Indistinguishability and identifiability of kinetic models for the MurC reaction in peptidoglycan biosynthesis

J.G. Hattersley[a], J. Pérez-Velázquez[a], M.J. Chappell[a], D. Bearup[b], D. Roper[b], C. Dowson[b], T. Bugg[c], N.D. Evans[*,a]

[a] *School of Engineering, University of Warwick, Coventry, CV4 7AL, UK (e-mail: Neil.Evans@warwick.ac.uk).*
[b] *School of Biological Sciences, University of Warwick, Coventry, UK*
[c] *School of Chemistry, University of Warwick, Coventry, UK*

Abstract

An important question in Systems Biology is the design of experiments that enable discrimination between two (or more) competing chemical pathway models or biological mechanisms. In this paper analysis is performed between two different models describing the kinetic mechanism of a three-substrate three-product reaction, namely the MurC reaction in the cytoplasmic phase of peptidoglycan biosynthesis. One model involves ordered substrate binding and ordered release of the three products; the competing model also assumes ordered substrate binding, but with fast release of the three products. The two versions are shown to be distinguishable; however, if standard quasi steady-state assumptions are made distinguishability can not be determined. Once model structure uniqueness is ensured the experimenter must determine if it is possible to successfully recover rate constant values given the experiment observations, a process known as structural identifiability. Structural identifiability analysis is carried out for both models to determine which of the unknown reaction parameters can be determined uniquely, or otherwise, from the ideal system outputs. This structural analysis forms an integrated step towards the modelling of the full pathway of the cytoplasmic phase of peptidoglycan biosynthesis.

Key words: Indistinguishability, Identifiability, Experiment design, MurC, Biomedical systems, Parameter identification

*Corresponding author

1. Introduction

In the biological sciences it is becoming increasingly common to collect data in high-throughput experiments on several scales: genomic, proteomic, or metabolic (Snoep and Westerhoff [30], Sauer et al. [27]). These data hold the promise of identifying the mechanisms of interactions that comprise large-scale regulatory biochemical networks. An important step in this work is the design of experiments to allow discrimination between two (or more) competing pathway models or biological mechanisms. Structural indistinguishability provides a formal approach to distinguish between competing model mechanisms (Kholodenko et al. [19]).

In Systems Biology mathematical models that are generated invariably include large numbers of state variables with numerous model parameters, many of which are unknown, or cannot be directly measured. With such highly complex systems there are often few direct measurements that can be made and limited access for input perturbation to elucidate system dynamics. These limitations cause problems when investigating the existence of hidden pathways or attempting to estimate unknown parameters. Identifiability analysis provides a formal approach to determine what additional inputs and/or measurements are necessary in order to reduce, or remove, these limitations and permit the derivation of models that can be used for practical purposes with greater confidence (Snoep and Westerhoff [30], Sauer et al. [27], Kholodenko et al. [19]).

Structural indistinguishability for system models is concerned with determining the uniqueness between possible candidates for the model (or mechanism) structure (Evans et al. [13]). The analysis is concerned with whether the underlying possibilities for the parameterised mathematical model can be distinguished using the inputs (perturbations or interventions) and observations (or measurements) available for the biological system under investigation.

In chemical kinetics it is key to characterise reaction mechanisms, however there are often several different process models that are consistent with the available data. These mechanisms may be described by the same mathematical representation (see Érdi and Tóth [9], Espenson [10]) but without formal analysis of the mathematical model a reaction mechanism's validity is only disproved by showing inconsistency with available data. Whilst this problem has been recognised [5, 6], structural indistinguishability is not routinely applied to chemical kinetics experiments and model

development. For linear systems structural indistinguishability analysis is generally exhaustive with all competing mechanisms generated from a given one (see Godfrey and Chapman [16]). For nonlinear systems, approaches are generally only for pairs of candidate models, though in some cases a parameterised family of such candidates can be generated. There has been limited application of structural indistinguishability analysis to the chemical kinetic models. In Schnell et al. [28] the issues of distinguishability with respect to biochemical kinetics was considered via application of structural indistinguishability to classical models for a single-enzyme, single-substrate reaction. In addition, simple kinetic models are incorporated into studies of structural analysis methods, often with Michaelis-Menten type reparameterisation (see Saccomani et al. [25], Bellu et al. [3]; however, to the authors' knowledge this is the first time indistinguishability and identifiability have been applied to a full (three-substrate/three-product) enzyme kinetic model.

Structural identifiability arises in the inverse problem of inferring from the known, or assumed, properties of a biological system, estimates for the corresponding rate constants and other parameters; as such it can be considered as a special case of the structural indistinguishability problem. Structural identifiability analysis considers the uniqueness of the unknown model parameters from the input-output structure corresponding to proposed experiments to collect data for parameter estimation. This is an important, but often overlooked, prerequisite to experiment design, system identification and parameter estimation, since estimates for structurally unidentifiable parameters are effectively meaningless. If parameter estimates are to be used to inform intervention or inhibition strategies, or other critical decisions, then it is essential that the parameters be uniquely identifiable. Numerous techniques for performing a structural identifiability analysis on linear parametric models exist (see Godfrey and DiStefano III [17], Walter [34]). In comparison, there are relatively few techniques available for nonlinear systems such as the Taylor series approach (Pohjanpalo [23]), similarity transformation based approaches (Tunali and Tarn [31], Vajda et al. [32], Evans et al. [12]) and differential algebra techniques (Ljung and Glad [20], Saccomani, Audoly, and D'Angio [26]). Unfortunately for systems with a complex structure significant computational problems can arise even for relatively low dimensional models. At present there has been relatively little work on techniques for large-scale, highly complex systems, which are typical in Systems Biology. As shall be

shown, the analytic identifiability approaches can generate computationally intractable solutions; therefore, an alternative numerical approach that uses the sensitivity of the observation functions to changes in the parameters is implemented to suggest the local identifiability of the parameters and associated model (Fisher [15], Jacquez [18]).

The purpose of this paper is to explore the possible effectiveness of using structural indistinguishability and identifiability techniques in model discrimination within Systems Biology networks, using MurC as a case study; to this end, a structural indistinguishability analysis is performed between two different models describing the kinetic mechanism of the MurC (Ter-ter) reaction in the cytoplasmic phase of bacterial peptidoglycan biosynthesis (BPB). One model assumes the kinetic mechanism proposed by Emanuele et al. [8], which involves step-by-step release of the three products; the competing model assumes that the release of the products is simultaneous. A structural identifiability analysis is also carried out for both models to ensure that the model output uniquely determines the unknown parameters.

2. The Models

Prior to describing the mathematical models of the Ter-ter enzyme reaction, it is necessary to review the basic biology and context of the MurC reaction. The cell wall of many bacteria is composed of peptidoglycan, which is made up of a combination of peptide bonds and carbohydrates. Peptidoglycan serves a structural role in the bacterial cell wall, giving rigidity, as well as counteracting the osmotic pressure of the cytoplasm (El Zoeiby et al. [7]). The BPB pathway (shown in Fig. 1) is a significant target in the development of antibacterial agents (Walsh [33]). A detailed understanding of the biosynthesis pathway is essential for the development of new strategies for antibacterial action to compensate for the emergence of clinical resistance to penicillin antibiotics in *S. aureus, S. pneuminiae* and Gram-negative pathogens such as *P. aeruginosa* (Marmor et al. [21], Anderson et al. [1]), and the emergence of vancomycin resistance in Enterococci, together with the lack of new classes of antibacterial agent. Although this pathway is quite well known, characterisation, especially of the later lipid linked steps, has been hampered by difficulties in making the natural substrates and there remain some uncertainties within the wider reaction network. As such

this pathway is ideal for a feasibility study of the effectiveness of structural indistinguishability and identifiability in mechanism discrimination, model formulation and experiment design (Snoep and Westerhoff [30], Sauer et al. [27]).

Within the biosynthetic pathway for bacterial peptidoglycan there is already a reasonable (but not complete) understanding of the cytoplasmic pathways. Whilst a basic pathway scheme is recognised for this region, there are still a number of steps within the pathway where competing reaction schemes may exist (e.g., multiple isoforms of an enzyme within the same cell) and also where feedback inhibition may play an important (but not yet fully understood) role. The study of the cytoplasmic phase of the biosynthetic process comprises a comprehensive understanding of the MurA to MurF reactions of the full pathway (see Fig. 1). In this paper the focus is on MurC.

