Metabolic Pathway Inference from Time Series Data: A Non Iterative Approach

Laura Astola[1,2], Marian Groenenboom[1,2], Victoria Gomez Roldan[3], Fred van Eeuwijk[1,2], Robert D. Hall[3,4], Arnaud Bovy[3,4], and Jaap Molenaar[1,2]

[1] Biometris, Wageningen University and Research Centre, Droevendaalsesteeg 1, Wageningen, The Netherlands
[2] Netherlands Consortium for Systems Biology, Amsterdam, The Netherlands
[3] Bioscience, Plant Research International, Wageningen University and Research Centre, Wageningen, The Netherlands
[4] Centre for Biosystems Genomics, P.O. Box 98, Wageningen, The Netherlands

Abstract. In this article, we present a very fast and easy to implement method for reconstruction of metabolic pathways based on time series data. To model the metabolic reactions, we use the well-established setting of ordinary differential equations. In the present article we consider a network leading to the accumulation of quercetin-glycosides in tomato (*Solanum lycopersicum*). Quercetin belongs to a group of plant secondary metabolites, generally referred to as flavonoids, which are extensively being studied for their variety of important functions in plants as well as for their potentially health-promoting effects on human. We use time series measurements of metabolite concentrations of quercetin derivatives. In the present setting, the observed concentrations are the variables and the reaction rates are the unknown parameters. A standard method is to solve the parameters by reverse engineering, where the ordinary differential equations (ODE) are solved repeatedly, resulting in impractical computation times. We use an alternative method that estimates the parameters by least squares minimization, and which is, in the order of hundred times faster than the iterative method. Our reconstruction method can incorporate an arbitrary *a priori* known network structure as well as positivity constraints on the reaction rates. In this way we can avoid over-fitting, which is another often encountered problem in network reconstruction, and thus obtain better estimates for the parameters. We test the presented method by reconstructing artificial networks and compare it with the more conventional method in terms of residuals between the observed and fitted concentrations, computing times and the proportion of correctly identified edges in the network. Finally we exploit this fast method to statistically infer the kinetic constants in the flavonoid pathway. We remark that the method as such is not limited to metabolic network reconstructions, but can be used with any type of time-series data that is modeled in terms of linear ODE's.

Keywords: Metabolic network inference, flavonoid pathway reconstruction.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 97–108, 2011.

1 Introduction

Flavonoids are a class of secondary metabolites in plants, most commonly found in fruits and flowers. They are involved in various processes, for example in flower, fruit and seed pigmentation, plant growth, protection against UV radiation, and interaction with micro-organisms [1]. Daily dietary consumption of these compounds has been associated with human health promotion and disease prevention, in particular reducing cardiovascular diseases, certain cancers and other age related diseases [2,3]. In tomato, many genes involved in common flavonoid biosynthetic pathways have been identified. Nevertheless the molecular basis of the structural modifications in flavonoid glycosylation and methylation pathways is still relatively unknown. Glycosylation is an enzymatic process that modifies solubility, chemical stability and the biological properties of flavonoids. It is also crucial for flavonoid accumulation. Several glycosylated flavonoids have been reported in tomato fruits, most of them being derivatives of the flavonol quercetin [4]. In this work we consider the quercetin biosynthetic pathway in tomato seedlings.

Many popular models inferring metabolic reaction networks, rely on ordinary differential equations [5,6,7], although this approach has its limitations, especially when using the conventional approach [8]. In the conventional approach, one starts with an initial guess of the parameters (reaction rates), solves the ODE's and compares the resulting solution curves at discrete time points with the observed values at corresponding time points. If they are not sufficiently similar, one adjusts the parameters and repeats the comparisons until the solutions are close enough to the measurements. Although efficient optimization algorithms are available in most mathematical software packages, this approach is inherently time-consuming, due to the fact that one needs to solve ODE's repeatedly [9,6]. This poses a major problem, especially if one wants to perform a large number of simulations, e.g., to study the effect of perturbations or noise. In such case the computation time of a single reconstruction becomes critical. In this paper, we overcome this by presenting a method for fast reconstruction of metabolic networks from observed metabolite concentration data. In [10], Schmidt et al. introduced a method to infer interactions of a small genetic network via computing the Jacobian of the kinetic equations in the vicinity of a steady state. They build on an example given by Kholodenko et al. [11], proposing improvements by considering a series of constant perturbations. We apply a similar method to time series measurements of flavonoids in tomato seedlings. We adjust and extend their method to allow for constrains in the variables and so that a priori known non-existing interactions can be excluded from the network. This method can also be used to estimate unknown constant influxes from the ambient metabolic system.

