

Bayesian analysis of non-linear differential equation models with application to a gut microbial ecosystem

Daniel J. Lawson*¹, Grietje Holtrop^{1,2}, and Harry Flint²

¹ Biomathematics and Statistics Scotland James Clerk Maxwell Building, The King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, Scotland, UK

² Rowett Institute of Nutrition and Health at the University of Aberdeen, Bucksburn, Aberdeen, AB21 9SB, Scotland, UK

Received zzz, revised zzz, accepted zzz

Process models specified by non-linear dynamic differential equations contain many parameters, which often must be inferred from a limited amount of data. We discuss a hierarchical Bayesian approach combining data from multiple related experiments in a meaningful way, which permits more powerful inference than treating each experiment as independent. The approach is illustrated with a simulation study and example data from experiments replicating aspects of the human gut microbial ecosystem. A predictive model is obtained that contains prediction uncertainty caused by uncertainty in the parameters, and we extend the model to capture situations of interest that cannot easily be studied experimentally.

Key words: Bacteria; Convergence; Ecology; Hierarchical Model; Process Model;
Supporting Information for this article is available from the author or on the WWW under
<http://dx.doi.org/10.1022/bimj.XXXXXXX>

1 Introduction

Many biological and physical systems can in theory be described by a well understood process, but the process cannot be used directly in practical inference because parameters are inaccessible for direct measurement. A ‘reductionist’ strategy for inference (Levi-Montalcini and Calissano, 2006) by measuring each parameter in a separate experiment is often not feasible and ignores interactions between parameters.

We consider parameter inference for a non-linear differential equation model describing multiple experiments simultaneously. We will use a hierarchical Bayesian approach as it allows for combining data from different sources whilst accounting for differences between them. Previous Bayesian approaches to differential equation models include pharmacokinetic models (Gelman et al., 1996; Wakefield, 1996; Lunn and Aarons, 1997; Mezzetti et al., 2003), and Human Immunodeficiency Virus (HIV) modelling (Putter et al., 2002; Banks et al., 2005; Huang et al., 2006). These describe how a potentially large number of individuals are described by the same process but with different parameters. Our methodology aims to relate parameters between experiments that are limited in data. Parts of the underlying processes are assumed identical across experiments, whilst other aspects of the underlying processes are allowed to vary. Useful inference requires both knowledge of the process and a statistical model to relate the experimental data. Model parameters are not uniquely identifiable for each experiment individually but when multiple experiments are linked using a hierarchical statistical model, they can be identified.

We apply this approach to an ecological microbiology model with which we ultimately wish to understand and control bacterial populations inside the human gut. These are generally inaccessible for direct measurement. The gut has strong spatial structure, many nutrients are absorbed into the gut wall and

*Corresponding author: e-mail: dan.lawson@bristol.ac.uk. Current address: Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK. Phone: +44 1179287990

are therefore unavailable to bacteria, and flow rates are strongly time dependent (Louis et al., 2007). In contrast, the *in-vitro* experiments that provide our data are designed to have limited spatial structure, nonabsorbent walls and simple continuous flow. The parameterised biological model obtained from such studies, with quantified uncertainty, can then be used to predict how bacteria might behave under conditions that more closely resemble the human gut. Of specific interest to us is whether the bacterial dynamics change when absorption of nutrients and metabolic products is taken into account. This can not be investigated easily by experiments *in vitro*, but is straightforward to build into the statistical model which can then be used to predict how bacterial behaviour is influenced.

We will examine the qualitative benefits of a Bayesian approach using Markov-Chain Monte Carlo (MCMC) (Gamerman, 1997) techniques using simulated data, followed by experimental data. With limited data the posterior distribution is ‘large’ and complex meaning that MCMC algorithms converge very slowly. However, simultaneous inference restricts the posterior distribution size and simplifies its shape, making it feasible to explore the posterior parameter space. In Section 2 we discuss an ecological modelling problem for microbiology data to motivate the method. In Section 3 we define the statistical model, which we explore using a simulation study in Section 4. We then apply the method to the microbiology data in Section 5 before our concluding remarks in Section 6.

2 Ecological gut microbiology problem

A diverse range of microbes live in the human colon by fermenting nutrients (called substrates) that pass through our stomach without being digested. Two types of substrate are of interest to us: starch and non-starch polysaccharides (NSP). Starch is easily fermented by the bacteria, whereas NSP, consisting mainly of fibre, is harder to ferment. As a consequence, growth on starch is of the order 10 times faster than growth on NSP, and essentially all starch is used by the bacteria, whilst some NSP is not fermented and passes out of the colon in feces. Bacteria produce short-chain fatty acids (SCFA) as a by-product of fermentation. The majority of the SCFA will be absorbed by the human host, but can also be further processed by other bacterial types. SCFA concentrations are of primary concern for health, as each has unique health implications.

For simplicity, the colon can be divided into two main sections. The first section is closest to the stomach, hence plenty of starch and NSP are available to the bacteria allowing high fermentation rates and high acidity (pH of approximately 5.5). In the second section of the colon, closer to the rectum, starch supply is reduced although NSP supply may still be high. Fermentation activity is lower and the pH will tend more towards neutral (pH of approximately 6.5). Some bacteria are more tolerant of high acidity than others, and as a consequence bacterial composition will vary depending on local conditions. For more details, see e.g. Cummings (1997).

