# 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

### Abstract

We present algorithms for parameter estimation and model selection that identify both the structure and the parameters of an ordinary differential equation model from experimental data. We mainly focus on the case of unknown structure and some time course information available for every variable to be analysed, and we exploit this to make the algorithms as efficient as possible.

The algorithms are designed to handle problems of realistic size, where reactions can be non-linear in the parameters and where data can be sparse and noisy. To achieve computational efficiency, parameters are mostly estimated for one equation at a time, giving a fast and accurate parameter estimation algorithm compared to other algorithms in the literature. The model selection is done with an efficient heuristic search algorithm, where the structure is built incrementally.

We use two test systems that previously have been used to evaluate identification algorithms, a metabolic pathway and a genetic network. We successfully identify both test systems, using a reasonable amount of simulated data. Besides, measurement noise of realistic levels can be handled. In comparison to other methods that were used for these test systems, the main strengths of our algorithms are that a fully specified model, and not only a structure, is identified, and that they are considerably faster compared to other identification algorithms.

**Keywords**: model identification, biological modelling, parameter estimation, model selection, ordinary differential equations, S-system.

# 1 Introduction

A commonly studied identification problem is that the structure, i.e. the form of the equations is assumed known, but with unknown parameters. In this paper, we consider the problem where not only the parameters but also the structure is unknown. Finding the structure is referred to as **model selection**, and finding the parameters as **parameter estimation**. We have developed algorithms for these tasks that work together to determine a fully specified ordinary differential equation (ODE) model, and which have been carefully matched to each other for speed and accuracy.

The algorithms require that the problem has been modelled and specified mathematically as an optimization problem. We note that when the structure is unknown, this is not a trivial task. For parameter estimation we need suitable experimental data, possibly from several experiments including e.g. different input signals and gene deletions, and an error function. For model selection we additionally require a discrete model space, in our case a set of possible reactions, and an error function that takes model complexity into account. The desired solution is a parsimonious ODE model that has a good fit to the data, according to the error function. Depending on the circumstances, additional information such as parameter bounds, partly known or guessed initial structure etc. may be part of the definition of a given problem. The specification of the model selection problem is further discussed in Sect. 3.

Another observation is that when the structure is unknown, it is natural to require that some time course data is available for all variables that we wish to analyse. Otherwise, since no structural information is available, there would be no source of information for the variable, and no conclusions could be drawn about it. We therefore focus on this case in our presentation. In Sect. 5.1 we discuss the mixed case when the structure is only partially unknown.

Main elements of our algorithmic approach are as follows:

- Given that some time course information is known for all variables to be analysed, we exploit this to make the algorithms as efficient as possible. This makes it possible to solve non-trivial problems on a standard computer, rather than a super-computer, and provides a natural base case for further development.

- Most of the time, parameters are estimated separately for each equation. In this way we obtain several parameter estimation problems with low dimension instead of one problem with high dimension. This decomposition gives some stability and efficiency benefits by itself, and also enables the model selection algorithm to rapidly evaluate many different local structures with local computations only.

- The model structure is built incrementally with a heuristic search algorithm where single new reactions are added and sometimes deleted. Parameter estimation is done simultaneously during the incremental progress of the structure search.

The problem of system identification with unknown structure is very challenging. A solution method will therefore also have limitations, and results should be properly assessed. This is further discussed in Sect. 5. For us, an important goal has been to see if we in practice can obtain meaningful results with such small amounts of data that can usually be available from experiments. This has been done by evaluating our approach for two test systems that previously have been used to evaluate such algorithms.

The first test system is a metabolic pathway considered by Arkin et al. [1], see Fig. 1. Arkin infers the non-trivial structure of this pathway from time series measurements of compound concentrations. The method is based on factor analysis of the pairwise correlations between the variables, and the result of the analysis is then interpreted manually to a possible structure. In contrast to Arkin's method, our algorithms attempt to find a fully specified dynamic model, i.e. with both structure and parameters.

The second test system that we consider is a genetic network studied by Kikuchi et al. [2], see Fig. 2. This model is defined in the S-system formalism [4, 5] and it was almost perfectly identified using a genetic algorithm and super-computing [2].

There are also other methods than Arkin's method for inferring the structure of a system without also inferring its dynamic behaviour. The input to such methods is either steady-state data [6, 7, 8] or time series data [9, 10]. A unified framework for both steady-state and temporal data is given in [11]. We generally note that the output from methods that infer the structure only, may be used as initial input to our algorithm.

Other methods than Kikuchi's method for identification of fully specified dynamic biomolecular models have been proposed, mainly in the area of evolutionary computation [12, 13, 14]. The strength of such methods is that any desired objective function can be chosen and then directly optimised given available data. The drawback is the high or possibly extreme requirement for computational power.

