Self-assembly models of variable resolution

Andrzej Mizera*[1], Eugen Czeizler†[2], and Ion Petre[1]

[1]Department of Information Technologies,
Åbo Akademi University, Finland
[2]Department of Information and Computer Science,
Aalto University, Finland

Abstract

Model refinement is an important aspect of the model-building process. It can be described as a procedure which, starting from an abstract model of a system, performs a number of refinement steps in result of which a more detailed model is obtained. At the same time, in order to be correct, the refinement mechanism has to be capable of preserving already proven systemic quantitative properties of the original model, e.g. model fit, stochastic semantics, etc. In this study we concentrate on the refinement in the case of self-assembly models. Self-assembly is a process in which a disordered ensemble of basic components forms an organized structure as a result of specific, local interactions among these components, without external guidance. We develop a generic formal model for this process and introduce a notion of model resolution capturing the maximum size up to which objects can be distinguished individually in the model. All bigger objects are treated homogenously in the model. We show how this self-assembly model can be systematically refined in such a way that its resolution can be increased and decreased while preserving the original model fit to experimental data, without the need for tedious, computationally expensive process of parameter refitting. We demonstrate how the introduced methodology can be applied to a previously published model: we consider the case-study of *in vitro* self-assembly of intermediate filaments.

1 Introduction

The great complexity of biological systems enforces the need for representing them in formal models in order to investigate them and make specific predictions about their behaviour that can be tested in subsequent experiments. Starting from a model abstracting a biological system, the iterative process of hypothesis generation, experimental design, experimental analysis, and model refinement lies at the core of systems biology ([4, 16, 22]). Even more, this approach is

*Address correspondence to: Andrzej Mizera, Department of Information Technologies, Åbo Akademi University, Joukahaisenkatu 3-5 A, FIN-20520 Turku, Finland. Phone: +358 2 215 4045; Fax: +358 2 215 4732; E-mail: amizera@abo.fi

†Work done while the author was at Åbo Akademi University.

proposed as the only logical way for biology to advance ([19]). Development and refinement of a mathematical model of a biochemical process proceeds, in general, in accordance with the following scenario. First, an abstraction of the process is made by identifying a relatively small set of biochemical reactions which are capturing the main features of the process' machinery. The chosen biochemical reactions may be very abstract themselves, i.e. one reaction may in fact encapsulate many real reactions which constitute a whole subprocess in a living organism. Second, the molecular model formed of the chosen reactions is transformed into an associated mathematical model. This usually involves two steps: obtaining equations describing the dynamics of the system by assuming some proper kinetic law, e.g. mass-action law, Michaelis-Menten kinetics, etc., and then identifying the model parameter values so that the model fits some experimental data.

During the process of model development some simplifications and abstractions are introduced. With time, there may be a necessity for them to be refined and modelled in a more detailed, accurate way. However, some carefulness is required on this stage. For example, one could take all the intended changes into consideration while simply repeating the whole model development procedure. But this solution involves repeating from scratch the time-consuming, computationally-intensive model fitting, see [5]. Another approach, not much investigated in the literature, is to refine the model in such a way that the previously obtained fit is preserved. This basically implies deriving the parameter values of the refined model from the ones of the original model.

In this study we concentrate on the step of model refinement in the iterative cycle of systems biology, which is an important aspect of the model-building process. In particular, we develop a refinement procedure for a family of ordinary differential equation (ODE) models describing the process of self-assembly. Self-assembly is a process in result of which some pronounced structures emerge out of an ensemble of scattered basic elements. Important is the fact that the arrangements take place based just on local interactions between the building blocks, without any external guidance. In our work we develop a generic formal model for self-assembly. It consists of an ensemble of all possible objects that can potentially appear in the course of the self-assembly, a composition operation and a mapping from objects of the ensemble to positive integers. The number is interpreted as the size of the considered object. The generic model allows us to further introduce the notion of model resolution. We continue by discussing the refinement of such models, i.e. we formally show how the resolution of a self-assembly model can be increased and decreased while preserving the original model fit to experimental data. We demonstrate how our methodology of self-assembly model refinement can be applied to an existing model. To this aim we utilize the previously published model of the *in vitro* assembly of intermediate filaments from tetrameric vimentin, see [6, 15].

Our methodology of self-assembly model refinement is a particular instance of *formal model refinement*. This topic has been extensively studied in Computer Science, see, e.g., [3, 23, 24], especially in connection to formal software specifications. The method we propose is an instance of *data refinement*, where one replaces a variable with a set of other variables in a way that introduces more details into the model, while keeping the model constraints unchanged.

The paper is organized as follows. First, a general, formal characterization of the self-assembly process is presented. Then, the notion of model resolution

is introduced and the model refinement procedure consisting in increasing and decreasing the model resolution while preserving the fit to experimental data is described. Finally, the technique is applied in a case study where the self-assembly of intermediate filaments is considered.

2 A generic model for self-assembly

Self-assembly is a term coined to name processes in which a disordered ensemble of basic components forms an organized structure as a result of specific, local interactions among these components, without external guidance. In a general case, the process of self-assembly can be formalized as follows. We consider an ensemble $\mathcal{E}$ of all possible objects that can potentially appear in the course of the self-assembly process, including the initial ones. Each object O from the ensemble has a scalar value $size(O)$ associated with it and determined through a mapping $size : \mathcal{E} \to \mathbb{N}_+$. Moreover, the objects from $\mathcal{E}$ can combine with each other to form another object from $\mathcal{E}$ in such a way that the sum of the sizes of the objects equals the size of the resulting object. More formally, if we denote the composition operation with + , then

$$O_1 + O_2 = O_r \quad \Rightarrow \quad size(O_1) + size(O_2) = size(O_r), \tag{1}$$

where O_r is the object assembled from component objects O_1 and O_2. The ensemble $\mathcal{E}$ together with the binary operation + forms a structure $(\mathcal{E}, +)$, which in abstract algebra is named a *semigroup*. Furthermore, this structure is homomorphic with the $(\mathbb{N}_+, +)$ semigroup by the *size* map.

Our generic model for self-assembly is on a high level of abstraction, focusing on the *size* of the emerging structures, while ignoring all details of the topology of such structures. *Size* here can mean any semigroup homomorphism between $(\mathcal{E}, +)$ and $(\mathbb{N}_+, +)$, as noted above. Intuitively, the *size* map would count the number of elementary blocks forming the self-assembled structure under consideration. This approach is applicable to any type of self-assembly processes: uni-dimensional (such as the elongation of intermediate filaments, the case-study investigated in this paper), branched two-dimensional structures, three-dimensional assemblies, etc. However, extending the dynamics of the *size* distribution of the self-assembled structures with some of their topological details would require a very different type of modelling, which goes beyond the scope of our approach.

Through the map *size*, for a fixed $n \in \mathbb{N}_+$ we define a family of object classes $\mathcal{S}^{(n)} = \{\mathcal{S}_1^{(n)}, \ldots, \mathcal{S}_n^{(n)}, \mathcal{S}_{\geq n+1}^{(n)}\}$: $\mathcal{S}_i^{(n)}$ contains all the objects from $\mathcal{E}$ with size i for $i = 1, \ldots, n$ and $\mathcal{S}_{\geq n+1}^{(n)}$ consists of all objects with size greater than n. Each object from $\mathcal{E}$ belongs to exactly one of these classes. Notice that for $m > n$ we have $\mathcal{S}_k^{(n)} = \mathcal{S}_k^{(m)}$ for all $k \in \{1, \ldots, n\}$.

The composition of objects in $\mathcal{E}$ is described by a system of rules. For the general characterization of self-assembly we will assume that the rules are at

the level of abstraction of $\mathcal{S}^{(n)}$, i.e. that the system of rules is of the form

$$\begin{aligned}
\mathcal{S}_i^{(n)} + \mathcal{S}_j^{(n)} &\to \mathcal{S}_{i+j}^{(n)}, && \text{for all } 1 \le i \le j \le n,\ i+j \le n;\\
\mathcal{S}_i^{(n)} + \mathcal{S}_j^{(n)} &\to \mathcal{S}_{\ge n+1}^{(n)}, && \text{for all } 1 \le i \le j \le n,\ i+j \ge n+1;\\
\mathcal{S}_i^{(n)} + \mathcal{S}_{\ge n+1}^{(n)} &\to \mathcal{S}_{\ge n+1}^{(n)}, && \text{for all } 1 \le i \le n;\\
\mathcal{S}_{\ge n+1}^{(n)} + \mathcal{S}_{\ge n+1}^{(n)} &\to \mathcal{S}_{\ge n+1}^{(n)}. &&
\end{aligned} \tag{2}$$

In the case of biochemical systems these rules are usually referred to as (biochemical) reactions and we will use this terminology in the following. The semantics of the reactions in the above form can be described as: an object from class $\mathcal{S}_i^{(n)}$ combines with an object from class $\mathcal{S}_j^{(n)}$ to form an object of class $\mathcal{S}_{i+j}^{(n)}$ if $i+j \le n$ or $\mathcal{S}_{\ge n+1}^{(n)}$ if $i+j \ge n+1$. Notice that any reaction of this form automatically satisfies the self-assembly condition (1).