2.1 Experimental system

The results of Walker et al. (2005) describe an experimental system using human fecal matter as a seed for a ‘realistic’ bacterial community, which is placed into a fermentor vessel and fed on a continuous supply of substrates with consistency chosen to represent many nutrients available in the human colon. For inference purposes, the substrates are combined into starch and NSP. This experimental set up represents a simplification of the human colon as characteristics such as spatial structure, absorption of nutrients and products, and fluctuations in flow, are not considered. A stirring mechanism ensures that the contents of the vessel are well-mixed, and the material that leaves the fermentor via an overflow outlet is therefore assumed to be representative of the vessel contents. The bacteria and SCFA are measured from samples taken from the outlet. Substrate input is known but is not measured in the output. The pH in the vessel is controlled by the operator and is held constant at either pH 5.5 (representing the first section of the colon) or pH 6.5 (representing the later section of the colon). Details of how our data were extracted from (Walker et al., 2005) are given in Web Appendix A.

2.2 Mathematical model

Modelling approaches for microbial systems range from simple differential equation models (Wilkinson, 1998; Coleman et al., 1996) to spatially explicit individual based simulations (Kreft et al., 1998; Beretta and Kuang, 2000). As the simplest relevant model, we use a differential equation model for bacterial growth and competition at the level of functional groups. Our mathematical model is an abstraction of the commonly used experimental fermentor system described in Section 2.1. We consider a single well-mixed compartment, although extension to multiple compartments is trivial. The input contains substrates which are converted to bacterial mass, waste, and a number of SCFA. Some bacterial strains are capable of gaining extra energy from SCFA. Examples are shown in Fig. 1, which are used in Sections 4 and 5.

Bacteria belonging to group i are associated with a substrate j at concentration $B_{ij}(t)$, and we observe the total concentration $B_i(t) = \sum_j B_{ij}(t)$. It is assumed that bacterial groups are well mixed between substrates j , such that B_{ij} is proportional to the concentrations $S_j(t)$. Growth is modelled using Michaelis-Menten kinetics (Wilkinson, 1998), such that substrate must be first captured by bacteria, and then processed at a finite rate. At high substrate levels, growth becomes limited to a maximum rate G_{ij}^S . At low substrate levels, growth rate is proportional to the amount of substrate according to $S_j(t)G_{ij}^S/K_i^S$, where K_i^S is the Michaelis-Menten factor defining the substrate level for which the growth rate is half the maximum. For every unit of bacterial growth on substrate, an amount E_{ij}^S of substrate is used; this is therefore an ‘inefficiency’ of conversion.

SCFA of type k is produced by some strains of bacteria at a rate O_{ik} for every unit of bacterial growth. Some bacteria can utilise SCFA, and this is modelled according to Michaelis-Menten dynamics with maximum rate G_{ik}^A (referred to as utilisation of SCFA) and Michaelis-Menten factor K_i^A . The growth on SCFA is proportional to the growth already achieved on substrate, and occurs at inefficiency E_{ik}^A . In models of the human gut, SCFA k is absorbed at a rate $R_k A_k(t)$, but experimental fermentors have $R_k = 0$. Flow through the vessel (or compartment) occurs at a specified and known rate $F(t) = F$. Substrate j and SCFA k flow into the fermentor at known rates of FS_j^0 and FA_k^0 respectively. Bacteria i , Substrate j and SCFA k leave the fermentor at a rate of $FB_i(t)$, $FS_j(t)$ and $FA_k(t)$ respectively.

The initial values of the system are known making this an initial value problem. Our assumptions directly relate to the widely used fermentor experimental setup, and are in accord with those frequently made in bacterial gut modelling. Other possible bacterial behaviours are reviewed by Wilkinson (1998). Web Appendix B contains a table of used terms.

This leads to the following system of equations:

$$\frac{dB_i(t)}{dt} = -FB_i(t) + \sum_{j=1}^{n_S} U_S(i, j; t) + \sum_{k=1}^{n_A} U_A(i, k; t) \quad (1)$$

$$\frac{dS_j(t)}{dt} = F[S_j^0 - S_j(t)] - \sum_{i=1}^{n_B} E_{ij}^S U_S(i, j; t) \quad (2)$$

$$\frac{dA_k(t)}{dt} = F[A_k^0 - A_k(t)] - R_k A_k(t) + \sum_{i=1}^{n_B} [P_A(i, k; t) - E_{ik}^A U_A(i, k; t)]. \quad (3)$$

with the intermediate variables:

$$B_{ij}(t) = B_i(t) \frac{S_j(t)}{\sum_{j'=1}^{n_S} S_{j'}(t)} \quad (4)$$

$$U_S(i, j; t) = B_{ij}(t) G_{ij}^S S_j(t) / (S_j(t) + K_i^S) \quad (5)$$

$$U_A(i, k; t) = B_i(t) \left(\sum_{j=1}^{n_S} U_S(i, j; t) \right) G_{ik}^A A_k(t) / (A_k(t) + K_i^A) \quad (6)$$

$$P_A(i, k; t) = O_{ik} \left[\sum_{j=1}^{n_S} U_S(i, j; t) + \sum_{k=1}^{n_A} U_A(i, k; t) \right]. \quad (7)$$

2.3 Inference goals

The human gut is not understood well enough to usefully define a process model for which we can fully infer the parameters. Additionally, no non-trivial single fermentor experiment has uniquely identifiable parameters. Models can be parameterised either by using strong prior information or using data from multiple experiments. In the experiments we can exactly choose and measure the flow rate F , the input substrate S_j^0 , the input SCFA A_k^0 and the host absorption rate R_k (which is zero). These parameters are highly uncertain in the human colon, which additionally has spatial structure and non-continuous flow.

The experiments we consider were run at two pH levels, which were chosen to reflect the acidity of the two main compartments of the human colon. For some bacterial groups, the growth rates G_{ij}^S and inefficiencies E_{ij}^S on substrate are expected to change with pH, and therefore between experiments. Parameters that are assumed unaffected by pH are SCFA usage inefficiency E_{ik}^A , growth rates on SCFA G_{ik}^A , the produced (i.e. output) SCFA O_{ik} , and the Michaelis-Menten factors K_i^S for substrate and K_k^A for SCFA.

