Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

Model-supported Patient Stratification Using Set-based Estimation Methods

Nadine Rudolph * **Petar Andonov** * **Heinrich J. Huber** *
Rolf Findeisen *

* *Laboratory for Systems Theory and Automatic Control
Otto-von-Guericke University Magdeburg, Germany*

Abstract: Stratification of patients into different risk subcategories for disease development plays an important role in medical treatments. It sets the basis for physicians to decide upon personalized interventions. This patient-specific therapy design increasingly becomes supported by mathematical models that describe the underlying disease processes on a detailed molecular level. However, the mathematical description of disease development is challenging. Often the underlying processes act on different time scales. Furthermore, the biomedical data and measurements have different quantities, qualities and uncertainties. New methods are required to address this heterogeneity in the data landscape and to integrate measurements on different time scales in order to extract meaningful information over the disease process. We devise an approach for integrating biological signals for short and long-term molecular processes into a coherent framework. To this end, we combine set-based estimation methods for short-term molecular pathways with classification approaches of long-term disease development. The developed framework is demonstrated by means of IL-6-induced Jak-STAT3 and MAPK trans-signaling.

Keywords: Patient stratification; short-term signal transduction; long-term cellular responses; combining time scales; set-based methods; classification methods

1. INTRODUCTION

Signal transduction pathways are important for transmitting initial signals at the receptor level into a cascade of various protein-protein interactions, eventually resulting in cellular responses such as cell differentiation, cell death or growth. Deregulation of such signal transduction processes, potentially induced by altered protein levels or missense mutations, can eventually lead to the development of severe diseases. To this end, several inflammation-related diseases, such as Inflammatory Bowel Diseases, Rheumatoid Arthritis and even cancer have been associated to altered signaling (Roy et al. (2008); Kamimura et al. (2003); Veltman et al. (2017)).

Mathematical models to understand disease progression over time can, often in combination with model-driven experiments, help to gain insight into the key players of misbalanced signaling and unravel drug targets to mitigate or reverse disease progression. Yet, the provision of reliable mathematical models faces several challenges. Firstly, such models contain often unknown model parameters such as kinetic parameters and initial protein concentrations, preventing the establishment of quantitative and reliable models. Secondly, measurement data are almost ever subject to large uncertainties and are often even unreliable (Streif et al. (2016)). Finally, the cellular and molecular processes that govern disease development over time operate on different time scales and can cover fast ones, such as pathway initialization, receptor formation and activation, or slower ones such as cell growth and differentiation. This heterogeneity gets aggravated by the fact that the presence of these different time scales is reflected by a different landscape in data quality. Specifically, data for short-term measurements can be obtained with high temporal resolution (such as protein levels over time), while only sparse (such as measurements of tumour growth) or categorical (survival/death, stage of inflammation) are often available for long-term processes. The combination of these both time scales is not trivial, but nevertheless indispensable for the understanding of how and why diseases develop at the molecular level.

To identify model parameters for short-term processes from given, uncertain data, set-based and probability-based estimation methods have been developed (Kreutz et al. (2012); Rumschinski et al. (2010); Milanese and Novara (2004); Bemporad et al. (2005)). Set-based methods account for data uncertainties by finding sets of possible model parametrizations such that the model predictions fit the uncertain experimental data over time. Set-based methods have been applied in the context of parameter estimation, model invalidation, and experimental design for biochemical problems. For reliably predicting disease progression, however, extension of such set-based frameworks to identify parameters for short and long-term behavior are needed.

The previous work Rudolph et al. (2015) presented a set-based approach for combining both time scales for signaling of the Jak-STAT3 pathway to relate early signaling events with long-term responses. However, an appropriate

[1] NR, PA, and RF are members of the International Max Planck Research School (IMPRS) for Advanced Methods in Process and System Engineering, Magdeburg.

[2] The authors acknowledge partial funding by the BMBF in the frame of the research project InTraSig, grant number 031A300A and the research center of dynamic systems (CDS) funded in the frame of the ERDF (European Regional Development Fund).

[3] Corresponding author: rolf.findeisen@ovgu.de

classification of long-term patient risk for the development of inflammatory diseases was not considered. We therefore here provide an unified framework combining set-based estimation methods and classification approaches for patient stratification into risk subcategories based on long-term outcomes.

This contribution is structured as follows: In Section 2, the set-based framework for combining short- and long-term scales as well as an overview of the proposed framework are outlined. In Section 3, our approach using IL-6-induced Jak-STAT3 and MAPK trans-signaling is employed, followed by a brief conclusion in Section 4.