In mathematical modelling it is common to associate a variable (understood as a function) $F : \mathbb{R}_+ \to \mathbb{R}_+$ with each of the sets in $\mathcal{S}^{(n)}$. We denote with $F_i^{(n)}$ the variable corresponding to the set $\mathcal{S}_i^{(n)}$ for $i \in \{1, \ldots, n, \ge n+1\}$. The value of the variable $F_i^{(n)}$ is interpreted as the concentration of objects from the associated set $\mathcal{S}_i^{(n)}$, present in the system undergoing self-assembly at a particular point in time. Further, we assume that the kinetics of the reactions is based on the *law of mass action* ([17]). This law is a mathematical model of reaction dynamics: it states that the reaction rate is proportional to the probability of collision of the reactants, while the probability itself is proportional to the product of concentrations of reactants raised to the number in which they enter the reaction ([17]). We use $k_{i,j}$, $1 \le i \le j \le n+1$ to denote the respective proportionality factor, the so-called *rate constant*, of the reaction with the left-hand side containing $\mathcal{S}_i^{(n)}$ (or $\mathcal{S}_{\ge n+1}^{(n)}$ if $i = n+1$) as one and $\mathcal{S}_j^{(n)}$ (or $\mathcal{S}_{\ge n+1}^{(n)}$ if $j = n+1$) as the other term. For example,

$$\mathcal{S}_2^{(n)} + \mathcal{S}_3^{(n)} \xrightarrow{k_{2,3}} \mathcal{S}_5^{(n)}$$

and

$$\mathcal{S}_2^{(n)} + \mathcal{S}_{\ge n+1}^{(n)} \xrightarrow{k_{2,n+1}} \mathcal{S}_{\ge n+1}^{(n)}.$$

The change of concentrations in time of the objects undergoing self-assembly can be described using ordinary differential equations (ODEs). By the law of mass action, the system of ODEs associated with the self-assembly system determined by the reactions in (2) is

$$\begin{cases}
\dfrac{\mathrm{d}F_i^{(n)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} k_{i,j}\, F_i^{(n)}\, F_j^{(n)}\, [i \ne j] - 2\, k_{i,i}\, {F_i^{(n)}}^2 - k_{i,n+1}\, F_i^{(n)}\, F_{\ge n+1}^{(n)} \\
\qquad\qquad + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} k_{j,i-j}\, F_j^{(n)}\, F_{i-j}^{(n)} \qquad \text{for all } 1 \le i \le n, \\
\dfrac{\mathrm{d}F_{\ge n+1}^{(n)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \le i \le j \le n,\\ i+j \ge n+1}} k_{i,j}\, F_i^{(n)}\, F_j^{(n)} - k_{n+1,n+1}\, {F_{\ge n+1}^{(n)}}^2,
\end{cases} \tag{3}$$

where $[\ldots]$ are used as the Iverson brackets ([14, 18]), i.e. $[i \neq j]$ is 1 if $i \neq j$ and 0 otherwise. The negative term in the equation for $\mathrm{d}F^{(n)}_{\geq n+1}/\mathrm{d}t$ originates from the last rule in (2), where two objects from the set $\mathcal{S}^{(n)}_{\geq n+1}$ combine to form a bigger object belonging to the same class. In consequence, in $\mathcal{S}^{(n)}_{\geq n+1}$ two objects are consumed and one is produced, thus the net result is that one object disappears from $\mathcal{S}^{(n)}_{\geq n+1}$.

3 A notion of model resolution

When considering the dynamics of the self-assembly process, one of the main concerns is the distribution of the number of components of different sizes in time. To this aim we introduce the notion of *model resolution* in the context of self-assembly. We say that *a self-assembly model is of resolution n* if it consists of the set of reactions describing the interactions between the classes of objects $\mathcal{S}^{(n)}_1, \ldots, \mathcal{S}^{(n)}_n, \mathcal{S}^{(n)}_{\geq n+1}$, i.e. the set of rules of the form in (2). The associated mathematical model (ODE-based or not), comprising variables $F^{(n)}_1, \ldots, F^{(n)}_n, F^{(n)}_{\geq n+1}$ is also referred to as an n-resolution model. Thus, the system in (3) is a self-assembly ODE model of resolution n. Intuitively, a self-assembly mathematical model is of resolution n if it allows for capturing the dynamics of the number (or concentration) of components that are exactly of size i, where $0 \leq i \leq n$.

In light of this definition the superscript (n) obtains a new meaning: it indicates the resolution of the considered model, i.e. $F^{(n)}_j$ determines the concentration of objects of size j in time in the model of resolution n and $\mathcal{S}^{(n)}_j$ refers to the class of objects of size j which appears in the set of reactions of the n-resolution self-assembly model. This will be useful when considering the relationships between models of various resolutions in the subsequent subsections.

When setting the resolution of our generic self-assembly model we effectively partition the set of possible emerging structures into two, depending on their size:

(i) the set of *visible assemblies* whose size is at most the resolution level, and

(ii) the set of *invisible assemblies* whose size is larger than the resolution level.

The self-assembly process can be modelled in all of its combinatorial details on the set of visible assemblies, including the assembly of all possible pairs of visible assemblies and even their disassembly (disassembly is however not covered in our case-study). For the invisible assemblies (size larger than the resolution level) we only specify a number of generic reactions covering their elongation. The idea here is that the details of the dynamics of such assemblies are beyond the scope (or beyond the experimental measuring capabilities) of our current model.

Choosing the resolution of a self-assembly model should be done in a careful way, so that it includes in its visible assemblies that part of the species space that is important for the model. Changing the resolution of a model may be needed during the modelling process, depending on the application. For example, a model of relatively low resolution may be enough in the early stage of the process, when no (or very few) assemblies of large size exist. Later on however,

as the size of the existing self-assembled structures grows, the modeller may need to increase the resolution level to be able to track the details of the interactions involving larger structures. We discuss in the next section a method to increase the model resolution in such a way that the model's numerical fit to experimental data is preserved. Note also that the resolution may be fixed *a priori* to a level that is higher than the number of available molecules, thus making the whole species space visible, with the price that the manipulation of the model (such as the model fit and validation) may become computationally expensive.

3.1 Increasing the model resolution while preserving the model fit

In this section we concentrate on the refinement in the case of the self-assembly models. The aim is to increase the range of component sizes for which the distribution in time is captured by the model, i.e. increase the model resolution, while preserving the data fit of the original model. In the context of the associated mathematical models, we say that a model of resolution $n+1$ is a *quantitative refinement* of a model of resolution n if and only if the following quantitative refinement conditions are satisfied:

$$F_i^{(n+1)}(t) = F_i^{(n)}(t), \quad 1 \leq i \leq n \tag{4}$$

and

$$F_{n+1}^{(n+1)}(t) + F_{\geq n+2}^{(n+1)}(t) = F_{\geq n+1}^{(n)}(t), \tag{5}$$

for all $t \geq 0$.

In the case of the self-assembly ODE models of the form in (3), the quantitative refinement from resolution n to $n+1$ involves appropriate setting of the rate constants and the initial values of the model of resolution $n+1$ given the rate constants and the initial values of the model of resolution n. We show in the following how this should be performed.

We start our considerations with the statement of a lemma concerning the existence and uniqueness of solutions of the self-assembly ODE system of any fixed resolution.

