Computers in Biology and Medicine 106 (2019) 91–96

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Refinement-based modeling of the ErbB signaling pathway

Bogdan Iancu[a,b,1], Usman Sanwal[a,b,1], Cristian Gratie[a,b], Ion Petre[a,c,d,*]

[a] Computational Biomodeling Laboratory, Turku Centre for Computer Science, Finland
[b] Department of Computer Science, Åbo Akademi University, Finland
[c] Department of Mathematics and Statistics, University of Turku, Finland
[d] National Institute for Research and Development in Biological Sciences, Romania

ARTICLE INFO

Keywords:
Computational modeling
Model construction
Refinement
ErbB signaling pathway
ODE-Based models
Event-B
Invariant

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. Refitting a model at every step of its development is time consuming and computationally intensive. The concept of *model refinement* brings about an effective alternative by providing adequate parameter values that ensure the preservation of its quantitative fit at every refinement step. We demonstrate this approach by constructing the largest-ever refinement-based biomodel, consisting of 421 species and 928 reactions. We start from an already fit, relatively small literature model whose consistency we check formally. We then construct the final model through an algorithmic step-by-step refinement procedure that ensures the preservation of the model's fit.

1. Introduction

Mechanistic control of cellular activity is intricate and making predictions about its system-level behavior is highly difficult. Our ability to make such predictions can be essential not only in reversing the dynamics of cellular impairment, but also in directing cellular activity towards a more favorable behavior. Mathematical modeling 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 Ref. [1].

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 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 Ref. [2].

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. 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 Refs. [3,4]. For reaction-based models with a quantitative dynamic described by ODEs, the method was referred to as *quantitative model refinement*, see Ref. [5] and then extended and called *fit-preserving data refinement* [6].

We discuss in this paper the implementation of the largest-ever model built through model refinement, describing the ErbB signaling 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 in the refined model for their corresponding 'parent' species in the initial model. We started with a model of the EGFR (ErbB1) signaling pathway proposed in Refs. [7,8]. Throughout the paper, the model from Ref. [7] is 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 needed to build it with that of [9]. We used logic-based formal methods support based on modeling with Event-B [10] to

* Corresponding author. Computational Biomodeling Laboratory, Turku Centre for Computer Science, Finland.
E-mail address: ion.petre@utu.fi (I. Petre).
[1] Authors with equal contribution.

https://doi.org/10.1016/j.compbiomed.2019.01.016
Received 10 October 2018; Received in revised form 18 January 2019; Accepted 19 January 2019
0010-4825/

make sure that the basic model is consistent. An Event-B specification is written as an abstract machine that consists of variables and events. An event has a logical guard, which specifies under what condition the event will occur, and some actions. This allowed us to ensure that all 110 species (variables in Event-B) are properly defined and that the 148 reactions (events in Event-B) are consistently written. We then refined this basic model to its full version, consisting of 421 species and 928 reactions. This part of the modeling was done based on our methods on quantitative model refinement and implemented in COPASI [11], resulting in a mass-action, ODE-based model. We describe in this article the modeling process and discuss its implications.

2. Background

2.1. Quantitative model refinement

The refinement of reaction-based models was proposed in Ref. [5] and later extended in Ref. [6] 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 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:

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

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 Ref. [5], refers to introducing a way to distinguish between several variants of a species. This is reflected in the refined model by replacing the considered species with subspecies corresponding to its variants. 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. We consider that all the B-species are new variables, distinct from the A-species.

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 the rewritings of r_j that use corresponding subspecies instead of species from the original model in all possible ways. Each reaction r_j of the form (1) is thus replaced with reactions $r_j^{(l)}$ as follows:

$$\sum_{q=1}^{p_1} c_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} c_{jmq}^{(l)} B_{mq} \xrightarrow{k_j^{(l)}} \sum_{q=1}^{p_1} d_{j1q}^{(l)} B_{1q} + \ldots + \sum_{q=1}^{p_m} d_{jmq}^{(l)} B_{mq}, \tag{2}$$

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, from a structural point of view, the refined model M_R captures the same kind of interactions as the original model M, only written in terms of the refined species B_{iq} and allowing for different values of the kinetic rate constants.

