A Method to Calibrate Metabolic Network Models with Experimental Datasets

Octavio Perez-Garcia[1], Silas Villas-Boas[2], and Naresh Singhal[1]

[1] Department of Civil and Environmental Engineering, University of Auckland, New Zealand
[2] Centre for Microbial Innovation, School of Biological Sciences,
University of Auckland, New Zealand

Abstract. A method to calibrate stoichiometric coefficients values related to uncharacterized or lumped reactions of metabolic network models is presented. The method finds coefficients values that produce a model version that best fits multivariable experimental data. The method was tested with a metabolic network of 44 metabolites and 49 stoichiometric reactions, with four reactions having undetermined stoichiometric coefficients values. A total of 1320 model versions with different combinations of stoichiometric coefficient values were generated. Experimental data was used to produce a calibration curve and different fitness scores were used to evaluate the accuracy of flux balance analysis (FBA) simulations of these model versions to reproduce the experimental data. The model version with highest fitness to the experimental data was found using Mean Relative Error (MRE) scores and auto-scaled transformation of estimated datasets.

Keywords: Metabolic network model, biochemical reaction stoichiometry, flux balance analysis, model calibration.

1 Introduction

Stoichiometric metabolic network (SMN) modeling and flux balance analysis (FBA) are emerging techniques in systems biology that can be used to quantify the rate of reactions within the network formed by chemical compounds and sequenced chemical reactions in cells' metabolism (Oberhardt, et al. 2009; Orth, et al. 2010). The accuracy of these techniques in estimating observation within real cells is dependent on rigorous model calibration (or model curation) (Edwards, et al. 2001). Model calibration involves contrasting the model's estimated values with experimental data (obtained from measurements in real systems) in order to refine model structure until differences between the datasets are minimal (Feist, et al. 2009). When experimental data is limited or does not contain information appropriate for model calibration, it may limit the ability of SMN models in producing relevant and accurate estimates.

The lack of experimental data for model calibration is a common problem in SMN modeling. For example, the general procedure for SMN model calibration involves comparing estimated and experimental growth rates observed under different organic carbon sources (Durot, et al. 2009; Edwards, et al. 2001). Thus, de facto this approach may not be applicable to calibrating models of organism or microbial communities

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_22, © Springer International Publishing Switzerland 2014

that don't growth using organic molecules (i.e. autotrophs). Another scenario arises with 13C-flux analyses, which provide metabolic reaction rate measurements that can be directly compared with FBA simulations; however, experimental data is limited by expensive nature of the technique and it ability to provide rate measurements only for the central carbon pathway (Sauer 2006). Moreover, the most valuable experimental data for model calibration and validation come from transcript-, proteo- and metabol–omic analyses (Kümmel, et al. 2006; Lewis, et al. 2010). This however requires pre-existing analytical and expertise capability to obtain relevant -omic data. Because of these limitations, it is desirable to implement computational methods to calibrate SMN models using common bioprocess performance data.

In this work we introduce a method to calibrate stoichiometric coefficients values related to uncharacterized or lumped reactions of SMN models based on minimizing model estimation errors to reproduce experimental datasets composed of process variables such as culture yield, compound uptake and production rates, and other variables describing the steady state metabolism of the cell culture under specific experimental conditions. This multivariate experimental dataset is used to measure accuracy of model simulations via various goodness to fit scores.

2 Methods

2.1 Metabolic Network

A metabolic network of the nitrogen respiration pathway of *Nitrosomonas europaea* was used to develop the calibration method. The metabolic network, constructed using (Thiele and Palsson 2010) protocol, consisted of 44 metabolites, 49 stoichiometric reactions and 3 compartments. The COBRA toolbox 2.0 (Schellenberger, et al. 2011) together with the GLPK solver running in Matlab®7 R2010b software (MathWorks Inc., Natick, MA, USA) was used to convert the metabolic network into its mathematical form and perform FBA simulations. Mathematically the network was represented as a stoichiometric matrix, $S\ (m \times n)$, of m metabolites and n reactions. A non-zero entry $s_{i,j}$ in S indicates the participation of metabolite i in reaction j. All reactions within the network were mass-balanced such that $S * v = 0$, where v was the vector of reaction rates (or fluxes) (Feist, et al. 2009; Varma and Palsson 1994). The reaction rate limits (or constraints) were defined in the form $\alpha_j \leq v_j \leq \beta_j$, where α_j and β_j are the lower and upper limits placed on the reaction rate v_j (Varma and Palsson 1994).

