

Experimental design and parameter estimation applied to the DREAM6 challenge: team crux



B. Steiert, C. Kreutz, A. Raue, J. Timmer

Center for Biosystems Analysis, Physics department, Freiburg university

bernhard.steiert@frias.uni-freiburg.de, ckreutz@fdm.uni-freiburg.de, araue@fdm.uni-freiburg.de

Abstract

The goal of the DREAM6 parameter estimation challenge was to perform experimental design considerations to estimate parameters of gene regulatory networks and to be able to extrapolate the systems' behavior, i.e. time courses of dynamic variables are predicted under perturbed conditions.

Here, the methodology to approach this issue is summarized. The maximum likelihood parameter estimates have been calculated by local nonlinear optimization with sensitivity equation based gradients in combination with latin hypercube sampling of the initial parameter guess. In addition, the profile likelihood has been utilized for identifiability analyses and to determine informative experiments. Because we found a bias for the estimation of Hill-coefficients, we had to carefully choose the upper boundaries of the parameter domain for the Hill-coefficients.

Our experimental design strategy was guided by the following major aspects:

1. Obtaining a sufficient amount of "cheap" data, i.e. WT measurements, to have some knowledge about the underlying system.
2. In case of several similarly likely local minima, experiments have been designed to discriminate such ambiguities.
3. Improving the knowledge of practically non-identifiable parameters.
4. Performing the experiments which reduce the bulk of the variance in the demanded extrapolation settings.

Methodology

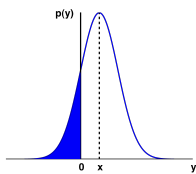


Fig. 1: Error model for measurements $y < 0$ set to 0. The probability of noise realizations $y < 0$ is considered by the cumulative density.

Maximize the logarithm of the likelihood

$$L_{cut}(y|\theta) = \prod_{i|y_i > 0} \frac{1}{\sqrt{2\pi}\sigma_{total}} e^{-\frac{(y_i - \pi_i)^2}{2\sigma_{total}^2}} \cdot \prod_{i|y_i \leq 0} \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}\sigma_{total}} e^{-\frac{(y' - \pi_i)^2}{2\sigma_{total}^2}} dy'$$

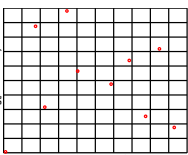


Fig. 2: Since the fitting algorithm we use is a local nonlinear optimization approach with sensitivity equation based gradients, it has to be ensured that the resulting local optimum is also globally optimal. For this purpose, a latin hypercube sampling of the initial parameter guesses was used. This ensures a wide variety of starting parameter sets. After several purchases of data, only one optimum was left that could describe the data.

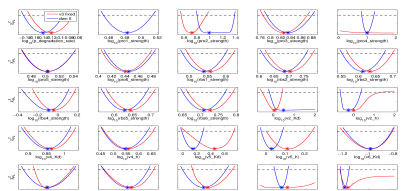


Fig. 3: In this example for step 5 of model M1, the parameter profile-likelihood (red lines) indicates large confidence intervals for $pro4_strength$, $v8_h$, $v8_K_d$, and $v2_h$. After purchasing p5 siRNA data these parameters became identifiable (blue lines).

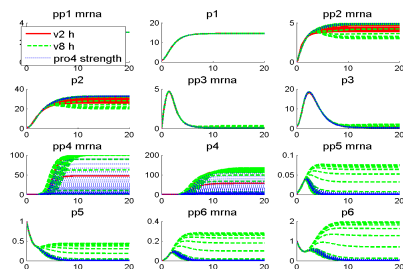


Fig. 4: The model predictions for all possible perturbation experiments have been evaluated for parameters along the profiles shown in Fig. 3. The different colors indicate the impact of the badly identifiable parameters $pro4_strength$, $v8_h$, and $v2_h$. Here, we decided to buy p2 and p4 protein data.

