INFERRING THE GENES UNDERLYING FLAVONOID PRODUCTION IN TOMATO

Laura Astola[1,2], *Victoria Gomez-Roldan*[2,3] *and Jaap Molenaar*[1,2]

[1]Biometris, Wageningen University and Research Centre,
P.O. Box 100, 6700 AC Wageningen, The Netherlands
[2]Netherlands Consortium for Systems Biology, Amsterdam, The Netherlands
[3]Bioscience, Plant Research International,
Wageningen University and Research Centre, The Netherlands
laura.astola@wur.nl, victoria.choserot@wur.nl, jaap.molenaar@wur.nl

ABSTRACT

Flavonoids are plant secondary metabolites that are extensively studied for their proposed positive effects on human health. They are the end products of a cascade of enzymatic reactions that convert initially toxic substances to glycosylated forms. To determine which enzymes are precisely responsible for which conversions is by far not trivial, since hundreds of candidate genes are in principle capable of performing the transformation of interest. In this paper we propose a method to solve this problem for the glycosylation of flavonoids by coupling gene expression data to the metabolic pathway underlying glycosylation. The core of the method is to estimate time dependent coefficients in a highly efficient way. To show how this approach performs, we apply this method to study the flavonoid glycosylation pathway in tomato (*Solanum lycopersicum*) seedlings.

INTRODUCTION

In tomato seedlings, over 200 putative glycosyl transferases [1] constitute the set of potential enzymes that catalyze the reactions of interest. The experimental validation of each glycosylation process using purified target proteins is costly and time consuming. Therefore, we want to limit the number of enzyme candidates by mathematical modeling, using both, the metabolite concentration and the gene expression data. In order to simulate and analyze the glycosylation processes, we first need to have a sufficiently descriptive model system.

Whereas gene and signaling networks require the inference of the network architecture as well as the estimation of the network parameters, in metabolic networks one typically has some *a priori* information on the possible network configurations. This shifts the emphasis from structural inference methods as Boolean networks [2], Bayesian and statistical inference [3,4] towards kinetic parameter estimation methods [5,6]. Since we have time series data for metabolites and gene expressions (measured from same sample material), a reasonable choice is to use ordinary differential equations (ODEs) as a model system [7]. When (as in our case) the kinetic parameters are not known, one may find suitable models in general biochemical systems theory [8]. Typically the identifiability of the parameters is not guaranteed [9]. One approach towards improving the identifiability of the parameters is the so-called dynamic flux estimation (DFE) [10].

The initial set up of our approach is similar to DFE in that the slopes/derivatives of the measured metabolites are estimated directly from the data and also in that the kinetic rates are being solved at each time point. In this paper we discuss first how to estimate the time dependent kinetic rates from a time series of metabolite concentration data, and then how to extract the corresponding potential gene candidates from the time series microarray data. For clarity, we begin by briefly sketching the inference procedure for constant kinetic rates and then generalize this to the time dependent case.

CONSTANT PARAMETER ESTIMATION

We recall that any network can be represented as a graph, where nodes are connected by directed or undirected edges when there is some interaction between these nodes. In a metabolic network a node represents a substrate or a product, and a directed edge from node i to node j means that i can be converted to j by enzymatic activity. To an edge from node i to j, we assign a weight, i.e., the kinetic rate $k_{ij} \geq 0$. This indicates the rate of product formation. In network reconstruction one may find as a result of an estimation procedure that $k_{ij} = 0$. Then we may conclude that there is no edge connecting nodes i and j.

A general time-invariant linear ODE model with constant coefficients and nonhomogeneous source terms, satisfying the mass conservation law, can be written as

$$\dot{X}_i(t) = -\sum_{j \neq i} k_{ij} X_i(t) + \sum_{j \neq i} k_{ji} X_j(t) + b_i \,, \qquad (1)$$

for $i = 1, \ldots, n$. The first summation stands for the edges leaving X_i, the second for the incoming edges, while b_i

represents a possible constant in or outflow. To simplify the notation, we introduce a matrix A with components given by

$$\begin{cases} A_{ij} & = k_{ji}, \qquad i \neq j \\ A_{ii} & = -\sum_{j \neq i} k_{ij}, \end{cases} \tag{2}$$

Then, (1) becomes

$$\dot{X}_i(t) = \sum_{j=1}^{n} A_{ij} X_j(t) + b_i \ , i = 1, \ldots, n \ . \tag{3}$$

To reconstruct the network from time-series measurements, we have to estimate the reaction rates k_{ij}, i.e., the weights of the edges in the network. Due to (2), it is sufficient to estimate the matrix A.

