Appendix

Efficient algorithms for ordinary differential equation model identification of biological systems

Peter Gennemark and Dag Wedelin

Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Göteborg, Sweden.
+46-(0)31-772 10 00
{peterg,dagw}@chalmers.se

1 Detailed information about the test systems and experiments

In this section the details of the two test systems and the corresponding experiments are presented.

1.1 The metabolic test system

The metabolic test system is taken from Arkin et al. [1], and represents a biochemical NAND gate. Mechanisms of this type are common in biochemical systems, such as glycolysis. The system of Arkin has two input variables I_1 and I_2 and five measured variables $S_3 - S_7$. It can be described by the following equations:

$$S_3'(t) = -v_1 - v_2 + v_3 + v_4 \tag{1}$$

$$S_4'(t) = v_1 - v_3 \tag{2}$$

$$S_5'(t) = v_2 - v_4 \tag{3}$$

$$S_6'(t) = -S_7'(t) = v_5 - v_6 \tag{4}$$

The kinetic equations all follow Michaelis-Menten kinetics and inhibition is non-competitive. They are specified according to

$$v_1 = \frac{S_3(t)V_{max1}}{(S_3(t) + K_{D1})\left(1 + \frac{I_1(t)}{K_{I1}}\right)} \tag{5}$$

$$v_2 = \frac{S_3(t)V_{max2}}{(S_3(t) + K_{D2})\left(1 + \frac{I_2(t)}{K_{I2}}\right)} \tag{6}$$

$$v_3 = \frac{S_4(t)V_{max3}}{S_4(t) + K_{D3}} \tag{7}$$

$$v_4 = \frac{S_5(t)V_{max4}}{S_5(t) + K_{D4}} \tag{8}$$

$$v_5 = \frac{S_7(t)V_{max5}}{(S_7(t) + K_{D5})\left(1 + \frac{S_3(t)}{K_{I3}}\right)} \tag{9}$$

$$v_6 = \frac{S_6(t)V_{max6}}{S_6(t) + K_{D6}}. \tag{10}$$

The following rate constants, taken from Arkin et al., were used: $V_{max1-2} = 5$, $V_{max3-4} = V_{max6} = 1$, $V_{max5} = 10$, $K_{D1-6} = 5$ and $K_{I1-3} = 1$.

To evaluate our algorithms, we used 12 simulated experiments with different combinations of manually created input steps for I_1 and I_2, see Table 1. Each experiment was simulated with high accuracy from $t = 0$ to $t = 150$. The standard numerical integration method dlsode has been used in all implementations (see *http://www.llnl.gov/CASC/odepack*).

To create the equivalent of real experimental data, the data of these simulations were then reduced to 7 uniformly sampled data-points per variable, giving a total of 84 data-points per variable over 12 experiments. For experiment 1 the time courses are depicted in Fig. 1.

As a second data-set we included 13 data-points per variable and experiment (11 data-points were uniformly sampled and 2 data-points were collected before the applied input signals at time zero). In this data set, we also simulated measurement noise by adding Gaussian noise with a 3.5% standard deviation relative to the particular experimental value.

A third data-set of 25 data-points per variable and experiment (21 uniformly sampled and 4 before time zero) and with added noise with constant 10% was also created. Similarly, a fourth data-set of 25 data-points with 20% noise was created.

Compared to the data used by Arkin the total number of data-points is similar, although we distribute our data-points over more experiments. Over all experiments, Arkin uses a total of 360 and 55 data-points per variable in the two different studies [1, 2], while our test cases include between 84 and 300 data-points per variable. The use of several experiments allow us to investigate a range of interesting values of the input functions and at the same time design individual input functions to give relatively smooth time series for the variables.

1.2 The genetic network test system

The other test system is taken from Kikuchi et al. [3] - a small genetic network originally considered in Hlavacek et al. [4]. This model is defined as a so called S-system model. The S-system formalism [5, 6] is based on approximating kinetic laws with multivariate power-law functions. A model consists of n non-linear ODEs and the generic form of equation i reads

$$X_i'(t) = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}}(t) - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}(t) \tag{11}$$

where X is a vector (length n) of dependent variables, α and β are vectors (length n) of non-negative rate constants and g and h are matrices ($n \times n$) of kinetic orders, that can be negative as well as positive. The parameters of the genetic network are given in Table 2.

From this model we simulated data for ten experiments with different initial conditions, see Table 3. For each experiment the ODEs were simulated from $t = 0$ to $t = 0.5$.

