Bioinformatics Advance Access published May 25, 2009

Systems Biology

Genetic Modification of Flux (GMF) for Flux Prediction of Mutants

Quanyu Zhao[1] and Hiroyuki Kurata[1,*]

[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Gene deletion and over-expression are critical technologies for designing or improving the metabolic flux distribution of microbes. Some algorithms including flux balance analysis (FBA) and minimization of metabolic adjustment (MOMA) predict a flux distribution from a stoichiometric matrix in the mutants in which some metabolic genes are deleted or non-functional, but there are few algorithms that predict how a broad range of genetic modifications, such as over-expression and under-expression of metabolic genes, alters the phenotypes of the mutants at the metabolic flux level.

Results: To overcome such existing limitations, we develop a novel algorithm that predicts the flux distribution of the mutants with a broad range of genetic modification, based on elementary mode analysis. It is denoted as Genetic Modification of Flux (GMF), which couples two algorithms that we have developed: Modified Control Effective Flux (mCEF) and Enzyme Control Flux (ECF). mCEF is proposed based on CEF to estimate the gene expression patterns in genetically modified mutants in terms of specific biological functions. GMF is demonstrated to predict the flux distribution of not only gene deletion mutants but also the mutants with under-expressed and over-expressed genes in *Escherichia coli* and *Corynebacterium glutamicum*. This achieves breakthrough in the a priori flux prediction of a broad range of genetically modified mutants.

Contact: kurata@bio.kyutech.ac.jp

Supplementary information: supplementary file and programs are available at the journal's website or http://www.cadlive.jp

1 INTRODUCTION

A cell is a sophisticated factory with lots of the physiological functions including synthesis, transport, storage, and degradation of biological molecules. Systems biology aims at rationally designing cellular functions at the molecular interaction levels (Joyce and Palsson, 2006). Its goals are to understand the mechanism of how biochemical networks generate particular cellular functions and to rationally design the molecular processes to meet an engineering purpose (Nishio *et al.*, 2008). There have been many studies that aim at producing useful metabolites using genetically engineered organisms, which often require a mathematical strategy of how the molecular processes of complex and robust cells are designed to achieve enhanced production.

Metabolic engineering in microorganisms is among promising methodologies for the synthesis of biochemical compounds, such as biofuels (Atsumi *et al*., 2008; Stephanopoulos, 2007) and amino acids (Park *et al*., 2007). Since more than ten genome-scale metabolic networks in microorganisms have been established and the flux distributions for hundreds of mutants could be determined in large-scale (Fischer and Sauer, 2005), mathematical models are required to integrate such biological information and experimental data, and contribute to the strain improvements (Lee *et al*., 2005). *In silico* modeling is a challenge to accurately predict the cellular physiological behaviors.

Constraint-based flux analysis is used for predicting the steady-state intracellular fluxes from the stoichiometric matrices with specific objective functions by optimization algorithms (Kauffman *et al*., 2003). FBA is based on the structural or topological information of metabolic networks, while definition of some objective functions such as maximum biomass formation or growth rate is critical. These days several algorithms, including Linear Programming (LP), Quadratic Programming (QP) (Segrè *et al*., 2002) and Mixed Integer Linear Programming (MILP) (Shlomi *et al*, 2005), have been applied to estimate a change in the flux distribution in gene knockout mutants. In regulatory-FBA (rFBA) information of gene expression is incorporated by the Boolean logic formalism (Akesson *et al*., 2004; Covert *et al*., 2004) that uses a binary system, where the flux of one reaction is set to be zero if the relative gene is not expressed. The flux distribution of such gene knockout mutants could be optimized by LP under this additional constraint. In Minimization Of Metabolic Adjustment (MOMA), the flux distributions of gene knockout mutants can be estimated by the QP-based minimization of the Euclidian distance from those of wild type to those of a mutant(Segrè *et al*., 2002). Regulatory on/off minimization (ROOM) uses MILP to predict the fluxes of gene deletion mutants in which the number of significant flux changes is minimized compared with wild type (Shlomi *et al*, 2005).

On the other hand, an alternative way by network-based pathway analysis emerges for constructing a mathematical model that accesses the properties and functions for metabolic networks. It has focused on two approaches, elementary modes (EMs) (Schuster *et al*., 1999; Schuster *et al*, 2000) and extreme pathways (Schilling *et*

[*]To whom correspondence should be addressed.

al., 2002). Both employ a convex set of vectors used to characterize all steady-state flux distributions of a biochemical network. Elementary mode or extreme pathway analysis is suggested to provide the correlations between metabolic pathways and transcriptional patterns (Çakır *et al*., 2007; Çakır *et al*., 2004a; Stelling *et al*., 2002).

Control Effective Fluxes (CEFs) had been developed based on EM analysis to predict the transcriptional regulations, e.g., the gene expression patterns of *Escherichia coli* (Stelling *et al*., 2002) and *Saccharomyces cerevisiae* (Çakır *et al*., 2007; Çakır *et al*., 2004a) grown on different substrates. The CEF algorithm was modified to analyze the erythrocyte enzymopathies of human red blood cells (Çakır *et al*., 2004b), but its application was restricted to mutants with partially deficient enzymes, where the available range of enzyme activity was from zero for the gene deletion condition to one for a normal state. This algorithm has not been validated by experimental data yet in a quantitative way.

Enzyme Control Flux (ECF) (Kurata *et al*., 2007b; Zhao and Kurata, 2009) has been proposed based on EM analysis to predict the correlation between the relative enzyme activity profile of a mutant to wild type and its associated flux distribution. ECF estimated the flux distributions of a variety of genetically modified mutants of *E. coli* and *Bacillus subtilis* by using their associated enzyme activity profiles.

Those EM-based methods are critically responsible for linking an EM matrix to transcriptional patterns or for connecting an enzyme activity profile to its associated flux distribution, but few algorithms predict the flux distributions of genetically modified mutants by estimating changes in their transcript or enzyme profiles.

