The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 373–380

Impacts of Different Metabolite Measurement Protocols on Estimating Parameters in Complex Photosynthetic Carbon Metabolism Models

Wei Pan[1,2] John Fettig[3] Eric de Sturler[4]
Xin-Guang Zhu[1,5,*]

[1]Plant Systems Biology Group, Partner Institute of Computational Biology,
Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China
[2]Department of Electronic Science and Technology,
University of Science and Technology of China, Hefei, 230027, China
[3]Department of Theoretical and Applied Mechanics,
University of Illinois at Urbana Champaign, Urbana, Illinois 61801
[4]Department of Mathematics, Virginia Tech, Blacksburg, Virginia, 24601-0123
[5]Institute of Plant Physiology and Ecology,
Chinese Academy of Sciences, Shanghai, 200032, China

Abstract Obtaining kinetic constants used in rate equations is a key component of building systems model of metabolism. The availability of high throughput metabolomics measurements provides an opportunity to use reverse engineering approach to estimate kinetic parameters such as maximal rate of enzyme $V_{\max}$ especially. We presented a method of using two optimization algorithms respectively to estimate $V_{\max}$ using time series measurements of incomplete metabolites combined with a metabolism model of photosynthetic carbon metabolism. The impact of choosing different experiment protocols on the accuracy of parameter estimation was evaluated. The result showed that using steady state initial condition and multiple measurements at each points can give the best estimates of the estimation results. In addition, the choice of time intervals of time series data does influence the estimation, i.e. focus on transient response or steady state part of time series data will be better for estimation.

1 Introduction

Metabolism simulation is important for understanding the function of metabolism in the post-genomic era. Different metabolism models have been built, see review [1, 2]. Metabolism models can be used to testing hypothesis regarding certain dynamic behaviors, asking “what if”questions, exploring better designs for achieving certain target function, or modification of metabolic pathway for generation of new functions [3]. The model of photosynthetic carbon metabolism is one of such models. The individual steps of photosynthesis process have been studied in great detail, which have been used to build

[*] Corresponding Author. Email: zhuxinguang@picb.ac.cn

complete mathematical model of photosynthesis [4, 5]. The complete model of photosynthesis has wide applications, such as identifying enzymes for improved crop yield, identifying enzymes limiting photosynthetic efficiency under different conditions [6], studying the mechanisms of transient signals of CO_2 uptake and fluorescence emission [7]. However, it is difficult to differentiate the activities of enzymes in the chloroplast stroma or in the cytosol. In addition, the activities of different enzymes are under constant modifications in the field, and therefore, are different among different leaves [8]. So the measured enzyme activities from large quantity of leaf material, as required for enzyme activity measurement, will inevitably only represent averaged enzyme activities and not represent any individual leaf performance. The recent development of metabolomics technology provides a new opportunity to solve the problem of lack of kinetic information [3, 9], i.e. using the measured time series measurements of metabolite concentrations to estimate the enzyme concentrations based on the kinetic models.

Parameter estimation in biochemical pathways has attracted increasing attention in recent years [10]. However, many issues related to parameter estimation using kinetic models still need to be addressed. First of all, the time series data for parameter estimation are inherently noisy since the samples for metabolites measurements have to be taken from different parts of the leaves, or even different leaves. In addition, most metabolites involved in the photosynthetic carbon metabolism exist in cytosol and participate in other metabolism as well. Furthermore, multiple measurements of metabolite concentrations, which are required to gain more accurate estimate of enzyme concentrations, need to be taken for leaves at the same physiological status, e.g. at 10 seconds for dark-adapted leaves.

Here presented a method of using two optimization algorithms respectively to estimate $V_{\max}$ using time series of incomplete metabolite measurements combined with a metabolism model of photosynthetic carbon metabolism. The impact of choosing different experiment protocols on the accuracy of parameter estimation was evaluated. We tested the relationship between estimate accuracy and measure protocol, then came up with an optimal combination of protocols for better estimation. Finally, the new approach incorporates the experiments error in addition to the possible error in time recording in constructing the target functions, and correspondingly has the ability to use different measurement simultaneously for parameter estimation.