Lemma 1. *The system of ODEs for a self-assembly model of resolution n, where $n \in \mathbb{N}$, admits exactly one solution for any fixed initial condition.*

Proof. Let us rewrite (3) in the form

$$\mathbf{F}' = \mathcal{F}(\mathbf{F}),$$

where $\mathbf{F}(t) = [F_1^{(n)}(t), \ldots, F_n^{(n)}(t), F_{\geq n+1}^{(n)}(t)]^{\mathrm{T}}$ and $\mathcal{F} : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ defines a vector field on $\mathbb{R}^{n+1}$. A solution of this system is a function $\mathbf{F} : J \to \mathbb{R}^{n+1}$ defined on some interval $J \subset \mathbb{R}$ such that, for all $t \in J$, $\mathbf{F}'(t) = \mathcal{F}(\mathbf{F}(t))$. Now, it is enough to observe that the right-hand sides of the equations in (3) are continuously differentiable with respect to the coordinates of $\mathbf{F}$. Thus, the mapping $\mathcal{F}$ is Lipschitz continuous on a bounded domain ([8]) and by the Picard-Lindelöf theorem ([8]) it follows that for any initial conditions the considered system has a unique solution $\mathbf{F}(t)$. □

Equipped with Lemma 1, we continue to show how the refinement of a self-assembly model can be effectively achieved. This is the content of the following theorem, where $l_{i,j}$, $1 \leq i \leq j \leq n+2$ denote the rate constants of the $(n+1)$-resolution model and $k_{p,q}$, $1 \leq p \leq q \leq n+1$ are the rate constants of the n-resolution model. A discussion about the biological basis for the numerical choices made in Theorem 1 is included after its proof.

Theorem 1. *Setting the kinetic rate constants of the $(n+1)$-resolution model in the following way*

$$\begin{cases} l_{i,j} \quad := k_{i,j} & 1 \leq i \leq j \leq n, \\ l_{i,n+1} := k_{i,n+1} & 1 \leq i \leq n, \\ l_{i,n+2} := k_{i,n+1} & 1 \leq i \leq n, \\ l_{n+1,n+2} := 2\, k_{n+1,n+1}, & \\ l_{n+1,n+1} := k_{n+1,n+1}, & \\ l_{n+2,n+2} := k_{n+1,n+1}, & \end{cases} \tag{6}$$

and its initial values so that they satisfy

$$F_i^{(n+1)}(0) = F_i^{(n)}(0), \quad 1 \leq i \leq n, \tag{7}$$

$$F_{n+1}^{(n+1)}(0) + F_{\geq n+2}^{(n+1)}(0) = F_{\geq n+1}^{(n)}(0) \tag{8}$$

ensures that the self-assembly ODE model of resolution $n+1$ is a quantitative refinement of the self-assembly ODE model of resolution n.

Proof. Let us write the system of ODEs for the model of resolution $n+1$:

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\, [i \neq j] - 2\, l_{i,i}\, {F_i^{(n+1)}}^2 \\ \qquad\qquad - l_{i,n+1}\, F_i^{(n+1)}\, F_{n+1}^{(n+1)} - l_{i,n+2}\, F_i^{(n+1)}\, F_{\geq n+2}^{(n+1)} \\ \qquad\qquad + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} l_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \qquad \text{for all } 1 \leq i \leq n, \\ \dfrac{\mathrm{d}F_{n+1}^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{j,n+1}\, F_j^{(n+1)}\, F_{n+1}^{(n+1)} - 2\, l_{n+1,n+1}\, {F_{n+1}^{(n+1)}}^2 \\ \qquad\qquad - l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{\geq n+2}^{(n+1)} + \displaystyle\sum_{j=1}^{\lceil \frac{n}{2} \rceil} l_{j,n+1-j}\, F_j^{(n+1)}\, F_{n+1-j}^{(n+1)} \\ \dfrac{\mathrm{d}F_{\geq n+2}^{(n+1)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \leq i \leq j \leq n, \\ i+j \geq n+2}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} + \sum_{j=1}^{n} l_{j,n+1}\, F_j^{(n+1)}\, F_{n+1}^{(n+1)} \\ \qquad\qquad + l_{n+1,n+1} {F_{n+1}^{(n+1)}}^2 - l_{n+2,n+2} {F_{\geq n+2}^{(n+1)}}^2. \end{cases} \tag{9}$$

Let us further denote by $G^{(n+1)}$ the sum of $F_{n+1}^{(n+1)}$ and $F_{\geq n+2}^{(n+1)}$, i.e.

$$G^{(n+1)}(t) = F_{n+1}^{(n+1)}(t) + F_{\geq n+2}^{(n+1)}(t).$$

With use of the expressions for $\mathrm{d}F_{n+1}^{(n+1)}/\mathrm{d}t$ and $\mathrm{d}F_{\geq n+2}^{(n+1)}/\mathrm{d}t$ in (9), we can compute the derivative of $G^{(n+1)}$

$$\begin{aligned}\frac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} &= \frac{\mathrm{d}F_{n+1}^{(n+1)}}{\mathrm{d}t} + \frac{\mathrm{d}F_{\geq n+2}^{(n+1)}}{\mathrm{d}t} = \\ &= \sum_{i=1}^{\lceil \frac{n}{2} \rceil} l_{i,n+1-i}\, F_i^{(n+1)}\, F_{n+1-i}^{(n+1)} + \sum_{\substack{1\leq i\leq j\leq n,\\ i+j\geq n+2}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} \\ &- l_{n+1,n+1}\, {F_{n+1}^{(n+1)}}^2 - l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{\geq n+2}^{(n+1)} - l_{n+2,n+2} {F_{\geq n+2}^{(n+1)}}^2.\end{aligned} \tag{10}$$

By substituting the rate constants in the above expression for $\mathrm{d}G^{(n+1)}/\mathrm{d}t$ in accordance with (6) we obtain that

$$\begin{aligned}\frac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} &= \sum_{\substack{1\leq i\leq j\leq n,\\ i+j\geq n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})^2 = \\ &= \sum_{\substack{1\leq i\leq j\leq n,\\ i+j\geq n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\, {G^{(n+1)}}^2.\end{aligned} \tag{11}$$

Now, by substituting the rate constants also in the equations for $\mathrm{d}F_i^{(n+1)}/\mathrm{d}t$ in (9) for all $1 \leq i \leq n$ and combing with (11) we have that

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\, [i \neq j] - 2\, k_{i,i}\, {F_i^{(n+1)}}^2 \\ \qquad\qquad - k_{i,n+1}\, F_i^{(n+1)}\, G^{(n+1)} + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} k_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \\ \qquad\qquad \text{for all } 1 \leq i \leq n, \\ \dfrac{\mathrm{d}G^{(n+1)}}{\mathrm{d}t} = \displaystyle\sum_{\substack{1\leq i\leq j\leq n,\\ i+j\geq n+1}} k_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} - k_{n+1,n+1}\, {G^{(n+1)}}^2. \end{cases} \tag{12}$$

The above system is identical with (3) modulo the renaming of variables, i.e. $F_i^{(n+1)}$ is in place of $F_i^{(n)}$ for all $1 \leq i \leq n$ and $G^{(n+1)}$ is in place of $F_{\geq n+1}^{(n)}$. Hence, if the initial values are set up as stated in the theorem, then (3) and (12) constitute the same initial value problem. By the existence and uniqueness stated in Lemma 1, there exists exactly one solution to this problem and thus we have that $F_i^{(n)}(t) = F_i^{(n+1)}(t)$ for all $1 \leq i \leq n$ and $G^{(n+1)}(t) = F_{n+1}^{(n+1)}(t) + F_{\geq n+2}^{(n+1)}(t) = F_{\geq n+1}^{(n+1)}(t)$. □

Notice that what is important for the refinement is that the initial values of the $(n+1)$-resolution model satisfy (8), however how the initial value of $F_{\geq n+1}^{(n)}$ is split into $F_{n+1}^{(n+1)}(0)$ and $F_{\geq n+2}^{(n+1)}(0)$ is irrelevant, i.e. any partition of this value in accordance with (8) leads to a quantitative refinement of the model of resolution n into a model of resolution $n+1$.

The choice of the kinetic rate constants in Theorem 1 for the refined model is consistent with the following basic principle:

by distinguishing several subtypes of a reactant, we do not change the kinetics of the reactions they participate in.

In other words, whenever we distinguish several subspecies A_1, A_2, ..., A_m of a species A, we consider in the refined model that each subspecies A_i participates in the same reactions in which A was participating in the original model and moreover, their kinetics is unchanged. (Extra biological knowledge about kinetic differences among A_1, ..., A_m may be included in the model in a subsequent step; we only focus here on setting up the more detailed model as a quantitative refinement of the original model.) Our reasoning about the model refinement is discrete, in terms of a finite number of subspecies of a given species. Consequently, our reasoning about the reaction kinetics and its changes is also discrete, in terms of collision-based reactions.

