Statistical Modeling of Biochemical Pathways

Robert B. Burrows*†, Gregory Warnes‡[1], and R. Choudary Hanumara§

March 1, 2007

$Id: paper.Snw 1039 2006-12-06 19:20:46Z warnes $

†New England Biometrics, North Scituate, RI, rbb@nebiometrics.com

‡Pfizer, Inc., Groton, CT, gregory.r.warnes@pfizer.com

§University of Rhode Island, Kingston, RI, rch@cs.uri.edu

∗Corresponding author. Email: rbb@nebiometrics.com

We have examined the usefulness of Bayesian statistical methods for the modeling of biochemical reactions. With simulated data, it is shown that these methods can effectively fit mechanistic models of sequences of enzymatic reactions to experimental data. These methods have the advantages of being relatively easy to use and producing probability distributions for the model parameters rather than point estimates, allowing more informative inferences to be drawn.

Three Markov chain Monte Carlo algorithms are used to fit models to data from a sequence of 4 enzymatic reactions. The algorithms are evaluated with respect to the goodness-of-fit of the fitted models and the time to completion. It is shown that the algorithms produce essentially the same parameter distributions but the time to completion varies.

1 Introduction

$Id: introduction.Snw 1039 2006-12-06 19:20:46Z warnes $

We know a great deal about the individual components of living organisms but much less about the properties of the systems of which they are a part. Our ignorance in this regard prevents us from developing a deeper theoretical understanding of living systems and from developing more effective methods for altering the system behavior in beneficial ways.

In order to gain some insight into the properties of biological systems, we would like to use quantitative models that describe the processes at the molecular level. Specifically, we would like to develop methods that can be used

[1]current address: Dept. of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, gregory_warnes@urmc.rochester.edu

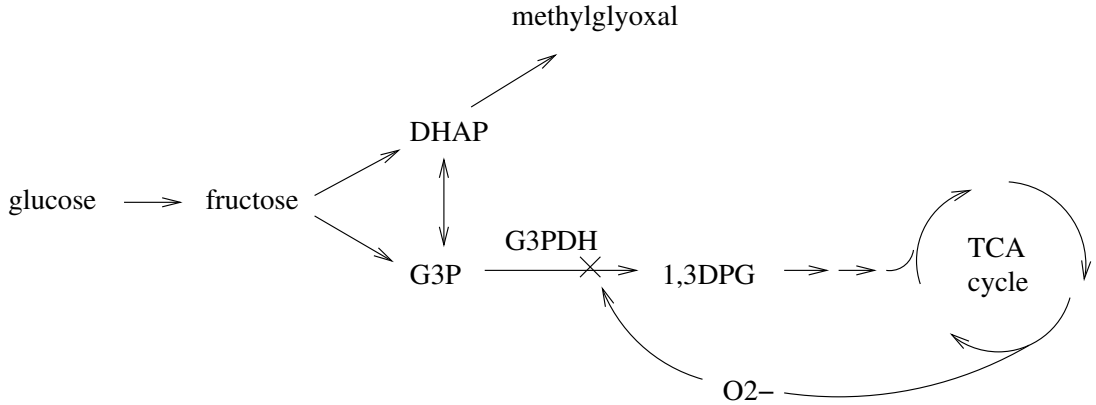


Figure 1: Production of methylglyoxal in hyperglycemia

to deduce the kinetic parameters of enzyme-catalyzed reactions. With this information, it should be possible to identify the most effective control points in a pathway and modify the pathway behavior in predictable ways.

An example of how a model of a biochemical pathway could be used is in the elucidation of a mechanism for the pathology seen with diabetes. There are thought to be four mechanisms that explain hyperglycemic toxicity in diabetes [1, 2]. One mechanism is illustrated in Figure 1.