As a first, tentative, step towards developing a more realistic model of the gut, we will use the inferred parameter distributions to investigate how absorption of SCFA affects the bacterial populations and SCFA concentrations.

3 Methodology

Consider a system of M related experiments indexed by $a = 1 \dots M$. Observations y_{abc} are made at time points t_{abc} (with $c = 1 \dots m_{ab}$ indexing the m_{ab} observations of variable b in experiment a). The vector θ_a contains the p_a unknown parameters for experiment a .

The statistical model is specified in three hierarchical levels. The bottom level describes the observations, which are modelled as

$$y_{abc} | \theta_a, t_{abc}, \sigma_{abc}^2 \sim N(f_a(\theta_a, t_{abc}), \sigma_{abc}^2), \quad (8)$$

where $f_a(\cdot)$ is defined by Equations 1–7 parameterised for experiment a . The second level of the hierarchy describes the experiment-specific parameter vector θ_a which follows a normal distribution with mean μ (length p) and between-experiment variance Σ ($p \times p$ matrix), modelled as

$$\theta_a | \mu \sim N(X_a \mu, X_a \Sigma X_a^T)$$

The ‘mapping matrix’ X_a is a $p_a \times p$ matrix, with element $(X_a)_{ij}$ defined as 1 if element i of θ_a corresponds to element j of μ , and zero otherwise (hence each row of X_a will contain exactly one 1). Pre- and post-multiplication of Σ by X_a ensures that the appropriate elements of Σ are selected. Web Appendix C shows an example of constructing this matrix.

The top level of the hierarchy describes the prior distribution for μ , specified independently for each element d as a Normal prior:

$$\mu_d \sim N(\beta_d, \tau_d^2)$$

where the prior parameters β_d and τ_d^2 are known constants; details of how we calculated them for the data are in Web Appendix D.

Let Θ denote all parameters, i.e. $\Theta = \{\mu, \theta_1 \dots \theta_M\}$. The posterior distribution is then given as

$$p(\Theta|y) \propto \left\{ \prod_{a,b,c} N(y_{abc}|f_a(\theta_a), \sigma_{abc}^2) \right\} \times \left\{ \prod_a N(\theta_a|X_a\mu, X_a\Sigma X_a^T) \right\} \times \left\{ \prod_d (N(\mu_d|\beta_d, \tau_d^2)) \right\}. \quad (9)$$

As the number of data sets available for estimation of the biological parameters of interest is limited, we treat the within-study variance σ^2 and between-study variance Σ as known.

3.1 Posterior evaluation

A Metropolis-Hastings MCMC sampling (Gamerman, 1997) algorithm is used to evaluate the posterior. A random subset of elements of Θ , denoted by $\Theta_{\{s\}}$, is chosen where each individual element of Θ has probability $p_e = 0.1$ to be included in the subset. The subset is updated as follows: $\Theta'_{\{s\}} \sim N(\Theta_{\{s\}}, \delta_{\{s\}})$ where δ is a pre-defined matrix reflecting the ‘step size’ for each of the elements of Θ (we use a diagonal matrix in practice). Let Θ' denote Θ with subset $\{s\}$ replaced by $\Theta'_{\{s\}}$. The probability of accepting Θ' is based on the usual Metropolis-Hastings ratio given by $\min\left(1, \frac{p(\Theta'|y)}{p(\Theta|y)}\right)$ with $p(\Theta|y)$ given in Equation 9. As usual in MCMC, we discard an initial ‘burn-in’ portion of the chain before evaluating the posterior density. Web Appendix D gives the hyperparameters used and Web Appendix E fully specifies the algorithm. Convergence is assessed by examining the potential scale reduction factor $\sqrt{\hat{R}}$ (Gelman and Rubin, 1992) or effective sample sizes accounting for autocorrelation, calculated with the ‘CODA’ package (Plummer et al., 2008) for R (R Development Core Team, 2008). We refer to a chain as ‘converged’ if for all parameters the effective sample size is greater than 100 and (for the inference on experimental data only) $\sqrt{\hat{R}} < 1.4$ and the 97.5% quantile for $\sqrt{\hat{R}}$ is less than 2.

To address concerns of poor mixing in the low posterior regions, and to perform a limited sensitivity analysis, we additionally consider maximum a-posteriori (MAP) estimates obtained by numerically maximising the posterior starting from multiple random point samples from the prior distribution as described in Web Appendix D.

4 Simulation studies

The simulation studies are based around the fermentor system described in Section 2.2 and aim to reflect data collection as observed in practice. We start with a simple model (Section 4.1) and extend it to a more complex model structure (Section 4.2).

4.1 Inference for one bacterial type

We first simulate data a typical dataset from a fermentor system for the simplest possible scenario. It is assumed that a single type of substrate (starch) is consumed by a single bacterial strain (nominally *Bacteroides*, or ‘Bac’) producing a single SCFA type (Acetate) as is shown in Figure 1 (left). Two fermentor experiments are simulated: Experiment 1 in an acidic environment (pH 5.5) and Experiment 2 in an acid

neutral (pH 6.5) environment. All model parameters are assumed to be the same for both pH values, except for the maximum growth rate, which at pH 5.5 is set at 1/10th that of acid neutral environment. The simulated data, shown in Web Figure 1, (details given in Web Appendix F) consist of the concentration of bacteria and SCFA measured once every 20 hours, for 7 time points. Furthermore, it is assumed that the starch input is known. We will refer to these two experiments as Study 1. Based on the (simulated) measurements of SCFA and bacteria we attempt to infer the Michaelis-Menten factor, the SCFA production rate, the inefficiency of growth, and the maximum growth rate in the low acidity environment. The maximum growth rate in the acid neutral environment is assumed known, as is the variation between (Σ) and within experiments (σ^2).