2 Results and Discussion

2.1 The mathematical model of photosynthetic carbon metabolism

The photosynthetic carbon metabolism includes the Calvin cycle, photorespiratory pathway, starch synthesis, and triose phosphate export process [4], which is one of the most important metabolic process on this planet due to its role of generating carbohydrate, which is the basis of much of the wants of the human society, such as food, energy, fiber etc.. A detailed mathematical model of this process has been developed [4] and is briefly described here. The reaction diagram is shown in Figure 1 and Figure 2. In these figures, the numbers represent the reaction numbers, which are used in the notation of the rate equations as subscripts. The double-headed arrows represent reversible reactions. The single-headed arrows represent essentially irreversible reactions. The space between two dashed lines represents the chloroplast membrane. Most of the reactions in this system were assumed to follow Michaelis-Menten type kinetics. For those that were not following Michaelis-Menten type kinetic, the best available rate equations from literature were used. The Michaelis-Menten constants for each substrate in each reaction were obtained from literature. For the complete description of the model, see [4, 5].

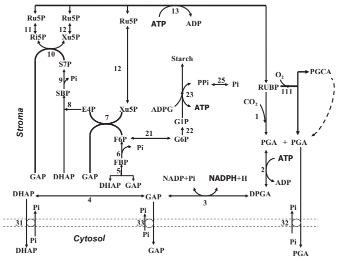

Figure 1: The circles in the chloroplast membrane represent phosphate translocators, which mediate the export of three-carbon metabolites (i.e. DHAP, PGA, and GAP) from the stoma to the cytosol with a counter-import of phosphate.

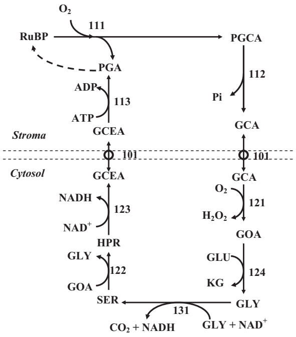

Figure 2: The steps of the photorespiratory reactions represented in the model. The circles in the chloroplast membrane represent glycerate/glycollate translocators, which mediate the transport of glycerate and glycollate through the thylakoid membrane between stroma and cytosol.

2.2 Estimate $V_{\max}$

Here, we apply two different popular optimization algorithms, Levenberg-Marquardt algorithm (gradient algorithm) and Nelder-Mead simplex algorithm (simplex algorithm), to estimate different enzymes' maximum rate $V_{\max}$ in photosynthetic carbon metabolism model. One one hand, we propose an hypothesize that the changing direction and general trend of results in different experiments should be similar though the estimate accuracy differs due to the optimization route. Then the results derived from different algorithms will be plausible. One the other hand, with such a comparison, we could choose the relative optimal algorithm that has lower estimate error for future application in estimating $V_{\max}$. Details on the algorithms and implementation will be discussed in Method section.

In our experiment, we try to minimize the estimated errors for each estimates using different choices of parameters, the number of initial conditions, the number of time points, the number of measurement at each time point, the noise intensity of experiment data, and time intervals among the time span of one experiment. It should be noted that the target value in the target function should be set to the observed value under noise.

2.2.1 Initial condition number

To set different initial conditions means to adopt various initial concentration of metabolites in photosynthetic carbon metabolism model. To vary initial conditions means to set different experiment protocols which will result in the difference in initial conditions. But the Michaelis-Menten constants will remain unchanged even if different initial conditions are selected. This metabolic network could achieve the same steady state regardless of the selection of initial value. However, the transient response would vary at the starting period of time span.