The metabolic network of *N. europaea* respiration pathways had to be calibrated (curated) as it contained four reactions involving compounds with undetermined stoichiometric coefficients. These four reactions (presented in Table 1) corresponded to electron transport chain reactions that translocate protons across the cellular membrane and the ATP stoichiometric coefficient in the biomass reaction. The precise stoichiometry of these reactions was ambiguous or not found in the literature.

2.2 Generation of Model Versions and Simulations

Candidate stoichiometric coefficients were defined in the above four reactions so that mass and energy balance was preserved, as presented in Table 1. Although Table 1 shows that fraction values were not assigned to candidate coefficients, if needed, these can be directly assigned without method modification.

Table 1. Reactions of the SMN model calibrated in this study

Rx. ID	Stoichiometric equation	Candidate coefficients “*s*”	Number of coefficients
A	nh3[p] + o2[p] + q8h2[c] + *s* h[c] → nh2oh[p] + h2o[c] + q8[c] + *s* h[p]	0, 1, 2	3
B	q8h2[c] + 2 cyt552[p] + ***s*-2** h[c] → *s* h[p] + q8[c] + 2 cyt552e[p]	2, 4	2
C	atp[c] + 10 nadh[c] + 0.25 protein[c] + *s* m[c] ←→ adp[c] + 10 nad[c] + pi[c] + 0.26 h[c] + biomass[c]	From 0 to 50 each 5	20
D	*s* atp[c] + 10 nadh[c] + 0.25 protein[c] + m[c] ←→ *s* adp[c] + 10 nad[c] + *s* pi[c] + *s* h[c] + biomass[c]	From 0 to 100 each 10	11

By systematically combining the candidate stoichiometric coefficients for each reaction with those for the remaining three reactions we obtained a total of 1320 combinations (3 x 2 x 20 x 11 = 1320) that gave 1320 model versions. These model versions and their corresponding FBA simulations were automatically generated with the following Matlab® script:

```
%% Run a FBA simulation for all model versions generated by changing the
S matrix.
%Define column vectors “coeffsA”, “coeffsB”, “coeffsB” and “coeffsD” of
coefficient values of each reaction. All column vectors must be of the
same length.
coeffsA=[];
coeffsB=[];
coeffsC=[];
coeffsD=[];
%Assign the values of “coeffs” to the corresponding stoichiometric
coefficient with coordinates (i,j) in model.S matrix and run a FBA
simulation. Values of column vectors are assigned row by row until the
length of “coeff1”.
for j=1:length(coeff1);
    model.S(43,50)=coeff1(j);
    model.S(42,50)=coeff2(j);
    model.S(26,50)=coeff2(j);
    model.S(11,50)=coeff2(j);
    model.S(11,33)=coeff4(j);
    model.S(20,33)=coeff5(j);
    model.S(11,41)=coeff6(j);
    model.S(20,41)=coeff7(j);
    solution=optimizeCbModel(model);    %FBA simulation
    M(:,j)=solution.x         %Generation of “M” matrix
end
```

The above script generated a M matrix of n number of reactions and d number of FBA solutions (d=1320). Note that FBA simulations can be substituted by other methods to estimate network fluxes, such as random sampling or extreme pathways.

2.3 Defining Experimental Datasets for Calibration

The compound concentration curves, biomass concentration, reactor volume, inflow rate of growth medium, and other biochemical information reported in experiments previously published (Vadivelu, et al. 2006; Whittaker, et al. 2000) on *N. europaea* growing in aerobic conditions without substrate limitation (oxygen and ammonium) were used to define an experimental dataset X_l of 28 mean values of l variables that describe the metabolism of this organisms under specified conditions. Table 2 defines the categories to which the 28 variables of the experimental dataset belong.