The purposes of the genetic modifications for micro-organisms are to increase in the yields of bio-compounds or to decrease in the productions of by-products. Gene deletion is not the unique methodology for the strain improvement. Over-expression or under-expression of genes are very important technologies to increase the productivities of target compounds (Becker *et al*., 2007; Becker *et al*., 2005; Nicolas *et al*., 2007; Ohnishi *et al*., 2005). However, the existing methods including MOMA and rFBA only predict the flux distribution of the mutants that completely lack the genes coding metabolic enzymes. They are not applicable to the mutants that over-express or partially synthesize specific enzymes. A broad range of genetic modification would present a suitable optimization strategy for the strain improvements.

To overcome this limitation, we propose an EM-based algorithm that couples our original algorithms: modified CEF (mCEF) and ECF, which is denoted as Genetic Modification of Flux (GMF). GMF predicts the flux distributions for not only gene knockout mutants but also the mutants with over-expressed or under-expressed genes using the topological structures of metabolic networks. The feasibility of GMF is demonstrated by applying it to genetic mutants of *E. coli*, and *Corynebacterium glutamicum*.

2 METHODS

2.1 Control-Effective Flux (CEF)

The EM is the minimal set of enzymes that can operate at steady-state with all the irreversible reactions operating properly (Schuster *et al*., 1999). The EM matrix $\mathbf{P} = (p_{ij})$ is determined from the stoichiometric matrix and the flux vector $\mathbf{v} = (v_1, v_2, ..., v_n)^t$ is represented as:

$$\mathbf{v} = \mathbf{P} \cdot \boldsymbol{\lambda}\,, \tag{1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)^t$ is the Elementary Mode Coefficient (EMC) vector. $\mathbf{P}$ is the $n \times m$ matrix in which n is the number of reactions in a metabolic network model and m is number of the EMs. Each column of $\mathbf{P}$ was normalized by the element of substrate uptake in each EM (If the element of substrate uptake is zero under sole carbon source conditions, EMs form internal loops. It is out of this study.).

The original algorithm of CEF was developed to estimate the change in transcriptional regulations based on the topology of metabolic networks with specific biological reactions, when the substrate changes, e.g., from glucose to acetate, ethanol or glycerol (Çakır *et al*., 2007; Çakır *et al*., 2004a; Stelling *et al*., 2002). The efficiency of the *j*-th EM for each cellular objective, $\varepsilon_{j,CELLOBJ}$, is defined as the ratio of EM's output (reaction involving the objectives) to the investment required to form each EM (the sum of the absolute elements in the EM):

$$\varepsilon_{j,CELLOBJ} = \frac{p_{CELLOBJ,j}}{\sum_i |p_{i,j}|} \tag{2}$$

where $p_{i,j}$ is the normalized element of the *i*-th reaction in the *j*-th EM and *CELLOBJ* is the reaction number for the specific biological function (biomass production and ATP generation). CEF for the *i*-th reaction, which is associated to the flux of the *i*-th reaction, is represented by the weighted sum of the *i*-th elements of all EMs using their associated efficiency $\varepsilon_{j,CELLOBJ}$:

$$cef_i = \sum_{CELLOBJ} \frac{1}{p_{CELLOBJ}^{\max}} \frac{\sum_j (\varepsilon_{j,CELLOBJ} \cdot |p_{i,j}|)}{\sum_j \varepsilon_{j,CELLOBJ}} \tag{3}$$

where $p_{CELLOBJ}^{\max}$ is the maximum element in the row of biological functions.

The theoretical transcript ratio for the *i*-th reaction under different substrate conditions, s_1 and s_2, is provided by:

$$\Theta_i(s_1, s_2) = \frac{cef_i(s_2)}{cef_i(s_1)}\,. \tag{4}$$

Details of the CEF algorithm are described elsewhere (Çakır, *et al*., 2004; Stelling, *et al*., 2002).

2.2 Modified algorithm of Control-Effective Flux (mCEF)

To apply CEF to a broad range of genetic mutants that over-express, under-express or lack a metabolic gene, the CEF algorithm is modified. The efficiency of the *j*-th EM for such a genetic mutant is defined by:

$$\varepsilon_{j,CELLOBJ}^{m} = \frac{p_{CELLOBJ,j} \cdot EA_j}{\sum_i \left(|p_{i,j}| \cdot \eta_i \right)}\,, \tag{5}$$

$$\eta_i = \begin{cases} EAP_i & \text{(if reaction } i \text{ is modified)} \\ 1 & \text{(if reaction } i \text{ is not modified)} \end{cases}\,,$$

where EAP_i is the relative gene expression (enzyme activity) responsible for the *i*-th reaction of a mutant to wild type. EAP_i is equivalent to zero if the gene of the *i*-th reaction is deleted. If it is over-expressed or under-expressed, EAP_i is more than 1 or less than 1, respectively. η_i is the correction factor for calculating the investment for genetic mutants. The term $\sum_i\left(|p_{i,j}|\cdot\eta_i\right)$ can be regarded as the EM investment that accounts for the cost required for the modified gene of the *i*-th reaction. When a gene belonging to an EM is over-expressed or under-expressed, the investment term increases or decreased, respectively. EA_j is the correction factor that incorporates the change in the modified reaction into each EM's output, as defined by:

$$EA_j = \prod_{i=1}^{n} ge_{i,j} \quad . \tag{6}$$

$$ge_{i,j} = \begin{cases} EAP_i & (\text{if the } i\text{-th reaction is involved in the } j\text{-th EM}) \\ 1 & (\text{if the } i\text{-th reaction is not involved in the } j\text{-th EM}) \end{cases}$$

where $ge_{i,j}$ is the parameter representing the gene expression state for the *i*-th reaction in the *j*-th EM. When a gene within an EM is over-expressed or under-expressed, its output (the numerator of Equation (5)) increases or decreases, respectively. This multiplication form (Equation (6)) reflects the biological fact that metabolic flux distributions would be determined not by a few rate limiting steps than by overall gene expression profiles (Heinrich and Rapoport, 1974, Small and Kacser, 1993). This idea is also employed by ECF. These factors would be empirically or intuitively derived from biological behaviors rather than based on rigorous physical mechanisms.