Some of the damage to tissue components seen in diabetes is thought to be the result of reactions with glycolytic intermediates such as methylglyoxal. Nishikawa et al. [3] have hypothesized that this is caused by the increased catabolism of glucose in diabetes which leads to an increase of superoxide anion (O_2^-). Superoxide is known to inhibit the enzyme glyceraldehyde-3-phosphate dehydrogenase (G3PDH) by about 50% under these conditions. The inhibition of this enzyme then increases the concentration of glyceraldehyde-3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP) which breaks down to form methylglyoxal, an active aldehyde which readily reacts with and inactivates proteins. To test this hypothesis we need to be able to answer questions such as: Is the observed inhibition of glyceraldehyde-3-phosphate dehydrogenase by 50% sufficient to explain the observed increase in methylglyoxal concentration? If not, then what other mechanisms might be operative? What is the most effective way to modulate these reactions? We address these questions by developing analytical methods that will allow us to characterize these types of biochemical systems at the molecular level.

We have used a Bayesian modeling approach with fitting via Markov chain Monte Carlo (MCMC) methods [4, 5] to estimate the kinetic parameters of the enzymes in a metabolic pathway. With this approach we can combine known information about the pathway with newly-generated experimental data to produce the joint probability density of all of the parameters. This joint probability density can then be used to answer various questions about the behavior of the system.

We have explored the utility of the the Bayesian approach using a simulation technique: First we generated data for hypothetical sequences of enzymatic reactions, and then we fit models to this data using Bayesian methods. The biochemical data is generated using Gillespie's stochastic simulation algorithm [6]. This algorithm is based on well-established chemical kinetic theory and thus has a firmer physical basis than alternative methods based

on differential equations. It also naturally incorporates the fluctuations that result from the probabilistic nature of chemical reactions. The algorithm has been validated with data from many chemical reactions and has been been used in many biochemical simulators, e.g., [7, 8]. While the algorithm appears to give good results with biochemical reactions, it should be borne in mind that the underlying kinetic theory assumes well-mixed, freely diffusing reactants, an assumption that will not always be valid with biochemical systems.

We simulate perturbation experiments. In these experiments, the concentration of the first reactant in a pathway is transiently increased and the concentrations of all the reactants is monitored as they return to a steady state. This type of experiment is quite feasible with cultured animal cells or suspensions of microbial cells such as yeast, and it yields information on the behavior of the pathway as a single system.

MCMC simulation was found to be a very useful method for fitting systems of equations to data. It is capable of fitting several equations simultaneously, and does not require an excessive number of data points. A very attractive feature of the method is that it produces realistic estimates of the variance and covariance of the parameters which are useful for inference.

2 Methods

`$Id: methods.Snw 1039 2006-12-06 19:20:46Z warnes $`

2.1 The Pathway

While we examined several pathway structures, in this paper we present the results for a 5-reaction sequence that is sampled at 12, 16, or 25 time points with 3 replicates at each time point. The units for time and concentration are arbitrary; for the purpose of illustration we will use minutes as the unit of time and micromoles/liter (μM) as the unit of concentration. The sequence is

$$source \longrightarrow R1 \longrightarrow R2 \longrightarrow R3 \longrightarrow R4 \longrightarrow R5 \longrightarrow sink \tag{1}$$

and the the equations that were used in the Gillespie simulatiuon are

$$\begin{aligned}
\text{source} &\xrightarrow{k_1} \text{R1} \\
\text{R1} + \text{E1} &\underset{k_3}{\overset{k_2}{\rightleftharpoons}} \text{R1E1} \xrightarrow{k_4} \text{R2} + \text{E1} \\
\text{R2} + \text{E2} &\underset{k_6}{\overset{k_5}{\rightleftharpoons}} \text{R2E2} \xrightarrow{k_7} \text{R3} + \text{E2} \\
\text{R3} + \text{E3} &\underset{k_9}{\overset{k_8}{\rightleftharpoons}} \text{R3E3} \xrightarrow{k_{10}} \text{R4} + \text{E3} \\
\text{R4} + \text{E4} &\underset{k_{12}}{\overset{k_{11}}{\rightleftharpoons}} \text{R4E4} \xrightarrow{k_{13}} \text{R5} + \text{E4} \\
\text{R5} &\xrightarrow{k_{14}} \text{sink}
\end{aligned}$$

2.2 The Experiment

To estimate the kinetic parameters of the enzymes in a sequence of reactions, we need to find expressions of the form

$$\text{reaction velocity} = f([substrate], [enzyme])$$

These expressions are found by perturbing a reaction network and observing the metabolite concentrations as the system relaxes back to a steady state. Specifically, the Gillespie simulation was run with the initial values for the reactant concentrations until a steady state was achieved. The steady state appeared to have been reached by *time* = 15 min. At *time* = 20 min the concentration of $R1$ was increased to 10000 μM. At the indicated time points the metabolite concentrations are measured, and each reaction velocity is calculated as the slope of the curve of metabolite concentration as a function of time.

