A qualitative method for parameter estimation in gene regulatory networks using extended Kalman filtering

Mattia Petroni[1], **Luigi Canciani**[2], **Miha Mraz**[1] **and Miha Moškon**[1]

[1]*University of Ljubljana,*
Faculty of Computer and Information Science,
Tržaška 25, 1000 Ljubljana, Slovenia

[2]*Società Italiana di Medicina Generale,*
Health Search Institute,
Via Sestese 61, 50141 Firenze, Italia

E-mail: mattia.petroni@gmail.com

Abstract

Gene regulatory networks present a complex logic control of the cellular metabolism in all living organisms. Over the last decade, study of their synthetic implementation has made an intensive use of accurate modeling and simulation approaches. These are essentially based on the models of cellular activity as dynamical systems defined by complex biochemical processes and reactions, which are described by various numerical parameters. Unfortunately, most of these parameters are usually unknown. In this paper we present a parameter estimation method using extended Kalman filtering and χ^2 statistical test for model validation. The extended Kalman filtering approach is suitable for many improvements in the meaning of computational complexity and model verification in the contrast to other state of the art methods. As a reference model, we propose a synthetic gene regulatory network, which exhibits an oscillatory behavior for a different range of parameters.

1 Introduction

In the last decade, synthetic biology has reached impressive achievements, due to the increasing interest in using its techniques in applied clinical sciences and embedded therapy such as clinical immunology, gene therapy and many other applications from various scientific areas. A central point in synthetic biology is the engineering approach of the system's objectives by an accurate mathematical modeling of gene expression processes and gene regulatory networks (GRNs). The goal of modeling is to design a biological system by imitating its functionality in a formal mathematical way. An impressive development in this modeling stage came up from control theory, which is based on the presentation of GRNs as complex control systems. Many mathematical constructs can be used for modeling and simulating the reaction networks and gene expression in this context. Typical examples are deterministic and stochastic approaches, which were studied vastly in the literature [15, 16].

Deterministic models are commonly used by biologists, where the biochemical reactions, which define the quantitative aspects of the underlying system, are precisely described with deterministic (non)linear ordinary differential equations (ODE). With this approach the problem of values of the reaction-kinetic rates may arise, because parameters that define ODE models are often unknown. Hence parameter fitting is required for a successful model definition. A correct estimation of these values can result in a more complete model for the validation of several biological applicable systems such as molecular drug delivery. Furthermore, a precise model affinity is crucial, if we want to develop future computational platform inside the cellular environment [15, 16].

In this paper we present the basic approach of extended Kalman filtering and χ^2 test for estimation and evaluation of unknown parameters in biological systems models. Section 2 presents the related work of parameter estimation techniques. Section 3 describes the extended Kalman filtering approach and the χ^2 statistical test for models validation. An example of synthetic gene network with some unknown parameters values is then presented in section 4 as a possible model for investigation of described method. Finally, possible improvements and suggestions for further work are introduced in section 5.

2 Related Work

The extended Kalman filter has become a *de-facto* standard for parameter estimation in control theory and in this paper we propose its application for estimating reaction rates and kinetic constants in synthetically designed GRNs. The main disadvantage of the state-of-the-art methods is their computational complexity, when applied to models with high number of unknown parameters. Typically, a model representation of DNA based logic gates, implemented with GRNs, could hide hundred of unknown kinetic constants. An approach with extended Kalman filtering seems to decrease computational complexity, but a

model validation is still necessary [10]. An additional improvement of the computational complexity of parameter estimation may come from modeling synthetic GRNs with synthetic DNA binding proteins such as zinc-fingers or TAL effectors, which own high similarity in kinetics and therefore could drastically reduce the unknown parameter space.

