2 Methodology of Model Development

Mathematical models have been developed to explain biological phenomena for centuries. Most classical interest areas modeled by continuous equations are related to population dynamics [3]. For instance, predator-prey models were proposed based on population balance to predict the species population in ecology. Similar ideas were further applied to describe the dynamic relationship between patients and carriers to estimate the outburst of infectious diseases in epidemiology. It was after the development of Michaelis-Menten equations that mathematical models begin to be widely applied to microscopic biological system. As a breakthrough in enzymology, Michaelis-Menten kinetic model precisely described enzyme-substrate interaction and successfully matched the enzymatic experimental data [4]. It is still commonly used today to describe enzymatic reactions and other biological reactions at a molecular level [2].

Molecular scale biological sciences involve studies of complex living systems with difficulty in measuring arrays of interacting species. The sizes of most biomolecules are smaller than the resolution of common microscopes. Electrophysiological procedure measuring the signal response of neuronal tissues requires multi-step treatment and instrumentation before experiment. Also quantification of common molecules depend on complicated assays usually involve multiple experimental steps such as cell culturing, purification, and staining. Thus, mathematical modeling provides valuable theoretical hypothesis to guide the experimental study and maximize the value of each experiment. Successful examples include theoretical models to explain molecular events such as signal transduction, neuronal spiking, and gene expression, etc. It is through systematic modeling that the hierarchical structure and subsystem interactions in biological systems can be unraveled and understood.

2.1 Recognize biological systems of interest

The methodology listed here is applicable to different kinds of biological systems. Yet the systems of interest in this thesis are related to memory pathology and memory formation mechanisms. The most serious causes of senile dementia and memory loss are Alzheimer's disease and Parkinson's disease both of which are found to be related to amyloid fibril formation.

2.2 Propose biochemical reaction mechanisms

Depending of the types of systems being studied, the biochemical reactions involved differ. Yet the chances are that for most biological systems having been studies

previously, some relevant reaction pathways can be recognized. By cross referencing literature, doing experiments, or drawing influence diagram, it is possible to propose the principal reaction mechanisms.

It is likely that the biological system of interests involve a number of subsystems. The boundaries of different subsystems and hierarchical structure need to be well defined. Ideally each of the subsystem should be individually analyzed to ensure the validity or the proposed model and the robustness of the overall system.

2.3 Convert reactions into differential equations

Material balance and reaction kinetics are the two principal guidelines in converting chemical reactions into differential equations. The variables involved can be classified into input variables, state variables, and response variables. The response variable can take the form of algebraic equation:

$$\bar{Y} = \bar{F}(\bar{X}, \bar{\theta}) \tag{1.1}$$

It can also be expressed as ordinary differential equations:

$$\frac{d\bar{X}}{dt} = \bar{f}(\bar{X}, \bar{\theta}) \tag{1.2}$$

The number of sate variables has to be equal to the number of independent differential equations. In addition, the initial conditions of all the state variables need to be specified.

$$@t = 0 \qquad \bar{X} = \bar{X}_{initial} \tag{1.3}$$

2.4 Least square parameter estimation

There are usually parameters embedded in the model such as reaction rate constants, physical or chemical properties, etc. Parameter estimation is therefore necessary to determine the values of those parameters. Parameter estimation is more straightforward for linear models [5] but require iteration processes for nonlinear ones [6].

2.4.1 Linear least square regression (LLSR)

Least squared errors are usually adopted for estimating parameters. Other methods include maximum likelihood estimation, Bayesian estimator, etc. For linear models, the minimization of sum of squared errors only involves solving sets of linear equations.

$$\bar{Y} = X\bar{\beta} + \bar{\varepsilon} \tag{1.4}$$

Or the expanded version assuming there are n experimental data points and p parameters:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1p} \\ 1 & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{1.5}$$

The prerequisite conditions for using linear least square regression are called Gauss-Markov assumptions which are listed below.