2.3 The data

For the 5-reaction sequence of reactions the perturbation of $R1$ at $time = 20$ results in the time courses plotted in Figure 2. For each time point there are 5 values for the reactant concentrations and 5 values for the estimated reaction rates. The data at each time point was generated by sampling 3 values for each reactant from a normal distribution centered on the Gillespie output and with a standard deviation of 50 μM, i.e., the measurement error is 50 μM. Three sets of time points containing 12,16, and 25 points were used. The reaction velocities were estimated with the *smooth.spline()* function in *R* [9].

Note that we are dealing with two different types of distribution here. One is the distribution of data values that arises from experimental error and the other is the distribution of model parameter values that are generated by MCMC simulation. The uncertainty in the parameter values is reduced by increasing the number of data points (Figure 5) and by decreasing experimental error but in general there is not a closed form solution for these relationships.

2.4 The Biochemical Model

Individual reactions were fit using the Michaelis-Menten equation [10] for individual enzymes. This is a reaction of the form

$$E + S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_3} E + P \tag{2}$$

where

$$\begin{aligned} S &= \text{substrate concentration} \\ E &= \text{free enzyme concentration} \\ ES &= \text{concentration of the enzyme-substrate complex} \\ P &= \text{product concentration} \end{aligned}$$

It has been shown [11] that in a steady state the rate of the reaction v is

$$v = \frac{V_{max} S}{K_m + S} \tag{3}$$

where

$$\begin{aligned} V_{max} &= (E + ES)k_3 = \text{maximum reaction velocity} \\ K_m &= \frac{k_2 + k_3}{k_1} = \text{substrate concentration at half-maximal velocity} \end{aligned}$$

This form of the equation is very useful because v and S are usually measureable and V_{max} and K_m can be obtained by fitting equation 3 to the data. In contrast, modeling v as a function of the individual rate constants is less useful because that requires the measurement of the concentration of the enzyme-substrate complex which is technically difficult.

In this application we are dealing with sequences of reactions which are not in a steady state, so we cannot use the Michaelis-Menten equation directly. Instead, we use equations of the form

$$v = \frac{aS}{b + S} - \frac{cP}{d + P} \tag{4}$$

The coefficients a, b, c, and d in the equations can be estimated with data obtained following a change in the concentration of one of the reactants. For example, four equations can be fit to the data plotted in Figure 2:

$$\frac{\mathrm{d}R2}{\mathrm{d}t} = v_2 = \frac{d_1 R1}{d_2 + R1} - \frac{d_3 R2}{d_4 + R2}$$

$$\frac{\mathrm{d}R3}{\mathrm{d}t} = v_3 = \frac{d_3 R2}{d_4 + R2} - \frac{d_5 R3}{d_6 + R3}$$

$$\frac{\mathrm{d}R4}{\mathrm{d}t} = v_4 = \frac{d_5 R3}{d_6 + R3} - \frac{d_7 R4}{d_8 + R4}$$