In [11], we experimented with a fast parameter estimation method, where the efficiency was based on the fact that we avoided iterative solving of ODEs by directly substituting the measurements into the ODEs and by approximating the derivatives with finite differences. An alternative and often better approach to obtain approximations for the time derivatives $\dot{X}_i(t_j)$, is to fit splines to the time series data $\mathbb{X}_i(t_j)$. For each metabolite, we have 9 replicates of averaged metabolite concentrations measured per given time point. To obtain curves that represent the data faithfully, we require that the distance between the curves and the measurements are minimal and that at the same time the curves are smooth. To achieve this we fit P-splines, which are B-splines with a penalization for non-smoothness [12]. The coefficient λ of the penalty term can be chosen, e.g., using leave-one-out cross validation.

From these splines, we evaluate the derivative estimates at time points t_j. These estimates are then used as entries in the matrix $\dot{\mathbb{X}}$. In this formulation, the problem of network inference comes down to solving the set of equations given by

$$\dot{\mathbb{X}} = A\mathbb{X} \ . \tag{4}$$

Solving the parameters directly would be fast since it involves only matrix manipulations. However, it often results in over-fitting, since all possible edges are included in the modeled network. Another serious weakness of such a matrix (pseudo-) inversion approach is the fact that we cannot control the positivity of the reaction rates. Although in [13], positive(negative) coefficients were interpreted as activation(inhibition) of the compounds, in many biological pathways, negative coefficients are not allowed. Thus we take a more general approach that allows 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 reformulate the equation as a minimization problem:

$$\arg\min_{A} \left(||\dot{\mathbb{X}} - A\mathbb{X}|| \right) \ . \tag{5}$$

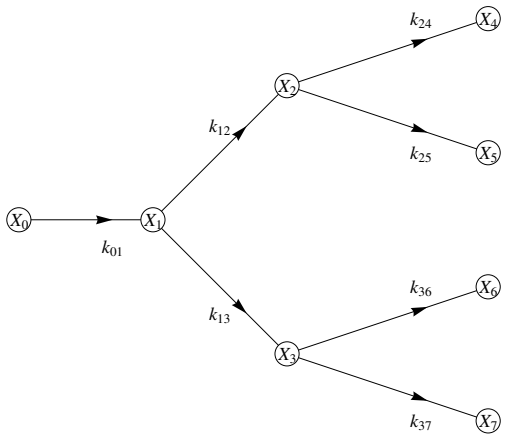

Figure 1. A putative graph for quercetin glycosylation pathway, used as the minimum spanning tree for the networks in the simulations. This is an example of a graph with rooted tree structure.

This alternative formulation allows inclusion of expert knowledge in a simple way. We put $A_{ij} = 0$, when an edge from node i to node j cannot exist.

TIME DEPENDENT PARAMETER ESTIMATION

A shortcoming of the model in the previous section is that it cannot capture the trends in enzyme concentrations which are naturally time varying. To take the enzyme dynamics into account we extend the previous in a straightforward fashion as follows:

- Scheme 1: Fit first natural- or B-splines to data and evaluate estimates for derivatives $\dot{X}_i$. Substitute these estimates and the measurement data $\mathbb{X}_i(t_k)$ into the ODE-system obtaining a set of algebraic equations at each separate time point t_k. Solve first, the constant parameters $k_{ij}(t_k)$, obtaining a set of estimates. Fit a function of choice to these sets over time range.

For example, a second order polynomial may describe the trend of the enzyme activity sufficiently for a relatively short time. We remark that in case the metabolic network in question has a structure of a rooted tree graph, the estimated parameters at each time point are unique. This is an advantage in terms of identifiability of the parameters. The glycosylation pathways for flavonoids such as quercetin and kaempferol are in fact expected to be of this type.

We compare scheme 1 to an alternative standard method:

- Scheme 2: Iteratively solve the ODEs with varying kinetic rates $k_{ij}(t) = \alpha_{ij}t^2 + \beta_{ij}t + \gamma_{ij}$ (or another suitable function of choice), until the solutions

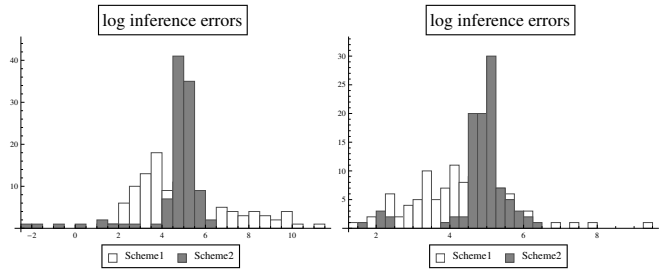

Figure 2. We have compared two different reconstruction schemes versus their errors in 100 simulations. By errors we mean here the sum of squared differences between the original kinetic rates used in the simulations and the reconstructed kinetic rates. On the left: The (logarithmic) errors in the inferred networks using scheme 1 (proposed method) and scheme 2 (iterative method). Right: as in the left hand side but with 10% uniformly sampled noise added to sample data.

