculture to enhance the yield of interested products [4-7]. The concept of FBA is to optimize a defined objective function subjected to a large set of constraints mainly based on the known stoichiometry of the biochemical reactions. A typical objective function of e.g. microorganisms or tumor cells is to maximize biomass production. Assuming a steady state, FBA enables us to circumvent embedding of detailed knowledge about the reaction kinetics, which is difficult to obtain, particularly in a large, genome scale. However, we often see that utilizing only stoichiometric data of cell reactions is inadequate to achieve good phenotype predictions for a specific condition. Therefore, transcriptional regulation has been in the spotlight for

computational biologists to utilize this information [8-11], as this experimental data is much easier to obtain than metabolic flux data from ^{13}C isotopic tracer analyses [12-14]. However, using gene expression profiles for estimating metabolic fluxes has been a matter of debate and discrepancies compared to ^{13}C metabolic flux data have been reported [15, 16].

The concept of systems biology together with constraint-based modeling has also been introduced in health sciences. Systemic approach has been adapted to aid researchers to study pathogenicity of diseases to find new drug targets [17-20]. In every organism, having adequate nutrients is significant to maintain their metabolic functions. This also applies to pathogens. If access to nutrition is inhibited or enzymes involved in nutrient utilization are blocked, pathogens will perish. Although the idea is simple, the practice is challenging; especially, in a complex environment. For example, osteomyelitis, which is a bone infection, it is mainly caused by *Staphylococcus aureus* (*S. aureus*) [21-23]. The bacteria express adhesins to promote adherence to bone structures such as osteoblasts or collagen. This leads to biofilm formation and bone destruction [23-25]. Due to the structure of bone and biofilm, the infection is very persistent which causes problems for orthopedics and negatively impacts on quality of life of the patient. Furthermore, antibiotic resistance also contributes to worsening the situation [26, 27]. Understanding host-pathogen interaction will certainly lead us to a way to a successful treatment. To achieve a novel treatment strategy, a good model is required. As mentioned, although ^{13}C metabolic data is very detailed, it is difficult to achieve and interpret results from ^{13}C isotopic tracer analysis; particularly, under this complicated environment [24, 28]. Meanwhile, gene expression data is easy to acquire in clinical settings. It is possible to integrate this information into a metabolic network to study host-pathogen interactions to

find interesting targets. Thus, we propose a novel mechanistic approach employing FBA prediction by integrating gene expression profiles.

As a case study, we demonstrated our concept by using gene expression data from Gram-positive bacterium *Bacillus subtilis* (*B. subtilis*) growing in different carbon sources [29, 30] to predict carbon sources based on the transcriptional profiles of the investigated metabolic enzymes. We validated our flux prediction results by comparing them with ^{13}C metabolic flux data from the same condition [15, 29]. As shown in previous studies, transcriptional regulation generally played a major role in enzymes involved in substrate uptakes [15]; this was seen from agreement between gene expression data and ^{13}C metabolic flux data. With our method, we were able to predict carbon sources and investigate the metabolic network of *B. subtilis* by using only gene expression profiles. We show that our method is promising, generalizable, and versatile. Moreover, from a technical aspect, our method performed faster to remove thermodynamically infeasible loop (TIL) as compared to loopless COBRA (ll-COBRA) [31], which is the well-established method in the community. We believe that our approach will be a good starting point to explore cell metabolism to generate new hypotheses for further studies; especially, in the environment that ^{13}C metabolic flux data is hard to achieve.

Materials and Methods

1. Data assembly

1.1. Gene expression data

Published microarray gene expression datasets of *B. subtilis* strain BSB1, which is a derivative strain of 168 trpC2 strain, were used in this study. The tilling arrays were

designed to cover the whole genome of *B. subtilis* (GenBank: AL009126) [29, 30]. Gene expression data of *B. subtilis* grown in LB medium was used in generating a list of gene-reaction pairs for mapping gene expression profiles to fluxes process [30].

For a training procedure, gene expression data of *B. subtilis* grown in minimal medium in eight different carbon source conditions (glucose, fructose, gluconate, glutamate/succinate, glycerol, malate, malate/glucose, pyruvate) [30] was used to train a metabolic network of the same organism. For gene expression data of *B. subtilis* grown in LB medium and minimal medium in eight different carbon source conditions, the data was provided as an online supplementary (Table S2) from the publication [30].

Besides, we applied a time-series gene expression data [29] with two nutrient shift scenarios (glucose to glucose plus malate, malate to glucose plus malate) as a validation dataset. *B. subtilis* were grown in minimal medium on a single substrate until an OD_{600} of 0.5. Then, the other substrate (glucose or malate) was added to the culture to assess the bacteria behavior after the nutrient shift. The data was provided on BaSysBio database (https://basysbio.ethz.ch/openbis/basysbio_openbis.html).

1.2. ^{13}C metabolic flux data

^{13}C metabolic flux data were used to validate flux prediction results from our approach. For the training dataset, we used ^{13}C metabolic flux data from Chubukov *et al* [15] using the similar eight conditions (glucose, fructose, gluconate, glutamate/succinate, glycerol, malate, malate/glucose, pyruvate) as in gene expression data [30]. The data was downloaded from Supplementary Table S4 as supporting information for the

publication. For the validation dataset, ^{13}C metabolic flux data from Buescher *et al* [29] was also generated under the same settings (glucose to glucose plus malate, malate to glucose plus malate). The data was accessible from the same repository as in gene expression data from the same publication (see 1.1 Gene expression data). In both datasets, ^{13}C isotopic tracer experiments were performed to obtain metabolite abundance. Later, the data were fitted to the metabolic network of *B. Subtilis* to find the best-fitted set of fluxes as mentioned in the original publications [15, 29].

2. Data pre-processing

Pre-processed gene expression datasets were provided in log2 transformation. We rechecked and matched BSU number with Gene ID using data from Uniprot [32], KEGG [33-35], bioDBnet [36] and the literature [37]. For the training data, each condition had three biological replicates. We averaged gene expression values for each gene per condition and used the averaged values in our mapping procedure. However, for the validation dataset, we performed an additional step before calculating average values since some timepoints contained only two biological replicates, while the majority of timepoints contained three biological replicates. We clustered all timepoint by calculating Euclidian's distances to select only two closely related replicates for each timepoint per scenario. Then, we averaged gene expression values for each timepoint per scenario.

^{13}C metabolic flux data from Chubukov *et al* [15] and Buescher *et al* [29] were used as published without pre-processing step.

3. Model building

3.1. Metabolic network and work environment transfer

B. subtilis 168 metabolic network [38] was acquired from BiGG Models database (BiGG ID iYO844) [39]. To be able to work on the same environment for every dataset and use established tools from our group, we transferred the metabolic model (SBML format) from MATLAB programming environment (www.mathworks.com) to R programming environment (www.r-project.org).

In our model, we allowed a small amount of sum flux value from other external reaction fluxes, for example, amino acids, to enter the system to provide relaxation to the optimization and avoid infeasible solution. Based on ^{13}C metabolic flux data from Chubukov *et al* [15], we set the value to the minimum possible sum flux value from all eight conditions to 0.6878778.

Apart from this step, all computation in this study was performed only in R. Gurobi optimizer (www.gurobi.com) was used as a numerical solver to solve any optimization problem occurring in the study.

3.2. Selection of gene-reaction pairs

In this study, we mainly focused on the central energy metabolism (glycolysis, tricarboxylic acid cycle (TCA cycle), pentose phosphate pathway (PPP), urea cycle). We included all 40 reactions and genes mentioned in the publication of Chubukov *et al* [15]. These 40 reactions which had ^{13}C metabolic flux data, were our core reactions. We also extended our model by including neighbor reactions, which are related to

these 40 reactions but contained no ^{13}C metabolic flux data in our study. In the end, we had 98 reactions and 140 gene-reaction pairs. It was possible to pair one reaction to more than one gene if genes were co-expressed.