The MurC reaction to be modelled is a three-substrates, three-products enzyme catalysed reaction. It is assumed that the kinetic mechanism is as proposed by Emanuele et al. [8], see Fig. 2, which involves step-by-step release of the three products, with a competing model that assumes that the release of the products is simultaneous. Note however that the mathematical model does not apply to MurC alone, but to any general three-substrates three-products ordered mechanism (potentially MurD, E, and F, if they are ordered).

2.1. Model 1: Ordered release of products

From Emanuele et al. [8] it is known that the third substrate (L-Ala) binds before the reaction starts. Thus the reaction can be written as the following set of chemical steps:

$$
\begin{aligned}
E + S_1 &\underset{r_1}{\overset{k_1}{\rightleftharpoons}} C_1 \\
C_1 + S_2 &\underset{r_2}{\overset{k_2}{\rightleftharpoons}} C_2 \\
C_2 + S_3 &\underset{r_3}{\overset{k_3}{\rightleftharpoons}} C_3 \underset{r_4}{\overset{k_4}{\rightleftharpoons}} C_4 \underset{r_5}{\overset{k_5}{\rightleftharpoons}} C_5 + P \\
C_5 &\underset{r_6}{\overset{k_6}{\rightleftharpoons}} C_6 + Q \\
C_6 &\underset{r_7}{\overset{k_7}{\rightleftharpoons}} E + R
\end{aligned}
$$

where:

1. Enzyme (E), E
2. Substrate 1 (ATP), S_1
3. Substrate 2 (UNAM), S_2
4. Substrate 3 (L-Ala), S_3
5. Complex E-ATP, C_1
6. Binary Complex E-ATP-UNAM, C_2
7. Ternary Complex E-ATP-UNAM-L-Ala, C_3
8. Complex E-ADP-UNAMA-Pi, C_4
9. There are three products: ADP (R), UNAMA and Pi

Note that the order of the release of the products is not determined in this scheme, it is only known that ADP releases at the end, the release order of the other two products (UNAMA and Pi) is not known and therefore they have been generically called P and Q. The complexes ERP and ER are C_5 and C_6, respectively.

2.2. Model 2: Fast release of products

The second model assumes that the release of the products is fast (instantaneous and simultaneous). The corresponding set of chemical steps is as follows:

$$
\begin{aligned}
E + S_1 &\underset{r_1}{\overset{k_1}{\rightleftharpoons}} C_1 \\
C_1 + S_2 &\underset{r_2}{\overset{k_2}{\rightleftharpoons}} C_2 \\
C_2 + S_3 &\underset{r_3}{\overset{k_3}{\rightleftharpoons}} C_3 \underset{r_4}{\overset{k_4}{\rightleftharpoons}} C_4 \underset{r_5}{\overset{k_5}{\rightleftharpoons}} E + P + Q + R
\end{aligned}
$$

2.2.1. Equations

The time evolution of the reaction is obtained by applying the Law of Mass Action to yield a set of 13 coupled nonlinear ordinary differential equations with 14 parameters for Model 1 and 11 coupled nonlinear differential equations with 10 parameters for Model 2. The system equations for both models are defined below.

These are the core equations for both models

$$
\begin{aligned}
S_1'(t) &= -k_1 E(t) S_1(t) + r_1 C_1(t), \\
C_1'(t) &= k_1 E(t) S_1(t) - r_1 C_1(t) - k_2 C_1(t) S_2(t) + r_2 C_2(t), \\
S_2'(t) &= -k_2 C_1(t) S_2(t) + r_2 C_2(t), \\
C_2'(t) &= k_2 C_1(t) S_2(t) - r_2 C_2(t) - k_3 C_2(t) S_3(t) + r_3 C_3(t), \\
S_3'(t) &= -k_3 C_2(t) S_3(t) + r_3 C_3(t) \\
C_3'(t) &= k_3 C_2(t) S_3(t) - r_3 C_3(t) - k_4 C_3(t) + r_4 C_4(t)
\end{aligned}
$$

Model 1 core equations plus:

$$
\begin{aligned}
E'(t) &= -k_1 E(t) S_1(t) + r_1 C_1(t) + k_7 C_6(t) - r_7 E(t) R(t), \\
C_4'(t) &= k_4 C_3(t) - r_4 C_4(t) - k_5 C_4(t) + r_5 C_5(t) P(t), \\
C_5'(t) &= k_5 C_4(t) - r_5 C_5(t) P(t) - k_6 C_5(t) + r_6 C_6(t) Q(t), \\
C_6'(t) &= k_6 C_5(t) - r_6 C_6(t) Q(t) - k_7 C_6(t) + r_7 E(t) R(t), \\
P'(t) &= k_5 C_4(t) - r_5 C_5(t) P(t), \\
Q'(t) &= k_6 C_5(t) - r_6 C_6(t) Q(t), \\
R'(t) &= k_7 C_6(t) - r_7 E(t) R(t).
\end{aligned}
$$

Model 2 core equations plus:

$$
\begin{aligned}
E'(t) &= -k_1 E S_1(t) + r_1 C_1(t) + k_5 C_4(t) - r_5 EPQR(t), \\
C_4(t)' &= k_4 C_3(t) - r_4 C_4(t) - k_5 C_4(t) + r_5 EPQR(t), \\
P'(t) &= k_5 C_4(t) - r_5 EPQR(t), \\
Q'(t) &= k_5 C_4(t) - r_5 EPQR(t), \\
R'(t) &= k_5 C_4(t) - r_5 EPQR(t).
\end{aligned}
$$

2.2.2. Experimental Observations

In biochemical progress curve experiments, concentrations are measured by the absorbance of light at one or more wavelengths (see Fersht [14], Nölting [22]). If the molar absorptivities of the measured components at these wavelengths are known, then biochemists use Beer's law to determine the concentrations.

Thus, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{\epsilon}\,\boldsymbol{x} \tag{1}$$

where $\boldsymbol{y}$ is the vector of observed absorbances at different wavelengths, $\boldsymbol{x}$ is the vector of concentrations of species absorbing at these wavelengths, and $\boldsymbol{\epsilon}$ is a matrix of molar absorptivities. In these experiments, the kinetic parameters are determined from expressions for the species concentrations as a function of time. The concentration of the substrate or product is recorded in time after the initial fast transient and for a sufficiently long period to allow the reaction to approach equilibrium. While they are less common now, progress curve experiments were widely used in the early period of enzyme kinetics when Victor Henri was active in the field (early 20th Century). It is assumed that progress curves are available and therefore can be related to the models presented in the previous section.

3. Methods

3.1. Structural Indistinguishability

Since an indistinguishability analysis can be seen as a generalisation of the identifiability problem, it can be studied by modifying existing approaches for identifiability. Here a modification of the Taylor series approach for identifiability is used.

Consider two uncontrolled systems of the form:

$$\Sigma(\boldsymbol{p}) \begin{cases} \dot{\boldsymbol{x}}(t,\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{x}(t,\boldsymbol{p}),\boldsymbol{p}), & \boldsymbol{x}(0,\boldsymbol{p}) = \boldsymbol{x}_0(\boldsymbol{p}) \\ \boldsymbol{y}(t,\boldsymbol{p}) = \boldsymbol{h}(\boldsymbol{x}(t,\boldsymbol{p}),\boldsymbol{p}) \end{cases} \tag{2}$$

$$\tilde{\Sigma}(\tilde{\boldsymbol{p}}) \begin{cases} \dot{\tilde{\boldsymbol{x}}}(t,\tilde{\boldsymbol{p}}) = \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}(t,\tilde{\boldsymbol{p}}),\tilde{\boldsymbol{p}}), & \tilde{\boldsymbol{x}}(0,\tilde{\boldsymbol{p}}) = \tilde{\boldsymbol{x}}_0(\tilde{\boldsymbol{p}}) \\ \tilde{\boldsymbol{y}}(t,\tilde{\boldsymbol{p}}) = \tilde{\boldsymbol{h}}(\tilde{\boldsymbol{x}}(t,\tilde{\boldsymbol{p}}),\tilde{\boldsymbol{p}}) \end{cases} \tag{3}$$

