Bogdan Iancu | Usman Sanwal | Cristian Gratie | Ion Petre

# Fit-preserving refinement of the ErbB signalling pathway

Turku Centre *for* Computer Science

TUCS Technical Report
No 1187, August 2017

# Fit-preserving refinement of the ErbB signalling pathway

Bogdan Iancu
Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500 Finland
biancu@abo.fi

Usman Sanwal
Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500 Finland
usanwal@abo.fi

Cristian Gratie
Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500 Finland
cgratie@abo.fi

Ion Petre
Computational Biomodeling Laboratory
Åbo Akademi University and Turku Centre for Computer Science
Turku 20500 Finland
ipetre@abo.fi

TUCS Technical Report

No 1187, August 2017

**Abstract**

The construction of large scale biological models is a laborious task, which is often addressed by adopting iterative routines for model augmentation, adding certain details to an initial high level abstraction of the biological phenomenon of interest. However, refitting a model at every step of its development is time consuming and computationally intensive. In this context, *fit-preserving data refinement* brings about an effective alternative by providing adequate parameter values that ensure fit preservation at every refinement step. We address here the implementation of fit-preserving data refinement for a model of the ErbB signalling pathway, which is extended to include four different types of receptor tyrosine kinases, ErbB1-4, and two types of ligands, EGF and HRG. We build an extensive model, which ensures a good fit by construction with notably less effort than what a parameter estimation routine would require.

**Keywords:** Computational modelling; model construction; refinement; ErbB signalling pathway; ODE-based models; Event-B; invariant.

**TUCS Laboratory**
Computational Biomodeling Laboratory (Combio Lab)

# 1 Introduction

Traditional biology brought about major contributions to the understanding of biochemical foundations of living cells. The reductionist view that characterized biology for over a century accounts only for the functionality of individual molecular components, see [7]. However, biological systems are highly complex and certain aspects of their behaviour cannot even be predicted, let alone characterised by analysing individual components, see [24]. The mechanistic control of cellular activity is intricate and making predictions about its system-level behaviour is highly difficult. Our ability to make such predictions can be essential not only in reversing the dynamics of cellular impairment, but also in manipulating cellular activity towards auspicious behaviour. Moreover, a precise prediction of behavioural patterns of a large model indicates the ability to understand its elemental constituents and properties. Mathematical modelling is essential in making such predictions, but its use as a standard procedure in the field of practical applications is severely limited due to large numbers of parameters that are required either to be fixed or estimated, see [15].

A massive number of parameters to estimate requires the availability of a large volume of data and makes model fitting computationally intensive. For this reason, we focus on *refinement-based model construction* as an intermediary step preceding parameter estimation in the model development cycle. Stepwise refinement emerged from the field of software engineering. It was introduced at first as a concept in parallel computing and it expanded quickly, giving rise to the framework of refinement calculus, where it is promoted as a refinement method to ensure correctness preservation, see [6].

In the field of systems biology, model refinement becomes crucial in the model development cycle. Model fit is greatly affected by changes in the number of reactants, reactions, modules, etc. As argued above, the entire process of model fitting for considerably large models is not only a tedious task for the modeler as such, but it is computationally intensive since most parameter estimation routines take considerable time to complete and require massive amounts of computational resources. Hence, an iterative approach which relies on the conventional reiteration of the entire model fitting procedure is not feasible for large models. As an alternative, we consider an approach which ensures model fit preservation at every refinement step. The approach was discussed in the literature for rule-based models, see [11, 20]. For reaction-based models, the method was referred to as *quantitative model refinement*, see [18] and, more recently, it was extended and called *fit-preserving data refinement* [13].

We discuss in this paper the implementation of fit-preserving data refinement for a model describing the ErbB signalling pathway. Our refinement approach is based on *data refinement*, where a finite set of subspecies of a given species in the initial model are substituted for their corresponding parent species in the refined model. We start with a model of the EGFR (ErbB1) signalling pathway

proposed in [17] and [23]. Throughout the paper, the model from [17] will be referred to as the *basic model*. We refined this model to include four different types of receptor tyrosine kinases, ErbB1-4, structurally related to the epidermal growth factor receptor, EGFR, and two types of ligands, EGF and HRG, and we compared the computational effort with the one of [10]. The computational model we used was built based on the mass-action kinetic law, using an ODE-based formulation. The model was written and analysed in Copasi [16], including its parameter estimation.

Building further on the idea of constructing biomodels through model refinement, we also investigated the use of formal method support to make sure that our model have built in a correct way. For this, we built the basic model of ErbB signalling pathway using Event-B [3]. An Event-B system is a formal method approach, which is based on the B-Method and it is widely used for developing mathematical models of complex discrete systems. An Event-B specification is written as an abstract machine that consists of variables and these variables defines the state and some events. An event consists of predicates, the guard, which specify under what condition event will occur, and some actions. Event-B is supported by an Eclipse-based software package called the Rodin platform. With the help of Rodin platform one can edit the model and can generate all the required proof obligations. Modelling with Event-B has many applications in software engineering, see [3], and its advantages for biomodeling have only recently been investigated in [22].

