Parameter estimation in models combining signal transduction and metabolic pathways: the dependent input approach

N.A.W. van Riel and E.D. Sontag

Abstract: Biological complexity and limited quantitative measurements pose severe challenges to standard engineering methodologies for modelling and simulation of genes and gene products integrated in a functional network. In particular, parameter quantification is a bottleneck, and therefore parameter estimation, identifiability, and optimal experiment design are important research topics in systems biology. An approach is presented in which unmodelled dynamics are replaced by fictitious 'dependent inputs'. The dependent input approach is particularly useful in validation experiments, because it allows one to fit model parameters to experimental data generated by a reference cell type ('wild-type') and then test this model on data generated by a variation ('mutant'), so long as the mutations only affect the unmodelled dynamics that produce the dependent inputs. Another novel feature of the approach is in the inclusion of *a priori* information in a multi-objective identification criterion, making it possible to obtain estimates of parameter values and their variances from a relatively limited experimental data set. The pathways that control the nitrogen uptake fluxes in baker's yeast (Saccharomyces cerevisiae) have been studied. Well-defined perturbation experiments were performed on cells growing in steady-state. Time-series data of extracellular and intracellular metabolites were obtained, as well as mRNA levels. A nonlinear model was proposed and was shown to be structurally identifiable given data of its inputs and outputs. The identified model is a reliable representation of the metabolic system, as it could correctly describe the responses of mutant cells and different perturbations.

1 Introduction

Biomolecular circuits such as regulatory networks and metabolic pathways, play a fundamental role in ongoing research in cell biology. There is increasing awareness that biological processes should be understood integrated in their system environment (systems biology). Although the identification of genes and proteins and the description of metabolic pathways are very important issues, the next step is to understand the dynamics and the function of biomolecular networks. These networks cannot simply be described as an assembly of genes, proteins and metabolites. Mathematical modelling and dynamic simulation are important constituents of systems biology. Systems biology inspires new developments in relevant exact sciences, such as system and control theory [1]. The biological complexity and limited quantitative measurements impose major challenges for the methodologies that are being developed for modelling and simulation. One of the important bottlenecks is the estimation of model parameters from experimental time-series data [2, 3].

This paper originated in our interest in the interaction between metabolic and genetic regulatory networks. In many human diseases, such as type 2 diabetes and heart failure, there are delicate imbalances in these dynamic interactions. In mammalian cells, the amino acids glutamine and glutamate, besides glucose, are the primary nutrients for cell functioning [4]. Glutamine is the most abundant amino acid and an important precursor for peptide and protein synthesis. It serves as a nitrogen transporter in the body and can be used as fuel for different tissues and cell types.

In the low eukaryote *Saccharomyces cerevisiae* (baker's yeast) the structure of the metabolic network of glutamine and glutamate (referred to as the central nitrogen metabolism, CNM; [5]) is similar to that in mammalian cells (Fig. 1). As *S. cerevisiae* has important biotechnological applications for the industrial production of (heterologous and/or engineered) proteins, understanding and rational manipulation of its amino acid and protein metabolism (metabolic engineering) are of direct economical interest.

Cellular metabolism is highly adaptive, which enables cells to select for the 'most optimal' substrate and survive large differences in nutrient availability. Most unicellular organisms regulate the uptake of nutrients via so called catabolic repression: if the cell senses the availability of a preferred substrate, the systems involved in the uptake and processing of 'bad' nutrients will be down-regulated; enzymes are degraded and gene transcription is repressed. In *S. cerevisiae*, the preferred nitrogen sources are glutamine, ammonia and, to a lesser extent, glutamate [6]. The selectivity for these substrates is called nitrogen catabolic repression (NCR). A surplus of glutamine or ammonia also represses its own uptake and metabolism.

IEE Proceedings online no. 20050076

doi:10.1049/ip-syb:20050076

Paper first received 15th October 2005 and in revised form 10th February 2006

N.A.W. van Riel is with the Department of Biomedical Engineering and the Department of Electrical Engineering, Eindhoven University of Technology, PO Box 513, Eindhoven MB 5600, The Netherlands

E.D. Sontag is with the Department of Mathematics and the BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854-8019, USA

E-mail: n.a.w.v.riel@tue.nl

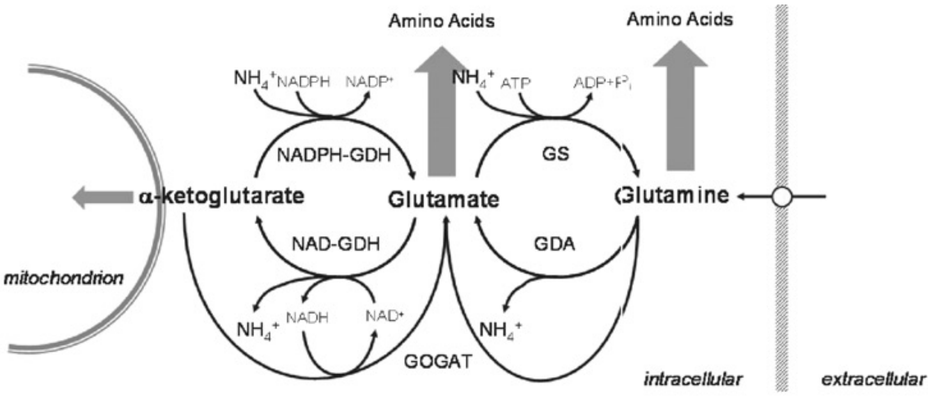

Fig. 1 *Metabolic network of CNM*

GS, glutamine synthetase; GDA, glutaminases; GOGAT, glutamate synthase; NAD/NADPH-GDH, NAD- and NADPH-dependent glutamate dehydrogenase; (NAD(P)(H) are redox cofactors)

One of the primary aims of this study was to illustrate how concepts and methods from system identification and parameter estimation can be used in the development of a dynamic model of a system that integrates signal transduction, gene regulation and metabolic pathways. The model should be able to describe the in vivo behaviour of NCR, such as observed in chemostat experiments [7]. In a chemostat, cells can be grown in a quantitatively well-defined steady-state that is determined by the balanced inflow of fresh culture medium (nutrients, minerals etc.) into the fermenter and the outflow of fermentation broth [8]. A synthetic medium is used of which the composition has been designed such that all but one of the necessary substrates are present in surplus. The inflow of the limiting substrate determines the growth rate of the culture. Glutamine limited chemostats were run and the steady-state was perturbed by pulsing nitrogen substrates to the culture. A model structure with five state variables was derived on the basis of the known molecular mechanisms. The model parameters of the genetic circuit were unknown and those of the substrate kinetics had only been determined by classic biochemical experiments (i.e. 'in vitro'). On the basis of experimental profiles of extracellular and intracellular metabolites measured after excitation of the system, estimates of the parameter values were obtained.

A second major goal of our work was to identify new directions for the extension of system identification theory motivated by the requirements of systems biology. Besides independent inputs u, state variables x and observed outputs y, the mathematical framework was extended with 'dependent inputs' v. These artificial inputs v represent model variables for which the values during a simulation are imposed by the corresponding experimental data (a problem sometimes referred to as 'closed loop identification'), but cannot be manipulated by the experimentator, in contrast to the (classical) independent inputs u. This framework allows study of the processes of interest in a modular fashion. One of the intracellular metabolites, which was measured during the perturbation experiments, was treated as dependent input. Hereby, (the regulation of) the transport systems for the uptake of glutamine and ammonia and the genetic control circuit could be analysed without the need to model all the downstream metabolic pathways. A significant reduction in the complexity of the system to be described is achieved. The dependent input approach is particularly useful in validation experiments, because it allows one to fit model parameters to experimental data generated by a reference ('wild-type') cell type and then testing this model on data generated by a variation ('mutant'), so long as the mutations only affect the unmodelled dynamics that produce the dependent inputs. We justify the approach using the theory of universal inputs for distinguishability [16–22].

We developed a model of *S. cerevisiae*, with seven unknown model parameters that were estimated by an output error approach. As usual, the model was optimised for its predictive power (i.e. the fit of the data) by minimising the difference between the data and the model output in a least squares criterion. However, the model was also optimised for typical, *a priori* known (biological) characteristics of the dynamics of the system. Owing to the experimental setup with a chemostat, the cells were assumed to be in steady-state before each perturbation. In the numerical algorithm, the Least Squares criterion was combined with constraints derived from this experimental steady-state condition in a multi-objective optimisation criterion.

2 Mathematical model

The following general model structure is proposed to describe the dynamics and the model output of a general nonlinear input–output system parameterised by a vector $\boldsymbol{\theta}$ which represents the unknown constants in the system. The state equation is

$$\dot{\boldsymbol{x}}(t, \boldsymbol{\theta}) = f(\boldsymbol{x}(t, \boldsymbol{\theta}), \boldsymbol{u}(t), \boldsymbol{\theta}) \quad \text{with } \boldsymbol{x}(t_0, \boldsymbol{\theta}) = \boldsymbol{x}_0 \qquad (1)$$

(dot indicates time-derivative) and the output equation is

$$\boldsymbol{y}(t, \boldsymbol{\theta}) = \boldsymbol{C}\boldsymbol{x}(t, \boldsymbol{\theta}) \qquad (2)$$

where $\boldsymbol{x} \in \mathbb{R}^n_{\geq 0}$ is the state vector, $\boldsymbol{u} \in \mathbb{R}^r_{\geq 0}$ the input vector and $\boldsymbol{y} \in \mathbb{R}^m_{\geq 0}$ the output vector. We assume in the general formulation that the equations have the property that solutions remain non-negative, when starting with initial states that have non-negative coordinates, and using inputs that are also non-negative. This is a property that is verified when a model represents concentrations, and it can be easily verified for our *S. cerevisiae* model. (Sometimes, quantities of interest in biological models may represent currents in ion channels, metabolic fluxes or other signed quantities; the general considerations apply equally well if we consider systems whose states, inputs and outputs take arbitrary real values.) The components of the vector field f are (nonlinear) functions that describe the structure of the system, parameterised by a vector $\boldsymbol{\theta} \in \mathbb{R}^p_{\geq 0}$. The matrix $\boldsymbol{C}$ selects the states that are observed.