This paper is organized as follows. In Sect. 2 we derive our reconstruction framework that modifies and extends the ideas given in [10], using only elementary calculus. In Sect. 3 we perform reconstructions using both, the conventional

reverse engineering and the proposed method. As data for this comparison we take time series generated from artificial networks. We compare the differences in terms of residuals, computing times and the accuracy of the network topology. In Sect. 4, we statistically infer the quercetin glycosylation network by exploiting the fast reconstruction scheme. We finish with conclusions in Sect. 5.

2 Metabolic Network Reconstruction

In this section we specify the mathematical model for the dynamics of metabolic reactions, and derive a fast method for the reconstruction of metabolic networks.

2.1 Modeling Metabolic Interactions

Metabolic pathways are often visualized as graphs, where each node or vertex represents the molar concentration of the substrate participating in the reactions, and the edges represent the mass fluxes between the nodes. To reconstruct such a graph, i.e., to infer the metabolic pathway, we estimate the reaction rates from time-series measurements of concentrations of the compounds involved. A popular and powerful mathematical model for metabolic networks consists of a set of ordinary differential equations, depending on the initial concentrations and the reaction rates [5,6,10,12]. Our present task is to find estimates for the reaction rates such that:

- The model yields a good fit to the observations
- The model is not too sensitive to perturbations/noise
- The number of parameters is as small as possible

We note that in the case of flavonoid pathways, we cannot explicitly measure the concentrations of some boundary(input/output) nodes, due to the extremely fast conversion of one substrate into another. This hampers for example the use of graphical models for initial analysis, since we have missing data. Here, we show that we can still estimate these hidden substrates by including them as constants in the ODE system.

Let us first look at an example of a putative flavonoid network (see Fig. 1) and the corresponding mathematical model. Denoting the concentration of substrate i at time t as $X_i(t)$, $(i = 1, \ldots, 6)$, we can mathematically model this as

$$\begin{aligned}
\dot{X}_1(t) &= -k_{10}X_1(t) - k_{12}X_1(t) - k_{13}X_1(t) + k_{21}X_2(t) + k_{31}X_3(t) + k_{01} \\
\dot{X}_2(t) &= -k_{21}X_2(t) - k_{24}X_2(t) - k_{25}X_2(t) + k_{12}X_1(t) + k_{42}X_4(t) + k_{52}X_5(t) \\
\dot{X}_3(t) &= -k_{31}X_3(t) - k_{36}X_3(t) + k_{13}X_1(t) + k_{63}X_6(t) \\
\dot{X}_4(t) &= -k_{42}X_4(t) + k_{24}X_2(t) \\
\dot{X}_5(t) &= -k_{52}X_5(t) + k_{25}X_2(t) \\
\dot{X}_6(t) &= -k_{63}X_6(t) + k_{36}X_3(t)
\end{aligned} \tag{1}$$

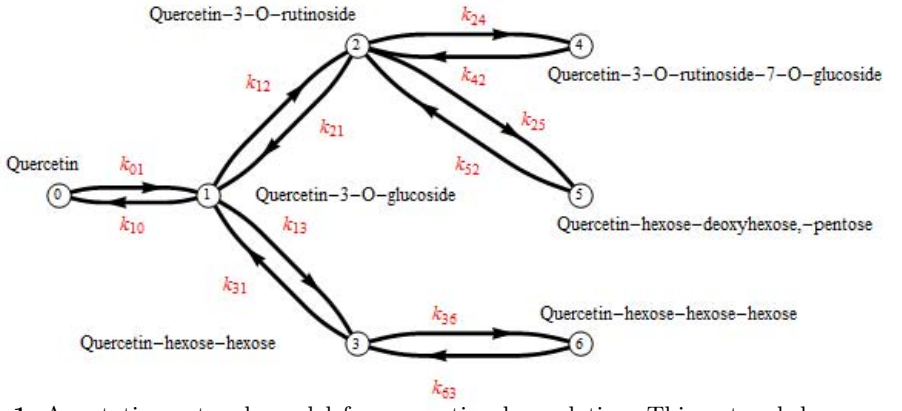

Fig. 1. A putative network model for quercetin glycosylation. This network has one vertex with number 0, which is connected to a larger metabolic network and from which a chemical precursor flows into the network.