For neighbor gene-reaction pairs, we carefully selected gene-reaction pairs which showed a tendency to aid the mapping process. We performed T-Test between gene expression values of *B. Subtilis* grown in LB medium and in minimal medium with specific carbon sources (glucose, fructose, gluconate, glutamate/succinate, glycerol, malate, malate/glucose, pyruvate) to select only gene-reaction pairs, which were at least significantly up-regulated or down-regulated in one condition to be included in the gene-reaction list. P-value cutoff was 0.05. BH method was used to adjust p-value for the analysis [40]. The list of gene-reaction pairs in this study is provided in TableS.1 in Supplementary information.

3.3. Flux balance analysis (FBA)

In constraint-based modeling, FBA is a well-known analytical method to study a flux state in the metabolic network. At a steady state, FBA employs only stoichiometric data to seek an optimal solution lied within a solution space by optimizing (minimization or maximization) a defined objective function [2]. The objective function is formulated as a linear optimization problem as shown below.

$$\text{Maximize/Minimize } c_r^T v_r \qquad (1)$$

subjected to

$$\sum_r S_r \cdot v_r = 0 \quad (2)$$

$$lb_r \leq v_r \leq ub_r \quad (3)$$

where S_r represents a stoichiometric matrix of metabolite and reaction, and v_r is a flux variable for each reaction r with lower bound lb_r and upper bound ub_r. c_r^T is a vector of optimization coefficients. For the interested reaction, c_r^T is set to 1, while c_r^T for other reactions is set to 0. In growth simulation, c_r^T for biomass reaction is normally set to 1.

Besides, after acquiring the optimal solution, it is possible to explore suboptimal states by constraining the objective function.

$$c_r^T v_r \geq \gamma \cdot Z \quad (4)$$

where $Z = c_r^T v_r$ is an optimal solution to equation (1) and γ is a parameter that forces the analysis to be done in suboptimal states ($0 \leq \gamma < 1$). However, when information is available, for example, a growth rate for a biomass constraint, it is possible to replace $\gamma \cdot Z$ with this value to confine the solution space to obtain more realistic flux distribution.

3.4. Flux variability analysis (FVA)

Apart from FBA, FVA is another widely used technique in constraint-based modeling. Instead of optimizing biomass production or substrate reaction like in FBA, FVA minimizes or maximizes flux in each reaction at a time for the entire metabolic network

to determine a range of the minimum and maximum possible flux for each reaction. It serves as a powerful method to determine robustness of the metabolic network and identify important reactions [2, 41, 42].

For each reaction r in a metabolic network, FVA minimizes or maximizes flux v_r to find the range of flux while satisfies all constraints:

$$\text{Maximize/Minimize } v_r \tag{5}$$

subjected to

$$\sum_r S_r \cdot v_r = 0 \tag{6}$$

$$lb_r \leq v_r \leq ub_r \tag{7}$$

where S_r is a stoichiometric matrix of the metabolic network with metabolites and reactions, lb_r is a lower bound of v_r, and ub_r is an upper bound of v_r. By assuming a steady state as shown in equation (6), FVA solves the optimization problem and obtains the minimum and maximum possible flux for each reaction r in the metabolic network.

In this case, we applied FVA approach to reduce a solution space by computing possible range for neighbor reactions (see 3.2 Selection of gene-reaction pairs) and used this value later as a boundary to map gene expression profiles to flux. We specified the minimum biomass production for different conditions as reported in Chubukov *et al* [15]. Since the flux range from FVA is stricter and more realistic than

upper and lower bounds, gene expression constraints are appropriately estimated (see 5. Mapping gene expression profiles onto a metabolic network).

4. Machine learning approach

To systemically develop our approach, we implemented machine learning concept in our method development process. We started by employing our method to train the metabolic network with gene expression data, and then validated the method with another independent dataset. The overview of the entire development process was illustrated in Fig.1.

With machine learning procedure, we were able to find the best parameter setting from the training dataset [30] by comparing flux prediction results from each different parameter setting with ^{13}C metabolic flux data [15]. Then, we applied the selected parameter setting with the validation dataset [29] and evaluate prediction performance of the method.

5. Mapping gene expression profiles onto a metabolic network

Under a mathematical framework of FBA, we developed our approach based on linear programming (LP). With a hypothesis that gene expression data should aid the algorithm to find a correct solution, we linearly mapped gene expression values to predicted fluxes, formulated within an optimization problem:

Let v_{fit_r} represents a normalized gene expression constraint for each investigated reaction r, v_{fit_r} is derived by using information from gene expression data and flux range as shown in equation (8) below.

$$v_{fit_r} = V_{min_r} + \left(\bar{g}_r - g_{min_r}\right)\left[\frac{(V_{max_r} - V_{min_r})}{(g_{max_r} - g_{min_r})}\right] \quad (8)$$

where $\bar{g}$ is a gene expression value from a specific condition, g_{min_r} is a minimum gene expression value and g_{max_r} is a maximum gene expression value across all condition, V_{min_r} is a minimum possible flux and V_{max_r} is a maximum possible flux from ^{13}C metabolic flux (core reaction) or FVA calculation (neighbor reaction).

In the framework of FBA, v_r represents flux for reaction r in the metabolic network. Our optimization problem was built on the hypothesis that gene expression should reflect metabolic flux in reality. As shown in equation (9), we tried to minimize the distance between v_r and v_{fit_r}. Moreover, weight w_r was computed to adjust the optimization problem through equation (10). We obtained V_{weight_r} by selecting the highest magnitude of absolute values of V_{min_r} and V_{max_r} in each reaction r. We assumed that the optimization should favor the narrow flux range reaction because this reaction should reflect the real situation more. This resulted in higher w_r for the narrow flux range reaction.

$$\text{Minimize } \sum_r w_r \cdot \left|v_r - v_{fit_r}\right| \quad (9)$$

$$w_r = \frac{1}{V_{weight_r} + 1} \quad (10)$$

subjected to

$$\sum_r S_r \cdot v_r = 0 \quad (11)$$

$$lb_r \leq v_r \leq ub_r \quad (12)$$

where S_r is a stoichiometric matrix of the metabolic network with metabolites and reactions, lb_r is a lower bound of v_r, and ub_r is an upper bound of v_r. A biomass constraint was set for each different condition or scenario based on the publications of Chubukov *et al* and Buescher *et al* [15, 29]. In our implementation, we opened the lower bound for all eight carbon source uptake reactions at the same time to allow fluxes to move freely based on demand from given gene expression profiles. The bounds for all eight carbon sources were taken from Chubukov *et al* [15].

Moreover, to prevent a high level of flux in reactions outside our mapped reactions, we constrained these fluxes by varying a coefficient between 0 to 1. The value of 0.01 was acquired from our training procedure. By assuming a steady state as shown in equation (11), our mapping method solves the optimization problem and obtains the predicted fluxes for reactions in the metabolic network.

6. Search space reduction

In general, flux range from FVA is narrower than original bounds. Nevertheless, we often observe that even though the flux range is smaller, but it does not decrease substantially, which in turn a search space is still large. This gives the algorithm flexibility to provide an unrealistic optimal solution; especially, the calculation of gene expression constraint v_{fit_r} (see 5. Mapping gene expression profiles onto a metabolic network). To solve the issue, we developed an algorithm to circumvent this arisen issue. The algorithm is a part of the training scheme to narrow the flux ranges for neighbor reactions. We only implemented this in the training step. A scheme for the algorithm is explained from step a) to g).

a) The algorithm generates a list of search space reduction reactions from the neighbor reactions.

b) The list is ranked based on flux range value. The highest flux range is ranked as the first order.

c) The algorithm selects neighbor reaction r based on the ranking from the list.

d) For selected neighbor reaction r from the ranked list, maximum V_{maxPR_r} and minimum V_{minPR_r} possible flux from FVA approach or the previous run are reduced.

$$V_{maxCR_r} = nr\left(V_{maxPR_r}\right) \quad (13)$$

$$V_{minCR_r} = nr\left(V_{minPR_r}\right) \quad (14)$$

where V_{maxCR_r} is modified maximum possible flux for the current run, V_{minCR_r} is modified minimum possible flux for the current run, and nr is a user-defined reduction value ($0 < nr \leq 1$). In this study, we used 0.5 for each iteration. However, if V_{maxPR_r} and V_{minPR_r} are positive values, only equation (13) is applied to ensure that the search space is reduced properly. Similar to a case of negative V_{maxPR_r} and V_{minPR_r}, only equation (14) is used.

e) V_{maxCR_r} and V_{minCR_r} are applied as V_{max_r} and V_{min_r} in equation (8) to calculate gene expression constraint in the mapping procedure (see 5. Mapping gene expression profiles onto a metabolic network).

f) Total model mapping errors are compared between the previous run and current run. If the total model mapping error from the previous run is greater than the

current run, the algorithm goes back to step d). If not, the algorithm moves to the next reaction in the list by repeating step c).

g) The algorithm stops when it progresses through the entire list.