The states in a biomolecular circuit model are typically the levels of messenger RNA (mRNA), proteins and metabolites. Typically, such a network contains regulation loops in which the feedback action is a function of the state variables, as illustrated in Fig. 2. If the state variable(s) acting in the feedback loop can be measured, then the measured signal(s) could be used to drive the system, while the actual feedback is removed

$$\dot{\boldsymbol{x}}(t, \boldsymbol{\theta}) = f(\boldsymbol{x}(t, \boldsymbol{\theta}), \boldsymbol{u}(t), \boldsymbol{v}(t), \boldsymbol{\theta}) \tag{3}$$

with $\boldsymbol{v} \in \mathbb{R}^q_{\geq 0}$, which we will refer to as the 'dependent inputs' (or 'driving function'). Especially for (complex) network systems this concept can be advantageous because it can significantly reduce the model size, as the subsystem comprising the feedback loop does not need to be described. In systems biology, one is often interested in only a subsystem of the total cellular network; it is usually neither feasible nor necessary to model the entire system at the level of the molecular players. Moreover, the use of 'dependent inputs' is a powerful tool in model validation, as we discuss later. The drawback is a loss of predictive power, because the 'open-loop' model can only be used to simulate situations for which the dependent inputs $\boldsymbol{v}$ have been measured in the real system.

We now discuss the model of *S. cerevisiae*, which was developed on the basis of a previously published, more extensive simulation model of CNM in yeast [5]. The model structure was derived from mass balances of the different species in the system. The model should be able to describe the growth of yeast when ammonia and glutamine are used as a nitrogen source (see experiments described in Section 3). For the extracellular metabolites ammonia (x_1) and glutamine (x_2) (both in mM), the chemostat setup yields a description of the inflow of substrate through the medium and the outflow of the fermentation broth (cells, residual medium and metabolites produced by the cells), both at a rate equal to the dilution rate D [h^{-1}] of the chemostat. In (4) and (5), $D(u_i - x_i)$, $i = 1, 2$, represents the net inflow of ammonia and glutamine into the fermenter. A third term in the mass balance represents the substrate uptake by the cells according to Michaelis–Menten kinetics, scaled to the concentration of cells [expressed as the dry cell weight (DCW), in (g × L^{-1})]. Parameters $V_{\max}$ represent the maximum (limiting) specific activity (μmol × g^{-1} × min^{-1}) and K_m the substrate affinity (μmol × g^{-1}).

$$\dot{x}_1 = D(u_1 - x_1) - (\mathrm{DCW})x_4 \frac{V_{\max\mathrm{Mep}} x_1}{K_{\mathrm{Mep}} + x_1} \tag{4}$$

$$\dot{x}_2 = D(u_2 - x_2) - (\mathrm{DCW})x_3 \frac{V_{\max\mathrm{Gapgln}} x_2}{K_{\mathrm{Gapgln}} + x_2} \tag{5}$$

Glutamine and many other amino acids are mainly transported into the cell via the General Amino acid Permease, encoded by the gene *GAP1* and subject to NCR. x_3 is the relative level of the active protein Gap1p. Three genes were identified that encode for ammonia permeases, *MEP1,2,3*. In the model, the three permeases have been lumped as x_4 (the relative level of Mepp). The ammonia uptake system is also catabolically repressed. Repression occurs via inactivation of the transcription factor Gln3 (x_5) that binds to the promoters of the NCR sensitive genes to initiate their transcription [6]. Gln3 is fully active in the nucleus under nitrogen limitation, that is the experimental condition of the chemostat experiments. NCR is triggered when intracellular glutamine, v (μmol × g^{-1}), reaches a critical value, indicated as gln$_T$. Steep sigmoidal functions (Hill equations) have been used to model gene regulation and protein (in)activation (6)–(8). At the protein level, the parameter n represents the 'cooperativity coefficient'. Especially protein activation and inactivation via (de)phosphorylation can occur with relatively high cooperativity [16] such that the sigmoid relation becomes switch-like

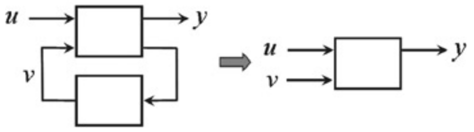

Fig. 2 *Opening the loop of the feedback system by threating a measured variable in the feedback loop as a dependent system input* $\boldsymbol{v}$

$$\dot{x}_3 = k_{\mathrm{s}} x_5 - k_{\mathrm{i}} x_3 \frac{v^n}{gln_T^n + v^n} - k_{\mathrm{d}} x_3 \tag{6}$$

$$\dot{x}_4 = k_{\mathrm{s}} x_5 - k_{\mathrm{i}} x_4 \frac{v^n}{gln_T^n + v^n} - k_{\mathrm{d}} x_4 \tag{7}$$

$$\dot{x}_5 = k_{\mathrm{im}}(1 - x_5) - k_{\mathrm{ex}} x_5 \frac{v^n}{gln_T^n + v^n} \tag{8}$$

where k_{s} is the rate constant of protein synthesis (min^{-1}), k_{i} is the rate constant of NCR triggered inactivation (min^{-1}) and k_{d} the rate constant of protein degradation (min^{-1}). It was assumed that the rate constants are equal for both proteins. k_{im} and k_{ex} are the translocation rate constants of Gln3 to and from the nucleus, respectively, (min^{-1}). In the steady-state of the glutamine limited chemostat, NCR is not active: the sigmoidal function in (6)–(8) is 0, Gap1p (x_3) and Mepp (x_4) are fully expressed (equal to 1) and Gln3 (x_5) is fully active in the nucleus (equal to 1). Therefore $k_{\mathrm{s}} = k_{\mathrm{d}}$ to fulfil the steady-state condition for the permeases. For simplicity, it was assumed that the inactivation rate constant of Gln3 (which is the export rate constant k_{ex}) is equal to the inactivation rate constant of the permeases, k_{i}. Moreover, it was assumed that the translocation rate constants of Gln3 to and from the nucleus are equal. In Fig. 3, the system is shown as a three-compartment system.

The model is reformulated as follows

$$\dot{x}_1 = D(u_1 - x_1) - \alpha_4 x_4 \frac{x_1}{\alpha_1 + x_1} \tag{9}$$

$$\dot{x}_2 = D(u_2 - x_2) - \alpha_3 x_3 \frac{x_2}{\alpha_2 + x_2} \tag{10}$$

$$\dot{x}_3 = \alpha_5(x_5 - x_3) - \beta x_3 \frac{v^n}{\gamma + v^n} \tag{11}$$

$$\dot{x}_4 = \alpha_5(x_5 - x_4) - \beta x_4 \frac{v^n}{\gamma + v^n} \tag{12}$$

$$\dot{x}_5 = \beta(1 - x_5) - \beta x_5 \frac{v^n}{\gamma + v^n} \tag{13}$$

with $\alpha_4 = [\mathrm{DCW}]V_{\max\mathrm{Mep}}$, $\alpha_3 = [\mathrm{DCW}]V_{\max\mathrm{Gapgln}}$, $\alpha_5 = k_{\mathrm{s}} = k_{\mathrm{d}}$, $\beta = k_{\mathrm{i}} = k_{\mathrm{ex}} = k_{\mathrm{im}}$ and $\gamma = gln_T^n$.

In summary, the states x_i are

x_1 = extracellular ammonia concentration $[\mathrm{NH}_4^+]_{\mathrm{ex}}$ in mM,
x_2 = extracellular glutamine concentration $[gln]_{\mathrm{ex}}$ in mM,

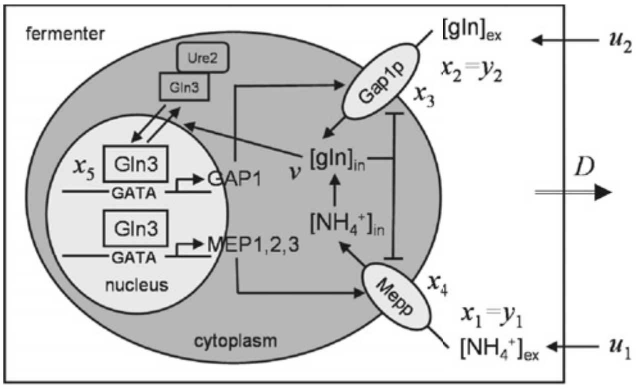

Fig. 3 *Experimental system of a glutamine limited chemostat in which NCR is triggered by pulsing ammonia (u_1) or glutamine (u_2)*

Arrows indicate activation, T-bars represent repression
GATA is the DNA sequence to which transcription factor Gln3 binds
It is assumed that binding of Gln3 results in the same rate of transcription and translation for the different proteins
If intracellular glutamine reaches a threshold, Gln3 translocates into the cytoplasm where it is kept inactive by binding to Ure2
Intracellular ammonia $[NH_4^+]_{in}$ is not included in the model, instead, the measured profile of intracellular glutamine $[gln]_{in}$ is applied as a dependent input v

x_3 = relative level of general amino acid permease Gap1p [−],
x_4 = relative level of ammonia permease Mepp [−],
x_5 = relative level of nuclear transcription activator Gln3 [−],

the independent inputs u_i are

u_1 = ammonia concentration in culture medium feed $[NH_4^+]_{feed}$ in mM,
u_2 = glutamine concentration in feed $[gln]_{feed}$ in mM,

and the dependent input v is:

v = intracellular glutamine concentration $[gln]_{in}$ in [μmol × g^{-1}].