where $p \in \Omega \subseteq \mathbb{R}^q$ and $\tilde{\boldsymbol{p}} \in \tilde{\Omega} \subseteq \mathbb{R}^{\tilde{q}}$, both open subsets consisting of the admissible parameter vectors for the two systems respectively; $\boldsymbol{f}(\cdot,\boldsymbol{p})$ and $\boldsymbol{h}(\cdot,\boldsymbol{p})$ are analytic on $M(\boldsymbol{p})$, an open and connected subset of $\mathbb{R}^n$ such that $\boldsymbol{x}_0(\boldsymbol{p}) \in M(\boldsymbol{p})$; $\tilde{\boldsymbol{f}}(\cdot,\boldsymbol{p})$ and $\tilde{\boldsymbol{h}}(\cdot,\boldsymbol{p})$ are analytic on $M(\tilde{\boldsymbol{p}})$, an open and connected subset of $\mathbb{R}^{\tilde{n}}$ such that $\tilde{\boldsymbol{x}}_0(\tilde{p}) \in M(\tilde{\boldsymbol{p}})$; $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$ are constant parameter vectors; Ω and $\tilde{\Omega}$ are the sets of admissible parameter vectors for the two models (2) and (3), respectively; $\boldsymbol{x}(t,\boldsymbol{p})$ and $\tilde{\boldsymbol{x}}(t,\tilde{\boldsymbol{p}})$ are the state variables for each model, which are the different species concentrations whose values are governed by the system of differential equations comprising the model, (2) and (3), respectively. These kinetics, and hence the solutions $\boldsymbol{x}(t,\boldsymbol{p})$ and $\tilde{\boldsymbol{x}}(t,\tilde{\boldsymbol{p}})$, are dependent on the particular parameter vectors $\boldsymbol{p} \in \Omega$ and $\tilde{\boldsymbol{p}} \in \tilde{\Omega}$ used in the models.

The indistinguishability problem arises because, in general, it is not possible to measure all reactants in a given chemical reaction. An experiment that is used to collect measurements of the process gives rise to an output structure for the model, the resulting output, or measurement vectors are $\boldsymbol{y}(t,\boldsymbol{p}) = (y_1(t,\boldsymbol{p}),\ldots,y_r(t,\boldsymbol{p}))^T$ and $\tilde{\boldsymbol{y}}(t,\tilde{\boldsymbol{p}}) = (y_1(t,\tilde{\boldsymbol{p}}),\ldots,y_r(t,\tilde{\boldsymbol{p}}))^T$, respectively, and it is these vectors that are compared with the collected experimental data during subsequent parameter estimation. Suppose that there exists a $\boldsymbol{p} \in \Omega$ and a $\tilde{\boldsymbol{p}} \in \tilde{\Omega}$ such that $\boldsymbol{y}(t,\boldsymbol{p}) = \tilde{\boldsymbol{y}}(t,\tilde{\boldsymbol{p}})$ for all $t \geq 0$. Then it is not possible to distinguish between the model given by (2) with parameter vector $\boldsymbol{p}$ (i.e., $\Sigma(\boldsymbol{p})$) and the model given by (3) with parameter vector $\tilde{\boldsymbol{p}}$ (i.e., $\tilde{\Sigma}(\tilde{\boldsymbol{p}})$) from their outputs. Therefore, even with perfect data (continuous measurements that are noise-free and error-free) it is not possible to distinguish between the reaction schemes modelled by $\Sigma(\boldsymbol{p})$ and $\tilde{\Sigma}(\tilde{\boldsymbol{p}})$ from the proposed experiment. In this case the models $\Sigma(\boldsymbol{p})$ and $\tilde{\Sigma}(\tilde{\boldsymbol{p}})$ are said to be *indistinguishable*, written as $\Sigma(\boldsymbol{p}) \sim \tilde{\Sigma}(\tilde{\boldsymbol{p}})$. The models are therefore *structurally indistinguishable* ($\Sigma \sim \tilde{\Sigma}$) if for almost all $\boldsymbol{p}$,there exists a $\tilde{\boldsymbol{p}}$ such that ($\Sigma \sim \tilde{\Sigma}$) and for almost all $\tilde{\boldsymbol{p}}$ such that $\tilde{\Sigma}(\tilde{\boldsymbol{p}}) \sim \Sigma(\boldsymbol{p})$.

3.2. Structural Identifiability

Prior to discussing the methods of determining identifiability, it is necessary to formally define identifiability of a single parameter and the relationship between parameter identifiability and model structure identifiability. A review of identifiability techniques can be found in Godfrey and DiStefano III [17]. For a generic parameter vector ($\boldsymbol{p}$) which is a member of the set of all admissible parameter vectors Ω, the individual parameter, p_i, is referred to as *locally identifiable* if there exists a neighbourhood of vectors around $\boldsymbol{p}$, $N(\boldsymbol{p})$, such that, if $\tilde{\boldsymbol{p}} \in N(\boldsymbol{p}) \subseteq \Omega$ and

$$\boldsymbol{y}(t, \boldsymbol{p}) = \boldsymbol{y}(t, \tilde{\boldsymbol{p}}) \; \forall \, t \geq 0, \tag{4}$$

then $p_i = \tilde{p}_i$. If the neighbourhood of the locally defined set equates to the set of all admissible vectors ($N(\boldsymbol{p}) = \Omega$) then the parameter p_i is said to be *globally identifiable*. Finally, if the parameter p_i is not locally identifiable, then it is *unidentifiable* and the experiment that this input-output relation corresponds to cannot be used to distinguish between p_i and $\widetilde{p}_i$.

The identifiability of an individual parameter determines the identifiability of the model structure; the following definitions succinctly describe this relationship [11]:

Definition 1. A model is said to be *structurally globally identifiable* (SGI) if, for generic $p \in \Omega$, all of the parameters p_i (in $\boldsymbol{p}$) are globally identifiable.

Definition 2. (SLI). A model is said to be *structurally locally identifiable* (SLI) if, for generic $p \in \Omega$, all of the parameters p_i (in $\boldsymbol{p}$) are locally identifiable and at least one is not globally identifiable.

Definition 3. A model is said to be *structurally unidentifiable* (SUI) if, for generic $p \in \Omega$, any of the parameters p_i (in $\boldsymbol{p}$) is unidentifiable.

For a more complete definition, please refer to the papers by Chappell et al. [4] and Evans et al. [12]. There are several techniques available to determine the identifiability of nonlinear systems (e.g. differential algebra methods [25] or the Similarity Transformation approach [12]); however, the Taylor series approach of Pohjanpalo [23] can be used for experiments which can produce time series data. The basis of the Taylor series approach is that the components of the output or observation function $y_i(t, \boldsymbol{p})$ and its successive time derivatives are evaluated at some known time point (usually an initial condition), i.e.

$$y(t, \boldsymbol{p}) = y(0, \boldsymbol{p}) + y^{(1)}(0, \boldsymbol{p})t + ... + y^{(n)}(0, \boldsymbol{p})\frac{t^n}{n!} + ...$$

where

$$y^{(n)}(0, \boldsymbol{p}) = \frac{d^n}{dt^n} y(0, \boldsymbol{p}).$$

These derivatives are thus expressed solely in terms of the system parameters $\boldsymbol{p}$ (and $\tilde{\boldsymbol{p}}$ for the indistinguishability case). Since the coefficients in the Taylor series expansion are unique and, in principle, measurable, the identifiability problem reduces to determining the number of solutions for the system parameters in a set of algebraic equations that are, in general, non-linear in the parameters. Failure to obtain a result using this approach does not necessarily imply that the reaction schemes are distinguishable, and further tests may be required.

