2015 European Control Conference (ECC)
July 15-17, 2015. Linz, Austria

Analysis-based Parameter Estimation of an *in vitro* Transcription-Translation System

Zoltán A. Tuza[1], Dan Siegal-Gaskins[3], Jongmin Kim[4] and Gábor Szederkényi[1,2]
[1]Pázmány Péter Catholic University, Faculty of Information Technology and Bionics,
Práter u. 50/a, H-1083 Budapest, Hungary
[2]Institute for Computer Science and Control, Hungarian Academy of Sciences,
Kende u. 13-17, H-1111 Budapest, Hungary
[3]California Institute of Technology, Biology and Biological Engineering, Pasadena, CA, USA
[4]Wyss Institute for Biologically Inspired Engineering, Harvard University Boston, MA, USA
e-mail: tuza.zoltan@itk.ppke.hu

Abstract— **Recent advances in measurement technology provide us with rich source of data for estimating parameters in biomolecular circuit models, particularly in simplified *in vitro* transcription-translation systems, so-called molecular "breadboards". In this paper, we elaborate on a mass action type dynamic model for such an *in vitro* system and detail a parameter estimation procedure that may be used with time series data containing information about both transcriptional and translational stages of gene expression. The identification process is supported by structural identifiability analysis to ensure proper model structure. Statistical analysis and validation of the estimated parameter set help us to understand the characteristics of point estimation results.**

I. Introduction

While the use of quantitative models in biology has become commonplace in recent decades, the amount and types of experimental data available for model parameter estimation are often severely limited. In many cases, temporal resolution and/or the sensitivity of the measurement technique present significant obstacles for effective parameter estimation. Additionally, structural non-identifiability of the model is also a possibility [4], [5], [24]. However, recent developments in real-time mRNA reporter technology have provided new, powerful tools, which together with fluorescent proteins allow for the concurrent tracking of the concentrations of mRNA and protein species of interest. As a result, we can now measure transcription and translation simultaneously with sufficiently high frequency and specificity to directly use the obtained time series data for parameter estimation [15]. This is particularly useful in the rapidly expanding field of synthetic biology, wherein biological 'parts' (e.g., promoters, terminators, genes) can be rapidly combined into 'biocircuits' [23] that may not otherwise exist.

Several examples for modeling of the transcription and translation processes can be found in the literature, ranging from coarse grain [11], [21] to very detailed [9], [1] focusing on different aspects of gene expression. Our aim here is to analyze and improve a previously proposed ODE-based model describing transcription and translation in a cell-free experimental environment using real measurement data. The desired purpose of the model is not only to fit the existing data and showing the dynamics of unmeasured states, but also making predictions that can be validated experimentally.

Previously, we developed a mass action-based model that included a number of experimentally-validated parameters [25], then tested the *in vitro* system under a variety of experimental conditions with improved measurement methodology [20]. However, neither identification related model analysis and nor detailed evaluation of the estimates were provided in these previous papers. Therefore, in this paper we perform further model analysis—such as time-scale separation and structural identifiability—to improve our model. Also, utilizing the comprehensive measurements from [20], we will estimate and validate some of the parameters of the process model.

II. Experimental background

A. *In vitro transcription and translation system*

Cell-free gene expression systems are popular platforms for biocircuit design. Biocircuits are proprietary DNA segments assembled into one or several circular DNA and added to either living cell or some kind of cell extract. These externally supplied genes are expressed in the host environment where they may interact with each other and/or the host environment via protein-protein, protein-DNA, etc. interactions. With this interaction network, various tasks can be performed such as computation, sensing, and actuation.

All experimental data described in this paper were obtained in a cell-free environment derived from *Escherichia coli* crude extract. This extract contains all the endogenous system components necessary for transcription and translation (e.g., ribosomes, RNA polymerase, translation initiation and elongation factors, etc.) but is free of structural components (e.g, cell wall) and genomic DNA. The processed extract is supplemented with molecular energy sources: nucleotides, amino acids, and tRNAs. In this simple form the crude extract with energy source mixture is fully capable of transcription/translation-based biocircuit operation. The detailed description of the system and the preparation steps of the crude extract can be found in [22].

978-3-9524269-4-4 © EUCA

From 'biocircuits' prototyping point of view *in vitro* experiments have advantages over *in vivo* ones, e.g. a short incubation time, repeatability, usage of linear DNA, thus avoiding time-consuming cloning and *in vivo* propagation [23]. Recently, a paper-based solution emerged, further increasing the potential of *in vitro* prototyping [14].

One drawback of current implementations of *in vitro* breadboards is the lack of a "continuous mode" in which gene expression could be sustained for an extended period of time. Furthermore, necessary biomolecular resources are supplied externally in fixed amounts only at the outset of the experiment, and no internal source is available for the resource replenishment. Since resources are limited in this environment, resource competition can arise [27]. A model that explicitly accounts for resources can reveal dynamics of resource allocation that are difficult or impossible to measure.