The general problem of identifying both structure and parameters of S-systems is discussed by Voit [5, 15]. For identification of parameters in a known structure, we refer to [16, 17, 18].

To our knowledge, our approach as outlined by the combination of items above together with their detailed implementation, is new for this kind of problem. On a general level, there is some similarity to Wedelin [19], who identifies graphical models (also known as Bayesian networks). Some details of the algorithms are arbitrary or are defined by external subroutines that may be replaced. We therefore throughout emphasise the overall approach and computational structure of the algorithms, rather than specific details.

## 2 Parameter estimation in a given structure

An important part of identifying a dynamic system is the way in which parameters are estimated. This is a non-trivial problem in itself [20, 21]. In our approach, the model selection algorithm requires repeated evaluation of many different tentative structures in which parameters need to be estimated, so the computational efficiency of the parameter estimation becomes especially crucial. We use a combination of new and standard ideas to get the best performance.

For parameter estimation we need one or several sets of time-series experiments. Each experiment does not need to be complete but every variable should be measured in at least one experiment. In every call to the parameter estimation algorithm the structure is given, and each unknown parameter is free within a lower and an upper bound. The solution is a parameter vector $\mathbf{k}$ that maximises the likelihood of the observed data $\hat{X}$.

In practice, we work with the log-likelihood. By assuming independent and normally distributed measurement errors and disregarding constant terms we can express the log-likelihood for one time series as

$$L(\hat{X}_j|\,\mathbf{k}) = -\frac{1}{2}\sum_i \left(\frac{X_j(t_i) - \hat{X}_j(t_i)}{\sigma_j(t_i)}\right)^2 \tag{1}$$

where $i$ indexes the measurement points, and where $X_j$, $\hat{X}_j$ and $\sigma_j$ denotes simulated data, experimental data and standard deviation for variable $j$, respectively. The total log-likelihood $L(\hat{X}|\,\mathbf{k})$ is defined by summing over all variables and all experiments.

Our algorithm decomposes the parameter estimation per ODE as indicated in the following outline. At all times, current parameter values and the most recent simulation of each variable are maintained.

```
REPEAT
        1. Select input for the other variables, to be used in step 2: experimental
           data or simulated data.
        2. FOR each equation DO
```

(a) Make a rough estimate of the parameters in the equation with the derivative method [5, 22].

(b) Improve the estimate of the parameters in the equation with **local** simulation based error calculation (using standard parameter estimation subroutines).

3. Improve the estimate of all parameters with **global** simulation based error calculation (using standard parameter estimation subroutines).

UNTIL convergence

The decomposition in step 2 has several advantages. It gives an opportunity to better exploit the structure of the problem, giving sub-problems with significantly lower dimensions than the global problem. It also improves stability by creating a more direct connection between the known time courses and the parameter estimation, forcing the parameter search in the direction of the global optimum. Even disregarding the possibilities it creates for model selection, decomposition can be powerful for parameter estimation in isolation.

In order to estimate the parameters in any single equation, we must supply given time functions for any other variables occurring in the equation. In step 1 above, this is done in one of two ways. Initially we use the given time series for the other variables interpolated at any intermediate points needed using cubic smoothing spline interpolation [23] (pppack, see *http://www.netlib.org*). This gives good stability, but may give poor accuracy when we have sparsely sampled data-points.

If at some stage the model has become correct enough it may therefore happen that the simulated time series created by the algorithm are more accurate than the interpolated data. These can then replace the input of interpolated time series to our single equation parameter estimation problems. As a consequence, the subproblems become dependent of each other, and have to be solved repeatedly over and over again until convergence. One way to decide between interpolated data or simulated data, is simply to calculate the likelihood for the two alternatives and apply the method that gives the best result.

The actual estimation is done in steps 2 and 3. In step 2a, a first estimate is done with the very fast but inaccurate derivative method (dn2gb, see *http://www.netlib.org*). This estimate is then used as a starting point for step 2b where standard parameter estimation by non-linear least squares is employed (dn2gb). The required error function used by the parameter estimation routines is calculated by simulating the single ODE to evaluate the error function and residuals. We have used single shooting [21]. However, multiple shooting [24, 25] can be used as needed. How well parameters can be estimated in step 2b depends on the accuracy of the known time series for the variables, but much less so than for the derivative method. In order to avoid local maxima and thereby improve stability, several random starting points can be evaluated in both step 2a and 2b. In later iterations, it makes sense to continue from the best estimates so far rather than random starting points. We allow the simultaneous use of several experiments. All experiments are then treated simultaneously in 2a, while simulation is repeated for each individual experiment in 2b. For further information of simulation of a single variable, see the Appendix.