For $EAP_i = 0$, the EM containing it is neglected ($\varepsilon^m_{j,CELLOBJ} = 0$), which is consistent with EM analysis of gene deletion mutants. For $EAP_i = 1$, *i.e.*, when gene expressions are not changed at all, Equation (5) is consistent with Equation (2). Equation (5) is an extension of the original efficiency (Equation (2)) to genetic mutants.

The modified CEF (mCEF) for the mutant is defined as:

$$mCEF_i(mut) = \sum_{CELLOBJ} \frac{1}{p^{\max}_{CELLOBJ}} \frac{\sum_j\left(\varepsilon^m_{j,CELLOBJ}\cdot|p_{i,j}|\cdot\eta_i\right)}{\sum_j \varepsilon^m_{j,CELLOBJ}}, \tag{7}$$

where η_i is used to weight its associated element of each EM.

Since mCEF for wild type corresponds to that of the original CEF:

$$mCEF_i(w) = \sum_{CELLOBJ} \frac{1}{p^{\max}_{CELLOBJ}} \frac{\sum_j(\varepsilon_{j,CELLOBJ}\cdot|p_{i,j}|)}{\sum_j \varepsilon_{j,CELLOBJ}}, \tag{8}$$

the CEF ratio for a mutant to wild type, the relative change in a gene expression profile of a mutant to wild type, is provided by:

$$\Theta_i(w,mut) = \frac{mCEF_i(mut)}{mCEF_i(w)} \tag{9}$$

Details of the algorithm are illustrated in the supplementary file (Figure S1, Table S1-S4). In this method, *CELLOBJ* is the reaction number for the biomass formation and ATP generation.

2.3 Enzyme-Control Flux (ECF)

Enzyme-Control Flux (ECF) was developed to estimate the mathematical correlation between an enzyme activity profile and its associated flux distribution, based on the EM matrix $\mathbf{P}$ (Kurata *et al*., 2007b). The EMCs of wild type $\boldsymbol{\lambda}^w = (\lambda_1^w, \lambda_2^w, \ldots \lambda_m^w)^t$ are calculated by quadratic programming (Schwartz and Kanehisa, 2005; Schwartz and Kanehisa, 2006) from the flux distribution of the wild type as follows:

$$\min \sum_j (\lambda_j^w)^2$$

$$\textit{subject to } \mathbf{P}\cdot\boldsymbol{\lambda}^w = \mathbf{v} \tag{10}$$

$$\lambda_j^w \geq 0$$

The EMCs of a mutant are provided by the multiplication form:

$$\lambda_j^{mut} = \gamma\cdot\lambda_j^w \prod_{i=1}^{n} a_{i,j} \tag{11}$$

$$a_{i,j} = \begin{cases} a_i & (\text{if the } i-\text{reaction is involved in the } j\text{ - th EM}) \\ 1 & (\text{if the } i-\text{reaction is not involved in the } j\text{ - th EM}) \end{cases}$$

where $\boldsymbol{\lambda}^{mut} = (\lambda_1^{mut}, \lambda_2^{mut}, \ldots \lambda_m^{mut})^t$, $a_{i,j}$ is the relative enzyme activity of a mutant to wild type for the *i*-th reaction in the *j*-th EM, a_i is the enzyme activity ratio of the mutant to wild type for the *i*-th reaction. $\boldsymbol{\lambda}^{mut}$ is normalized by a factor of γ, so that the substrate uptake flux is the same as that of wild type. The flux distribution of the mutant is provided by:

$$\mathbf{v}^{mut} = \mathbf{P}\cdot\boldsymbol{\lambda}^{mut} \tag{12}$$

Details of the algorithm are described elsewhere (Kurata *et al*., 2007b).

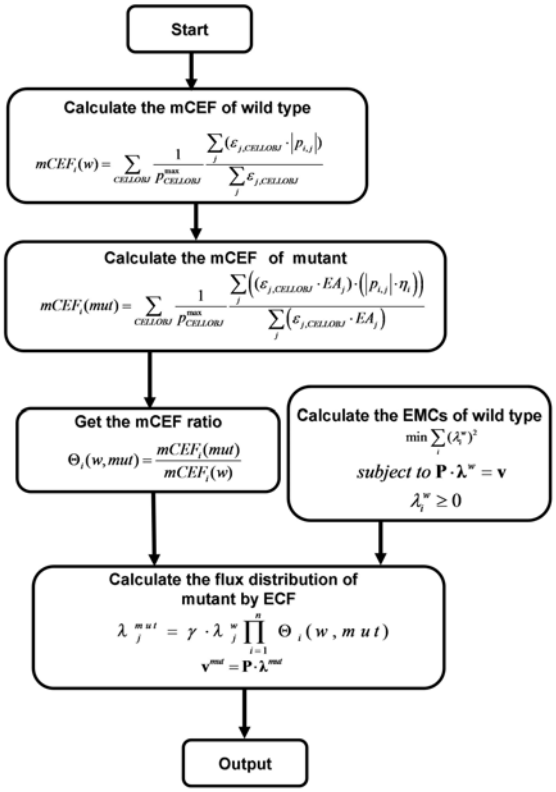
Start
Calculate the mCEF of wild type
$mCEF_i(w) = \sum_{CELLOBJ} \frac{1}{p^{\max}_{CELLOBJ}} \frac{\sum_j(\varepsilon_{j,CELLOBJ}\cdot|p_{i,j}|)}{\sum_j \varepsilon_{j,CELLOBJ}}$
Calculate the mCEF of mutant
$mCEF_i(mut) = \sum_{CELLOBJ} \frac{1}{p^{\max}_{CELLOBJ}} \frac{\sum_j\left((\varepsilon_{j,CELLOBJ}\cdot EA_j)\cdot(|p_{i,j}|\cdot\eta_i)\right)}{\sum_j\left(\varepsilon_{j,CELLOBJ}\cdot EA_j\right)}$
Get the mCEF ratio
$\Theta_i(w,mut) = \frac{mCEF_i(mut)}{mCEF_i(w)}$
Calculate the EMCs of wild type
$\min \sum_i (\lambda_i^w)^2$
$\textit{subject to } \mathbf{P}\cdot\boldsymbol{\lambda}^w = \mathbf{v}$
$\lambda_i^w \geq 0$
Calculate the flux distribution of mutant by ECF
$\lambda_j^{mut} = \gamma\cdot\lambda_j^w \prod_{i=1}^{n} \Theta_i(w,mut)$
$\mathbf{v}^{mut} = \mathbf{P}\cdot\boldsymbol{\lambda}^{mut}$
Output