B. Measurements

Transcription was monitored using the Malachite Green aptamer (MGApt), a short RNA segment that binds the Malachite Green (MG) dye (triphenylmethane) [8] and enhances the dye's fluorescence. The MGApt was placed in the $3'$ untranslated region (UTR) of a gene encoding Green Fluorescent Protein (GFP), which serves as the translational reporter. Production of the combined MGApt-GFP construct is driven by a strong constitutive promoter.

Experiments took place in a 10 μl reaction volume at 29 °C over 14 hours, with each experiment repeated three times. DNA concentration was varied between 0.01 nM and 20 nM (see Figure 1 in [20]). Fluorescence was measured every 3 minutes for both MGApt (excitation: 610 nm; emission: 650 nm) and GFP (excitation: 485 nm; emission: 525 nm) in a Biotek plate reader. All measurements were background corrected to account for fluorescence of the MG dye. Further details of the measurements and sample preparation can be found in [20].

III. Process Model

We chose a mass action kinetics (MAK) framework for modeling transcription and translation in the *in vitro* system. Unlike many of the common models used in synthetic biology, our model explicitly accounts for resource consumption in order to cover resource limits and resource sharing effects. Mass action–based modeling has a number of advantages, including the ease with which stochastic solvers can be applied (e.g., to investigate the dynamics in the non-deterministic regime [12]), as well as certain strong statements that may be made about mass action systems using Chemical Reaction Network Theory, even without knowledge of the model parameters [7], [6], [19].

The model of transcription-translation is written in the following general non-linear form:

$$\dot{x} \;=\; f(x, P),\; x(0) = x_0,$$

where $x : \mathbb{R} \to \mathbb{R}_+^n$ is the state vector, $P \in \mathbb{R}^m$ denotes the vector of model parameters, in our case $n = 14$ and $m = 18$. Table I lists the non-zero initial values used in this paper.

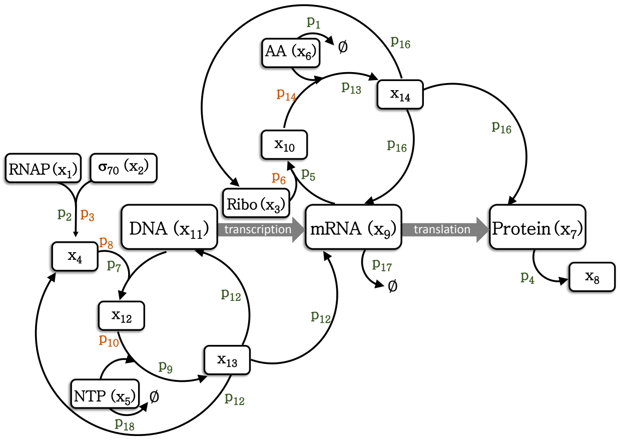

Fig. 1. Overview of the process model. The model is built around the central dogma of molecular biology with additional steps accounting for resource consumptions and degradations. Forward and reverse reaction rate coefficients are denoted with green and orange colors, respectively. The $x_1, \ldots, x_{14}$ are the species concentrations. The parameters p_{11} and p_{15} are not shown, see Appendix for details.

The dynamics of the transcription and translation is given by the following system of ODEs:

$$\begin{aligned}
\dot{x}_1 &= -F_1 & \dot{x}_2 &= -F_1 \\
\dot{x}_3 &= -F_3 + F_{10} & \dot{x}_4 &= F_1 - F_4 + F_7 \\
\dot{x}_5 &= -F_5 - F_{12} & \dot{x}_6 &= -F_8 - F_{13} \\
\dot{x}_7 &= -F_2 + F_{10} & \dot{x}_8 &= F_2 \\
\dot{x}_9 &= -F_3 + F_7 + F_{10} - F_{11} & \dot{x}_{10} &= F_3 - F_8 + F_9 \\
\dot{x}_{11} &= -F_4 + F_7 & \dot{x}_{12} &= F_4 - F_5 + F_6 \\
\dot{x}_{13} &= F_5 - F_6 - F_7 & \dot{x}_{14} &= F_8 - F_9 - F_{10}.
\end{aligned} \tag{1}$$

The $F_i, i = 1, \ldots, 13$ appearing in the above ODEs are the following:

$$\begin{aligned}
F_1 &= p_2 x_1 x_2 - p_3 x_4, & F_2 &= p_4 x_7, \\
F_3 &= p_5 x_9 x_3 - p_6 x_{10}, & F_4 &= p_7 x_{11} x_4 - p_8 x_{12}, \\
F_5 &= p_9 x_{12} x_5 - p_{10} x_{13}, & F_6 &= p_{11} x_{13}, \\
F_7 &= p_{12} x_{13}, & F_8 &= p_{13} x_{10} x_6 - p_{14} x_{14}, \\
F_9 &= p_{15} x_{14}, & F_{10} &= p_{16} x_{14}, \\
F_{11} &= p_{17} x_9, & F_{12} &= p_{18} x_5, \\
F_{13} &= p_1 x_6,
\end{aligned} \tag{2}$$

where $p_i \in \mathbb{R}_+, i = 1, \ldots, 18$ are the positive model parameters (reaction rate coefficients). The elements of the state vector x are the species concentrations. The operation of the modeled process with the roles of the individual species is briefly described below.

a) Transcription dynamics: Transcription of DNA (x_{11}) in *E. coli* begins when the sigma factor (x_2) and the core RNA polymerase (x_1) forms a complex (x_4) and thus facilitates the binding of the RNAP holoenzyme to the promoter region on the DNA (x_{12}). The RNAP then builds the mRNA (x_9) through a process of transcription initiation and elongation, taking from the pool of available RNA nucleotides (ATP, GTP, CTP, UTP), denoted as x_5. This process is depicted in Figure 1. It should be noted that, as a modeling assumption we do not distinguish between the functional segments of the DNA (e.g, promoter, terminator, etc.) and instead model it as just a single species.