To get a consistent refined model, M_R has to capture the same dynamics as M, in the following sense:

$$[A_i](t) = [B_{i1}](t) + \ldots + \left[B_{ip_i}\right](t). \tag{3}$$

This means that 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. To ensure that (3) holds, it was shown in Ref. [6] that it is enough to impose a simple linear constraints that relate the kinetic rate constants of M_R to those or M. This condition can be formulated as follows. We will use vectors $\boldsymbol{c_j}$ and $\boldsymbol{d_j}$ to denote the stoichiometric coefficients of reaction r_j. 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 Ref. [6], can be written as:

$$\sum_{l\ s.t.\ \boldsymbol{c_j^{(l)}} = \boldsymbol{c_j^{(s)}}} k_j^{(l)} = \binom{\boldsymbol{c_j}}{\boldsymbol{c_j^{(s)}}} k_j \,, \text{ where } \binom{\boldsymbol{x}}{\boldsymbol{y}} = \frac{\prod_i x_i!}{\prod_j y_j!} \,, \tag{4}$$

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 (4) 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. Two examples demonstrating the use of this constraint to build some models can be found in Ref. [6].

Condition (4) gives the modeler a wide range of options for how to set the parameters of the refined model. Any choice of a numerical setup that satisfies (4) will ensure that the refined model is consistent with the basic model and in particular, that it remains data-fit to the same extent that the basic model was. Some of the parameters of the refined model may be known, either from the literature, or from direct measurements. They can be integrated into the setup of the refined model simply by setting their known values in (4) and using them to determine suitable values for the remaining unknown parameters. Even more, if the equation thus obtained has no solution for the remaining parameters, this provides the modeler a clear sign of inconsistency over the available parameter values. This can be very valuable in terms of localizing an inconsistency early on in the model building process, rather than carrying it through and aiming to identify it in the final model. The approach may even be used when the available knowledge of the parameters is approximate, in terms of a range of possible values rather than a precise value. This makes the model refinement approach a powerful technique to integrate partial and approximate knowledge into building a large model.

2.2. Model construction with Event-B

Event-B [10] is a formal method approach to modeling complex systems, which was developed from classical B [12] and action systems [13]. Its modeling language is based on set theory and first order logic. An Event-B model consists of two types of modules: *contexts* and *machines*. The context is the *static* part of model, and consists of *types*, *constants*, and *axioms*. The machine is the *dynamic* part of model, and may contain *variables*, *invariants* and *events*. The initial state of system is specified with a specific event called *Initialization*. 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.

Rodin [14] is an open source Eclipse-based tool, which provides the modeling and proving support in Event-B. Rodin implements a powerful logical engine making sure that the model is consistently written, both syntactically and semantically. If errors are detected, for example in variable names, or in event specifications, leading to undeclared variables or to invariants not holding, the source of the errors is precisely indicating. This is an excellent support for the modeler when writing large models such as the ones discussed in this study.

Here we recall briefly a general scheme to build a chemical reaction network in Event-B, discussed in details in Ref. [15] to build an Event-B model for the heat shock response. Each species of the reaction network is modeled 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 $2A \to B$ and $A + B \to 3A$, 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

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

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

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 consistent construction of the model. Rodin checks this automatically and flags up if an error occurred.

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

We have an event in Event-B for each of the reactions. For each event, the guards of the event make sure that there should be enough of each of the reactants 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 Ref. [15].

3. Case-study: the ErbB signaling pathway

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

The ErbB signaling 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, [9,16,17]. Due to its intrinsic complexity and association with the progression of various cancer types, the pathway was extensively analyzed and found to be key to the unremitting growth and development of carcinoma cells.

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 signaling 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 signaling molecules. There are two signaling 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) signaling cascade through the Raf, MEK and ERK kinases, see Refs. [7,8]. The effect brought about by signaling 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 Ref. [8].

The initial model, introduced in Ref. [7], 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 Refs. [8,18]. 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 Refs. [19,20]. Protein isoform specificity (multiple forms for the same protein) is not accounted for in the model in Ref. [7]. The system described by the model in Ref. [7] 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 Ref. [7].