3.3. Quasi-Steady State Assumptions

In Model 1 there are 13 state variables corresponding to the concentrations of the enzyme (E), three substrates (S_1, S_2, S_3), three products (P, Q, R) and complexes (C_1, C_2, C_3, C_4, C_5, C_6). In Model 2 there are 11 state variables corresponding to the concentrations of the enzyme (E), the three substrates (S_1, S_2, S_3), three products (P, Q, R) and complexes (C_1, C_2, C_3 and C_4). After simplifying the models, by assuming the rate the complexes are formed are significantly faster than other reactions, the right-hand side of both sets of model equations is given by a set of 6 differential equations (all of them are defined by the same function) for the substrates and products. The quasi-steady-state equations for the two models are:

QSS Model 1: Let

$$f(S_1, S_2, S_3, P, Q, R) = \frac{A}{B} \tag{5}$$

where

$$A = E_0(r_1r_2r_3r_4r_5r_6r_7PQR - k_1k_2k_3k_4k_5k_6k_7S_1S_2S_3)$$

$$\begin{aligned}B = {} & k_4k_5k_6k_7r_1r_2 + k_5k_6k_7r_1r_2r_3 + k_6k_7r_1r_2r_3r_4 + k_1k_4k_5k_6k_7r_2S_1 + k_1k_5k_6k_7r_2r_3S_1 \\ & + k_1k_6k_7r_2r_3r_4S_1 + k_1k_2k_4k_5k_6k_7S_1S_2 + k_1k_2k_5k_6k_7r_3S_1S_2 + k_1k_2k_6k_7r_3r_4S_1S_2 \\ & + k_3k_4k_5k_6k_7r_1S_3 + k_1k_3k_4k_5k_6k_7S_1S_3 + k_2k_3k_4k_5k_6k_7S_2S_3 + k_1k_2k_3k_4k_5k_6S_1S_2S_3 \\ & + k_1k_2k_3k_4k_5k_7S_1S_2S_3 + k_1k_2k_3k_4k_6k_7S_1S_2S_3 + k_1k_2k_3k_5k_6k_7S_1S_2S_3 \\ & + k_1k_2k_3k_6k_7r_4S_1S_2S_3 + k_1k_2k_3k_4k_5r_6QS_1S_2S_3 + r_7(k_6 + r_6Q)R(r_1r_2(k_4k_5 + r_3(k_5 + r_4)) \\ & + k_3k_4k_5(r_1 + k_2S_2)S_3) + r_5P(k_7r_1r_2r_3r_4 + r_1r_2r_3r_4r_7R + k_1k_7r_2r_3r_4S_1 + k_1k_2k_7r_3r_4S_1S_2 \\ & + k_1k_2k_3k_4k_7S_1S_2S_3 + k_1k_2k_3k_7r_4S_1S_2S_3 + r_6Q(r_1r_2r_3r_4 + k_1S_1(r_2r_3r_4 + k_2S_2(r_3r_4 + k_3(k_4 + r_4)S_3)) \\ & + r_7R(k_4r_1r_2 + r_1r_2r_3 + r_1r_2r_4 + r_1r_3r_4 + r_2r_3r_4 + k_3r_1(k_4 + r_4)S_3 + k_2S_2(r_3r_4 + k_3(k_4 + r_4)S_3))))\end{aligned}$$

the reduced system consists of six equations, each with right-hand side given by f:

$$S_i' = f(S_1, S_2, S_3, P, Q, R)$$

and

$$P' = Q' = R' = -f(S_1, S_2, S_3, P, Q, R).$$

QSS Model 2: Let

$$g(S_1, S_2, S_3, P, Q, R) = \frac{C}{D} \tag{6}$$

where

$$C = E_0(r_1r_2r_3r_4r_5PQR - k_1k_2k_3k_4k_5S_1S_2S_3)$$

$$\begin{aligned}D = {} & k_4k_5r_1r_2 + k_5r_1r_2r_3 + r_1r_2r_3r_4 + k_3k_4k_5r_1S_3 + k_2k_3k_4k_5S_2S_3 + r_5PQR(k_4r_1r_2 + r_1r_2r_3 \\ & + r_1r_2r_4 + r_1r_3r_4 + r_2r_3r_4 + k_3r_1(k_4 + r_4)S_3 + k_2S_2(r_3r_4 + k_3(k_4 + r_4)S_3)) + k_1S_1(k_4k_5r_2 + k_5r_2r_3 \\ & + r_2r_3r_4 + k_3k_4k_5S_3 + k_2S_2(k_4k_5 + r_3(k_5 + r_4) + k_3(k_4 + k_5 + r_4)S_3))\end{aligned}$$

with the QSS system described by six equations, given by g (S_1, S_2 and S_3) or $-g$ (P, Q and R).

The quasi-steady assumption is a particular case of a simplification which relies on conservation laws; finding all conservation laws applicable for Model 1 and 2 it was possible to find reduced versions of both models. Using this method it is possible to choose to keep certain variables, for example only substrates, or only complexes.

3.4. Numerical Identifiability

For complex systems the symbolic analysis, as described in Section 3.2 is often difficult to perform manually, and as is the case for the two models presented (Section 2), intractable to symbolic computational. In such cases an alternative method is presented by Pohjanpalo [24] which employs the derivatives of the observations functions to determine the local identifiability of the unknown parameter vector by generating a potentially infinite Jacobian matrix, evaluated at a known time point (t_0) i.e.

$$\mathbf{J} = (J_{ij}(t_0)) = \begin{bmatrix} \frac{\partial y_1(t_0,\boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_1(t_0,\boldsymbol{p})}{\partial p_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m(t_0,\boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_m(t_0,\boldsymbol{p})}{\partial p_q} \\ \frac{\partial y_1^{(1)}(t_0,\boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_1^{(1)}(t_0,\boldsymbol{p})}{\partial p_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1^{(n)}(t_0,\boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_1^{(n)}(t_0,\boldsymbol{p})}{\partial p_q} \\ \vdots & \ddots & \vdots \end{bmatrix} \tag{7}$$

where y_i is the i^{th} observation, p_j the j^{th} element of the parameter vector p, and the superscript (n) donates the n^{th} derivative. For the nonlinear models presented above (Sections 2.1 and 2.2) due to the number of parameters to be estimated and the complexity of the model, the Jacobian matrix, as defined by (7) could not be calculated to sufficient dimensions to obtain an appropriate rank measure.[1]

A less general approach that is applicable to large complex systems is to use numerical methods

[1] To determine the rank a minimum of the 10th Taylor series co-efficient is required for the fast model and 14th for the ordered release model; however, evaluating at time $t = 0$ results in zero initial conditions for all but the enzyme and substrates, therefore higher order co-efficients are required.

to estimate the sensitivity matrix, where the sensitivity matrix is defined as the matrix of first derivatives, with respect to the parameter vector $\boldsymbol{p}$, of the observation at each measurement time-point

$$S = (S_{ij}) = \left(\frac{\partial y_i(\boldsymbol{t}, \boldsymbol{p})}{\partial p_j} \right) = \begin{bmatrix} \frac{\partial y_1(t_1, \boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_1(t_1, \boldsymbol{p})}{\partial p_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m(t_r, \boldsymbol{p})}{\partial p_1} & \dots & \frac{\partial y_m(t_r, \boldsymbol{p})}{\partial p_n} \end{bmatrix} \tag{8}$$

where t_r is the r^{th} observation sample. The sensitivity matrix can be estimated using finite-differencing methods; however, a more robust and accurate result can be obtained with Automatic Differentiation. The local identifiability of the system can be determined from the rank of the numerical sensitivity matrix by the following rules:

- if rank(S) = number of parameters, consistently, then the model is structurally identifiable, else
- if rank(S) $<$ number of parameters, the model is unidentifiable.

If the model is unidentifiable the unidentifiable parameters are indicated by the dependent columns of the sensitivity matrix.

4. Results

4.1. Structural Indistinguishability Analysis

Using the Taylor series approach it can be shown that in order for the outputs to be equal at least one of the parameters would genuinely need to be zero and therefore it can be safely concluded that the two models are distinguishable. This is the case if S_1 and P_1 are measured simultaneously, or other combinations of the products P, Q and R. Due to the complexity of the resulting equations it is not possible to provide a detailed list of these results; however, a worked example can be found in Appendix A.

For the reduced quasi steady-state model, as described in Section 3.3, it was not possible to perform the required symbolic manipulation to offer any information with respect to the model's indistinguishability due to the increased equation complexity.

4.2. Structural Identifiability Analysis

A purely analytical structural identifiability analysis using the Taylor series approach was carried out for both models to ensure that the model output (i.e. continuous, perfect and noise-free data) uniquely determines the unknown parameters. The two models yield the following parameter vectors:

- 14 parameters for Ordered Model (Section 2.1), $\boldsymbol{p_o} = (k_1,\ \ldots,\ k_7,\ r_1,\ \ldots,\ r_7)^T$.
- 10 parameters for Fast Model (Section 2.2), $\boldsymbol{p_f} = (k_1,\ \ldots,\ k_5,\ r_1,\ \ldots,\ r_5)^T$.