The crude extract contains the whole spectrum of mRNA degradation enzymes, thus they attack the mRNA in each complexes (x_9, x_{10}, x_{14}) and render functional mRNAs non-functional, but we can only measure the aggregated effect of these degradation pathways. Thus, a model with multiple degradation pathways for mRNA can lead to structural identifiability issues, since many parameter combinations of the pathways can produce the same rate of overall mRNA degradation. Our recent investigation of the saturation of degradation enzyme capability suggest that we can describe the mRNA degradation by the a first order reaction, see Figure S2 in [20].

b) Translation dynamics: As in the previous step, translation initiation and elongation are modeled as a one-step process. The ribosome (x_3) binds the free mRNA (x_9) and incorporates amino acids (x_6) carried by tRNAs into the growing polypeptide chain. Since the tRNA and the AAs are supplemented externally to the crude extract in high quantity, charged tRNA is always abundant in the system and we can neglect the tRNA charging dynamics. When translation is terminated, x_{14} dissociates and the translated protein (x_7) is released as shown in Figure 1. The variant of GFP (x_7) that we used in the experiments requires 5-7 mins (p_4) to develop the fluorophore and become visible (x_8) [17]. The maturation is modeled as a first order reaction (Flux: F_2).

The protein level does not achieve a steady state via balance of production and degradation; the final protein concentration—due to lack of protein degradation—is fixed when the system runs out of resources.

c) Resource degradation: The dynamics of *in vitro* systems are largely influenced by the finite amount of resources and the change of conditions (e.g., waste accumulation, pH change, etc.) during biocircuit operation. This has been known for some time; an early paper on cell-free expression highlighted how ATP degradation leads to a decrease in protein production [13]. Recent experimental and computational studies on resource effects have shown that the operational lifetime of a system can be extended by maintaining optimal pH and replenishing resources in such a way that enzymes in the system remain functional over long periods of time [21], [17], [18]. These earlier findings led us to incorporate degradation of transcriptional and translational resources as necessary components of our process model.

Because of these findings, we model the degradation of transcriptional resource (x_5) as a first order reaction (Flux: F_{12}) and it is sufficient to capture the decay of transcriptional activity observed in the measurements.

Similar modeling assumption can be made for the translational resources. In our previous work we showed that with additional nucleotides, transcription produces significantly more mRNA, but translation output is roughly the same (Figure 3 in [20]). To accommodate this observation in our model, we include a same type of degradation (Flux: F_{13}) for translational resources (x_6).

TABLE I

SPECIES IN THE MODEL WITH NON-ZERO INITIAL CONCENTRATIONS. RESOURCE (R) TYPE SPECIES ARE ESTABLISHED BY THE CRUDE-CELL EXTRACT PROTOCOL [22]. THE VALUES IN CASE OF ENZYME (E) TYPE SPECIES ARE TAKEN FROM THE LITERATURE. WE TOOK THE AVERAGE CONCENTRATION OF THE NUCLEOTIDES (ATP, CTP, GTP, UTP) AND DENOTED THE AVERAGE AS NTP. (THE SAME GOES FOR AMINO ACIDS)

Species	State	Initial value		Source
[NTP]	x_5	1.2 mM	R	Protocol in [22]
[AA]	x_6	1.5 mM	R	Protocol in [22]
[RNAP]	x_1	100 nM	E	Table S5 in [18]
[Ribo]	x_3	1000 nM	E	Table S5 in [18]
[σ_{70}]	x_2	35 nM	E	Table S5 in [18]

1) Measured outputs: The measured MGApt signal is the experimental measure of mRNA concentration and thus the total concentration of mRNA within the system can be calculated via

$$[mRNA]^{tot} = x_9 + x_{10} + x_{14}. \tag{3}$$

The matured GFP protein is a final product in the system that exists only in one single-specie complex (x_8). Thus observed outputs can be written as

$$\begin{aligned} h_1(x) &= S_1[mRNA]^{tot}, && (4)\\ h_2(x) &= S_2 x_8. && (5) \end{aligned}$$

Each output was converted to nM via scaling factors $S_1 = 7.75$ a.u./nM and $S_2 = 1.723$ a.u./nM. These conversion factors were calculated from the calibration curves using purified MGApt and deGFP as described in [20].

IV. ANALYSIS OF THE PROCESS MODEL

The goal of this section is twofold. First, we briefly examine the dynamics of the process model. Then, we check the model structure itself whether it is theoretically possible to uniquely determine the parameters of the process model.

