Modeling and Estimation of Dynamic EGFR Pathway by Data Assimilation Approach Using Time Series Proteomic Data

Shinya Tasaki1,3*, Masao Nagasaki2*, Masaaki Oyama1
stasaki@ims.u-tokyo.ac.jp masao@ims.u-tokyo.ac.jp moyama@ims.u-tokyo.ac.jp

Hiroko Hata1 hata@ims.u-tokyo.ac.jp
Kazuko Ueno2 uepi@ims.u-tokyo.ac.jp
Ryo Yoshida2 yoshidar@ims.u-tokyo.ac.jp

Tomoyuki Higuchi4 higuchi@ism.ac.jp
Sumio Sugano3 ssugano@ims.u-tokyo.ac.jp
Satoru Miyano2 miyano@ims.u-tokyo.ac.jp

1 Medical Proteomics Laboratory, Institute of Medical Science, the University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
2 Human Genome Center, Institute of Medical Science, the University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
3 Department of Medical Genome Science, Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-8562, Japan
4 Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan
* These authors contributed equally.

Abstract

Cell Illustrator is a model building tool based on the Hybrid Functional Petri net with extension (HFPNe). By using Cell Illustrator, we have succeeded in modeling biological pathways, e.g., metabolic pathways, gene regulatory networks, microRNA regulatory networks, cell signaling networks, and cell-cell interactions. The recent development of tandem mass spectrometry coupled with liquid chromatography (LC/MS/MS) technology has enabled researchers to quantify the dynamic profile of a wide range of proteins within the cell. The proteomic data obtained by using LC/MS/MS has been considerably useful for introducing dynamics to the HFPNe model. Here, we report the first introduction of the time-series proteomic data to our HFPNe model. We constructed an epidermal growth factor receptor signal transduction pathway model (EGFR model) by using the biological data available in the literature. Then, the kinetic parameters were determined in the data assimilation (DA) framework with some manual tuning so as to fit the proteomic data published by Blagoev et al. (Nat. Biotechnol., 22:1139–1145, 2004). This in silico model was further refined by adding or removing some regulation loops using biological background knowledge. The DA framework was used to select the most plausible model from among the refined models. By using the proteomic data, we semi-automatically constructed a well-tuned EGFR HFPNe model by using the Cell Illustrator coupled with the DA framework.

Keywords: EGFR pathway model, proteome data, HFPNe, Petri net, data assimilation, model selection, parameter estimation

1 Introduction

The epidermal growth factor receptor (EGFR) pathway is one of the best-studied signal transduction systems, and regulates important cellular processes, e.g., DNA synthesis and cell proliferation, migration, and adhesion. In many cases, a defective EGFR pathway is known to be a cause of cancer. Recently, the comprehensive EGFR pathway map was constructed based on the information obtained
from the literature [28]. This map revealed the entire structure of the EGFR pathway network with its high complexity.