In the last decade many parameter estimation approaches were proposed, but no standard has been actually defined, because of the intrinsic uncertainty of the underlying biological systems. All presented methods have to face with nonlinear constraints implicit in the ODE models. In [2], some known parameters are used for the first approximation, based on assumption of correctness of the underlying model. Some parameter estimations come out with several order of magnitude of difference, but the fitting evaluation is still acceptable. This "sloppiness" is a common result in many techniques and some authors believe that this is the reason for the difficulties in the nonlinear parameter estimation [9, 10]. In [2, 14] the Levenberg-Marquardt method is used to estimate the parameters' values and a statistical verification is performed to check the model congruence with the experimental measures. In addition, the promoter stability behavior is considered. A novel approach using a Bayesian inference technique with an *a posteriori* analysis of parameter contribution to the model, is shown in [3, 20]. An interesting statistical inference of parameters, that uses noisy microarray data is presented in [4]. An adapted Prediction Error Method for kinetic parameter identification is used in [5]. But the main contributions from control theory in this matter comes from the state estimation techniques by state observers methods and extended Kalman filtering approach [6, 10, 11, 17]. Recently, a comparison of the performance of the extended Kalman filter and the nonlinear least squares fit is studied in [7]. A novel approach with particle filter in contrast with Kalman is described in [12]. A comparison between Gauss-Newton iteration method and the weighted linear least squares, considering rational reaction rates, is performed in [21]. At last a comprehensive review of metaheuristics approaches such as simulated annealing and evolutionary genetic algorithms is described in [19].

3 Extended Kalman Filtering

Our problem domain can be generally formulated in a system of differential equations of the form:

$$\begin{cases} \dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, \theta) + \mathbf{w} \\ \dot{\theta} = 0 \\ \mathbf{x}(t_0) = \mathbf{x}_0 \\ \theta(t_0) = \theta_0 \\ \mathbf{y}_k^{(1)} = h^{(1)}(\mathbf{x}(t_k)) + \mathbf{v}_k^{(1)} \\ \dots \\ \mathbf{y}_k^{(p)} = h^{(p)}(\mathbf{x}(t_k)) + \mathbf{v}_k^{(p)} \end{cases} \tag{1}$$

where $\mathbf{x}$ is the state vector, commonly of chemical species, $\mathbf{u}$ is the input signal vector that may represent environmental constraints, θ is an unknown vector of parameters' constants, f is the derivate function of state variation, $\mathbf{w}$ is the process noise, $\mathbf{y}_k^{(p)}$ is a vector of experimental measurements performed for p different quantities at discrete time $t_k (k = 1, 2, ..., s)$, h is the output function and $\mathbf{v}$ represents the measurements' noise. θ_0 and $\mathbf{x}_0$ are referred to the initial estimation of parameters and states at time t_0. Both $\mathbf{w}$ and $\mathbf{v}$ are assumed to be Gaussian random variables with zero mean and covariance matrix Q and R respectively. Unknown parameters θ are commonly the protein degradation rates, kinetic rates of transcription factors and the association (k_{on}) and dissociation (k_{off}) rates of enzymatic compounds.

The extended Kalman filter is a recursive method for estimating the state of a nonlinear system. According to [10, 18][1], we need to define the initial conditions of the system, precisely the initial estimate of $\mathbf{x}_0$ and error covariance P

$$\begin{aligned} \hat{\mathbf{x}}_0^+ &= \mathbb{E}\{\mathbf{x}_0\}, \\ P_0^+ &= \mathbb{E}\left\{(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+)(\mathbf{x}_0 - \hat{\mathbf{x}}_0^+)^T\right\}, \end{aligned} \tag{2}$$

where the exponent sign $+$ is referred to the evaluation *a posteriori*. Now we can start to evaluate the equations of the extended Kalman filter throughout discrete time intervals $\{t_1, t_2, ..., t_s\}$

$$P_k^- = F_{k-1} P_{k-1}^+ F_{k-1}^T + Q_{k-1} = P(t_k), \tag{3a}$$
$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1}, \tag{3b}$$
$$\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}(t_k), \tag{3c}$$
$$\hat{\mathbf{x}}_k^+ = \hat{\mathbf{x}}_k^- + K_k(\mathbf{y}_k - H_k \hat{\mathbf{x}}_k^-), \tag{3d}$$
$$P_k^+ = (I - K_k H_k) P_k^- (I - K_k H_k)^T + K_k R K_k^T, \tag{3e}$$