Two inference scenarios, outlined in Table 1, are considered for Study 1: in Scenario A we consider the simultaneous inference scheme as discussed above. In Scenario B, we first perform inference on the data from the acid neutral experiment alone, summarise the resulting posterior distribution as univariate means and standard deviations (ignoring correlations), and use these results as a prior for analysing the data from the acidic experiment.

Figure 2 illustrates the posterior probabilities of each after 100000 MCMC iterations for Scenario A, and 100000 iterations for each stage in Scenario B. Scenario A allows for strong inference on all parameters, and the chain has converged. However, under Scenario B, the chain converges very poorly (at both stages), resulting in incorrect inference due to the inefficiency of our algorithm to explore the flat posterior mode. This flatness can be understood using the discussion on the Michaelis-Menten factor in Section 2.2. At neutral acidity (pH 6.5) the substrate level is low so that the maximum growth rate is confounded with the Michaelis-Menten factor. At high acidity (pH 5.5) the substrate level is high and the Michaelis-Menten factor does not have much effect on growth rates. For each pH level a wide range of parameter combinations leads to an (approximately) constant likelihood, and only when the data from each of the two pH levels are combined in Scenario A can the Michaelis-Menten factor be estimated. The information gained from each experiment individually is insufficient for standard inference procedures but parameter estimates are extractable via hierarchical modelling.

We also investigated the effect of Σ on the posterior distribution as all parameters simultaneously covary. Web Appendix F shows that there is little difference between parameters inferred with perfect correlation and those inferred with small Σ . As Σ increases above 0.5, the Michaelis-Menten factor in Experiment 1 and the inefficiency in Experiment 2 become increasingly less certain and harder to identify.

4.2 Inference for two bacterial types

To investigate how the method relates to differing data qualities, we apply it to a more realistic problem as outlined in Figure 1 (middle). Two bacterial types compete for a single substrate (starch) and two SCFA types (Acetate and Butyrate) are produced. As in Study 1, the *Bacteroides* bacterial type consumes only substrate to produce Acetate. The second type (nominally *Roseburia rectalis*, or ‘Rrec’) consumes starch and is a producer of Butyrate. Furthermore, it achieves additional growth from consuming Acetate. Experiment 3 is performed at pH 5.5 and experiment 4 is performed at pH 6.5, i.e. the same pH values as in Study 1. These two experiments are referred to as Study 2, and the simulated data points and general behaviour of this model are shown in Web Figure 2. Four inference scenarios, summarised in Table 1, are considered. In Scenario C a strong prior for inferred values (Normal with correct mean and standard deviation of 0.1 of the mean) is used to infer for Study 2. The other scenarios attempt to manage with a *uniform* prior on parameters. In Scenario D simultaneous inference is performed on the simulated data (consisting of substrate, SCFA and bacterial counts) from both Study 1 and 2. In Scenario E we do not measure substrate output as this is not commonly available. All were run for 200000 MCMC steps. The 10 parameters to be inferred are: Michaelis-Menten factor for growth of Bac on starch, Michaelis-Menten factor for Rrec for growth on starch and acetate (assumed the same), the growth rates on starch of both Bac and Rrec bacterial strains in the low pH environment, their inefficiencies on starch, Bac production

of Acetate, Rrec production of Butyrate, Rrec inefficiency of Acetate usage and utilisation of Acetate (for extra growth of Rrec on SCFA).

Convergence of the MCMC chain is acceptable for Scenario C and D. Scenario E led to poor MCMC convergence with 5 parameters having effective sample sizes of less than 50 (minimum 11). We may therefore need 20 times (or more) the number of samples in this chain to achieve the same convergence as Scenario C. Figure 3 shows box-plots for each parameter under each Scenario. Scenarios C and D have roughly equivalent posterior distributions.

Examination of the posterior distribution shows that in Scenario E there is a lack of identifiability between the Rrec SCFA inefficiency (i.e. use of Acetate for additional growth) and Bac Acetate production.

Only scenarios C and D provide useful one-dimensional summaries for further analysis, indicating that it is necessary to assume a strong prior on parameter values or to use sufficiently diverse experimental conditions in order to avoid identifiability problems during inference.

5 Inference for microbiology data

Having demonstrated that despite limited data, our approach is capable of recovering posterior information on parameters of interest, we will now continue to analyse some real data of (Walker et al., 2005) analogously to Scenario C. These data were obtained from two experiments run under neutral (pH 6.5) or acidic (pH 5.5) conditions. Each fermentor received two substrates, namely starch and NSP. Bacterial growth is thought to be considerably slower on the less degradable NSP than on starch. In addition to Bac and Rrec, we will also consider a group of acetate producing bacteria, AcProd, that are thought to be more resistant to an acidic environment than the Bac group. The relationships between bacteria, substrates and SCFA are illustrated in Figure 1 (right). Our choice of bacterial and SCFA groupings provide a reasonable representation of the data, with approximately 75% of the total bacteria and total SCFA covered; see Table 3 of (Walker et al., 2005).