The EGFR pathway has been well analyzed by using in silico modeling [37, 38]. The in vivo dynamics of several proteins introduced in these models were obtained from results of biochemical experiments, e.g., western blotting. These protein data were used for parameter estimation to refine the in silico model. Both data acquisition and parameter fitting are the basic steps for creating an in silico simulation model. Thus, improvements of these processes are crucial for achieving precise in silico simulation. One of the major challenges is the acquisition of in vivo data on signal transduction pathways. During intracellular signal transduction, a large number of molecules are regulated by multiple mechanisms, e.g., binding, phosphorylation, dephosphorylation, localization, degradation, and Ca signaling. Thus, more comprehensive measurement methods are required for obtaining global profiles of cellular signaling. Another challenge is the establishment of a link between the in silico model and the in vivo data. As the amount of in vivo data increases, it becomes more difficult to estimate the kinetic parameters of in silico models that represent the in vivo dynamics. Studies aimed at addressing these problems will be challenging because both experimental and computational problems act as bottlenecks in creating a superior in silico EGFR model. In recent years, proteome analysis with tandem mass spectrometry coupled with liquid chromatography (LC/MS/MS) technology has improved greatly [16, 31, 46]. LC/MS/MS can identify a large number of proteins and peptides from a highly complex sample. Moreover, the establishment of stable isotope labeling methods has made it possible to quantitatively measure proteins and peptides in samples [29, 30]. The quantitative results from LC/MS/MS experiments are very useful for creating more complex and precise EGFR models. Since 1999, we have been developing a software tool called Cell Illustrator [24, 25, 26, 47, 49, 51]. For this, we defined a modeling architecture known as the Hybrid Functional Petri net (HFPN) and its extension (HFPNe). By using this software, we have been analyzing models of pathways such as metabolic pathways, signaling pathways, gene regulatory networks, and microRNA regulatory networks, [8, 9, 20, 21, 22, 35, 41, 44, 47, 49, 51]. Kinetic parameters in these models were tuned manually based on biological knowledge available in the literature. However, with the increase in knowledge about the elements involved in a biological pathway, it becomes extremely difficult to manually introduce these parameters into the models. To overcome this difficulty, Geoffey et al. [17] proposed separating the MAPK pathway model with HFPN into six independent sub-pathways and estimated their parameters by using a genetic algorithm. However, it must be noted that their method can be applied only if all reactions conform to Michaelis-Menten kinetics. As an alternate computational approach, we have developed data assimilation (DA) framework that can automatically estimate these parameters by using in vivo data. In our previous study, we demonstrated that parameter estimation and model selections based on the DA function very well when coupled with the HFPN model of the mammalian circadian clock [27]. However, since this circadian clock model is an extremely simple model that consists of only 12 entities and 28 processes, it would be a major challenge to apply the DA method effectively to more complex systems. In this study, we have compiled the information available in the literature and modeled the EGFR pathway. By using the extensive time-series quantitative data on proteins reported by Blagoev et al. [2], we semi-automatically performed the estimation of the fine-tuned parameters based on the DA framework. This method also allowed us to select the best-estimate model that matched the in vivo data from among several hypothesized EGFR models. Our model is the first HFPNe model that is compatible with time-series proteomic data measured by LC/MS/MS, and our study indicates that the DA method is applicable to relatively complex systems.

This paper is organized as follows. In Section 2.2, we present the EGFR pathway model for the transduction of epidermal growth factor (EGF) signals; this model has been constructed based on biological data available in the literature. In Section 2.3, as in vivo data, we use the time-series data of tyrosine-phosphorylated proteins in EGF-stimulated HeLa cells with LC/MS/MS technology as reported by Blagoev et al. [2]. In Section 2.4, we demonstrate how to combine the above mentioned in
Modeling and Estimation of Dynamic EGFR Pathway by DA

in vivo data and the parameter estimation method using the DA. The components of the EGFR model are larger than those of the previous circadian model presented in the study by Nagasaki et al. [27]. To overcome the problems caused by this difference, we divide the EGFR model into three sub-pathways and combine these sub-pathways after applying parameter estimations based on the DA framework. The procedures used for division and merging are also described in Section 2.4. In Section 3.1, we compare the in silico model and the in vivo experimental results and discuss its performance. One of the merits of the DA method is that the best model among various hypothesized models can be selected. In Section 3.2, we present the developed in silico models in which hypothesized regulations are added or removed from the original model. The best model is selected from among these models using the DA method and its performance is discussed.