The model in Ref. [7] accounts for a set of 12 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 Ref. [7]. We have imported the model in COPASI [11] and made it available at [21].

4. Results

4.1. Ensuring the consistency of the basic model

The basic model of [7] consists of 148 reactions and 103 reactants. Ensuring the consistency of such a large model is non-trivial as small errors in the variable names or in the stoichiometric coefficients are often only indicated by standard softwares such as COPASI after the model is fully implemented; in such cases it is typically difficult to trace the source of the error. We decided to use Event-B to check the consistency of the basic model. We took advantage of the Rodin platform's powerful logic-based engine to verify on-the-fly the consistency of each reaction. To do this, the Rodin platform uses the features of Event-B asking that each event (corresponding in our case to a reaction) specifies the pre-conditions of that event being enabled (in our case, that the variable corresponding to each reactant has a value at least as large as its reaction coefficient) and the effect of that event being triggered (in our case, the variables corresponding to the reactants and the products being updated with the corresponding stoichiometric coefficient). Any error in writing a variable name and/or a stoichiometric coefficient is indicated immediately and the source of the error indicated based on the local logical checkup applied to every event as it is being written. This was an important tool to make sure that the initial model we started from was consistent. To write the events corresponding to each of the model reactions, we applied the general scheme presented in Section 2.2. The resulting Event-B model can be downloaded at [21]. We only show here two of the events in Table 2.

4.2. The refined model: its species and reactions

We implemented the refinement-based extension of the EGFR signaling pathway model from Ref. [7] by distinguishing some of the details between the four receptor members of the ErbB family: ErbB1 (EGFR), ErbB2 (HER2), ErbB3, ErbB4. We also considered two types of

Table 2
Two events modeling the forward and reverse directions of the first reaction of the ErbB signaling pathway.

Rec1f
WHERE
@grd1 EGFR $\geq 1 \wedge$ EGF ≥ 1
THEN
@act1 EGFR := EGFR -1
@act2 EGF := EGF -1
@act3 EGF-EGFR := EGF-EGFR $+1$
END

Rec1r
WHERE
@grd1 EGF-EGFR ≥ 1
THEN
@act1 EGF-EGFR := EGF-EGFR -1
@act2 EGFR := EGFR $+1$
@act3 EGF := EGF $+1$
END

ligands: EGF and HRG. An overview of our refinement strategy is illustrated in the supplementary material.

The refined model comprises species divided in two categories: *atomic* or *complex*, see Ref. [22]. An atomic species represents a single protein (or gene, etc.), while a complex species is 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 Ref. [7] is refined. All complex species present in the model from Ref. [7] 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:

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

The entire signaling 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 Ref. [7] is the following:

$$\left.\text{EGF} + \text{EGFR}\right\} \underset{k_b^-}{\overset{k_b^+}{\rightleftharpoons}} \text{EGF} - \text{EGFR}. \tag{5}$$

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:

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

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

We need to fix the kinetic parameters of the refined model so that it is a fit-preserving refinement of the model in Ref. [7], i.e. the sufficient condition (4) for our refined model is met. One simple way to do this is by setting $k_{i,j}^- = k_b^-$ and $k_{i,j}^+ = k_b^+$, for all $i = \overline{1,2}$, $j = \overline{1,4}$. Other possibilities exist as well, including in the case some of these constants are to be taken from literature or from experimental data, rather than freely fixed. In such a case, the known values are used in equation (4) to inform the setting of the unknown values.

The ligand-binding receptor activation reaction is followed by a dimerization reaction in the basic model of [7]:

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

Since EGF − EGFR is refined to 8 different species, the left hand side of this dimerization reaction will have $8 + \binom{8}{2} = 36$ possible combinations in the refined model: 8 where the two dimerizing monomers are identical, and $\binom{8}{2}$ where they are different. We model the dimerization of the receptor-bound ligands to include only homo-dimers as products in the final model, for simplicity, as dimerization is one of the core reactions of the model, right at the beginning of the transduction pathway. This is done also so that the model we obtain is as close as possible to the one in Ref. [9] (that we aim to compare our model against). These dimers interact with a number of species, resulting into reactants formed as dimers bound to chains of atomic species, downstream the ErbB signaling pathway, whose structure is of the following form: $(L - R)_j^2 - AC$, where AC is the chain of bound atomic species. We include in dimerization monomers of different structures to emphasize the mathematical derivation of the refinement and satisfy its numerical constraints.