Intuitively, more different initial conditions mean higher estimation accuracy due to the introduction of more sufficient data. However, we notice that number of the random chosen initial conditions have no monotonic relationship with the estimation accuracy. This means that we should not arrange the experiment at bench randomly but following some rules. However, from the results in Figure 3, there are no such rules for the selection of initial conditions. As mentioned, steady state holds the same for the photosynthetic carbon metabolism model regardless of transient response.

Then we suppose the observed steady state as initial condition. To test the hypothesis, estimated errors from random chosen initial conditions and steady state initial condition are compared. From the result in Figure 3, estimated error is relatively low when initial condition number equals 2 in both algorithms. Then we will take this number as default in the following experiment for comparison.

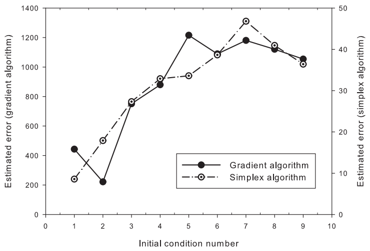

Figure 3: Estimated error under different initial condition numbers. Estimated errors by gradient algorithm and simplex algorithm are showed in left and right axis respectively. The ticks are scaled along y axis.

2.2.2 Measurement number

Because the intrinsic noise in biochemical reactions as well as the perturbation from external environment, measurements at each time point for one initial condition should not be recorded for only one time. However, more measurements means more repeats of experiments thus more cost. So it is necessary to test an optimal measurement number that could guarantee an optimal accuracy. We generate a set of synthetic data under the same initial condition. Then we add noise to the synthetic data at each measurement. The default intensity of noise is set to 0.2 (20% fluctuation around the original data). This noise intensity could vary according to the practical condition of experiment.

The results suggest that the estimated error will decrease when the measure number increases. The result in Figure 4 suggests that the estimated error with only one measurement is much higher than that with more than one measurement. Also, it should be noted in Figure 4 that more measurements indicate more cost in experiment arrangement while achieving slightly higher accuracy. Then a compromise between accuracy and cost should be made. We set 4 as the default measurement number in the following experiment for comparison.

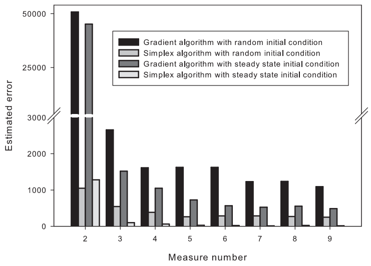

Figure 4: Estimated error under different measurement numbers. The measurement number ranges from 2 to 9. The estimated errors are showed considering type of initial condition and selection of algorithms simultaneously.

2.2.3 Noise intensity

Noise intensity of the data will affect the estimation accuracy. In this experiment, we study which algorithm is more robust against the noise while achieving higher estimation accuracy simultaneously. We assume the data is corrupted by Gaussian white noise with zero mean and variance to be tuned. The results in Table 1 suggest that both algorithms have limitations in coping with relatively high intensity noise. They also show that the estimation accuracy decreases when the noise intensity enhances.

We adopt noise with five different intensities to test the robustness of the two algorithms and determine feasible range of noisy data for estimation. The result in Table 2 suggest that both algorithms have limitation in dealing with noisy data with noise intensity exceeding 1 (that is 100% fluctuation around the original data).

Table 1 - Estimated error under different noise intensity for both algorithms

Noise Intensity/Algorithm	RG	RS	SG	SS
0.0500	209.0381	138.5700	51.9407	15.4618
0.5000	311.3802	640.3600	260.0898	45.6568
1.0000	24452.5218	63626.2000	542007.6761	10268.4733
5.0000	inf	inf	inf	inf
10.0000	inf	inf	inf	inf

Here is a notation for the abbreviation of row *algorithm* in Table 2 . The former letter represents the type of initial conditions: R is for **R**andom initial condition, S is for **S**teady state initial condition; the latter letter represents the selection of optimization algorithms: G is for **G**radient algorithm, S is for **S**implex algorithm.