2.1 Hybrid Functional Petri Net with Extension

Petri net is a network that consists of the following elements: place, transition, arc, and token. In this paper, we use entity, process, and connector for intuitive notations. An entity (depicted as a circle) can hold tokens as its content. A process (depicted as a rectangle) has connectors originating from some entities and connectors exiting the process to enter some entities. A process connected with these connectors defines a firing rule in terms of the contents of the entities to/from which the connectors are attached. When modeling with the Petri net, an entity represents the amount/density of a biological molecule/object, and a process defines the speed/condition/mechanism of interaction/reaction/transfer among the entities connected by the connectors. The conventional Petri net can be used to model only the discrete features in biological pathways, e.g., logical regulatory relationships between genes. To overcome this limitation, we have defined the concept of Hybrid Functional Petri net with extension (HFPNe) [24]. When the notion of object corresponds to the Java class and method, a more detailed biological pathway modeling can be realized using systematic methods, e.g., alternative splicing, ribosomal frameshifting, subcellular localization information [24]. HFPNe has three types of entities - discrete, continuous, and generic - and three types of processes - discrete, continuous, and generic. The concepts of discrete entity and discrete process are identical to those in the traditional discrete Petri net. A continuous entity can hold a real number as its content. A continuous process fires continuously at the speed of the parameter assigned to it. A generic entity can hold any object, e.g., an mRNA sequence, or the phosphorylation state of a protein; moreover, a generic process handles complex reactions by updating the state of connected entities, e.g., degradation, translation, and phosphorylation, for complex pathway modeling. Three types of connectors are used in HFPNe, and a specific value is assigned to each connector as a threshold script. When a process connector (depicted as a solid line connector in Figure 1) with a threshold script is attached to a discrete/continuous/generic process (depicted as a single circle/double circle/a single circle with a cross, respectively), a certain number of tokens are transferred throughout the process connector only if the evaluated result of the threshold script is true. The activity rule of an association connector is identical to that of a process connector in terms of the threshold, but the content of the entity at the source of the association connector is not changed by activation. An association connector (depicted as a dashed line connector in Figure 1) can be used to represent enzyme activity since the enzyme itself is not consumed in a reaction. An inhibitory connector (depicted as a line terminating with a small bar in Figure 1) with a threshold script enables the process to remain active only if the evaluated result of the threshold script is false. For example, an inhibitory connector can be used to represent repressive activity in gene regulation. The formal definition is available in the published study of Nagasaki et al. [24].

2.2 In silico EGFR Model Based on Information in Published Literature

Figure 2 shows an HFPNe model that has been constructed using the compiled and interpreted information on the EGFR signal transduction pathways available in the literature [4, 5, 6, 7, 10, 11,
Figure 1: Essentially, Petri nets are constructed using three types of symbols for entities, processes, and connectors. In the Cell Illustrator, both sets of entities and processes are classified into discrete, continuous, and generic types, and entities and processes can be replaced with pictures reflecting the biological images. This replacement makes the HFPNe model of a biological pathway more comprehensible for biologists.

Figure 2: HFPNe model of EGFR pathway. For entities and processes, pictures reflecting biological images are used (see Figure 1). Biological meanings of transitions $T_1$–$T_{31}$ are summarized in Table 1.
Table 1: Biological facts obtained from the literature and assigned to processes in the HFPNe model in Figure 2. #1: Corresponding processes in the HFPNe. #2: Corresponding sub-pathways in Figure 2.