When seen as the result of a collision between the reactants, the kinetics of a reaction depends on a biochemical constant (whose value depends on the specifics of the reactants and of the environment) and on the number of possible combinations of reactant molecules, see [9, 10] for a detailed presentation of this approach. The number of such combinations in the case of a collision $A + B$ (say, type 1) is $[A] \cdot [B]$, but in the case of a collision $A + A$ (say, type 2), it is $[A] \cdot ([A] - 1)/2$, where $[A]$, $[B]$ denote the number of molecules of species A and B, respectively. This is the fundamental reason why $l_{n+1,n+2}$ is set in Theorem 1 to a value that is twice as large as the kinetic rate constant of its corresponding reaction in the original model. Indeed, reaction

$$\mathcal{S}^{(n)}_{\geq n+1} + \mathcal{S}^{(n)}_{\geq n+1} \xrightarrow{k_{n+1,n+1}} \mathcal{S}^{(n)}_{\geq n+1} \tag{13}$$

is replaced in the refined model with reactions

$$\mathcal{S}^{(n+1)}_{n+1} + \mathcal{S}^{(n+1)}_{n+1} \xrightarrow{l_{n+1,n+1}} \mathcal{S}^{(n+1)}_{\geq n+2}, \tag{14}$$

$$\mathcal{S}^{(n+1)}_{n+1} + \mathcal{S}^{(n+1)}_{\geq n+2} \xrightarrow{l_{n+1,n+2}} \mathcal{S}^{(n+1)}_{\geq n+2}, \tag{15}$$

$$\mathcal{S}^{(n+1)}_{\geq n+2} + \mathcal{S}^{(n+1)}_{\geq n+2} \xrightarrow{l_{n+2,n+2}} \mathcal{S}^{(n+1)}_{\geq n+2}. \tag{16}$$

When reasoning about the kinetic rate constants of the refined reactions, we preserve the same biochemical constants as in the case of the original reaction (no changes in the biochemical details of the subspecies as compared to the original species, as formulated in our basic principle). The number of combinations of reactants in the various reactions is however different: whereas reactions (13), (14), and (16) are of type 2 (as defined above), reaction (15) is of type 1. If we chose a discrete mathematical model formulation in terms of stochastic processes, then the kinetic rate constants of reactions (14)-(16) would be set to be equal to that of reaction (13). Translating such a model into a continuous, ODE-based model involves a change in the kinetic rate constants, where that of reaction (15) is set to twice that of reactions (13), (14), and (16) to account for the different way of reasoning about collisions in discrete and in continuous terms. Indeed, an ODE-based model considers the kinetic of a reaction of type 2 to be proportional to $[A]^2$, unlike in the case of a discrete model, where it is proportional to $[A] \cdot ([A] - 1)/2$. We refer to [9] for a detailed discussion on the relationship between the stochastic and the deterministic version of a biomodel. We also note that similar choices for the kinetic rate constants were made in [7]

when dealing with the refinement of rule-based models. Finally, we remark that the calculations in the proof of Theorem 1 show that our choice of kinetic rate constants, justified by the biochemical arguments above, lead to a numerically-correct quantitative model refinement.

Now, let us consider a more general case, namely the refinement of a model of resolution n to a model of resolution $n+m$. In this case the refinement conditions that need to be satisfied for all $t \geq 0$ are the following:

$$F_i^{(n+m)}(t) = F_i^{(n)}(t), \quad 1 \leq i \leq n$$

and

$$\sum_{j=1}^{m} F_{n+j}^{(n+m)}(t) + F_{\geq n+m+1}^{(n+m)}(t) = F_{\geq n+1}^{(n)}(t).$$

We start our considerations by a simple lemma.

Lemma 2. *The property of a self-assembly ODE model to be the quantitative refinement of another model of lower resolution is transitive, i.e. if the model $\mathcal{M}^{(n+m)}$ of resolution $n+m$ is the refined version of the model $\mathcal{M}^{(n)}$ of resolution n and $\mathcal{M}^{(n+m+k)}$ of resolution $n+m+k$ is the refined version of the model $\mathcal{M}^{(n+m)}$, then $\mathcal{M}^{(n+m+k)}$ is a quantitative refinement of $\mathcal{M}^{(n)}$, where n, m, k are positive integers.*

Proof. By the refinement conditions we have that for all $t \geq 0$

$$\begin{cases} F_i^{(n)}(t) = F_i^{(n+m)}(t), \quad 1 \leq i \leq n, \\ \sum_{i=1}^{m} F_{n+i}^{(n+m)}(t) + F_{\geq n+m+1}^{(n+m)}(t) = F_{\geq n+1}^{(n)(t)} \end{cases}$$

and

$$\begin{cases} \forall_{1 \leq i \leq n+m} \quad F_i^{(n+m)}(t) = F_i^{(n+m+k)}(t), \\ \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = F_{\geq n+m+1}^{(n+m)(t)}. \end{cases}$$

This implies that

$$F_i^{(n)}(t) = F_i^{(n+m+k)}(t), \quad 1 \leq i \leq n$$

and

$$\begin{aligned} F_{\geq n+1}^{(n)}(t) &= \sum_{i=1}^{m} F_{n+i}^{(n+m)}(t) + \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = \\ &= \sum_{i=1}^{m} F_{n+i}^{(n+m+k)}(t) + \sum_{i=1}^{k} F_{n+m+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t) = \\ &= \sum_{i=1}^{m+k} F_{n+i}^{(n+m+k)}(t) + F_{\geq n+m+k+1}^{(n+m+k)}(t). \end{aligned}$$

Thus it follows that the model of resolution $n+m+k$ constitutes a refinement of the model of resolution n. □

In the next theorem we show how the quantitative refinement of the model of resolution n to the one of resolution $n+m$ can be effectively achieved. We denote by $l_{i,j}$, $1 \leq i \leq j \leq n+m+1$ the rate constants of the $(n+m)$-resolution self-assembly model $\mathcal{M}^{(n+m)}$ and by $k_{p,q}$, $1 \leq p \leq q \leq n+1$ the ones of the n-resolution self-assembly model $\mathcal{M}^{(n)}$.

Theorem 2. *Setting the kinetic rate constants of the $(n+m)$-resolution self-assembly ODE model $\mathcal{M}^{(n+m)}$ in accordance with the rate constants of the n-resolution self-assembly ODE model $\mathcal{M}^{(n)}$ in the following way*

$$\begin{cases} l_{i,j} := k_{i,j} & 1 \le i \le j \le n+1, \\ l_{i,n+j} := k_{i,n+1} & 1 \le i \le n,\ 2 \le j \le m+1, \\ l_{n+i,n+i} := k_{n+1,n+1} & 2 \le i \le m+1, \\ l_{n+i,n+j} := 2\,k_{n+1,n+1} & 1 \le i < j \le m+1, \end{cases} \tag{17}$$

and its initial values so that they satisfy

$$F_i^{(n+m)}(0) = F_i^{(n)}(0), \quad 1 \le i \le n, \tag{18}$$

$$\sum_{i=1}^{m} F_{n+i}^{(n+m)}(0) + F_{\ge n+m+1}^{(n+m)}(0) = F_{\ge n+1}^{(n)}(0) \tag{19}$$

ensures that $\mathcal{M}^{(n+m)}$ is a quantitative refinement of $\mathcal{M}^{(n)}$.