In our example pathway, the substrate X_0 cannot be directly measured and has to be estimated. This substrate corresponds to the vertex number 0 connecting the quercetin pathway to the rest of the ambient metabolic network. In Fig. 1 we have drawn those edges that are considered to be relevant from a biological point of view. This putative network is based on the work of Iijima et al. [13] and the pathway model in KEGG [14].

Network inference is concerned with finding those edges that are most consistent with the given data. This may imply that one starts with the assumption that all possible edges are present and subsequently concludes that some rates k_{ij} are zero. A more general formulation of a linear ODE model is

$$\dot{X}_i(t) = -\sum_{j\neq i} k_{ij} X_i(t) + \sum_{j\neq i} k_{ji} X_j(t) + b_i \ , \tag{2}$$

where $(i = 1, \ldots, n)$. To simplify the notation, we introduce a matrix A with components given by

$$\begin{cases} A_{ij} & = k_{ji}, \qquad i \neq j \\ A_{ii} & = -\sum_{j\neq i} k_{ij}, \end{cases} \tag{3}$$

Then, (2) becomes

$$\dot{X}_i(t) = \sum_{j=1}^{n} A_{ij} X_j(t) + b_i \ , \tag{4}$$

with corresponding homogeneous system

$$\dot{X}_i(t) = \sum_{j=1}^{n} A_{ij} X_j(t) \ . \tag{5}$$

For the present reconstruction algorithm we need the concentrations $X_i(t)$ at equidistant time points $t = t_0, t_1, \ldots, t_n$, with $n \geq N$, where N is the number of nodes in the network.

2.2 Derivation of the Objective Function

To reconstruct a metabolic network from time-series measurements, we have to estimate the reaction rates k_{ij}, which give the weights of the edges in the network. Due to (3), it is sufficient to estimate A. In what follows, we present a step by step derivation leading to the minimization problem (16), whose minimizer gives the required estimate for A. We denote the data, i.e., measured concentrations of substrate i at time point t_j, as $\mathbb{X}_{i,j}$.

We start from the well known property that for any solution of a homogeneous linear ODE with constant coefficients, such as the one in (5), it holds that

$$X(t + \Delta t) = \exp(A\Delta t)X(t) \ , \tag{6}$$

where $\exp(M)$ denotes the matrix exponential of M and Δt is some time step.

Now we construct matrices $\mathbb{X}_{\text{new}}$ and $\mathbb{X}_{\text{old}}$ as follows

$$\mathbb{X}_{\text{new}} = \begin{pmatrix} \mathbb{X}_{1,n} & \mathbb{X}_{1,n-1} & \cdots & \mathbb{X}_{1,1} \\ \mathbb{X}_{2,n} & \mathbb{X}_{2,n-1} & \cdots & \mathbb{X}_{2,1} \\ \vdots & & \cdots & \vdots \\ \mathbb{X}_{n,n} & \mathbb{X}_{n,n-1} & \cdots & \mathbb{X}_{n,1} \end{pmatrix} \ , \qquad \mathbb{X}_{\text{old}} = \begin{pmatrix} \mathbb{X}_{1,n-1} & \mathbb{X}_{1,n-2} & \cdots & \mathbb{X}_{1,0} \\ \mathbb{X}_{2,n-1} & \mathbb{X}_{2,n-2} & \cdots & \mathbb{X}_{2,0} \\ \vdots & & \cdots & \vdots \\ \mathbb{X}_{n,n-1} & \mathbb{X}_{n,n-2} & \cdots & \mathbb{X}_{n,0} \end{pmatrix} \ . \tag{7}$$

If the data would perfectly follow the model, we would have that

$$\mathbb{X}_{\text{new}} = \exp(A\Delta t)\mathbb{X}_{\text{old}} \ , \tag{8}$$

where $\Delta t = t_{i+1} - t_i$. We assume the measurement times to be equidistant. Taking the matrix logarithm we find an estimate for A

$$A = \frac{1}{\Delta t} \log\left(\mathbb{X}_{\text{new}}\mathbb{X}_{\text{old}}^{-1}\right) \ . \tag{9}$$