The paper is organized as follows: we start with a description of the fit-preserving data refinement approach in Section 2. We describe the implementation of fit-preserving refinement on a case study in Section 3: we briefly introduce the main biological processes involved in the ErbB signalling pathway, as described in [17] and [23], then we discuss the numerical setup of the refined model. We discuss the Event-B-based model for the ErbB pathway in Section 4. We conclude the paper with an analysis of our results.

# 2 Quantitative model refinement

In this section we discuss the refinement of reaction-based models, as proposed in [18] and later extended in [13] to address both the construction of the refined model and the assignment of its kinetic rate constants in such a way that it captures the same dynamics as the original model. We aim to provide a self-contained presentation of the main result from [13] and, to this end, we adjusted the formal notation in order to avoid abstractions that are unnecessary for the work presented in this paper.

## 2.1 Data refinement

We consider a mass-action reaction-based model $M$ consisting of $m$ species, $\Sigma = \{A_1, \ldots, A_m\}$ and $n$ reactions $R = \{r_1, \ldots, r_n\}$, where $m, n \in \mathbb{N}^*$. We write the reactions of $M$ as:

$$r_j \ : \ c_{j1}A_1 + c_{j2}A_2 + \ldots + c_{jm}A_m \xrightarrow{k_j} d_{j1}A_1 + d_{j2}A_2 + \ldots + d_{jm}A_m \ , \qquad (1)$$

where, for each $k = \overline{1,n}$, $c_{jk}$ and $d_{jk}$ stand for the stoichiometric coefficients corresponding to reaction $r_j$'s left- and right-hand sides, respectively. Furthermore, $k_j$ is the kinetic rate constant of $r_j$.

The model $M$ can be refined to include more details about its species and/or reactions. Data refinement, as proposed in [18], refers to drawing a distinction between several variants of the same species. This is reflected in the refined model by replacing the considered species with subspecies corresponding to its variants. Whereas the initial distinction might refer to a single atomic species, one needs to account for it throughout the model and consider the corresponding differentiation of all complex species that contain the initially refined atomic element. The particular composition of complex species is relevant for deciding additional refinement that is triggered by an initial atomic refinement, but this composition plays no further role in the construction of the refined reactions and the choice of rate constant values. Without loss of generality, we can assume that each species $A_i$ is refined into $p_i$ subspecies, $\{B_{i1}, \ldots, B_{ip_i}\}$, where $p_i = 1$ for species $A_i$ which are not actually refined.

To obtain the reactions of the refined model, we assume that subspecies take part in the same interactions as the corresponding parent species, but possibly with different kinetic rate constants. More precisely, for each reaction $r_j$, the corresponding refinements are all possible rewritings of $r_j$ to use corresponding subspecies instead of species from the original model. Each reaction $r_j$ of the form (1) is thus replaced with reactions $r_j^{(l)}$ as follows:

$$\sum_{q=1}^{p_1} c_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} c_{jmq}^{(l)} B_{mq} \xrightarrow{k_j^{(l)}} \sum_{q=1}^{p_1} d_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} d_{jmq}^{(l)} B_{mq},$$

where $k_j^{(l)}$ is the kinetic rate constant of reaction $r_j^{(l)}$. The stoichiometric coefficients $c_{jiq}^{(l)}$ and $d_{jiq}^{(l)}$ are non-negative integers such that $c_{ji1}^{(l)} + \ldots + c_{jip_i}^{(l)} = c_{ji}$ and $d_{ji1}^{(l)} + \ldots + d_{jip_i}^{(l)} = d_{ji}$.

Note that, in the form presented so far, the refinement of the original model $M$ is only structural, as the rate constant values assignment is not addressed yet.

## 2.2 Fit-preserving refinement

Given a mass-action reaction-based model $M$ and its structural refinement $M_R$ constructed as above, we aim to assign appropriate values for the rate constants in

$M_R$. We follow the approach of [13] and aim to choose values that enable $M_R$ to capture the same dynamics as $M$, in the following sense:

$$[A_i](t) = [B_{i1}](t) + \ldots + [B_{ip_i}](t) \ , \tag{2}$$

i.e., at any time $t$, the concentration predicted by $M$ for any species $A_i$ is the same as the sum of concentrations predicted by $M_R$ for the subspecies of $A_i$. These constraints define what is referred to as *fit-preserving refinement* in [13]. To ensure the satisfaction of (2), the authors show that it is enough to impose simple linear constraints that relate the kinetic rate constants of $M_R$ to those or $M$.