Table 2. Definition of variables used to produce the experimental dataset and number of dataset variable values found in previously published experiments

Variable category	Formula for variable estimation with model simulation results	Number of dataset variables
Growth rate	$= v_{biomass}$	1
Specific substrate uptake rate	$= v_{substrate}$	2
Specific compound production rate	$= v_{product}$	2
Molar yield ratio of product	$= \frac{v_{product}}{v_{substrate}}$	4
Net amount of compound used in reaction	$= v_j * s_i$	2
ATP molar yield ratio	$= \frac{\sum_{j=1}^{J}(v_{ATP\ consumption} * s_{ATP})_j}{\sum_{j=1}^{J}(v_{ATP\ synthezis})_j}$	2
Proton translocation yield ratio	$= \frac{\sum_{j=1}^{J}(v_{H+\ produced} * s_{H+})_j}{v_{substrate}}$	2
Pivot compound reaction yield ratio	$= \frac{v_{consumption\ of\ i\ in\ reaction\ j}}{v_{synthezis\ of\ i}}$	6
Percentage yield ratio of element	$= \frac{v_{production\ of\ i} * s_i * a}{v_{consumption\ of\ i} * s_i * a} * 100$	4
Element mass balance	$= \sum_{j=1}^{J}(v_j * s_i * a)_j$	3
TOTAL		28

Note: v_j is the rate of reaction j; s_i is the stoichiometric coefficient of compound i; J is the total number of reactions that consume or/and produce the compound i; and a is the element's number of atoms in compound i.

The 28 variables apply to the same specific steady state condition of *N. europaea* growth. In the case of experiments with batch cultures a steady state was assumed for time periods where the change in substrate and product concentrations maintained a linear trend, therefore indicating a constant rate of consumption and production of compounds. The rate variables were normalized by the total biomass in bioreactor (expressed as grams of dry weight (gDW)).

2.4 Evaluation of Goodness to Fit

By applying the generic formulas presented in Table 2, the 28 variables were estimated using the simulation results of the M matrix to produce a x_l dataset for each of the 1320 model versions. Experimental and estimated datasets ($X_l$and x_l respectively) were $\log_{10}$ or auto-scale transformed because dataset values had different order of magnitude and dimensions (e.g. $v_{O2\,uptake}$ = 2.5mmol/gDW*h while the $v_{biomass}/v_{O2\,uptake}$ yield = 0.012gDW/mmol-O_2). Data transformation was necessary for capturing the deviation between the observed and estimated values in absolute terms and to minimize the effect of varying scales for different variables (Schuetz, et al. 2007; van den Berg, et al. 2006). The goodness to fit was evaluated after data transformation.

The goodness to fit of a model describes the degree to which model predictions fit experimental data (Makinia 2010). The overall fitness between experimental and estimated datasets was evaluated using the fitness scores presented in Table 3. Model version with lower fitness scores was considered to have the highest accuracy in reproducing the experimental data, and therefore considered to yield a calibrated model.

Table 3. Some formulas to evaluate goodness to fit of a model. Modified from (Makinia 2010)

Fitness score	Formula
Mean relative error (MRE)	$MRE = \frac{1}{n}\sum_{l=1}^{n}\frac{\lvert(X_l - x_l)\rvert}{X_l}$
Mean absolute error (MAE)	$MAE = \frac{1}{n}\sum_{l=1}^{n}\lvert(X_l - x_l)\rvert$
Root mean squared error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(X_l - x_l)^2}$
Root mean squared scaled error (RMSSE)	$RMSSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{m_l(X_l - x_l)^2}{(std)_l^2}}$

Note: n is the total number of variables in datasets (observed and estimated); X_l is the observed data (measured in experiments) in variable l; x_l is the estimated data (estimated in simulations) in variable l; m_l is the number of data points contributing to X_l; std is the standard deviation of the observed data (measured in experiments) in variable l.

3 Results

Fig 1 presents the fitness scores obtained for the 1320 model versions. MRE and RMSSE scores similarly ranked the fitness of model versions and identified model version number "757" as showing the best fitness to experimental data. Model version "757" had the following stoichiometric coefficient values: $s_{i,A}$ = 1, $s_{i,B}$ = 2, $s_{i,C}$ = 5, $s_{i,D}$ = 80. On other hand, according to MRE and RMSSE scores, model version "24" had the worst fitness to the experimental data. MAE and RMSE scores gave different ranks for the model versions from each other as well as from MRE and RMSSE.