One may often encounter difficulties in inverting $\mathbb{X}_{\text{old}}$. As a remedy one may regularize the matrix using Tikhonov regularization (or ridge regression) [15]. For this, one solves for some small $\alpha > 0$

$$A = \frac{1}{\Delta t} \log\left(\mathbb{X}_{\text{new}}\left(\mathbb{X}_{\text{old}} + \alpha\mathbf{I}\right)^{-1}\right) \ . \tag{10}$$

For an optimal choice of parameter α one may consult, e.g., [16].

We now turn to estimate A from the nonhomogeneous system (4). We append the scalar one to the vector X:

$$X^*(t) = \begin{pmatrix} X_1(t) \\ \vdots \\ X_n(t) \\ 1 \end{pmatrix} \ , \tag{11}$$

and to matrix A, we append the column of influx vectors $\mathbf{b}$.

$$A^* = \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} & b_1 \\ \vdots & & \dots & & \vdots \\ A_{n,1} & A_{n,2} & \dots & A_{n,n} & b_n \end{pmatrix} . \tag{12}$$

Then we may concisely write (4) as

$$\dot{X}_i(t) = \sum_j A^*_{ij} X^*_j(t) \ . \tag{13}$$

The essence of the approach is that we are incorporating the data directly into our expression and that we have a homogeneous structure. Usually, the number of measurements is not equal to the number of unknowns. Thus having a square matrix is more the exception than the rule. As a consequence, typically we cannot solve linear ODE's using (9) or (10). Therefore, we approximate the derivatives with finite differences.

$$\dot{X}_{i,j} \approx \frac{X_{i,j+1} - X_{i,j}}{\Delta t} \approx \sum_k A^*_{i,k} X^*_{k,j} \ , \tag{14}$$

thus in terms of the data matrices introduced in (7) we get the estimate for A^* using pseudo-inverse:

$$A^* = \frac{1}{\Delta t} \left(\mathbb{X}_{\text{new}} - \mathbb{X}_{\text{old}} \right) \mathbb{X}^T_{\text{old}} \left(\mathbb{X}_{\text{old}} \mathbb{X}^T_{\text{old}} \right)^{-1} \ . \tag{15}$$

It goes without saying that this is very fast since it involves only matrix manipulations. On the other hand it can result in over-fitting, since all possible edges are included in the modeled network. Another serious shortcoming of this approach is the fact that we cannot control the positivity of the reaction rates. Although in [10], positive(negative) coefficients were interpreted as activation(inhibition) of the compounds, in many biological pathways, negative coefficients are not allowed. This also holds for the example we will give in Sect. 4. Thus we need a more general approach that does allow sparse networks, where one can exclude all irrelevant edges that are not contained in any biologically feasible model, and in which one can constrain the reaction rates to be positive, without substantially compromising computation time.

To this end, we note that the formula in (15) provides in fact an explicit solution of the following minimization problem

$$\arg\min_{A^*} \left(||A^* \mathbb{X}_{\text{old}} - \frac{1}{\Delta t} \left(\mathbb{X}_{\text{new}} - \mathbb{X}_{\text{old}} \right) ||^2 \right) \ . \tag{16}$$

This alternative formulation allows inclusion of expert knowledge in a simple way. E.g., we can at will put $A^*_{ij} = 0$, when an edge from node i to node j can not exist. Nearly all mathematical software packages (Mathematica, Matlab, Maple etc.) can numerically find the minimizer A^* (and thus the reaction rates k_{ij}) with the constraint that $k_{ij} \geq 0$.

3 Experiments with Artificial Data

In the conventional reverse engineering method the parameters k_{ij} are estimated, using optimization algorithms to minimize the sum of squares between the ODE solutions and the concentration measurements. This involves repeated solving of the ODE's, which is the major time consuming part in the process [6]. We compare our direct inference method with this conventional reverse engineering method. As an example we present here the case with six nodes, where five nodes X_1, X_2, X_3, X_4, X_5 correspond to measurable concentrations and one node X_0 is a boundary node that connects the network to the surroundings. We generated artificial networks in which both, the positions and weights of the edges were randomly chosen. For such a random network the ODE's in (2) were solved. We sampled these solution curves and subsequently reconstructed the original network based on these samples.