In step 3, we perform a global estimation of all parameters in order to fine tune the parameters to the nearest maximum (dn2gb). The error calculation is here based on a global simulation. Global estimation is generally significantly slower than the decomposition approach, but appears to have better convergence properties close to a maximum, and so reduces the number of iterations needed in the outer loop. However, it can in principle be skipped if estimation based on global simulation is not feasible.

The stopping criterion of the outer loop can be set in various ways. In our implementation we have simply used a small constant number of iterations, typically 2.

We finally note that by selecting simulated data in step 1, we actually free ourselves from the initial assumption that all times series have to be known. On the other hand, using interpolated data greatly improves the stability of convergence in the beginning which may explain why we in our experiments have not had any significant problems with divergence and local maxima.

### 2.1 Evaluation of the parameter estimation algorithm

The parameter estimation algorithm is a result in its own right, so we evaluate it separately using a known hard test case involving simulated data from a model with 36 parameters [21, 26, 27]. In [21], seven different parameter optimisation methods are evaluated using this test case and only two of them give reasonably good estimates. The result, see Table 1, indicates that our algorithm is more accurate and faster compared to the methods considered in [21, 26, 27]. We note that [26] uses the same standard routine for non-linear least squares (dn2gb) as we do, indicating that the specific features of our approach are significant for the result.

## 3 Model selection

The problem of model identification can generally be seen as the problem of finding the best model in some **model space**, according to an error function. Since we divide the model into structure and parameters, it is natural to consider the following algorithmic approach: On top of the parameter estimation algorithm we add a model selection algorithm that picks different tentative model structures, and for each structure that we consider, the parameters are estimated as described in

Sect. 2, and the error is calculated. Thus, it is not the case that the final structure is determined first before any parameters are estimated, but the parameter estimation algorithm is used as a subroutine called many times by the model selection algorithm. We note that an exhaustive search of the model space is not feasible for problems of realistic size.

Before we look at algorithmic details we will clarify what we need to properly specify also the model selection problem. In addition to the time course data used for parameter estimation, we use the following:

- **A discrete model space defined by a set of reaction types**. The model space of the structure search algorithm is defined by a collection of reaction types. See Table 2 for the model space of our first example, and the Appendix for the second example.
- **An error function**. The maximum likelihood criterion of Sect. 2 needs to be extended to consider structural complexity. How to do this is a well investigated but non-trivial problem, and it is difficult to have a definite opinion on a best choice. If no problem specific information is available a common approach is to include a penalty term, typically a function of the number of parameters and the number of data-points [28, 29, 30]. If prior information is available, customised error functions could be useful in low-data situations in the biological domain. For our test systems we have minimised a quite simple error function of the form

  $$-L + \lambda K \tag{2}$$

  where $\lambda$ is a constant and $K$ is the number of parameters. See Sect. 4 and the Appendix for further information.
- **Parameter bounds for reaction types**. For each reaction type in the model space, lower and upper bounds are assigned to each parameter. See Table 2 for the bounds of our first example, and the Appendix for the second example.
- **An initial model**. The initial structure consists of all compounds and any number of reactions of the model corresponding to the established knowledge of the structure of the system. Some parameters may be known or partially known. For our main test problems no reactions have been considered known.

In practice, the first two items can be expected to require customisation of the implementation of the algorithm, and the last two and the time course data can typically be provided in a data file. In this paper, we choose to consider all of this as input. We acknowledge that the specification of all these items in a particular context involves non-trivial applied and theoretical issues. However, we think that we for our purposes have drawn a sensible line between the task of modelling and defining problems, and on the other hand the algorithmic task of solving them, where the latter has been our focus.

The principle of the model selection algorithm is a local search heuristic that reconstructs the model structure incrementally. A **current model** is maintained at all times. New reactions are tested, and if they are found to decrease the error function they are added to the current model. Reactions may also be removed from the current model at a later stage if this improves the error function, so the algorithm is not greedy, although optimality cannot be guaranteed in the general case.

The search for new reactions is decomposed per variable in order to match the parameter estimation algorithm which works in the same way. We can then quickly test many different possible equations for a variable, with only local operations on the model, and without making global simulations.