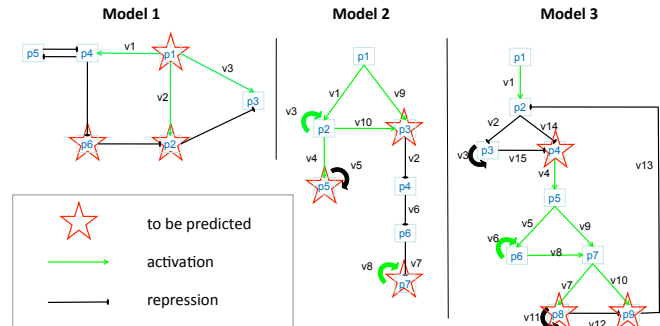
Evaluating the performance

Since there was an error in the data production step, there is no objective way of scoring the challenge. However, a rough estimate of the performance can be obtained as follows:

1. Choose exactly the same data sets without noise as bought during the challenge
2. Choose the same error model as applied during the challenge
3. Fit the model to estimate "effectively true parameters" for the data purchased
4. Calculate the parameter score, i.e. the distance of the submitted parameters to the "effectively true parameters"
5. Simulate the predictions with the "effectively true parameters"
6. Calculate the respective "effectively true" prediction score

	Model 1	Model 2	Model 3
Parameters	0.002	0.020	0.014
This yields the following scores for our team: Predictions	0.344	4.475	0.285

Models and extrapolation setting



Data purchase decisions

Step	Action	Arguments	Remaining credits
1	WT protein data of p1-p6	(WT-P), (WT), (E-data)	8800
2	WT high-density MA	(WT), (hd-MA), (E-data)	7800
3	v1_h, v1_K ₀	(ID), (E-para)	6200
4	v3_h, v3_K ₀	(ID), (E-para)	4600
5	siRNA pp5, measurement of p2&p4	(OptPerPL), (siRNA), (Budget)	3850
6	v2_h, v2_K ₀	(ID)	2250
7	rbcs of p4	(OptPerPL), (BI), (Extra)	1400
8	siRNA pp5 high-density MA	(OptPerPL), (Budget)	50

Tab. 1: Summary of the decision to spend the budget for model M1. The arguments are provided in the order of their priority. If an argument dominated, it is display in bold-face.

Step	Action	Arguments	Remaining credits
1	WT protein data of p1-p3, p5-p7	(WT-P), (WT), (E-data)	8800
2	v3_h, v3_K ₀	(ID), (Extra)	7200
3	siRNA pp6, measurement of p4&p7	(1st), (lv7), (Pred), (siRNA)	6450
4	v4_h, v4_K ₀	(ID), (MC)	4850
5	v8_h, v8_K ₀	(ID)	3250
6	siRNA pp1, measurement of p3&p4	(OptPerPL), (siRNA)	2500
7	siRNA pp2, measurement of p4&p6	(OptPerF), (siRNA)	1750
8	v10_h, v10_K ₀	(lv10)	150

Step	Action	Arguments	Remaining credits
1	WT protein data of p2-p9	(WT-P), (WT), (E-data)	8400
2	pp9 KO high-density MA	(BI), (hd-MA), (MALarge)	6600
3	v12_h, v12_K ₀	(Extra), (LocMin), (ID)	5000
4	v8_h, v8_K ₀	(LocMin), (ID), (Module)	3400
5	v9_h, v9_K ₀	(ID), (LocMin), (Extra)	1800
6	v6_h, v6_K ₀	(ID), (lv5)	200

Tab. 2 and 3: Summary of the decision to spend the budget for models M1 and M2.

Abbreviation	Explanation	Abbreviation	Explanation
(WT)	Wild-type data: cheap & informative	(OptPerF)	Maximal improvement of Fisher-Information in the asymptotic setting
(WT-P)	Initial step: WT-data for all proteins	(MC)	Monte-Carlo evaluation confirmed the guess
(hd-MA)	High density Microarray data contains information about fast processes	(ID)	Practical non-identifiability ((semi-)indefinite confidence-interval) resolved
(P>mRNA)	Ratio of data-points / credits always better for protein data	(BI)	Perturbation switches between qualitatively different behaviors
(MALarge)	Ratio better for larger models + data for all species → more robust	(Extra)	Improves accuracy of extrapolation
(E-para)	Extreme strategy: Buy most parameters	(1st)	A protein was not measured at all
(E-data)	Extreme strategy: Buy most data	(lxyz)	Informative for parameters x, y, z
(Budget)	Try to use credits almost completely	(Pred)	Informative dynamic behavior
(siRNA)	Cheaper than knock-outs, but often as informative	(Module)	Improve identifiability of a sub-module of bad estimates
(OptPerPL)	Maximal informative according to parameter profiles (see Fig. 4)	(LocMin)	Discriminate between different local minima which describe the data

Tab. 4: Explanation of the abbreviations used in Tab. 1-3.

Acknowledgments

The authors acknowledge financial support by the FP7 EU project CancerSys [HEALTH-F4-2008-223188].