2. SET-BASED STRATIFICATION APPROACH

We are interested in the combination of biological processes that act on two different time scales and for which uncertain measurement data are available. Furthermore, we are interested in how to use the uncertain data and dynamical model describing the fast and short-term processes to make a prediction about the slow and long-term outcomes.

A classical approach for the combination of different time scales would be to consider a dynamical model for each time scale and to look for a relation between the parameters in each model. Although this could work in general, a deep knowledge about the underlying processes at each time scale, their interconnection and also insight in the parameter variation due to the uncertainties would be required. The challenge becomes even bigger if we consider only sparse data on the long-term scale. We consider the extreme case of a single data point on the long-term scale which will be further used for the stratification of a simulated patient cohort into risk subcategories.

We propose a framework that requires only one dynamical model for the short-term scale processes as well as uncertain patient data for processes on the short- and long-term scale, and a classification method. We continue with the mathematical formulation of the uncertain data, parameters and dynamical processes. Afterwards we present the framework and two algorithms, which illustrate how the method learns the correct stratification and explain how the method is applied.

2.1 Set-based Problem Setup

In the following, we focus on the first step to determine uncertain parameters from short-term biochemistry experiments that are used as inputs for the classification methods. To this end, we utilize the set-based approach presented in Rumschinski et al. (2010).

We consider a class of polynomial discrete-time systems of the form

$$f(x(k+1), x(k), u(k), p) = 0 \tag{1a}$$
$$h(y(k), x(k), u(k), p) = 0, \tag{1b}$$

where $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ are polynomial or rational functions, $x(k) \in \mathbb{R}^{n_x}$ is the time-variant state vector, $u(k) \in \mathbb{R}^{n_u}$ the time-variant input vector, and $p \in \mathbb{R}^{n_p}$ the time-invariant parameter vector. The model output equations are given by $h : \mathbb{R}^{n_y} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}$, which are assumed as polynomial functions, and $y(k) \in \mathbb{R}^{n_y}$ denote the time-variant model output vector. Time is indexed by $k \in \mathbb{N}$ and considered within a finite horizon $\mathcal{T}$, $k \in \mathcal{T}$ for parameter estimation. Measurements are considered to be uncertain at each time instance k, i. e. $y(k)$ and $u(k)$, and lie inside the compact sets $\mathcal{Y}_k$ and $\mathcal{U}_k$, respectively. Furthermore, the initial conditions $x(0)$ and the model parameters p are contained inside the sets $\mathcal{X}_0$ and $\mathcal{P}$:

$$\begin{aligned} y(k) \in \mathcal{Y}_k,\ u(k) \in \mathcal{U}_k, \\ x(0) \in \mathcal{X}_0,\ p \in \mathcal{P}. \end{aligned} \tag{2}$$

Considering the introduced uncertainty descriptions, our goal is to estimate consistent and patient-specific parameter sets $\mathcal{P}_c$, such that a model for short-term and dynamic processes is consistent with experimental patient data, i. e. $y(k) \in \mathcal{Y}_k$ and $u(k) \in \mathcal{U}_k$, $\forall k \in \mathcal{T}$. Therefore, we reformulate the systems dynamics and uncertainties into a so-called feasibility problem FP:

$$\mathrm{FP} : \begin{cases} \text{find} & \xi \\ \text{subject to} & g_i(\xi), \end{cases} \tag{3}$$

with $\xi \in \mathbb{R}^{(n_x+n_y+n_u)n_t+n_p}$ being the vector containing all time-variant and time-invariant variables of (1) and the constraints $g_i(\xi)$ represent the nonlinear dynamics in (1) as well as the set-based uncertainties from (2).

Due to the nonlinearities in (1), the stated FP becomes nonconvex and is hard to solve. As shown in Borchers et al. (2009), it is possible to relax the FP into a convex semidefinite feasibility problem (SDP). To deal with larger problems the SDP can be further relaxed into a linear program (LP), which can be solved efficiently using state-of-the art solvers, e. g. CPL (2007).

To obtain an outer-bounding of the feasible parameter space, the FP in (3) is replaced by an optimization problem in which the single parameters are minimized or maximized (cf. Fig. 1, Rumschinski et al. (2010)).