The structure search algorithm can be outlined as follows:

```
REPEAT
    FOR every variable
        1. Based on the current model, the unknown parameters of the ODE
           are locally estimated and the error is calculated.
        2. Every possible reaction in the model space is temporarily added
           to the ODE one by one. For each reaction that is tried, the parameters
           of the resulting ODE are locally estimated and the error is
           calculated.
        3. If a better model was found in 2, the best reaction is added to the
           ODE, and all parameters in the entire model are re-estimated.
        4. Weak reactions are removed. All possible sub-models of a particular
           ODE are evaluated and if the error decreases by removing any
           single reaction this is done. If the current model is modified, all
           parameters in the entire model are re-estimated.
UNTIL error function does not improve
```

The local parameter estimation in steps 1 and 2 above involves only the parameter estimation steps 2a and 2b for one variable as described in Sect. 2. In steps 3 and 4 above the full parameter estimation algorithm is used.

It is difficult to explicitly give the time complexity of the algorithm, but we note that thanks to the decomposition our implementation in practice behaves polynomially with respect to the number of variables and amount of experimental data. Almost all computational time is spent in the parameter estimation subroutines. See the Appendix for further information.

Finally, we note that the structure search algorithm can be optimised in different ways. For instance, we can monitor variables that have reached a stable structure, and restrict further search to the more uncertain parts of the structure. From the user's perspective, one can significantly reduce the running time by constraining the model space by prior information.

# 4 Test results

In this section we present test results for identification of both structure and parameters for the metabolic and genetic test systems, respectively. All tests were run on an ordinary PC (Intel Xeon, 2.8 GHz).

## 4.1 Identification of the metabolic system

For the metabolic test system we considered the model space and parameter bounds given in Table 2. We first considered exact data. In this case the results are almost entirely insensitive to the values of $\sigma_j$ and $\lambda$ in the error function (2). To avoid division by zero, we artificially set the $\sigma_j$ to the arbitrary value 1, and $\lambda$ was arbitrarily set to 1 as well. Given an empty model structure and a data set including 12 experiments and 7 data-points per variable in each experiment (see Appendix), we identified the correct model structure, and reasonable parameter values in about 60 minutes. The full test details and results are given in Table 3. We note that conservation of mass has not been assumed. If this constraint is included, the model space becomes smaller and there is a significant decrease in running time. For fewer than 7 data-points per variable and experiment we are not able to correctly identify the metabolic test system.

Compared to Arkin et al. [1], our data-set is smaller, and has been adapted to be better suited for parameter estimation. We identify not only the dependencies, but a fully specified ODE model.

For real systems similar to the metabolic test system, measurement errors of *in vitro* data are reported in the range 1-7% [31]. We added measurement noise from a normal distribution with standard deviation 3.5%, 10% and 20% to our artificial data. To successfully identify the model we used the same 12 experiments as before, but we increased the number of data-points per experiment and variable. In the error function the true standard deviations were considered known for all data-points, and $\lambda$ was manually set as described in the Appendix. See Table 3 for full test details and results. We note that the models include a few false positive reactions but these have significantly smaller rate constants, so they could in principle be filtered out with a low threshold (see also the Appendix for a discussion on the error function). The running time was about 100 minutes. Repeating with data simulated with different random seeds gave similar results.

The test results indicate that the algorithm can handle real biological experiments: the number of measurement points can be reduced to acceptable levels and the algorithm can handle data that have measurement noise of realistic levels added to them. We note that the we use a model space with four reaction types, although only two types are required for successful identification.

### 4.2 Identification of the genetic network

For the genetic network we considered the same model space, parameter bounds and data-set as Kikuchi et al. [2] (see Appendix). These data are exact and consists of 10 different experiments with 11 uniformly sampled data-points per variable. As for the metabolic system, the results for exact data are almost entirely insensitive to the values of $\sigma_j$ and $\lambda$, and we used the arbitrary values 0.01 and 1, respectively. The smoothing parameter was set to 0.001.

With these settings we identified the correct model (relative error in the parameters in the order of machine precision) in 6 minutes. Kikuchi et al. [2] obtained one false positive interaction ($h_{53} = 0.7$) and the true parameters within 18% relative error using 70 hours on a super-computer with a cluster of 1040 processors (Pentium 3, 933 MHz). This result shows that our approach is more accurate and significantly faster compared to the genetic algorithm approach by Kikuchi et al. Moreover, there are only a few method-specific parameters to adjust in our approach.

In addition, we could decrease the number of data-points per variable and experiment to 3 and still obtain a perfect model (smoothing parameter=0.001, $\lambda$ was set to 10 since this improves computational speed, but the final answer is still insensitive to the value of $\lambda$). We also ran the genetic network test system with 20% noise. Using 11 non-uniformly sampled data-points per variable and experiment (smoothing parameter=0.01, $\lambda = 2$), the correct structure and reasonable parameter values were identified. The running time was about 6 minutes for both these tests.