Intracellular glutamine (v) was measured in the experiments in combination with the concentrations of extracellular ammonia (x_1) and extracellular glutamine (x_2) (see Section 3). The output matrix is $\boldsymbol{C} = \mathrm{diag}(1, 1, 0, 0, 0)$ and $\boldsymbol{y} = [x_1, x_2]^T$. The fixed parameters and initial conditions (indicated with superscript 0) are given in Table 1.

3 Experiments

We studied the metabolic and genetic regulations involved in NCR using glutamine limited chemostat cultures of several strains of *S. cerevisiae* (Σ1278b and VWk43) and two different mutants (Δ*gln1* and Δ*glt1*, deficient in glutamine synthetase and glutamate synthase, respectively; [7, 17]). Cells were grown aerobically at 30°C and pH 5.0 in working volumes ranging from 0.5 to 1.5 l. To perturb the steady-state of the glutamine limited cells (growing at a specific growth rate equal to the dilution rate of 0.1 h^{-1}) and trigger the metabolic and genetic regulation, pulses of different type and quantity of nitrogen sources were subsequently added to the fermenter. Between the different pulses, the cells were allowed to recover to steady-state. The profiles of certain intra- and extracellular metabolites and mRNA of NCR sensitive genes (specifically, extracellular ammonia and glutamine concentrations as well as intracellular glutamine concentration) were measured. Data were obtained during steady-state and after the pulses at 0, 1, 2, 3, 4, 6, 8, 10, 15, 20, 30, 45, 60, 90 and 120 min (Fig. 4).

To obtain the data, we performed the following experiment. Fermentation broth was rapidly withdrawn from the fermenter using a syringe. The sample was divided into four fractions. Extracellular metabolites were measured in the supernatant of the first fraction obtained after rapid separation from the cells through filtration. Intracellular metabolites were measured in cell extracts obtained from the second fraction after quenching of protein and metabolic activity in cold buffered methanol below −20°C and subsequent extraction in boiling buffered ethanol. Metabolites were determined by HPLC and/or enzymatic assay. Intracellular metabolite levels were expressed as μmol per gram biomass. Biomass was determined from the third fraction as the dry cell weight in (g × L^{-1}) after overnight drying of the cells at 100°C. The fourth fraction was immediately frozen in liquid nitrogen and used for RNA extraction. Labelled oligonucleotides were used for northern blot analysis. *ACT1* is expressed at constant levels and was used as internal control in the northern blots for the amount of RNA blotted. The blots were scanned and digitised with imaging software. Quantitative expression levels were obtained by calculating the intensity ratio between the gene of interest and *ACT1*, with the observed maximum expression level defined as 100%. Samples were processed and analysed repeatedly (3–5 replicates) to obtain an average value and standard deviation per time sample.

Table 1: Model constants

Symbol	Value	Unit	Description	Source
x_1^0	0	mM	$[NH_4^+]_{ex}^0$	experiment
x_2^0	0.02 ± 0.02	mM	$[gln]_{ex}^0$	experiment
x_3^0	1	[−]	Gap1p^0	experiment
x_4^0	1	[−]	Mepp^0	experiment
x_5^0	1	[−]	Gln3^0	experiment
u_1^0	0	mM	$[NH_4^+]_{feed}$	experiment
u_2^0	20.5	mM	$[gln]_{feed}$	experiment
v^0	8.4 ± 2.5	μmol × g^{-1}	$[gln]_{in}^0$	experiment
D	0.1	h^{-1}	dilution rate	experiment
(DCW)	9.4 ± 0.4	g × L^{-1}	dry cell weight	experiment
n	20	[−]	cooperativity	assumption

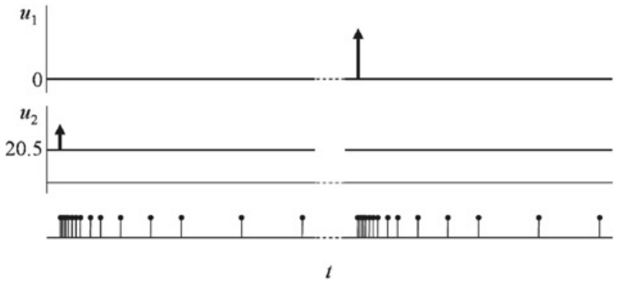

Fig. 4 *Input profile of a 18 mM glutamine pulse added to a 20.5 mM constant feed (u_2) and a 40 mM ammonia pulse (u_1) after steady-state was recovered*

Lower line indicates the timing of the 30 samples

Data of two experiments with wild-type strain Σ1278b have been used for system identification: a 18 mM glutamine pulse and a 40 mM ammonia pulse (Fig. 5). The values reported in Table 1 represent the averages (± standard deviation) as obtained from all steady-state samples. In the model the average values have been used. To validate the identified model, six different experiments were done in either different yeast strains or with different perturbation levels: (1) 18 mM glutamine pulse to a Δ*gln1* mutant, (2) 40 mM ammonia pulse to a Δ*gln1* mutant, (3) 10 mM glutamine pulse to wild-type strain VWk43, (4) 20 mM glutamine pulse to wild-type VWk43, (5) 10 mM glutamine pulse to a Δ*glt1* mutant and (6) 20 mM glutamine pulse to a Δ*glt1* mutant. The Δ*gln1* mutant cannot synthesise glutamine from ammonia after an ammonia pulse and the Δ*glt1* mutant lacks a pathway to degrade glutamine (Fig. 1).

The mutant strains differ from the wild-type strains in precisely the parts of the system that have been left unmodelled and whose effect is represented by the 'dependent inputs'. This means that the identified model should remain the same for the mutant strains, although the 'dependent inputs' used in testing the model when applied to the mutant strains will be different than in the wild-type case. In this manner, the introduction of 'dependent inputs' provides a powerful mechanism for model validation.

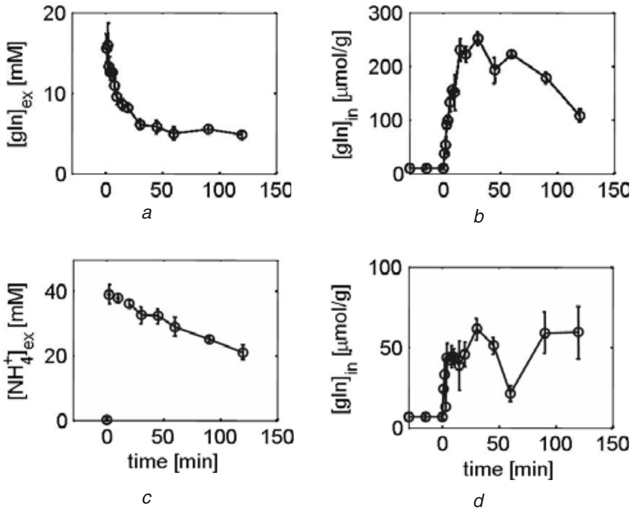

Fig. 5 *Data of two nitrogen pulse experiments with wild-type strain Σ1278b used for identification*

a extracellular glutamine (y_2)
b intracellular glutamine (v) after an 18 mM glutamine (u_2) pulse
c extracellular ammonia (y_1)
d intracellular glutamine (v) after a 40 mM ammonia (u_1) pulse
Bars indicate the standard deviation of the data

4 Identification

4.1 Structural identifiability

As a first step, we investigated if in the ideal, theoretical case, the seven unknown model parameters ($\boldsymbol{\theta} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta, \gamma]$) could be estimated given the two independent system inputs, two outputs and one dependent input. We were able to show that this is, indeed, the case. Moreover, we showed that a small number of combinations of constant values for inputs and dependent inputs suffice for identification. Appendix 9.1 provides a mathematical proof of this fact.

Also in the appendix, we explain how the mathematical theory of 'universal inputs' in control theory [9–15] guarantees that a 'generic' input will be enough for identification in the ideal case, lending considerable support to the whole concept of using dependent inputs.

4.2 Maximum likelihood

Owing to the presence of unmodelled dynamics, modelling errors and measurement noise, the measured data are assumed to be obtained from a stochastic process. The discrete time measurement models are described by

$$\boldsymbol{z}(t_k) = \boldsymbol{y}(t_k) + \boldsymbol{\varepsilon}(t_k) \quad k = 1, \ldots, N \tag{14}$$

$$\boldsymbol{w}(t_k) = v(t_k) + \boldsymbol{\varepsilon}(t_k) \quad k = 1, \ldots, N \tag{15}$$

where $\boldsymbol{z}$ are the measurements of the outputs, $\boldsymbol{w}$ the measurements of the dependent inputs (both sampled at the same, non-equidistant N discrete times t_k) and $\boldsymbol{\varepsilon}$ is the measurement error, assumed to be additive zero mean white noise with known variance $\sigma^2(t_k)$.

The difference between the measurements aligned in z and the simulated time-discrete model output aligned in $\boldsymbol{y}$, that is the model error $\boldsymbol{e}_k$, was weighted in a quadratic criterion J_N

$$\boldsymbol{e}_k = \boldsymbol{z}(t_k) - \hat{\boldsymbol{y}}(t_k, \boldsymbol{u}, \boldsymbol{w}, \hat{\boldsymbol{\theta}}) \quad k = 1, \ldots, N \tag{16}$$

$$J_N(\hat{\boldsymbol{\theta}}) = \boldsymbol{e}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{e} \tag{17}$$

where $\hat{\boldsymbol{\theta}}$ is the vector of estimated parameters, $\hat{\boldsymbol{y}}$ is the model output for the parameter realisation $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{W}$ is a $[m \cdot N \times m \cdot N]$ positive definite symmetric weighting matrix (the weighted least squares algorithm), where $m = 2$ is the dimension of the output space. Then

$$\hat{\boldsymbol{\theta}}_N = \arg \min_{\hat{\boldsymbol{\theta}} \geq 0} J_N(\hat{\boldsymbol{\theta}}) \tag{18}$$

$$\text{subject to} \quad \dot{\boldsymbol{x}}(t_0, \hat{\boldsymbol{\theta}}) = 0 \tag{19}$$

which imposes the steady-state requirement of the chemostat before each pulse experiment. As the parameters have a physiological interpretation, they were bounded to ≥ 0 ($\boldsymbol{\theta} \in \mathbb{R}^p_{\geq 0}$).