Fig. 1 A schematic diagram for the GMF Algorithm

2.4 ECF with mCEF (GMF)

GMF is presented to predict the flux distributions of the mutants that overexpress the specific genes encoding metabolic enzymes, or partially or fully lack them. A schematic diagram of the algorithm is depicted as shown in Figure 1. This EM-based algorithm consists of two fundamental algorithms: mCEF and ECF.

The CEF ratios of a mutant to wild type are calculated from the metabolic network topology by the mCEF algorithm presented in this study. Assuming that a gene expression profile is linearly correlated to its associated enzyme activity profile, the EMCs of a mutant are estimated from the flux distribution of wild type by quadratic programming (Equation (10)). In some cases, there is a quantitative correlation between mRNA expression and protein levels (Ideker *et al*., 2001; Siddiquee *et al*., 2004). When the enzyme activity ratios can be replaced by the CEF ratios, the EMCs for the mutant are provided by Equation (11):

$$\lambda_j^{mut} = \gamma \cdot \lambda_j^{w} \prod_{i=1}^{n} \Theta_i(w, mut) \tag{13}$$

Finally, the flux distribution of the mutant is provided by:

$$\mathbf{v}^{mut} = \mathbf{P} \cdot \boldsymbol{\lambda}^{mut} \tag{14}$$

2.5 Test of GMF

The GMF algorithm is tested by using experimental data. The prediction accuracy is evaluated by the prediction error:

$$Prediction\ \ error = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(v_{i,GMF} - v_{i,exp}\right)^2} \tag{15}$$

where, $v_{i,GMF}$ is the GMF-predicted flux for the *i*-th reaction and $v_{i,\exp}$ is the experimental flux for the *i*-th reaction; *n* is the number of the fluxes.

2.6 FBA and MOMA

FBA typically employs linear programming for the prediction of fluxes with the maximum growth rate as the objective function (Kauffman *et al*., 2003):

$$\begin{gathered} \max \mathbf{v}_{growth} \\ subject\ to\ \mathbf{S} \cdot \mathbf{v} = 0 \\ v_{i,\min} \le v_i \le v_{i,\max}\ \left(i = 1, \ldots, n\right) \end{gathered} \tag{16}$$

where $\mathbf{S}$ represents the stoichiometric matrix and $\mathbf{v}$ is the column vector representing the fluxes. In matrix $\mathbf{S}$, there are *n* reactions. The maximum biomass formation is selected as the objective function. $v_{i,\min}$ and $v_{i,\max}$ are the lower and higher boundaries for each v_i. In the irreversible reaction, $v_{i,\min}$ is equal to zero.

In MOMA, the flux distribution of mutants could be estimated by the minimization of the Euclidian distance from that of wild type (Segrè *et al*., 2002).

$$D(\mathbf{v}^w, \mathbf{v}^m) = \sqrt{\sum_{i=1}^{N}(v_i^m - v_i^w)^2} \tag{17}$$

where v_i^w is the flux of the wild type for the *i*-th reaction and v_i^m is the calculated flux of the mutants for the *i*-th reaction. The equation could be converted to the standard form for quadratic programming.

$$\begin{gathered} \min\left(\mathbf{v}^m - \mathbf{v}^w\right)^T\left(\mathbf{v}^m - \mathbf{v}^w\right) \\ subject\ to\ \mathbf{S} \cdot \mathbf{v} = 0 \\ v_{i,\min} \le v_i \le v_{i,\max}\ \left(i = 1, \ldots, n\right) \\ v_d = 0\ \text{ (if the } d\text{-th reaction is deleted).} \end{gathered} \tag{18}$$

2.7 Implementation

All the calculations are performed by Matlab (The Mathworks Inc.). The EMs are calculated by CellNetAnalyzer (Klamt, *et al*., 2007). The application program of GMF is freely available at the journal's web site. Generally, it is hard to find the global optima in nonlinear systems with a huge space of search parameters due to calculation complexity. In this study, quadprog, a function in Matlab, is employed to solve quadratic programming. In medium-scale quadprog algorithm, an active set method is adopted.

2.8 Metabolic network models

The metabolic networks for *E. coli*, *S. cerevisiae* and *C. glutaminum* were reconstructed and analyzed by CADLIVE 2.75 (Kurata *et al*., 2003; Kurata *et al*., 2007a) and CellNetAnalyzer (Klamt *et al*., 2007). The reactions and metabolites of the metabolic network model for *E. coli* and *C. glutamicum* were presented in Supplementary file (Figure S2 and S3, Table S5-S8). The reactions of biomass formation were cited from references for *E. coli* (Wiback *et al*., 2004) and *C. glutamicum* (Gayen *et al*., 2006). Those of *S. cerevisiae* were from Förster's paper (Förster *et al*., 2002). The latest version and manual of CADLIVE could be freely downloaded from our webpage (http://www.cadlive.jp)

3 RESULTS

3.1 GMF algorithm

In GMF, CEF is modified to predict the gene expression patterns for a broad range of the mutants with a deleted, under-expressed or over-expressed gene. mCEF predicts how a specific metabolic gene modification affects the change in the gene expression profile of a mutant to that of wild type. ECF estimates how a change in the enzyme activity profiles between a mutant and wild type alters the flux distribution of wild type. ECF integrates the enzyme activity profile into EMCs in a multiplication form, calculating the flux distributions (Kurata *et al*., 2007b). mCEF is directly connected to ECF to predict the flux distribution, where ECF uses the ratios of gene expressions instead of the associated enzyme activity profile.