We first introduce some notations. We will use vectors $\boldsymbol{c_j}$ and $\boldsymbol{d_j}$ to denote the stoichiometric coefficients of reaction $r_j$. Similarly, the corresponding coefficients of the refined reactions $r_j^{(l)}$ will be denoted by $\boldsymbol{c_j^{(l)}}$ and $\boldsymbol{d_j^{(l)}}$, respectively. The sufficient conditions for $M_R$ to be a fit-preserving refinement, according to [13], can be written as:

$$\sum_{l \text{ s.t. } \boldsymbol{c_j^{(l)}} = \boldsymbol{c_j^{(s)}}} k_j^{(l)} = \binom{\boldsymbol{c_j}}{\boldsymbol{c_j^{(s)}}} k_j \ , \ \textit{where} \ \binom{\boldsymbol{x}}{\boldsymbol{y}} = \frac{\prod_i x_i!}{\prod_j y_j!}, \ , \tag{3}$$

for any reaction $r_j$ and any selected refinement of its left-hand side $\boldsymbol{c_j^{(s)}}$. The sum is taken over all refined reactions $r_j^{(l)}$ that have the selected left hand side $\boldsymbol{c_j^{(s)}}$. The intuitive interpretation of (3) is that the refined rate constants depend on the rate constant of the original reaction $r_j$, as well as on the left-hand side stoichiometric coefficients of both the original reaction and the refined one.

The approach presented for building the structural refinement of a reaction-based model will typically add a large number of reactions that are not chemically meaningful, for example because they violate conservation laws for atomic species. In order to remove such reactions from the fit-preserving refinement, we set the corresponding rate constants to zero. For the refinement presented in this paper, we have chosen equal values for the remaining constants, by dividing the sum prescribed by (3) to the number of non-zero rate constants.

# 3 Case-study: the ErbB signalling pathway

In this section we describe our case study, the EGFR signalling network, and its fit-preserving data refinement.

## 3.1 Biological background

The ErbB signalling pathway is an evolvable pathway, responsible for the regulation of various physiological responses of the mammalian cell, such as growth,

survival, proliferation, differentiation and motility, see [8, 10, 21]. Due to its intrinsic complexity and association with the progression of various cancer types, the pathway was extensively analysed. However, being highly robust as a result of a notable modularity and redundancy, it exhibits in turn an unavoidable vulnerability, being key to the unremitting growth and development of carcinoma cells. Since recent findings brought about vast amounts of information regarding its organization and its signalling compartmentalisation, detecting any type of aberrant behaviour in the activation of the pathway entails a system-level perspective in its analysis, [8, 25].

The network involves a number of extracellular ligands, four receptor tyrosine kinases (RTKs), ErbB1-4: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4, and various intracellular proteins (cytoplasmic adapters, scaffolds and enzymatic proteins). Following a process of homo- and hetero-dimerization, the receptors bind to multiple ligands, leading to the activation of the downstream Ras/Raf/MEK/ERK cascades.

We introduce in the following the functional properties of the signalling pathway, focusing solely on the influence of one of the receptor tyrosine kinases of the ErbB family: the epidermal growth factor receptor (EGFR). The epidermal growth factor (EGF) binds to the extracellular domain of the transmembrane epidermal growth factor receptor (EGFR). The ligand-bound receptor undergoes a process of dimerization, which precedes an accelerated auto-phosphorylation of its intracellular domain. The activated ligand-bound receptors recruit a number of cytoplasmic enzymes and adapter proteins, initiating signal propagation down the Ras/Raf/MEK/ERK cascades.

The activation of Ras-GTP through the hydrolization of Ras-GDP is promoted by the internalization and dissociation of a suite of signalling molecules. There are two signalling pathways that entail the activation of the Ras-GTP protein: the Shc-dependent and Shc-independent pathways. The Shc-dependent pathway commences promoted by the binding of Shc to the autophosphorylated, ligand-bound, dimerized receptor and is sustained through the binding to the growth factor receptor-binding protein 2, Grb2. The Shc-independent pathway in turn is sustained by a direct binding of the autophosphorylated, ligand-bound, dimerized receptor to Grb2. Both the Shc-dependent and Shc-independent pathways involve the recruitment of Sos, protein Ras being docked onto the membrane and its association with Sos promoting the formation of Ras-GTP. The activated Ras-GTP triggers the mitogen activated protein kinase (MAPK) signalling cascade through the Raf, MEK and ERK kinases, see [17, 23]. The effect brought about by signalling is the activation (phosphorylation) of ERK, which in turn regulates the dynamics of multiple cellular proteins and transcription factors involved in cellular growth and differentiation, see [23].

The initial model, introduced in [17], is a reaction-based model of the EGF-induced signal transduction through the Ras/Raf/MEK/ERK cascades and it consists of 148 reactions, 103 reactants and 90 kinetic rate constants. It is an updated

version of two previous models shown in [19] and [23]. The model includes a negative feedback loop from the doubly phosphorylated ERK (ERK-PP) to the Sos protein, leading to the unbinding of Grb2-Sos from the receptor complex, see [9, 12]. Protein isoform specificity (multiple forms for the same protein) is not accounted for in the model in [17]. The system described by the model in [17] is characterized by a stable steady state in the absence of stimulus (EGF), corresponding to a state of inactive (unphosphorylated) ERK. The model specifies two pools of doubly phosphorylated ERK, one located in the cytoplasm and one correlated with the internalization process, see [17].