The covariance matrix of unbiased parameter estimates $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ has the inverse of the Fisher information matrix (FIM) $\boldsymbol{V_\theta}$ as lower bound ($\mathrm{cov}(\hat{\boldsymbol{\theta}}) \geq \boldsymbol{V_\theta}^{-1}$, the so called Cramér–Rao bound [18, 19]). The FIM is based on the weighted sum of squared residuals $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{e}$ and the Jacobian $\boldsymbol{J}$ of the cost function with respect to the parameters for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and the number of data points N

$$\boldsymbol{V_\theta} = N(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{e})^{-1}\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{20}$$

This accommodates Gaussian model residuals under maximum likelihood estimation and is asymptotically correct for arbitrary distribution of the residuals under weighted least-squares estimation [20, 21]. The weighting matrix $\boldsymbol{W}$ was chosen as the inverse of the data covariance matrix cov($\boldsymbol{z}$). Under this condition $\text{cov}(\hat{\boldsymbol{\theta}}) = \boldsymbol{V}_{\boldsymbol{\theta}}^{-1}$ [18]. Now $\hat{\boldsymbol{\theta}}_N$ is the minimum variance, unbiased estimate and the diagonal elements of the matrix $\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}$ are approximations of the variance of the estimated parameters $(\hat{\sigma}_{\boldsymbol{\theta}}^2)$. Alternatively, the distribution of the parameter estimates could also have been obtained using a bootstrap approach, which does not require the model residuals to be Gaussian distributed.

In $\boldsymbol{z}$ and $\hat{\boldsymbol{y}}$, the data and model output, respectively, of the 18 mM glutamine pulse and 40 mM ammonia pulse to wild-type strain Σ1278b were combined. The data variance was obtained by processing and analysing the same samples multiple times.

4.3 Technical information

The simulations and parameter estimation were carried out in MATLAB 6.5 (The Mathworks, Inc.), running under Microsoft Windows XP Pro on a 2.4 GHz IBM compatible PC with 1 GB RAM. During simulation the independent inputs u_1 and u_2 were defined according to the experimental conditions. Dependent input v was a measured profile. Linear interpolation of the input signals was used to obtain values for each simulation time sample. For parameter estimation the Levenberg–Marquardt algorithm `lsqnonlin` was used from MATLAB's Optimisation Toolbox version 2.2. The termination tolerance for the objective function was set to 10^{-4}. Parameters were estimated with lower bounds equal to zero and the termination tolerance for the parameter estimates was 10^{-5}. The steady-state condition (19) was implemented by augmenting the output error criterion (17) with the sum of the squared vector of the differential equations (9)–(13) at $t_1 = 0$, penalising deviations from steady-state and resulting in a two-objective criterion. Convergence to the global minimum of the objective function cannot be guaranteed. The algorithm was started with different initial values for the unknown parameters to verify potential local minimums.

5 Results and discussion

The estimated parameter values can be found in Table 2. Standard deviations as high as 70% (k_s) have been obtained. On the basis of analysis of the FIM, it was concluded that the system was not sufficiently excited to allow identification of K_{Mep} from these data. This could be explained because $y_1 = [\text{NH}_4^+]_{\text{ex}}$ was zero in steady-state (Table 1) and after the 40 mM ammonia pulse the Mepp system became immediately saturated (i.e. uptake flux equal to V_{maxMep}) and this state was maintained for the following 2 h during which samples were taken.

Simulation results of the identified model are shown in Fig. 6. The estimated NCR threshold level of intracellular glutamine (gln_T) has been included together with experimental data. The response to glutamine showed a 50%

Table 2: Estimated model parameters (mean ± standard deviation)

Parameter	Value	Unit
V_{maxMep}	0.0153 ± 0.0070	mmol × g^{-1} × min^{-1}
K_{Mep}	n.i.	mmol × g^{-1}
$V_{\text{maxGapgln}}$	0.148 ± 0.014	mmol × g^{-1} × min^{-1}
K_{Gapgln}	5.01 ± 0.12	mmol × g^{-1}
gln_T	60.1 ± 1.2	μmol × g^{-1}
k_s	0.0107 ± 0.0074	min^{-1}
k_i	0.0910 ± 0.0103	min^{-1}
k_{im}	0.0910 ± 0.0103	min^{-1}
k_{ex}	0.0910 ± 0.0103	min^{-1}

n.i.: not identifiable, a value of 0.2 was used for simulation

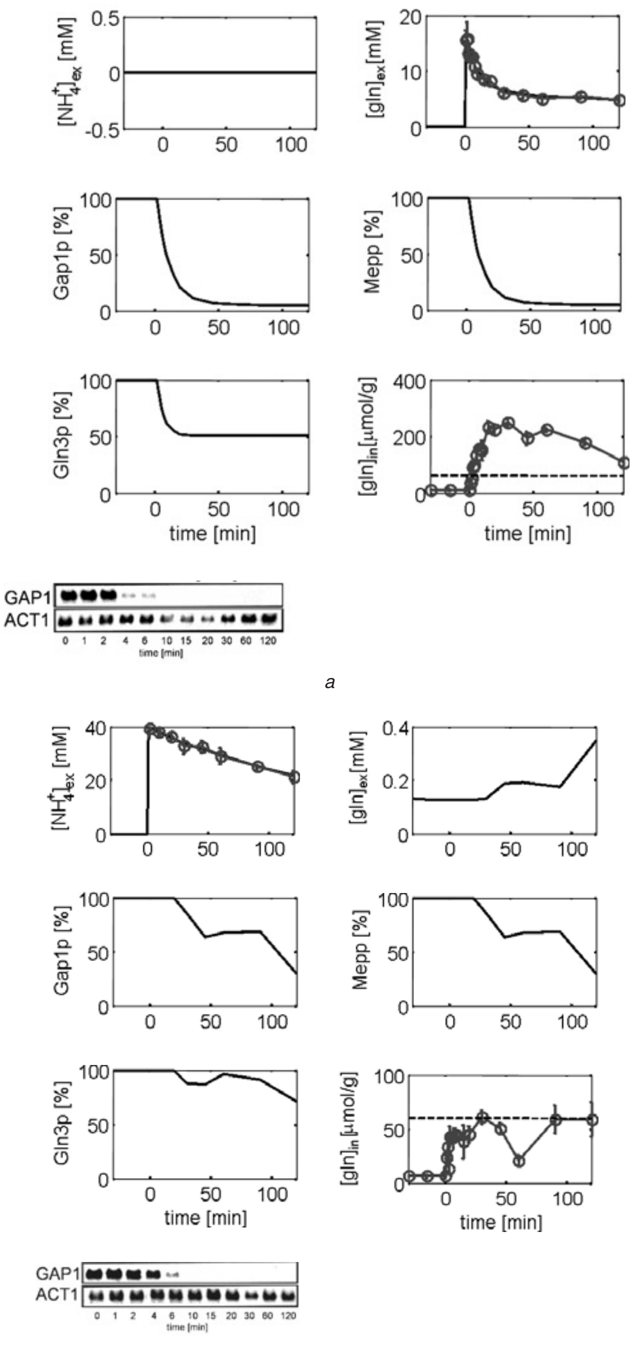

Fig. 6 *Simulation results with identified model*

a 18 mM glutamine pulse

b 40 mM ammonia pulse to wild-type strain Σ1278b

Experimental data (circles) have been included as the dependent input (intracellular glutamine) and to verify the simulated output profiles (extracellular ammonia and extracellular glutamine)

Horizontal dashed line indicates estimated NCR threshold level of intracellular glutamine (gln$_T$)

For GAP1 experimental mRNA levels (northern blot analysis) have been shown

Actine mRNA (ACT1) was used as an internal control for the amount of RNA blotted repression of the transporters Mepp and Gap1p within 10 min after the pulse and only less than 10% activity after 50 min (Fig. 6*a*). The maximal repression of Gln3p activity in the nucleus is only 50%. The model predicted that reactivation occurs 2.5 to 3 h after the pulse, when intracellular glutamine decreased below the threshold level. The response after the ammonia pulse was less trivial (Fig. 6*b*). Initially, the CNM maximally used the increased availability in nitrogen. After 20 min, NCR was activated and, apparently, the system was regulated at an intracellular glutamine concentration close to the threshold level.

A qualitative validation of the model was obtained by comparison of the predicted profiles of Gap1p with the experimental mRNA profiles of *GAP1* included in Fig. 6. The model correctly described a rapid decrease in *GAP1* after the glutamine pulse and a somewhat delayed repression after addition of ammonia. After both pulses, the decrease in the measured transcription levels was faster than the predicted repression of protein activity. Moreover, after the ammonia pulse, *GAP1* was completely repressed after 6 min, whereas Gap1p was never completely repressed according to the model. The residual activity of Gap1p and Mepp allows the cells to maintain growth, and preventing an intracellular overload that might be toxic. For a quantitative validation, the identified model was used to describe the uptake profiles of glutamine and ammonia in six different experiments.