<table>
<thead>
<tr>
<th>Biological phenomena from experimental data in literature</th>
<th>#1</th>
<th>#2</th>
<th>Process type</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding of EGF to EGFR induces the dimerization of the receptors resulting in autophosphorylation of the receptors.</td>
<td>T1</td>
<td>i</td>
<td>Association</td>
<td>[28]</td>
</tr>
<tr>
<td>c-Cbl binds to the tyrosine-phosphorylated EGFR and simultaneously c-Cbl is tyrosine phosphorylated.</td>
<td>T4</td>
<td>i</td>
<td>Association</td>
<td>[42]</td>
</tr>
<tr>
<td>The ubiquitinated EGFR is degraded by the proteasome/lysosome.</td>
<td>T7</td>
<td>i</td>
<td>Degradation</td>
<td></td>
</tr>
<tr>
<td>Vav2 binds to tyrosine-phosphorylated EGFR via its SH2 domain and is tyrosine phosphorylated by EGFR.</td>
<td>T8</td>
<td>ii</td>
<td>Association</td>
<td>[23, 32, 41]</td>
</tr>
<tr>
<td>Vav2 activates Rac1 by promoting the exchange of bound GDP for GTP.</td>
<td>T9</td>
<td>ii</td>
<td>Phosphorylation</td>
<td>[7]</td>
</tr>
<tr>
<td>Activated MEKK phosphorylates MKK3/4/6/7. Phosphorylated MKK3/4/6/7 protein p38MAPK. We modeled MKK3/4/6/7 as one protein for simplification.</td>
<td>T12</td>
<td>ii</td>
<td>Phosphorylation</td>
<td>[10, 11, 18, 42]</td>
</tr>
<tr>
<td>Grb2 associates with tyrosine-phosphorylated EGFR.</td>
<td>T14</td>
<td>iii</td>
<td>Association</td>
<td>[18, 19]</td>
</tr>
<tr>
<td>Shc binds to the tyrosine-phosphorylated EGFR.</td>
<td>T15</td>
<td>iii</td>
<td>Association</td>
<td>[34, 36]</td>
</tr>
<tr>
<td>Shc is tyrosine phosphorylated and interacts with Grb2</td>
<td>T16</td>
<td>iii</td>
<td>Phosphorylation</td>
<td>[39, 40, 45]</td>
</tr>
<tr>
<td>Sos1 binds to Grb2.</td>
<td>T17</td>
<td>iii</td>
<td>Association</td>
<td>[4]</td>
</tr>
<tr>
<td>Complex of Sos1 associated with EGFR catalyzes Ras GTP/GDP exchange.</td>
<td>T19</td>
<td>iii</td>
<td>Exchange</td>
<td>[20]</td>
</tr>
<tr>
<td>Ras/GTP recruits Raf to the plasma membrane and promotes Raf activation process.</td>
<td>T20</td>
<td>iii</td>
<td>Association</td>
<td>[5]</td>
</tr>
<tr>
<td>Activated Raf activates both MKK1 and MKK2.</td>
<td>T21</td>
<td>iii</td>
<td>Phosphorylation</td>
<td>[5]</td>
</tr>
<tr>
<td>Activated MKK1 and MKK2 phosphorylate Erk2.</td>
<td>T23</td>
<td>iii</td>
<td>Phosphorylation</td>
<td>[39]</td>
</tr>
<tr>
<td>Sos1 is phosphorylated by phosphorylated Erk2, resulting in decreased binding affinity to Grb2.</td>
<td>T25</td>
<td>iii</td>
<td>Phosphorylation</td>
<td>[6]</td>
</tr>
<tr>
<td>Dok binds to phosphorylated EGFR and is phosphorylated.</td>
<td>T27</td>
<td>iii</td>
<td>Association</td>
<td>[1, 15]</td>
</tr>
<tr>
<td>RasGAP binds to phosphorylated Dok and upregulates the RasGTP dephosphorylation.</td>
<td>T28</td>
<td>iii</td>
<td>Phosphorylation</td>
<td></td>
</tr>
<tr>
<td>Activated Rac/Cdc42 induces phosphorylation of MKK1 and MKK2 indirectly.</td>
<td>T30</td>
<td>iii</td>
<td>Phosphorylation</td>
<td>[10, 11, 18, 42]</td>
</tr>
</tbody>
</table>