Several parameters are thought to be influenced by pH or type of substrate, and need to be specified accordingly. There are now 29 parameters to be inferred in total, specified below (number of parameters in parentheses). Previous data indicate that Rrec grow just as well at high substrate levels at both pH levels considered, so we infer their maximum growth rate on starch and NSP as constant between experiments (2), and similarly for the inefficiency of starch and NSP usage (2). The 'AcProd' community is not well defined and so the specific bacteria involved might differ with acidity; they are parameterised by their growth rates (4) and efficiencies (4). Similarly the Bacteroides group are known a-priori to have different growth rates (4) and efficiencies (4). The remaining parameters are all constant between experiments: Michaelis-Menten factors for growth on substrate (3), Rrec Michaelis-Menten factor for growth on acetate (1), Acetate production by Bacteroides and Acetate Producers (2), Acetate utilisation of Rrec and its efficiency (2), and Butyrate production of Rrec (1).

The prior distributions for these parameters are reasonably informative, and are based on knowledge on metabolic pathways as well as results from studies on single bacterial strains *in vitro*. Growth on NSP is less well-understood than on starch, and the priors are suitably weaker. It is assumed that the variation within experiments σ_{abc}^2 (see Walker et al., 2005, Table 2 and 3) and variation between experiments Σ are known, based on data from a repeated experiment with bacteria seeded by a second human donor. Web Appendix D gives details for parameter values and a sensitivity analysis for Σ ; additional implementation details are Web Appendix E.

In practice, it is not easy to choose a suitably simple model and set of parameters for inference. The current model was obtained after an iterative process of model simplification and testing, using a similar analysis to Figure 3, by comparing MCMC posterior distributions. This involved close collaboration between modellers and microbiologists to ensure that the main characteristics of the microbial processes were still represented in the model. Convergence was satisfactory and the results are summarised in Figure 4. The growth rate on starch in acidic conditions of the AcProd group differed strongly from our prior, indicating that the AcProd group is indeed diverse and the composition at different pH changes. Additionally, the production of Acetate by the Bac group did not meet our prior expectation based on efficient conversion and therefore our prior needs to account for inefficiency.

The posterior prediction for the growth curves is given by Figure 5. This model is a simple attempt to describe bacterial ecology in an inference setting and therefore should be considered cautiously; particularly, not all bacterial data points are simultaneously satisfied by the model predictions, perhaps implying that some important feature is omitted from the model.

Biological interpretation of our results must be cautious as the model is simple and the prior used is strong. Despite this, our findings are biologically plausible. The inefficiency levels show that a higher level of cell maintenance is necessary for both Bacteroides and other Acetate Producers under acidic conditions. Rrec utilisation of Acetate appears to come at an efficiency cost for use of substrate, as energy is required to do this. However, it may be necessary to model the metabolic pathways in more detail to generalise to Acetate-poor conditions. Under starvation conditions Bac outperforms the other groups as shown from its Michaelis-Menten factor. Rrec finds Acetate difficult to obtain at low densities which is probably due to

SCFA being free in the medium whereas cells can adhere to substrate (and therefore the Michaelis-Menten factor for substrate is smaller). As expected, the maximum growth rate for *Bacteroides* is lower at high pH, and it does not appear to grow on NSP. This may be due to phenotypic plasticity allowing specialisation to the more abundant starch.

The continuous fermentor environment provides a highly simplified experimental model of the human gut. One of these simplifications is that absorption of metabolites is not allowed for. In our mathematical model formulation, however, it is easy to relax this assumption and Table 2 shows what happens to the SCFA and bacteria when the fractional absorption rate $R_k \sim \text{Normal}(0.72, 0.072^2)g/d$, a realistic value representing high SCFA concentrations (Jr. et al., 1977) with 10% uncertainty. Fecal Butyrate and Rrec are significantly reduced at all pH values. Similarly, we can run our model under 'high fibre' conditions with input consisting of 5.0g/d starch (i.e. 0.6g/d less), 2.4g/d NSP (i.e. 0.6g/d more) and flow rate of 1 turnover in 18 hours (compared to 1 turnover per 24 hours). This allows us to compare the host absorption of SCFA: Acetate decreases from 1.24g/d (95% confidence interval, $CI = 0.98 - 1.66$) by a factor of 0.53 (95% $CI = 0.41 - 0.72$) and Butyrate increases from 0.63g/d (95% $CI = 0.37 - 0.93$) by a factor of 1.68 (95% $CI = 1.26 - 2.72$). Although these findings should be interpreted with caution, the results clearly show how absorption of metabolites results in a bacterial shift in favour of Acetate Producers which comes at the cost of *Bacteroides*. A high fibre diet results in increased Butyrate and decreased Acetate absorption for the human host.

As an additional check we have performed maximum a posteriori (MAP) estimation from starting conditions sampled from the prior using standard numerical maximisation techniques, which confirms that our estimate is indeed a local maxima (See Figure 4). Additionally we performed a sensitivity analysis using the MAP procedure on the variation within experiments Σ as this is not well known. The MAP estimates change by less than 1% when the magnitude of Σ is scaled by a factor of 2 each way (Web Appendix D).

6 Discussion

Despite a limited amount of data, we have shown how integration of knowledge on the underlying model structure in combination with informative prior information on model parameters allows for inference of model parameters of interest. Like a prior distribution for a set of parameters, the functional form of the model (which forms the summary of our knowledge of the process) should be updated if it does not agree with the data. In this sense, we view our process model as a type of prior, providing restrictions in the parameter space.

Although our model assumptions will likely be quantitatively incorrect, they do allow us to investigate what would happen (in a relative sense) if some of the experimental conditions are relaxed. Relaxations such as allowing for absorption of metabolites are straightforward to incorporate in the mathematical model but are not easily implemented in experimental studies. The approach discussed here therefore provides a potentially powerful tool for microbiologists to explore such 'what if' scenarios. Our predictions are consistent with observations of human colons, adding to the validity of using experimental fermentors and theoretical models to understand gut bacterial ecology.

Data collection and analysis is an expensive exercise, even for the relatively simple fermentor studies used here. Our findings indicate that more time samples should be analysed for bacterial composition (at present only at start and end of experiment), and replication of fermentors should be considered to allow for inferring a posterior distribution of the variance components.