Glutamine limited cultures of mutant strain Δ*gln1*, which lacks glutamine synthetase, were perturbed by pulsing 18 mM glutamine and 40 mM ammonia. As the mutant strain was constructed in the genetic background of the Σ1278b wild-type, it was assumed that the kinetic parameters and rate constants had approximately the same value. In Fig. 7, the model error is plotted (together with the experimental standard deviation in the data). It has to be noted that in the original study, the Δ*gln1* mutant was used to show that NCR is not triggered by intracellular glutamine only, but also by ammonia [6, 7]. The latter mechanism was not incorporated in the model as presented here. This 'undermodelling' could explain the error in the description of the validation data. In both cases, the initial uptake phases (the first 10 min after the pulses before NCR was activated) were predicted correctly. In Appendix 9.2, the model predictions and the data are shown.

The validation of the model with data of the mutant (in which glutamine-triggered NCR after an ammonia pulse was impaired because of the lack of *GLN1*) indicated that in the wild-type, intracellular glutamine caused the main repression, whereas an additional repressive mechanism was apparent in the mutant. This mechanism is probably a signal derived from intracellular ammonia [6]. The activation of this second repressive mechanism was most profound after the ammonia pulse (Fig. 7*b*), when intracellular glutamine did not rise to a level that could trigger NCR (maximum of 15 μmol × g^{-1}; Appendix 9.2). This observation was in agreement with the previous conclusion. The non-white residuals in Fig. 7*a* suggested that this second trigger was also activated after the glutamine pulse to the mutant, in contrast to the response of the wild-type. The analysis with the model showed that, among others, the interaction between glutamine and ammonia metabolism has to be incorporated for a more realistic representation of NCR [5].

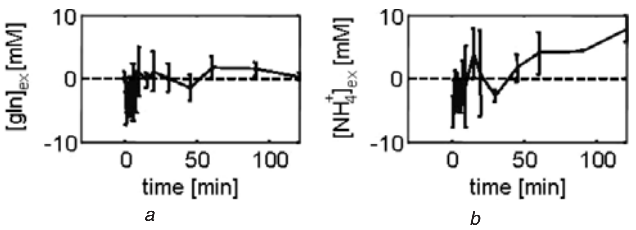

Fig. 7 *Model error in the description of the uptake profiles*
a glutamine
b ammonia after injection to a glutamine limited culture of mutant strain Δgln1 to validate the model
Bars indicate the standard deviation in the experimental data

6 Conclusion

The aim of this paper was to illustrate an approach for data-driven modelling and parameter estimation in combined signal transduction and metabolic systems. The genetic control of nitrogen uptake in *S. cerevisiae* was used as case study.

Six of the unknown parameter values could be estimated, given the limited data set of only 30 samples. The identified model was shown to be a reliable representation of the biological system, because it could correctly describe the responses of different yeast strains and different perturbations. The model clearly showed that intracellular glutamine cannot be the only signal triggering NCR, which is still an ongoing discussion between cell biologists [6]. This systems biology approach to study NCR, provides important insights on how yeast can optimally control nitrogen uptake.

Two generic, highly valuable concepts were introduced for parameter estimation in biomolecular networks. The concept of 'dependent inputs' allows the opening of some of the feedback loops that connect the pathways of interest to the rest of the cell and its environment. The model can focus on a smaller part of the network, as long as this subsystem is still integrated in its in vivo system environment through the measured input signals. If the intracellular level of some of the proteins or metabolites can be (accurately) measured in time, these signals can be used as forcing functions for these inputs. The idea of forcing functions has been applied to other areas of physiological modelling since the 1970s. It is a well-known concept in pharmacokinetic compartmental modelling in whole-body metabolic and endocrine systems, such as glucose homeostasis in which the measured insulin blood plasma profile after a meal or intravenous injection of glucose is used as input to describe the blood glucose levels and estimate physiological parameters such as insulin sensitivity [20]. Other applications are modelling of haemodynamics, in which measured blood flow is used as an input for a model of blood pressure, or vice versa and functional imaging of tumours with dynamic contrast enhanced MRI, in which the arterial profile of an injected contrast agent is used to describe the dynamics of contrast enhancement in the surrounding (tumour) tissue to quantify the endothelial permeability [22]. To our knowledge, this approach has so far not been extended to the biomolecular networks typical in systems biology.

Secondly, *a priori* knowledge was used to improve the *a posteriori* identifiability of the model, that is the model conditioned on the available experimental data. In many biological and biomedical systems, the possibilities to perturb the inputs to excite the system dynamics are limited. Furthermore, when samples are taken from body fluids or tissue and/or are (biochemically) analysed off-line, the number of samples in a time-series data set will be (extremely) limited. A possibility to obtain unique and accurate parameter estimates in sparsely-sampled systems is to include *a priori* information, both quantitative and qualitative, on the system behaviour in the identification criterion. Here, we applied the basal steady-state of the chemostat experiments as additional information to restrict the feasible parameter space. Also this concept can readily be applied to other systems biology applications, although the translation of (qualitative) *a priori* information into a numerical identification criterion and the relative importance of the different objectives in a multi-objective criterion will always be somewhat subjective.

7 Acknowledgments

The authors' collaboration on this work started during a visit by them to the Mathematical Biosciences Institute of Ohio State University, in December 2004. The authors gratefully acknowledge the MBI for an environment conducive to interdisciplinary interactions. Dr. J.M. Guillamón and Dr. E.G. ter Schure (Utrecht University, The Netherlands) are acknowledged for experimental work. E.D.S. also acknowledges partial support from NSF Grants DMS-0504557 and EIA-0205116.

8 References

1 Sontag, E.D.: 'Some new directions in control theory inspired by systems biology', *Syst. Biol.*, 2004, **1**, pp. 9–18

2 Cho, K.H., Shin, S.Y., Kolch, W., and Wolkenhauer, O.: 'Experimental design in systems biology based on parameter sensitivity analysis with Monte Carlo simulation: a case study for the TNFalpha mediated NF-kappaB signal transduction pathway', *Simulation*, 2003, **79**, pp. 726–739

3 Kutalik, Z., Cho, K.-H., and Wolkenhauer, O.: 'Optimal sampling time selection for parameter estimation in dynamic pathway modeling', *Biosystems*, 2004, **75**, pp. 43–55

4 Newsholme, P., Procopio, J., Lima, M.M.R., Pithon-Curi, T.C., and Curi, R.: 'Glutamine and glutamate – their central role in cell metabolism and function', *Cell Biochem. Funct.*, 2003, **21**, pp. 1–9

5 van Riel, N.A.W., Giuseppin, M.L.F., Ter Schure, E.G., and Verrips, C.T.: 'A structured, minimal parameter model of the central nitrogen metabolism in *Saccharomyces cerevisiae*: the prediction of the behaviour of mutants', *J. Theor. Biol.*, 1998, **191**, pp. 397–414

6 Ter Schure, E.G., van Riel, N.A.W., and Verrips, C.T.: 'The role of ammonia metabolism for nitrogen catabolite repression in *Saccharomyces cerevisiae*', *FEMS Microbiol. Rev.*, 2000, **24**, pp. 67–83

7 Ter Schure, E.G., Silljé, H.H.W., Vermeulen, E.E., Kalhorn, J., Verkleij, A.J., Boonstra, J., and Verrips, C.T.: 'Repression of nitrogen catabolic genes by ammonia and glutamine in nitrogen-limited continuous cultures of *Saccharomyces cerevisiae*', *Microbiology*, 1998, **144**, pp. 1451–1462

8 Weusthuis, R.A., Pronk, J.T., van den Broek, P.J., and van Dijken, J.P.: 'Chemostat cultivation as a tool for studies on sugar transport in yeasts', *Microbiol. Rev.*, 1994, **58**, (4), pp. 616–630

9 Sontag, E.D.: 'Mathematical control theory: deterministic finite dimensional systems' (Springer, New York, 1998, 2nd Edn.)

10 Sontag, E.D.: 'On the observability of polynomial systems, I: finite-time problems', *SIAM J. Control Optim.*, 1979, **17**, pp. 139–151

11 Grasselli, O.M., and Isidori, A.: 'Deterministic state reconstruction and reachability of bilinear control processes'. Proc. Joint Automatic Control Conf., San Francisco, June 22–25, 1977, (IEEE, New York), pp. 1423–1427

12 Sussmann, H.J.: 'Single-input observability of continuous-time systems', *Math. Syst. Theory*, 1979, **12**, pp. 371–393

13 Wang, Y., and Sontag, E.D.: 'Orders of input/output differential equations and state space dimensions', *SIAM J. Control Optim.*, 1995, **33**, pp. 1102–1127

14 Sontag, E.D.: 'Spaces of observables in nonlinear control'. Proc. Int. Congress of Mathematicians 1994, (Birkhäuser Verlag, Basel, 1995), vol. 2, pp. 1532–1545

15 Sontag, E.D.: 'For differential equations with r parameters, $2r+1$ experiments are enough for identification', *J. Nonlinear Sci.*, 2002, **12**, pp. 553–583

16 Ferrell, J.E. Jr: 'Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs', *Trends Biochem. Sci.*, 1996, **21**, pp. 460–466

17 Guillamón, J.M., van Riel, N.A.W., Giuseppin, M.L.F., and Verrips, C.T.: 'The glutamate synthase (GOGAT) of *Saccharomyces cerevisiae* plays an important role in the central nitrogen metabolism', *FEMS Yeast Res.*, 2001, **1**, pp. 169–175

18 Ljung, L.: 'Parameter estimation methods' in 'System identification – theory for the user' (PTR Prentice-Hall, Upper Saddle River, NJ, 1999, 2nd Edn.), Ch. 7

19 Walter, E., and Pronzato, L.: 'Qualitative and quantitative experiment design for phenomenological models a survey', *Automatica*, 1990, **26**, (2), pp. 195–213

20 Carson, E.R., Cobelli, C., and Finkelstein, L.: 'The mathematical modeling of metabolic and endocrine systems' (Wiley, New York, 1983)