2.2.4 Sampling intervals

The time intervals determine the amount of data information for parameter estimation. Obviously, taking more points means increase in computation cost. In this study, we try to test the relationship between sampling intervals and estimate accuracy of the optimization.

We focus on two parts, transient response and steady state, that constitute the time series data. Generally, we could hardly get the initial value when $t = 0$ but several discrete sampling points. There are four types of sampling time selection in this test, defined as *dtimes*. The four selections are (a). dtimes = [10 30 60 120 300 600]; (b). dtimes = [500 1000 1500 2000 2500 3000]; (c). dtimes = [10 30 60 120 300 600 900 1200 1500]; (d). dtimes = [10 30 60 120 300 600 900 1200 1500 2000 3000], respectively. It should be noticed that steady state will be achieved at around 900 min. Then the four selections focus on different preference for transient response and steady state.

The results in Figure 5 suggest that emphasis only on the transient response (selection (a)) or steady state (selection (b)) will achieve higher estimate accuracy rather than to cover a wide range of sampling selections (selection (c) and selection (d)). We notice that selection (b) need five times observe duration of selection (a) while not increasing considerable estimate accuracy. We therefore adopt selection (a) as default in all the tests throughout the paper.

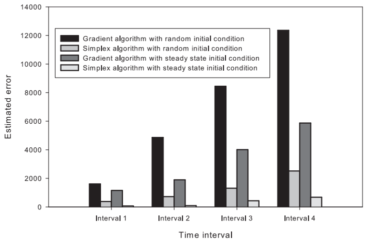

Figure 5: Estimated error under different time intervals. The measurement number ranges from 2 to 9. The estimated errors are showed considering type of initial condition and selection of algorithms simultaneously. Interval 1-4 in the figure represent (a). dtimes = [10 30 60 120 300 600]; (b). dtimes = [500 1000 1500 2000 2500 3000]; (c). dtimes = [10 30 60 120 300 600 900 1200 1500]; (d). dtimes = [10 30 60 120 300 600 900 1200 1500 2000 3000], respectively.

3 Conclusions

In this paper, we estimated maximal rate of enzyme $V_{\max}$ in photosynthetic carbon metabolism model. We tested the relationship between the estimate accuracy and several measurement protocols, including initial condition number, measurement number, noise intensity and time intervals, in order to come up with an optimal selection of raw data for parameter estimation. To guarantee the generality of the selection, two popular optimization algorithms were utilized for comparison. The result showed that higher estimation accuracy could be achieved under the following conditions: 1. random choose initial condition was not helpful and setting initial condition as the steady state; 2. measuring each time point at least 2 times; 3. trying to lower the noise perturbation to the data; 4. emphasising only on the transient response or steady state of the time course. The results corresponded to our hypothesis and a set of combination of setups were optimized.

4 Methods

4.1 Target Function

In our implementation of the target function, we considered standard deviation from both the time axis and the measurement axis. The standard deviations of the measurements were used to normalize the difference in magnitude between concentrations of different compounds. If the simulated value for ith compound is at jth time point is Y_{ijs}, and the mean and standard deviation of measurements at jth time point for ith compound is Y_{ijf} and σ_{ij}, then the contribution of measurements axis in the target function is expressed as:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{Y_{ijf}-Y_{ijs}}{\sigma_{ij}}\right)^2.$$

