A method for model comparison based on the parameter sensitivity measures

Ekaterina Myasnikova

Peter the Great St.Petersburg Polytechnical University, 29 Politekhnicheskaya, 195251, St.Petersburg, Russia,
e-mail: myasnikova@spbcas.ru

Alexander Spirov

Sechenov Institute of Evolutionary Physiology and Biochemistry, 44 Toreza, 194223, St. Petersburg, Russia,
Computer Science and CEWIT, State University of New York Stony Brook, Stony Brook, NY, USA
e-mail: alexander.spirov@gmail.com

Introduction. In modeling of complex biological systems one often faces a dilemma of trade-off between over-simplification of mechanisms underlying the modelled biological processes and the model over-parameterization. In the former case the model may turn to be unrealistic while in the latter case the fitting to experimental data may lead to non-identifiable parameter estimates due to insufficient or too noisy data. This kind of uncertainties constitute the overfitting problem that results in unreliable and non-unique parameter estimates and a poor predictive power of the model. Methods for analysis of parameter sensitivity and identifiability [1] may give a clue to the correct choice of the level of model detail.

Fairly often the whole parameter set is in a natural way subdivided into subsets each characterizing a certain biological input or any biological feature. For instance, gene circuit models [2,3] that dynamically reconstitute the set of interactions within a genetic network are defined by several parameter subsets each corresponding to a specific ligand, transcription factor (TF). In certain cases there is no need to check each individual parameter for identifiability but it is just sufficient to consider parameter subsets in total. For example, such a situation arises when the model is used for predictions of gene expression patterns in null-mutants by setting to zero the parameter subsets corresponding to absent genes. In our previous work [4] we have introduced quantitative measures of prediction power based on estimation of relative sensitivity to parameter subsets and successfully applied them to the gene circuit model presented in [2,3] describing the dynamics of segmentation gene expression in *Drosophila melanogaster*. The measures took into account two sources of non-identifiability of the parameter subset: 1) strong correlations between parameters from the subset and the other ones; and 2) high sensitivity to the rest of parameters while

parameters from the subset do not essentially affect the quality of fit. The measures will be refferred to as Measure 1 and 2, respectively.

We propose a modified version of the method based on the similar principles and designed to compare models of the same biological system but at different levels of detail. To make sure that the model complication is practically reasonable it is necessary to check the identifiability of the subset of parameters additionally introduced into the model. Examples of such models are numerous: in gene circuits one and the same genetic system can be described by different number of TFs, thereby adding to the model a whole subset of parameters; or in models of transcriptional control a TF can be characterized by different number of binding sites (BS) or even a set of BS clusters in *cis*-regulatory modules [5-7] which are characterized by subsets of parameters; extra parameters can correspond to additional non-linear terms; etc.

Method for model comparison. An idea underlying the method is to analyze instead of model sensitivity to each individual parameter the sensitivity to those combinations of parameters that make a maximum impact on the model solution. Such combinations are found for both versions of the model; then the values of sensitivities of the more detailed model to both kinds of combinations are computed, and a ratio of these values is used to construct the criteria. The sensitivity to parameter combinations is characterized by the confidence area of parameter estimates in the vicinity of the model solution.

We define two measures characterizing each of two sources of non-identifiability specified above. The measures have a clear geometrical interpretation: they reflect properties of the confidence area shape, its oblongness and inclination as described in [4]. Very low values (close to zero) of Measure 1 testify for a high correlation between parameters, whereas high values of Measure 2 (close to 1) demonstrate the low sensitivity of the larger model to the subset of additional parameters and hence their poor identifiabilty. In this case one may suspect that the larger model is over-parametrized while experimental data used for fitting contain insufficient information about the additional parameters. Thus we face the overfitting problem.