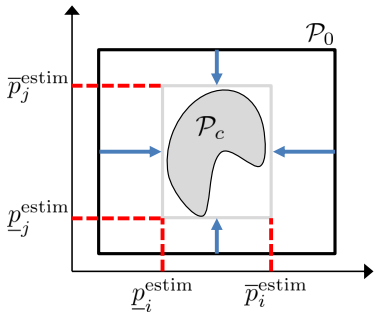

Fig. 1. Outer-bounding of uncertain parameters p_i and p_j. The algorithm allows to estimate tight lower ($\underline{p}_j^{\text{estim}}$, $\underline{p}_i^{\text{estim}}$) and upper ($\overline{p}_j^{\text{estim}}$, $\overline{p}_i^{\text{estim}}$) bounds for the consistent parameter set $\mathcal{P}_c$ (light grey rectangle) by sequentially and iteratively tightening the initial parameter bounds $\mathcal{P}_0$ (blue arrows and black rectangle).

In the second step of the analysis, we predict the long-term behavior from model parameters determined in the first step. Therefore, we arrange the sequence of long-term data into ordered tupels for each patient, which we associate to classes of long-term patient responses and which serve as output categories for the classification approach. If the long-term data only consists of a single endpoint (as in our example below), its value are directly associated with a group such as providing a stratification for low, medium or high risk patient based on predefined thresholds for the respective data. For this, the patients are

split into groups for training, for validation and for test, in line with standard procedures of classification approaches. The analysis is performed for each patient of each group separately.

2.2 Overview of the Stratification Framework

Our approach combines the set-based estimation framework presented in Rumschinski et al. (2010) for data on the fast, short-term scale with classification methods for patient stratification using data on the slow, long-term scale as detailed in Figure 2. We first assume that short-term data obtained from biochemical surrogates of patients (such as cells extracted from biopsies or blood serum) are available. These surrogates are probed by biochemical stimulation (e. g. excitation of a certain pathway) and point us to changes in proteins or genes indicative of disease status within seconds to minutes. We use dynamical computational models to understand the short-term signaling processes and apply set-based estimation methods to obtain model parameter sets that can describe the biochemical data. These parameter sets are subsequently transformed (piped) through a classification algorithm (e. g. an artificial neural network, support vector machines, etc.), resulting in a set of transformed parameters (in our example: weights and thresholds in a neural network) to match the long-term response (e. g. disease status over days).

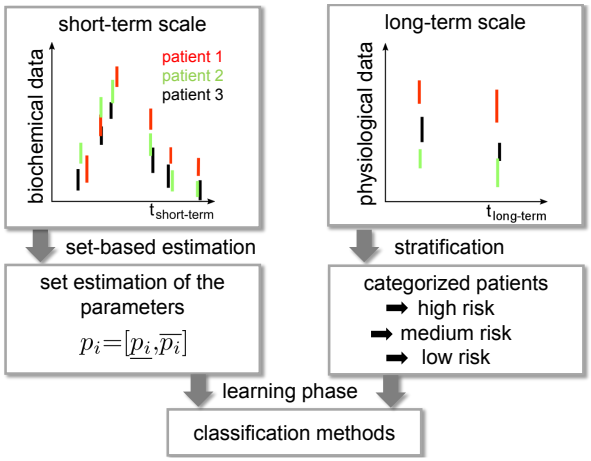

Fig. 2. Proposed workflow for combining fast and uncertain biochemical processes on the short-term scale with slow and uncertain physiological responses on the long-term scale. The short-term scale data often describe the patient diagnosis over longer time and may be seen as a means for patient strata for disease development (e. g. low, medium and high risk).

The presented workflow is formalized in Algorithm 1. The algorithm is used to obtain a trained classifier based on short-term patient data. Once Algorithm 1 has produced a classifier that has satisfying performance, the obtained result is used to stratify new patients according their long-term profiles as described in Algorithm 2.

In the following section, we illustrate applicability of Algorithm 1 and 2.

3. EXAMPLE

In the following, we use simulated measurements for the short- and long-term scale for a group of 50 patients.