- The random error terms have expected value of zero
- The random error terms are uncorrelated
- The random error terms have the same variance

That is, the distribution of error can be written as: $\varepsilon \sim N_n(0, \sigma^2 I_n)$

One of the tasks included in linear least square regression is to find the best estimate of parameter vector $\hat{\beta}$. The geometric analysis shown in Fig helps visualize the algorithms while algebraic argument can also be done to reach the same results. The objective is to minimize the sum of squared errors ($\varepsilon^T\varepsilon$) with respect to vector β. The projection of vector β onto Ω, the column space of matrix X, is called θ. That is, $\theta = X\beta$. Now the sum of squared error becomes the square of the length of error vector Y- θ since $\varepsilon^T\varepsilon = (Y-\theta)^T(Y-\theta) = \| Y-\theta \|^2$. If we let θ vary in Ω, the optimal $\hat{\theta}$ would be the one that makes (Y- θ) perpendicular to surface Ω, according to Fig. Therefore $X^T(Y-\hat{\theta}) = 0$ which is the same as the following equation:

$$X^T\hat{\theta} = X^TY \tag{1.6}$$

Once the optimal vector $\hat{\theta}$ on space Ω is found, it can be converted back to vector $\hat{\beta}$ given the columns of matrix X are linearly independent. This becomes the normal equations of linear least square regression.

$$X^TX\hat{\beta} = X^TY \tag{1.7}$$

The best estimate of parameter vector $\hat{\beta}$ can now be found by simply multiplying the reverse matrix of (X^TX) to both sides of the normal equations.

$$\hat{\beta} = \left(X^TX\right)^{-1}X^TY \tag{1.8}$$

Then the error terms can be calculated by taking the difference between experimental data and predicted values:

$$\varepsilon = Y - \hat{Y} = Y - X\hat{\beta} \tag{1.9}$$

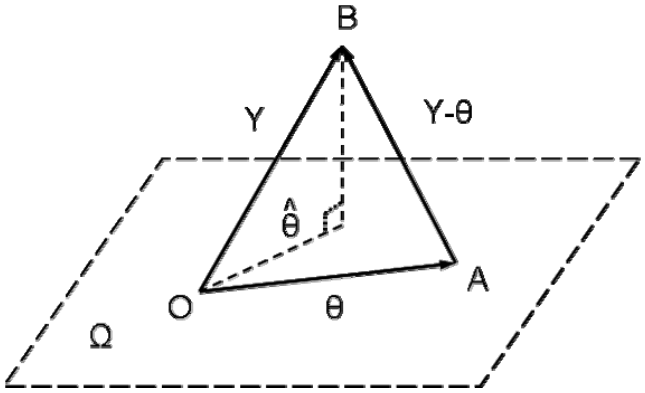

Figure 2-1 The geometrical scheme used to find the optimal vector that minimizes the sum of squared errors (SSE).

Under Gauss-Markov assumption, the measurement vector Y follows the n×n normal distribution: $Y \sim N_n(X\beta, \sigma^2 I_n)$. As a result, the distribution of estimated parameter vector $\hat{\beta}$ follows $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$ and normalized sum of squared errors follows χ^2 distribution: $SSE/\sigma^2 \sim \chi^2_{n-p}$ where S^2 is the unbiased estimator of σ^2:

$$S^2 = \frac{SSE}{n-p} = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n-p} \tag{1.10}$$

Finally, the 100(1- α)% confidence interval of parameter β_i can be computed using the following equation where *t* stands for *t*-distribution with *n-p* degrees of freedom at the α level of significance. The proofs in details can be found in Seber and Lee, 2003 [5].

$$\hat{\beta}_i \pm t_{\alpha/2, n-p} S \sqrt{(X^T X)^{-1}_{ii}} \tag{1.11}$$

2.4.2 Nonlinear least square regression (NLSR)

The least square regression for nonlinear model does not have straightforward formula just as linear regression does. Instead solving NLSR problems requires optimization algorithms. The general form of nonlinear model can be expressed as the following:

$$y_i = f(x_i; \overline{\theta}) + \varepsilon_i \quad (i = 1, 2, ..., n) \tag{1.12}$$

The sum of squared errors is the sum of the squared differences between experimental data and the model.

$$S(\bar{\theta}) = \sum_{i=1}^{n} \left[y_i - f(x_i; \bar{\theta}) \right]^2 \tag{1.13}$$

In order to find the minimum of the sum of squared errors, we need to differentiate $f(x_i; \bar{\theta})$ with each parameter θ_r for r from 1 to p.

$$\left. \frac{\partial S(\theta)}{\partial \theta_r} \right|_{\hat{\theta}} = 0 \quad (r = 1, 2, ..., p) \tag{1.14}$$

If we denote $f_i(\bar{\theta}) = f(x_i; \bar{\theta})$, we can use the define F(θ) as the design matrix for nonlinear model:

$$F(\bar{\theta}) = \left[\left(\frac{\partial f_i(\bar{\theta})}{\partial \theta_j} \right) \right] \tag{1.15}$$

Then Eq. (1.14) can be rewritten as:

$$\sum_i \left\{ y_i - f_i(\bar{\theta}) \right\} \left. \frac{\partial f_i(\bar{\theta})}{\partial \theta_r} \right|_{\theta = \hat{\theta}} = 0 \tag{1.16}$$

or

$$0 = \hat{F}^T \left\{ y - f(\hat{\theta}) \right\} = \hat{F}^T \hat{\varepsilon} \tag{1.17}$$

The geometrical meaning of Eqs. (1.17) is that the optimal error vector should be perpendicular to the column space of design matrix F. Therefore, they are called normal equations for nonlinear model. The subsequent algorithms for parameter estimation and model prediction become similar to those of linear least square regression.