## 5 Summary and discussion

The presented algorithms reconstruct both the structure and the parameters of two test systems given simulated data. For the metabolic test system the algorithm identifies both structure and parameters using a similar amount of data with a higher level of noise than Arkin et al. For the genetic network test system the algorithms reconstruct the system to a higher degree of accuracy using less data than Kikuchi et al. Furthermore, the computational effort is several orders of magnitude lower. Worst-case model identification scenarios have been tested, since only an empty initial structure has been assumed. Moreover, the approach is not dependent on the specific reaction types used and can be applied also in other areas.

An important reason for the efficiency of both algorithms is the decomposition to one equation at a time, which makes it makes it easy to stabilise the parameter estimation with known experimental data and to search many models by local operations. Another reason is the heuristic search. Because of the time complexity characteristics of the algorithms, we have reason to expect that the approach works also for larger systems.

From a general perspective, we see our results as a proof of concept that identification can be done in practice, with realistic input data, and with reasonable computational effort.

## 5.1 An extensible approach

To better fit a particular application, the algorithms can be extended in several straightforward ways. One example is that a number of reaction types can be implemented to be easily available, as well as different common system constraints. The search strategy can also be extended. For instance, by using beam search (see e.g. [32]) we can if necessary make a more thorough search and at the same time avoid a combinatorial explosion. We note that it is possible to adopt any one of our two algorithms and replace the other with some other algorithm, e.g. with a more global search.

We consider the present result to be a natural step towards exploring methods with partially unknown structure and incomplete data sets. We note that there is no fundamental conflict between using our approach where the structure is unknown, and using ordinary parameter estimation approaches where the structure is known. Here note that our fast parameter estimation is only critically needed in the unknown parts where many possible structures need to be tested, i.e. where a properly specified identification problem has to include time course data anyway. The bottom line is that we can expect extensions of our approach to be useful as a basis for solving also these problems. As a simple example, after some modifications of our implementations we are able to identify the genetic network test system when one variable is completely unobserved but the interactions (but not the strengths of the interactions) of that variable are known. Using the same data set as in [2] the running time is a couple of minutes.

We finally remark on the basic problem that variables that are important in the real system may be missing in the mathematical problem we analyse. Depending on the role of such variables, and other factors such as the chosen model space, meaningful results may be obtained subject to proper interpretation. For example, a pathway A-B-C where data for B is missing will typically result in a proposed reaction between A and C. We have not considered the algorithmic discovery of hidden variables, whose existence is determined using the model space and the error function as the only sources of information. This is a difficult problem, the feasibility of which is highly sensitive to the nature and accuracy of available data.

### 5.2 Evaluation and use of identification algorithms

Given the diverse nature of different systems and problems, and the many components of the proposed algorithms, it is difficult to give general statements about performance and reliability. Still, to bring some clarity to this, we will outline the fundamental reasons why an identification algorithm may fail to perform as expected:

1. **The error function**, which defines the mathematical objective of the identification problem, can have a strong influence on the final result. Especially model complexity makes the choice of error function an intricate issue.

2. **Insufficient data**. There may be models with the same or lower error than the correct model. Even a perfect algorithm will then give an incorrect answer. Of course, the data required is a function of the system to be identified and the chosen model space.

3. **Algorithmic difficulties**. The algorithm may fail to find the best solution due to its heuristic nature. Generally, it is not possible to know if the best solution was found.

It may be difficult to assess if a specific result is unreliable and if so, which of the reasons above are really the causes of the problem. The situation is further complicated by the fact that also an incorrect model can have some explanatory value as well as predictive power.

As a simple example of a case when the algorithm fails we consider the genetic network test system. In Sect. 4 we reported that 3 (non-uniformly sampled) datapoints per experiment is enough for identification when using all 10 experiments. Now, reducing the number of experiments to 6 still results in perfect identification, while 5 experiments results in 1 false positive and 1 false negative interaction. The error of the proposed model is slightly higher than that of the correct model, indicating that the failure is due to point 3 above.

A pragmatic approach to algorithm reliability is to explore it on a case by case basis as we have done in this paper. Even if the real system of interest is unknown, one can define similar systems, simulate realistic experimental data and test if they can be reconstructed. It is also possible to provide additional data such as parameter sensitivity w.r.t. $L$, as an aid in interpreting the result. A user can also counter reliability problems by supplying better input data (experiments, initial model, parameter bounds), or by reducing the model space. Better data is not just a matter of volume or noise but also of the right kind of data. For example, one may include data measured on a disturbed system, where for instance certain genes are deleted to break up feedback loops or isolate sub-systems. Another possibility is to make experiments with different input data as in our examples.