In the end, we obtained the final V_{maxCR_r} and V_{minCR_r} for the current parameter setting. For each different parameter setting, we compared flux prediction results from the mapping procedure to ^{13}C metabolic flux data to select the best parameter setting. Then, we applied the selected setting to the validation dataset to evaluate prediction performance of the mapping approach as shown in Fig.1.

7. Reducing the number of thermodynamically infeasible loops (RED-TIL)

In constraint-based modeling, we usually neglected the loop law in order to reduce computational complexity. The loop law is similar to Kirchhoff's second law for electrical circuits [43]. It is stated that at a steady state there should be no net flux around a closed cycle in the metabolic network. This resulted in the loop problem still occurs inside the network. The consequence of having such loops, which are thermodynamically infeasible, is that FBA simulation becomes less realistic as compared to experimental data. The problem of a thermodynamically infeasible loop (TIL) can be solved by imposing thermodynamics constraints such as standard-state free energy of reaction into the optimization. However, it is very challenging to acquire information for the whole metabolic network as well as to integrate the information in such a simple way to prevent it from turning into a non-linear problem, which is computationally intensive.

To solve this problem, Schellenberger *et al* [31], introduced a method called loopless-COBRA (ll-COBRA) to remove TILs from the network. The method does not require additional thermodynamics information. It utilizes a direction of flux distribution, which exists in every metabolic network, to generate a simpler mixed-integer linear programming (MILP) problem. Although the problem becomes less complex, it still takes time. Here, we adapted the same concept as ll-COBRA, but tackled the similar problem in the novel iterative approach to speed up the process. After obtaining flux prediction results from the mapping procedure (see 5. Mapping gene expression profiles onto a metabolic network), the results were used as an input for a MILP problem to identify TILs and exclude them.

After removing external reactions and applying a flux value threshold (default = 0.01) on the flux prediction results, we received $supp(v)$ as the support of v, which contains a subset of the reactions (internal reactions), where v has the flux value greater than or equal to 0.01 ($v \geq 0.01$). Then, we generated an optimization problem to determine the length of a minimum-containing TIL in the solution as shown in equation (11):

$$\text{Minimize } \sum_r \lambda_r \tag{15}$$

subject to

$$\sum_r S_r \cdot \lambda_r = 0 \tag{16}$$

$$\lambda_r \geq inFC_r \tag{17}$$

$$\sum_r inFC_r \geq 2 \tag{18}$$

$$inFC_r \in \{0, 1\} \tag{19}$$

where λ_r is a flux of reaction r ($\forall r \in supp(v)$), S_r is a stoichiometric matrix of the metabolic network with metabolites and reactions, $inFC_r$ is a binary variable which equals to 1 for a reaction which involves in TIL. In the system that contains TIL, there must be at least two reactions involved as shown in equation (18). If equation (15) to (19) are satisfied, it means that TIL is detected in the system, and the algorithm proceeds further to remove TIL. If not, the MILP problem is infeasible, and the algorithm stops.

As mentioned above, it is required at least two reactions taken part in TIL. Corresponding variables $inFC_{r_1}, inFC_{r_2}, \dots, inFC_{r_k}$ are all equal to 1 and the total number of these variables is greater than or equal to 2 ($k \geq 2$). It is possible to exclude TIL by enforcing an inequality as shown in equation (20) in the next optimization when the mapping procedure is re-optimized in equation (8) to (12) (see 5. Mapping gene expression profiles onto a metabolic network).

$$\sum_{i=1}^{k} inFC_{r_i} \leq k - 1 \qquad (20)$$

Equation (20) forces the algorithm to search for the solution that puts at least one of these variables $inFC_{r_1}, inFC_{r_2}, \dots, inFC_{r_k}$ to 0. It reduces the number of involved reactions which leads to TIL being discarded from the solution. The process of TIL detection and solution re-optimization are performed iteratively until TIL is undetectable in the system. Thus, we received the loopless flux prediction results based on the threshold we set.

Results

Search space reduction improving flux prediction

To ensure that we considered only flux distribution within thermodynamically feasible subspace from our mapping approach, we started by implementing ll-COBRA method together with our method to train the metabolic network of *B. subtilis* [31]. For neighbor reactions that had no flux ranges from ^{13}C metabolic flux data [15], we used flux ranges from FVA to compute gene expression constraints (see 3.4. Flux variability analysis (FVA) and 5. Mapping gene expression profiles onto a metabolic network in Materials and Methods). However, we observed that in many reactions, flux ranges from FVA did not substantially differ from original lower and upper bounds. These high flux ranges did not reflect the real situation and should have been reduced. As a result of this, we developed a new iterative algorithm to modify the flux ranges from FVA to solve this problem and find the best parameter setting for the training step (see 6. Search space reduction in Materials and Methods). The full flux prediction results from before and after search space reduction are provided in Table.S2 and Table.S3 in Supplementary information.

To compare the flux prediction results from before and after applying search space reduction, we computed Pearson's correlation coefficient (PCC) values using flux prediction results and available ^{13}C metabolic flux data [15]. We observed that after applying the search space reduction algorithm, the flux prediction results were improved. An average PCC value was increased from r = 0.58 (before) to r = 0.63 (after). The lists of PCC values from before and after applying search space reduction are listed in Table.S4 and Table.S5 in Supplementary information.

Reducing thermodynamically infeasible loops (RED-TIL) employing an iterative novel method

Although ll-COBRA is a well-known method to efficiently remove TILs in constraints-based modeling, the main drawback is computation speed. The method has to generate one big MILP problem and searches for the loopless optimal solution within the thermodynamically feasible (loopless) region [31]. Because of this, we propose a novel TIL removing approach to solve the same problem but requires less runtime. Under a different strategy from ll-COBRA, RED-TIL directly solves an FBA problem first, and then use this optimal solution as an input to identify TIL and remove it from the system when the FBA problem is re-optimized. The process is done iteratively until no TIL detected in the solution under a certain threshold (see 7. Reducing the number of thermodynamically infeasible loops (RED-TIL) in Materials and Methods). We implemented both methods under the same R programming environment using the same numerical solver to compare these approaches. Because of different algorithms, it is possible that the outputs from RED-TIL and ll-COBRA are not completely identical, but they are still comparable. Majority of the predicted fluxes from mapped reactions (98 reactions) from both methods in glucose and malate conditions, which are the most preferred carbon sources for *B. subtilis*, were generally similar (Fig.2). The dots formed a straight line in both cases. We also found the same trend in other six conditions in Fig.S1 in Supplementary information. The full flux prediction results from both approaches are also provided in Table.S3 and Table.S6 in Supplementary information.

As mentioned above, speed is the highlight feature of RED-TIL. We compared runtime and runtime ratio between RED-TIL and ll-COBRA (Table.3). In general, the average runtime of RED-TIL is three times faster than ll-COBRA. To be noted, speed can

differ under different conditions, which is due to complexity of the problem. As shown in Table.3, the runtime of RED-TIL varies from two to six times faster than ll-COBRA.

Gene expression profile based flux balance model enables identifying the carbon source for *B. subtilis*

To predict the uptake of the major carbon sources, we restricted our model to central energy metabolism (glycolysis, tricarboxylic acid cycle (TCA cycle), pentose phosphate pathway (PPP), urea cycle) as these biochemical pathways are mainly involved in catabolizing the potential carbon sources. The metabolic network of *B. subtilis* was trained with gene expression profiles from eight conditions (glucose, fructose, gluconate, glutamate/succinate, glycerol, malate, malate/glucose, pyruvate) [(30)]. As we aimed to correctly predict the carbon sources based on the given gene expression profiles, we allowed fluxes from all eight carbon source transporter reactions to move freely at once (see 5. Mapping gene expression profiles onto a metabolic network in Materials and Methods). The amount of flux for each carbon source was adjusted automatically based on relevant gene expression level (Fig.3). To assess our flux prediction results, we assembled a confusion matrix of the flux prediction results showing the predicted transporter with the highest flux (most used carbon sources) *versus* the experimental conditions (carbon sources). In summary, six out of six single carbon source conditions were predicted correctly. For two carbon source conditions, two out of two conditions were also predicted correctly (Table.2). Although the glutamate transporter showed the first highest peak in several conditions, the succinate transporter showed only one first highest peak in glutamate/succinate condition. The flux full prediction results are provided in Table.S6 in Supplementary information.