To demonstrate the performance and feasibility of GMF, we compare it with other predictive methods such as FBA and MOMA, and show the outstanding performance of GMF.

3.2 Prediction of gene expression patterns for gene deletion and over-expressing mutants by mCEF

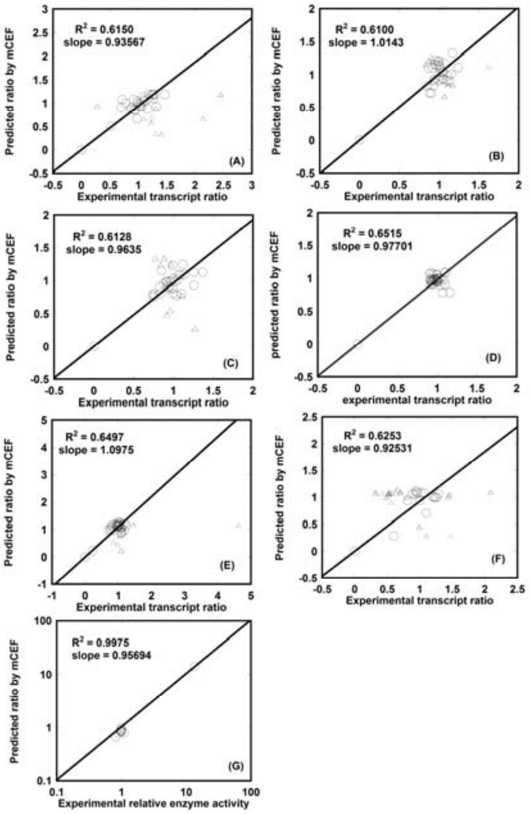

Fig. 2 Comparison of the gene expression ratios of a mutant to wild type predicted by mCEF with the experimental gene expression ratios. (A) *pgi*, (B) *fbp*, (C) *fba*, (D) *pyk*F, (E) *gnd* and (F) *rpe* knockout mutants in *E. coli*. The open triangles were omitted from statistical analysis, which corresponds to the reactions with small flux values. Comparison of the simulated gene expressions with the experimental enzyme activities in a pyruvate decarboxylase (*pdc*) overproduction mutant in *S. cerevisiae* (G). The point whose relative enzyme activity is more than 10 corresponds to the over-expressed gene.

To demonstrate the feasibility of mCEF, the predicted results by mCEF were compared with the experimental data (Ishii *et al*., 2007), as shown in Figure 2 (A-F).

To show a correlation between the predicted results and experimental data, we performed a statistical analysis using a linear regression model, while some points with large prediction errors were omitted in the same manner as the related studies for CEF (Çakır *et al*., 2007; Çakır *et al*., 2004a; Stelling *et al*., 2002). The predicted transcript ratios were correlated with the experimental ones for *pgi*, *fbp*, *fba*, *pykF*, *gnd* and *rpe* knockout mutants of *E. coli*. The coefficients of determination, R^2, are more than 0.6 and the slopes are between 0.9253 and 1.0975. These statistical analyses demonstrate that mCEF provides a significant correlation between the predicted gene expression and experimental data.

Here, we investigate how the omitted gene expression data affect the accuracy of the subsequent prediction of a flux distribution by ECF. If a large error occurs at a high flux reaction, the prediction error for the flux distribution would increase. However, most of the gene expressions with large errors are related to the reactions with a small flux (with less than 10), such as *ppc* and *pck* in a *pgi* mutant, *edd*, *rpi*, *taka*, *tktb* and *zwf* in a *gnd* mutant, and *ack*, *eda*, *edd*, *pps*, *rpi*, *tkta* and *tktb* in a *rpe* mutant. Thus, we would not expect that the errors in such gene expressions significantly affect the subsequent ECF-based prediction of a flux distribution.

The enzyme activities ratios in a pyruvate decarboxylase (*pdc*) over-expression mutant of *S. cerevisiae* to wild type (van Hoek *et al*., 1998) were calculated by mCEF, as shown in Figure 2(G). The value of the relative enzyme activity of *pdc* was approximately fourteen. mCEF is effective in predicting gene over-expression mutants, although the algorithm proposed by Çakır and co-workers is applicable only to the partially or fully deficiency of enzymes in metabolic networks (Çakır *et al*., 2004b). If their algorithm is used to calculate the gene expression profile in the *pdc* overexpression mutant of *S. cerevisiae*, the CEF ratio for the *pdc* gene is one so that the gene expression profile of the mutant is the same as that of wild type.

To investigate how over-expression of a specific gene affects the gene expression profile according to mCEF, the profile for the *C. glutamicum* mutant that over-synthesizes G6P dehydrogenase is calculated by mCEF, as shown in Figure 3. The relative enzyme activity was changed from zero to fifteen. The gene expression profile was changed greatly when G6P dehydrogenase was deficient or over-expressed less than the 2-fold compared to those of wild type, while more than 2-fold over-expression of the gene did not remarkably change the expression profile. It suggests that over-expressing mutants do not so greatly change the flux distribution as the deficient and knockout mutants.