$$\frac{\mathrm{d}R5}{\mathrm{d}t} = v_5 = \frac{d_7 R4}{d_8 + R4} - d_9 R5$$

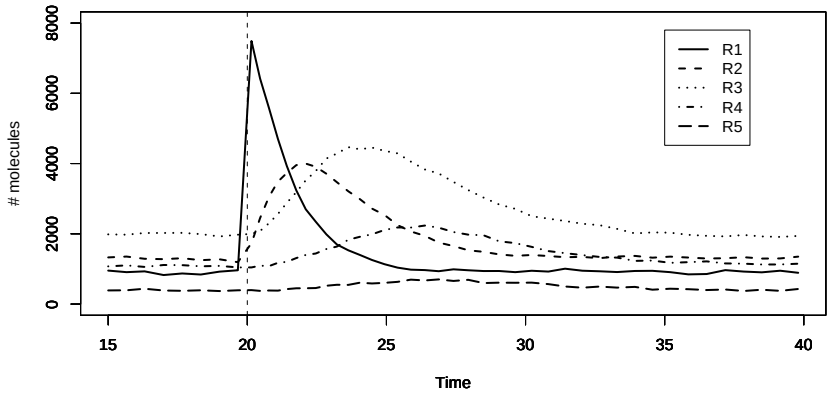


Figure 2: Reactant concentrations following a pulse of R1 at $time = 20$ for the sequence of reactions $R1 \rightarrow R2 \rightarrow R3 \rightarrow R4 \rightarrow R5 \rightarrow sink$.

2.5 Statistical Models

The statistical model of the parameters is the product of a prior distribution and a likelihood distribution, where the prior expresses our current estimate for the parameter values and the likelihood is the probability of the parameters given the data.

The parameters to be estimated, i.e., the modes and the standard deviation of the rate constant distributions, are necessarily non-negative and have maximum values that are bounded by the physical nature of the systems. We model this using statistical model for each parameter of the form

$$\frac{3d_i}{\mu_i} \sim \chi^2_5$$

where the μ_i are the estimates of the parameter values before any data are collected. With this distribution, the d_is will have a mode of μ_i and large values will be unlikely. For this situation, we selected $\mu_i \equiv 50$ units for the rate constants and $\sigma^2 \equiv 2$ for the data standard deviation of the estimates.

The likelihood is the probability of the estimated reaction velocities v_j given the model. The error in the velocity estimates is assumed to be $N(0,\sigma^2)$ so that the reaction velocities v_j have a normal distribution:

$$v_j \sim N(\mu_j, \sigma^2)$$

where

$$\mu_j = \frac{aS_j}{b+S_j} - \frac{cP_j}{d+P_j}$$

The observed data consist of values for v_j, S_j, and P_j. For example, for the 5-reaction model we have

$$v_2 \sim N\left(\frac{d_1 R1}{d_2 + R1} - \frac{d_3 R2}{d_4 + R2},\ \sigma^2\right)$$
$$v_3 \sim N\left(\frac{d_3 R2}{d_4 + R2} - \frac{d_5 R3}{d_6 + R3},\ \sigma^2\right)$$
$$v_4 \sim N\left(\frac{d_5 R3}{d_6 + R3} - \frac{d_7 R4}{d_8 + R4},\ \sigma^2\right)$$

and

$$v_5 \sim N\left(\frac{d_7 R4}{d_8 + R4} - d_9 R5,\ \sigma^2\right)$$

In this case there are 10 parameters to be estimated: the 9 coefficients $d_1 - d_9$ and σ^2.