Results and Conclusion. The method is tested on a nonlinear ODE model imitating gene circuits fitted to synthetically simulated data. Then the method performance is demonstrated

on the model of transcriptional control of the *Drosophila* melanogaster *even-skipped* gene published in [5]. The model is based on the positions and sequence of individual BS on the DNA and quantitative, time-resolved expression data at cellular resolution.

To use the model, a set of BS had to be specified. Initial fitting to data was done using 17 BS of four TFs found by footprint experiments. All TFs were coded by genes belonging to the gap segmentation gene family. While the fit quality was rather high, there were notable patterning defects. The conclusion was made that 17 sites were not sufficient to produce the correct pattern. Thus the model was supplemented by input expression data and 17 BS of 3 additional ligands (Cad, Kni and Tll). The 34-site model showed the improved quality of solutions. In summary, it included Kr, Gt, Kni and Tll as repressors and Bcd, Hb and Cad as activators.

Our method was applied to analyze the identifiability of parameters added to the 34-site model. Two measures were computed for each of three TFs separately. Measure 1 takes on the value close to zero ($< 10^{-5}$) for Cad-associated parameter estimates, that means their strong correlation with the rest of parameters. As for Measure 2, its value for Kni- and Tll-parameter subsets is almost equal to 1 (>0.99999), while for Cad takes on the lower value (0.91). Thus it was shown that the only TF that essentially contributes to the model improvement is the activator Cad, whilst repressors Kni and Tll practically don't affect the model fit. This conclusion is confirmed by directly setting to zero parameters associated with Kni and Tll, that in both cases do not lead to worsening of the model fit quality. In particular, this is true for Kni-parameters, which provide the highest value of Measure 2 and whose exclusion from the model even decreases the value of the cost functional.

The analysis results bring us to the folllowing conclusions. Additional parameters are poorly identifiable. Parameters associated with an activator Cad are correlated with the smaller (4 TF) set of parameters; the parameters related to added repressors, especially to Kni, are of low necessity within the model as almost don't affect the model fit, hence the model is over-parameterized.

The predictive power problem is vividly discussed in recent publications devoted to new series of the *eve* models (e.g.[7]) and we believe that our approach will shed some light on

the questions raised in these articles.

We conclude that the method presented in the paper is really helpful in the choice of a reasonable level of detail for sound gene modeling. It allows to prevent model over-parameterization, and not only reveal the overfitting but to detect the sources of uncertainties in parameter estimation.

The research is supported by RFBR grant 15-04-07800-a.

1. M.Ashyraliyev, J.Jaeger, J.G. Blom (2008) On Parameter Estimation and Determinability for Drosophila Gap Gene Circuits, *BMC Systems Biology*, **2**:83.

2. J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz (2004) Dynamic control of positional information in the early Drosophila embryo, *Nature*, 430:368-371.

3. Kozlov K, Surkova S, Myasnikova E, Reinitz J, Samsonova M (2012) Modeling of Gap Gene Expression in *Drosophila* Kruppel Mutants. *PLoS Comput Biol* 8(8): e1002635

4. E.Myasnikova, K.Kozlov (2014), Statistical method for estimation of the predictive power of a gene circuit model, *Journal of Bioinformatics and Computational Biology*, 12(2) DOI: 10.1142/S0219720014410029

5. H. Janssens, S. Hou, J. Jaeger, A. Kim, E. Myasnikova, D. Sharp, and J. Reinitz **(**2006**)**. Quantitative and predictive model of transcriptional control of the *Drosophila* melanogaster even-skipped gene. *Nature Genetics*, 38:1159–1165

6. K.Kozlov, V.Gursky, I.Kulakovskiy and M.Samsonova (2014). Sequence-based model of gap gene regulatory network. *BMC genomics* 12/2014 15 Suppl 12:S6.

7. Ilsley, G. R., Fisher, J., Apweiler, R., DePace, A. H., & Luscombe, N. M. (2013). Cellular resolution models for even skipped regulation in the entire *Drosophila* embryo. *eLife*, 2, e00522. doi:10.7554/eLife.00522