We obtain the following refined reactions:

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

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

$$\mathscr{B} = \{\text{EGF} - \text{ErbBp},\ \text{HRG} - \text{ErbBq} | p, q = \overline{1,4}\}\,.$$

The constraint (4) for ensuring fit-preservation translates to the following relations for the dimerization reaction:

$$\begin{aligned}&\textstyle\sum_{l=1}^{8} k_{i,j,l}^+ = \begin{cases} k_d^+, & \text{if } i = j;\\ 2k_d^+, & \text{if } i < j;\end{cases}\\ &\textstyle\sum_{1\leq i\leq j\leq 8} k_{i,j,l}^- = k_d^-.\end{aligned}$$

The total number of refined reactions for the dimerization is $36 \times 8 = 288$. In line with our choice to consider only homo-dimers $(\text{ L}- \text{R})_i^2$ or $(\text{ L}- \text{R})_j^2$, we set the kinetic rate constants as follows:

$$k_{i,j,l}^+ = \begin{cases} k_d^+, & \text{if } l = i \text{ or } l = j;\\ 0, & \text{otherwise;}\end{cases}$$

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

It is easy to check that this choice of values satisfy equation (4) and thus, we obtain a fit-preserving refinement. As observed before, other choices for setting these values are obviously also possible, but for simplicity we made this ‘symmetrical’ choice here.

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

$$(\text{EGF} - \text{EGFR})^2 \underset{k_f^-}{\overset{k_f^+}{\rightleftharpoons}} (\text{EGF} - \text{EGFR}^*)^2.$$

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

$$(\text{L}- \text{R})_l^2 \underset{k_l^-}{\overset{k_l^+}{\rightleftharpoons}} (\text{L}-\text{R}^*)_l^2,$$

where for any $k = \overline{1,8}$ the refined species $(\text{ L}- \text{R})_k$ are (in some arbitrary, fixed, order) the elements of the set $\mathscr{B}$ 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:

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

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

$$\overset{[k_p]}{\rightarrow} \text{EGFR},$$

Which is refined to:

$$\overset{[k_i]}{\rightarrow} R_i, \quad i = \overline{1,4},$$

with $R_i \in \{\text{ErbB1, ErbB2, ErbB3, ErbB4}\}$. The kinetic rate constants $k_1, ..., k_4$ for receptor production must comply (4) so we preserve the model fit: $k_1 + k_2 + k_3 + k_4 = k_p$. To set the values of the constants we can integrate experimental observations of [23] that reported a different number of tyrosine phosphorylation sites for the four receptors: 12 sites for ErbB1, 6 for ErbB2, 11 for ErbB3, and no information on the sites for ErbB4. The higher the number of phosphorylation sites, the higher the activation rate for the receptor. Accordingly, we may set the activation rates of the receptors proportional to their number of tyrosine phosphorylation sites (with ErbB4 set to having by default one such site). The combination of these two constraints (one mathematical, the other experimental) leads to the following choice of the kinetic rate constants:

$$k_1 = \frac{12}{30}k_p = \frac{2}{5}k_p, \quad k_2 = \frac{6}{30}k_p = \frac{1}{5}k_p, \quad k_3 = \frac{11}{30}k_p, \quad k_4 = \frac{1}{30}k_p.$$

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 such as $(\text{EGF}-\text{EGFR}^{*})^2 - \text{AC}$, where AC represents a chain of bound atomic species (e.g. GAP − Grb2 − Sos − Ras − GDP − Prot). This species is refined as follows:

$$(\text{EGF}-\text{EGFR}^{*})^2 - \text{AC} \rightarrow \{(\text{LRst})^2_j - \text{AC} | j = \overline{1,8}\}, \tag{7}$$

where $(\text{L}-\text{R})_j \in \mathscr{B}$ 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 $(\text{LRst})^2_j - \text{AC}$ are involved in reactions of the following types:

$$(\text{LR}^{*})^2_j - \text{AC} + AS \underset{k_g^-}{\overset{k_g^+}{\rightleftharpoons}} (\text{LR}^{*})^2_j - \text{AC} - \text{AS},$$

$$(\text{LR}^{*})^2_j - \text{AC} \overset{k_g}{\rightarrow} (\text{LR}^{*})^2_j - \text{AC} - \text{deg},$$

$$(\text{LR}^{*})^2_j - \text{AC} \overset{k_{g_i}}{\rightarrow} (\text{LR}^{*}_i)^2_j - \text{AC},$$

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 reaction in the initial model, 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. The obtained refined model can be found at [21].