12, 18, 19, 20, 23, 32, 34, 36, 39, 40, 41, 42, 43, 45]. We have focused on the pathways that are mediated by tyrosine-phosphorylation with EGF stimulation. Since, tyrosine-phosphorylation is the major mechanism of EGFR signal transduction and the proteomic quantification methods for tyrosine-phosphorylated proteins have been established (see Section 2.3). Our EGFR model consists of three major sub-pathways: (i) EGFR activation and degradation pathway, (ii) p38 MAPK activation pathway, and (iii) ERK2 activation pathway. The pathway of p38MAPK activation affects the activation of Erk2 but has been poorly-studied in other EGFR simulation models [37, 38]. Therefore, we have
modeled sub-pathway (ii) as well as other sub-pathways. Sub-pathway (i) contains processes \( T1 \)–\( T7 \) as shown in Table1. After the phosphorylation of EGFR by EGF stimulation, c-Cbl binds to the phosphorylated EGFR and is phosphorylated. EGFR is ubiquitinated by c-Cbl and then degraded [\( T1, T2, T3, T4, T5, T6, \) and \( T7 \)]. Sub-pathway (ii) contains processes \( T8 \)–\( T13 \). After the phosphorylation of EGFR, Vav2 also binds to the phosphorylated EGFR and activates Rac1/Cdc42 [\( T8, T9, \) and \( T10 \)]. Rac/Cdc42 activates the MAPK cascade, MEKKs, and p38 MAPK [\( T11, T12, \) and \( T13 \)]. Sub-pathway (iii) contains processes \( T14 \)–\( T31 \). After the phosphorylation of EGFR, related proteins - Shc, Grb2, and Sos1 - bind to this receptor and change their state, i.e., into the phosphorylated form [\( T14, T15, T16, T17, \) and \( T18 \)]. Downstream of Sos1, Ras activates the MAPK cascade, Raf, Mkk1/2, and Erk2 [\( T19, T20, T21, T22, T23, \) and \( T24 \)]. The activated Rac/Cdc42 also induces phosphorylation of Erk2 [\( T30 \) and \( T31 \)]. Activated Erk2 then inactivates Sos1 by phosphorylation [\( T25 \) and \( T26 \)]. After the phosphorylation of EGFR, Dok is also recruited to the phosphorylated EGFR and is activated. The activated Dok reduces the Ras activity via dephosphorylation [\( T27, T28, \) and \( T29 \)]. In the EGFR pathway, activated EGFR binds to many components (Shc, Grb2, Dok, Vav, and c-Cbl in this model). Thus, if all possible interactions among the components are presented in the model, the pathway model needs many processes and will become too complicated. To overcome this difficulty, we have applied three hypotheses to simplify the EGFR model without compromising the necessary information. First, each adaptor protein, with the exception of c-Cbl, competitively binds to EGFR. Second, the phosphorylated EGFR is ubiquitinated by c-Cbl and degraded in the proteasome, regardless of the EGFR binding states. Third, the activity of EGFR is not affected by its ubiquitinated state. In this model, phosphorylated EGFR entities are merged to form one entity whether they are ubiquitinated or not. The ubiquitinated EGFR ratio and the degradation speed of the merged entity are calculated independently by using a special continuous entity that sums the total amount of phosphorylated EGFR using a special generic process. This simplification resulted in the reduction of the total elements in our model.

2.3 \textit{In vivo} Data: Proteome Data Obtained by LC/MS/MS using the SILAC method

Several modified LC/MS/MS approaches have been developed for quantitative proteomics using protein labeling methods, e.g., stable isotope labeling with amino acids in cell culture (SILAC), isotope-coded affinity tags (ICAT) [13], isobaric tags for relative and absolute quantitation (iTRAQ) [33] and CDIT (culture-derived isotope tags) [14]. These approaches allow researchers studying signal transduction pathways in a cell to measure the global time-dependent profiles of tyrosine-phosphorylated proteins [2, 48]. From these methods, we used the data from the time-series data based on the SILAC method used by Blagoev et al. [2] as an input for parameter estimations and model selections with DA. In general, antibody-based methods have been used for the quantification of protein amount. Compared to these conventional methods, LC/MS/MS-based quantification methods have various merits with regard to their comprehensiveness and data accuracy. First, LC/MS/MS-based methods can quantify a large number of proteins including novel proteins during a single measurement. In contrast, in antibody-based methods, the number of measurable proteins depends on the number of antibodies. Since the optimal conditions for each antibody are different, several experiments under different conditions are essential. Second, the quantitative data obtained using LC/MS/MS are less biased than those obtained using antibody -based techniques. The quality of the data largely depends on the antibody’s affinity and specificity for its antigen. With regard to the western blotting technique, it is difficult to eliminate researchers’ subjectivity to determine the band range for quantification. On the other hand, quantification by LC/MS/MS is based on the common properties of protein; hence, its data is more reliable.

The basis of the SILAC method is the labeling of all cellular proteins by growing cells in media containing stable isotope-containing amino acids [29, 30]. The SILAC method is a better protein
Figure 3: Summary of the method used for calculating the relative values of phosphorylated proteins with HFPNe. The lower box shows an example pathway model that consists of seven entities $[e_1$–$e_7]$ and four processes $[T_1$–$T_4]$. With the $e_1$ stimulation, $e_2$ is phosphorylated. The phosphorylated $e_2$ binds to $e_4$. Using the experimental method described in Section 2.3, the amounts of $e_3$ and $e_5$ are measured relative to the unstimulated state. To calculate the corresponding value with HFPNe, we introduce two continuous entities $[e_6, e_7]$ and two generic processes $[T_3, T_4]$. The values of $e_3$ and $e_5$ at 0 pt are stored in $e_6$ by the $T_3$ process that fires only at the stimulated time point. The $T_4$ process continuously calculates the values of $e_3$, $e_5$, and $e_6$ entities at every time point and stores the ratio of the sum of $e_3$ and $e_5$ to $e_6$ into $e_7$.