We generated these artificial networks as follows. We chose a (uniformly distributed) random integer to determine the number of zeros in an adjacency matrix for nodes X_2, X_3, X_4, X_5, X_6. After determining the topology of the network in this way, we assigned a (uniformly distributed) random real number between zero and one as the weight for each edge independently.

We compared the reconstructions using both, the conventional method and our proposed method, first by using exact samples and then by adding $\pm 10\%$ (uniformly distributed) noise to the samples. Finally we did reconstructions assuming that the topology of the network is known *a priori*. This can be compared to the situation when reconstructing real metabolic networks, since one usually has some putative information on the possible connections between substrates. A typical result using 20 sample points is plotted in Fig. 2. We observe that although the conventional method is tuned to closely approximate the solution curves, the resulting networks are not necessarily closer to the original. While it is obvious that the fast reconstruction method based on first order approximation will generally give larger residual with respect to the original data, another question is, what does this mean in terms of reconstructed networks. To answer this question experimentally, we generated random networks as described before. Then, to simulate a typical reconstruction situation, where only a minimum amount of data is available, we did repeated reconstructions using only six sample points. From each reconstruction, we recorded the residuals (i.e., the distances between the reconstruction and the original function at sample points) and the computation times in seconds. We plotted them in logarithmic scale to be able to include large values in the picture. In addition to this we compared the topologies of the network adjacency matrices. That is we counted all those edges that were missing or redundant compared to the adjacency matrix of the original network. The results for 100 reconstructions are shown in Fig. 3. It seems that the iterative method, while demanding a lot of computing time and indeed resulting in better fit, does not necessarily deliver better results in terms of network reconstruction. For illustration of this matter see Fig. 4.