2.5 Summary chart of model development method

The standard procedure of model development is illustrated in
Figure 2-2. Once the system of interest has been identified, the next step in divided into two pathways: the performance of experiments and development of theoretical models. Eventually the model outputs the results of goodness-of-fit, parameter estimation, and model prediction. Yet experiments and models should not be independent of one another. Instead, it is through the interaction between these two that the mechanisms underlying specific biological phenomena can be unraveled efficiently. For instance, model-based experimental design potentially lowers the number of experiments and increases the

accuracy of measurements [1]. That is, more information can be extracted from one single set of experiment.

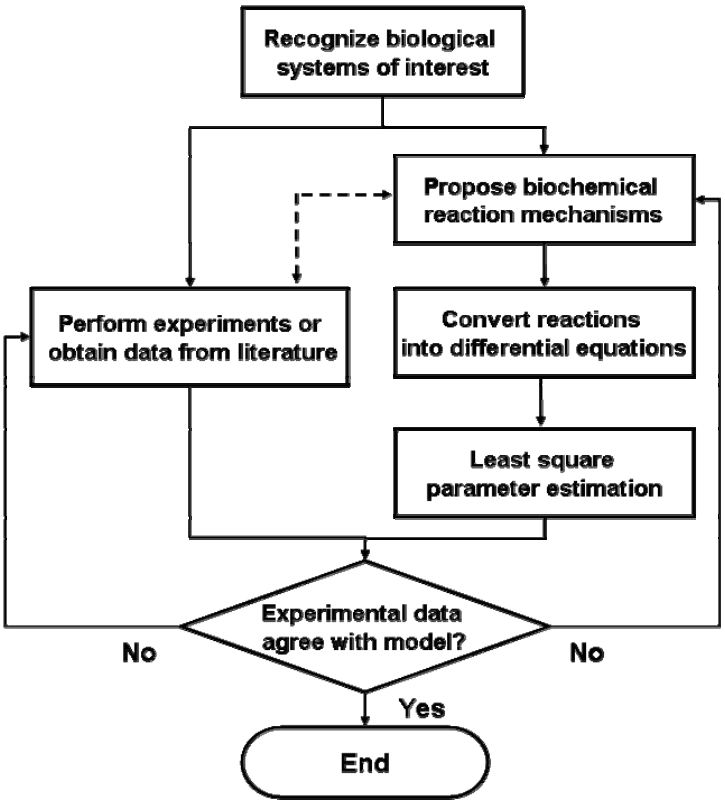

Figure 2-2 The summary chart of model development process. The model is developed together with execution of experiments.

Lastly, one essential statistical concept is worth clarifying. The agreement between the experimental data and model is the necessary but not sufficient condition to confirm the validity of the model. That is, the satisfactory goodness-of-fit between one set of data and the model means the specific data cannot reject the proposed model. Yet it does not exclude the possibility that alternative models can explain the same set of data equally well or the possibility that other sets of data can refute the model. Therefore, cautions need to be taken when asserting the validity of the model. That is, models that have only been tested against certain sets of data are not universally true; instead, each model usually has its own limitations.

1. Berkholz R, Guthke R (2001) Model based sequential experimental design for bioprocess optimisation—an overview. Springer, Brussels, Belgium
2. Bhalla US, Iyengar R (1995) Emergent properties of networks of biological signaling pathways. Science 2:381-387
3. Edelstein-Keshet L (2005) Mathematical Models in Biology. Society for Industrial and Applied Mathematics, Philadelphia, PA

4. Michaelis M, Menten ML (1913) Kinetics of invertase action. Z Biochem 49:333-369
5. Seber GAF, Lee AJ (2003) Linear Regression Analysis. American Statistical Association, Hoboken, NJ
6. Seber GAF, Wild CJ (2003) Nonlinear Regression. Wiley, Hoboken, NJ