We believe that properly used, algorithms for system identification can be a useful tool in biological research. At the very least the results can be used as hints for further exploration of the real system of interest. Even without biological data, a system including simulation and identification should be useful for experimental planning e.g. to find out what experiments and other knowledge are needed to identify a system.

**Acknowledgements**

Thanks to our collaborators Bodil Nordlander, Per Sunnerhagen and Stefan Hohmann at Cell and Molecular Biology, Göteborg University, Olle Nerman at Mathematical Statistics, Chalmers University of Technology and Edda Klipp at BCB, Max-Planck Institute for Molecular Genetics.

This work was supported by Chalmers Bioscience Program.

# 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] 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), pp. 643-50

[3] 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

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

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

[6] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L.: 'Integrated genomic and proteomic analyses of a systematically perturbed metabolic network', *Science*, 2001, **292**, pp. 929-34

[7] Kholodenko, B.N., Kiyatkin, A., Bruggeman, F.J., Sontag, E., Westerhoff, H.V. and Hoek, J.B.: 'Untangling the wires: a strategy to trace functional interactions in signaling and gene networks', *Proc Natl Acad Sci U S A.*, 2002, **99**(20), pp. 12841-6

[8] Gardner, T.S., di Bernardo, D., Lorenz, D. and Collins, J.J.: 'Inferring genetic networks and identifying compound mode of action via expression profiling', *Science*, 2003, **301**, pp. 102-5

[9] Lee, T.I., Rinaldi, N.J., Robert, F. et al.: 'Transcriptional regulatory networks in Saccharomyces cerevisiae', *Science*, 2002, **298**, pp. 799-804

[10] Sontag, E., Kiyatkin, A. and Kholodenko, B.N.: 'Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data', *Bioinformatics*, 2004, **20**(12), pp. 1877-86

[11] Cho, K.H., Choo, S.M., Wellstead, P. and Wolkenhauer, O.: 'A unified framework for unraveling the functional interaction structure of a biomolecular network based on stimulus-response experimental data', *FEBS Lett.*, 2005, **579**(20), pp. 4520-8

[12] Gilman, A. and Ross, J.: 'Genetic-algorithm selection of a regulatory structure that directs flux in a simple metabolic model', *Biophys. J.*, 1995, **69**(4), pp. 1321-33

[13] Koza, J.R., Mydlowec, W., Lanza, G., Yu, J. and Keane, M.A.: 'Reverse engineering of metabolic pathways from observed data using genetic programming', *Pac. Symp. Biocomput.*, 2001, pp. 434-45

[14] Cho, D.Y., Cho, K.H. and Zhang, B.T.: 'Identification of biochemical networks by S-tree based genetic programming', *Bioinformatics*, 2006, **22**(13), pp. 1631-40

[15] Voit, E.O. and Almeida, J.: 'Decoupling dynamical systems for pathway identification from metabolic profiles', *Bioinformatics*, 2004, **20**(11), pp. 1670-81

[16] Sorribas, A., Samitier, S., Canela, E.I. and Cascante, M.: 'Metabolic pathway characterization from transient response data obtained in situ: parameter estimation in S-system models', *J. Theor. Biol.*, 1993, **162**(1), pp. 81-102

[17] Almeida, J.S. and Voit, E.O.: 'Neural-network-based parameter estimation in S-system models of biological networks', *Genome Informatics*, 2003, **14**. pp. 114-123

[18] Polisetty, P.K., Voit, E.O. and Gatzke, E.P. 'Identification of metabolic system parameters using global optimization methods', 2006, *Theor Biol Med Model.*, **3**(4)

[19] Wedelin, D.: 'Efficient estimation and model selection in large scale graphical models', *Stat. Comp.*, 1996, **6**, pp. 313-323

[20] Schittkowski, K.: 'Numerical data fitting in dynamical systems : a practical introduction with applications and software', (Dordrecht, Kluwer Academic, 2002)

[21] 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

[22] Englezoz, P. and Kalogerakis, N.: 'Applied parameter estimation for chemical engineers' (Marcel Dekker, Inc., New York, 2001)

[23] de Boor, C.: 'A practical guide to splines' (Springer-Verlag, New York, 1978), pp. 235-43

[24] Bock, H.G., 'Recent advances in parameter identification techniques for ODE', in: P. Deuflhard, P. and Hairer, E. (Eds.), 'Numerical Treatment of Inverse Problems in Differential and Integral Equations', Vol. 2, Progress in Scientific Computing, Birkhuser, Basel, 1983

[25] Timmer J.: 'Modeling noisy time series: Physiological tremor', *Int. J. Bifurcation Chaos*, 1998, **8**(7), pp. 1505-16