Bayesian hierarchical modelling of process model parameters is a powerful and under-utilised tool for inference. Our results show that this methodology can describe sparse data with a process model incorporating a large amount of knowledge of the structure of the data. Because we focus on methods that treat the process model as a black box, they are applicable to many other problems.

Acknowledgements We are grateful to Mark Brewer and Glenn Marion for detailed comments and to Tony Travis with assistance with using the High Performance Computing cluster. This work was funded by the Scottish Government Rural and Environment Research and Analysis Directorate (RERAD).

Conflict of Interest

The authors have declared no conflict of interest.

Supplementary Materials

Web Appendices A-F and Web Figures 1-2 referenced in Sections 2-5 are available online at the Biometrical Journal website. These provide a more explicit description and specific details required for replication of our result as discussed in the text.

References

Banks, H. T., S. Grove, S. Hu, and Y. Ma (2005). A hierarchical bayesian approach for parameter estimation in hiv models. *Inverse Problems* 21, 18031822.

Beretta, E. and Y. Kuang (2000). Global stability in a well known delayed chemostat model. *Communications in Applied Analysis* 4, 147–155.

Coleman, M. E., D. W. Dreesen, and R. G. Wiegert (1996). A simulation of microbial competition in the human colonic ecosystem. *Applied and Environmental Microbiology* 62, 3632–3639.

Cummings, J. H. (1997). The large intestine in nutrition and disease. Technical report, Institut Danone. http://www.danoneinstitute.org/publications/book/pdf/large_intestine.pdf.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. 2-6 Boundary Row, London, UK. SE1 8HN: Chapman and Hall.

Gelman, A., F. Bois, and J. Jiang (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Stat. Ass.* 91, 1400–1412.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457–511.

Huang, Y., D. Liu, , and H. Wu (2006). Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics* 62, 413423.

Jr., M. G. S., K. H. Soergel, C. M. Wood, and J. J. Steff (1977). Absorption of short-chain fatty acids from the human ileum. *Digestive Diseases and Sciences* 22, 340–347.

Kreft, J.-U., G. Booth, and J. W. T. Wimpenny (1998). Bacsim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology* 144, 3275–3287.

Levi-Montalcini, R. and P. Calissano (2006). The scientific challenge of the 21st century: from a reductionist to a holistic approach via systems biology. *BMC Neuroscience* 7, S1.

Louis, P., K. Scott, S. Duncan, and H. Flint (2007). Understanding the effects of diet on bacterial metabolism in the large intestine. *J. Applied Microbiol.* 102, 1197–1208.

Lunn, D. J. and L. J. Aarons (1997). Markov Chain Monte Carlo techniques for studying interoccasion and intersubject variability: Application to pharmacokinetic data. *Appl. Statist.* 46, 73–91.

Mezzetti, M., J. G. Ibrahim, F. Y. Bois, L. M. Ryan, L. Ngo, and T. J. Smith (2003). A Bayesian compartmental model for the evaluation of 1,3-butadiene metabolism. *Appl. Statist.* 52, 291–305.

Plummer, M., N. Best, K. Cowles, and K. Vines (2008). *CODA: Output analysis and diagnostics for MCMC*. R package version 0.13-2.

Putter, H., S. H. Heisterkamp, J. M. A. Lange, and F. de Wolf (2002). A bayesian approach to parameter estimation in hiv dynamical models. *Statistics in Medicine* 21, 21992214.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Wakefield, J. (1996). The bayesian analysis of population pharmacokinetic models. *J. Am. Stat. Ass.* 91, 62–75.

Walker, A. W., S. H. Duncan, E. C. M. Leitch, M. W. Child, and H. J. Flint (2005). pH and peptide supply can radically alter bacterial populations and short-chain fatty acid ratios within microbial communities from the human colon. *Applied and Environmental Microbiology* 71, 3692–3700.

Wilkinson, M. H. (1998). Model intestinal microflora in computer simulation (MIMICS) technical report: Ordinary Differential Equations for modelling bacterial interactions in the gut. Technical report, Centre for High Performance Computing, University of Groningen.