The model in [17] accounts for a set of 13 biochemical processes: EGFR activation, Shc, Grb2, Sos recruitment, activation and inactivation of Ras, activation of Raf, dephosphorylation of Raf, phosphorylation and dephosphorylation of MEK, ERK dephosphorylation, negative feedback from ERK to Sos, internalization of complexes involving EGFR and degradation reactions. For more details, we refer the reader to [17]. We have imported the model in COPASI [16], it is available at [1].

## 3.2 Fit-preserving data refinement

This subsection focuses on the augmentation of the EGFR signalling pathway model from [17] by implementing fit-preserving data refinement. In this regard, we take into account all four receptor members of the ErbB family: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4. We also consider two types of ligands: EGF and HRG.

The refined model comprises species divided in two categories: *atomic* or *complex*, see [14]. We define accordingly as atomic species those species which have an autonomous structure and representing an irreducible unit within the model. A complex species is then composed of at least two atomic species bound together. All the four members of the ErbB family, ErbB1-4, as well as both ligands, EGF and HRG, are atomic species. These species are to be refined in the model and none of the other atomic species present in the model from [17] is refined. All complex species present in the model from [17] comprising ErbB1 (EGFR) and/or EGF are refined to include all four receptor members of the ErbB family, and the types of ligands: EGF and HRG. We also take into account all dimer and receptor-ligand binding combinations. We describe formally the above data refinements as follows:

$$\begin{aligned} &\mathsf{EGFR} \rightarrow \{\mathsf{ErbB1}, \mathsf{ErbB2}, \mathsf{ErbB3}, \mathsf{ErbB4}\}; \\ &\mathsf{EGF} \rightarrow \{\mathsf{EGF}, \mathsf{HRG}\}. \end{aligned}$$

In what follows we describe the fit-preserving refinement of the model by examining the various types of reactions from the original model and explaining

how numerical values were chosen for the rate constants of the corresponding refined reactions.

The entire signalling process is triggered by receptor-activation: the ligand (EGF or HRG in the refined model) binds to the receptor (in the refined model: ErbB1, ErbB2, ErbB3, ErbB4). The initiating reaction for receptor-activation in the model from [17] is the following:

$$\mathsf{EGF} + \mathsf{EGFR} \underset{k_b^-}{\overset{k_b^+}{\rightleftharpoons}} \mathsf{EGF\text{-}EGFR}. \tag{4}$$

The reaction above is refined in our model to account for both types of ligands ($L_i, i = \overline{1,2}$) and all four types of receptors ($R_j, j = \overline{1,4}$), leading to the following reactions:

$$L_i + R_j \underset{k_{i,j}^-}{\overset{k_{i,j}^+}{\rightleftharpoons}} L_i - R_j, \text{ for all } i = \overline{1,2}, j = \overline{1,4}, \tag{5}$$

where $L_i \in \{\mathsf{EGF}, \mathsf{HRG}\}$ and $R_j \in \{\mathsf{ErbB1}, \mathsf{ErbB2}, \mathsf{ErbB3}, \mathsf{ErbB4}\}$.

Our objective here is to determine the kinetic parameters for the refined model so that it is a fit-preserving refinement of the model in [17], i.e. the sufficient conditions (3) for our refined model are met. Let us now look at the refinements of the ligand-binding reaction (4). If we focus on the forward direction, there is a single reaction for each left-hand side (i.e. for all other reactions sharing the same left-hand side we implicitly set the rate constant to zero). The same is true for the reverse direction. Thus, we can ensure that constraints (3) are fulfilled by setting $k_{i,j}^- = k_b^-$ and $k_{i,j}^+ = k_b^+$, for all $i = \overline{1,2}$, $j = \overline{1,4}$.

The ligand-binding receptor activation reaction is followed by a dimerization reaction, whose products are associated with a considerable number of reactants in the *ErbB* signalling pathway. We list below the dimerization reaction from the basic model of [17]:

$$2\,\mathsf{EGF\text{-}EGFR} \underset{k_d^-}{\overset{k_d^+}{\rightleftharpoons}} (\mathsf{EGF\text{-}EGFR})^2 \ .$$

We refine this reaction to account for all possible combinations of monomers, thus obtaining $8 + \binom{8}{2} = 36$ possible left-hand sides. For the right-hand side, we only consider homo-dimers as refinements of the original dimer. We obtain the following refined reactions:

$$(\mathsf{L\text{-}R})_i + (\mathsf{L\text{-}R})_j \underset{k_{i,j,l}^-}{\overset{k_{i,j,l}^+}{\rightleftharpoons}} (\mathsf{L\text{-}R})_l^2 \ , \tag{6}$$

where $i,j,l = \overline{1,8}$ such that $i \leq j$. For any $k = \overline{1,8}$ the refined species $(\mathsf{L\text{-}R})_k$ are (in some arbitrary, fixed, order) the elements of the following set:

$$\mathcal{B} = \{\mathsf{EGF-ErbBp, HRG-ErbBq} \mid p,q = \overline{1,4}\} \ .$$