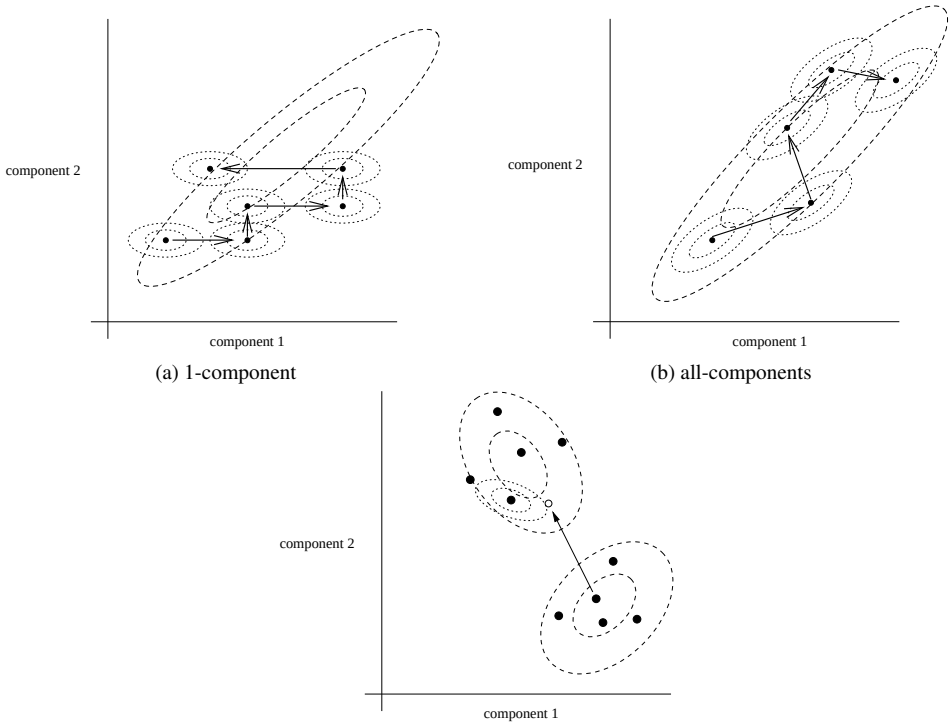
2.5.1 MCMC sampling algorithms

The efficiency of a MCMC simulation [4, 12] depends heavily on the method used to find candidate points to add to the Markov chain [$q(\cdot|\cdot)$]. For this reason different methods have been developed to increase the efficiency of sampling from the sample space. We have used three algorithms from the Hydra library [13]: component-wise Metropolis, all-components Metropolis, and Normal Kernel Coupler [14].

The component-wise Metropolis and the all-components Metropolis algorithms both operate on single Markov chains. The component-wise algorithm (Figure 3a) generates candidate points by changing the value of only one component (dimension) of the current state at each iteration by sampling from a univariate normal distribution centered at the current point.

The all-components algorithm (Figure 3b) changes all the components simultaneously by sampling from a multivariate normal distribution centered at the current point. The component-wise Metropolis algorithm has the advantage of simplicity but may move very slowly if the components are highly correlated. The all-components Metropolis avoids the problems with correlation, if an appropriate covariance matrix is supplied, by updating all components simultaneously but may perform poorly in high dimensions. A problem with both of these algorithms is that if the component distributions have two or more modes separated by areas of low probability, the simulator can get stuck in the vicinity of one mode and fail to visit the other. The NKC algorithm is designed to avoid this difficulty.

The Normal Kernel Coupler (NKC) algorithm operates on multiple chains, so that there are several "current" states. The NKC uses a normal kernel density estimate as the proposal distribution for choosing candidate states. The estimate is generated using the entire set of current points. Since the algorithm uses the entire set of current points it can move over areas of low probability in the parameter space, especially if the user takes care to seed each mode with a few points in the starting set.



(a) 1-component (b) all-components

(c) NKC

Figure 3: Movement of Markov chains with the component-wise and all-components Metropolis algorithms. Movement in a 2-dimensional space is illustrated, so each point has 2 components. The dotted lines are contours of equal probability density for the proposal distributions and the dashed lines are probabilty contours of the target distribution.

3 Convergence

The MCMC simulations are run until the Markov chains have reached stable distributions as assessed by the *mcgibbsit()* algorithm [14]. *mcgibbsit()* calculates the number of iterations necessary to estimate a user-specfied quantile q to within $\pm r$ with probability s, i.e., *mcgibbsit()* indicates when the MCMC sampler has run long enough to provide good confidence interval estimates. The defaults, which are used in this paper, are $q = 0.025$, $r = 0.0125$, and $s = 0.95$. These values generate estimates of the 2.5% and 97.5% quantiles to $\pm$ 0.0125 quantiles with probability 0.95.

4 Results

`$Id: results.Snw 1039 2006-12-06 19:20:46Z warnes $`

All three algorithms converged to similar distributions and produced essentially the same metabolic pathway model. The marginal distributions for the parameters were mound-shaped (Figure 4), were skewed to the right and had thicker tails than normal distributions.