Algorithm 1 Set-based classifier training

Input:
Short- and long-term scale data for each patient;
A dynamical model describing the behavior of the short-term scale data;
Threshold values defining each risk group;
An untrained classifier;

Output:
A trained classifier;

1: Perform set-based parameter estimation using the short-term scale data and the dynamical model
2: Process the data for each time instance on the long-term scale into the risk subcategories
3: Stratify the patients into the corresponding risk subcategories
4: Arrange the corresponding parameter bounds of each patient to the corresponding risk subcategory for each long-term time instance
5: Split patients into groups for training, validation and test
6: Train the chosen classifier
7: Verify the quality of the classifier and if needed adjust and re-train the classifier

Algorithm 2 Patient stratification

Input:
Short-term scale data for each patient;
The same dynamical model describing the behavior of the short-term scale data as in Algorithm 1;
The trained classifier from Algorithm 1;

Output:
A prediction of the patient-specific risk category;

1: Perform set-based parameter estimation using the short-term scale data and the dynamical model
2: Input the set-based parameter estimation results into the trained classifier
3: Obtain the stratification results

Simulated short-term data consist of a biochemical data set, assumed to be provided by biochemically testing the response of extracted patient tissue to a certain stimulation. In particular, we assume to avail of protein changes of patient tissue associated with two important signaling pathways, IL-6-induced Jak-STAT3 and MAPK trans-signaling. Thereby, trans-signaling refers to a specific branch of the respective pathway inducing a large number of genes, coding for proteins, that are involved in long-term pathophysiology, such as cancer, Rheumatoid Arthritis, and Multiple Sclerosis (Yu et al., 2009; Rose-John, 2012; Lo et al., 2011).

As long-term response, we assume to avail of a single endpoint at a later time point for each patient, which we together associate to low, medium and high risk for developing an inflammatory disease.

3.1 Mathematical Pathway Models

In the following, the implemented models for IL-6-induced Jak-STAT3 and MAPK trans-signaling are introduced (cf. Fig. 3 and Heinrich et al. (2003)).

Jak-STAT3 pathway model: IL-6 initiates trans-signaling by binding to the soluble receptor subunit glycoprotein 80 (sIL-6R). Two entities of the resulting receptor complex are then bound to two receptor subunits glyco-

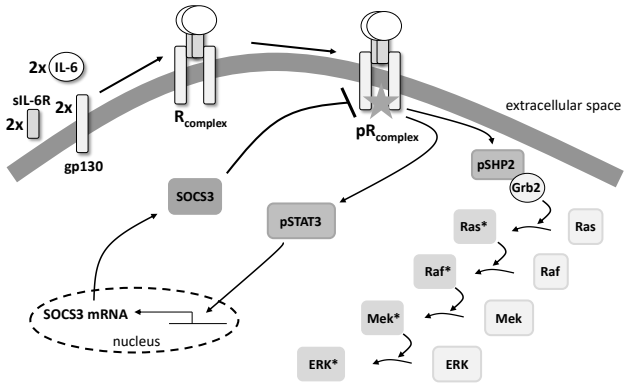

Fig. 3. Schematic representation of IL-6-induced receptor complex formation during trans-signaling, Jak-STAT3 pathway activation and MAPK signaling.

protein 130 (gp130) forming a hexameric receptor complex ($\mathrm{R_{complex}}$). Due to the constitutive binding of tyrosine kinases of the Jak family to the intracellular domain of gp130, Jaks become activated and in turn phosphorylate gp130 ($\mathrm{pR_{complex}}$). To this active hexameric complex, Signal Transducer and Activator of Transcription 3 (STAT3) proteins are recruited to phosphorylated gp130. STAT3s are phosphorylated (pSTAT3) by Jaks leading to the formation of active STAT3 dimers. Phosphorylated STAT3 dimers function as nuclear transcription factors regulating several target genes, including its own negative regulator Suppressors of Cytokine Signaling 3 (SOCS3).

The above described reaction mechanisms can be described as follows:

$$\begin{aligned}
x_1(k+1) &= x_1(k) + \Delta t\big(p_1 x_7(k) u - p_2 x_1(k) \\
&\quad - 2p_3 x_2(k)^2 x_1(k)^2 + 2p_4 x_8(k)\big) \\
x_2(k+1) &= x_2(k) + \Delta t\big(2p_4 x_8(k) - 2p_3 x_2(k)^2 x_1(k)^2\big) \\
x_3(k+1) &= x_2(k) + \Delta t\Big(\frac{p_5 x_8(k)}{1 + p_{13} x_6(k)} - p_6 x_3(k)\Big) \\
x_4(k+1) &= x_4(k) + \Delta t\big(p_7 x_3(k) x_9(k) - p_8 x_4(k)\big) \\
x_5(k+1) &= x_5(k) + \Delta t\big(p_9 x_4(k) - p_{10} x_5(k)\big) \\
x_6(k+1) &= x_6(k) + \Delta t\big(p_{11} x_5(k) - p_{12} x_6(k)\big).
\end{aligned} \tag{4}$$