A. mRNA dynamics

During the experiments we observed that a peak in mRNA production occurs around 150 min (see Figure 2., left panel). By doing simple calculations, we can find a relation between reaction rates that is valid in that time instant. The total mRNA concentration is given by Equation (3). From this, we can calculate that an extremum in mRNA concentration may occur when $\dot{x}_9 + \dot{x}_{10} + \dot{x}_{14} = 0$. Then, we obtain

$$p_{12}x_{13} - p_{17}x_9 = 0. \tag{6}$$

In Equation (6) the first term is the transcription rate and the second part is the degradation rate of mRNA. On the other hand, the value of x_{13} depends on the concentration of NTP (x_5), which decreases over time. Therefore, the maximum of mRNA level occurs when the two terms in Equation (6) are equal.

B. Steady state assumption

We can somewhat simplify the model by considering the fact that the sigma factor (x_2) binds with the RNAP (x_1) on a time-scale that is much faster than those of other reactions. Thus, we assume that $\dot{x}_1 = 0$ and $\dot{x}_2 = 0$, and then the differential equation for state x_4 becomes

$$\dot{x}_4 = -F_4 + F_7. \tag{7}$$

This way, we do not have to consider the dynamics of x_1 and x_2 and estimate or find values from literature for their parameters (p_2, p_3). However, we need to estimate the initial value of x_4 denoted by x_{4init}.

C. Structural Identifiability

At this point we can check whether it is theoretically possible to determine the model parameters based on the model structure and the observables ($h_1(x), h_2(x)$). The parameterized model in (1) is structurally identifiable if

$$h(x(t), P'_{id}) = h(x(t), P''_{id}) \quad \forall t \implies P'_{id} = P''_{id}, \tag{8}$$

where $h(x, P_{id})$ is the output function with parameter vector P_{id}. According to this definition, a structurally non-identifiable system may produce exactly the same output for different parameterizations.

There are many approaches to check structural identifiability of a nonlinear system [4]. We choose the generating series approach for structural identifiability analysis of the process model.

A generating series can be created when the output functions are expanded in a series w.r.t inputs and time. The coefficients of the series are $h(x_0, P_{id})$ and Lie derivatives the output functions along vector field f evaluated at the initial time point. Then, a vector, denoted by $s(P_{id})$, can be formulated with the coefficients of the generating series. A system of nonlinear equations is defined as $s(P_{id}) = c$, where c is an arbitrary constant. Finally, we try to solve this system of nonlinear equations for P_{id}. The existence and uniqueness of the solution defines the structurally identifiability of the model, see [4] for details. We used this approach and checked the model structure and its output with the GenSSI toolbox [4].

Calculating the generating series is computationally intensive, and the computation time rapidly grows by increasing of the number of parameters checked by the algorithm. Thus, we had to limit the number of parameters in the identifiability analysis by assuming that p_{17} (mRNA degradation) and the initial value of x_3 (Ribosome concentration) are known. After that, we checked a model with 12 reaction rate coefficients and one initial value denoted by vector $P_{id} \in \mathbb{R}^{13}_{+}$. According to the report generated by the GenSSI toolbox the set of nonlinear equations had more than one solution, we can thus conclude that our model is at least locally identifiable.

V. Parameter Estimation

Our parameter estimation procedure is based on the commonly applied minimization of the distance between the measured and model computed output. The statistical evaluation and validation of the parameters is also carried out to investigate the quality of the parameter estimation and to identify parameters where further experiments may need to decrease uncertainty.

A. Prediction error minimization

Figure 2 shows the mRNA and GFP measurements with four different initial DNA concentrations, each measurements was repeated three times.

After model reduction and taking the parameter values from the literature into account we can formulate the $P_{est} \in \mathbb{R}^{15}_{+}$ parameter vector, which consists of 13 reaction rate coefficients $[p_1, p_5, \ldots, p_{10}, p_{12}, \ldots, p_{14}, p_{16}, \ldots, p_{18}]$ and two initial values ($x_{3_{init}}$ and $x_{4_{init}}$). Since a kinetic system requires positive parameter values, we restrict the range of possible parameters onto the positive orthant with appropriate lower limits. Different starting points for the parameter estimation were generated with hypercube sampling from a uniform distribution.

The model has two measured outputs $h_1(t)$ and $h_2(t)$ for the mRNA and GFP concentrations, respectively. To incorporate the multiple outputs into the cost function we normalize each term with the maximum of the corresponding time series data $\bar{y}_k(t) = y_i(t)^{(k)}/max(y_i^{(k)}(t))$, where k is the index of consecutive experiments with different initial DNA concentrations and i denotes the measured outputs. We did the same normalization with the model output $\bar{h}_i^{(k)}(t) = h_i^{(k)}(t)/max(y_i^{(k)}(t))$. This leads to the following cost function

$$C(P_{est}) = \sum_{k=1}^{M} \sum_{t=1}^{T} \left[\bar{y}_1^{(k)}(t) - \bar{h}_1^{(k)}(x(t), P_{est})\right]^2 + \left[\bar{y}_2^{(k)}(t) - \bar{h}_2^{(k)}(x(t), P_{est})\right]^2, \tag{9}$$

where $M = 4$ is the number of different experiments we consider and T=280 is the number of samples and samples are taken in every 3 min.