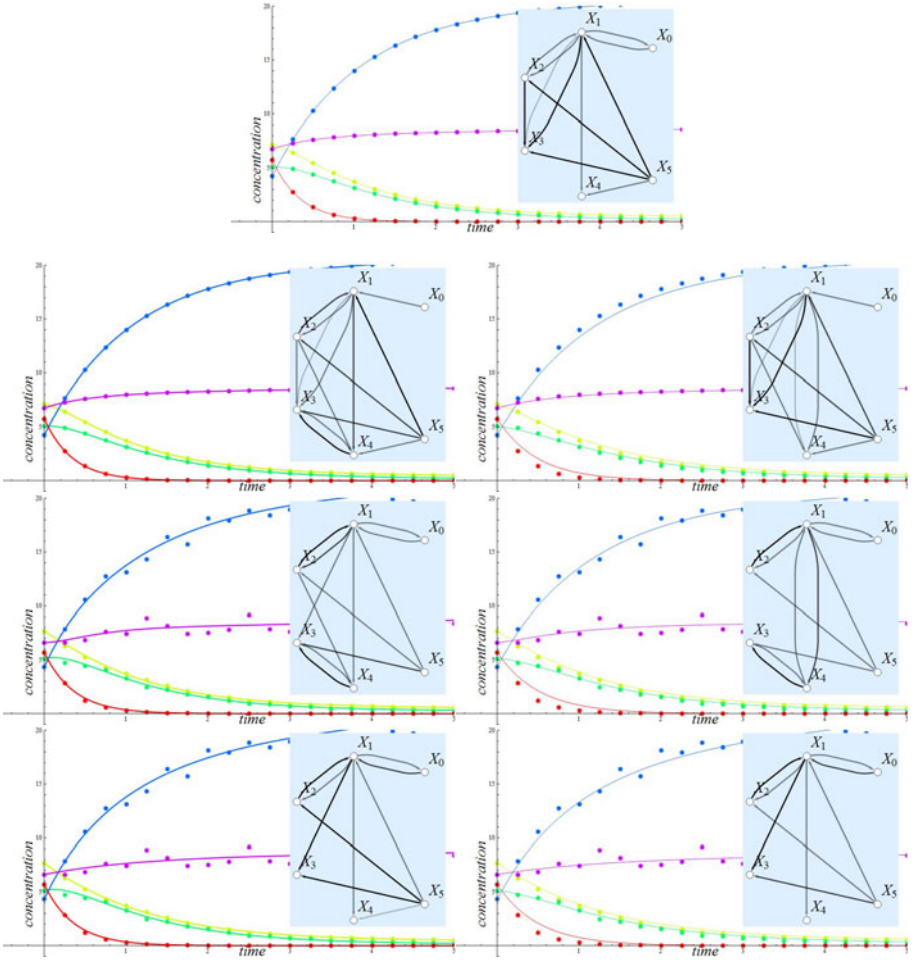

Fig. 2. On the top: the original artificial network and the corresponding ODE solution curves. Left column: reconstructions with the iterative method. Right column: reconstructions with the fast method described in this paper. Top row, reconstructions from exact samples. Middle row, reconstructions from samples with ±10% noise. Bottom row, reconstructions from the same noisy data, when the network topology is known *a priori*.

4 Experiments with Flavonoid Data

The high efficiency of the present method allows a statistical strategy to discriminate between relevant and redundant edges. The idea is to perform repeated

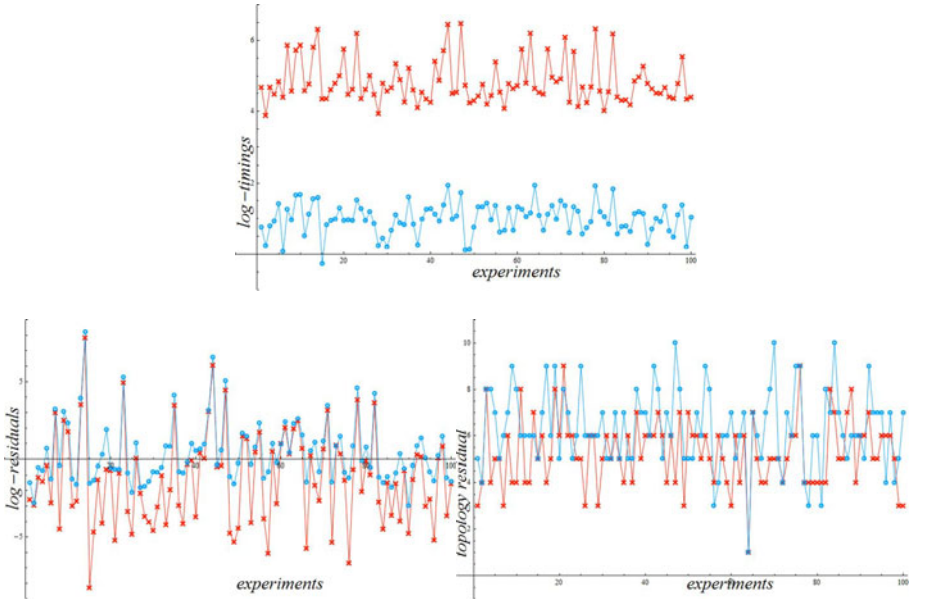

Fig. 3. Red crosses correspond to the iterative method and the blue circles to the fast method described in this paper. Top: logarithms of the computation times, in seconds. Bottom left: logarithms of the residuals (in substrate concentrations) with respect to the original concentration curves. Bottom right: the numbers of missing/redundant edges compared to the original network.

network reconstructions using the putative network in Fig. 1, meanwhile adding random noise to the measurements. If the reconstructions consistently assign a zero value to a parameter k_{ij}, we can suspect that the corresponding edge is not likely to exist in a network derived from an ODE model. In our experiments we took the substrate concentration data of the metabolites involved in the putative quercetin pathway. These concentrations were measured from tomato seedlings during days 5 to 9 after germination [17]. Subsequently, we performed 1000 reconstructions using formula (16), while adding $\pm 10\%$ random noise to the data. The resulting distributions for parameters k_{ij} can be seen in Fig. 5. The number of bars in the histograms is approximately the square root of the number of reconstructions. This kind of simulation can give a significant clue to whether the nodes i and j are connected or not and also provide insight on how sensitive the parameters are w.r.t. noise. From this result we could for example conclude that the edge from node 1 to node 0 and the edge from node 3 to node 6 are redundant. The exact criteria for discarding edges depend on the context of the network.