Thereby, the variables $x_1(k)$, $x_2(k)$, $x_3(k)$, $x_4(k)$, $x_5(k)$, $x_6(k)$ and u denote IL-6~sIL-6R, gp130, $\mathrm{pR_{complex}}$, pSTAT3, SOCS3 mRNA, SOCS3, and IL-6. Furthermore $x_7(k)$, $x_8(k)$, $x_9(k)$ describe the entities sIL-6R, $\mathrm{R_{complex}}$ and STAT3, respectively which can be extracted from the following conservation laws:

$$\begin{aligned}
\mathrm{sIL{-}6R^{Total}} &= \mathrm{sIL{-}6R + IL{-}6{\sim}sIL{-}6R} \\
&\quad + \mathrm{2R_{complex} + 2pR_{complex}} \\
\mathrm{gp130^{Total}} &= \mathrm{gp130 + 2R_{complex} + 2pR_{complex}} \\
\mathrm{STAT3^{Total}} &= \mathrm{STAT3 + pSTAT3}.
\end{aligned}$$

Furthermore, Δt denotes the sampling time. Note that for model simplification, we assumed Jak kinases to be represented by gp130 species.

MAPK pathway model: Due to the activation of Jaks and the subsequent phosphorylation of gp130 akin to Jak-STAT3 signaling, the SH2-containing protein tyrosine phosphatase 2 (SHP2) is recruited and phosphorylated. Phosphorylated SHP2 (pSHP2) acts as an adaptor protein for several proteins, including Growth factor receptor-bound protein 2 (Grb2). Grb2 is constitutive associated with SOS (Son Of Sevenless), which is a guanine nucleotide exchange factor activating the small G-protein Ras which is bound to the nucleotide guanosine diphosphate (GDP). SOS forces Ras to release GDP and subsequently, Ras binds to nucleotide guanosine triphosphate resulting in Ras activation ($\mathrm{Ras^*}$). $\mathrm{Ras^*}$ interacts with and stimulates downstream signaling effectors, including the kinase Raf. Raf is activated to $\mathrm{Raf^*}$ and stimulates its downstream target, the MAP kinase ERK through the intermediate kinase Mek. Stimulated ERK ($\mathrm{ERK^*}$) activates a number of transcription factors, which play an important role in cell proliferation and differentiation.

The reaction mechanisms for the MAPK pathway can be described as follows:

$$\begin{aligned}
x_1(k+1) &= x_1(k) + \Delta t\big(p_1 x_7(k) u - p_2 x_1(k) \\
&\quad - 2p_3 x_2(k)^2 x_1(k)^2 + 2p_4 x_8(k)\big) \\
x_2(k+1) &= x_2(k) + \Delta t\big(2p_4 x_8(k) - 2p_3 x_2(k)^2 x_1(k)^2\big) \\
x_3(k+1) &= x_2(k) + \Delta t\big(p_5 x_8(k) - p_6 x_3(k)\big) \\
x_4(k+1) &= x_4(k) + \Delta t\big(p_7 x_3(k) x_8(k) - p_8 x_4(k)\big) \\
x_5(k+1) &= x_5(k) + \Delta t\big(p_9 x_4(k) x_9(k) - p_{10} x_5(k)\big), \\
x_6(k+1) &= x_6(k) + \Delta t\big(p_{11} x_5(k) x_{10}(k) - p_{12} x_6(k)\big) \\
x_7(k+1) &= x_7(k) + \Delta t\big(p_{13} x_6(k) x_{11}(k) - p_{14} x_7(k)\big).
\end{aligned} \tag{5}$$

In (5), the variables $x_1(k)$, $x_2(k)$ and $x_3(k)$ are similar to (4), whereby the same conserved moieties hold for sIL-6R and $\mathrm{R_{complex}}$. Moreover, the variables $x_4(k)$, $x_5(k)$, $x_6(k)$, and $x_7(k)$ denote $\mathrm{Ras^*}$, $\mathrm{Raf^*}$, $\mathrm{Mek^*}$ and $\mathrm{ERK^*}$, respectively. The inactive forms Ras, Raf, Mek and ERK denoted as $x_8(k)$, $x_9(k)$, $x_{10}(k)$, and $x_{11}(k)$ can be extracted from the conservation laws:

$$\mathrm{Ras^{Total} = Ras + Ras^*},\ \mathrm{Raf^{Total} = Raf + Raf^*}$$
$$\mathrm{Mek^{Total} = Mek + Mek^*},\ \mathrm{ERK^{Total} = ERK + ERK^*}.$$

We note that Grb2/SOS was not explicitly modeled but considered as an integral part of the phosphorylated receptor.