$$\begin{aligned} &\operatorname*{argmin}_{P_{est} \in \mathcal{P}} C(P_{est}) \\ &0 \leq P_{est} \leq UB, \end{aligned} \tag{10}$$

where $\mathcal{P}$ is the set of feasible parameters and UB is the vector of upper bounds. The optimization stated in Equation (10) was performed with a gradient-free global optimizer implemented as a pattern-search [3]. Mainly, to avoid interference between the accuracy of the ODE solver and the finite differentiation for gradient calculation which commonly occur when gradient based optimization is applied in this setup.

During the parameter estimation we found numerous local minima, where the mRNA degradation varied over several orders of magnitude. In our previous study [20], we conducted independent measurement of mRNA degradation in the *in vitro* system (see Section II for the details). From that study, the measurement puts the mRNA half-life in a range of 12–16 min [18]. Thus, we used that information to restrict

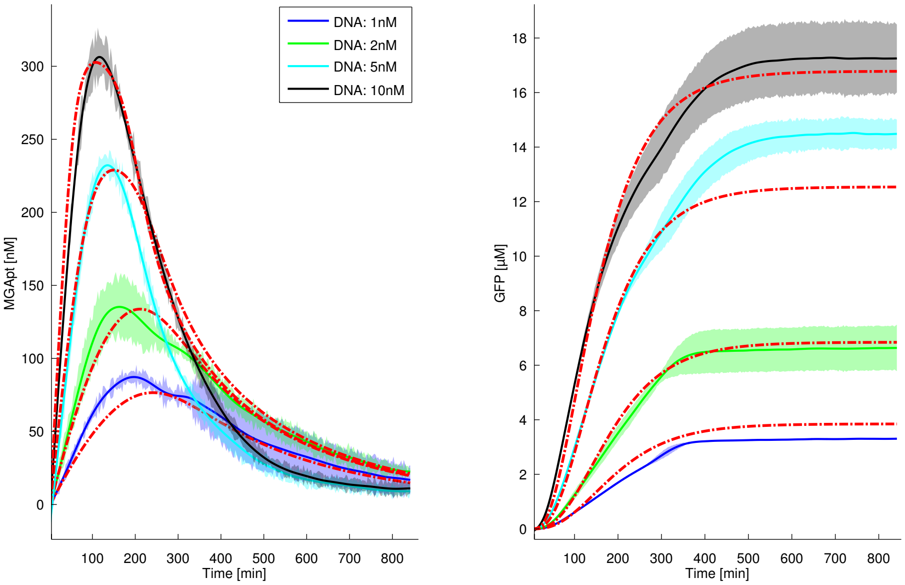

Fig. 2. Simulations with the estimated parameter set is shown in red. The Figure contains time series measurement for both channels with different initial concentration of plasmid DNA (1 nM, 2 nM, 5 nM and 10 nM plasmid DNA concentration, shown in green, red, cyan and black respectively). The left panel shows the dynamics of MGApt, which is proportional to the mRNA concentration. The GFP dynamics is shown on the right panel. The fluorescent counts for each channel have been converted with Equations (4) and (5) to nM and μM, respectively.

TABLE II
NUMERICAL RESULT OF THE PARAMETER ESTIMATION. THE TABLE HAS 13 REACTION RATE COEFFICIENTS AND TWO INITIAL CONCENTRATIONS.

	Name	Value	Confidence Interval	Unit
1	p_1	1.90×10^{-4}	$[1.53 \times 10^{-4}\ 2.27 \times 10^{-4}]$	1/s
2	p_5	6.94×10^{-2}	$[2.87 \times 10^{-2}\ 1.10 \times 10^{-1}]$	$1/(\mathrm{s} \cdot \mathrm{nM})$
3	p_6	8.43×10^{-1}	$[3.42 \times 10^{-1}\ 1.34 \times 10^{0}]$	1/s
4	p_7	7.00×10^{-3}	$[2.89 \times 10^{-3}\ 1.11 \times 10^{-2}]$	$1/(\mathrm{s} \cdot \mathrm{nM})$
5	p_8	1.38×10^{2}	$[-7.72 \times 10^{1}\ 3.54 \times 10^{2}]$	1/s
6	p_9	6.20×10^{-2}	$[1.31 \times 10^{-2}\ 1.11 \times 10^{-1}]$	$1/(\mathrm{s} \cdot \mathrm{nM})$
7	p_{10}	3.55×10^{-1}	$[-4.02 \times 10^{-1}\ 1.11 \times 10^{0}]$	1/s
8	p_{12}	8.90×10^{-3}	$[7.79 \times 10^{-3}\ 1.00 \times 10^{-2}]$	1/s
9	p_{13}	5.95×10^{-2}	$[2.63 \times 10^{-2}\ 9.26 \times 10^{-2}]$	$1/(\mathrm{s} \cdot \mathrm{nM})$
10	p_{14}	2.33×10^{5}	$[-1.10 \times 10^{5}\ 5.76 \times 10^{5}]$	1/s
11	p_{16}	3.02×10^{-1}	$[-7.24 \times 10^{-2}\ 6.76 \times 10^{-1}]$	1/s
12	p_{17}	1.07×10^{-3}	$[9.32 \times 10^{-4}\ 1.21 \times 10^{-3}]$	1/s
13	p_{18}	2.98×10^{-4}	$[2.48 \times 10^{-4}\ 3.47 \times 10^{-4}]$	1/s
14	x_{3_0}	375.58	[253.22 497.95]	nM
15	x_{4_0}	276.13	[242.70 309.55]	nM

the mRNA degradation rate (p_{17}) in the parameter estimation process.