As a first data-set, eleven data-points were uniformly sampled for each variable and experiment. This is exactly the same data as used by Kikuchi et al. [3]. See Fig. 2 for data from experiment 1. As a second data-set we sampled three data-points non-uniformly ($t = 0$, 0.025 and 0.50) per variable and experiment. The non-uniform sampling gives slightly higher precision in the transient part of the curves and thereby allows us to use fewer data points.

1.3 Test case for parameter estimation

A detailed description of the test case from [7] used for our parameter estimation algorithm is available at *http://www.iim.csic.es/∼julio/GR03_statement.txt*.

2 Additional comments on the algorithms

2.1 Example of estimation in a single equation with simulation

We decompose the parameter estimation problem to single equations by assuming the time series for the other variables known. For example, considering the metabolic test system, we obtain one parameter estimation problem for S_3 (Eq. 1) that contains ten parameters, one for S_4 (Eq. 2) that contains five parameters etc. For biomolecular networks that are sparse, which is typically the case, we note that the number of parameters in each equation is independent of the total number of variables in the system.

In step 2b of the parameter estimation we use simulation of the single ODE in order to find the parameters of that ODE that maximise the likelihood function. For Eq. 2, step 2b works as follows. Values for the parameters V_{max1}, K_{D1}, K_{I1}, V_{max3} and K_{D3} are proposed by the parameter estimation algorithm and the time series for I_1 and S_3 are given, either as interpolated data or as simulated data from the current best model. We simulate S_4 with high accuracy, beginning with an initial point taken from the data. The result is compared to the given data points for S_4 and the likelihood function is calculated.

2.2 The model space

The model space of the structure search algorithm is defined by a set of reaction types. The collection of reaction types given in Table 2 of the main text is large enough to let us build interesting and non-trivial test models resembling real systems. For instance, the metabolic test system is composed of two of the reaction types. Naturally, this test system is more difficult to identify when the model space includes all four reaction types.

In general, it is not known which reaction types that occur in a system and, consequently, one must carefully specify a plausible collection of reaction types for a particular area of application. Once implemented, this collection or parts of it can be reused in future applications. The set of reaction types also determines which kinds of parameter estimation algorithms are required, e.g. linear or non-linear.

We note that it is natural to have bounds for the parameters and, in reality, reasonable guesses can usually be made from the literature. In particular, given sufficient prior knowledge of the system one may be able to set the lower bounds of k and V_{max} greater than zero and thereby reducing the risk of over-fitting. For the metabolic test system we used wide bounds as given in Table 2 of the main text.

Most biomolecular ODE models in the literature can be described in terms of reactions taken from a set of reaction types. One exception is the S-system formalism, where the kinetics are described in the form of Eq. 11. However, for S-systems, a non-zero element in g or h can be considered as the correspondence to a reaction. This is easiest understood by comparing the interactions in Figure 2 of the main text to the non-zero elements of Table 2. Therefore, to identify a S-system, we need to find the best combination of non-zero kinetic orders, instead of finding the best combination of reactions from a collection of reaction types as in the more common case. In practice, all elements in g and h corresponding to a 'reaction' are bound in an user-specified interval, e.g. $[-3, 3]$, while all other kinetic orders are bound to zero. To add a 'reaction', we simply switch the bounds from zero to the user-specified interval, and the other way around to remove a 'reaction'.

For the genetic network we considered the same bounds as Kikuchi et al.[3]: $\forall i$: $\alpha_i, \beta_i \in [0, 15]$ and $\forall i, j : g_{ij}, h_{ij} \in [-3, 3]$.

2.3 Model complexity and error function

An important issue in model identification is how to keep complexity low and avoid over-fitting. In general, a model with a complex structure is more likely to have high likelihood, because the parameter space is large and the model can be fine-tuned to fit experimental data.

There are several ways to avoid high complexity [8, 9]. One common way is to

consider model complexity implicitly in some termination criterion. Another more elegant way to explicitly punish high complexity is to include a penalty term in the error function. This term is typically a function of the number of parameters and the number of data-points. Common variants include AIC [10], BIC [11] and MDL [12]. Here we wish to point out that any termination criterion in model identification algorithms either includes method-specific parameters or relies on the assumptions made for the suggested penalty function. This is an open research area [13].

Naturally, our first attempt was to use one of the existing methods for penalising structural complexity. However, all above mentioned methods gave an unacceptably large number of false positive reactions for our test systems. Only when constraining the model space (e.g. by assuming mass balance) we obtained reasonably good output. There are several possible reasons for this effect, one being that we cannot assume that the term $-L$ is really minimised, because of imperfections in the parameter estimation algorithm.