The results from all the symbolic identifiability analysis can be found in Table 1. It was found that by measuring only one of the substrates and one of the products it is possible to identify all parameters of Model 2, if it is assumed that the penultimate step is irreversible (i.e. $r_5 = 0$); if this is assumption is not made it is not possible to determine the identifiability of the model. If only one substrate is measured it is still possible to identify a number of the model parameters, from Table 1, 7 parameters are identifiable for Model 1, 5 for model 2. Similarly for Model 1, assuming that release of UNAMA and Pi are irreversible (i.e. $r_5 = r_6 = 0$), all of the remaining parameters can be uniquely determined when measuring one substrate and one product. All parameters not listed in Table 1 must therefore be considered undetermined with regard to their identifiability, as, if it were possible to compute additional Taylor series co-efficients, it may yield the identifiability of the remaining parameters.

Identifiability analyses for the two models were also performed for other combinations of measurements: all three products, all three substrates, etc. but measuring S_1, P_1 proved to be the minimal number of feasible observations required in order to identify uniquely the maximum number of model parameters.

This analysis was also performed on the reduced versions of the models (QSS and using conservation equations) but computational problems were again encountered due to the complex form of the simplified equations. The results did not improve when attempting to include more measurements, this was due to the fact that for the QSS version all six equations are the same and the

system is reduced to just one system equation. Therefore if more measurements are added then effectively this does not result in the addition of any additional information for the analysis.

4.3. Kinetic Constants and Parameter Models

Identifiability results for the QSS models appear to be unpromising with even the second Taylor series co-efficient being intractable to algebraic manipulation by symbolic computation. However, although several parameters of the full model can be identified this does not imply that this information can be used for the QSS case. It is possible to reparameterise the QSS model by using chemical kinetic constants which relate substance reaction rates (e.g. Michaelis-Menten) to rate constants (e.g. k_i, r_i)), see Segel [29]. Reparameterising in terms of kinetics constants introduces an additional set of mathematical relationships into already intractable system equations; therefore, to simplify the QSS model the equations can be reparameterised to incorporate the rate constants into macro parameters by replacing the rate constant co-efficients of (5) and (6) with a single macro parameters (e.g. $K_1 = r_1r_2r_3r_4r_5r_6r_7$ in the numerator of (5)). If these macro parameters are identifiable and there is a unique mapping between rate constants and macro parameters the identifiability of the underlying forward and reverse rate constants can be determined.

The process of performing the identifiability of the QSS model can be simplified by noting the replication inherent in the system equations. As an example consider the ordered release model (Model 1, (5)), if the first substrate is measured the following relationships can be used to eliminate variables

$$
\begin{aligned}
S_2(t) &= S_1(t) - (S_1(0) - S_2(0)) \\
S_3(t) &= S_1(t) - (S_1(0) - S_3(0)) \\
P(t) &= S_1(0) - S_1(t) \\
Q(t) &= S_1(0) - S_1(t) \\
R(t) &= S_1(0) - S_1(t)
\end{aligned}
$$

Substituting these conservation laws into the functions f and g, ((5) and (6), respectively, results

in a function of $S_1(t)$ only, i.e.

$$S_1'(t, \boldsymbol{K}) = F(S_1(t), \boldsymbol{K}), \tag{9}$$

where $\boldsymbol{K}$ is a vector of the macro parameters K_i. If an alternative vector is considered ($\tilde{\boldsymbol{K}}$) for the system to be identifiable, when measuring only S_1, it is the case that

$$S_1'(t, \boldsymbol{K}) = F(S_1(t), \boldsymbol{K}) = F(S_1(t), \tilde{\boldsymbol{K}}) = S_1'(t, \tilde{\boldsymbol{K}}). \tag{10}$$

As F is a rational function it can be re-arrange to give a univariate polynomial (in $S_1(t)$), which can be equated to zero. For this equation to be equal to zero on an open interval of time it must be the case that the coefficients are identically zero. For the ordered model this results in nine non-zero coefficients, indicating that a maximum of nine rate constants of the QSS model can be identifiable, five less than the required fourteen, whilst for the fast QSS Model only 9 parameters where shown to be identifiable.

The relationship between macro parameters and rate constant can be determined by examining the rank of the macro-rate constant Jacobian matrix, i.e.

$$J = (J_{ij}) = \left(\frac{\partial K_i}{\partial p_j}\right) = \begin{bmatrix} \frac{\partial K_1}{k_1} & \dots & \frac{\partial K_1}{\partial r_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial K_m}{\partial k_1} & \dots & \frac{\partial K_m}{\partial r_n} \end{bmatrix}. \tag{11}$$

where K_i is the i^{th} element of the macro parameter vector $\boldsymbol{K}$ and p_j is the j^{th} element of the rate constant vector $\boldsymbol{p}$. For the ordered model the rank of this matrix is thirteen, one less than the required number of parameters, and 9 for the fast model, also one less than the total rate constant. In both QSS models there is therefore a loss of information with regards to the macro parameter reparameterisation which should be considered when designing the experiment. Therefore, this has important biological implications since it is the quasi steady-state reparameterisation that is often used in order to understand the enzymes that catalyse chemical reactions, it is fundamental to have quantitative measurement of the rate of the catalysed reaction.

4.4. Numerical Identifiability Analysis

In order to conduct the numerical identifiability analysis it must be possible to simulate the output of the system equations, this requires a parameter vector ($p_0 \in \mathbb{R}^q$) around which the sensitivity can be assessed. As initial estimates were not available from the literature, the initial values were taken from experimental measurement of MurB, a complex also synthesised within the peptidoglycan reaction scheme (see Figure 1). Due to the inaccuracy inherent in this approach a Monte-Carlo approach was used, extracting a set of parameter values from a log normal distribution $p_0 \times 10^{-3} \leq p_0 \leq p_0 \times 10^3$, where p_0 is the mean value specified in Table 2. This range covers the rate constants specified by Brown et al. [2] for MurA (see Figure 1).

Even with a large parameter space the results from a numeric identifiability analysis must still considered 'local', as it can not be guaranteed that uniqueness exists for the entire theoretical parameter space (i.e. on $\mathbb{R}^+$), unlike the symbolic identifiability analysis shown. The time-points for the measurements were taken as those achievable by typical stop-flow analysis (sampling frequency of 1000 Hz); however, a range of sampling frequencies where tried (500-10,000 Hz), with no difference in the numerical identifiability from those shown in Table 3.

The results of the numerical identifiability analysis can be seen in Table 3, where the minimum and maximum rank of the sensitivity matrix (8) is shown for one thousand different sets of initial parameters drawn from the parameter space, as defined in Table 2. As in the analytical identifiability analysis it is evident that measuring both product (Pi, variable P or Q) and substrate (ATP, variable S_1) together yields a greater number of identifiable parameters than measuring either in isolation; in fact, if both are measured the system is locally identifiable for the entire parameter space defined. However, the numerical analysis offers an improvement upon the analytical results in relation to performing parameter estimation when the product or substrate are measured separately. In the structural analysis no results are complete regarding the identifiability of any of the parameters measuring product (P) alone, and a minimal set of parameters is categorised for substrate only measurements. The numerical analysis suggests that for the ordered release model, in a worst case scenario, five parameters are unidentifiable if the ATP substrate (S_1) is measured, with four being unidentifiable if the first product is observed (P). Similar results are seen for the

fast release model, with the maximum of four parameters remaining unidentifiable. For the ordered release model the unidentifiable parameters appear consistently from the reverse rate constants, $(r_4, \ldots, r_7)$, irrespective of the observed state. However, if the fast release structure is assumed, measuring ATP consistently results in the reverse rate constants being unidentifiable (r_4 and r_5); measuring Pi in isolation, yields the forward rate constants (k_2 and k_3) as unidentifiable.

Numerical identifiability analysis was also performed on the two QSS models. The results are shown in the last two rows of Table 3 and three parameters were seen to be consistently numerically identifiable for the ordered model, and two for the fast release model. Both of these numbers are significantly below (nine for both models) the maximum possible number of identifiable parameters.