The constraints (3) for ensuring fit-preservation translate to the following relations for the dimerization reaction:

$$\sum_{l=1}^{8} k^{+}_{i,j,l} = \begin{cases} k^{+}_{d}, & \text{if } i = j; \\ 2k^{+}_{d}, & \text{if } i < j; \end{cases}$$

$$\sum_{1 \leq i \leq j \leq 8} k^{-}_{i,j,l} = k^{-}_{d} \, .$$

The total number of refined reactions for the dimerization is $36 \times 8 = 288$. In setting the values for kinetic rate constants we aim to reduce this number by assuming that the interaction of $(\mathsf{L\text{-}R})_i$ and $(\mathsf{L\text{-}R})_j$ can only produce $(\mathsf{L\text{-}R})^2_i$ or $(\mathsf{L\text{-}R})^2_j$, i.e. we set the rate constants of the other reactions to zero. Just as before, the remaining rate constants are set to equal values that add up to the corresponding sum required by the above constraints. We obtain the following assignments:

$$k^{+}_{i,j,l} = \begin{cases} k^{+}_{d}, & \text{if } l = i \text{ or } l = j; \\ 0, & \text{otherwise;} \end{cases}$$
$$k^{-}_{i,j,l} = \begin{cases} \frac{k^{-}_{d}}{8}, & \text{if } l = i \text{ or } l = j; \\ 0, & \text{otherwise.} \end{cases}$$

The dimerization reaction is followed down the signalling pathway by a phosphorylation reaction, which facilitates the process of receptor activation:

$$(\mathsf{EGF\text{-}EGFR})^2 \underset{k^{-}_{f}}{\overset{k^{+}_{f}}{\rightleftharpoons}} (\mathsf{EGF\text{-}EGFR^{*}})^2.$$

The phosphorylation of the ligand-bound receptor reaction is refined into the following reactions:

$$(\mathsf{L\text{-}R})^2_l \underset{k^{-}_{l}}{\overset{k^{+}_{l}}{\rightleftharpoons}} (\mathsf{L\text{-}R^{*}}_l)^2, \tag{7}$$

where for any $k = \overline{1,8}$ the refined species $(\mathsf{L\text{-}R})_k$ are (in some arbitrary, fixed, order) the elements of the set $\mathcal{B}$ as defined above. The kinetic rate constants of the refined phosphorylation reactions are set to equal the kinetic rate constants of the

original reaction, taking into account only reactions which have on the right hand side the phosphorylated counterpart of the left hand side:

$$\begin{aligned} k_l^+ &= k_f^+; \\ k_l^- &= k_f^-. \end{aligned} \tag{8}$$

The process of receptor activation is sustained by the following receptor production reaction:

$$\xrightarrow{k_p} \mathsf{EGFR} \tag{9}$$

The corresponding refined reactions are:

$$\xrightarrow{k_i} R_i, \quad i = \overline{1,4}, \tag{10}$$

with $R_i \in \{\mathsf{ErbB1}, \mathsf{ErbB2}, \mathsf{ErbB3}, \mathsf{ErbB4}\}$. We set the kinetic rate constants for receptor production so as to comply with (3):

$$k_i = \frac{k_p}{4}. \tag{11}$$

All reactions which have exactly one substrate comprising an EGF-EGFR dimer in its conformation or possibly its internalized equivalent, EGF-EGFRi, are refined into 8 different reactions, corresponding to each type of dimer-derived refined complex. Let us take for example a complex species, $(\mathsf{EGF\text{-}EGFR*})^2\text{-}\mathsf{AC}$, where AC represents a chain of bound atomic species (e.g. GAP-Grb2-Sos-Ras-GDP-Prot). This species is refined as follows:

$$(\mathsf{EGF\text{-}EGFR*})^2\text{-}\mathsf{AC} \rightarrow \{(\mathsf{L\text{-}R}^*)_j^{\,2} - AC \mid j = \overline{1,8}\}, \tag{12}$$

where $(\mathsf{L\text{-}R})_j \in \mathcal{B}$ and the $*$ character stands for the phosphorylation of the respective ErbB molecule. The refinement of species manifests in the refinement of reactions: all reactions involving a complex species in the initial model are to be refined accordingly.