21 Fedorov, V.V.: 'Theory of optimal experiments' (Academic, New York, 1972)

22 Port, R.E., Knopp, M.V., Hoffmann, U., Milker-Zabel, S., and Brix, G.: 'Multicompartment analysis of gadolinium chelate kinetics: blood-tissue exchange in mammary tumors as monitored by dynamic MR imaging', *J. Magn. Reson. Imaging*, 1999, **10**, pp. 233–241

9 Appendix

9.1 Structural identifiability of the continuous model

We will refer to this system as the model system

$$
\begin{aligned}
\dot{x}_1 &= D(u_1 - x_1) - \alpha_4 x_4 \frac{x_1}{\alpha_1 + x_1}\\
\dot{x}_2 &= D(u_2 - x_2) - \alpha_3 x_3 \frac{x_2}{\alpha_2 + x_2}\\
\dot{x}_3 &= \alpha_5(x_5 - x_3) - \beta x_3 \frac{u_3}{\gamma + u_3}\\
\dot{x}_4 &= \alpha_5(x_5 - x_4) - \beta x_4 \frac{u_3}{\gamma + u_3}\\
\dot{x}_5 &= \beta(1 - x_5) - \beta x_5 \frac{u_3}{\gamma + u_3}
\end{aligned}
$$

where α_i's, β, γ are seven unknown positive constants, states and inputs are non-negative and the outputs are $y_1 = x_1$ and $y_2 = x_2$. Note that, for convenience, we have replaced v^n by u_3. For the purposes of showing identifiability, this replacement is valid because, given any (non-negative) input u_3 for the current system, we may use $v = \sqrt[n]{u_3}$ (recall that $n = 20$ is assumed known) and obtain the same behaviour for the original system.

Assuming that D is known, that the inputs u_1, u_2, u_3 can be manipulated experimentally and that initial condition $\boldsymbol{x}(0) = \boldsymbol{x}^0$ is as in Table 1, we will show that the parameters α_i, β, γ are identifiable from $y_1(t)$ $(t \geq 0)$ and $y_2(t)$ $(t \geq 0)$. (We will show that four constant inputs are enough.)

Of course, in our application, the input u_3 cannot in reality be manipulated, as it represents signals which arise from ignored parts of the system. The assumption that u_3 is a free input is merely a convenience for the identifiability argument. Later, appealing to a theorem from control theory, we show that, in fact, a single generic ('randomly chosen') input function $(u_1(t), u_2(t), u_3(t))$ suffices as well, so that $u_3(t)$ could be the signal arising from the unmodelled subsystem. This theorem from control theory assumes that one has already proved identifiability.

We provide a precise mathematical statement next. It says that if two parameters sets are such that the same outputs result when certain four input functions are applied, then the parameters must coincide.

First, we introduce a notation for outputs. For a system

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta}), \quad y = h(\boldsymbol{x})$$

any initial state $\boldsymbol{x}^0$, any parameter set

$$\boldsymbol{\theta} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta, \gamma)$$

(vector of non-zero numbers), and any time-dependent input function $\boldsymbol{u} = \boldsymbol{u}(\cdot)$, we denote by $F(\boldsymbol{x}^0, \boldsymbol{u}, \boldsymbol{\theta})$ the function $\boldsymbol{y}(\cdot)$, where $\boldsymbol{y}(t) = h(\boldsymbol{x}(t))$ and $\boldsymbol{x}(t)$ is the solution of the initial value problem $\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta})$ with initial condition $\boldsymbol{x}(0) = \boldsymbol{x}^0$. (We assume that this solution is unique and is defined for all $t \geq 0$, for each input that is admissible in the sense of e.g. [9], as is the case with our model.)

The coordinates of the state $\boldsymbol{x}$ in our case are x_1, x_2, x_3, x_4, x_5, and they are always non-negative. In particular, we denote by x_1^0, x_2^0, x_3^0, x_4^0, x_5^0 the coordinates of the initial state $\boldsymbol{x}^0$. Moreover, we assume that $x_4^0 = x_5^0$. (We could assume, instead, that $x_3^0 = x_5^0$. Note that Table 1 gives the values that apply to our model, in particular $x_3^0 = x_4^0 = x_5^0 = 1$.)

Lemma: Consider the model system, and a fixed initial state $\boldsymbol{x}^0$.

Pick any six scalar non-zero real-analytic (for example, constant) inputs U, $\overline{U}$, V, $\overline{V}$, W, $\overline{W}$, such that $U \neq \overline{U}$, $V \neq \overline{V}$ and $W(0) \neq \overline{W}(0)$, and consider the following four vector inputs

$$\boldsymbol{u}^1 = (u_1^1, u_2^1, u_3^1) = (U, V, W)$$

$$\boldsymbol{u}^2 = (u_1^2, u_2^2, u_3^2) = (U, V, \overline{W})$$

$$\boldsymbol{u}^3 = (u_1^3, u_2^3, u_3^3) = (\overline{U}, \overline{V}, W)$$

$$\boldsymbol{u}^4 = (u_1^4, u_2^4, u_3^4) = (\overline{U}, \overline{V}, \overline{W})$$

Suppose that $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ are two parameter vectors with the following property

$$F(\boldsymbol{x}^0, \boldsymbol{u}^i, \boldsymbol{\theta}) = F(\boldsymbol{x}^0, \boldsymbol{u}^i, \tilde{\boldsymbol{\theta}}), \quad i = 1, 2, 3, 4$$

then, $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Note that writing equality $v = \boldsymbol{w}$, for two functions of time, means that $v(t) = \boldsymbol{w}(t)$ for all t (we sometimes write $v \equiv \boldsymbol{w}$ if in order to emphasise that the functions are identical). Thus, for example, in the lemma statement, U being non-zero means that $U(t) \neq 0$ for some t, and $U \neq \overline{U}$ means that $U(t) \neq \overline{U}(t)$ for some t.

The lemma says that the mapping from parameters to possible observations is one-to-one, or in other words, that the parameters are, at least theoretically, reconstructible from the observations. (After the proof, we remark that powerful theorems in control theory imply that, then, a single 'generic' input time function suffices.)

Proof: We prove the lemma through several steps.

Step 1. We first consider the outputs that result from applying the input $\boldsymbol{u} = \boldsymbol{u}^1$. As $F(\boldsymbol{x}^0, \boldsymbol{u}^1, \boldsymbol{\theta}) = F(\boldsymbol{x}^0, \boldsymbol{u}^1, \tilde{\boldsymbol{\theta}})$, and the coordinates x_1 and x_2 are part of the output, in particular we have that $x_1(t) = \tilde{x}_1(t)$ for all $t \geq 0$, where we denote by $\boldsymbol{x}(t)$ (resp. $\tilde{\boldsymbol{x}}(t)$) the solution when the parameter vector is $\boldsymbol{\theta}$ (resp. $\tilde{\boldsymbol{\theta}}$). Therefore it also holds that

$$\dot{x}_1(t) - D(\boldsymbol{u}^1(t) - x_1(t)) = \dot{\tilde{x}}_1(t) - D(\boldsymbol{u}^1(t) - \tilde{x}_1(t)) \quad (21)$$

for all $t \geq 0$. Let us introduce the functions

$$\varphi(t) = \alpha_4 x_4(t) \frac{x_1(t)}{\alpha_1 + x_1(t)} \text{ and } \tilde{\varphi} = \alpha_4 \tilde{x}_4(t) \frac{\tilde{x}_1(t)}{\alpha_1 + \tilde{x}_1(t)}$$
$$= \alpha_4 \tilde{x}_4(t) \frac{x_1(t)}{\alpha_1 + x_1(t)}$$

Then, from the form of the differential equation for x_1 and using (21)

$$\varphi(t) = \tilde{\varphi}(t)$$

for all $t \geq 0$.

Step 2. Next, we consider the output that results from applying $\boldsymbol{u} = \boldsymbol{u}^3$. Let us denote by $\psi(t)$ and $\tilde{\psi}(t)$ the functions $\alpha_4 x_4(t)(z_1(t))/(\alpha_1 + z_1(t))$, and analogously for $\tilde{z}$, that result from the solutions z and $\tilde{z}$ of the model system when using this new input. It is important to observe that we have the same coordinates $x_4(t)$ as earlier, because both $\boldsymbol{u}^1$ and $\boldsymbol{u}^3$ have the same third coordinate, and x_4 is not affected by the first two input coordinates. By an argument as earlier, $\psi \equiv \tilde{\psi}$.

Claim: For a generic time $t = \tau$, the following four properties hold

1. $x_1(\tau) \neq 0$
2. $z_1(\tau) \neq 0$
3. $x_1(\tau) \neq z_1(\tau)$
4. $x_4(\tau) \neq 0$

(By 'generic' in this claim we mean 'except at most for a countable subset of $[0, \infty)$.')