The result of the parameter estimation is summarized in Table II and simulations of the estimated parameter set (solid red lines) overlapped with the measurements is shown in Figure 2. Then, we validated the estimated parameter set over a different range of data (0.1 nM, 0.2 nM, 0.5 nM of plasmid DNA), on average we have 20% error in the final GFP production, but the qualitative features of the simulations are still acceptable in comparison with the measurements.

a) Numerical Implementation: The simulated ODE model is stiff — most likely as a result of the NTP (x_5) and AA (x_6) consumption — thus we used the efficient CVODES solver [16] to solve the model ODEs, also ODEmex software was applied for further speed gain [26].

B. Statistical Analysis of the Point Estimation

To evaluate the quality of the parameter estimation we used a Markov-Chain Monte Carlo (MCMC) implementation to generate the joint posterior distribution of the parameters. The MCMC procedure was initiated at the result of the point estimation with uniform prior and with the same lower and upper bounds that was used in Equation (10). The detailed version of the algorithm can be found in [10]. From the results of the MCMC algorithm, either through counting statistics or through the covariance matrix we can establish the confidence intervals for the parameters [2].

We ranked and listed the cross-correlations larger than 0.5 in Table III. It shows three groups of cross-correlations. In the first one, the translation rate (p_{16}), the translational resource binding (p_{14}) and the ribosome bindings (p_5, p_6) are highly correlated. This may suggest that there is not enough information in the measurement data to determine the correct parameters for all stages of translation. In the second group, there is a cross-correlation between the promoter strength (p_8) and translational reactions (p_{14}, p_{16}). In the third group, the same translation reaction coefficients (p_{14}, p_{16}) are grouped with the initial concentration of the sigma factor activated RNA polymerase ($x_{4_{init}}$).

The result of the cross-correlation analysis will certainly help in design future experiments to improve parameter estimations. Most of the parameters in Table III are related to translation. In order to get a better estimate of these parameters, we have to manipulate ribosome binding strength and/or ribosome concentration in the *in vitro* system. Considering the confidence intervals of the parameters we can see that some of the parameters are accurately estimated, e.g. p_1, p_{18} (resource degradation), p_{12} (transcription rate). This suggests

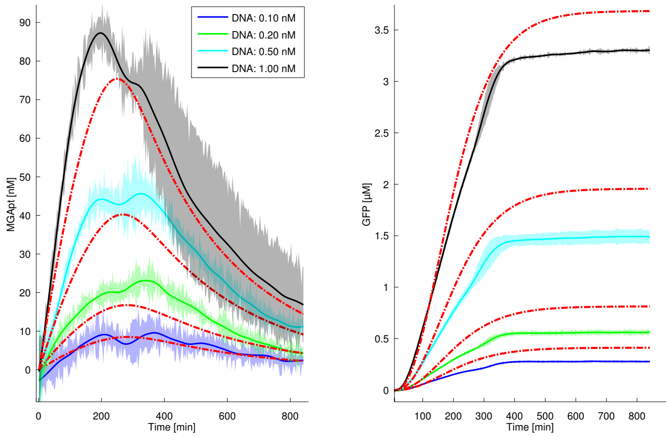

Fig. 3. Validation of the estimated parameter set over another set of data (0.1 nM, 0.2 nM, 0.5 nM of plasmid DNA). The red curves show the corresponding simulations with the parameters from Table II. On average there is 20% error in the final value of GFP, the simulation matches the time series data. The 1 nM data (black curve) was used for estimation, shown here only for comparison.

TABLE III
THE TABLE CONTAINS PARAMETER PAIRS WITH THE STRONGEST CROSS-CORRELATION.

Parameter	Parameter	Correlation
p_{14}	p_{16}	0.942
p_5	p_6	0.929
p_8	p_{14}	0.702
p_8	p_{16}	0.645
p_{14}	$x4_{init}$	0.574
p_{16}	$x4_{init}$	0.551

that we have a good estimate for the resource degradation and some of the transcription related parameters, e.g. p_{12}, p_7 and p_9. On the other hand, confidence intervals for p_8 (promoter strength), p_{10} (NTP binding), p_{14} (AA binding) and p_{16} (translation rate) are very large. Therefore, we need another way to increase our confidence in these parameters. One possibility was highlighted at the cross-correlation analysis, but these resource binding related parameters can be measured with tedious experiments (if measurement is possible at all). Therefore, it may necessary to do some sort of model reduction to improve the estimation result.