Proof. The proof is by induction on m. The basis of the induction which is the step from resolution n to $n+1$ ($m=1$) is given by Theorem 1. The statement of Theorem 2 clearly holds in this case and we proceed to the inductive step. We assume that the statement is true for $m = z$ for some $z \ge 2$ and we consider the case where $m = z+1$. Theorem 1 assures that setting

$$\begin{cases} l_{i,j}^{(n+z+1)} := l_{i,j}^{(n+z)} & 1 \le i \le j \le n+z, \\ l_{i,n+z+1}^{(n+z+1)} := l_{i,n+z+1}^{(n+z)} & 1 \le i \le n+z, \\ l_{i,n+z+2}^{(n+z+1)} := l_{i,n+z+1}^{(n+z)} & 1 \le i \le n+z, \\ l_{n+z+1,n+z+1}^{(n+z+1)} := l_{n+z+1,n+z+1}^{(n+z)}, & \\ l_{n+z+1,n+z+2}^{(n+z+1)} := 2\,l_{n+z+1,n+z+1}^{(n+z)}, & \\ l_{n+z+2,n+z+2}^{(n+z+1)} := l_{n+z+1,n+z+1}^{(n+z)} & \end{cases} \tag{20}$$

and the initial values of $F_{n+z+1}^{(n+z+1)}$ and $F_{\ge n+z+2}^{(n+z+1)}$ in such a way that

$$F_{n+z+1}^{(n+z+1)}(0) + F_{\ge n+z+2}^{(n+z+1)}(0) = F_{\ge n+z+1}^{(n+z)}(0) \tag{21}$$

is satisfied results in a refinement from the self-assembly model $\mathcal{M}^{(n+z)}$ of resolution $n+z$ to the model $\mathcal{M}^{(n+z+1)}$ of resolution $n+z+1$ (the subscripts of the kinetic rate constants in (20) indicate the reactions and the superscripts the models in terms of their resolution). By the induction hypothesis setting

$$\begin{cases} l_{i,j}^{(n+z)} := k_{i,j} & 1 \le i \le j \le n+1, \\ l_{i,n+j}^{(n+z)} := k_{i,n+1} & 1 \le i \le n,\ 2 \le j \le z, \\ l_{n+i,n+i}^{(n+z)} := k_{n+1,n+1} & 2 \le i \le z, \\ l_{n+i,n+j}^{(n+z)} := 2\,k_{n+1,n+1} & 1 \le i \le j \le z, \\ l_{i,n+z+1}^{(n+z)} := k_{i,n+1} & 1 \le i \le n, \\ l_{n+i,n+z+1}^{(n+z)} := 2\,k_{n+1,n+1} & 1 \le i \le z, \\ l_{n+z+1,n+z+1}^{(n+z)} := k_{n+1,n+1} & \end{cases} \tag{22}$$

and the initial values of $F_{n+i}^{(n+z)}$ and $F_{\ge n+z+1}^{(n+z)}$, where $1 \le i \le z$ in such a way that

$$\sum_{i=1}^{z} F_{n+i}^{(n+z)}(0) + F_{\ge n+z+1}^{(n+z)}(0) = F_{\ge n+1}^{(n)}(0) \tag{23}$$

is satisfied gives a refinement of $\mathcal{M}^{(n)}$ to $\mathcal{M}^{(n+z)}$. Combining (20) with (22) results in

$$l_{i,j}^{(n+z+1)} := k_{i,j} \qquad 1 \le i \le j \le n+1, \tag{24}$$

$$l_{i,n+j}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n,\ 2 \le j \le z, \tag{25}$$

$$l_{n+i,n+i}^{(n+z+1)} := k_{n+1,n+1} \qquad 2 \le i \le z, \tag{26}$$

$$l_{n+i,n+j}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i < j \le z, \tag{27}$$

$$l_{i,n+z+1}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n, \tag{28}$$

$$l_{n+i,n+z+1}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i \le z, \tag{29}$$

$$l_{i,n+z+2}^{(n+z+1)} := k_{i,n+1} \qquad 1 \le i \le n, \tag{30}$$

$$l_{n+i,n+z+2}^{(n+z+1)} := 2\,k_{n+1,n+1} \qquad 1 \le i \le z, \tag{31}$$

$$l_{n+z+1,n+z+1}^{(n+z+1)} := k_{n+1,n+1}, \tag{32}$$

$$l_{n+z+1,n+z+2}^{(n+z+1)} := 2\,k_{n+1,n+1}, \tag{33}$$

$$l_{n+z+2,n+z+2}^{(n+z+1)} := k_{n+1,n+1}. \tag{34}$$

Putting together (25), (28) and (30) gives $l_{i,n+j}^{(n+z+1)} := k_{i,n+1}$ for $1 \le i \le n$ and $2 \le j \le z+2$; combining (26), (32) and (34) results in $l_{n+i,n+i}^{(n+z+1)} := k_{n+1,n+1}$ for $2 \le i \le z+2$; finally, (27), (29), (31) and (33) can be simply written as $l_{n+i,n+j}^{(n+z+1)} := 2\,k_{n+1,n+1}$ for $1 \le i \le j \le z+2$. Together with (24) this coincides with (17). Moreover, (23) together with (21) gives (19). By Lemma 2, since $\mathcal{M}^{(n+z)}$ refines $\mathcal{M}^{(n)}$ and $\mathcal{M}^{(n+z+1)}$ refines $\mathcal{M}^{(n+z)}$, we have that $\mathcal{M}^{(n+z+1)}$ is a refinement of $\mathcal{M}^{(n)}$. This proves the induction hypothesis. □

3.2 Decreasing the model resolution while preserving the model fit

Let us now consider the reverse problem. Given a self-assembly model of certain resolution, say $n+1$, we want to obtain a self-assembly model of resolution n such that the model of resolution $n+1$ constitutes its quantitative refinement. We refer to this problem as the problem of decreasing model resolution. As in the case of increasing model resolution, the ODE systems of these two models are (3) and (9). However, now the known rate constants are the ones of the model of resolution $n+1$, i.e. $l_{i,j}$ for all $1 \le i \le j \le n+2$, and the task is to set appropriately the values of the rate constants $k_{i,j}$, $1 \le i \le j \le n+1$ of the model of resolution n.

In this presentation we restrict our considerations to the particular case where $k_{i,j} := l_{i,j}$ for all $1 \le i \le j \le n$. This is in accordance with the biological motivation of the model: species that were modelled explicitly in the original model and continue to be so in the new model should not see their kinetics changed. From a mathematical point of view, one could also consider a general approach where the constants $k_{i,j}$, $1 \le i \le j \le n$ are part of the unknowns. In this case, a similar approach would be applicable, leading however to more complicated equations.

We investigate how to set the remaining constants, i.e. $k_{i,n+1}$, $1 \leq i \leq n+1$, so that the quantitative refinement conditions are satisfied. Since we want for the two models to satisfy (4) and (5), based on (3) and the fact that $k_{i,j} := l_{i,j}$ for all $1 \leq i \leq j \leq n$ the derivatives of $F_i^{(n+1)}$, $1 \leq i \leq n$ and $(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})$ can be expressed as

$$\begin{cases} \dfrac{\mathrm{d}F_i^{(n+1)}}{\mathrm{d}t} = -\displaystyle\sum_{j=1}^{n} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)}\, [i \neq j] - 2\, l_{i,i}\, {F_i^{(n+1)}}^2 \\ \qquad\qquad - k_{i,n+1}\, F_i^{(n+1)}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}) + \displaystyle\sum_{j=1}^{\lceil \frac{i-1}{2} \rceil} l_{j,i-j}\, F_j^{(n+1)}\, F_{i-j}^{(n+1)} \\ \qquad\qquad \text{for all } 1 \leq i \leq n, \\ \dfrac{\mathrm{d}(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})}{\mathrm{d}t} = \displaystyle\sum_{\substack{1 \leq i \leq j \leq n, \\ i+j \geq n+1}} l_{i,j}\, F_i^{(n+1)}\, F_j^{(n+1)} \\ \qquad\qquad - k_{n+1,n+1}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})^2. \end{cases}$$

Now, we equalize the right-hand sides in the above system with the respective right-hand sides in the model of resolution $n+1$, i.e. (9), where the expressions for the derivatives of $F_{n+1}^{(n+1)}$ and $F_{\geq n+2}^{(n+1)}$ are added up to obtain an expression for $\mathrm{d}(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})/\mathrm{d}t$. After simplifying we obtain that the rate constants $k_{i,n+1}$, $1 \leq i \leq n+1$ have to satisfy

$$\begin{gathered} l_{i,n+1}\, F_i^{(n+1)}\, F_{n+1}^{(n+1)} + l_{i,n+2}\, F_i^{(n+1)}\, F_{\geq n+2}^{(n+1)} \\ = \\ k_{i,n+1}\, F_i^{(n+1)}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}) \end{gathered} \tag{35}$$

and

$$\begin{gathered} l_{n+1,n+1}\, {F_{n+1}^{(n+1)}}^2 + l_{n+1,n+2}\, F_{n+1}^{(n+1)}\, F_{n+2}^{(n+1)} + l_{n+2,n+2}\, {F_{\geq n+2}^{(n+1)}}^2 \\ = \\ k_{n+1,n+1}\, (F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})^2 \end{gathered} \tag{36}$$

independently of time, i.e. at any time point t, where $t \geq 0$. We do not reduce (35) by dividing its sides by $F_i^{(n+1)}$ since the variable for a particular i may be identically zero. In such case the rate constant $k_{i,n+1}$ can admit an arbitrary value. At the same time we notice that if for all $1 \leq i \leq n$ the variables $F_i^{(n+1)}$ are not identically zero, then such reduction can be done without loss of generality and in this case all $k_{i,n+1}$ admit the same value.