labeling method for LC/MS/MS measurement because of its high labeling efficiency and specificity as compared to other methods that are based on chemical incorporation. The actual scheme of measuring the relative quantitative data of tyrosine-phosphorylated proteins by using the SILAC method is as follows. Cells are grown in a medium containing only one form of an amino acid until it is completely incorporated into all the cellular proteins. Each cultured cell is stimulated by EGF for the indicated time interval and lysed. The lysates are mixed and the tyrosine-phosphorylated proteins are affinity purified using an anti-phosphotyrosine antibody. The purified tyrosine-phosphorylated proteins are fragmented by a protease, e.g., trypsin, followed by LC/MS/MS analysis. By using this approach, the time-series data of 81 tyrosine-phosphorylated proteins related to EGFR signaling was measured by Blagoev et al. [2].

Hence, the quantitative data obtained by LC/MS/MS using the SILAC method appears promising. Thus, in this paper, unlike the procedures followed in other studies [37, 38], the results of the SILAC method are used.

2.4 How to Combine the in vivo Data and the in silico EGFR Model

In our previous study, we performed parameter estimations of the mammalian circadian clock model by using the simulation data with the absolute quantitative data on mRNAs and proteins [27]. On the other hand, the time-series data of tyrosine-phosphorylated proteins in the EGFR pathway measured by the SILAC method is the relative quantitative data. Thus, in this simulation model, we needed to calculate the relative quantitative value for all entities, each of which represents the quantity of samples in vivo. In addition, each value of the in vivo data denotes the relative ratio of tyrosine-phosphorylated protein among samples. This implies that the same tyrosine-phosphorylated protein having different states, e.g., single or complex, is mixed into one value. Since it is not possible to directly compare an entity of the in silico pathway model and a value of the in vivo data for applying the parameter estimation framework, we had to introduce two types of special continuous entities ($e_6$ and $e_7$) with generic processes ($T_3$ and $T_4$), as shown in Figure 3. The first type of continuous entity retains all related entity values under an unstimulated steady state ($e_6$ in Figure 3). The other stores the ratio of steady state values, retained in the first type of special entities, to the entity values in the current time-step after EGF stimulation ($e_7$ in Figure 3). When carrying out in vivo and in vitro experiments for cell stimulation with growth factors, it is important to adjust cells to a constant
Figure 4: Comparison of simulation results and corresponding in vivo data. The blue diamonds represent the in silico simulation result. The red squares (five points for each plot) represent the result observed in vivo.

state before stimulation. The in vivo data in this paper also conform to the cell condition of culturing in serum-free medium for 16 h before EGF stimulation. This condition is usually called the steady state. It is considered that in the steady state, the amount of each molecule, except for the cell cycle elements, involved in the signal transduction pathway is constant. For emulating the steady state of an in vivo model, we applied 5000 pt simulation before EGF stimulation in the in silico model (pt is the virtual time of the HFPNe model; here, 60 pt = 0.1 h).

3.1 Parameter Estimation Procedure and Results

The in vivo data we used as input for parameter estimation with the DA contained six phosphorylated protein profiles (EGFR, Shc, c-Cbl, Vav2, Erk2, and p38MAPK). The procedure of parameter estimation by our DA method and its algorithmic details are described in our previous paper [27]. The EGFR model shown in Figure 2 consists of 53 entities and 115 processes. Because the number of free parameters are large, it is difficult to estimate the parameters of the entire network at once. As described in Section 2.2, we have divided the model into three sub-pathways with respect to the biological aspects and network structure: (i) EGFR regulation pathway, (ii) p38 MAPK activation pathway, (iii) ERK2 activation pathway. These sub-pathways have following features. Sub-pathway (i) can be treated separately from sub-pathway (ii) and (iii), because all the signal molecules in sub-pathway (i) are not regulated by the molecules in sub-pathway (ii) and (iii). Sub-pathway (ii) is regulated only by the molecules in sub-pathway (i). Sub-pathway (iii) is subject to regulation by both sub-pathway

Table 2: Fitting scores of the observed phosphorylated proteins.