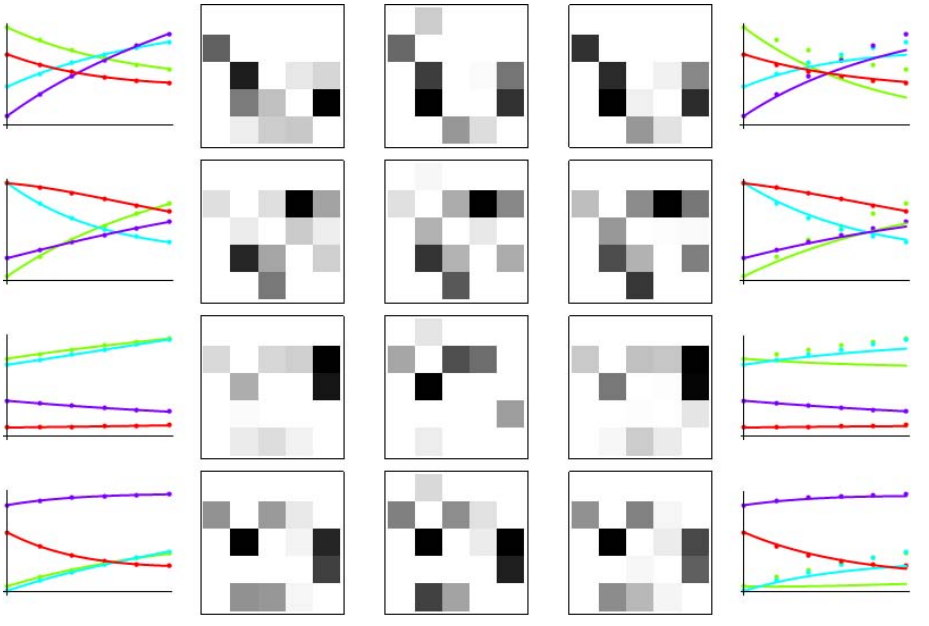

Fig. 4. Columns from left to right: Reconstructed solution curves from samples using iterative method, reconstructed network matrices using iterative method, original network matrices with which the data was created, reconstructed network matrices using the fast method described in this paper, and correspondingly the solution curves reconstructed with the fast method.

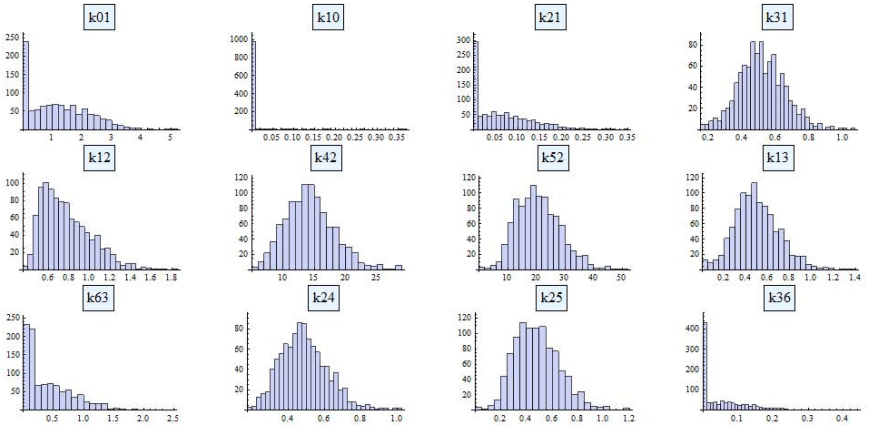

Fig. 5. Histograms showing the distributions of reaction rates k_{ij} estimated in a series of 1000 reconstructions. A pre-selection of relevant edges was done based on the putative model in Fig. 1. In each reconstruction the concentration data were randomly perturbed with $\pm 10\%$ noise. From the results, we can immediately distinguish those coefficients k_{ij} that are distributed around zero and those which in turn accumulate around a positive value.

5 Conclusions

We have experimented with a method for network reconstruction that is very fast compared to the conventional approach. The only requirements are that time-series data are available, the dynamics of the network can be modeled with ODE's, and that the number of measurements $n \geq N + 1$, where N is the number of nodes. Although our approach is inspired by [10], the application is different, since their work considers the approximation of the Jacobian of kinetic equations such as those in [11], in the vicinity of steady state and their time series consists of *in silico*, constant rate perturbations to a maximal enzyme rate. We, on the other hand model *in vivo* measurements, where the unknown influx rates correspond to the constant rate perturbation. In either case the formula (15) to estimate Jacobian is well known in numerical mathematics. We have modified this to a minimization problem (16) to adjust it to our model, where the kinetic constants have to be positive and where we have to be able to exclude nonsensical edges from the network.