where F_{k-1} and H_k are *Jacobians* of f and h respectively, evaluated in previous *a posteriori* estimations $\hat{\mathbf{x}}_{k-1}^+$. To prevent the output of the filter $\hat{\mathbf{x}}_k^+$ in being used outside its values' range, some constraint must be assumed during filtering evaluations [10]. This is a typical instance when we are dealing with concentrations of certain proteins inside the cell. A physical limit is defined and must be considered [1]. In [10], this consideration is presented as an optimization problem. We omit the formulas derivation and we just present the final result, which we will use further in this paper:

$$\begin{aligned} \hat{\mathbf{x}}_{k+1}^+ = \; & \arg\min (\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}^+)^T \\ & (P_k^+)^{-1} (\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}^+), \\ & \text{subject to } D\mathbf{x}_{k+1} \leq d_{k+1}. \end{aligned} \tag{4}$$

At this point an algorithm can be written by following the stages from equations 2 to 3e repeatedly for every time

[1] We refer to [10, 18] for the complete equations' descriptions and derivations of the extended Kalman filter.

step t_k, $k = 1, 2, ..., s$. If the estimation $\hat{\mathbf{x}}_k^+$ suddenly does not satisfy the required constraints, the equation 4 may be used instead of 3d.

Unfortunately the extended Kalman filter can sometimes diverge or divert results estimations. Therefore an empirical test must be performed to check the correctness and reliability of estimated values. The most simple approach is suggested in [10]. Let $\mathbf{x}_{\hat{\theta}_0}(t_k)$ be the solution of the system described in equation 1. Then a statistical test χ^2 can be constructed from the following rearrangement:

$$\hat{\mathbf{v}}_k^{(i)} = \mathbf{y}_k^{(i)} - h_k^{(i)}(\mathbf{x}(t_k)), \tag{5}$$

$$\hat{\sigma}_i^2 \approx \xi_i = \frac{1}{s}\sum_{k=1}^{s}\left(\hat{\mathbf{v}}_k^{(i)}\right)^2, \tag{6}$$

$$\frac{s\,\xi_i}{\chi_{s,1-\frac{\delta}{2}}} \leq \hat{\sigma}_i^2 \leq \frac{s\,\xi_i}{\chi_{s,\frac{\delta}{2}}}. \tag{7}$$

If the real variance σ_i^2, which is equal to the diagonal value of the covariance matrix R, belongs to the same interval as its constructed estimation $\hat{\sigma}_i^2$ in equation 7, then we have a high probability that the result estimation $\mathbf{x}_{\hat{\theta}_0}(t_k)$ also presents a legitimate solution for unknown parameter set θ.

4 Results

As a reference model we propose a synthetic gene regulatory network, which results in an oscillatory behavior, similarly as implemented in [8]. A brief scheme of the network is shown in Fig. 1.

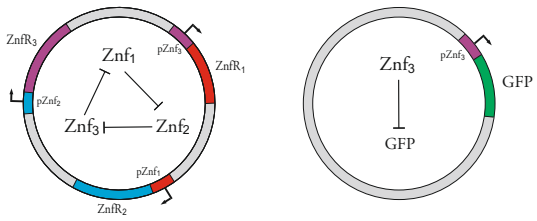

Figure 1: A repressilator circuit implemented with a regulatory repressor chain of three synthetic DNA binding proteins (repressors) i.e. zinc fingers, and a reporter protein i.e. green fluorescent protein - GFP. Its main advantage compared to the model presented in [8], is its scalability potential, i.e. number of repressors can be increased straightforwardly. A successful oscillatory response is though achieved only with odd number of repressors binding sites.