The variables $F_i^{(n+1)}$s are in fact functions of time which constitute a solution to the system of nonlinear, first-order differential equations in (9). Having the explicit solutions, one could easily check whether there exist $k_{i,n+1}$, $1 \leq i \leq n+1$ such that (35) and (36) are satisfied at any time point $t \geq 0$. However, to the best of our knowledge, obtaining an analytical solution to (9) in a general case, i.e. for arbitrary n, is infeasible. Thus, we consider numerical integration of the system and propose the following procedure for checking whether the reduction of resolution in the discussed case can be performed and, if yes, how the rate

constants should be set. First, we numerically integrate the ODE system for the model of resolution $n+1$ in (9) to identify all i, $1 \leq i \leq n$, for which the product $F_i^{(n+1)}\,(F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)})$ is identically zero. In all these cases any arbitrary value of the rate constant $k_{i,n+1}$ satisfies (35). For the remaining is we pick a time point at which the product is non-zero and simply solve (35) for $k_{i,n+1}$ at the chosen time point. Similarly, we solve (36) for the value of $k_{n+1,n+1}$ at a time point at which $F_{n+1}^{(n+1)} + F_{\geq n+2}^{(n+1)}$ is non-zero. Second, in order to be correct, the values of the rate constants have to satisfy the refinement conditions without exception at any arbitrary time point. The correctness can be checked numerically by setting the initial values of the n-resolution model as follows

$$\begin{cases} F_i^{(n)}(0) := F_i^{(n+1)}(0), \quad 1 \leq i \leq n, \\ F_{\geq n+1}^{(n)}(0) := F_{n+1}^{(n+1)}(0) + F_{\geq n+2}^{(n+1)}(0) \end{cases}$$

and investigating whether the dynamics of the two considered models satisfy (4) and (5). The numerical check provides the ultimate answer whether the resolution decrease is realizable or not in the discussed case. Notice that if the values of the rate constants of the model of resolution $n+1$, say $\mathcal{M}^{(n+1)}$, are such that $l_{n+1,n+1} = l_{n+2,n+2}$, $l_{n+1,n+2} = 2\,l_{n+1,n+1}$ and $l_{i,n+1} = l_{i,n+2}$, for all $1 \leq i \leq n$, then the decrease of resolution can be simply achieved by changing the sides of the assignments in (6). In particular, if $\mathcal{M}^{(n+1)}$ were the result of applying Theorem 1 to a model of resolution n $\mathcal{M}^{(n)}$, then this way of decreasing the resolution of $\mathcal{M}^{(n+1)}$ recovers $\mathcal{M}^{(n)}$.

4 A case study: the self-assembly of intermediate filaments

One of the characteristics of eukaryotic cells is the existence of the cytoskeleton – an intricate network of protein filaments that extends throughout the cytoplasm. It enables the cells to adopt a variety of shapes, interact mechanically with the environment, organize the many components in their interior, carry out coordinated and directed movements. It also provides the machinery for intracellular movements, e.g. transport of organelles in the cytoplasm and the segregation of chromosomes at mitosis ([1, 2]). There are three kinds of protein filaments that form the cytoskeleton: actin filaments, intermediate filaments (IFs) and microtubules. Each kind has different mechanical properties and is assembled from an individual type of proteins. Actin filaments and microtubules are formed from *globular* proteins (*actin* and *tubulin* subunits, respectively), whereas *fibrous proteins* are the building blocks of intermediate filaments ([2, 11]). Thousands of these basic elements assemble into a construction of girders and ropes that spreads throughout the cell.

One of the main functions of intermediate filaments is to provide cells with mechanical strength and they are especially prominent in the cytoplasm of cells that are exposed to such conditions. For example, IFs are abundantly present along nerve cells axons where they provide crucial internal reinforcement of these long cell extensions. They can also be observed in great number in muscle cells and epithelial cells. IFs are characterized by great tensile strength. By stretching and distributing the effect of locally applied forces, they protect cells

and their membranes against breaking due to mechanical shear. Compared with microtubules and actin filaments, IFs are more stable, tough and durable, e.g. remain intact during exposure of cells to salt solutions and nonionic detergents, while the rest of the cytoskeleton is mostly destroyed ([1]).

Intermediate filaments can be grouped into four classes: (1) *keratin filaments* in epithelial cells; (2) *vimentin filaments* in connective-tissue cells, muscle cells and supporting cells of the nervous system; (3) *neurofilaments* in nerve cells; and (4) *nuclear lamins*, which strengthen the nuclear membrane of all eukaryotic cells, see [1]. In [15] a quantitative kinetic model for the *in vitro* self-assembly of intermediate filaments from tetrameric vimentin was considered. The authors introduced two molecular models (the so-called *simple* and *extended* models) of this process. In general, the *in vitro* assembly of vimentin IF proteins can be described as a process consisting of three major phases: (i) formation of the unit-length filaments (ULFs); (ii) longitudinal annealing of ULFs and growing filaments; (iii) radial compaction of immature (16 nm diameter) filaments into mature (11 nm diameter) IFs ([12, 13]). However, in both models of [15] the last, third phase was excluded from consideration.

In the case of the simple model from [15], ULFs are treated as ordinary filaments. Moreover, as discussed in [6, 15], the extension of filaments with tetramers plays an insignificant numerical role. This correlates with an experimental observation that *in vitro*, starting from an initial pool of tetramers, tetramers quickly turn into ULFs. Thus, the filament elongation by tetramers is inhibited in the beginning by the lack of filaments and later by the lack of free tetramers. In consequence, the assembly process is described through the following sequence of molecular events:

$$\begin{gathered} 2\,T_1 \xrightarrow{k_1} T_2 \qquad 2\,T_2 \xrightarrow{k_2} T_4 \qquad 2\,T_4 \xrightarrow{k_3} T_{\geq 8} \\ 2\,T_{\geq 8} \xrightarrow{k_f} T_{\geq 8} \end{gathered} \tag{37}$$

where T_1 is interpreted as a tetramer, T_2 as an octamer, T_4 as a hexadecamer and, finally, $T_{\geq 8}$ is an emerging filament, having at least one ULF.

In [6] and [15] the model is fit to experimental data of [15]. The raw data consists of four sets, each containing the length distributions of growing filaments at distinct time points up to 20 min. The data sets are obtained for two initial concentrations of tetramers, i.e. 0.45μM and 0.9μM, in two cases: first, with adsorption onto carbon-coated copper grids and second, with adsorption onto mica support. The filament length distributions are determined from electron microscopy (EM) images and atomic force microscopy (AFM) images in the first and second case, respectively. For each set the time-dependent mean filament length (MFL) is calculated and only the processed data are reported in [15]. The models in [6, 15] are capable of reproducing the experimental data on time-dependent dynamics of the mean filament length, however are unsuitable for capturing the time-dependent distribution of the filament lengths. In consequence, the information carried by the available experimental data is not utilized to the full extent. The high resolution of the data is not incorporated into the models, the predictive power of the models is significantly limited since no predictions about the length distributions in time are possible, and the models cannot be fully validated against the available biological knowledge. This highlights the necessity for high-resolution models as a tool for better understanding of the still little-known process of filament self-assembly. In order to

meet this requirement, we apply our methodology of quantitative model refinement to (37). By increasing the resolution with two in two steps we get the following models: first

$$\begin{aligned} 2\,T_1 &\xrightarrow{k_1} T_2 & 2\,T_2 &\xrightarrow{k_2} T_4 \\ 2\,T_4 &\xrightarrow{k_3} T_8 & 2\,T_8 &\xrightarrow{k_f} T_{\geq 9} \\ T_8 + T_{\geq 9} &\xrightarrow{2\,k_f} T_{\geq 9} & 2\,T_{\geq 9} &\xrightarrow{k_f} T_{\geq 9} \end{aligned} \tag{38}$$

and next

$$\begin{aligned} 2\,T_1 &\xrightarrow{k_1} T_2 & 2\,T_2 &\xrightarrow{k_2} T_4 & 2\,T_4 &\xrightarrow{k_3} T_8 \\ 2\,T_8 &\xrightarrow{k_f} T_{\geq 10} & T_8 + T_9 &\xrightarrow{2\,k_f} T_{\geq 10} & T_8 + T_{\geq 10} &\xrightarrow{2\,k_f} T_{\geq 10} \\ 2\,T_9 &\xrightarrow{k_f} T_{\geq 10} & T_9 + T_{\geq 10} &\xrightarrow{2\,k_f} T_{\geq 10} & 2\,T_{\geq 10} &\xrightarrow{k_f} T_{\geq 10}. \end{aligned}$$