In our experiments we have kept the spirit of all these variants by using the error function (2) in the main text, but simply chosen to set λ manually for the problem at hand. We see this as a practical approach to circumvent the difficulty of determining it automatically. In practice, one can first try values of λ resulting in an error function similar to AIC, BIC or MDL.

For exact data any positive value for λ should in theory give the correct result. In practice, it can be arbitrarily set in a very wide range, but extremely low and high values should be avoided.

2.4 Computational complexity of the algorithm

In our implementation and for the kind of problems we have been considering, almost all running time of the model selection algorithm is spent in step 2, which in turn is dominated by calls to step 2 in the parameter estimation algorithm. Ignoring the other steps, we can sketch the time complexity, T, as

$$T = N_{loops}\ n\ N_{tr}\ T_{eval} \tag{12}$$

where N_{loops} is the maximum number of loops for one variable in the algorithm, n is the number of variables, N_{tr} is the number of test reactions and T_{eval} is the time-complexity of the local parameter estimation and error calculation. We note that T_{eval} is dependent on the complexity of the model, such as non-linearities and number of parameters, as well as the number of experimental data-points.

It is difficult to estimate N_{loops}, since it depends on the iterative behaviour of the algorithm. In particular, N_{loops} is affected by the error function. Ideally, N_{loops}

equals the maximum number of reactions that are not included in the initial structure but belong to the correct structure of the variable plus one for the terminating loop. On the other hand, if all possible reactions are added $N_{loops} = N_{tr}$. In practice, N_{loops} tends to be close to its ideal value. Besides, we note that biomolecular networks are typically sparse and, therefore, the number of correct reactions to add is bound by some constant.

In practice, it seems that also N_{tr} and T_{eval} can be kept relatively constant for different problem sizes, implying that the algorithm can be expected to behave polynomially with respect to the number of variables and amount of experimental data. For very large models it cannot be excluded that the global parameter estimation of steps 3 and 4 contribute significantly to the running time, potentially by a non-polynomial time complexity with respect to the number of variables and amount of experimental data. However, the global parameter estimation can actually be skipped without losing too much precision as indicated in Sect. 2.

Finally, we note that the algorithm always requires a model with smaller error than the error of the best model in the previous iteration, and hence, there is no possibility that the algorithm shows a cyclic behaviour.

References

[1] Arkin, A.P. and Ross, J.: 'Statistical construction of chemical reaction mechanisms from measured time-series', *J. Phys. Chem.*, 1995, **99**, pp. 970-979

[2] Arkin, A.P., Shen, P. and Ross, J.: 'A test case of correlation metric construction of a reaction pathway from measurements', *Science*, 1997, **277**, pp. 1275-79

[3] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K. and Tomita, M.: 'Dynamic modeling of genetic networks using genetic algorithm and S-system', *Bioinformatics*, 2003, **19**(5), 643-50

[4] Hlavacek, W.S. and Savageau, M.A.: 'Rules for coupled expression of regulator and effector genes in inducible circuits', *J Mol Biol.*, 1996, **255(**1), pp. 121-39

[5] Savageau, M.A.: 'Biochemical systems analysis: a study of function and design in molecular biology' (Addison-Wesley, Reading, Mass, 1976)

[6] Voit, E.O.: 'Computational analysis of biochemical systems. A practical guide for biochemists and molecular biologists' (Cambridge University Press, Cambridge, 2000)

[7] Moles, C.G., Mendes, P. and Banga, J.R.: 'Parameter estimation in biochemical pathways: a comparison of global optimization methods', *Genome Res.*, 2003, **13**(11), pp. 2467-74

[8] Brooks, R. and Tobias, A.: 'Choosing the best model: level of detail, complexity and model performance', *Math. Comp. Mod.*, 1996, **24**(4), pp. 1-14

[9] Zucchini, W.: 'An introduction to model selection', *J. Math. Psych.*, 2000, **44**, pp. 41-61

[10] Akaike, H.: 'Information theory and an extension of the maximum likelihood principle', In: Petrov, B.N. and Csaki, F. (Eds.) '2nd Int. Symp. Inform. Theory', Akademiai Kiado (Budapest, 1973), pp. 267-281

[11] Schwarz, G.: 'Estimating the dimension of a model', *Annals of Stat.*, 1978, **6**, pp. 461- 464

[12] Rissanen, J.: 'Modeling by shortest data description', 1978, *Automatica*, **14**, pp. 465-471