Gene expression profiles illustrate flux behavior inside the metabolic network

We compared our flux prediction results to the fluxes derived by ^{13}C labeling experiments from Chubokov *et al* [15]. This comparison was performed for all 40 reactions for which ^{13}C metabolic flux data was available. For each of these reactions, PCC value was calculated across all eight investigated carbon sources. The results are illustrated in Fig.4 and the exact values of PCC values are listed in Table.S7 in Supplementary information.

Moreover, we found distinct flux behavior from these eight different growing conditions. For example, *B. subtilis* possess two glyceraldehyde-3-phosphate dehydrogenases operating in opposite functions, glycolytic NAD-dependent GapA and gluconeogenic NADP-dependent GapB enzymes [44]. When we compared glucose condition with malate condition, the two main carbon sources for *B. subtilis* [45, 46], we saw that when *B. subtilis* was solely fed with glucose, the flux from glucose traveled downward from glycolysis pathway to TCA cycle. GapA enzyme was activated as we saw non-zero flux in this condition. By contrast, when *B. subtilis* was fed with malate, the flux from malate uptake moved upward from TCA cycle; it activated GapB enzyme and entered glycolysis and PPP pathway. The maps of traveling fluxes from our flux predicted results in both conditions were illustrated in Fig.5. The flux full prediction results are provided in Table.S6 in Supplementary information.

Gene expression profile based flux balance model detects changes in carbon source uptakes

By far, we showed how our approach utilized gene expression data to predict carbon sources for *B. subtilis* and how gene expression data displayed the regulation of flux inside the central energy metabolism. In these eight different conditions, *B. subtilis* were grown in one or two substrates all the time during the study. However, in a natural environment, there is a chance that the bacteria must switch from one carbon source to the other carbon sources, for example, from glucose to malate or the other way around. It is interesting to find out what happens after a shift of nutrients. Particularly, glucose and malate are preferred carbon sources for *B. subtilis* [45, 46]. We applied our approach to another time-series validation dataset [29] with the same parameter setting from the training data. This independent dataset consisted of two nutrient shift scenarios (glucose to glucose plus malate, malate to glucose plus malate). In these scenarios, *B. subtilis* were grown on a single substrate, and the other substrate was added later to see the shift in gene expression and metabolic fluxes. For both experiments, we processed data from the same eight time points (before addition of the other nutrient, 5 min, 10 min, 15 min, 25min, 45 min, 60 min, and 90 min after the addition). When we compared our predicted results with ^{13}C metabolic flux data provided within this work, average uptakes from malate and glucose came up as first and second highest peaks, i.e. our approach correctly predicted the most prominent carbon sources (Table.3). The flux full prediction results are provided in Table.S8 and Table.S9 in Supplementary information.

In the time-series data by Buescher *et al*., besides predicting the correct carbon sources, we also focused our extended analysis on glucose and malate transporters in two scenarios (glucose to glucose plus malate, malate to glucose plus malate) as referred from Buescher *et al*. We compared and then correlated the flux prediction results with ^{13}C

metabolic flux data from eight different time points for each scenario by calculating PCC values (Fig.6).

As expected from transporter reactions, we obtained good positive correlations from glucose (r = 0.57) and malate (r = 0.84) transporters in malate to glucose plus malate scenario and malate (r = 0.83) transporter in glucose to glucose plus malate scenario. For glucose transporter in glucose to glucose plus malate scenario, we received a strong negative correlation (r = -0.74). This contradicted our assumption, which led us to arrive at another assumption that there must be other mechanisms involving in this event. In this case, it turned out to be post-transcriptional regulation (see Discussion). This is beyond the capability of our approach since our goal is to develop a simple method to be as informative as possible to explore the metabolism, which is why we integrate only gene expression data. Despite the limitation of the approach, our method still detected the shift of nutrients between glucose and malate in malate to glucose plus malate scenario. This event was mainly controlled by transcriptional regulation (see Discussion). While the flux from malate decreased over time, the flux from glucose increased. It means that the approach works well if the task is within its capability. So, this should be kept in mind that the results from the approach can become cryptic if different control mechanisms mediate adaptation in these scenarios.

Discussion

In this work, we established the novel approach to integrate gene expression profiles onto the metabolic network. With our approach, we demonstrated how gene expression data contributed to the metabolism of *B. subtilis* and how to utilize this data to gain insights

about the metabolism; especially, the central energy metabolism which is our focus in this study. It is significant to point out that our findings in this study should not be misinterpreted as suggesting that transcriptional regulation is the only mechanism involving in flux regulation. Instead, we agree that there are other mechanisms which play important roles in controlling the metabolism, for example, post-transcriptional mechanism, substrate change, or allosteric enzyme regulation. Nevertheless, gene expression data is beneficial since transcriptional regulation still involves in the regulation of metabolism. As shown in our study, we succeeded in identifying correct carbon sources for *B. subtilis* in each different condition according to gene expression profiles. Moreover, for many reactions in substrate uptakes, glycolysis, and TCA cycle, the flux prediction results correlated well with ^{13}C metabolic flux data ($r > 0.60$). Since these reactions are the major contribution of the central energy metabolism, the average PCC for all 40 reactions across eight conditions is 0.65. Our study supports the idea that combining gene expression profiles with the right approach can be an alternative tool to study metabolism; particularly when ^{13}C metabolic flux data is difficult to acquire.

Besides being the alternative method to investigate metabolism, our approach is generalizable, flexible and versatile. It is unbound to any specific organism if the interested organism has a metabolic network and gene expression profiles are generated for the condition under investigation. With our approach, we can open the lower bound for external uptake reactions at the same time to allow fluxes to move freely based on demand from given gene expression profiles. It is different from conventional practice for FBA simulation. Typically, it is mandatory to restrict the lower bound for non-related uptake reactions in order to correctly predict metabolic fluxes. However, by integrating gene

expression data onto the metabolic network, we can loosen this restriction and let the data decide on which profile belongs to which transporter. As shown in this study, we still obtained good flux prediction results. Also, it is possible to apply to other environments. In more complex situations where ^{13}C labeling experiments are challenging to achieve, it can be a great tool to provide *in silico* simulation. For example, *S. aureus* in the bone cell or *Plasmodium falciparum* (*P. falciparum*) in the erythrocyte, the method can be applied to perform an inhibition study on these virulent pathogens to discover interesting drug targets. This can be used in the preliminary step to generate a list of candidates, which later can be used as a reference to perform experimental validations. Moreover, since our approach adjusts flux level based on gene expression level, it generates flux predictions in continuous values and requires no minimum gene expression level threshold to set genes as “expressed” or “non-expressed” as in Integrative Metabolic Analysis Tool (iMAT) [(9)] or Boolean logic setting like Gene Inactivity Moderated by Metabolism and Expression (GIMME) [(10)].

