An Algebraic-Numeric Algorithm for the Model Selection in Kinetic Networks

Hiroshi Yoshida¹,*, Koji Nakagawa², Hirokazu Anai³, and Katsuhisa Horimoto²,*

¹ Faculty of Mathematics, Organization for the Promotion of Advanced Research, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581 Japan
Tel/Fax.: +81-92-642-7396
phiroshi@math.kyushu-u.ac.jp

² Computational Biology Research Centre (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan
Tel.: +81-3-3599-8711; Fax.: +81-3-3599-8081
{nakagawa-koji,k.horimoto}@aist.go.jp

³ IT Core Laboratories, FUJITSU LABORATORIES LTD./CREST, JST., Kamikodanaka 4-1-1, Nakahara-ku, Kawasaki 211-8588, Japan
anai.jp.fujitsu.com

Abstract. We propose a novel algorithm to select a model that is consistent with the time series of observed data. In the first step, the kinetics for describing a biological phenomenon is expressed by a system of differential equations, assuming that the relationships between the variables are linear. Simultaneously, the time series of the data are numerically fitted as a series of exponentials. In the next step, both the system of differential equations with the kinetic parameters and the series of exponentials fitted to the observed data are transformed into the corresponding system of algebraic equations, by the Laplace transformation. Finally, the two systems of algebraic equations are compared by an algebraic approach. The present method estimates the model’s consistency with the observed data and the determined kinetic parameters. One of the merits of the present method is that it allows a kinetic model with cyclic relationships between variables that cannot be handled by the usual approaches. The plausibility of the present method is illustrated by the actual relationships between specific leaf area, leaf nitrogen and leaf gas exchange with the corresponding simulated data.

1 Introduction

The knowledge-based approach to constructing a biological network model is recognized as one of the most promising approaches [4]. In this approach, the causal relations between biological molecules are described as a directed graph, based on the gene interaction information collected from a large number of previous reports. Since each relation identified by experimental studies is regarded as strong evidence for the existence of edges in the network model, biological network models have been constructed for various biological phenomena by a knowledge-based approach. On the other hand, it is well-known that the relationships between the molecules in a living cell change

* Corresponding authors.
dynamically, depending on the cellular environment. Thus, the molecular relationships in the literature represent the responses to the different conditions in the experimental studies, and in the network model generated from the biological knowledge, the consistency of the model with the data observed by experimental studies must be considered carefully. Actually, several distinctive models of the relationship between molecules for a biological phenomenon can be obtained from the large amount of information in the literature [2, 5]. In these cases, a model that is consistent with the data observed under particular conditions should be selected from the candidate models.

The consistency of a model with the observed data first reminds us of the identifiability problem in the compartmental models for tracer kinetics [1, 5, 6]. In the compartmental models, the unknown parameters are estimated from tracer data in the accessible pools. The identifiability problem addresses the issue of whether the unknown parameters can be determined uniquely or non-uniquely from the tracer data. This issue has usually been solved through the transformation of differential equations into algebraic equations, by the Laplace transformation. Although a systematic algorithm for the identifiability problem was proposed [3], its application is limited to the unrealistic context of an error–free model structure and noise–free tracer data. Thus, it still seems to be difficult to solve the identifiability problem for actually observed data, in spite of the mathematical studies.

The issue of the consistency of a model with the observed data is also well known in statistics, as the test for causal hypotheses by using the observed data. The origin of the test for causal hypotheses is attributed to path analysis [12]. Unfortunately, the importance of this cornerstone research has been ignored for a long time, but the natural extension of the path analysis has been established as the well-known structural equation model (SEM) [8]. Indeed, the SEM has been utilized recently in various fields, in accordance with increased computer performance. However, the SEM without any latent variables, which is the natural form for applying the SEM to the biological networks, frequently faces difficulty in the numerical calculation of the maximum likelihood for the observed data. To overcome the difficulty of this calculation, the d-sep test [11] has been developed, based on the concept of d-separation in a directed acyclic graph [10]. Notice that the graph consistency with the data in the d-sep test can consider only the directed acyclic graph (DAG), without any cyclic relationships.

In this study, we propose a new method for selecting models, by estimating the consistency of a kinetic model with the time series of observed data. Our method is described in the following section. First, the kinetics for describing a biological phenomenon is expressed by a system of differential equations, assumed that the relationships between the variables are linear. Simultaneously, the time series of the data are numerically fitted as a series of exponentials. Next, the differential equations with the kinetic parameters and the series of exponentials fitted to the observed data are both transformed into the corresponding system of algebraic equations, by the Laplace transformation. Finally, the two systems of algebraic equations are compared by an algebraic approach. Thus, the present method estimates the model’s consistency with the observed data and the determined kinetic parameters. In §3, the plausibility of the present method is illustrated by the actual relationships between specific leaf area, leaf