3.2 Simulation of Short- and Long-term Patient Data

To demonstrate applicability of the presented approach, we assume 50 patients and calculate simulated data for the proteins pSTAT3 and $\mathrm{ERK^*}$, acting as upstream surrogates for inflammatory diseases.

For generating data on pSTAT3, the input IL-6 was fixed to 0.2 and $\mathrm{STAT3^{Total}}$ was set to 10. Furthermore, we assumed the kinetic parameters $p_i^{\mathrm{Jak-STAT3}}$ with $i := \{1,\dots,13\}$ as $p_i^{\mathrm{Jak-STAT3}}$=(0.075,0.056,0.01,0.00015,0.25,0.09,1.5,0.01,... 0.1,0.1,1,0.1,5)$^{\mathrm{T}}$ and the initial conditions were set to $x(0)$=(0,$\mathrm{gp130^{Total}}$,0,0,0,0). Values for the total concentrations of $\mathrm{gp130^{Total}}$ and $\mathrm{IL\text{-}6R^{Total}}$ were randomly generated within the bounds of $\mathrm{gp130^{Total}}$=[1,5] and $\mathrm{sIL\text{-}6R^{Total}}$=[0.5,2], respectively.

For data generation on $\mathrm{ERK^*}$, we fixed IL-6 to 0.2, $\mathrm{gp130^{Total}}$ to 5 and $\mathrm{IL\text{-}6R^{Total}}$ to 2. The kinetic parameters p_i^{MAPK} with $i := \{1,\dots,14\}$ were assumed as p_i^{MAPK}=(0.075,0.056,0.01,0.00015,0.25,0.09,0.1,0.1,0.1,... 0.1,0.1,0.1,0.1,0.1)$^{\mathrm{T}}$ and the initial conditions were

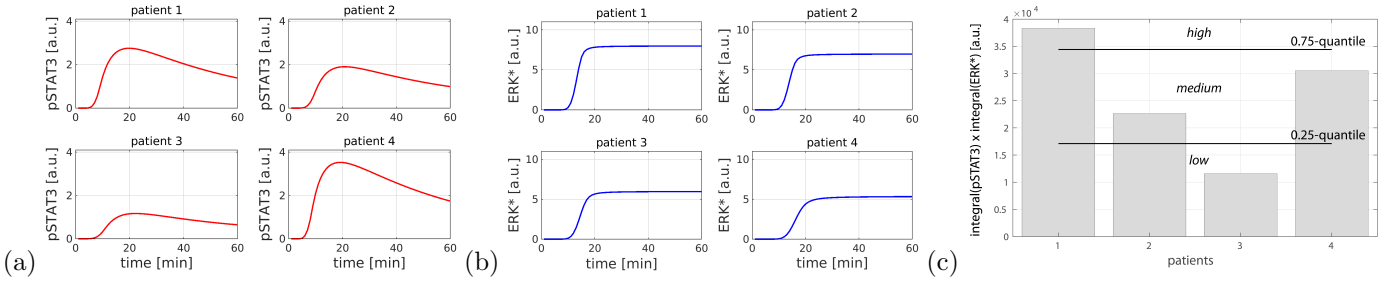

Fig. 4. Simulated data, exemplary for 4 patients. (a) Short-term profiles for pSTAT3 during Jak-STAT3 signaling and (b) ERK* during MAPK signaling. (c) Calculated long-term outcomes (bars), which are obtained by multiplying the integrals of the pSTAT3 and ERK* trajectories from (a) and (b), and the determined 75% and 25% quantiles (black bold lines) for patient stratification.

set $x(0)$=(0,gp130$^{\text{Total}}$,0,0,0,0,0). Values for the total concentrations of Ras$^{\text{Total}}$, Raf$^{\text{Total}}$, Mek$^{\text{Total}}$, and ERK$^{\text{Total}}$ were randomly generated within the bounds [1,10], respectively.

As long-term patient outcome, we assumed the integrated response of the activation of both pathways as a disease surrogate. We therefore first calculated the integrals of the pSTAT3 and ERK* trajectories for each patient using the Matlab function *trapz* for a time horizon of 60 minutes. Then, both integrals were multiplied to assess a collective effect for inflammation on the long-term scale. These results are then grouped over the patient cohort to classify them into patients with values lower than 25%-quantile "low risk patients", such between 25%- and 75%-quantiles "medium risk patients" and such with higher than the 75%-quantile "high risk patients".