However, like every other method, it has also limitations. Since our method based on data from gene expression profiles, the flux prediction results can be ambiguous in case of other flux control mechanisms. For example, in the time-series dataset from Buescher *et al*, although we predicted the average carbon sources correctly, we observed abnormality in glucose to glucose plus malate scenario. The flux prediction result from malate transporter in this scenario correlated well with ^{13}C metabolic flux data. However, for glucose transporter from the same scenario, we observed a negative correlation between our result and ^{13}C metabolic flux data. It is unusual since transporter reactions are regulated by transcriptional regulation. For glucose transporter, it is regulated by PTS

operon (ptsG, ptsI, ptsH). Although ptsG is more specific to glucose as compared to the other two genes, which are more general and non-sugar specific [21, 22], it should not be a factor in this case. We expected that our glucose transporter prediction from glucose to glucose plus malate scenario should have a positive correlation as compared to ^{13}C metabolic flux data. However, the result was different from our expectation. Together with the fact that this case was special since there was a shift of carbon sources from glucose to malate. This led us to another assumption that other mechanisms may regulate this scenario apart from transcriptional regulation. In the publication of Buescher *et al*, they performed multi-omics analysis using data from promoter activity, mRNA abundance, and protein abundance time profiles to identify post-transcriptional events [29]. After correlating gene expression level with protein level, they could identify high positive correlations in gene-protein pairs related to glycolysis such as phosphoglycerate mutase (r = 0.96), PTS glucose transporter (r = 0.88) and glyceraldehyde 3-phosphate dehydrogenase (r = 0.96) in malate to glucose plus malate scenario. However, they could not find correlations in gene-protein pairs related to glycolysis in glucose to glucose plus malate scenario. The absence of correlation between transcription level and protein level led to their conclusion that the translation initiation sequences play a role in this missing link. They concluded that glucose to glucose plus malate scenario was dominantly controlled by post-transcriptional mechanisms. On the other hand, malate to glucose plus malate scenario was regulated on a transcriptional level. This sets a good example of the limitation of the method. The method is possible to be used in this shifting scenario, but the prediction result can become ambiguous if the event is mainly controlled by other mechanisms. However, the prediction result still can be used as a sign to generate hypotheses to

investigate further, for example, at the protein level. This may lead to discovery of new insights that are different from what we first expect.

Apart from the mapping method and its application, we also present search space reduction as a novel approach to reduce flux ranges to limit the solution space. After implementing it with ll-COBRA, we showed that flux prediction results were improved, and we used it as a part of our training scheme to train the metabolic network with modified flux ranges. Furthermore, to improve the speed of ll-COBRA, we introduce RED-TIL as an alternative method. While ll-COBRA only searches for an optimal solution in predefined-thermodynamic feasible region, RED-TIL solves the same problem in the whole region, and later detects and forbids TIL, and re-optimizes the solution. As shown in our results, although flux prediction results from both methods were not identical due to different strategies, they were still comparable, and we achieved the same goal. In term of speed, RED-TIL averagely removed TILs three times faster than ll-COBRA. However, under different conditions and complexity of the problems, the speed varied from two to six times faster than ll-COBRA.

Conclusion

In summary, we have introduced a novel mechanistic approach to integrating gene expression profiles into the metabolic network including a new algorithm to reduce flux ranges and remove TILs. Our approach is promising, simple and generalizable. It serves as a great alternative tool to study fluxes inside the central energy metabolism. It can be applied to other organisms to predict carbon sources which later can be interesting targets; especially, pathogenic microorganisms. We believe that this approach can be a

bridge connecting missing links and pave a way to new insights, which eventually changes the way we understand metabolism.

Acknowledgements

We thank our research group members for helpful discussion and general support. We also thank Stefan Schuster for providing us useful suggestions on the project.

Author contributions

KT, FH and MO developed the approach. KT carried out the project. KT and RK wrote the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the project Center for Sepsis Control and Care (CSCC, 01EO1002 and 01EO1502), the Deutsche Forschungsgemeinschaft (https://www.dfg.de/) within the project KO 3678/5-1 and the Deutscher Akademischer Austauschdienst (DAAD). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, Cheng JK, et al. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nat Biotechnol. 2010;28(12):1279-85.
2. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Biotechnol. 2010;28(3):245-8.
3. Sharma AK, Konig R. Metabolic network modeling approaches for investigating the "hungry cancer". Semin Cancer Biol. 2013;23(4):227-34.
4. Bideaux C, Montheard J, Cameleyre X, Molina-Jouve C, Alfenore S. Metabolic flux analysis model for optimizing xylose conversion into ethanol by the natural C5-fermenting yeast Candida shehatae. Appl Microbiol Biotechnol. 2016;100(3):1489-99.
5. Chiewchankaset P, Siriwat W, Suksangpanomrung M, Boonseng O, Meechai A, Tanticharoen M, et al. Understanding carbon utilization routes between high and low starch-producing cultivars of cassava through Flux Balance Analysis. Sci Rep. 2019;9(1):2964.
6. Dang L, Liu J, Wang C, Liu H, Wen J. Enhancement of rapamycin production by metabolic engineering in Streptomyces hygroscopicus based on genome-scale metabolic model. J Ind Microbiol Biotechnol. 2017;44(2):259-70.
7. Kavscek M, Bhutada G, Madl T, Natter K. Optimization of lipid production with a genome-scale model of Yarrowia lipolytica. BMC Syst Biol. 2015;9:72.
8. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2010;107(41):17845-50.
9. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26(24):3140-2.
10. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4(5):e1000082.
11. Jerby L, Shlomi T, Ruppin E. Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. Mol Syst Biol. 2010;6:401.
12. Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu SH. Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput Biol. 2016;12(12):e1005244.
13. van den Esker MH, Koets AP. Application of Transcriptomics to Enhance Early Diagnostics of Mycobacterial Infections, with an Emphasis on Mycobacterium avium ssp. paratuberculosis. Vet Sci. 2019;6(3).
14. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.
15. Chubukov V, Uhr M, Le Chat L, Kleijn RJ, Jules M, Link H, et al. Transcriptional regulation is insufficient to explain substrate-induced flux changes in Bacillus subtilis. Mol Syst Biol. 2013;9:709.
16. Metallo CM, Vander Heiden MG. Understanding metabolic regulation and its influence on cell physiology. Mol Cell. 2013;49(3):388-98.
17. Sung J, Kim S, Cabatbat JJT, Jang S, Jin YS, Jung GY, et al. Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nat Commun. 2017;8:15393.

18. Shan M, Dai D, Vudem A, Varner JD, Stroock AD. Multi-scale computational study of the Warburg effect, reverse Warburg effect and glutamine addiction in solid tumors. PLoS Comput Biol. 2018;14(12):e1006584.
19. Gatto F, Miess H, Schulze A, Nielsen J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep. 2015;5:10738.
20. Chenard T, Guenard F, Vohl MC, Carpentier A, Tchernof A, Najmanovich RJ. Remodeling adipose tissue through in silico modulation of fat storage for the prevention of type 2 diabetes. BMC Syst Biol. 2017;11(1):60.
21. McNeil JC, Vallejo JG, Kok EY, Sommer LM, Hulten KG, Kaplan SL. Clinical and Microbiologic Variables Predictive of Orthopedic Complications Following S. aureus Acute Hematogenous Osteoarticular Infections in Children. Clin Infect Dis. 2019.
22. Schmitt SK. Osteomyelitis. Infect Dis Clin North Am. 2017;31(2):325-38.
23. Zimmerli W, Sendi P. Orthopaedic biofilm infections. APMIS. 2017;125(4):353-64.
24. Geraci J, Neubauer S, Pollath C, Hansen U, Rizzo F, Krafft C, et al. The Staphylococcus aureus extracellular matrix protein (Emp) has a fibrous structure and binds to different extracellular matrices. Sci Rep. 2017;7(1):13665.
25. Junka A, Szymczyk P, Ziolkowski G, Karuga-Kuzniewska E, Smutnicka D, Bil-Lula I, et al. Bad to the Bone: On In Vitro and Ex Vivo Microbial Biofilm Ability to Directly Destroy Colonized Bone Surfaces without Participation of Host Immunity or Osteoclastogenesis. PLoS One. 2017;12(1):e0169565.
26. Bouras D, Doudoulakakis A, Tsolia M, Vaki I, Giormezis N, Petropoulou N, et al. Staphylococcus aureus osteoarticular infections in children: an 8-year review of molecular microbiology, antibiotic resistance and clinical characteristics. J Med Microbiol. 2018;67(12):1753-60.
27. McNeil JC, Kaplan SL, Vallejo JG. The Influence of the Route of Antibiotic Administration, Methicillin Susceptibility, Vancomycin Duration and Serum Trough Concentration on Outcomes of Pediatric Staphylococcus aureus Bacteremic Osteoarticular Infection. Pediatr Infect Dis J. 2017;36(6):572-7.
28. Kavanaugh JS, Flack CE, Lister J, Ricker EB, Ibberson CB, Jenul C, et al. Identification of Extracellular DNA-Binding Proteins in the Biofilm Matrix. MBio. 2019;10(3).
29. Buescher JM, Liebermeister W, Jules M, Uhr M, Muntel J, Botella E, et al. Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism. Science. 2012;335(6072):1099-103.
30. Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012;335(6072):1103-6.
31. Schellenberger J, Lewis NE, Palsson BO. Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophys J. 2011;100(3):544-53.
32. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46(5):2699.
33. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-D61.
34. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.

35. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62.
36. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. Bioinformatics. 2009;25(4):555-6.
37. van den Esker MH, Kovacs AT, Kuipers OP. YsbA and LytST are essential for pyruvate utilization in Bacillus subtilis. Environ Microbiol. 2017;19(1):83-94.
38. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J Biol Chem. 2007;282(39):28791-9.
39. King ZA, Lu J, Drager A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44(D1):D515-22.
40. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.
41. Gudmundsson S, Thiele I. Computationally efficient flux variability analysis. BMC Bioinformatics. 2010;11:489.
42. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5(4):264-76.
43. Price ND, Famili I, Beard DA, Palsson BO. Extreme pathways and Kirchhoff's second law. Biophys J. 2002;83(5):2879-82.
44. Fillinger S, Boschi-Muller S, Azza S, Dervyn E, Branlant G, Aymerich S. Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. J Biol Chem. 2000;275(19):14031-7.
45. Kleijn RJ, Buescher JM, Le Chat L, Jules M, Aymerich S, Sauer U. Metabolic fluxes during strong carbon catabolite repression by malate in Bacillus subtilis. J Biol Chem. 2010;285(3):1587-96.
46. Meyer FM, Stulke J. Malate metabolism in Bacillus subtilis: distinct roles for three classes of malate-oxidizing enzymes. FEMS Microbiol Lett. 2013;339(1):17-22.

Tables

Table.1 Runtime comparison between ll-COBRA and RED-TIL

Condition	ll-COBRA‡	RED-TIL‡	Ratio (ll-COBRA/RED-TIL)
Glucose	31.041	9.876	3.143
Fructose	29.7	8.736	3.4
Gluconate	29.447	12.392	2.376
Glutamate/Succinate	38.736	14.628	2.648
Glycerol	30.571	8.538	3.581
Malate	36.664	14.713	2.492
Malate/Glucose	28.163	4.783	5.888
Pyruvate	30.928	12.177	2.54
Average	31.906	10.73	2.974

‡ Runtime in seconds on a workstation with Intel® Xeon CPU E5-2630 v4 and 64 GB Ram

Table.2 Prediction of the carbon source

Condition	Transporter							
	Glucose	Fructose	Gluconate	Glutamate	Succinate	Glycerol	Malate	Pyruvate
Glucose	**1**	2	2	2				
Fructose	2	**1**		**1**				
Gluconate			**1**	**1**				
Glutamate/Succinate		3		**1**	**1**	2		
Glycerol						**1**		
Malate				**1**		3	**1**	
Malate/Glucose	**3**						**2**	
Pyruvate				3				**1**

Table.3 Prediction of the carbon source for the time series of the nutritional shift

Condition	Transporter							
	Glucose	Fructose	Gluconate	Glutamate	Succinate	Glycerol	Malate	Pyruvate
Glucose/Malate to glucose plus malate	**2**	3					**1**	

Figures

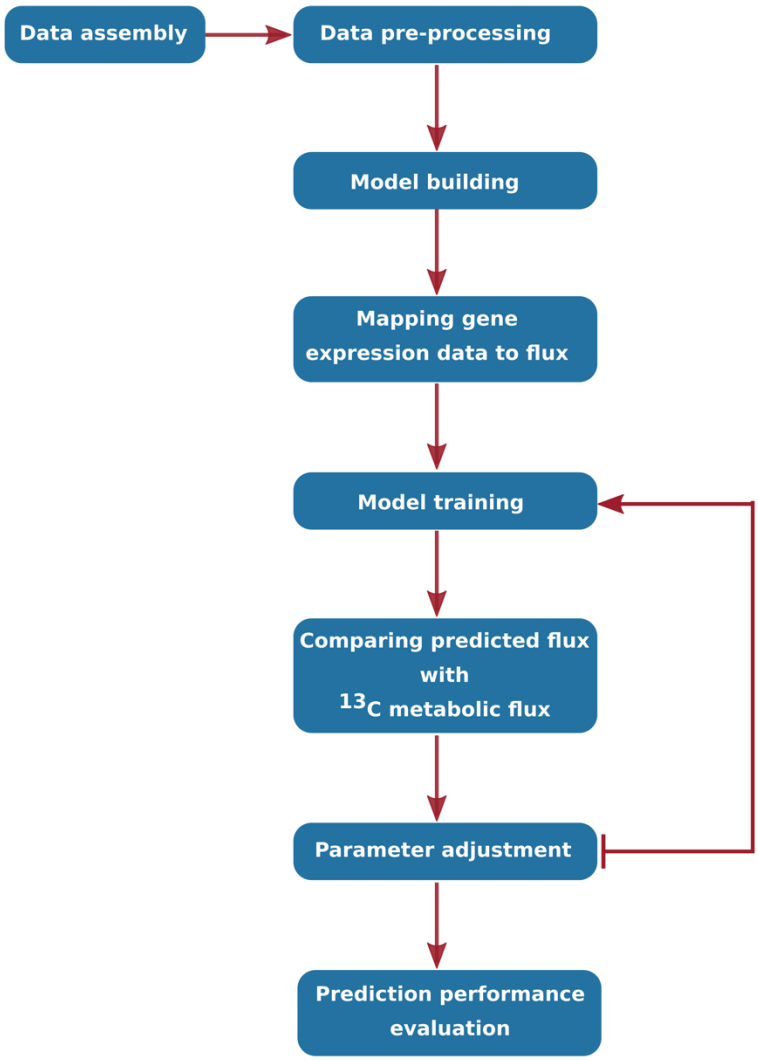
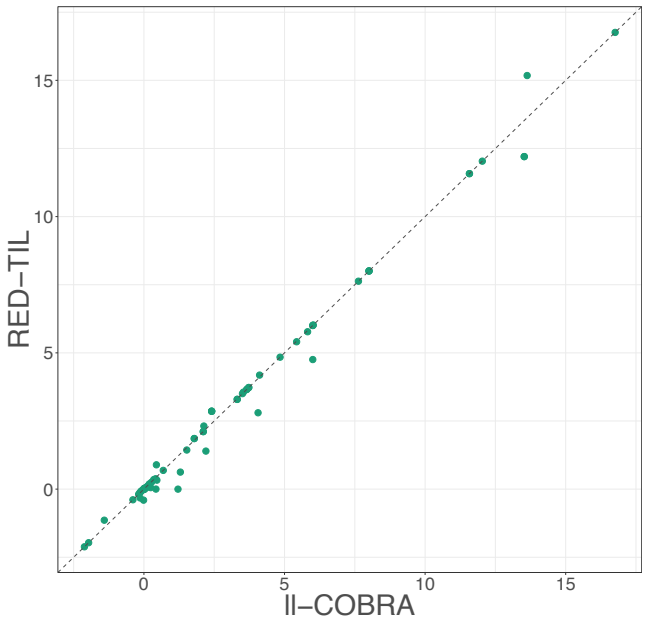


Fig.1 Overview of our method development. Arrows show a flow of the entire procedure. Machine learning approach was applied to train the metabolic network with gene expression data and find the best parameter setting.