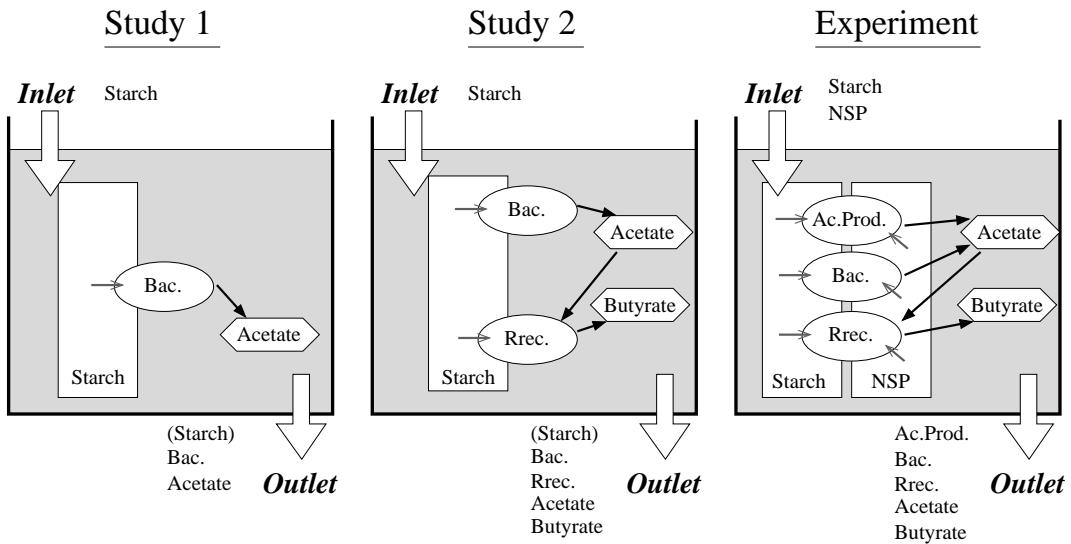


Figure 1 Flow diagrams for the fermentor models used in Sections 4 and 5. Substrate types (starch and NSP) are shown in squares. Bacterial types are shown in ellipses and are: Ac.Prod. (mixed acetate producers), Bac. (*Bacteroides*) and Rrec. (*Roseburia Rectalis*). Bacteria overlap substrate because they are assigned to substrates proportional to the substrate abundance. SCFA (Acetate and Butyrate) are shown in stretched hexagons. Input and output substances are listed; the brackets around starch indicated that this is not available in all cases.

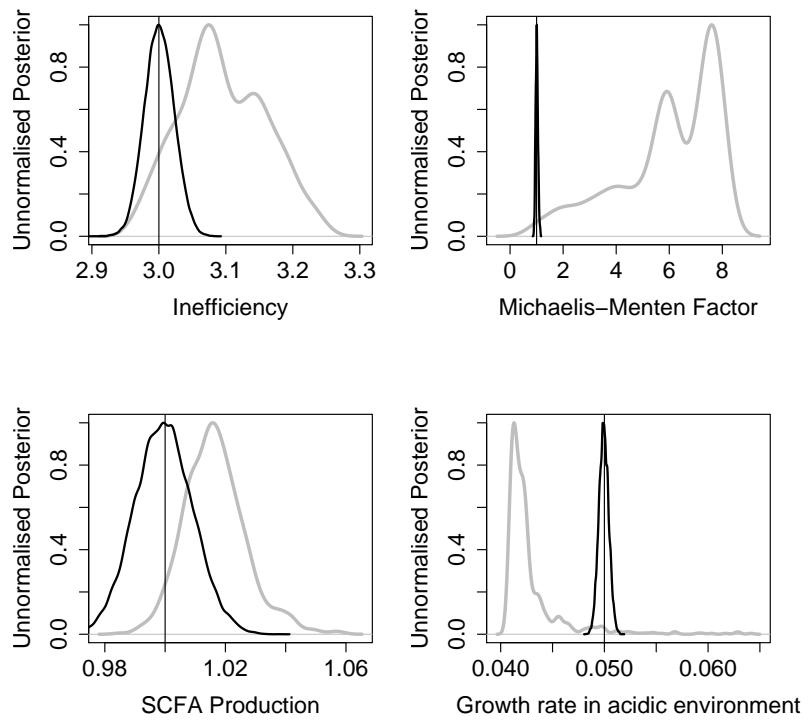


Figure 2 Densities for the posterior probability of parameters using simulated data (Study 1, single bacterial strain) for (black line) Scenario A using simultaneous inference, and (grey line) Scenario B using summary information and separate inference. The true parameter values are indicated as a vertical line. Scenario A has a converged MCMC chain (> 200 effective samples per parameter) whereas no parameters have a large effective sample size in Scenario B.

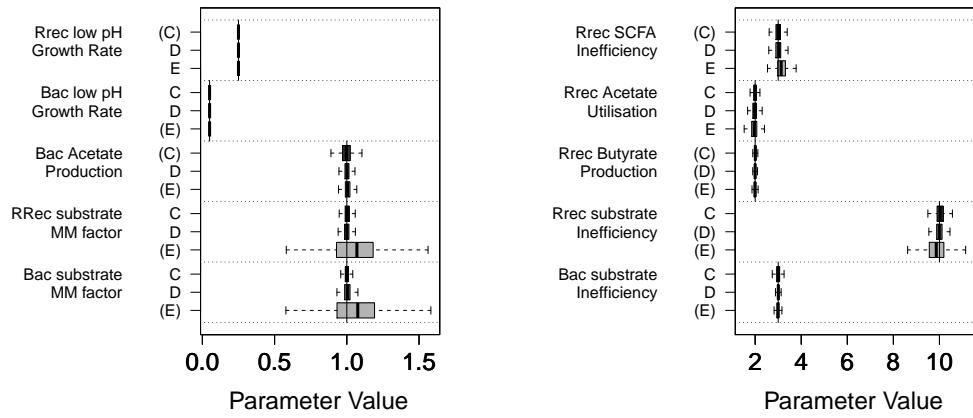


Figure 3 Posterior probability distribution for the parameters in Study 2 for the two bacterial strains experiments with true values shown as a vertical line. Each horizontal segment contains three box-plots, with Scenarios C, D, E running from top to bottom. ‘MM factor’ is the Michaelis-Menten factor. Scenario labels C and D are in brackets if the effective sample size is the range 100 – 200 (no brackets: > 200); Scenarios E is in brackets if the sample size is in the range 11 – 200.