Proof of Claim: Let S be the set of times t such that $x_1(t) = 0$, R the set of times t such that $z_1(t) = 0$, T the set of times t such that $x_4(t) = 0$, and Δ the set of times t such that $x_1(t) = z_1(t)$. The solutions of our differential equations are real-analytic functions of time ([9], Proposition C.3.12). So $x_1(t)$ is an analytic function of t, and therefore one of two cases must happen: either $x_1 \equiv 0$ or S is a discrete (countable, possibly finite or empty) set. If x_1 would vanish identically (first case), then the equation

$$0 \equiv \dot{x}_1 \equiv D(u_1^1 - x_1) - \alpha_4 x_4 \frac{x_1}{u_1^1 + x_1} \equiv Du_1^1$$

would imply that $u_1^1 \equiv 0$, which is a contradiction, because we assumed that U is non-zero. Thus S is a discrete set. Similarly, the set R is discrete. We claim that T is discrete too: if this were not the case, then $x_4(t) \equiv 0$, which contradicts the assumption that $x_4^0 \neq 0$. Finally, regarding Δ, suppose by way of contradiction that, instead, $x_1 \equiv z_1$. Then also $\dot{x}_1 \equiv \dot{z}_1$, which means that

$$D(u_1^1 - x_1) - \alpha_4 x_4 \frac{x_1}{\alpha_1 + x_1} \equiv D(u_1^3 - x_1) - \alpha_4 x_4 \frac{x_1}{\alpha_1 + x_1}$$

(recall that the same $x_4(t)$ appears in both terms). Therefore $u_1^1 \equiv u_1^3$, which contradicts $U \neq \overline{U}$. It follows that the union $S \cup \overline{S} \cup T \cup \Delta$ is discrete (a union of discrete sets is discrete), and therefore a generic τ has all the required properties, and the claim is proved. □

An analogous claim holds, clearly, for the parameter set $\tilde{\boldsymbol{\theta}}$. We now consider any generic point τ, so that the properties of the claim hold both for the system with parameters $\boldsymbol{\theta}$

and with parameters $\tilde{\boldsymbol{\theta}}$. We denote

$$A := \varphi(\tau) = \frac{qx_1(\tau)}{\alpha_1 + x_1(\tau)}$$

and

$$B := \psi(\tau) = \frac{qz_1(\tau)}{\alpha_1 + z_1(\tau)}$$

where we write for simplicity $q := \alpha_4 x_4(\tau) \neq 0$. Note that, as $\varphi(\tau) = \tilde{\varphi}(\tau)$ and $\psi(\tau) = \tilde{\psi}(\tau)$ we also have that

$$A = \tilde{\varphi}(\tau) = \frac{\tilde{q}x_1(\tau)}{\tilde{\alpha}_1 + x_1(\tau)}$$

and

$$B = \tilde{\psi}(\tau) = \frac{\tilde{q}z_1(\tau)}{\tilde{\alpha}_1 + z_1(\tau)}$$

where we write $\tilde{q} := \tilde{\alpha}_4\tilde{x}_4(\tau) \neq 0$ and we used that $x_1(t) = \tilde{x}_1(t)$ for all $t \geq 0$ (as $F(x^0, \boldsymbol{u}, \boldsymbol{\theta}) = F(x^0, \boldsymbol{u}, \tilde{\boldsymbol{\theta}})$, for each of the two inputs being considered), and therefore $x_1(\tau) = \tilde{x}_1(\tau)$, and similarly $z_1(\tau) = \tilde{z}_1(\tau)$.

We will next show that $\alpha_1 = \tilde{\alpha}_1$ and $q = \tilde{q}$.

We first remark that $Az_1(\tau) \neq x_1(\tau)B$. Indeed, if this is not the case, then the definitions of A and B would give us that

$$\frac{qx_1(\tau)z_1(\tau)}{\alpha_1 + x_1(\tau)} = \frac{qx_1(\tau)z_1(\tau)}{\alpha_1 + z_1(\tau)}$$

and therefore as $qx_1(\tau)z_1(\tau) \neq 0$ (recall our choice of τ), it would follow that $x_1 = z_1\ (\tau)$, which is a contradiction with the choice of τ.

Notice that

$$\begin{pmatrix} x_1(\tau) & -A \\ z_1(\tau) & -B \end{pmatrix}\begin{pmatrix} q \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} x_1(\tau) & -A \\ z_1(\tau) & -B \end{pmatrix}\begin{pmatrix} \tilde{q} \\ \tilde{\alpha}_1 \end{pmatrix} = \begin{pmatrix} Ax_1(\tau) \\ Bz_1(\tau) \end{pmatrix}$$

We showed that the determinant $Az_1(\tau) - x_1(\tau)B$ of the above matrix is non-zero, so it follows that $(\alpha_1, q) = (\tilde{\alpha}_1, \tilde{q})$.

The functions $\alpha_4 x_4(t)$ and $\tilde{\alpha}_4\tilde{x}_4(t)$ are both continuous, and we know that $\alpha_4 x_4(t) = \tilde{\alpha}_4\tilde{x}_4(t)$ for generic τ; it follows that $\alpha_4 x_4(t) = \tilde{\alpha}_4\tilde{x}_4(t)$ for all t. In particular

$$\alpha_4 x_4^0 = \alpha_4 x_4(0) = \tilde{\alpha}_4\tilde{x}_4(t) = \tilde{\alpha}_4 x_4^0$$

and therefore as $x_4^0 \neq 0$, $\alpha_4 = \tilde{\alpha}_4$. Also

$$x_4(t) = \tilde{x}_4(t) \quad \text{for all } t \geq 0$$

An argument entirely analogous to this, but considering the output y_2 instead of y_1, shows that $\alpha_2 = \tilde{\alpha}_2$ and $\alpha_3 = \tilde{\alpha}_3$, as well as

$$x_3(t) = \tilde{x}_3(t) \quad \text{for all } t \geq 0$$

We have shown the identifiability of α_i, $i = 1, 2, 3, 4$ from the outputs corresponding to the two inputs u^1 and u^3. We also showed that $x_3(t) = \tilde{x}_3(t)$ and $x_4(t) = \tilde{x}_4(t)$, provided that outputs coincide when u^1 and u^3 are applied to our model system. Therefore also $\dot{x}_4 \equiv \dot{\tilde{x}}_4$. Using $x_4^0 = x_5^0$, we have that $\alpha_5(x_4^0 - x_5^0) = 0$, and

$$\beta\frac{c}{\gamma + c} = \tilde{\beta}\frac{c}{\tilde{\gamma} + c} \tag{22}$$

where $c = u_3^1(0) = W(0)$ and $\bar{c} = u_3^2(0) = \overline{W}(0)$, which follows from $\dot{x}_4(0)/x_4(0) = \dot{\tilde{x}}_4(0)/\tilde{x}_4(0)$.

An entirely analogous argument, using the pair of inputs u^2 and u^4 instead of u^1 and u^3, gives

$$\beta\frac{\bar{c}}{\gamma + \bar{c}} = \tilde{\beta}\frac{\bar{c}}{\tilde{\gamma} + \bar{c}} \tag{23}$$

Dividing (22) by (23), we conclude that

$$\frac{\gamma + \bar{c}}{\gamma + c} = \frac{\tilde{\gamma} + \bar{c}}{\tilde{\gamma} + c}$$

and cross-multiplying and simplifying, we obtain $(\bar{c} - c)\gamma = (\bar{c} - c)\tilde{\gamma}$ and hence $\gamma = \tilde{\gamma}$, as $c \neq \bar{c}$. Using this last equality in (22), we conclude that also $\beta = \tilde{\beta}$. We are only left with showing the identifiability of α_5.

Observe that, as x_5^0 is fixed, and $\beta = \tilde{\beta}$ and $\gamma = \tilde{\gamma}$, then with the input $u = u^1$ (or with any other input, for that matter), the fifth coordinate of the solutions with parameter vectors $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ coincide: $x_5 \equiv \tilde{x}_5$. As we had proved that $x_4 \equiv \tilde{x}_4$, and so also $\dot{x}_4 \equiv \dot{\tilde{x}}_4$, we have, then, using $u = u^1$ and dropping the superscript

$$\alpha_5(x_5 - x_4) - \beta x_4\frac{u_3}{\gamma + u_3} \equiv \tilde{\alpha}_5(x_5 - x_4) - \beta x_4\frac{u_3}{\gamma + u_3}$$

and therefore

$$\alpha_5(x_5(t) - x_4(t)) = \tilde{\alpha}_5(x_5(t) - x_4(t)) \quad \text{for all } t \geq 0$$

We then conclude $\alpha_5 = \tilde{\alpha}_5$, unless it is the case that $x_5 \equiv x_4$. But this latter identity cannot hold, because it would imply

$$0 \equiv \dot{x}_4 - \dot{x}_5 \equiv \alpha_5(x_5 - x_4) - \beta x_4\frac{u_3}{\gamma + u_3}$$
$$- \beta(1 - x_5) - \beta x_5\frac{u_3}{\gamma + u_3} \equiv \beta(1 - x_5)$$

and therefore $x_5 \equiv 1$, which in turn would give, using the equation for $\dot{x}_5$ and the fact that $\dot{x}_5 \equiv 0$, that $\beta(u_3/\gamma + u_3) \equiv 0$, which is a contradiction as $u_3 = W \neq 0$. In summary, we showed that $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. □

Corollary: Consider the model system, and a fixed initial state $\boldsymbol{x}^0$. Then, for a generic smooth input u, the following property holds: if two parameter vectors $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ are such that

$$F(\boldsymbol{x}^0, u, \boldsymbol{\theta}) = F(\boldsymbol{x}^0, u, \tilde{\boldsymbol{\theta}})$$

then $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

The precise interpretation of the term generic in this corollary is as genericity with respect to the Whitney topology, as discussed in the citations given subsequently. For purposes of this paper, however, it is enough to think of ‘generic’ inputs as ‘random’ inputs: except for very special inputs, the identifiability condition holds. (As an example, consider the following two-dimensional system: $\dot{x}_1 = u$, $\dot{x}_2 = x_1$, with output $\boldsymbol{y} = \boldsymbol{\theta}(x_1 - x_2)$, where $\boldsymbol{\theta}$ is an unknown parameter and initial state $x = 0$. Clearly, if the input u is known, and thus $x_1(t) - x_2(t)$ is also known, the parameter $\boldsymbol{\theta}$ can be immediately obtained from $\boldsymbol{y}(t)$, unless it so happens that $x_1(t) = x_2(t)$ for all t. But $x_1 = x_2$ implies that $\dot{u} = \ddot{x}_1 = \ddot{x}_2 = u$, that is, $u = ce^t$. So, except for those very special inputs that have the form $u = ce^t$, every other input serves to identify parameters.)