The effect of the number of data points on the parameter distributions can be seen in Figure 5. We see some improved precision as the number of points increases from 16 to 25 but it is not pronounced. Overall, the reduction in width varied from 18-fold for d_9 and 9-fold for d_1 to 1.5-fold for d_4. Figure 6 is an example of a bivariate

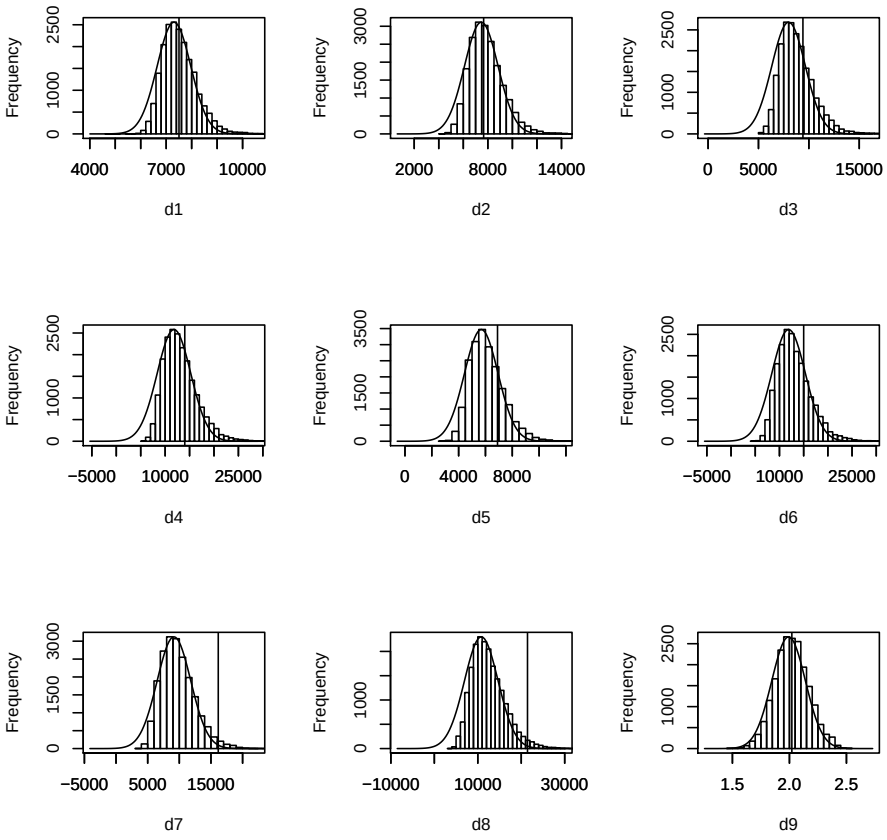


Figure 4: Histograms of the marginal probability distributions for the 5-reaction model generated with the all-components Metropolis algorithm and the 16-point dataset. The curves are normal densities with means equal to the medians of the distributions and variances equal to the variances of the distributions. Red vertical lines indicate the parameters values that minimize the mean squared residuals.