A



B

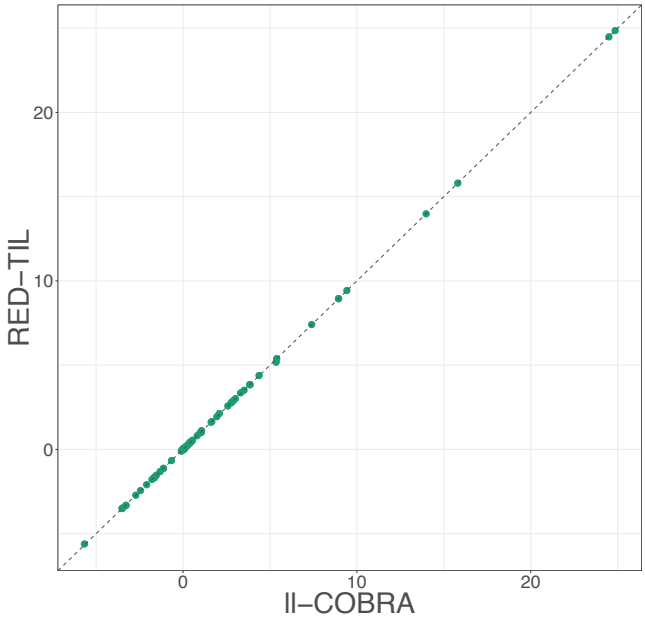


Fig.2 (A and **B)** Scatterplots of predicted fluxes (mmol h^{-1} $gcdw^{-1}$) from mapped reactions (98 reactions) between ll-COBRA and RED-TIL from glucose (A) and malate (B) conditions.

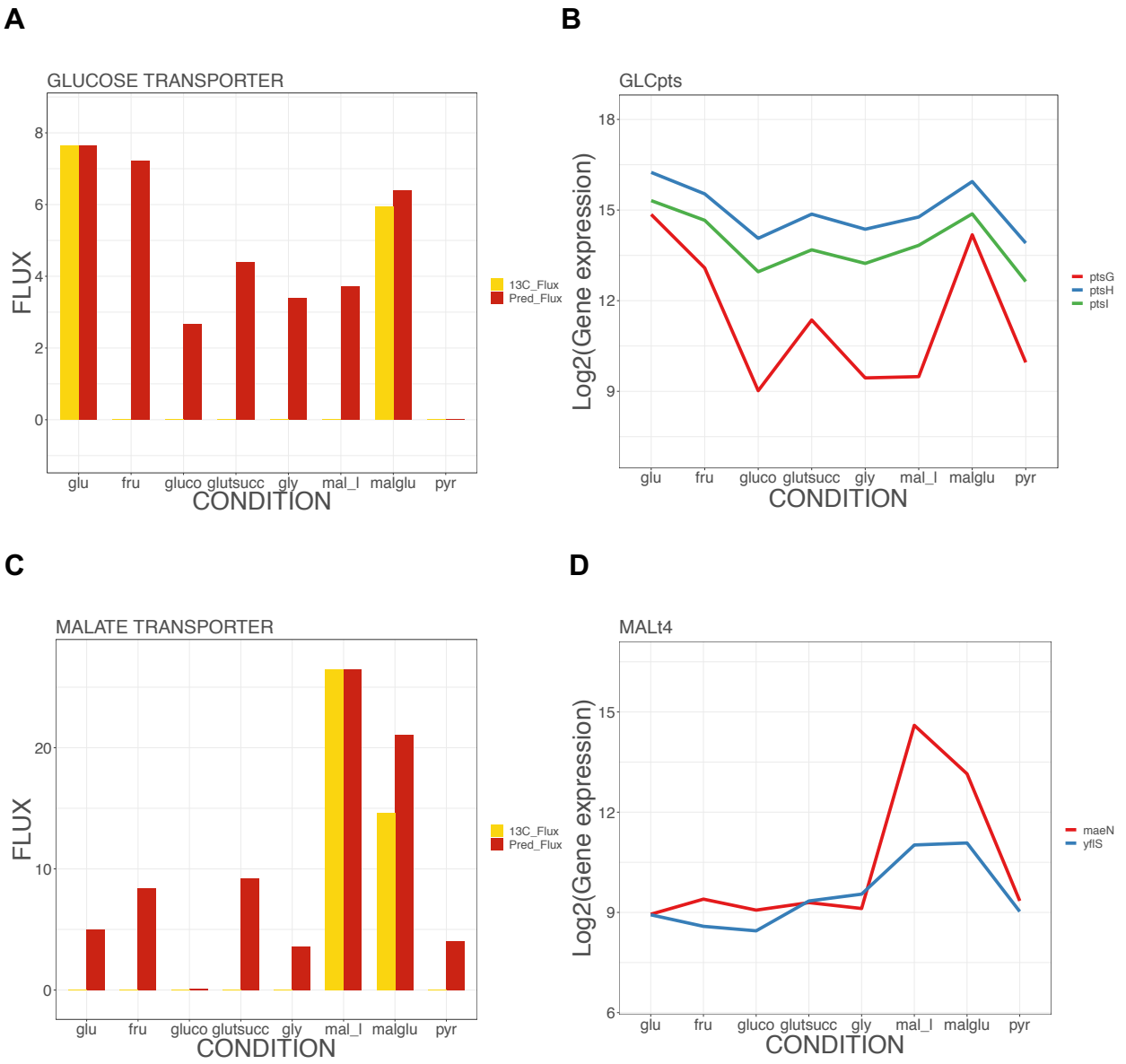


Fig.3 (A-D) Adjustment of carbon source uptakes by level of gene expression in glucose transporter (A and B) and malate transporter (C and D). Under eight different conditions, (A and C) present flux level (mmol h^{-1} $gcdw^{-1}$) from our prediction result (red) and ^{13}C metabolic flux data (yellow) while (B and D) show gene expression level in log2 scale from relevant genes. Flux is adjusted according to gene expression level in both transporters. The highest peak from each transporter identifies the carbon source.

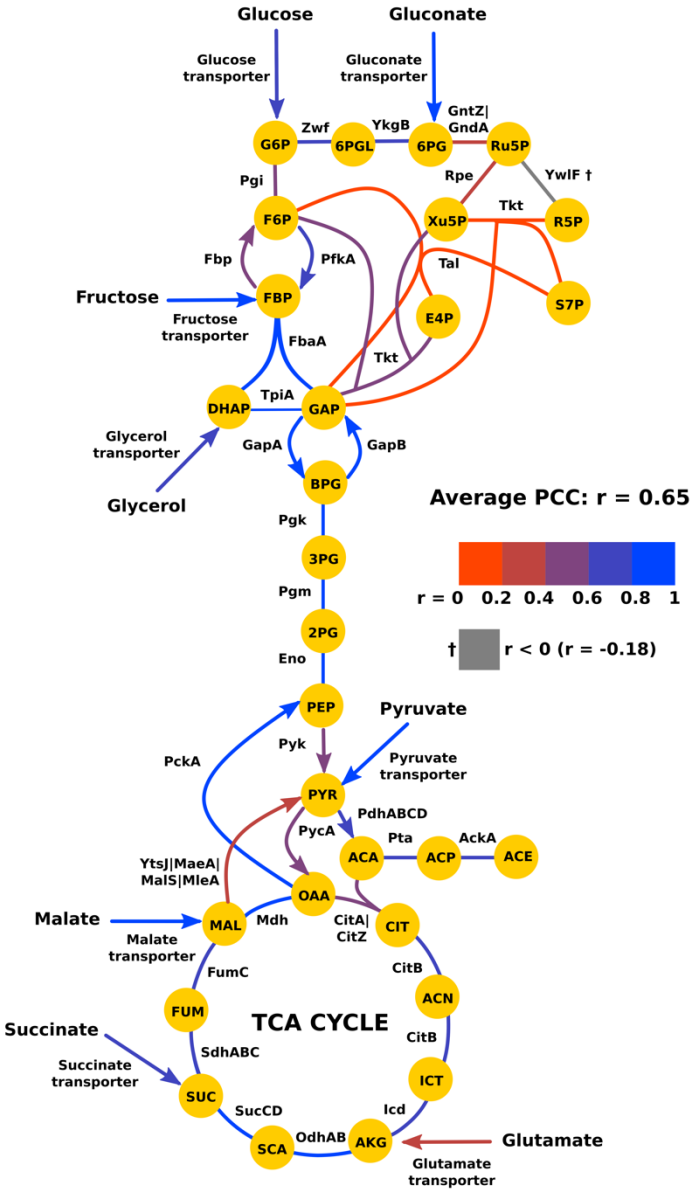


Fig.4 Color coded representation of the correlation between the flux prediction results and the fluxes derived by ^{13}C labeling experiments of Chubokov *et al* [15]. The value, r =1, represents the highest positive correlation between the flux prediction results and ^{13}C metabolic flux data. By contrast, r = 0, it interprets as no correlation.