<table>
<thead>
<tr>
<th>Entity Name</th>
<th>Fitting score [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-EGFR</td>
<td>99.9</td>
</tr>
<tr>
<td>p-c-Cbl</td>
<td>98.5</td>
</tr>
<tr>
<td>p-Shc</td>
<td>94.4</td>
</tr>
<tr>
<td>p-Vav2</td>
<td>98.2</td>
</tr>
<tr>
<td>p-Erk2</td>
<td>97.8</td>
</tr>
<tr>
<td>p-p38MAPK</td>
<td>5.4</td>
</tr>
</tbody>
</table>
Table 3: Description of the original model in Section 3.1 and the nine additionally hypothesized models.

<table>
<thead>
<tr>
<th>Name</th>
<th>Add/Remove</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>model 1</td>
<td>- (original)</td>
<td>Original model.</td>
</tr>
<tr>
<td>model 2</td>
<td>add</td>
<td>p38MAPK directly inactivates MKK3/4/6/7.</td>
</tr>
<tr>
<td>model 3</td>
<td>add</td>
<td>Erk2 directly inactivates MKK3/4/6/7.</td>
</tr>
<tr>
<td>model 4</td>
<td>add</td>
<td>p38MAPK itself activates the phosphatase of p38MAPK.</td>
</tr>
<tr>
<td>model 5</td>
<td>add</td>
<td>Erk2 directly activates the phosphatase of p38MAPK.</td>
</tr>
<tr>
<td>model 6</td>
<td>add</td>
<td>The combination of model 2 and model 4.</td>
</tr>
<tr>
<td>model 7</td>
<td>add</td>
<td>The combination of model 3 and model 5.</td>
</tr>
<tr>
<td>model 8</td>
<td>add</td>
<td>The combination of model 2 and model 5.</td>
</tr>
<tr>
<td>model 9</td>
<td>add</td>
<td>The combination of model 3 and model 4.</td>
</tr>
<tr>
<td>model 10</td>
<td>remove(control)</td>
<td>The activation process of Rac/Cdc42 by Vav2 is removed (T22 in Figure 2).</td>
</tr>
</tbody>
</table>

(i) and (iii) but it doesn’t affect other sub-pathways. This pathway decomposition based on network structure will be applicable for other models including many parameters. By effectively using the above features, the parameters of each sub-pathway were estimated and joined according to the following instructions (it must be noted that the estimation of the parameters in each sub-pathway was performed by the DA method after refining the range of each parameter by manual tuning). First, the parameters of sub-pathway (i) were estimated by the DA method and fixed. By fixing the parameters in sub-pathway (i), the influence of sub-pathway (i) on sub-pathway (ii) and (iii) became constant. Second, we estimated the parameters in sub-pathway (ii) and (iii) in that order since sub-pathway (ii) influences sub-pathway (iii) but no influence exists in the reverse direction. Regarding parameter estimation of sub-pathway (ii), some parameter sets provided results that were similar to the in vivo data. However, the influence of sub-pathway (ii) on sub-pathway (iii) is different among these estimated results. Thus, we treated each of these candidate parameter sets as the fixed parameter in the parameter estimation step of sub-pathway (iii), and selected the best parameter set among them.

By using the above processes, the estimation of the parameters in the entire model was completed. All the estimated parameters are shown in the supplementary information [50]. A comparison of the simulation results with these estimated parameters and corresponding in vivo data is shown in Figure 4. Table 2 shows the fitting scores calculated using \(1 - \frac{\sum (S_i - R_i)^2}{\sum R_i^2} \times 100\), where \(N\) is the number of observed points, and \(S_i\) and \(R_i\) (0 \(\leq i \leq N\)) are the values of simulated and observed data at time point \(i\), respectively. The fitting scores of all proteins except for p-p38MAPK were excellent. As in Figure 4, the discrepancy of p-p38MAPK between our simulation result and the in vivo data was the lack of down regulation at 10 to 20 min.