Complexes of the form $\{(\mathsf{L\text{-}R}^*)_j^{\,2} - AC\}$ are involved in reactions of the following types:

$$\begin{aligned} &(\mathsf{L\text{-}R}^*)_j^{\,2} - AC + AS \underset{k_g^-}{\overset{k_g^+}{\rightleftharpoons}} (\mathsf{L\text{-}R}^*)_j^{\,2} - AC - AS, \\ &(\mathsf{L\text{-}R}^*)_j^{\,2} - AC \xrightarrow{k_g} (\mathsf{L\text{-}R}^*)_j^{\,2} - AC - deg, \\ &(\mathsf{L\text{-}R}^*)_j^{\,2} - AC \xrightarrow{k_{g_i}} (\mathsf{L\text{-}R_i}^*)_j^{\,2} - AC, \end{aligned}$$

where *AS* is an atomic species as previously defined, *deg* is appended to illustrate the product of a degradation reaction, and *i* specifies an internalized counterpart of the respective reactant.

We set the kinetic rate constants for reactions of the above types to equal the kinetic rate constants of their corresponding initial reaction, setting to zero the kinetic rate constants of those reactions which have on the right hand side products originating from other reactants than those on the left hand side of the respective reaction.

### 3.3 Initial concentration values for the refined model

In the previous subsection we discussed the values of the kinetic rate constants for the refined model. In order to have a completely specified refinement, we need to also consider the initial concentration values. In the refined model the initial values are set so as to comply with the fit-preserving refinement relations, i.e. to reflect that the concentration of a species in the initial model equals the sum of the concentrations of all its subspecies present in the refined model. For example, consider again the complex species $(\mathsf{EGF\text{-}EGFR*})^2\text{-}\mathsf{AC}$. The initial concentration values should satisfy:

$$[(\mathsf{EGF\text{-}EGFR*})^2\text{-}\mathsf{AC}](0) = \sum_{j=1}^{8} [(\mathsf{L\text{-}R^*}_j)^2 - AC](0),$$

where $(\mathsf{L\text{-}R})_j \in \mathcal{B}$, the "$*$" character represents the phosphorylation status of the respective ErbB molecule and *AC* stands for a chain of bound atomic species.

Following the same approach as before, we choose not to favour any of the subspecies and thus assign equal values for the initial concentrations:

$$[(\mathsf{L\text{-}R^*}_j)^2 - AC](0) = \frac{1}{8}[(\mathsf{EGF\text{-}EGFR*})^2\text{-}\mathsf{AC}](0) \ .$$

The same pattern applies for the numerical setup of all initial concentrations for the refined species.

## 4 An Event-B model for the ErbB signalling pathway

In this section we discuss about building our Event-B-based model of the ErbB signalling pathway. Event-B [3] is a formal method approach which was developed from classical B [2] and action systems [5]. Event-B is used for modeling and analysis of systems and its modelling language is based on set theory and first order logic. An Event-B model consists of two types of modules, which are *contexts* and *machines*. The context has a *static* part of model, which can be *types*, *constants*, and *axioms* while machine has a *dynamic* part of model, which can be

*variables*, *invariants* and *events*. The initial state of system is specified with a specific event called *Initialisation*. An event consists of *guards* and *actions*. A guard is a condition for an event to be *enabled* and actions are assignments of various variables. A context can be *extended* by other contexts and a machine can be *refined* by other machines. A machine can also *see* one or more contexts.

In Event-B building a model usually starts from an abstract version of the model and it advances in successive refinements. In the initial abstraction level, a model may only focus on the main purpose of the system and details to the abstract model are added gradually through successive refinements. In Event-B, the concept of refinement is based on adding new features to the model, or adding details to current features of the model. This can be done by adding new events or refining the consisting events. Rodin [4] is an open source Eclipse-based tool, which provides the modelling and proving support in Event-B.

Here we recall briefly a general scheme to build a chemical reaction network in Event-B, discussed in details in [22] to build an Event-B model for the heat shock response. Each species of the reaction network is modelled by a variable of type $\mathbb{N}$, denoting the amount of that species. The model has one event associated to each reaction of the network. For example, for the two reactions $2A \rightarrow B$ and $A + B \rightarrow 3A$, we will have in the Event-B model two variables corresponding to these two species, as shown below. We introduce two invariants specifying their type, as well as an invariant expressing that the sum of $A$ and $B$ is conserved, as suggested by the two reactions. These invariants must be checked to hold after each refinement of the model, thus ensuring the correct construction of the model.

| **Variables** |
| --- |
| $A, B$ |
| **Invariants** |
| $A \in \mathbb{N}$ |
| $B \in \mathbb{N}$ |
| $A + B = constant$ |

We initialise each of these variables with their initial values corresponding to their reaction network values in Initialisation event. Apart from that, we will have two events for these two reactions in Event-B. For each event, the guards of the event make sure that there should be enough of each of the reactant for the reaction to be enabled while actions of event describe how the value of each variable changes. An Event-B model corresponding to these two reactions is shown in Table 1.

For a detailed discussion on building Event-B models for biochemical reaction networks we refer to [22].

We applied the general scheme presented above to the basic model of ErbB signaling pathway. The resulting model can be downloaded at [1]. We only show here two of the event in Table 2. All the proof obligations corresponding to these events were automatically discharged and these proof obligations serve us as a proof of correct construction of our model.