The simulated short (pSTAT3 and ERK*) and long time scale data (corresponding integrals) together with the patient stratification that is used for classification are shown exemplarily for 4 patients in Fig. 4.

3.3 Set-based Estimation of Patient-specific Parameters

The set-based approach described in Section 2.1 was used to remodel the short-term scale data for each pathway in order to render a set of outer-bounded parameters for each patient and pathway. For solving the feasibility problem of the dynamical system, Δt was set to 2 minutes for a time horizon of 60 minutes. To account for experimental uncertainties, errors of $\pm 10\%$ were added to the vectors, $p_i^{\text{Jak-STAT3}}$ and p_i^{MAPK} as well as the simulated short-term data for pSTAT3 and ERK*. From this generated short-term data, the patient-specific parameters gp130$^{\text{Total}}$ and IL-6R$^{\text{Total}}$ in the Jak-STAT pathway and Ras$^{\text{Total}}$, Raf$^{\text{Total}}$, Mek$^{\text{Total}}$, and ERK$^{\text{Total}}$ in the MAPK pathway were estimated for each patient using the outer-bounding approach as described in Section 2.1. Calculations were performed using the *Analysis, Design and Model Invalidation Toolbox* (ADMIT) (Streif et al., 2012) and the solver Cplex (CPL, 2007).

In Tables 1 and 2, the outer-bounding results for 4 simulated patients (cf. Fig. 4) are presented.

3.4 Training and Validation of the Classifier

For the classification part of our method, several approaches can be used (cf. Dougherty (2013); Lu and Weng (2007)). To demonstrate the proposed framework,

Table 1. Outer-bounding results for 4 patients for the Jak-STAT3 signaling pathway (cf. Fig. 4(a))

Patients	sIL-6R$^{\text{Total}}$	gp130$^{\text{Total}}$
1	[1.2, 1.8]	[3.7, 4.9]
2	[0.9, 1.6]	[3.2, 4.3]
3	[0.7, 1.1]	[2.7, 3.6]
4	[1.4, 2.0]	[4.1, 5.0]

Table 2. Outer-bounding results for 4 patients for the MAPK signaling pathway (cf. Fig. 4(b))

Patients	Ras$^{\text{Total}}$	Raf$^{\text{Total}}$	Mek$^{\text{Total}}$	ERK$^{\text{Total}}$
1	[7.6, 9.3]	[6.7, 9.2]	[7.2, 9.8]	[8.3, 9.6]
2	[6.8, 9.2]	[6.0, 9.0]	[6.4, 9.6]	[7.4, 8.5]
3	[5.9, 8.0]	[5.2, 8.7]	[5.6, 8.4]	[6.5, 7.4]
4	[3.9, 5.3]	[3.2, 5.4]	[5.7, 8.5]	[5.9, 6.8]

we opted to use Artificial Neural Networks (ANN), cf e. g. Picton (1994); Dougherty (2013). Yet, also Support Vector Machines (Abe, 2005) or boosting methods (Schapire, 2003) can be used.

Structurally, the chosen ANN consisted of a hidden layer with 10 neurons, to which the input information is fed, and one output layer, which provides the classification results. All neurons in the hidden layer have sigmoid activation functions, in contrast to the output layer having softmax activation functions (Sutton and Barto, 1998). The use of softmax neurons normalizes the outcome such that all outcomes add to 1, and hence the patient category as outcome can be interpreted as probability function (Goodfellow et al., 2016).

The input of the ANN is a vector of 12 inputs, which correspond to the upper and lower boundary value of each of the parameters of the two pathways, which preserves the set-based notion. Calculations were performed using the Neural Pattern Recognition app in MATLAB (Beale et al., 2017). The patients were split into training, validation, and test group according a split of 60%, 20% and 20%, respectively. The resulting confusion matrices are presented in Fig. 5. The results demonstrate a 83.3% correct classification of the training set, 90% of the cross-validation set, 80% of the test set, and an overall correctness of 84%.

To ensure the reproducibility of the results different runs were carried out. The used integrated random algorithm, for choosing which patient falls in which group, demonstrated that the results could only improve. Furthermore, increasing the size of the hidden layer did not produce improvement in the classification.

Fig. 5. Estimation results after applying Algorithms 1 and 2 on the patients data.