5. Conclusions

In this paper the application of a range of indistinguishability and identifiability analysis, both symbolic and numeric, have been applied to a kinetic reaction scheme. This analysis has highlighted the inherent difficulties in designing experiment protocols for such a complex non-linear system. However, using the proposed models for the MurC phase of the BPB pathway as an example, the benefits of performing this analysis have been shown. Through structural indistinguishability it has been shown that it is theoretically possible to discriminate between candidate mechanisms in order to ascertain the most feasible and appropriate model to describe the chemical process. Integral to the applicability of the analysis is the determination of necessary, and feasible, measurements/experiments required in order to achieve distinguishability. From this analysis it can be concluded that the common assumption of considering fast release of products as equivalent to a one-by-one step mechanism may not be accurate as these processes are intrinsically different (distinguishable) and in theory by observing a substrate and product simultaneously distinguishable.

Once it has been determined that the candidate models are distinguishable it is necessary to ensure the rate constants of the reaction are identifiable *a priori*, given the available experimental measurements. For the MurC reaction the minimal number of feasible observations required in order to identify, uniquely, the maximum number of model parameters proved to be the first substrate and one of the products of the reaction. However, using purely symbolic analysis it is not possible

to determine if all the parameters of both models are uniquely identifiable, but combining numeric and symbolic methods suggests that measuring the substrate and product in isolation will yield a model that is unidentifiable, whilst measuring these simultaneously will ensure that all parameters, and therefore the model in its entirety, will be at least locally identifiable. These results have implications in experiment design as it has been shown that relatively few measurable species are needed to be able to identify a great deal of the underlying parameters involved and therefore this reduces the amount of necessary experiments that may need to be performed.

An important issue addressed in this work are the difficulties encountered when extracting rate constant information from steady-state experiments. Unlike the models generated for the pre-steady state experiments, i.e. the full fast and ordered release models, the quasi steady-state models are intractable to symbolic analysis, with respect to the models indistinguishability or identifiability. Numerical results suggests that a small number of rate constants are identifiable but without the support of the symbolic results it is difficult to confidently identify which parameters the results relate to. Experimenters should therefore realise that multi-substrate reactions cannot be adequately described by velocity equations derived from rapid equilibrium assumptions alone as the distribution of enzyme species depends on the rate constants of all steps.

Conflicts of interest

None declared.

Acknowledgements

EPSRC funded project *Indistinguishability analysis for model discrimination in Systems Biology: A Feasibility Study applied to Bacterial Peptidoglycan Biosynthesis* (EP/E057535/1).

References

[1] Anderson, M. S., Eveland, S. S., Onishi, H. R., Pompliano, D. L., 1996. Kinetic mechanism of the Escherichia coli UDPMurNAc-Tripeptide D-Alanyl-D-alanine-adding enzyme: Use of a Glutathione S-Transferase Fusion. Biochemistry 35, 16264–16269.

[2] Brown, E. D., Marquardt, J. L., Lee, J. P., Walsh, C. T., Anderson, K. S., Sep 1994. Detection and characterization of a phospholactoyl-enzyme adduct in the reaction catalyzed by udp-n-acetylglucosamine enolpyruvoyl transferase, murz. Biochemistry 33 (35), 10638–10645.

[3] Bellu, G., Saccomani, M., Audoly, S., D'Angiò, L., Oct 2007. Daisy: a new software tool to test global identifiability of biological and physiological systems. Computer Methods and Programs in Biomedicine 88 (1), 52–61.

[4] Chappell, M. J., Godfrey, K. R., Vajda, S., Nov 1990. Global identifiability of the parameters of nonlinear systems with specified inputs: a comparison of methods. Math Biosci 102 (1), 41–73.

[5] Cobelli, C., Salvan, A., 1977. Parameter estimation in a biological two compartment model–ii: a computer experimental study of the influence of the initial estimate in the parameter space,of some sampling protocols and of weighting factors. Mathematical Biosciences 37 (3-4), 267 – 274.

[6] Franco, R., Gavaldà, M. T., Canela, E. I., Sep 1986. A computer program for enzyme kinetics that combines model discrimination, parameter refinement and sequential experimental design. Biochem J 238 (3), 855–862.

[7] El Zoeiby, A., Sanschagrin, F., Havugimana, P. C., Garnier, A., Levesque, R. C., 2001. In vitro reconstruction of the biosynthetic pathway of peptidoglycan cytoplasmic precursor in Pseudomonas aeruginosa. FEMS Microbiol. Lett. 201, 229–235.

[8] Emanuele, J. J., Jin, H., Yanchunas, J., Villafranca, J. J., 1997. Evaluation of the kinetic mechanism of Escherichia coli Uridine Diphosphate-N-acetylmuramate:L-Alanine Ligase. Biochemistry 36, 7264–7271.

[9] Érdi, P., Tóth, J., 1989. Mathematical Models of Chemical Reactions. Princeton University Press.

[10] Espenson, J. H., 1995. Chemical Kinetics and Reaction Mechanisms. McGraw-Hill, Singapore.

[11] Evans, N., Hattersley, J., Hutchison, C., Hu, Y., Godfrey, K., Bradwell, A., Mead, G., Chappell, M., September 2006. Modelling of haemodialysis in limiting serum free light chains in patients with renal failure. In: D. Feng, O. Dubois, J. Z., Carson, E. (Eds.), Proceedings of the 6th IFAC Symposium on Modelling and Control in Biomedical Systems. Elsevier, Oxford, pp. 75–80.

[12] Evans, N. D., Chapman, M. J., Chappell, M. J., Godfrey, K. R., 2002. Identifiability of uncontrolled nonlinear rational systems. Automatica 38, 1799–1805.

[13] Evans, N. D., Chappell, M. J., Chapman, M. J., Godfrey, K. R., 2004. Structural indistinguishability between uncontrolled (autonomous) nonlinear analytic systems. Automatica 40, 1947–1953.

[14] Fersht, A. R., 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. Freeman, New York.

[15] Fisher, F. M., 1959. Generalization of the rank and order conditions for identifiability. Econometrica 27 (3), 431–447.

[16] Godfrey, K. R., Chapman, M. J., 1990. Identifiability and indistinguishability of linear compartmental models. Math. Comput. Simulat. 32, 273–295.

[17] Godfrey, K. R., DiStefano III, J. J., 1987. Identifiability of model parameters. In: Walter, E. (Ed.), Identifiability of Parametric Models. Pergamon Press, Ch. 1, pp. 1–20.

[18] Jacquez, J. A., Mar 1984. Red blood cell as glucose carrier: significance for placental and cerebral glucose transfer. Am J Physiol 246 (3 Pt 2), R289–R298.

[19] Kholodenko, B. N., Bruggeman, F. J., Sauro, H. M., 2005. Mechanistic and modular approaches to modeling and inference of cellular regulatory networks. In: Alberghina, L., Westerhoff, H. (Eds.), Systems Biology: Definitions and Perspectives. Springer-Verlag.

[20] Ljung, L., Glad, T., 1994. On global identifiability for arbitrary model parametrizations. Automatica 30, 265–276.

[21] Marmor, S., Petersen, C. P., Reck, F., Yang, W., Gao, N., Fisher, S. L., 2001. Biochemical characterization of a phosphinate inhibitor of Escherichia coli MurC. Biochemistry 40, 12207–12214.

[22] Nölting, B., 1999. Protein Folding Kinetics: Biophysical Methods. Springer, Berlin.

[23] Pohjanpalo, H., 1978. System identifiability based on the power series expansion of the solution. Math. Biosci. 41, 21–33.

[24] Pohjanpalo, H., 1982. Identifiability of deterministic differential models in state space. research report no. 56. Tech. rep., Technical Research Centre of Finland, Espoo.

[25] Saccomani, M., Audoly, S., Bellu, G., D'Angio, L., 2001. A new differential algebra algorithm to test identifiability of nonlinear systems with given initial conditions. IEEE Conference on Decision and Control 4, 3108–3113.

[26] Saccomani, M. P., Audoly, S., D'Angio, L., 2003. Parameter identifiability of nonlinear systems: the role of initial conditions. Automatica 39, 619–632.

[27] Sauer, U., Heinemann, M., Zamboni, N., 2007. GENETICS: Getting Closer to the Whole Picture. Science 316 (5824), 550–551.

[28] Schnell, S., Chappell, M. J., Evans, N. D., Roussel, M. R., 2006. The mechanism distinguishability problem in biochemical kinetics: The single-enzyme, single-substrate reaction as a case study. Comptes Rendus Biologies 329, 51–61.