4.3. The refined model: numerical setup

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 $(\text{EGF}-\text{EGFR}^{*})^2 - \text{AC}$. The initial concentration values should satisfy:

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

where $(\text{L}-\text{R})_j \in \mathscr{B}$, 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 favor any of the subspecies and thus assign equal values for the initial concentrations:

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

Obviously, any other choice that satisfies the refinement condition is just as good from the computational point of view. This flexibility allows the modeler to include some values that may be known from the literature and set the other, unknown values accordingly. The same applies for the numerical setup of all initial concentrations for the refined species.

The refined model was implemented in COPASI and it is available at [21].

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 Ref. [9] consisting of 499 reactants and 828 reactions, a good fit was obtained by running about 100 times annealing methods, over 24 h on a cluster consisting of 100 nodes.

Our approach to building that same model was to refine the model from Ref. [7], 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 [9]. Our approach proved to be effective in building the refined model, by preserving its fit without any supplementary parameter estimation. The development of the refined model required little domain specific knowledge, in contrast with the efforts of [9], where knowledge about each of the detailed refined reactions had to be explicitly formulated.

In the approach to model refinement proposed in Ref. [6], that we follow in our paper, the decision on how to refine the species of a model is captured abstractly through a refinement relation between the set of species of the basic model and that of the refined model. This is treated in an abstract way by only imposing some light constraints so that each basic species is related to at least one new species (no species is lost), and any new species is related to a unique basic species (no ambiguity in the refinement relation). The advantage of this approach is that it allows a general solution for the fit-preserving problem based on linear algebra, leading to equation (4). In practice, the refinement of a species is in fact dependent on the refinement of others. For example, the refinement of a chemical complex will be determined by the way its components are refined. This situation is left as a modeling decision to be made outside of our approach, yielding the refinement relation that the modeler wishes to adopt. We illustrated this in our case study. Once the relation is fixed, the approach of [6] offers an elegant solution to

how the model parameters should be fixed so that the model fit is preserved in the refinement. This problem has also been addressed in other modeling formalisms, in particular rule-based modeling, by having it as an integral part of the model refinement process, rather than as an external decision. This leads to considerations regarding the structure of species in a model, captured through graphs and site-graph rewriting, see e.g. Refs. [3,4,24,25]. The key advantage that our method brings is that partial experimental knowledge about the model can be integrated with the mathematical constraints, leading to a model that is guaranteed to remain fit to the data, while integrating available experimental observations. We demonstrated this in how we set up the activation rates of the ErbB receptors in connection with the number of their tyrosine phosphorylation sites and with the mathematical constraint for the data fit.

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.

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. This 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.

Acknowledgments

This work was partially supported by Academy of Finland under grant 267915.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.01.016.