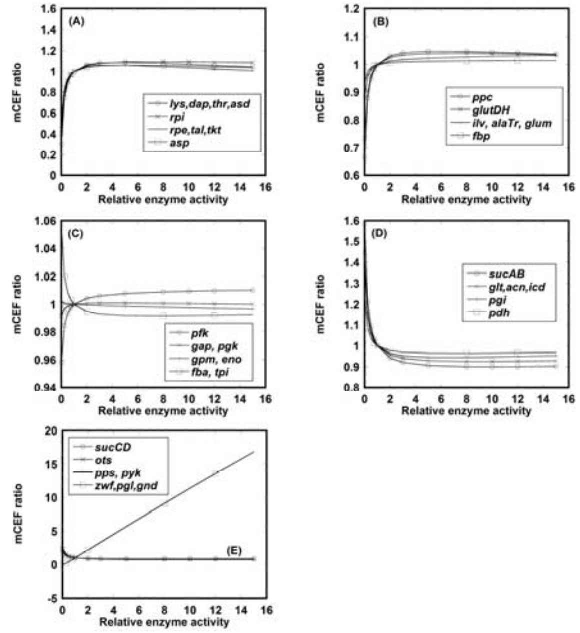

Fig. 3 mCEF ratios for each gene expression in a genetically modified mutant of *C. glutamicum*. The relative enzyme activity of G6P dehydrogenase is changed from 0 to 15.

3.3 Prediction of flux distributions of gene deletion mutants by GMF

To demonstrate the feasibility of GMF in comparison with FBA and MOMA, GMF is applied to predict the flux distribution of *E. coli* gene deletion mutants of *zwf*, *gnd* (Zhao *et al.*, 2004), *pyk*F (Siddiquee *et al*., 2004), *ppc* (Peng *et al*., 2004), or *pgi* (Hua *et al*., 2003) under aerobic conditions, as shown in Figure 4 and Table 1. In this comparison, FBA employed a typical objective function, biomass production, because biomass production is a most suitable objective function under these specific conditions (Schuetz *et al*., 2007). GMF predicts the intracellular fluxes for these five mutants more accurately than FBA and MOMA (Segrè *et al*., 2002), indicating that GMF takes advantages in estimating the change in gene expression pattern for flux predictions. Neither FBA nor MOMA explicitly considers any change in gene expression pattern caused by genetic modification.

Since there are generally no unique objective functions common to all of the physiological and genetic conditions of microorganisms in FBA (Schuetz *et al*., 2007), different objective functions (ATP yield per flux unit and biomass formation) are used for FBA to predict the fluxes of nine gene deletion mutants, as shown in Table S9 (Supplementary data). The prediction errors by GMF are less than those by FBA with two different objective functions, where the prediction accuracy of ATP yield per flux unit shows a similar tendency to that of maximum biomass formation. It supports that GMF can predict the flux distributions for genetic mutants more accurately than FBA.

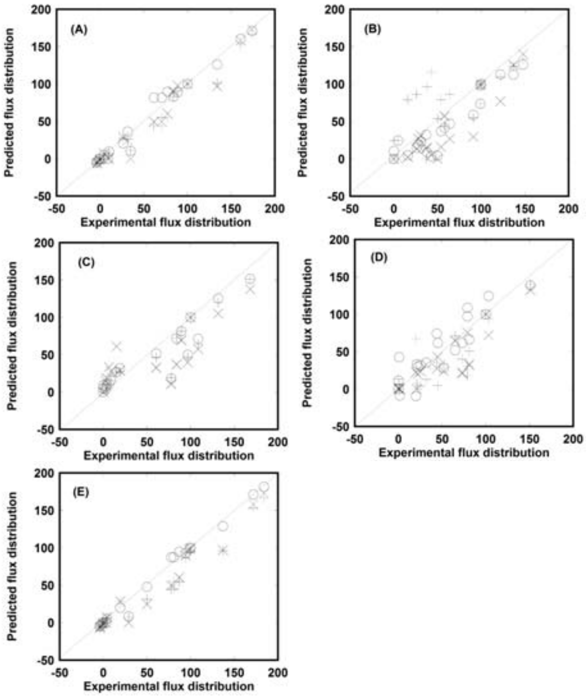

Fig. 4 Comparisons of the predicted and experimental flux distributions in *E. coli* gene deletion mutants. (A) *gnd* mutant, (B) *pgi* mutant, (C) *ppc* mutant, (D) *pyk*F mutant, (E) *zwf* mutant. Cultivation is carried out under aerobic conditions. The prediction is carried out by FBA (×), MOMA (+) and GMF (O).

Table 1 Prediction errors for five mutants of *E. coli* by FBA, MOMA and GMF

Method	*zwf*	*gnd*	*pgi*	*ppc*	*pyk*
FBA	18.38	14.76	23.68	29.92	21.10
MOMA	18.06	14.27	29.38	19.79	25.83
GMF	6.43	9.21	18.47	18.95	20.46

3.4 Applicability to over-expression and under-expression mutants

To show another advantage of GMF, GMF is applied to prediction of the flux distribution of gene over-expression (Becker *et al*., 2007; Becker *et al*., 2005; Nicolas *et al.*, 2007) and under expression mutants (Ohnishi *et al*., 2005), as shown in Figure 5, while existing algorithms including MOMA and rFBA are not applicable to them. GMF accurately predicts the flux distributions of these four mutants, indicating that GMF is feasible for a broad range of genetic modification.

To confirm the validity for GMF to over- or under-expression mutants, we compared the prediction errors by GMF with those by a control (mock) method where all the expression ratios are set to one. The prediction errors by GMF are 1.18, 9.37 and 10.00 for a *zwf* over-expressing mutant of *E. coli*, *zwf* and *fbp* over-expressing mutants of *C. glutamicum*, respectively. Those by the control method are 2.62, 9.64, and 10.40. This shows that accounting for a change in gene expression takes an advantage for enhanced prediction accuracy.