$X_i(t)$ are sufficiently close to the measurements at points t_k. Typically this means that an objective function such as $|| \sum_k X_i(t_k) - \mathbb{X}_i(t_k)||$, where $\mathbb{X}_i$ are the measurement vectors, is minimized.

We have compared these two parameter inference schemes using simulated data. In the simulation, to generate artificial data, we assigned pseudo random values to α_{ij}, β_{ij} and γ_{ij} in a range, such that the resulting ODE solutions have approximately same range as the biological data used in the application. The networks used in the simulations were random modifications of the graph in Fig. 1, but with at least one cycle, to make the inference a bit more challenging. In the first set of experiments, we used noiseless data sampled from the simulation results. In the second set, we added $\pm 10\%$ uniformly distributed noise to these same samples.

As can be seen from Fig. 2 the proposed method gives on average the best results, although scheme 1 occasionally succeeds in finding the most accurate estimates, when the data is noiseless. In computation time scheme 1 was on an average 700 times faster than the iterative scheme 2. The comparisons in Fig. 2 were done in a setting where no initial values nor parameter constraints (except for the positivity) were given to the solvers and the parameters were estimated using global search.

APPLICATION IN THE INFERENCE OF ACTIVE GENES

As an application, we consider the inference of the genes behind the enzymatic reactions in metabolic pathways. As an example we take the quercetin glycosylation pathway occurring during the development of a tomato seedling. Quercetin glycosides are a subset of flavonoids, which are plant secondary metabolites naturally produced by plants. Flavonoids are actively studied besides for their important role in protecting the growing plants from external stress, also for their proposed beneficial effects on prevention of

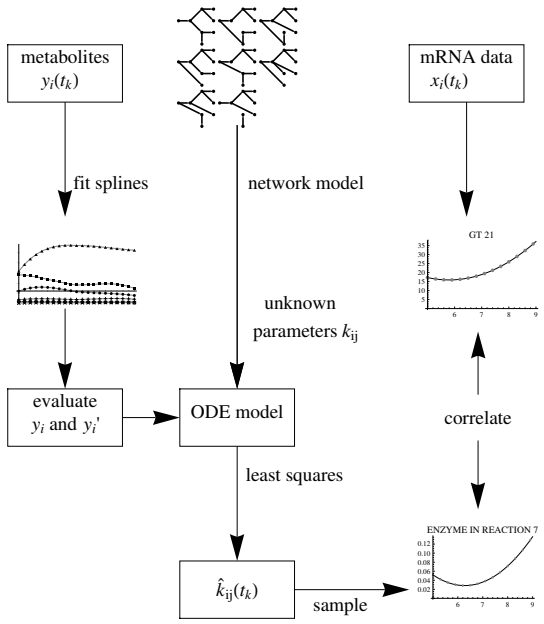

Figure 3. Schematic view of the gene inference procedure.

chronic diseases in humans [14].

In the experiments daily samples were extracted from the seedlings. The same time series sample material from seedlings were analyzed using liquid chromatography mass spectrometer for metabolite concentrations and on a mRNA microarray for the expression levels of glucosyl transferases (GTs).

We use the heuristics that in a mRNA microarray obtained from time series samples, the expression levels sufficiently correlate with the actual protein concentration. Correlation of sample vectors captures the similarity of the (finite) derivatives and curvatures, while ignoring the average values. This is indeed what we want to measure, since the mean values of the expression levels and enzyme concentrations are not likely to be similar, because the units are not physically related. The proposed work flow for the GT inference is briefly as follows:

1. Given the time series metabolite concentration data, estimate the time dependent parameters using all biologically relevant networks. Select the network that gives the best fit to measurements with respect to residual or goodness of fit etc. Save the kinetic rates estimated on the best networks.

2. Compute correlations between the time series of mean expression levels of each GT and the kinetic rates.

3. Select those GTs whose dynamics correlate most with kinetic dynamics.

For convenience, we have summarized this as a schematic diagram in Fig. 3.