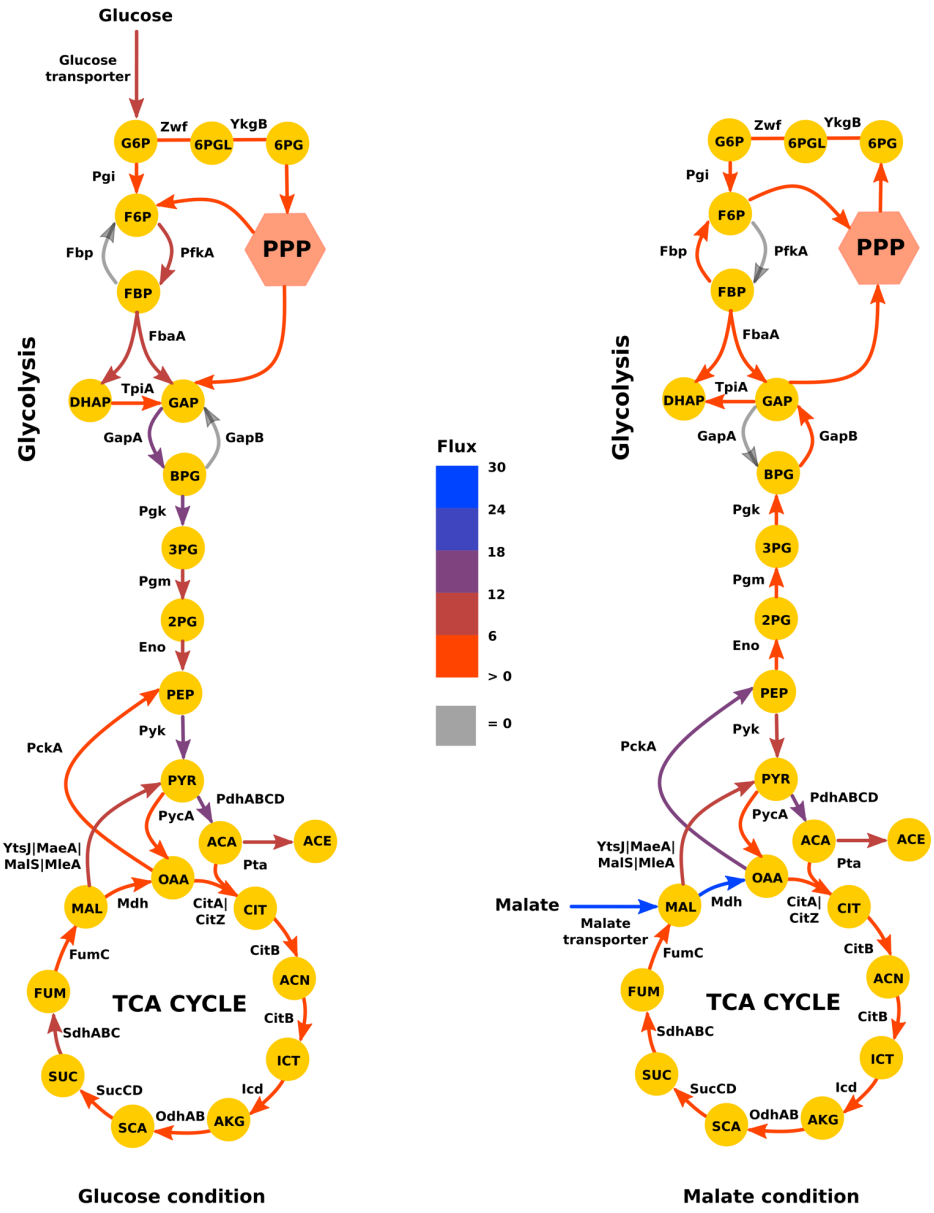
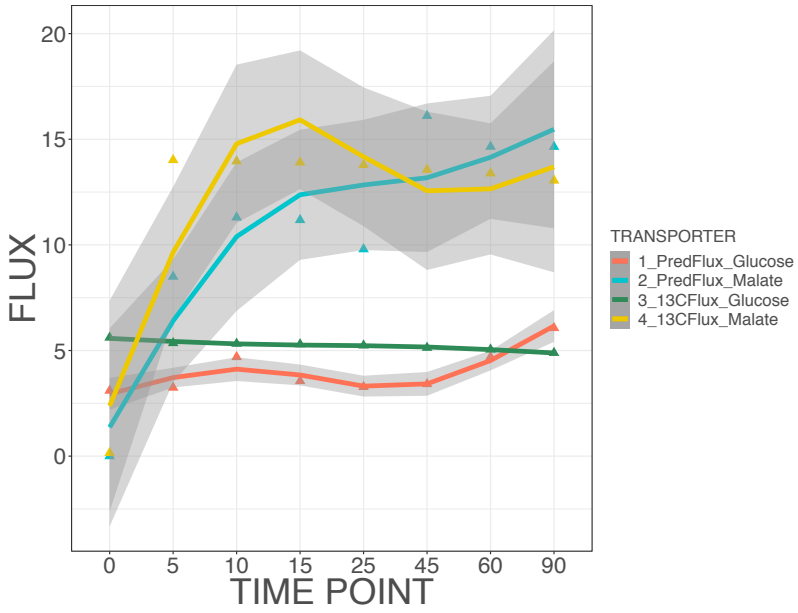


Fig.5 Graphical presentation of influence of different carbon sources on flux regulation from the flux prediction results. Level of flux is shown in mmol h^{-1} $gcdw^{-1}$. For GapA enzyme (GAPD), there is only flux from glucose condition (Left), while malate condition (Right) shows no flux at all. Nonetheless, for GapB enzyme, it solely shows flux from malate.

A



B

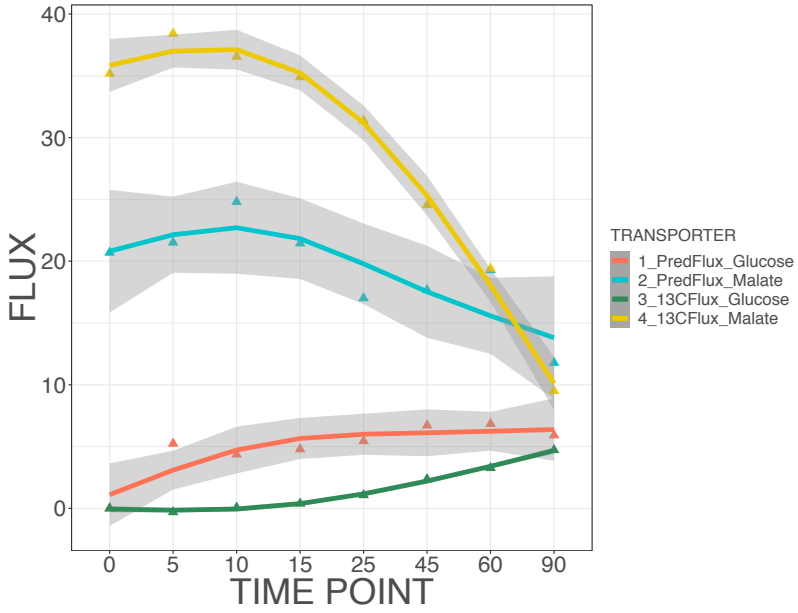


Fig.6 (A and **B)** Carbon source shifts between glucose and malate. For each plot, level of flux (mmol h^{-1} $gcdw^{-1}$) and timepoint (min) are given. While PredFlux_ represents the flux prediction results from glucose transporter (PredFlux_Glucose) and malate transporter (PredFlux_Malate), 13CFlux_ defines ^{13}C metabolic flux data from glucose transporter (13CFlux_Glucose) and malate transporter (13CFlux_Malate). Triangle spots show real

values while lines are the results from real values by fitting of splines. Shadings represent 95% confidence intervals. After adding the second substrate in glucose to glucose plus malate (A) or malate to glucose plus malate (B), the uptakes from glucose and malate transporter start to change which are the sign of carbon source shifts.