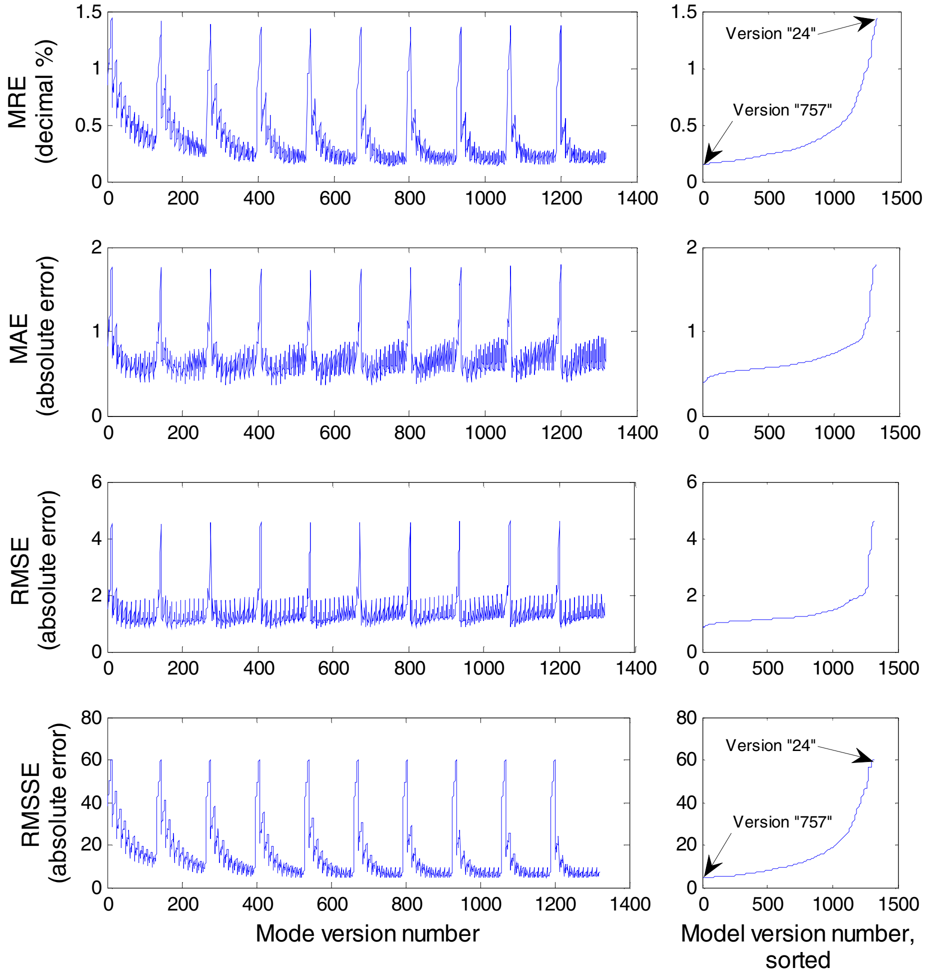

Fig. 1. Plots of fitness scores obtained for the 1320 model versions. Plots of the right column present the same scores but model versions were sorted from lower to higher fitness scores.

Fig 2 presents calibration curves generated by plotting experimental vs estimated datasets for model version "24" and "757". Fig. 2 also presents the effect of log transformed and auto-scaled datasets. Without transformation the deviation between estimated and experimental datasets cannot be visually evaluated during calibration because the large variation in the scale of different variables. The Log_{10} transformed and auto-scaled data gave less noisy calibration curves because all variables were re-scaled to the same units.

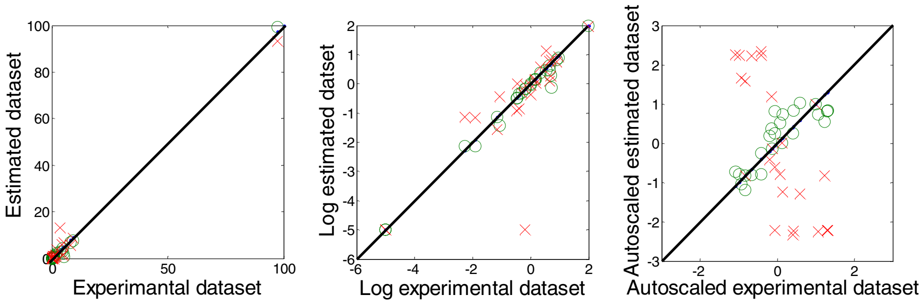

Fig. 2. Model calibration curves generated without treatment, $\log^{10}$ transformed treatment and auto-scaled treatment of datasets. Diagonal line represents a perfect fitness between experimental and estimated datasets. "X" markers represent the estimated dataset of model version "24". "O" markers represent the estimated dataset of model version "757".