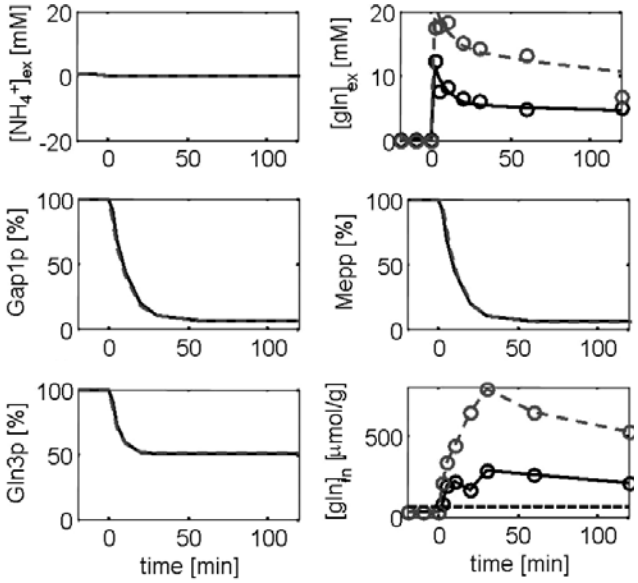

Fig. 8 *Model description of a 10 mM (solid black) and 20 mM (dashed grey) glutamine pulse to wild-type strain VWk43*

Experimental data (circles) have been included as the dependent input (intracellular glutamine) and to verify the simulated output profiles (extracellular glutamine)

Proof: We appeal to the universal input theorem for distinguishability, which says that generic smooth inputs are capable of separating any two distinguishable states in a system. More precisely, one applies the theorem to a parametric identification problem by viewing parameters as constant states, as described in the work of Sontag [10, pp.148]. The universal input theorem is one of the key results in control theory, and was proved first for bilinear systems in the work of Grasselli and Isidori [11], for polynomial as well as analytic nonlinear systems (restricted to compact subsets) [10] and in general form [12]; the work of Wang and Sontag [13] gave a relatively simple proof, and applications to controllability and other problems are described by Sontag [14].

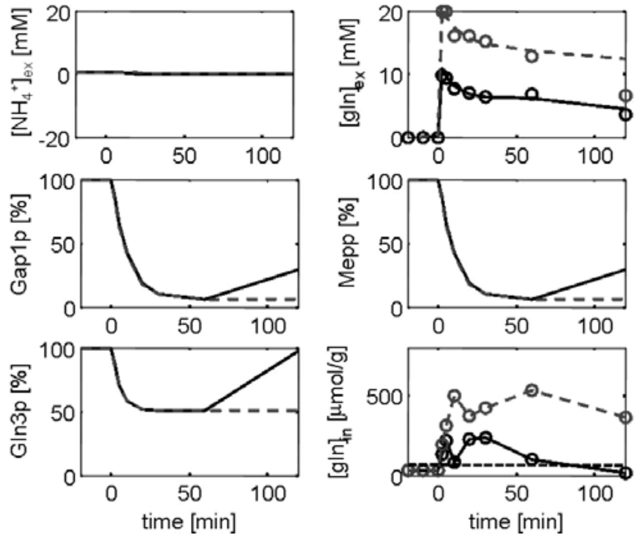

Fig. 9 *Model description of a 10 mM (solid black) and 20 mM (dashed grey) glutamine pulse to mutant strain Δglt1*

Experimental data (circles) have been included as the dependent input (intracellular glutamine) and to verify the simulated output profiles (extracellular glutamine)

Thus, as the lemma proves that any pair of parameters can be distinguished by some input/output experiments, the universal input theorem guarantees that generic inputs will suffice for this task. □

The corollary supports the use of the 'internal input' approach, because it asserts that generic inputs, such as those arising from measured data generated by unmodelled dynamics, are sufficient for identifiability. Although it is theoretically possible that the input $\boldsymbol{v}$ that appears in this fashion will happen to be one of the 'exceptional inputs' that appear non-generically, this is unlikely. A somewhat more serious gap between theory and practice is in the use of 'pulse' inputs in our experiments. In general, there is no theoretical guarantee that such inputs will be enough for identification. However, pulses may be approximated closely in an arbitrary manner by generic inputs, and, in

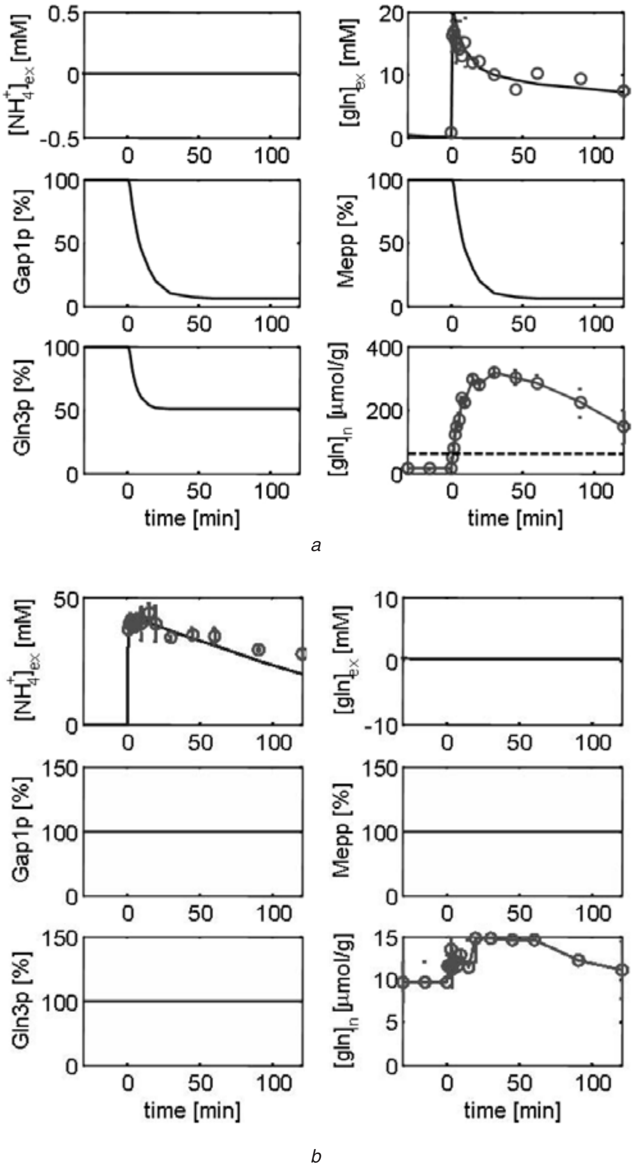

Fig. 10 *Model description*

a an 18 mM glutamine pulse

b 40 mM ammonia pulse to mutant strain Δgln1

Experimental data (circles) have been included as the dependent input (intracellular glutamine) and to verify the simulated output profiles (extracellular ammonia and extracellular glutamine).

Horizontal dashed line indicates estimated NCR threshold level of intracellular glutamine (gln$_T$)

any case, one may argue that, in practice, the applied inputs are never exactly pulses.

An interpretation of the conclusion of the corollary is as follows. Suppose that we pick a generic input $\boldsymbol{u}$, and we use it as an input to the 'true' system, measuring the outputs $\boldsymbol{y} = (y_1, y_2)$, which we write as $\boldsymbol{z}(t)$. For any parameter vector $\boldsymbol{\theta}$, we write $\boldsymbol{y}(\cdot, \boldsymbol{\theta}) = F(x^0, \boldsymbol{u}, \boldsymbol{\theta})$ and $\boldsymbol{e}(t, \boldsymbol{\theta}) = \boldsymbol{z}(t) - \boldsymbol{y}(t, \boldsymbol{\theta})$. Then, we set up the quadratic criterion

$$J(\boldsymbol{\theta}) = \int_0^T \boldsymbol{e}(t, \boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{W} \boldsymbol{e}(t, \boldsymbol{\theta})\, \mathrm{d}t$$

where T is time duration of the input and $\boldsymbol{W}$ is a positive definite symmetric weight matrix.

Provided that the true system is a model system for some (unknown) set of parameters $\boldsymbol{\theta}^0$, and that there is no noise in observations, then $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0$, and the minimum is achieved uniquely, at the true parameter set $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. This provides a theoretical justification for the use of the maximum likelihood approach, at least when there is no model mismatch and noise is small.

We do not provide details here, but it is also possible to prove that sampling at generic times will suffice for identification in this same theoretical sense; in other words, the error criterion could be stated for a sum over a certain number of samples instead of as an integral over the entire non-negative real axis, and the uniqueness result holds. Refer to the work of Sontag [15], and in particular the example provided in the introduction to that paper.

9.2 Validation

For a quantitative validation, the identified model was used to predict the uptake profiles of glutamine or ammonia in six different experiments. In Fig. 8 the results of a 10 mM and 20 mM glutamine pulse to wild-type strain VWk43 are shown. The results of the 20 mM glutamine pulse are comparable with those of the 18 mM glutamine pulse in the identification data (Fig. 6), despite the intracellular glutamine profile reaches significantly higher levels. Also in wild-type VWk43, NCR is rapidly triggered after both pulses and the uptake remained largely repressed during 2 h after the pulses. When the same pulses were applied to the Δ*glt1* mutant, the results were somewhat different (Fig. 9). The model predicted that NCR was released approximately 1 h after the 10 mM glutamine pulse, resulting in an increased glutamine uptake rate. The model-predicted extracellular glutamine concentration at 120 min matched the corresponding data point. Finally, in Fig. 10 the results of a 18 mM glutamine and 40 mM ammonia pulse to the Δ*gln1* mutant are shown. The residuals of the glutamine and ammonia uptake profiles have been shown in Fig. 7.