Besides analyzing the confidence intervals and evaluating the cross-correlation, we can take advantage of the fact that we calculated the full joint posterior distribution of the parameters. Hence, we can select regions from the parameter distribution (via confidence intervals) and take all the parameters from a selected region and simulate them. Roughly speaking, we can visualize how the dynamics is 'spreading' by the variations of the parameters. The results computed by using the parameters in Table II is shown in Figure 4. where the general shape of the dynamics is the same over different confidence intervals (99%, 95%, 90%, 50%). Although higher uncertainty arises around the mRNA peak time and peak value and around the steady-state level of GFP.

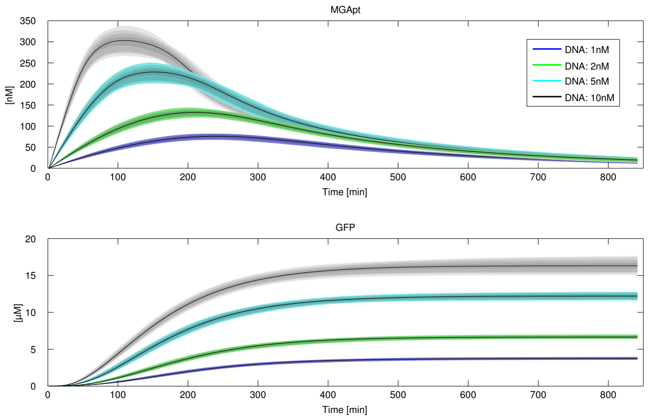

Fig. 4. Samples for the joint posterior distribution of parameter sets with different confidence intervals (99%, 95%, 90%, and 50%).

VI. Conclusion

Building upon our previous modeling work [25] and extensive experimental data collection [20], we further refined our model describing a transcription/translation process in a cell-free environment by introducing direct degradation of transciptional and translational resources and simplifying the mRNA degradation mechanism. The demostrated simple model analysis explained the observed mRNA dynamics and allowed steady state assumption-based model reduction. To ensure proper model structure, we checked—the often negleceted—structural identifiability of the improved model. This model turned out to be at least locally structurally identifiable, the provides a good foundation for parameter esetimation. For the parameter estimation, we applied a derivative-free pattern search method accommodating multi-channel multi-experiments data. The resulting parameter set was statistically evaluted and validated on a different data set. Statistical analysis revealed some uncertain parameters that we attempted to explain from biochemical and experimental points of view. Based on these results, we can focus our future work to comutationally assist the experiment design and possibly to make further reductions in the dynamical model.

VII. Appendix

A. Dealing with large stoichiometric coefficients

In both transcription and translation, elongation and termination steps are not modeled as individual reactions. In a naive implementation of these chemical reactions, problems arise due to the large stoichiometric coefficients for NTP (x_5) and AA (x_6), which depends on the length of transcript (around 1000 bases for our example), thereby increasing reaction order and yielding stiff ODEs.

To address this issue, we first lump NTP (x_5) and AA (x_6) into units of 100, bringing the order of the reaction down to ∼10. We then introduce an auxiliary reaction (F_6) with a rate coefficient $p_{11} = (b-1)p_{12}$, it is $b-1$ times higher than the transcription reaction rate coefficient, where b is one hundredth of the gene length. With this

TABLE IV

SPECIES - STATE VARIABLE

Species	State variables
[RNAP]	x_1
[σ]	x_2
[Ribo]	x_3
[RNAP:σ]	x_4
[NTP]	x_5
[AA]	x_6
[GFP]	x_7
[GFP*]	x_8
[mRNA]	x_9
[mRNA:Ribo]	x_{10}
[DNA]	x_{11}
[DNA:RNAP:σ]	x_{12}
[NTP:DNA:RNAP:σ]	x_{13}
[AA:mRNA:Ribo]	x_{14}

auxiliary reaction, the reaction order for translation is one. Analogously, the reaction order for translation becomes one with a similar auxiliary reaction (F_9) for AA consumption ($p_{15} = b/3 \cdot p_{16}$).

VIII. ACKNOWLEDGMENTS

This work was partially supported by the Hungarian Scientific Research Fund under grant no. OTKA NF104706 and by the DARPA Living Foundries Program under Contract HR0011-12-C-0065. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We would like to thank Vipul Singhal for his contribution on the biomolecular model. The authors also thank Dávid Csercsik and Petra Rovó for their helpful comments and suggestions.

REFERENCES

[1] Sabine Arnold, Martin Siemann, Kai Scharnweber, Markus Werner, Sandra Baumann, and Matthias Reuss. Kinetic modeling and simulation of *in vitro* transcription by phage T7 RNA polymerase. *Biotechnology and Bioengineering*, 72:548–561, 2001.

[2] Richard C. Aster, Brian Borchers, and Clifford H. Thurber. *Parameter Estimation and Inverse Problems*. Elsivier, 2012.

[3] Charles Audet and Jr. J. E. Dennis. Analysis of generalized pattern searches. *SIAM J. Optim.*, 13:889–903, 2006.