To incorporate the standard deviation of the time axis, assuming that we have measurement data for time t_m, we cans obtain the simulated value at different time points centered around t_m, say, five point earlier and five points later than t_m. The time points are labeled as: $t_{m-5}, t_{m-4}, \ldots, t_m, t_{m+1}, t_{m+1}, \ldots, t_{m+5}$. The Y value at each of the time point is correspondingly defined as: $Y_{m-5}, Y_{m-4}, \ldots, Y_m, Y_{m+1}, Y_{m+1}, \ldots, Y_{m+5}$. Calculate the absolute value of the difference between the simulated Y and the measurement data at the time (Y_d), and then take the smallest value of the difference. To construct the target function, first find the t corresponding to the smallest $|Y-Y_d|$ as t_{sd}; then calculate the probability density function of t_{sd} assuming that the average of the local t is t_m, the standard deviation is 0.1 following a *Gaussian* distribution. The value of the probability is represented as $PDF(t_{sd})$,

$$PDF(t_{sd})=\frac{1}{\sqrt{2\pi}}e^{-0.5u^2},$$

where in our case, we assume that σ is assumed to be 0.1 in this particular example. Take the integral of the probability *PDF* of all t with $|t-t_m|>|t_{sd}-t_m|$. The integral has a value of p. Use the $1/p$ directly in the target function. Therefore, the total target function is calculated as:

$$F=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{Y_{ijf}-Y_{ijs}}{\sigma_{ij}}\right)^2+\frac{1}{p}.$$

4.2 Optimization algorithms

Several optimization algorithms were applied in the search for the parameters. Due to the complicated nonlinearities of the network, the searching route and optimization mechanism for different algorithms, the performance and accuracy differed tremendously.

In this study, two methods were applied separately. Levenberg-Marquardt method is a gradient-based method for solving nonlinear least squares problems. This can be seen as Gauss–Newton with damping or as a combination of Gauss–Newton with steepest descent [11, 12]. Simplex algorithm is a direct search method that does not use numerical or analytic gradients [13]. It is based on the idea of an adaptive simplex: the simplest polytope of $n+1$ vertices in n dimensions.

Acknowledgements

The work is supported by the National Center for Supercomputing Applications and the National Science Foundation of China general program: 30970213

References

[1] Fell, D. (1997). Understanding the control of metabolism. London.

[2] Schilling, C., Edwards, J., Letscher, D., Palsson, B. (2000). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. Biotechnology and Bioengineering, 71(4), 286-306.

[3] Schilling, C., Schuster, S., Palsson, B., Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnology progress, 15(3), 296-303.

[4] Zhu, X. (2004). Computational approach to guiding biotechnological improvement of crop photosynthetic efficiency. PhD thesis, University of Illinois at Urbana-Champaign.

[5] Zhu, X., de Sturler, E., Long, S. (2007). Optimizing the Distribution of Resources between Enzymes of Carbon Metabolism can Dramatically Increase Photosynthetic Rate. A Numerical Simulation using an Evolutionary Algorithm. Plant physiology.

[6] Chen, C., Zhu, X., Long, S. (2008). The effect of leaf-level spatial variability in photosynthetic capacity on biochemical parameter estimates using the Farquhar model: a theoretical analysis. Plant physiology, 148(2), 1139.

[7] Zhu, X., Baker, N., de Sturler, E., Ort, D., Long S. (2005). Chlorophyll a fluorescence induction kinetics in leaves predicted from a model describing each discrete step of excitation energy and electron transfer associated with photosystem II. Planta, 223, 114-133.

[8] Moore, B., Cheng, S., Seemann, D. (1999). The biochemical and molecular basis for photosynthetic acclimation to elevated atmospheric CO_2. Plant, Cell & Environment, 22(6), 567-582.

[9] Weckwerth, W. (2003). Metabolomics in systems biology. Annual Review of Plant Biology, 54, 669-689.

[10] Crampin, E., McSharry, P., Schnell, S. (2004). Extracting biochemical reaction kinetics from time series data. In Knowledge-Based Intelligent Information and Engineering Systems, Springer, 329-336.

[11] Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. Quart. Appl. Math, 2(2), 164-168.

[12] Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics, 11(2), 431-441.

[13] Lagarias, J., Reeds J., Wright M., Wright P. (1999). Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM Journal on Optimization, 9, 112-147