The calibration curves in Fig. 2 show similar results to those obtained for fitness scores, especially with auto-scaled data. Model versions fitness can therefore be evaluated in both ways, with fitness scores or visually using calibration curves. MRE and RMSEE scores provided the most meaningful fitness scores because of two reasons: i) they corrected differences in the scales or units of variables and ii) they better reflect small deviations between experimental and estimated datasets. However MRE is a more meaning full score because error is measured on a scale from 0 to 1 (0 represents a perfect fit) which allows one to directly determine the percentage of accuracy with the formula: $accuracy\,(\%) = ((1 - MRE) * 100)$. Auto-scaled transformations provided the best way of visually representing the deviations of estimated datasets from experiments because auto-scaled datasets are more sensitive to numerical differences between the two datasets.

4 Conclusions

The method presented in this work is an easy to implement way of calibrating stoichiometric coefficients in SMN models. The essence of the method is to find those coefficients that produce a model version that best fits experimental data. In this sense, SMN model structure (coefficients of the S matrix) is defined by experimental data and this ensures realistic estimates for intracellular reaction rates. Extracting the maximum possible information on values of variables from experiments or literature is necessary to produce a robust experimental dataset with large number of variables l and thereby improve the fitness scores. This method can be applied to evaluate the fitness of multiple metabolic reactions constraints and can also be extended to fit models to transcriptomics, proteomics or metabolomics data by defining new variables (as in Table 3) using parsimonious enzyme usage FBA (pFBA) (Lewis, et al. 2010) and network embedded thermodynamic (NET) analysis (Kümmel, et al. 2006). However these applications are out of the scope of this research.

References

1. Durot, M., Bourguignon, P., Schachter, V.: Genome-scale models of bacterial metabolism: Reconstruction and applications. FEMS Microbiol. Rev. 33, 164–190 (2009)
2. Edwards, J.S., Ibarra, R.U., Palsson, B.: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19, 125–130 (2001)
3. Feist, A.M., Herrgård, M.J., Thiele, I., Reed, J.L., Palsson, B.: Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology 7, 129–143 (2009)
4. Kümmel, A., Panke, S., Heinemann, M.: Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Molecular Systems Biology 2 (2006a)
5. Lewis, N.E., Hixson, K.K., Conrad, T.M., Lerman, J.A., Charusanti, P., Polpitiya, A.D., Adkins, J.N., Schramm, G., Purvine, S.O., Lopez-Ferrer, D., Weitz, K.K., Eils, R., König, R., Smith, R.D., Palsson, B.Ø.: Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular Systems Biology 6 (2010)
6. Makinia, J.: Mathematical Modelling and Computer Simulation of Activated Sludge Systems. IWA Publishing, London (2010)
7. Oberhardt, M.A., Chavali, A.K., Papin, J.: Flux balance analysis: Interrogating genome-scale metabolic networks. Methods in Molecular Biology 500, 61–80 (2009)
8. Orth, J.D., Thiele, I., Palsson, B.: What is flux balance analysis? Nat. Biotechnol. 28, 245–248 (2010)
9. Sauer, U.: Metabolic networks in motion: 13C-based flux analysis. Molecular Systems Biology 2 (2006)
10. Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bord-bar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., Palsson, B.Ø.: Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. Nature Protocols 6, 1290–1307 (2011)
11. Schuetz, R., Kuepfer, L., Sauer, U.: Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular Systems Biology 3 (2007)
12. Thiele, I., Palsson, B.Ø.: A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5, 93–121 (2010)
13. Vadivelu, V.M., Keller, J., Yuan, Z.: Stoichiometric and kinetic characterisation of Nitrosomonas sp. in mixed culture by decoupling the growth and energy generation processes. J. Biotechnol. 126, 342–356 (2006)
14. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.: Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC Genomics 7 (2006)
15. Varma, A., Palsson, B.: Metabolic flux balancing: Basic concepts, scientific and practical use. Bio/Technology 12, 994–998 (1994)
16. Whittaker, M., Bergmann, D., Arciero, D., Hooper, A.: Electron transfer during the oxidation of ammonia by the chemolithotrophic bacterium Nitrosomonas europaea. Biochimica et Biophysica Acta - Bioenergetics 1459, 346–355 (2000)