A model of the gene network presented in Fig. 1, can be fully described by a system of reaction network

$$\mathrm{pZnf}_i + \mathrm{RNAP} \underset{\mathrm{k_{off}}}{\overset{\mathrm{k_{on}}}{\rightleftharpoons}} \mathrm{pZnf}_i^*, \tag{8}$$

$$\mathrm{pZnf}_i + \mathrm{ZnfR}_{i-1} \underset{\mathrm{k_d}}{\overset{\mathrm{k_r}}{\rightleftharpoons}} \mathrm{pZnf}_i^{'}, \tag{9}$$

$$\mathrm{pZnf}_i^* \xrightarrow{\mathrm{k_{tr}}} \mathrm{pZnf}_i + \mathrm{mRNA}_i + \mathrm{RNAP}, \tag{10}$$

$$\mathrm{mRNA}_i \xrightarrow{\mathrm{k_{tl}}} \mathrm{mRNA}_i + \mathrm{ZnfR}_i, \tag{11}$$

$$\mathrm{mRNA}_i \xrightarrow{\mathrm{k_{dm}}} \emptyset, \tag{12}$$

$$\mathrm{ZnfR}_i \xrightarrow{\mathrm{k_{dp}}} \emptyset, \tag{13}$$

$$\mathrm{pZnf}_i^{'} \xrightarrow{\mathrm{k_{dr}}} \mathrm{pZnf}_i, \tag{14}$$

where pZnf_i presents free, pZnf_i^* activated and $\mathrm{pZnf}_i^{'}$ repressed promoter state of the i-th zinc finger gene, mRNA_i presents the messenger RNA, ZnfR_i the synthesized zinc finger protein, which behaves as a repressor and RNAP RNA polymerase initiator. Operation $i-1$ is performed modulo n, where n is the number of stages (Fig. 1 presents a scenario where n equals 3).

In order to present the model in the form of equation 1, a system of differential equations must be derived from chemical reactions. An example of simple equation of a zinc finger protein could be

$$\begin{aligned}\frac{\mathrm{d}\left[\mathrm{ZnfR}_i\right]}{\mathrm{d}t} &= \mathrm{k_{tl}}\left[\mathrm{mRNA}_i\right] + \mathrm{k_d}\left[\mathrm{pZnf}_{i+1}^{'}\right] \\ &\quad - \mathrm{k_r}\left[\mathrm{pZnf}_{i+1}\right]\left[\mathrm{ZnfR}_i\right] - \mathrm{k_{dp}}\left[\mathrm{ZnfR}_i\right].\end{aligned} \tag{15}$$

Parameters that need to be evaluated are clearly seen in the ODE presentation. The association and dissociation kinetic constants of zinc fingers and RNA polymerase (RNAP) can be experimentally measured with surface plasmon resonance (SPR) technique (results are shown in Table 1). The transcription and translation rate constants (k_{tr} and k_{tl}) of gene expression are on the other hand, as many other factors such as dimerization and phosphorilation rates, very difficult or even impossible to evaluate experimentally. Estimation of such parameters can be performed with the procedure described in equations 3a-3e.

parameter	measured	reference
$\mathrm{k_{on}}$	$56\ \mu M^{-1}s^{-1}$	[13]
$\mathrm{k_{off}}$	$0.2\ s^{-1}$	[13]
$\mathrm{k_r}$	$0.031\ \mu M^{-1}s^{-1}$	[22]
$\mathrm{k_d}$	$0.0002\ s^{-1}$	[22]

Table 1: Experimentally measured parameter values of association and dissociation constants for T7 RNA polymerase and several zinc-finger DNA binding proteins in *E. coli*.

5 Conclusion

We opted for extended Kalman filtering estimation technique mostly because of its advantages in computational complexity and robustness of value estimation. A statistical test appears to be necessary when performing a reliability validation. Approach presented in Extended Kalman Filtering Section is used for the evaluation of unknown parameters. According to [10], it could also be

used to validate the acceptability of the entire model.

A complete analysis and comparison of implemented method by consideration of real experimental results from SPR, will be a part of further studies, but successful parameter estimations are on the other hand expected shortly. Nevertheless, a stability analysis of the established model should also be performed to avoid the value inconsistencies. Such analysis should help to define the limits of parameters acceptance, for which the model exhibits the expected behavior [16]. Computational complexity of extended Kalman filtering will also be investigated with model of repressilator, consisting of hundred or even more repressors. Complex models with hundreds of parameters may be in the future required for developing more complex logic structures for implementation of computational platforms based on biological systems [15].

Recent works presented many problems, which have to be solved in order to perform χ^2 test on the systems with high number of parameters [10, 11]. In order to overcome these problems, we expect to use novel approaches for parameter estimation, such as soft computing techniques, which are capable of dealing with high dimensional spaces and high number of variables.