As an example, in Fig. 4, we see the expression levels of the three GTs that correlate most with the estimated

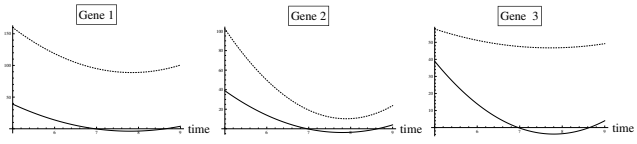

Figure 4. The expression levels (dotted line) of three different glucosyl transferase genes and the estimated kinetic rate (continuous line) for a reaction that converts quercetin to quercetin-3-O-glucoside. The mean expression levels of gene 1 (leftmost frame) correlate best with the predicted enzymatic activity.

kinetic rates for a reaction that glycosylates quercetin to quercetin-3-O-glucoside. Although the units for the predicted enzyme activities and gene expression levels differ, we observe an almost identical shape in the leftmost frame of Fig. 4.

To experimentally test whether the inferred genes are actually transcribing the enzymes that glycosylate the flavonols, a set of selected genes are currently being cloned.

As a computational validation, we tested whether substituting the data of the selected genes into the model will result in better likelihood (of observing the measurements) than when we substitute the other less correlated genes. In the simulations we ran Markov Chain Monte Carlo-algorithm [15] to ensure a rich set of gene combinations and scalings of expression levels. We ordered the genes into a sequence according to their correlation with the predicted enzyme concentration levels. We took two sets of genes according to their order number in the sequence: $1, 2, \ldots, 10$ and $11, 12, \ldots, 20$. We tested whether the residuals corresponding to the simulations using the data of these two sets have equal means and variances. For the mean test we obtained a P-value less than 0.00001 and for the variance test a P-value of less than 0.006. We may conclude that in the context of a dynamic kinetic reaction model, the gene set with high correlation is significantly more likely to have caused the observations.

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). Both the NCSB and CBSG are Centres of Excellence under the auspices of the Netherlands Genomics Initiative.

1. REFERENCES

[1] J. Wang, "Glycosyltransferases: key players involved in the modification of plant secondary metabolites," *Front. Biol. China*, vol. 4, no. 1, pp. 39–46, 2009.

[2] T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinformatics*, vol. 16, no. 8, pp. 727–734, 2000.

[3] D. Husmeier, R. Dybowski, and S. Roberts, *Probabilistic modeling in bioinformatics and medical informatics*, Springer, 2005.

[4] N. Price and I. Shmulevich, "Biochemical and statistical network models for systems biology," *Curr Opin Biotechnol*, vol. 18, no. 4, pp. 365–370, 2007.

[5] B. Palsson, *Systems Biology: Simulation of Dynamic Network States*, Cambridge University Press, 2011.

[6] E. Conrad and J. Tyson, "6. modeling molecular interaction networks with nonlinear ordinary differential equations," in *System Modeling in Cellular Biology, from concepts to nuts and bolts*, Z. Szallasi, J. Stelling, and V. Periwal, Eds. 2010, pp. 97–123, The MIT Press.

[7] W. Chen, M. Niepel, and P. Sorger, "Classic and contemporary approaches to modeling biochemical reactions," *Genes & development*, vol. 24, no. 17, pp. 1861–1875, 2010.

[8] E. Voit, S. Marino, and R. Lall, "Challenges for the identification of biological systems from in vivo time series data," *In Silico Biol.*, vol. 5, pp. 83–92, 2005.

[9] G. Craciun and C. Pantea, "Identifiability of chemical reaction networks," *J Math Chem*, vol. 44, pp. 244–259, 2008.

[10] G. Goel, *A Novel Framework for Metabolic Pathway Analysis*, Ph.d thesis, Wallace H. Coulter Dept. of Biomedical Engineering, Georgia Institute of Technology, Atlanta, December 2009.

[11] L. Astola, M. Groenenboom, V. Gomes Roldan, F. Eeuwijk, R. Hall, A. Bovy, and J. Molenaar, "Metabolic pathway inference from time series data: a non iterative approach," in *Lecture Notes in Bioinformatics*. 2011, vol. 7036, pp. 97–108, Springer.

[12] P. Eilers and B. Marx, "Flexible smoothing with b-splines and penalties," *Statistical Science*, vol. 11, no. 2, pp. 89–121, 1996.

[13] H. Schmidt, K.-H. Cho, and E. Jacobsen, "Identification of small scale biochemical networks based on general type system perturbations," *The FEBS Journal*, vol. 272, pp. 2141–2151, 2005.

[14] A. Bovy, E. Schijlen, and R. Hall, "Metabolic engineering of flavonoids in tomato (*Solanum lycopersicum*): the potential for metabolomics.," *Metabolomics*, vol. 3, no. 3, pp. 399–412, 2007.

[15] D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing: Ten Lectures on Subjective Computing*, vol. 2, Springer, 2007.