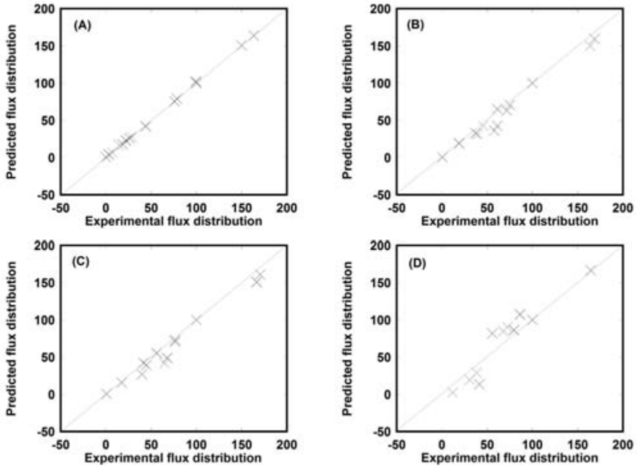

Fig. 5 GMF-predicted flux distribution for gene over-expression or under-expression mutants. (A) *zwf* mutant of *E. coli* (enzyme activity parameter, EAP: 15.11; Nicolas *et al*., 2007), (B) *zwf* mutant of *C. glutamicum* (EAP: 3.6; Becker *et al*., 2007),(C) FBP mutant of *C. glutamicum* (EAP: 9.3; Becker *et al*., 2005) ,(D) *gnd* mutants of *C. glutamicum* (EAP: 0.43; Ohnishi *et al*., 2005)

On the other hand, for a *gnd* under-expressing mutant of *C. glutamicum* (Figure 5D) the prediction error is 15.26 by GMF while it is 2.08 by the control method. In this mutant, the enzyme activity of 6-phosphogluconate dehydrogenase (*gnd*) decreased to about 43% of wild type, but the flux responsible for the *gnd* gene increased from 38.1 (wild type) to 41.4 (mutant). It seems some confliction with the predicted result by GMF. Since no Enter-Doudoroff pathway is seen in the metabolic reaction of *C. glutamicum* in KEGG (http://www.genome.jp/kegg/) and a genome-scale network model (Kjeldsen and Nielsen, 2009), the flux for *gnd* should be less than that of wild type. This inconsistency between the predicted and experimental data may need further studies.

As shown in Figure 6, GMF simulates the flux of the *zwf*-catalyzed reaction (G6P dehydrogenase reaction) in a *C. glutamicum* mutant (Nicolas *et al*., 2007), where the relative enzyme activity of *zwf* changes from zero to fifteen. The predicted fluxes are consistent with the experimental ones. The flux rapidly increases at a low value (1-5) of the relative enzyme activity and then saturated. Over-expression of a specific gene does not always lead to a linear increase in the flux, because the *in vivo* flux is constrained by the supply of substrates.

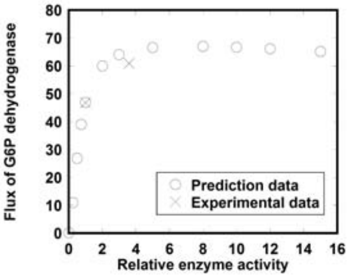

Fig. 6. The flux of the G6P dehydrogenase reaction with respect to the relative enzyme activity in a *zwf* mutant of *C. glutamicum*. The relative enzyme activity of G6P dehydrogenase is changed from zero to fifteen. Predicted flux (O) and experimental flux (×).

4 DISCUSSION

A breakthrough in GMF is that the mCEF algorithm enables the prediction for the gene expression patterns of genetically modified mutants that are required by ECF. The prediction by mCEF is performed based on the metabolic pathway architecture in terms of the ATP generation and biomass production. This reflects that a change in the gene expression pattern in metabolic networks is closely related to such specific biological functions.

ECF uses the gene expression profile estimated by mCEF to predict the flux distributions. In ECF, an enzyme activity profile is readily incorporated into the EMCs in the multiplication form, based on the fact that the flux distribution would be determined by many enzyme activities rather than a few rate limiting enzymes (Heinrich and Rapoport, 1974, Small and Kacser, 1993).

GMF shows a higher prediction accuracy of the flux distribution of a broad range of genetic mutants. To demonstrate the feasibility of GMF, it is important to compare it with typical algorithms of MOMA and FBA. The requirements of GMF are the same as MOMA: a metabolic network and a flux distribution of wild type, to predict the flux distribution of genetic mutants. Thus, we can directly compare their performance in terms of prediction accuracy and its applicability. On the other hand, FBA does not require the flux distribution of wild type but biological objective functions. The prediction accuracy of FBA generally depends on selection of objective functions (Schuster *et al*., 2008). Thus, different types of them are used to compare FBA with GMF (Table S9). Consequently, it is possible to say that GMF provides a high accuracy for prediction. Another advantage of GMF over these existing algorithms is that it is applicable to not only gene deletion mutants but also gene over-expression or under-expression mutants. This contributes to advances in design of biosynthesis by genetically engineered strains, because overexpression or under-expression of target genes is a promising strategy for enhanced production.

As mentioned above, GMF shows high predictive capability and applicability to a broad range of genetic mutants, while GMF is limited to small- or moderate-scale metabolic networks because it is based on EMs. To apply GMF to genome-scale networks it is necessary to develop new algorithms to avoid combinatorial explosion in the number of EMs.

ACKNOWLEDGEMENTS

This work was supported by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Systems Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan and partially by KAKENHI (Grant-in-Aid for Scientific Research (B), 2006, 18300098).

Conflict of Interest: None declared.

REFERENCES

Akesson,M. *et al*. (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.*, **6**, 285-293.

Atsumi,S. *et al*. (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, **451**, 86-89.

Becker,J. *et al*. (2005) Amplified expression of fructose 1,6-bisphosphatase in *Corynebacterium glutamicum* increases *in vivo* flux through the pentose phosphate pathway and lysine production on different carbon sources. *Appl. Environ. Microbiol.*, **71**, 8587-8596.