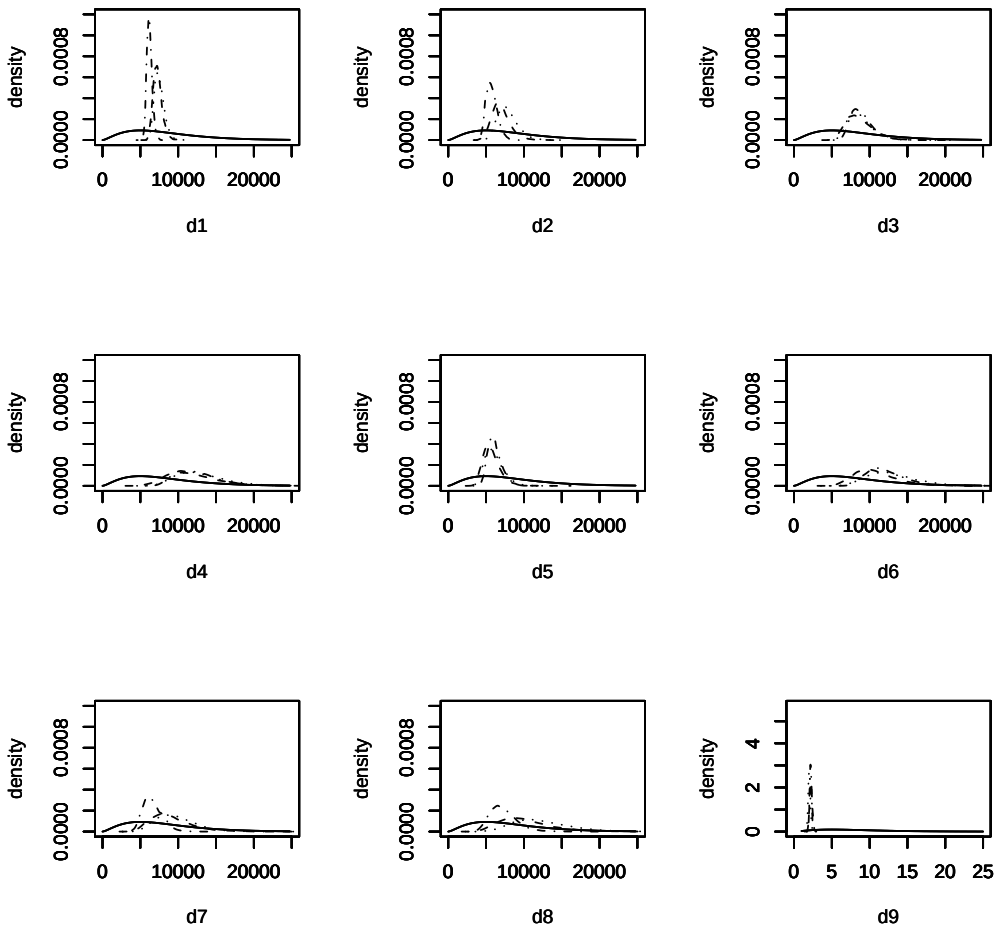


Figure 5: Posterior distributions from different numbers of data points for the all-components algorithm. (———) prior distribution; (- - - -) 12 points; (· · · · · ·) 16 points; (– · – · –) 25 points

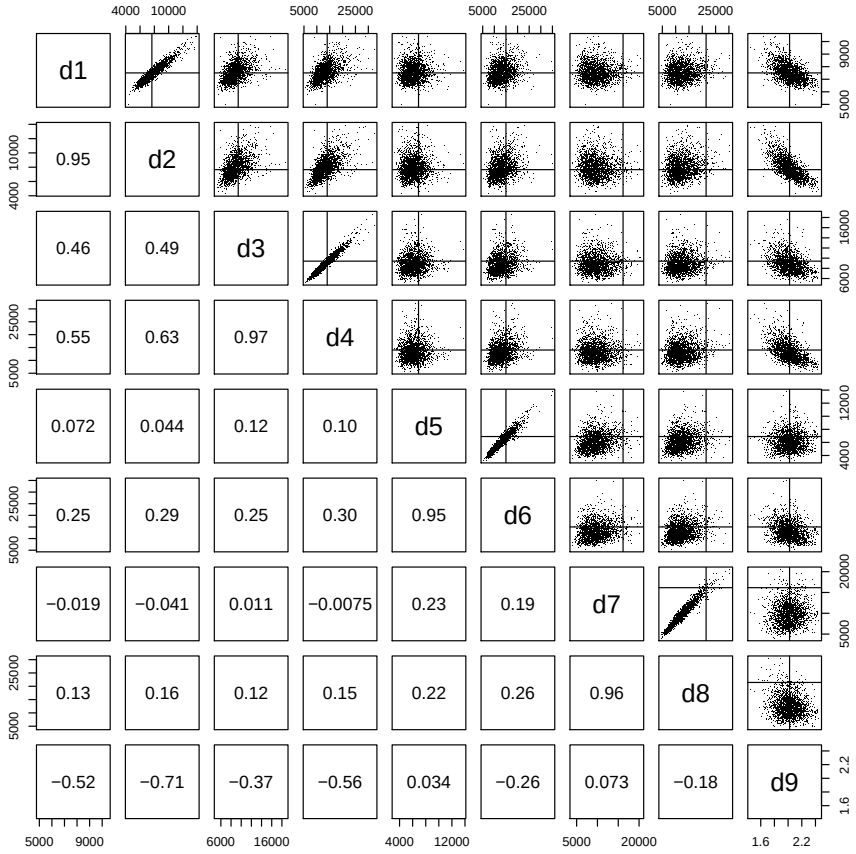


Figure 6: Bivariate scatter plots of the parameter distributions for the 5-reaction model found with the all-components Metropolis algorithm (upper triangle); correlation coefficients (lower triangle). The red lines indicate the maximum likelihood estimates of the parameters.