[26] Rodriguez-Fernandez, M., Mendes, P., Banga, J.R.: 'A hybrid approach for efficient and robust parameter estimation in biochemical pathways', 2006, *Biosystems*, **83**(2-3), pp. 248-65

[27] Rodriguez-Fernandez, M., Egea, J.A., Banga, J.R., 'Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems', 2006, *BMC Bioinformatics*, **7**, 483

[28] 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

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

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

[31] 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

[32] Witten, I.H. and Frank, E.: ’Data mining: practical machine learning tools and techniques with Java implementations’ (Morgan Kaufmann, U.S.A., 1999)

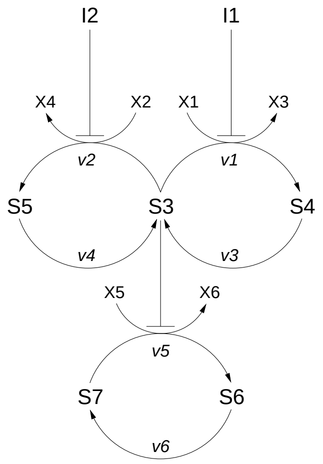

Figure 1: *The metabolic test system [1].* $I_1$ *and* $I_2$ *are input variables,* $S_3 - S_7$ *are measured variables,* $X_1 - X_6$ *are variables corresponding to metabolites assumed buffered at constant levels. The reactions* $v_1 - v_6$ *follow Michaelis-Menten kinetics (non-competitive inhibition) and are catalyzed by different enzymes which also are present at constant levels. The corresponding ODEs are given in the Appendix.*

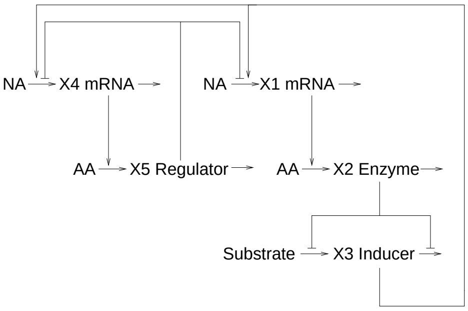

Figure 2: *The genetic network test system [2, 3]. The dependent variables* $X_1 - X_5$ *are measured, while NA (nucleic acid), AA (amino acid) and Substrate are assumed at constant levels. The corresponding ODEs are given in the Appendix.*

| Method | Processor frequency | Running time (and accuracy, relative parameter error) using data with | | |
|---|---|---|---|---|
| | | 0% noise | 3% noise | 5% noise |
| Moles [21] | 866MHz | 39h ($<$16%) | | |
| Rodriguez-F. [26] | 1.8GHz | 2-3h ($<$0.02%) | 2-3h | 2-3h |
| Rodriguez-F. [27] | 1.8GHz | $\approx$5min ($<$6x10$^{-3}$%) | | |
| Our method | 2.7 GHz | 1min$^a$ ($<$4x10$^{-9}$%) | 1.5min$^a$ | 35min$^b$ |

Table 1: *Comparison of running times and accuracies between different methods solving the parameter estimation test case. In our method, we set the smoothing parameter to 0.001 in the data interpolation, we used uniformly distributed starting points in logarithmic space for each variable in step 2a and 2b and we ran the parameter estimation for 2 iterations. All tests were repeated 5 times with similar results. For noisy data, we obtained an error smaller or equal to the error using the true parameters. a) 80 random starting points in step 2a (no random starting points in step 2b). b) 300 and 3000 random starting points in phase 2a and 2b, respectively.*

| Model space | | Parameter bounds |
|---|---|---|
| Reaction | Variables | |
| $kY_1(t)$ | $Y_1 \in \{S_{3-7}\}$ | $k \in [0, 50]$ |
| $kY_1(t)Y_2(t)$ | $Y_1 \in \{S_{3-7}\}$<br>$Y_2 \in \{I_{1-2}, S_{3-7}\}$ | $k \in [0, 50]$ |
| $\frac{Y_1 V_m}{Y_1 + K_D}$ | $Y_1 \in \{S_{3-7}\}$ | $V_m \in [0, 50]$<br>$K_D \in [0.1, 50]$ |
| $\frac{Y_1(t) V_m}{Y_1(t) + K_D} \cdot$<br>$\frac{1}{1 + Y_2(t)/K_I}$ | $Y_1 \in \{S_{3-7}\}$<br>$Y_2 \in \{I_{1-2}, S_{3-7}\}$ | $V_m \in [0, 50]$<br>$K_D \in [0.1, 50]$<br>$K_I \in [0.1, 50]$ |