[29] Segel, I. H., 1993. Enzyme Kinetics. Wiley Classics.

[30] Snoep, J. L., Westerhoff, H. V., 2005. From isolation to integration, a systems biology approach for building the Silicon Cell. In: Alberghina, L., Westerhoff, H. (Eds.), Systems Biology: Definitions and Perspectives. Springer-Verlag.

[31] Tunali, E. T., Tarn, T. J., 1987. New results for identifiability of nonlinear systems. IEEE T. Automat. Contr. 32, 146–154.

[32] Vajda, S., Godfrey, K. R., Rabitz, R., 1989. Similarity transformation approach to identifiability analysis of nonlinear compartmental models. Math. Biosci. 93, 217–248.

[33] Walsh, C., 2000. Molecular mechanisms that confer antibacterial drug resistance. Nature 406, 775–781.

[34] Walter, E., 1982. Identifiability of State Space Models. Springer-Verlag.

A. Indistinguishability Example

Due to the complexity of the models presented for describing the MurC reaction it is not possible to provide full details of all of the analysis. Therefore, two worked examples are shown describing the use of the Taylor series method to determine model indistinguishability and parameter identifiability. The system of equations for Model 1 (Ordered complex release) can be described by the following:

$$\begin{aligned}
E'(t) &= -k_1E(t)S_1(t) + r_1C_1(t) + k_7C_6(t) - r_7E(t)R(t) \\
S_1'(t) &= -k_1E(t)S_1(t) + r_1C_1(t) \\
S_2'(t) &= -k_2C_1(t)S_2(t) + r_2C_2(t) \\
S_3'(t) &= -k_3S_3(t)C_2(t) + r_3C_3(t) \\
C_1'(t) &= k_1E(t)S_1(t) - r_1C_1(t) - k_2C_1(t)S_2(t) + r_2C_2(t) \\
C_2'(t) &= k_2C_1(t)S_2(t) - r_2C_2(t) - k_3S_3(t)C_2(t) + r_3C_3(t) \\
C_3'(t) &= k_3S_3(t)C_2(t) - r_3C_3(t) - k_4C_3(t) + r_4C_4(t) \\
C_4'(t) &= k_4C_3(t) - r_4C_4(t) - k_5C_4(t) + r_5C_5(t)P(t) \\
C_5'(t) &= k_5C_4(t) - r_5C_5(t)P(t) - k_6C_5(t) + r_6C_6(t)Q(t) \\
C_6'(t) &= k_6C_5(t) - r_6C_6(t)Q(t) - k_7C_6(t) + r_7E(t)R(t) \\
P'(t) &= k_5C_4(t) - r_5C_5(t)P(t) \\
Q'(t) &= k_6C_5(t) - r_6C_6(t)Q(t) \\
R'(t) &= k_7C_6(t) - r_7E(t)R(t)
\end{aligned} \tag{12}$$

with the following initial conditions:

$$S_1(0) = S_{10}; S_2(0) = S_{20}; S_3(0) = S_{30}; C_1(0) = 0; C_2(0) = 0; C_3(0) = 0;$$
$$C_4(0) = 0; C_5(0) = 0; C_6(0) = 0; P(0) = 0; Q(0) = 0; R(0) = 0; E(0) = E_0.$$

Model 2 (Fast release) can be described by:

$$\begin{aligned}
\tilde{E}'(t) &= -\tilde{k_1}\tilde{E}(t)\tilde{S_1}(t) + \tilde{r_1}\tilde{C_1}(t) + \tilde{k_5}\tilde{C_4}(t) - \tilde{r_5}\tilde{E}(t)\tilde{P}(t)\tilde{Q}(t)\tilde{R}(t) \\
\tilde{S_1}'(t) &= -\tilde{k_1}\tilde{E}(t)\tilde{S_1}(t) + \tilde{r_1}\tilde{C_1}(t) \\
\tilde{S_2}'(t) &= -\tilde{k_2}\tilde{C_1}(t)\tilde{S_2}(t) + \tilde{r_2}\tilde{C_2}(t) \\
\tilde{S_3}'(t) &= -\tilde{k_3}\tilde{S_3}(t)\tilde{C_2}(t) + \tilde{r_3}\tilde{C_3}(t) \\
\tilde{C_1}'(t) &= \tilde{k_1}\tilde{E}(t)\tilde{S_1}(t) - \tilde{r_1}\tilde{C_1}(t) - \tilde{k_2}\tilde{C_1}(t)\tilde{S_2}(t) + \tilde{r_2}\tilde{C_2}(t) \\
\tilde{C_2}'(t) &= \tilde{k_2}\tilde{C_1}(t)\tilde{S_2}(t) - \tilde{r_2}\tilde{C_2}(t) - \tilde{k_3}\tilde{S_3}(t)\tilde{C_2}(t) + \tilde{r_3}\tilde{C_3}(t) \\
\tilde{C_3}'(t) &= \tilde{k_3}\tilde{S_3}(t)\tilde{C_2}(t) - \tilde{r_3}\tilde{C_3}(t) - \tilde{k_4}\tilde{C_3}(t) + \tilde{r_4}\tilde{C_4}(t) \\
\tilde{C_4}'(t) &= \tilde{k_4}\tilde{C_3}(t) - \tilde{r_4}\tilde{C_4}(t) - \tilde{k_5}\tilde{C_4}(t) + \tilde{r_5}\tilde{E}(t)\tilde{P}(t)\tilde{Q}(t)\tilde{R}(t) \\
\tilde{P}'(t) &= \tilde{k_5}\tilde{C_4}(t) - \tilde{r_5}\tilde{E}(t)\tilde{P}(t)\tilde{Q}(t)\tilde{R}(t) \\
\tilde{Q}'(t) &= \tilde{k_5}\tilde{C_4}(t) - \tilde{r_5}\tilde{E}(t)\tilde{P}(t)\tilde{Q}(t)\tilde{R}(t) \\
\tilde{R}'(t) &= \tilde{k_5}\tilde{C_4}(t) - \tilde{r_5}\tilde{E}(t)\tilde{P}(t)\tilde{Q}(t)\tilde{R}(t)
\end{aligned} \tag{13}$$

with the following initial conditions:

$$\tilde{S_1}(0) = S_{10}; \tilde{S_2}(0) = S_{20}; \tilde{S_3}(0) = S_{30}; \tilde{C_1}(0) = 0; \tilde{C_2}(0) = 0;$$
$$\tilde{C_3}(0) = 0; \tilde{C_4}(0) = 0; \tilde{P}(0) = 0; \tilde{Q}(0) = 0; \tilde{E}(0) = E_0; \tilde{R}(0) = 0;$$

If both the first substrate (S_1) and first product (P) are measured the observation functions for the two systems are

$$\begin{aligned}
\boldsymbol{y(t)} &= (S_1(t), P(t))^T \\
\boldsymbol{\tilde{y}(t)} &= (\tilde{S_1}(t), \tilde{P}(t))^T
\end{aligned}$$

For the models to be indistinguishable all terms in the Taylor series expansions for each system must be equal. Looking at the first term, evaluated at time $t = 0$ yields

$$\begin{aligned} \boldsymbol{y(0)} &= (S_{10}, 0)^T \\ \boldsymbol{\tilde{y}(0)} &= (S_{10}, 0)^T \end{aligned}$$

Evaluating the second Taylor series term

$$\begin{aligned} \boldsymbol{y^1(0)} &= (-E_0 k_1 S_{10}, 0)^T \\ \boldsymbol{\tilde{y}^1(0)} &= (-E_0 \tilde{k_1} S_{10}, 0)^T \end{aligned}$$

and therefore $\tilde{k_1} = k_1$. For the third Taylor series co-efficient, with k_1 substituted for $\tilde{k_1}$

$$\begin{aligned} \boldsymbol{y^2(0)} &= (E_0 k_1 S_{10}(E_0 k_1 + r_1 + k_1 S_{10}), 0)^T \\ \boldsymbol{\tilde{y}^2(0)} &= (E_0 k_1 S_{10}(E_0 k_1 + \tilde{r_1} + k_1 S_{10}), 0)^T \end{aligned}$$

resulting in $\tilde{r_1} = r_1$. This process continues until the eighth Taylor series co-efficient, where for the models to be indistinguishable the following must hold

$$E_0 k_1^2 k_2 k_3 k_4^2 k_5^2 S_{10}^2 S_{20}(-6E_0 k_1 + r_2 + k_2 S_{20}) S_{30} = 0. \tag{14}$$

Therefore, for the eighth terms to be equal, with a generic parameter vector, at least one of the forward rate constants (k_1, k_2, k_3, k_4 and k_5) must be zero, which for the reaction to take place, is not possible, thus indicating the models must, in theory, be distinguishable.