Table 1: The general form of an Event-B model for a reaction network

| **Event 1** |
| --- |
| **WHERE** |
| **@grd1** $A \geq 2$ |
| **THEN** |
| **@act1** $A := A - 2$ |
| **@act2** $B := B + 1$ |
| **END** |

| **Event 2** |
| --- |
| **WHERE** |
| **@grd1** $A \geq 1 \wedge B \geq 1$ |
| **THEN** |
| **@act1** $A := A - 1$ |
| **@act2** $B := B - 1$ |
| **@act3** $A := A + 3$ |
| **END** |

Table 2: Two events modeling the forward and reverse directions of the first reaction of the ErbB signalling pathway

| **Rec1f** |
| --- |
| **WHERE** |
| **@grd1** $\mathsf{EGFR} \geq 1 \wedge \mathsf{EGF} \geq 1$ |
| **THEN** |
| **@act1** $\mathsf{EGFR} := \mathsf{EGFR} - 1$ |
| **@act2** $\mathsf{EGF} := \mathsf{EGF} - 1$ |
| **@act3** $\mathsf{EGF\text{-}EGFR} := \mathsf{EGF\text{-}EGFR} + 1$ |
| **END** |

| **Rec1r** |
| --- |
| **WHERE** |
| **@grd1** $\mathsf{EGF\text{-}EGFR} \geq 1$ |
| **THEN** |
| **@act1** $\mathsf{EGF\text{-}EGFR} := \mathsf{EGF\text{-}EGFR} - 1$ |
| **@act2** $\mathsf{EGFR} := \mathsf{EGFR} + 1$ |
| **@act3** $\mathsf{EGF} := \mathsf{EGF} + 1$ |
| **END** |

# 5 Discussion

When building an extensive system-level biological model, refinement becomes a crucial step in the model development cycle. Starting with a high level abstraction of a biological process of interest, one very often needs to include more details regarding its reactants, reactions or constituent modules. A conventional approach which would involve a reiteration of the entire model development cycle is highly ineffective, since it involves running parameter estimation routines over large sets of parameters, requiring long-running intervals to complete and significant computational resources. For instance, for the model in [10] consisting of 499 reactants and 828 reactions, a good fit was obtained by running about 100 times annealing methods, over 24 hours on a cluster consisting of 100 nodes.

We refined the model from [17], considering two types of ligands: EGF and HRG and four different types of receptor tyrosine kinases: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4. This refinement brought about a massive augmentation in the number of reactants and, consequently, the number of reactions. While the

initial model comprises a number of 103 reactants and 148 reactions, the refined model consists of a number of 421 reactants involved in 928 reactions. The effort of fitting a model of this size could be commensurate with regards to the effort of model fitting to that of [10]. However, our approach proved to be effective in building the refined model, by exhibiting a good numerical behaviour without any supplementary model fit. The development of the refined model does not require domain specific knowledge.

Fit-preserving data refinement ensures a good fit by construction, starting from an already fit original model; further refinement steps can be applied after this original refinement so as to include more details regarding biological knowledge of the model. Our approach does not require any effort in model fitting. Moreover, if experimental data and computational resources are available, the fit-preserving refinement can be used as initialization for parameter estimation routines in order to improve the model fit. Note that in this case we are guaranteed to obtain at least as good a fit as the one of the original model, whereas starting the fitting process from scratch may lead to worse local optima. Moreover, all computational effort for model fitting goes into improving the initial model rather than randomly exploring the parameter space. This improves the scalability of compiling large models by stepwise refinement.

The scalability of fit-preserving refinement does not cover genome-scale models, however, it is very suited for the development of large-scale components and/or models and also in combination with model composition, either at the level of simulation or, even more, considering model composition as a meta-operation. We will visit this concept in future work.

Our methodology is versatile, as it is compatible with the integration of partial information regarding some parameters of the refined model. This makes it a suitable candidate for compiling large models, providing an algorithmic assignment of parameter values. The methodology only applies for mass-action models, we will consider other kinetic models in further studies.

Our refinement technique can describe several fit-preserving refinements for a given mass-action kinetic model based on the chosen values for the rate constants, allowing the modeler to subsequently filter out unreasonable reactions from the refined model.

Fit-preserving refinement can bring about considerable contribution to the construction of hierarchical models by allowing the migration from one level of resolution to another. An interesting aspect would be the possibility of retrieving a more general abstract representation of a model from a detailed one after the latter was refitted.

Building such a large model as the ErbB signalling pathway is quite a difficult and error-prone task when introducing new variables and parameters. Formal methods-based approaches with the concept of refinement as their core feature can be helpful to build such a large model. In particular, the formal method support is essential to guarantee that the model is built in a correct way, in the sense of

having all its invariant properties satisfied. For this, we provided Event-B/Rodin support, which in itself is very useful tool to build such a large model as the key feature of Event-B/Rodin is having the concept of refinement. Very little work has been done in this domain, with this line of research only recently being started in [22]. As compared to [22], our current model is considerably larger. In fact, the ErbB model turned out to be considerably larger than all models built so far with Rodin. This turned out to be beyond the current technical capabilities of Rodin and we could not in fact complete the full refinement of the model in Rodin. Theoretically/conceptually, it is possible to build such a big model with Rodin and the tool itself can be extended to allow for larger models. We are planning to work with developers of Rodin to overcome this problem with the explicit goal of extending the use of formal method-based biomodeling.