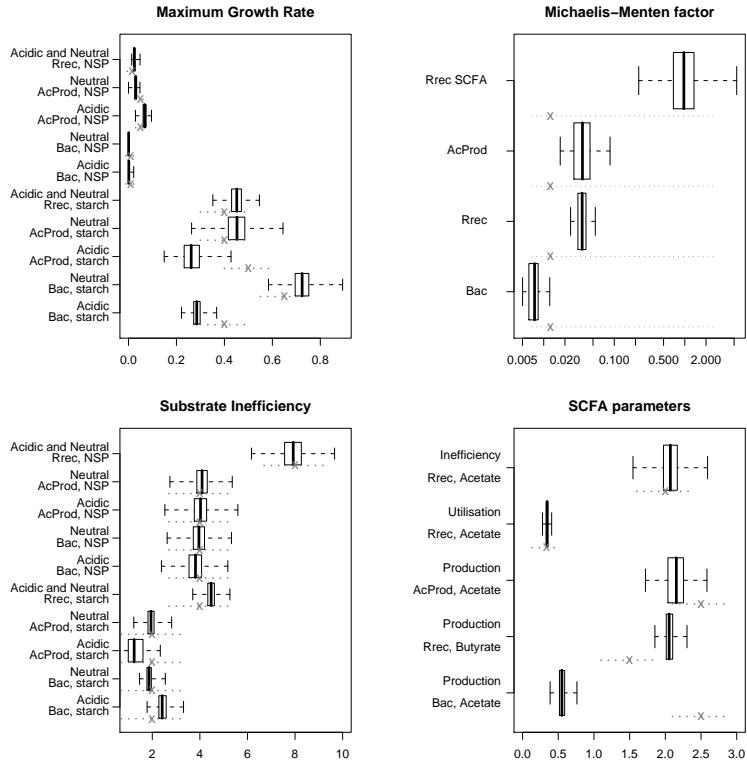


Figure 4 Posterior probabilities for the Walker et al. (2005) data, showing all parameters as a box-plot averaged over 3 separate MCMC chains. Grey crosses show prior means and grey dotted lines show twice the prior standard deviation (see Web Appendix D). The median value is the thick line, with the box representing the 50% confidence region and the dashed lines extend to the whole sample. The MAP estimate is the open circle. ‘Acidic’ and ‘Neutral’ refer to the pH 5.5 and pH 6.5 experiments respectively; see text for an interpretation of these results.

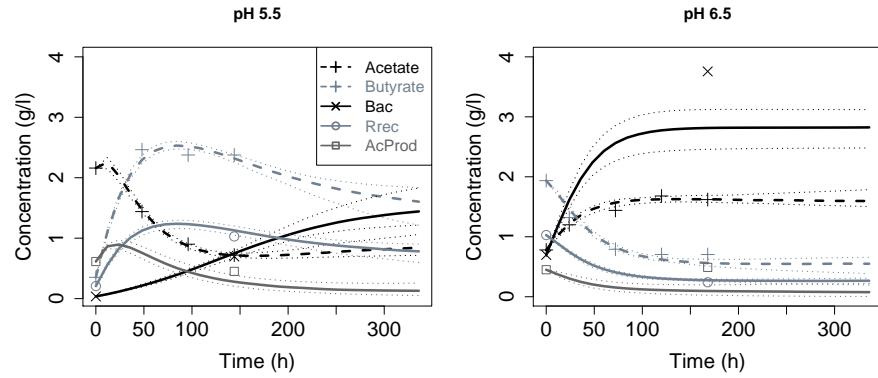


Figure 5 Prediction for the model based on parameter samples from the converged MCMC chain based on the Walker et al. (2005) data for (Left) pH 5.5 and (Right) pH 6.5. Shown are the median predicted value for bacteria (solid thick line), and for SCFA (broken line) along with the 95% credible intervals (thin dotted lines). Observed data for SCFA and for bacteria are shown from Web Appendix B. The model has all variables matching the data as initial conditions, and is run past the final data point to establish long term trends.

Table 1 Summary of scenarios considered for inference on data from simulation study 1 (one substrate, one bacterial strain and one SCFA), simulation study 2 (one substrate, two bacterial strains, two SCFA), and experimental data from Walker et al. (2005). Yes and No indicate whether a variable was measured, while - indicates that a variable did not form part of the models. Uniform priors on inferred parameters were used unless specified.

Scenario	Study	Bacteria			SCFA		Substrate		Analysis
		Bac	Rrec	AcProd	Ac.	Bu.	Starch	NSP	
A	1	Yes	-	-	Yes	-	Yes	-	Simultaneous inference of data from both pH.
B	1	Yes	-	-	Yes	-	Yes	-	Step 1: inference on data from pH 6.5 experiment. Step 2: use summary of posterior from Step 1 as prior for inference on data from pH 5.5.
C	2	Yes	Yes	-	Yes	Yes	Yes	-	Strong prior on inferred parameters; simultaneous inference of data from both pH.
D	1	Yes	-	-	Yes	-	Yes	-	Simultaneous inference of data from both studies at both pH.
D	2	Yes	Yes	-	Yes	Yes	Yes	-	Simultaneous inference of data from both studies at both pH.
E	1	Yes	-	-	Yes	-	No	-	Simultaneous inference of data from both studies at both pH.
E	2	Yes	Yes	-	Yes	Yes	No	-	Simultaneous inference of data from both studies at both pH.
Data	Walker	Yes	Yes	Yes	Yes	Yes	No	No	Simulated inference of data from both pH. Strong prior on inferred parameters.

Table 2 Predictions for total relative abundances at equilibrium under the assumption of host SCFA absorption, relative to the full posterior for the no absorption model from Figure 4 & 5.

pH	Quantity	Effect	Quantile		
			50%	2.5%	97.5%
5.5	Acetate	-	0.64	0.90	1.07
5.5	Butyrate	Down	0.24	0.42	0.54
5.5	Bac.	Up	1.06	1.18	1.33
5.5	Rrec.	Down	0.43	0.72	0.92
5.5	AcProd.	-	0.80	1.72	2.28
6.5	Acetate	Down	0.55	0.63	0.75
6.5	Butyrate	Down	0.22	0.44	0.60
6.5	Bac.	-	1.00	1.01	1.03
6.5	Rrec.	Down	0.40	0.75	0.98
6.5	AcProd.	-	0.03	1.43	2.14