4. CONCLUSIONS

We provided a modeling framework that allows for combining processes and data on the short- and long-term scale under the umbrella of an unified set-based approach. The framework can be seen as an extension of methods that provide a feasibility set for experimental data with inherent uncertainties. The sets obtained for a fast, short-term and often pathway-based description of the disease process are transformed by piping it through a classification algorithm to provide a prediction of long-term scale data. With this, we aimed to provide a shift in reasoning over feasibility sets to define them as a super-class for explaining data under uncertainty, while covering a process of detailed dynamical modeling and more abstract classification approaches at the same time.

REFERENCES

(2007). ILOG CPLEX 11.0 User's Manual. *Gentilly, France: ILOG SA*.

Abe, S. (2005). *Support vector machines for pattern classification*, volume 53. Springer.

Beale, M., Hagan, M.T., and Demuth, H.B. (2017). *Neural Network Toolbox - Getting Started Guide*. MathWorks Inc.

Bemporad, A., Garulli, A., Paoletti, S., and Vicino, A. (2005). A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50(10), 1567–1580.

Borchers, S., Rumschinski, P., Bosio, S., Weismantel, R., and Findeisen, R. (2009). A set-based framework for coherent model invalidation and parameter estimation of discrete time nonlinear systems. In *Proc. IEEE Conference on Decision and Control*, 6786–6792.

Dougherty, G. (2013). *Pattern Recognition and Classification: An Introduction*. Springer-Verlag New York.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Heinrich, P., Behrmann, I., Haan, S., Hermanns, H., Muller-Newen, G., and Schaper, F. (2003). Principles of Interleukin (IL)-6-type cytokine signalling and its regulation. *Journal of Biochemistry*, 374, 1–20.

Kamimura, D., Ishihara, K., and Hirano, T. (2003). IL-6 signal transduction and its physiological roles: the signal orchestration model. In *Reviews of Physiology, Biochemistry and Pharmacology*, 1–38. Springer.

Kreutz, C., Raue, A., and Timmer, J. (2012). Likelihood based observability analysis and confidence intervals for predictions of dynamic models. *BMC Systems Biology*, 6(1), 120.

Lo, C.W., Chen, M.W., Hsiao, M., Wang, S., Chen, C.A., Hsiao, S.M., Chang, J.S., Lai, T.C., Rose-John, S., Kuo, M.L., et al. (2011). IL-6 trans-signaling in formation and progression of malignant ascites in ovarian cancer. *Cancer Research*, 71(2), 424–434.

Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5), 823–870.

Milanese, M. and Novara, C. (2004). Set membership identification of nonlinear systems. *Automatica*, 40(6), 957–975.

Picton, P. (1994). *Introduction to Neural Networks*. Macmillan Education UK, London.

Rose-John, S. (2012). IL-6 trans-signaling via the soluble IL-6 receptor: importance for the pro-inflammatory activities of IL-6. *International Journal of Biological Sciences*, 8(9), 1237–1247.

Roy, P.K., Rashid, F., Bragg, J., and Ibdah, J.A. (2008). Role of the JNK signal transduction pathway in inflammatory bowel disease. *World Journal of Gastroenterology: WJG*, 14(2), 200.

Rudolph, N., Meyer, T., Franzen, K., Garbers, C., Schaper, F., Streif, S., Dittrich, A., and Findeisen, R. (2015). A two-level approach for fusing early signaling events and long term cellular responses. In *Proc. of the International Symposium on Advanced Control of Chemical Processes*, volume 48, 1228–1233. Elsevier.

Rumschinski, P., Borchers, S., Bosio, S., Weismantel, R., and Findeisen, R. (2010). Set-based dynamical parameter estimation and model invalidation for biochemical reaction networks. *BMC Systems Biology*, 4, 69–82.

Schapire, R.E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, 149–171. Springer.

Streif, S., Kim, K.K.K., Rumschinski, P., Kishida, M., Shen, D.E., Findeisen, R., and Braatz, R.D. (2016). Robustness analysis, prediction, and estimation for uncertain biochemical networks: An overview. *Journal of Process Control*, 42, 14–34.

Streif, S., Savchenko, A., Rumschinski, P., Borchers, S., and Findeisen, R. (2012). ADMIT: a toolbox for guaranteed model invalidation, estimation and qualitative–quantitative modeling. *Bioinformatics*, 28(9), 1290–1291.

Sutton, R.S. and Barto, A.G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

Veltman, D., Laeremans, T., Passante, E., and Huber, H.J. (2017). Signal transduction analysis of the NLRP3-inflammasome pathway after cellular damage and its paracrine regulation. *Journal of Theoretical Biology*, 415, 125–136.

Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. *Nature Reviews. Cancer*, 9(11), 798–809.