For the downregulation of an activated MAPK, some factors that promote dephosphorylation either inhibit phosphorylation or increase degradation are usually necessary. Thus, we assumed that the absence of such factors in our model was the cause of the difference in the dynamics of p-p38MAPK.

3.2 Model Selection

The important function of the DA framework is to objectively compare a number of hypothesized models and to extract the best model from among them.

The fitting score of p-p38MAPK was much worse than that of the other proteins examined. To extract models that are better than the previous model (model 1) and fit well with the experimen-
tal results for p-p38MAPK, we create eight EGFR models (model 2–9) by introducing additional hypothesized regulations that are related to the p38MAPK phosphorylation and dephosphorylation to the original model. In addition, as a control, one model (model 10) was created by removing the upstream regulator that influences p38MAPK phosphorylation. The details of these models are summarized in Table 3. For modeling regulations that inhibit the phosphorylation of p38MAPK, two models were created. Model 2 introduced the inhibitory regulation to MKK3/4/6/7 activity from p38MAPK. In contrast, model 3 introduced the same effect from Erk2. In order to model regulations that promote dephosphorylation of p38MAPK, two models were created. In model 4, the phosphatase that is directly activated by p38MAPK itself was modeled. Similarly, model 5 introduced the phosphatase directly activated by Erk2. Other models model 6–9 were combinations of these regulations. The best model was selected based on the approximated marginal likelihood [27]. The features of approximated marginal likelihood are as follows: if the p38MAPK fitting score is improved by adding parameters, due to the penalty for the newly added parameters, the score does not always improve. As summarized in Figure 5, five models (model 2, 4, 6, 8, 9) were better than model 1 (original model). The results showed that models 3, 5, and 7, which have additional regulation only from Erk2, were not acceptable. On the other hand, the models with additional regulation from p38MAPK are more applicable than the original one. Especially, model 6 including both direct inactivation of MKK3/4/6/7 and activation of phosphatase by p38 MAPK was the best among the 10 models. In summary, our in silico model and statistical framework suggested that some p38MAPK inhibitory regulation are derived not from Erk2 activation but from p38MAPK activation itself.

4 Conclusions and Future Work

We have concentrated on how to apply the high throughput proteomic data measured by LC/MS/MS with the SILAC method to in silico model parameter estimations with the DA. As a model system, we selected the EGFR pathway, which is one of the best-studied signal transduction pathways. We modeled the EGFR pathway by organizing the biological literature with the HFPNe architecture, and combined the in vivo data measured by the SILAC method with the parameter estimation method based on the DA. To perform efficient parameter estimation, we devised a method to simplify the complicated in silico model while retaining its generality. In summary, we divided the entire pathway into some sub-pathways (three in this model) in order to reduce interactions among them, and set the parameter estimation order among them depending on the directions of interactions that were shared by them. Our simulation results indicated that the presence of unknown additional regulations in the in silico model. Thus, we created eight models with additional hypothesized regulations and validated the hypothesized models by the DA method. The time-series data acquired by the proteomic approach includes many function-unknown proteins [2, 48]. This indicates that biological knowledge is still insufficient, even in the well-studied EGFR pathway. For modeling these proteins, it is essential to understand their roles in signal transduction. However, the experimental analysis of
each protein takes long time for obtaining its detailed functions in the pathway. Thus, based on high throughput computational methods, predicting protein function is one of the goals of in silico analyses and they have been studied extensively by using various approaches [47]. The Bayesian network inference is one of the promising approaches for the estimation of biological network causality. By the Bayesian network analysis, candidate signal transduction pathways including function-unknown proteins can be organized based on the proteomic data [3, 34]. Using information on these candidates, we will be able to refine our in silico model with the DA framework by performing the following steps. First, each candidate will be modeled with the HFPNe architecture and its dynamics will be refined by the parameter estimation function of the DA. Then, the best model from among these candidates will be selected by the model selection function of the DA. Theoretically, all the proteins measured by LC/MS/MS can be modeled with their dynamics through these steps. Our results showed that the DA could work with a relatively complex system. The next challenge would be to evaluate whether the DA framework can be applied to a highly complex system that reflects the events within the cell in detail. We believe that in the future it will be possible to extend our in silico model by the combination of the DA framework and the Bayesian technology.

References