The main advantage of this method is that, though it is slightly less accurate than the iterative method that minimizes the residual between the ODE-solutions and measurements, it is significantly faster allowing one to do statistical analysis that require large number of simulations. From the simulations in Sect. 3 we see that, it is around hundred times faster than the conventional iterative method and thus highly suitable for repeated reconstructions. We remark that the residual between the solution curves is not the best measure of successful network reconstruction. We have also experimentally observed (see Fig. 3) that in terms of the network structure, i.e., the adjacency matrix of the nodes, the proposed method performs similarly to the residual-based iterative method.

Acknowledgements. This work results from a collaboration between plant biologists, statisticians and mathematicians, initiated by the Netherlands Consortium for Systems Biology (NCSB) and Centre for Biosystems Genomics (CBSG) and financed by the Netherlands Genomics Initiative.

References

1. Koes, R., Quattrocchio, F., Mol, J.: The flavonoid biosynthetic pathway in plants: Function and evolution. The American Journal of Clinical Nutrition 16(2), 123–132 (1994)
2. Martin, C., Butelli, E., Petroni, K., Tonelli, C.: How can research on plants contribute to promoting human health? Plant Cell (May 2011), doi:10.1105/tpc.111.083279
3. Bovy, A., Schijlen, E., Hall, R.: Metabolic engineering of flavonoids in tomato Solanum lycopersicum: the potential for metabolomics. Metabolomics 3(3), 399–412 (2007)
4. Slimestad, R., Fossenn, T., Verheul, M.: The flavonoids of tomatoes. J. Agric. Food Chem. 56(7), 2436–2441 (2008)

5. Hatzimanikatis, V., Floudas, C., Bailey, J.: Analysis and design of metabolic reaction networks via mixed-integer linear optimization. AIChE Journal 42(5), 1277–1292 (1996)
6. Chou, I.-C., Voit, E.: Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math Biosci. 219(2), 57–83 (2009)
7. Zhan, C., Yeung, L.: Parameter estimation in systems biology models using spline approximation. BMC Systems Biology 5(14) (2011)
8. Hendrickx, D., Hendriks, M., Eilers, P., Smilde, A., Hoefsloot, H.: Reverse engineering of metabolic networks, a critical assessment. Molecular BioSystems 7, 511–520 (2011)
9. Kimura, S., Nakayama, S., Hatakeyama, M.: Genetic network inference as a series of discrimination tasks. Bioinformatics 25(7), 918–925 (2009)
10. Schmidt, H., Cho, K.-H., Jacobsen, E.: Identification of small scale biochemical networks based on general type system perturbations. The FEBS Journal 272, 2141–2151 (2005)
11. Kholodenko, B., Kiyatkin, A., Bruggeman, F., Sontag, E., Westerhoff, H., Hoek, J.: Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. Proc Natl. Acad. Sci. USA 99(20), 12841–12846 (2002)
12. Jha, S., van Schuppen, J.: Modelling and control of cell reaction networks. Pnar0116, CWI, Amsterdam (2001)
13. Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., Sakurai, N., Suda, K., Suzuki, T., Suzuki, H., Okazaki, K., Kitayama, M.: Metabolite annotations based on the integration of mass spectral information. The Plant Journal 54(5), 949–962 (2008)
14. Kyoto Encyclopedia of Genes and Genomes: Flavone and Flavonol Biosynthesis (2010), `http://www.genome.jp/kegg/pathway/map/map00944.html`
15. Golub, G., Hansen, P., O'Leary, D.: Tikhonov regularization and total least squares. SIAM J. Matrix Anal. & Appl. 21(1), 185–194 (1999)
16. Hansen, P., O'Leary, D.: The use of the l-curve in the regularization of discrete ill-posed problems. SIAM J. Sci. Comput. 14(6), 1487–1503 (1993)
17. Gomez-Roldan, M.V., Bovy, A., de Vos, R., Groenenboom, M., Astola, L.: LC-MS metabolite profiling on tomato seedlings in a systems biology approach. In: Metabomeeting, Helsinki, Finland (September 2011)