References

[1] W. S. Hlavacek, How to deal with large models?, Mol. Syst. Biol. 5.
[2] R.-J. Back, J. von Wright, Refinement Calculus, Springer, New York, 1998.
[3] V. Danos, J. Feret, W. Fontana, R. Harmer, J. Krivine, Rule-based modelling, symmetries, refinements, in: J. Fisher (Ed.), Formal Methods in Systems Biology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 103–122.
[4] E. Murphy, V. Danos, J. Féret, J. Krivine, R. Harmer, Rule-based modeling and model refinement, in: H.M. Lodhi, S.H. Muggleton (Eds.), Elements of Computational Systems Biology, John Wiley & Sons, Ltd, 2010, pp. 83–114.
[5] B. Iancu, E. Czeizler, E. Czeizler, I. Petre, Quantitative refinement of reaction models, Int. J. Unconv. Comput. 8 (5–6) (2012) 529–550.
[6] C. Gratie, I. Petre, Complete characterization for the fit-preserving data refinement of mass-action reaction networks, Theor. Comput. Sci. 641 (2016) 11–24.
[7] J.J. Hornberg, B. Binder, F.J. Bruggeman, B. Schoeberl, R. Heinrich, H.V. Westerhoff, Control of MAPK signalling: from complexity to what really matters, Oncogene 24 (36) (2005) 5533–5542.
[8] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, G. Müller, Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized egf receptors, Nat. Biotechnol. 20 (2002) 370 (EP –).
[9] W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, P. K. Sorger, Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data, Mol. Syst. Biol. 5.
[10] J.-R. Abrial, Modeling in Event-B: System and Software Engineering, first ed., Cambridge University Press, New York, NY, USA, 2010.
[11] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, U. Kummer, Copasi—a complex pathway simulator, Bioinformatics 22 (24) (2006) 3067–3074.
[12] J.-R. Abrial, The B-Book: Assigning Programs to Meanings, Cambridge University Press, New York, NY, USA, 1996.
[13] R.-J. Back, R. Kurki-Suonio, Decentralization of process nets with centralized control, Proceedings of the Second Annual ACM Symposium on Principles of Distributed Computing, ACM, New York, NY, USA, 1983, pp. 131–142.
[14] J.-R. Abrial, M. Butler, S. Hallerstede, T.S. Hoang, F. Mehta, L. Voisin, Rodin: an open toolset for modelling and reasoning in event-b, Int. J. Software Tool. Technol. Tran. 12 (6) (2010) 447–466.
[15] U. Sanwal, L. Petre, I. Petre, Stepwise construction of a metabolic network in event-b, Comput. Biol. Med. 91 (C) (2017) 1–12.
[16] K. Oda, Y. Matsuoka, A. Funahashi, H. Kitano, A comprehensive pathway map of epidermal growth factor receptor signaling, Mol. Syst. Biol. 1 (1).
[17] M. R. Birtwistle, M. Hatakeyama, N. Yumoto, B. A. Ogunnaike, J. B. Hoek, B. N. Kholodenko, Liganddependent responses of the erbb signaling network: experimental and modeling analyses, Mol. Syst. Biol. 3 (1).
[18] B.N. Kholodenko, O.V. Demin, G. Moehren, J.B. Hoek, Quantification of short term signaling by the epidermal growth factor receptor, J. Biol. Chem. 274 (42) (1999) 30169–30181.
[19] L. Buday, P.H. Warne, J. Downward, Downregulation of the Ras activation pathway by MAP kinase phosphorylation of Sos, Oncogene 11 (7) (1995) 1327–1331.
[20] D. Chen, S.B. Waters, K.H. Holt, J.E. Pessin, Sos phosphorylation and disassociation of the grb2-sos complex by the erk and jnk signaling pathways, J. Biol. Chem. 271 (11) (1996) 6328–6332.
[21] COPASI and Event-B Models for the Control of MAPK Signalling, (2016) https://bit.ly/2qXwGye.
[22] D.-E. Gratie, B. Iancu, S. Azimi, I. Petre, From Action Systems to Distributed Systems, Taylor & Francis, 2016, Ch. Quantitative Model Refinement in Four Different Frameworks, with Applications to the Heat Shock Response, (2014), p. 201.
[23] R.B. Jones, A. Gordus, J.A. Krall, G. MacBeath, A quantitative protein interaction network for the erbb receptors using protein microarrays, Nature 439 (2005) 168 (EP –).
[24] A. Basso-Blandin, W. Fontana, R. Harmer, A knowledge representation meta-model for rule-based modelling of signalling networks, in: C.A. Muñoz, J.A. Pérez (Eds.), Developments in Computational Models (DCM 2015), Vol. 204 of Electronic Proceedings in Theoretical Computer Science, 2015, pp. 47–59.
[25] R. Harmer, Y.-S.L. Cornec, S. Légaré, I. Oshurko, Bio-curation for cellular signalling: the kami project, in: J. Feret, H. Koeppl (Eds.), Computational Methods in Systems Biology, Springer International Publishing, Cham, 2017, pp. 3–19.