6 Acknowledgement

The research was supported by the scientific-research program Ubiquitous Computing (P2-0359) financed by Slovenian Research Agency in years from 2009 to 2012. Results presented here are in scope of PhD thesis that is being prepared by Mattia Petroni.

References

[1] Alon, U.: Introduction to System Biology, Chapman & Hall/CRC, (2007)

[2] Ashyraliyev, M., Jaeger, J., Blom, J.G.: Parameter estimation and determinability analysis applied to Drosophila gap gene circuits, BMC Syst. Biol. **2**:83 (2008)

[3] Barnes, C.P., Silka, D., Sheng, X., Stumpf, M.P.H.: Bayesian design of synthetic biological systems, PNAS **108**:37, 15190-15195, (2011)

[4] Cao, J., Zhao, H.: Estimating dynamic models for gene regulation networks. Bioinformatics, **24**:14, 1619-1624 (2008)

[5] Cinquemani, E., et al.: Stochastic dynamics of genetic networks: modelling and parameter identification. Bioinformatics, **24**:23, 2748-2754 (2008)

[6] Dochain, D.: State and parameter estimation in chemical and biochemical processes: a tutorial, J. Process Contr. **13**, 801-818 (2003)

[7] Dunlop, M.J., Murray, R.M.: Towards Biological System Identification: Fast and Accurate Estimates of Parameters in Genetic Regulatory Networks. Submitted to 45th IEEE DECIS. CONTR. P. (2006)

[8] Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators, Nature **403**:6767, 335-338 (2000)

[9] Gutenkunst, R.N., et al.: Universally sloppy parameter sensitivities in systems biology models, PLoS Comput. Biol. **3**:e189, (2007)

[10] Lillacci, G., Khammash, M.: Parameter Estimation and Model Selection in Computational Biology. PLoS Comput. Biol. **6**:3, (2010)

[11] Lillacci, G., Khammash, M.: Parameter Identification of Biological Networks Using Extended Kalman Filtering and χ^2 Criteria. 49th IEEE DECIS. CONTR. P., 3367-3372, Dec. 2010

[12] Lillacci, G., Khammash, M.: A distribution-matching method for parameter estimation and model selection in computational biology. Int. J. Robust. Nonlinear Control, (2012)

[13] Martin, C.T., A.Újvarí: Thermodynamic and Kinetic Measurements of Promoter Binding by T7 RNA Polymerase, Biochemistry **35**, 14574-14582 (1996)

[14] Mendes, P., Kell, D.B.: Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation, Bioinformatics **14**:10, 869-883 (1998)

[15] Moškon, M., Mraz, M.: Modeling As the Essential Step in the Construction of Biological Computer Structures, International Journal of Information and Education Technology, **1**:3, 185-189 (2011)

[16] Moškon, M.: Computer structures perspective on switching dynamics of simple biological systems, PhD. Dissertation, in preparation, University of Ljubljana (2012)

[17] Quach, M., Brunel, N., d'Alche-Buc, F.: Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference, Bioinformatics **23**:23, 3209-3216 (2007)

[18] Simon, D.: Optimal State Estimations, John Wiley & Sons, Inc., Hoboken, New Jersey, (2006)

[19] Sun, J., Garibaldi, J.M., Hodgman C.: Parameter Estimation Using Metaheuristics in Systems Biology: A Comprehensive Review, IEEE ACM T. COMPUT. BI. **9**:1, 185-202 (2012)

[20] Watanabe, Y., et al.:An estimation method for inference of gene regulatory net-work using Bayesian network with uniting of partial problems, BMC Genomics **13** (2012)

[21] Wu, F., Mu, L., Shi, Z.: Estimation of parameters in rational reaction rates of molecular biological systems via weighted least squares, Int. J. Syst. Sci. **41**:1, 73-80 (2010)

[22] Wei-Ping Yang, H. Wu, C.F. Barbas III: Surface plasmon resonance based kinetic studies of zinc finger-DNA interactions, J. Immunol. Methods. **183**, 175-182 (1995)