Table 2: *Model space and parameter bounds considered for the metabolic test system.*

| Reaction | Correct | Estimated parameters | | | |
|---|---|---|---|---|---|
| | param. | n=7 | n=13 | n=25 | n=25 |
| | | s=0.01 | s=0.01 | s=0.01 | s=0.1 |
| | | $\lambda$=1 | $\lambda$=5 | $\lambda$=5 | $\lambda$=5 |
| | | $\sigma_j$=1 | $\sigma_j$ = noise$\hat{X}_j$ | $\sigma_j$ = noise$\hat{X}_j$ | $\sigma_j$ = noise$\hat{X}_j$ |
| | | noise=0% | noise=3.5% | noise=10% | noise=20% |
| $\frac{S_3(t)V_m}{S_3(t)+K_D}\cdot$ | $\|V_m\|$=5 | 5.03 | 5.0 | 5.4 | 5.8 |
| $\frac{1}{1+I_1(t)/K_I}$ | $K_D$=5 | 5.00 | 4.8 | 4.8 | 4.5 |
| in $S_3'$ and $S_4'$ | $K_I$=1 | 0.993 | 1.0 | 0.9 | 0.9 |
| $\frac{S_3(t)V_m}{S_3(t)+K_D}\cdot$ | $\|V_m\|$=5 | 4.99 | 5.3 | 6.3 | 5.3 |
| $\frac{1}{1+I_2(t)/K_I}$ | $K_D$=5 | 5.00 | 4.9 | 4.8 | 4.6 |
| in $S_3'$ and $S_5'$ | $K_I$=1 | 1.00 | 0.9 | 0.7 | 0.8 |
| $\frac{S_4V_m}{S_4+K_D}$ | $\|V_m\|$=1 | 1.00 | 1.0 | 1.0 | 1.0 |
| in $S_3'$ and $S_4'$ | $K_D$=5 | 5.00 | 5.0 | 5.1 | 4.7 |
| $\frac{S_5V_m}{S_5+K_D}$ | $\|V_m\|$=1 | 1.00 | 1.0 | 0.9 | 0.8 |
| in $S_3'$ and $S_5'$ | $K_D$=5 | 5.00 | 5.0 | 4.8 | 4.2 |
| $\frac{S_7(t)V_m}{S_7(t)+K_D}\cdot$ | $\|V_m\|$=10 | 10.0 | 9.2 | 10 | 16 |
| $\frac{1}{1+S_3(t)/K_I}$ | $K_D$=5 | 5.00 | 5.0 | 5.0 | 5.1 |
| in $S_6'$ and $S_7'$ | $K_I$=1 | 0.995 | 1.1 | 0.9 | 0.6 |
| $\frac{S_6V_m}{S_6+K_D}$ | $\|V_m\|$=1 | 1.00 | 1.0 | 1.0 | 1.0 |
| in $S_6'$ and $S_7'$ | $K_D$=5 | 5.00 | 4.6 | 5.5 | 5.5 |
| FALSE POSITIVE REACTIONS: | | | | | |
| $kS_4S_6$ in $S_6'$ | $k$=0 | | -$10^{-6}$ | | |
| $kS_5S_6$ in $S_6'$ | $k$=0 | | -$10^{-6}$ | | |
| $kS_4S_1$ in $S_6'$ | $k$=0 | | $10^{-5}$ | | |
| $kS_2S_5$ in $S_5'$ | $k$=0 | | | -$10^{-5}$ | -$10^{-4}$ |
| $kS_6$ in $S_6'$ | $k$=0 | | | -$10^{-3}$ | -$10^{-3}$ |
| $kI_1S_4$ in $S_6'$ | $k$=0 | | | $10^{-4}$ | $10^{-4}$ |
| $kI_1S_7$ in $S_7'$ | $k$=0 | | | -$10^{-5}$ | -$10^{-4}$ |
| $kS_6S_7$ in $S_3'$ | $k$=0 | | | | $10^{-5}$ |
| $\frac{S_7(t)V_m}{S_7(t)+K_D}\cdot$ | $\|V_m\|$=0 | | | | $10^{-4}$ |
| $\frac{1}{1+S_2(t)/K_I}$ | $K_D$ | | | | 28 |
| in $S_5'$ | $K_I$ | | | | 0.1 |

Table 3: *Models identified from data simulated from the metabolic test system with noise levels of 0, 3.5, 10 and 20%. $n$ is the number of data-points per variable and experiment, $s$ is the smoothing parameter, and $\lambda$ and $\sigma_j$ are the constant in Eq. 2. Since every reaction in the test system belongs to two ODEs, absolute values of $V_m$ are presented.*