B. Identifiability Example

To determine the identifiability of the Model 2 (Fast release) when both the first substrate and product (S_1 and P_1) are measured simultaneously. The model is defined by (13); however, as the same model is being compared the observation function is given by $\boldsymbol{y(t,p)} = (S_1(t,\boldsymbol{p}), \mathrm{P}(t,\boldsymbol{p}))^T$.

In order to generate the generic solutions, two parameter sets are created ($\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$):

$$\boldsymbol{p} = (k_1, r_1, k_2, r_2, k_3, r_3, k_4, r_4, k_5, r_5)^T,$$

and

$$\tilde{\boldsymbol{p}} = (\tilde{k_1}, \tilde{r_1}, \tilde{k_2}, \tilde{r_2}, \tilde{k_3}, \tilde{r_3}, \tilde{k_4}, \tilde{r_4}, \tilde{k_5}, \tilde{r_5})^T.$$

As with the indistinguishability method seen above, successive Taylor series coefficients are calculated to yield a set of equations relating the model parameters, from which the identifiability classification (global or local) of each parameter can be determined. For example, the second Taylor series for $y(\boldsymbol{p}, t_0)$ are

$$\tilde{\boldsymbol{y}}^{\boldsymbol{1}}(\boldsymbol{0}, \boldsymbol{p}) = (-E_0 k_1 S_{10}, 0)^T$$
$$\tilde{\boldsymbol{y}}^{\boldsymbol{1}}(\boldsymbol{0}, \tilde{\boldsymbol{p}}) = (-E_0 \tilde{k_1} S_{10}, 0)^T$$

showing k_1 to be identifiable. For the third Taylor series term, with $\tilde{k_1}$ substituted for the known k_1

$$\tilde{\boldsymbol{y}}^{\boldsymbol{2}}(\boldsymbol{0}, \boldsymbol{p}) = (E_0 k_1 S_{10}(E_0 k_1 + r_1 + k_1 S_{10}), 0)^T$$
$$\tilde{\boldsymbol{y}}^{\boldsymbol{2}}(\boldsymbol{0}, \tilde{\boldsymbol{p}}) = (E_0 k_1 S_{10}(E_0 k_1 + \tilde{r_1} + k_1 S_{10}), 0)^T.$$

Indicating that the first rate constant r_1 is identifiable.

This process is repeated using increasing derivatives of the observation function ($y(\boldsymbol{p}, t)$ and $y(\tilde{\boldsymbol{p}}, t)$) until either all parameters, or as in the case for this model, computer resource limits the generation of further co-efficients. The eighth Taylor series coefficient giving the result as seen in Table 1, e.g. $k_1 = \tilde{k_1}, r_1 = \tilde{r_1}, k_2 = \tilde{k_2}, r_2 = \tilde{r_2}, k_3 = \tilde{k_3}, r_3 = \tilde{r_3}, k_4 = \tilde{k_4}, r_4 = \tilde{r_4}$ and $k_5 = \tilde{k_5}$, with the identifiability of rate constant r_5 unknown. The two subsequent Taylor series terms can be evaluated symbolically but the identifiability of the last reverse rate constant r_5 remains undetermined. The tenth derivative can not be evaluated due to memory resource allocation.

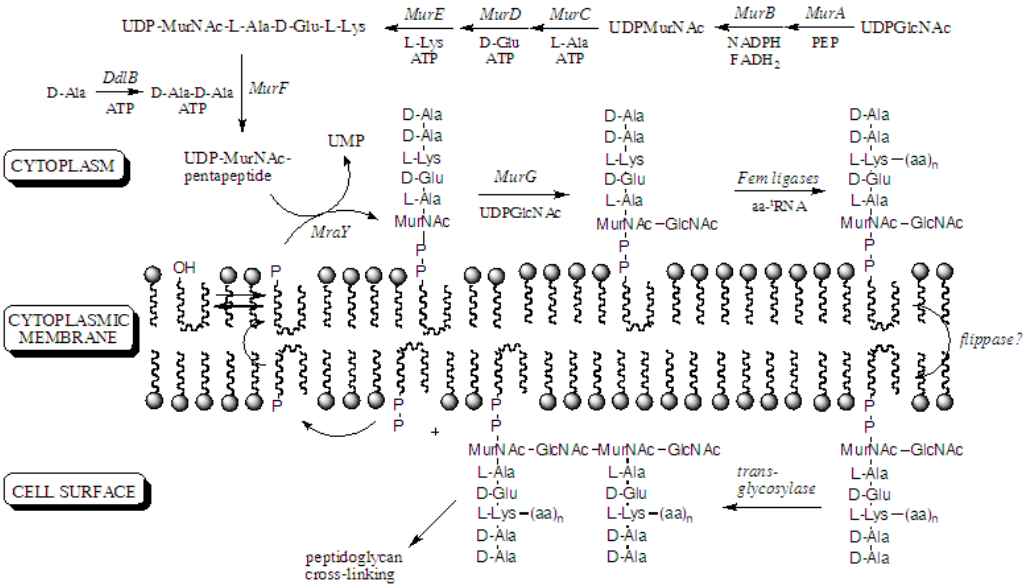

Figure 1: The biosynthetic pathway for bacterial peptidoglycan

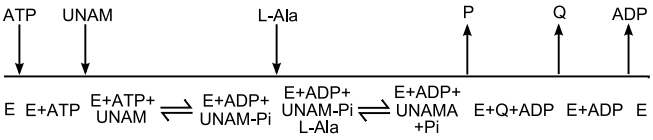

Figure 2: Kinetic Mechanism of Uridine Diphosphate-N-acetylmuramate:L-Alanine Ligase (MurC)

Table 1: identifiability results, only identifiable parameters are shown; all other parameters remain undetermined. The number in brackets indicates the number of Taylor series coefficients evaluated to achieve this result.

Observation	S_1	P	P, S_1
Model 1	k_1, r_1, k_2, r_2, k_3, r_3, k_4 (9)	- (9)	k_1, r_1, k_2, r_2, k_3, r_3, k_4, r_4, k_5, k_6, k_7, r_7 (9)
Model 2	k_1, r_1, k_2, r_2, k_3 (8)	- (7)	k_1, r_1, k_2, r_2, k_3, r_3, k_4, r_4, k_5 (8)
QSS 1	- (2)	- (2)	- (2)
QSS 2	- (4)	- (4)	- (4)

Table 2: mean parameter values for Monte-Carlo simulations and parameter range over which the initial value for the simulation are sampled.

Parameter (units)	**Value**	**Range**	**Parameter (units)**	**Value**	**Range**
$k_1(\mu M^{-1}s^{-1})$	2.5E-01	2.5E-4 to 2.5E2	$r_1(s^{-1})$	2.6E01	2.6E-3 to 2.6E4
$k_2(\mu M^{-1}s^{-1})$	2.4E01	2.4E-3 to 2.4E4	$r_2(s^{-1})$	2.6E01	2.6E-3 to 2.6E4
$k_3(\mu M^{-1}s^{-1})$	7.5E-01	7.5E-4 to 7.5E2	$r_3(s^{-1})$	1.9E01	1.9E-3 to 1.9E4
$k_4, k_5\ (s^{-1})$	1.1E01	1.1E-2 to 1.1E4	$r_4, r_5(s^{-1})$	2.6E01	2.6E-3 to 2.6E4
$k_6, k_7(\mu M^{-1}s^{-1})$	2.5E-01	2.5E-4 to 2.5E2	$r_6, r_7(s^{-1})$	2.6E01	2.6E-3 to 2.6E4

Table 3: numerical identifiability results; rank of the sensitivity matrix for the two models (Model 1 - ordered, Model 2 - fast) and three observations (P - Product Pi or UNAMA, S1 - Substrate ATP).

Observation	S_1	P	P, S_1
Model 1	9-13	10-11	14
Model 2	6-8	6-8	10
QSS 1	3	3	3
QSS 2	2	2	2