Note that T_9 is not a product of any reaction and it will not become one in any further refinement of the model. Since in our experimental set-up we have $T_9(0) = 0$, it follows that $T_9(t) = 0$ for all $t \geq 0$, i.e. reactions $T_8 + T_9 \to T_{\geq 10}$, $2\,T_9 \to T_{\geq 10}$ and $T_9 + T_{\geq 10} \to T_{\geq 10}$ can be eliminated. Thus, the model of resolution 8 coincides with the model of resolution 9. With the same reasoning, all models of resolution between 8 and 15 are identical. The model of resolution 16 is however different:

$$\begin{aligned} 2\,T_1 &\xrightarrow{k_1} T_2 & 2\,T_2 &\xrightarrow{k_2} T_4 & 2\,T_4 &\xrightarrow{k_3} T_8 \\ 2\,T_8 &\xrightarrow{k_f} T_{16} & T_8 + T_{16} &\xrightarrow{2\,k_f} T_{\geq 17} & 2\,T_{16} &\xrightarrow{k_f} T_{\geq 17} \\ T_{16} + T_{\geq 17} &\xrightarrow{2\,k_f} T_{\geq 17} & 2\,T_{\geq 17} &\xrightarrow{k_f} T_{\geq 17}. \end{aligned}$$

Thus, in a model of resolution n, for some arbitrary $n \geq 8$, the variables of the model are T_1, T_2, T_4, T_8, T_{16}, T_{24}, ..., T_{8k}, $T_{\geq n+1}$, where $k = \lfloor n/8 \rfloor$. The biological interpretation of the variable T_{8i}, $1 \leq i \leq k$, is the species of filament consisting of i complete ULFs. Using the terminology of [6] and [15], these are the filaments of length i. Thus, our model of resolution n is in fact the model of resolution $\lfloor n/8 \rfloor$ in terms of the number of complete ULFs included in the filament. This can be seen by rewriting the model (38) as follows (with some of the rate constants renamed):

$$\begin{aligned} 2\,T_1 &\xrightarrow{k_1} T_2 & 2\,T_2 &\xrightarrow{k_2} T_4 \\ 2\,T_4 &\xrightarrow{k_3} F_1 & 2\,F_1 &\xrightarrow{k_4} F_{\geq 2} \\ F_1 + F_{\geq 2} &\xrightarrow{k_u} F_{\geq 2} & 2\,F_{\geq 2} &\xrightarrow{k_f} F_{\geq 2}, \end{aligned} \tag{39}$$

where F_1 stands for filament of length 1 (denoted as T_8 in (38)), and $F_{\geq 2}$ stands for the longer filaments (denoted as $T_{\geq 9}$ in (38)). The refinement of this model

Rate constant	k_1	k_2	k_3	k_4	k_u	k_f
Value	3	30	30	0.25	0.95	0.11

Table 1: Kinetic rate constant values of the extended IF self-assembly model with fast ULF formation (39). The unit is $\frac{1}{\mu M \cdot s}$.

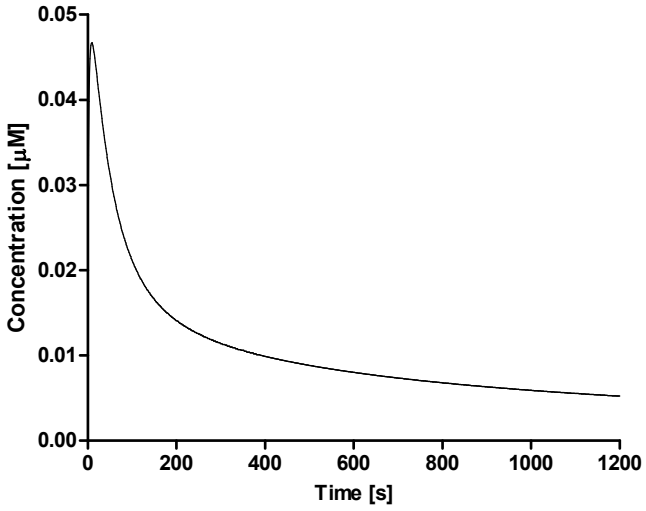
to a higher resolution level, say $n \geq 2$, can be done as follows:

$$\begin{aligned}
2\,T_1 &\xrightarrow{k_1} T_2 \\
2\,T_2 &\xrightarrow{k_2} T_4 \\
2\,T_4 &\xrightarrow{k_3} F_1 \\
F_1 + F_1 &\xrightarrow{k_4} F_2 \\
F_1 + F_i &\xrightarrow{k_u} F_{i+1} \qquad i \in \{2, \ldots, n, \geq n+1\} \\
F_i + F_i &\xrightarrow{k_f} F_{2i} \qquad 2 \leq i \leq n \\
F_i + F_j &\xrightarrow{2\,k_f} F_{i+j} \qquad 2 \leq i < j \leq n \\
F_i + F_{\geq n+1} &\xrightarrow{2\,k_f} F_{\geq n+1} \qquad 2 \leq i \leq n \\
F_{\geq n+1} + F_{\geq n+1} &\xrightarrow{k_f} F_{\geq n+1},
\end{aligned} \tag{40}$$

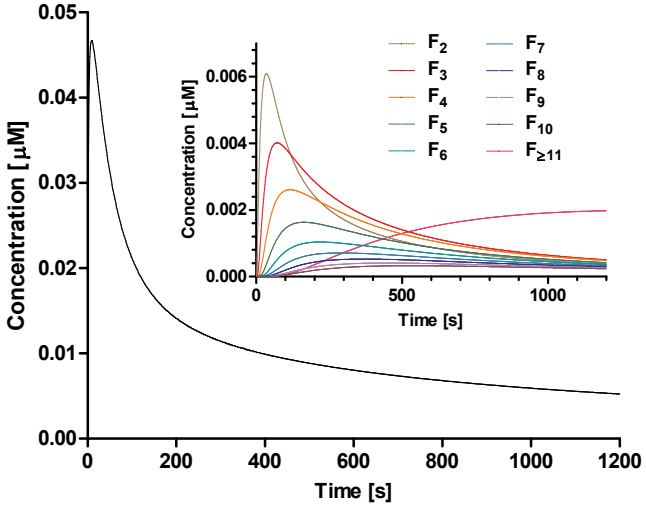
where we adopt the convention that all Fs with indices greater than n are identified with $F_{\geq n+1}$. Model (39) has been experimentally validated in [6]. Using the kinetic constants in Table 1, the numerical behaviour of the model correlates very well with experimental data in [15] on the *in vitro* assembly process of recombinant vimentin at 37 °C. Next, we refine the model in (39) by setting $n = 10$ in (40). In result we obtained a model of resolution 10 for the process of *in vitro* intermediate filament self-assembly that preserves the experimental data fit of the original model. In Figure 1 the dynamics of the overall concentration of filaments predicted by (39) and the model of resolution 10 are presented. Notice that the results are identical, which is in complete agreement with the theoretical deliberations, and there is no need for tedious parameter estimation during the construction of the high-resolution model.

5 Discussion

In this work we concentrated on model refinement, an important aspect of the model-building process. In general, the concept of model refinement can be described as a procedure which, starting with an abstract model of a system, carries out a number of refinement steps which lead to the construction of a more detailed model. At the same time, in order to be correct, the refinement mechanism should be capable of preserving already proven system properties of the original model, e.g. model fit, stochastic semantics, etc. In particular, in our study we focused on the issue of refining an ODE model describing the process of self-assembly. We introduced the notion of model resolution and showed how



(a)



(b)

Figure 1: Comparison between the dynamics of the extended model of IFs self-assembly with fast ULF formation originally introduced in [6] and the refined version of resolution 10. (a) The original extended model with fast ULF formation introduced in [6]. The curve shows the concentration of the intermediate filaments of any length in time. (b) The refined version with resolution 10. The colour curves of the subplot show the dynamics of IFs of lengths from the set $\{1, \ldots, 10\}$ and the overall concentration of filaments of length greater than 10. The black curve in the main plot is obtained by summing the concentrations in time of filaments of length 1 to 10 and those of length greater than 10. Notice that the two models predict identical overall concentration of IFs in time.

the resolution can be both increased and decreased while satisfying the condition of preserving the model fit. Moreover, we showed how the technique can be applied to an existing model: we considered the case-study of self-assembly of intermediate filaments.