[13] Crampin, E.J., Schnell, S. and McSharry, P.E.: 'Mathematical and computational techniques to deduce complex biochemical reaction mechanisms', *Prog. Biophy. Mol. Biol.*, 2004, **86**, pp. 77-112

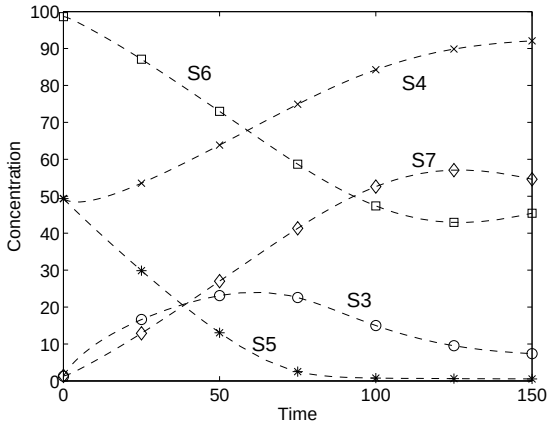

Figure 1: *Data for experiment 1 simulated from the metabolic test system. The markers indicate the data-points sampled.*

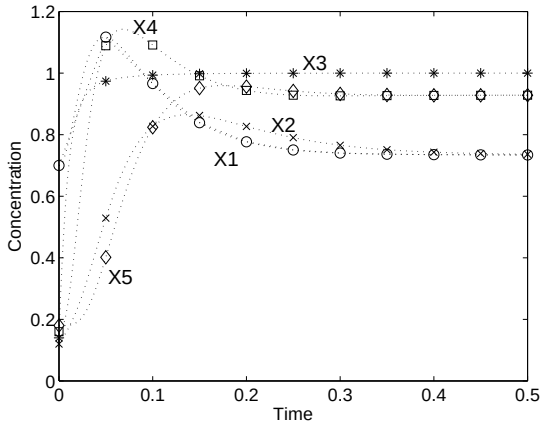

Figure 2: *Data for experiment 1 simulated from the genetic network test system. The markers indicate the data-points sampled.*

Exp.	I_1	I_2
1	$0.1 \rightarrow 2$	$0.1 \rightarrow 30$
2	$0.1 \rightarrow 30$	$0.1 \rightarrow 2.5$
3	$0.1 \rightarrow 20$	$0.1 \rightarrow 30$
4	$0.1 \rightarrow 0.1$	$0.1 \rightarrow 30$
5	$30 \rightarrow 0.1$	$0.1 \rightarrow 30$
6	$4 \rightarrow 1$	$3 \rightarrow 2$
7	$30 \rightarrow 1$	$30 \rightarrow 1$
8	$4 \rightarrow 2$	$5 \rightarrow 2$
9	$30 \rightarrow 10$	$30 \rightarrow 0.5$
10	$30 \rightarrow 0.5$	$30 \rightarrow 10$
11	$0.1 \rightarrow 1.5$	$0.1 \rightarrow 30$
12	$0.1 \rightarrow 0.5$	$10 \rightarrow 2.5$

Table 1: *Variables I_1 and I_2 for the different experiments of the metabolic test system. We use the notation $a_1 \rightarrow a_2$ where a_1 is the start value ($t \leq 0$) and a_2 is the value at $t > 0$.*

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	5.0			1.0		-1.0	10.0	2.0				
2	10.0	2.0					10.0		2.0			
3	10.0		-1.0				10.0		-1.0	2.0		
4	8.0			2.0		-1.0	10.0				2.0	
5	10.0				2.0		10.0					2.0

Table 2: *Parameters of the genetic network test system [3]. An empty element corresponds to 0.0. Each row corresponds to one ODE according to Eq. 11, e.g. the first row gives* $X_1'(t) = 5.0X_3(t)/X_5(t) - 10.0(X_1(t))^2$.

Exp.	X_1	X_2	X_3	X_4	X_5
1	0.70	0.12	0.14	0.16	0.18
2	0.10	0.70	0.14	0.16	0.18
3	0.10	0.12	0.70	0.16	0.18
4	0.10	0.12	0.14	0.70	0.18
5	0.10	0.12	0.14	0.16	0.70
6	0.70	0.70	0.14	0.16	0.70
7	0.10	0.70	0.70	0.16	0.18
8	0.10	0.12	0.70	0.70	0.18
9	0.10	0.12	0.14	0.70	0.70
10	0.70	0.12	0.14	0.16	0.70

Table 3: *Initial concentrations of the five variables of the genetic network model in each of the ten experiments [3].*