Becker,J. *et al*. (2007) Metabolic flux engineering of L-lysine production in *Corynebacterium glutamicum*--over expression and modification of G6P dehydrogenase. *J. Biotechnol.*, **132**, 99-109.

Çakır,T. *et al*. (2004a) Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnol. Bioeng.*, **86**, 251-260.

Çakır,T. *et al*. (2004b) Metabolic pathway analysis of enzyme-deficient human red blood cells. *Biosystems*, **78**, 49-67.

Çakır,T. *et al*. (2007) Effect of carbon source perturbations on transcriptional regulation of metabolic fluxes in *Saccharomyces cerevisiae*. *BMC Syst. Biol.*, **1**, 18.

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

Fischer,E. and Sauer,U. (2005) Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat. Genet.*, **37**, 636-640.

Förster,J. *et al*. (2002) A functional genomics approach using metabolomics and *in silico* pathway analysis. *Biotechnol. Bioeng.*, **79**, 703-712.

Gayen,K. and Venkatesh,K.V. (2006) Analysis of optimal phenotypic space using elementary modes as applied to *Corynebacterium glutamicum*. *BMC Bioinformatics*, **7**, 445.

Heinrich,R. and Rapoport,T.A. (1974) A linear steady state treatment of enzymatic chains. General properties, control and effector strength. *Eur. J. Biochem.*, **42**,89-95.

Hua,Q. *et al*. (2003) Responses of the central metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts. *J. Bacteriol.*, **185**, 7053-7067.

Ideker,T. *et al* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-934.

Ishii,N. *et al*. (2007) Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, **316**, 593-597.

Joyce,A.R. and Palsson,B.O. (2006) The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell. Biol.*, **7**, 198-210.

Kauffman,K.J. *et al*. (2003) Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, **14**, 491-496.

Kjeldsen,K.R. and Nielsen,J. (2009) *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.*,102,583-597.

Klamt,S. *et al* (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst. Biol.*, **1**, 2.

Kurata,H. *et al* (2003).CADLIVE for constructing a large-scale biochemical network based on a simulation-directed notation and its application to yeast cell cycle. *Nucleic Acids Res*. 2003;31(14):4071-4084.

Kurata,H. *et al* (2007a) Extended CADLIVE: a novel graphical notation for design of biochemical network maps and computational pathway analysis. *Nucleic. Acids Res.*, **35**, e134.

Kurata,H. *et al* (2007b) Integration of enzyme activities into metabolic flux distributions by elementary mode analysis. *BMC Syst. Biol.*, **1**, 31.

Lee,S.Y. *et al*. (2005) Systems biotechnology for strain improvement. *Trends Biotechnol.*, **23**, 349-358.

Nicolas,C. *et al*. (2007) Response of the central metabolism of *Escherichia coli* to modified expression of the gene encoding the glucose-6-phosphate dehydrogenase. *FEBS Lett.*, **581**, 3771-3776.

Nishio,Y. *et al*. (2008) Computer-aided rational design of the phosphotransferase system for enhanced glucose uptake in *Escherichia coli*. *Mol. Syst. Biol.*, **4**, 160.

Ohnishi,J. *et al*. (2005) A novel *gnd* mutation leading to increased L-lysine production in *Corynebacterium glutamicum*. *FEMS Microbiol. Lett.*, **242**, 265-274.

Park,J.H. *et al*. (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proc. Natl. Acad. Sci. USA*, **104**, 7797-7802.

Peng,L.F. *et al*. (2004) Metabolic flux analysis for a *ppc* mutant *Escherichia coli* based on C-13-labelling experiments together with enzyme activity assays and intracellular metabolite measurements. *FEMS Microbiol. Lett.*, **235**, 17-23.

Schilling,C.H. *et al*. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.*, **184**, 4582-4593.

Schuetz,R. *et al*. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol*., **3**, 119.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53-60.

Schuster,S. *et al*.(2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326-332.

Schuster,S. *et al*. (2008) Is maximization of molar yield in metabolic networks favoured by evolution? *J. Theor. Biol.*, **252**, 497–504.

Schwartz,J.M. and Kanehisa,M. (2005) A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics*, **21 Suppl 2**, ii204-ii205.

Schwartz,J.M. and Kanehisa,M. (2006) Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics,* **7**, 186.

Segrè,D. *et al*. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA*, **99**, 15112-15117.

Shlomi,T. *et al*. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. USA*, **102**, 7695-7700.

Siddiquee,K.A. *et al*. (2004) Effect of a pyruvate kinase (*pyk*F-gene) knockout mutation on the control of gene expression and metabolic fluxes in *Escherichia coli*. *FEMS Microbiol. Lett.*, **235**, 25-33.

Small,J.R. and Kacser,H. (1993) Responses of metabolic systems to large changes in enzyme activities and effectors. 1. The linear treatment of unbranched chains. *Eur. J. Biochem.*, **213**, 613-624.

Stelling,J. *et al*. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190-193.

Stephanopoulos,G. (2007) Challenges in engineering microbes for biofuels production. *Science*, **315**, 801-804.

van Hoek,P. *et al*. (1998) Effects of pyruvate decarboxylase overproduction on flux distribution at the pyruvate branch point in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **64**, 2133-2140.

Wiback,S.J. *et al*. (2004) Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the *Escherichia coli* spectrum. *Biotechnol. Bioeng.,* **86**, 317-331.

Zhao,J. *et al*. (2004) Global metabolic response of *Escherichia coli* to *gnd* or *zwf* gene-knockout, based on C-13-labeling experiments and the measurement of enzyme activities. *Appl. Microbiol. Biotechnol.*, **64**, 91-98.

Zhao,Q. and Kurata,H. (2009) Maximum entropy decomposition of flux distribution at steady state to elementary modes. *J. Biosci. Bioeng.*, **107**, 84-89.