Restricted sets of reactions There are two ways of restricting the set of reactions of a generic self-assembly model: either by considering just the intended subset of all possible reactions or by setting to zero the kinetic rate constants for those reactions that are not taking place. It is worth noticing that in both cases the refinement procedure will lead to the correct, expected model: in the first case none of the unwanted reactions will be introduced to the new model and in the second case all the new reactions related through the refinement to the original reactions with the rate constant set to zero will remain inactive, i.e. their rate constants will be zero as well.

Models of infinite resolution In this study we discussed the refinement of a self-assembly model of resolution n to the model of resolution $n+m$, where n and m are some fixed positive integers. One could however think of a refinement to the model of infinite resolution. Although we believe that our methodology would work also in this case, formal theoretical considerations of this issue are much more intricate: one would need to deal with a system of an infinite number of differential equations. Already at the stage of writing the differential equations of the model one would have to make sure that the appearing infinite function series are convergent. For example, let us consider a model of resolution 0, i.e $F + F \xrightarrow{k} F$, and refine it to a model of infinite resolution by assuming in accordance with our methodology that $k_{i,j} := 2k$ for $1 \leq i < j \leq \infty$ and $k_{i,i} := k$ for $1 \leq i \leq \infty$. The solution to the ODE model associated with the 0-resolution model, i.e. $\mathrm{d}F/\mathrm{d}t = -kF^2(t)$, can be obtained analytically: $F(t) = F(0)/(1+k\,t\,F(0))$. In the case of the infinite resolution model one already faces a problem of function series convergence while writing the differential equations for F_is. For each fixed i, the expression for the derivative $\mathrm{d}F_i/\mathrm{d}t$ contains a finite number of terms $k_{l,j}F_iF_j$ where $l + j = i$ with $1 \leq l \leq j < i$, and an infinite number of terms $-k_{i,j}F_iF_j$ where $j \geq 1$. The trouble is whether the infinite series $\sum_{j=1}^{\infty} k_{i,j}F_iF_j$ is convergent for all $t \geq 0$ or whether the terms can be reordered in such a way that the requirement of convergence is satisfied. The difficulty is increased by the fact that the explicit formulas for F_is are unknown. Further, in order for the refinement to be correct, the infinite function series $\sum_{i=1}^{\infty} F_i(t)$ has to be convergent to $F(t)$, i.e. $\sum_{i=1}^{\infty} F_i(t) = F(t)$. If $\sum_{i=1}^{\infty} \mathrm{d}F_i(t)/\mathrm{d}t$ were uniformly convergent, one could write

$$\mathrm{d}F/\mathrm{d}t = \sum_{i=1}^{\infty} \mathrm{d}F_i(t)/\mathrm{d}t. \tag{41}$$

In order to check whether the refinement condition is satisfied, it would be enough to verify (41) and make sure that $\sum_{i=1}^{\infty} F_i(0) = F(0)$. To this aim, by the refinement condition, the left-hand side in (41) could be written as

$$\mathrm{d}F/\mathrm{d}t = -k(\sum_{n=1}^{\infty}\sum_{i=1}^{n} F_{n-i}F_i),$$

where the Cauchy product of $(\sum_{i=1}^{\infty} F_i(t))^2$ is considered. Now, satisfiability of (41) could be checked by proper reordering of the terms on the both sides of (41). However, prior to this, one would need to make sure that all the convergence conditions required by such reorderings are fulfilled. We just signal this issue here without providing a solution to this interesting problem and leave it for further investigation.

Related work The discussed methods for decreasing and increasing the resolution of self-assembly ODE models can be viewed as examples of adaptations of formal model refinement techniques from the field of computer science to systems biology. To the best of our knowledge, formal model refinement has not been explored much in the context of systems biology and this is the first time that it is considered in relation to computational ODE-based models. Some attempts have been made previously in the case of the rule-based formalism, see [7, 21], where the authors consider a process called the *rule refinement*. It is a method to refine rule sets in such a way that the stochastic semantics, dictated by the number of different ways in which a given rule can be applied to a system, is preserved. It is shown how to refine rules and how to choose the refined rates so that the global dynamics of the original and refined systems are the same. For more details we refer to [7, 21].

In Section 3.1, we discussed the numerical choices for the rate constants of the refined self-assembly model and we presented the biological basis for them. However, in general, when considering refinement of reactions describing assembly of larger and lager complexes, one could think of deriving the rate constants based on physical deliberations, i.e. try to estimate how the size of the complexes influences the binding rates. Such an attempt was originally made in [20], where the collision probabilities in the stochastic approach to chemical kinetics were recalculated with taking into account the change in the masses of complexes under formation. However, the solution presented in [20] is not completely satisfactory due to the following two assumptions it is based on: i) reactants are shaped like balls, and, especially, ii) the diameter of the balls representing larger complexes is the same as the diameter of the balls representing small complexes. Nevertheless, this approach seems to have the potential to be developed further to correctly address the problem of relationship between rate constants of reactions involving reactants of same type but different sizes. We leave this interesting problem for further investigation.

Acknowledgments. The work of Eugen Czeizler, Andrzej Mizera and Ion Petre was supported by Academy of Finland, grants 129863, 108421, and 122426. Andrzej Mizera is on leave of absence from the Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland.

References

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology*. Garland Science, New York, 2nd edition, 2004.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

[3] R.-J. Back and J. von Wright. *Refinement Calculus*. Springer, 1998.

[4] F. J. Bruggeman and H. V. Westerhoff. The nature of systems biology. *Trends in Microbiology*, 15(1):45–50, 2007.

[5] 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(239), 2009.

[6] E. Czeizler, A. Mizera, E. Czeizler, R.-J. Back, J. E. Eriksson, and I. Petre. Quantitative analysis of the self-assembly strategies of intermediate filaments from tetrameric vimentin. *Manuscipt*, 2010.

[7] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In J. Fisher, editor, *Formal Methods in Systems Biology. First International Workshop, FMSB 2008, Proceedings*, volume 5054 of *Lecture Notes in Bioinformatics*, pages 103–122, Berlin Heidelberg, 2008. Springer-Verlag.

[8] G. de Vries, T. Hillen, M. Lewis, J. Müller, and B. Schönfisch. *A Course in Mathematical Biology: Quantitative Modelling with Mathematical and Computational Methods*. Monographs on Mathematical Modeling and Computation. SIAM, 2006.

[9] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[10] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[11] R. C. Henrikson, G. I. Kaye, and J. E. Mazurkiewicz. *NMS Histology*. National Medical Series for Independent Study. Lippincott Williams & Wilkins, 1997.

[12] H. Herrmann, M. Häner, M. Brettel, N.-O. Ku, and U. Aebi. Characterization of distinct early assembly units of different intermediate filament proteins. *Journal of Molecular Biology*, 286(5):1403–1420, 1999.

[13] H. Herrmann, M. Häner, M. Brettel, S. A. Müller, K. N. Goldie, B. Fedtke, A. Lustig, W. W. Franke, and U. Aebi. Structure and assembly properties of the intermediate filament protein vimentin: the role of its head, rod and tail domains. *Journal of Molecular Biology*, 264(5):933–953, 1996.

[14] K. E. Iverson. *A Programming Language*. Wiley, New York, 4th edition, 1962.

[15] R. Kirmse, S. Portet, N. Mücke, U. Aebi, H. Herrmann, and J. Langowski. A quantitative kinetic model for the in vitro assembly of intermediate filaments from tetrameric vimentin. *Journal of Biological Chemistry*, 282(52):18563–18572, 2007.

[16] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[17] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice*. Wiley-VCH, 2006.

[18] D. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, 1992.

[19] A. D. Lander. The edges of understanding. *BMC Biology*, 8:40, 2010.

[20] L. Lok and R. Brent. Automatic generation of cellular reaction networks with moleculizer 1.0. *Nat. Biotechnol.*, 23:131–136, 2005.

[21] E. Murphy, V. Danos, J. Feret, J. Krivine, and R. Harmer. Rule-based modeling and model refinement. In H. M. Lodhi and S. H. Muggleton, editors, *Elements of Computational Systems Biology*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.

[22] K. Raman and N. Chandra. Systems biology. *Resonance*, 15(2):131–153, 2010.

[23] W. L. Scherlis and D. S. Scott. First steps towards inferential programming. In R. E. A. Mason, editor, *Information Processing 83: Proceedings of the IFIP 9th World Computer Congress*, 1983.

[24] N. Wirth. Program development by stepwise refinement. *Communications of the ACM*, 14(4):221–227, 1971.