[4] O. Chis, J. R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 27:2610–2611, 2011.

[5] G. Craciun and C. Pantea. Identifiability of chemical reaction networks. *Journal of Mathematical Chemistry*, 44:244–259, 2008.

[6] G. Craciun, Y. Tang, and M. Feinberg. Understanding bistability in complex enzyme-driven reaction networks. *Proc Natl Acad Sci USA*, 103 (23):8697–8702, 2006.

[7] M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors - I. The deficiency zero and deficiency one theorems. *Chemical Engineering Science*, 42 (10):2229–2268, 1987.

[8] Dilara Grate and Charles Wilson. Laser-mediated, site-specific inactivation of RNA transcripts. *Proc Natl Acad Sci USA*, 96:6131–6136, 1999.

[9] Sandra J Greive, Jim P Goodarzi, Steven E Weitzel, and Peter H von Hippel. Development of a "modular" scheme to describe the kinetics of transcript elongation by rna polymerase. *Biophys J*, 101(5):1155–1165, Sep 2011.

[10] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7:191–379, 2001.

[11] Eyal Karzbrun, Jonghyeon Shin, Roy Bar-Ziv, and Vincent Noireaux. Coarse-grained dynamics of protein synthesis in a cell-free system. *Phys Rev Lett*, 106(4):048104, January 2011.

[12] Shev Macnamara, Alberto M Bersani, Kevin Burrage, and Roger B Sidje. Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation. *J Chem Phys*, 129(9):095105, Sep 2008.

[13] S. V. Matveev, L. M. Vinokurov, L. A. Shaloiko, C. Davies, E. A. Matveeva, and Alakhov YuB. Effect of the ATP level on the overall protein biosynthesis rate in a wheat germ cell-free system. *Biochim. Biophys Acta*, 1293(2):207–212, Apr 1996.

[14] Keith Pardee, Alexander A. Green, Tom Ferrante, D. Ewen Cameron, Ajay DaleyKeyser, Peng Yin, and James J. Collins. Paper-based synthetic gene networks. *Cell*, in press, 2014.

[15] Andreas Raue, Marcel Schilling, Julie Bachmann, Andrew Matteson, Max Schelke, Daniel Kaschek, Sabine Hug, Clemens Kreutz, Brian D. Harms, Fabian J. Theis, Ursula Klingmüller, and Jens Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLOS ONE*, 2013.

[16] R. Serban and A. C. Hindmarsh. Cvodes: the sensitivity-enabled ode solver in sundials. *Proceedings of IDETC/CIE 2005*, 2005.

[17] Jonghyeon Shin and Vincent Noireaux. Efficient cell-free expression with the endogenous e. coli rna polymerase and sigma factor 70. *J Biol Eng*, 4:8, 2010.

[18] Jonghyeon Shin and Vincent Noireaux. An e. coli cell-free expression toolbox: Application to synthetic gene circuits and artificial cells. *ACS Synth Biol*, 1(1):29–41, January 2012.

[19] G. Shinar and M. Feinberg. Structural sources of robustness in biochemical reaction networks. *Science*, 327:1389–1391, 2010.

[20] Dan Siegal-Gaskins, Zoltan A. Tuza, Jongmin Kim, Vincent Noireaux, and Richard M. Murray. Gene circuit performance characterization and resource usage in a cell-free "breadboard". *ACS Synth Biol*, 3:416–25, 2014.

[21] Tobias Stögbauer, Lukas Windhager, Ralf Zimmer, and Joachim O. Rädler. Experiment and mathematical modeling of gene expression dynamics in a cell-free system. *Integr. Biol.*, 4:494–501, 2012.

[22] Zachary Z Sun, Clarmyra A Hayes, Jonghyeon Shin, Filippo Caschera, Richard M Murray, and Vincent Noireaux. Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology. *J Vis Exp*, 79:e50762, 2013.

[23] Zachary Z Sun, Enoch Yeung, Clarmyra A Hayes, Vincent Noireaux, and Richard M Murray. Linear dna for rapid prototyping of synthetic biological circuits in an escherichia coli based tx-tl cell-free system. *ACS Synth Biol*, 3:387–397, 2014.

[24] G. Szederkényi, J. R. Banga, and A. A. Alonso. Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Systems Biology*, 5:177, 2011.

[25] Zoltan A Tuza, Vipul Singhal, Jongmin Kim, and Richard M Murray. In silico modeling toolbox for rapid prototyping of circuits in a biomolecular "breadboard" system. In *Decision and Control (CDC), 2013 51th IEEE Conference on*, pages 1404 – 1410, 2013.

[26] J. Vanlier, C. A. Tiemann, P. A J Hilbers, and N. A W van Riel. An integrated strategy for prediction uncertainty analysis. *Bioinformatics*, 28(8):1130–1135, Apr 2012.

[27] Enoch Yeung, Jongmin Kim, Ye Yuan, Jorge Goncalves, and Richard M Murray. Quantifying crosstalk in biochemical systems. In *51st IEEE Conference on Decision and Control (CDC)*, pages 5528–5535, Maui, HI, December 2012.