# References


[1] Copasi and eventb models for the control of mapk signalling. `http://combio.abo.fi/research/quantitative-model-refinement/refinement-of-the-egfr-signalling-pathway/`, 2013.

[2] Jean-Raymond Abrial. *The B-book: Assigning Programs to Meanings*. Cambridge University Press, New York, NY, USA, 1996.

[3] Jean-Raymond Abrial. *Modeling in Event-B: System and Software Engineering*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

[4] Jean-Raymond Abrial, Michael Butler, Stefan Hallerstede, Thai Son Hoang, Farhad Mehta, and Laurent Voisin. Rodin: an open toolset for modelling and reasoning in Event-B. *STTT*, 12(6):447–466, 2010.

[5] Ralph-Johan Back and Reino Kurki-Suonio. Decentralization of process nets with centralized control. In *Proceedings of the Second Annual ACM Symposium on Principles of Distributed Computing*, PODC '83, pages 131–142, New York, NY, USA, 1983. ACM.

[6] R.J. Back and J. von Wright. *Refinement Calculus*. Springer (New York), 1998.

[7] A.-L. Barabasi and Z.N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[8] M.R. Birtwistle, M. Hatakeyama, N. Yumoto, B.A. Ogunnaike, J.B. Hoek, and B.N. Kholodenko. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Molecular systems biology*, 3(1), 2007.

[9] L. Buday, P.H. Warne, and J. Downward. Downregulation of the Ras activation pathway by MAP kinase phosphorylation of Sos. *Oncogene*, 11(7):1327–1331, 1995.

[10] W.W. Chen, B. Schoeberl, P.J. Jasper, M. Niepel, U.B. Nielsen, D.A. Lauffenburger, and P.K. Sorger. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular systems biology*, 5(1), 2009.

[11] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In *Formal Methods in Systems Biology*, pages 103–122. Springer, 2008.

[12] C. Dong, S.B. Waters, K.H. Holt, and J.E. Pessin. Sos phosphorylation and disassociation of the Grb2-SOS complex by the ERK and JNK signaling pathways. *Journal of Biological Chemistry*, 271(11):6328–6332, 1996.

[13] C. Gratie and I. Petre. Fit-preserving data refinement of mass-action reaction networks. In *Language, Life, Limits*, pages 204–213. Springer, 2014.

[14] D.E. Gratie, B. Iancu, S. Azimi, and I. Petre. Quantitative model refinement in four different frameworks, with applications to the heat shock response. Technical Report 1067, TUCS, 2013.

[15] W.S. Hlavacek. How to deal with large models? *Molecular Systems Biology*, 5(1), 2009.

[16] S. Hoops, S. Sahle, R. Gauges, Ch. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. Copasi–a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[17] J.J. Hornberg, B. Binder, F.J. Bruggeman, B. Schoeberl, R. Heinrich, and H.V. Westerhoff. Control of MAPK signalling: from complexity to what really matters. *Oncogene*, 24(36):5533–5542, 2005.

[18] B. Iancu, El. Czeizler, Eu. Czeizler, and I. Petre. Quantitative refinement of reaction models. *IJUC*, 8(5-6):529–550, 2012.

[19] B.N. Kholodenko, O.V. Demin, G. Moehren, and J.B. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, 274(42):30169–30181, 1999.

[20] E. Murphy, V. Danos, J. Feret, J. Krivine, and R. Harmer. Rule based modeling and model refinement. *Elements of Computational Systems Biology*, pages 83–114, 2009.

[21] K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular systems biology*, 1(1), 2005.

[22] Usman Sanwal, Luigia Petre, and Ion Petre. Stepwise construction of a metabolic network in Event-B: the heat shock response. *TUCS Technical Reports*, (1160), 2016.

[23] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Müller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature biotechnology*, 20(4):370–375, 2002.

[24] Marc H.V. Van Regenmortel. Reductionism and complexity in molecular biology. *EMBO reports*, 5(11):1016–1020, 2004.

[25] Y. Yarden and M. Sliwkowski. Untangling the ErbB signalling network. *Nature reviews Molecular cell biology*, 2(2):127–137, 2001.

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 A, 20520 TURKU, Finland | www.tucs.fi

**University of Turku**
*Faculty of Mathematics and Natural Sciences*
- Department of Information Technology
- Department of Mathematics

*Turku School of Economics*
- Institute of Information Systems Sciences

**Åbo Akademi University**
- Department of Computer Science
- Institute for Advanced Management Systems Research

ISBN 978-952-12-3582-5
ISSN 1239-1891