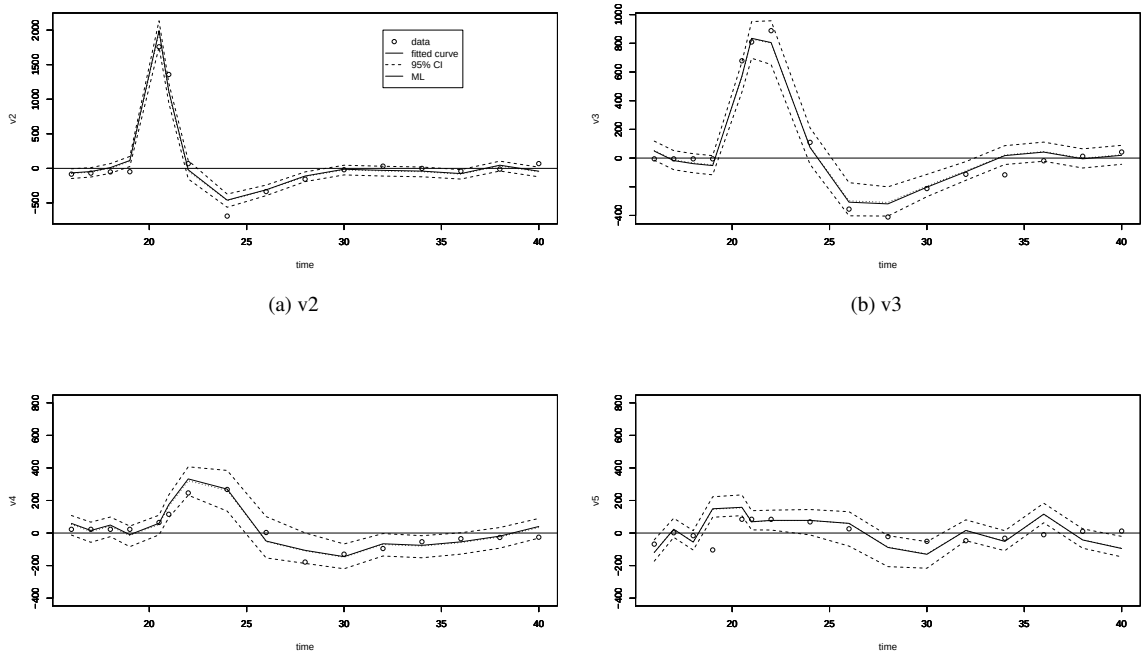
scatterplot of the distributions. There is correlation between some pairs of parameters, e.g., $d_1 - d_2$, but no evidence of multi-modality.

The value of the probability density for inference was assessed graphically. The probability density was used to find the maximum likelihood estimate for the model parameters and the 95% confidence intervals. The fits of the resulting models to the 16-point data set is shown in Figure 7. Quantitative measures of the fits for all the algorithms are given in Table 1. These measures of model fit could be used to compare the usefulness of different models though we have not done this for this paper.

Rates of convergence are illustrated in Figure 8. The mean sums of squared residuals (SSQ) are plotted vs.

Table 1: Mean squared residuals

	mean residual SSQ $\times 10^{-4}$			R^2_{adj}		
algorithm	12 pt.	16 pt.	25 pt.	12 pt.	16 pt.	25 pt.
1-comp	1.34	0.83	1.06	0.87	0.92	0.86
all-comp	0.74	0.93	0.74	0.93	0.90	0.90
NKC	0.86	0.73	0.71	0.92	0.92	0.91



(a) v2

(b) v3

(c) v4

(d) v5

Figure 7: Curves fit to the 16-point data with the all-components algorithm. The fitted curves are drawn with the maximum likelihood estimates for the parameters found with the L-BFGS-B algorithm [15] as implemented in the *R optim()* function.

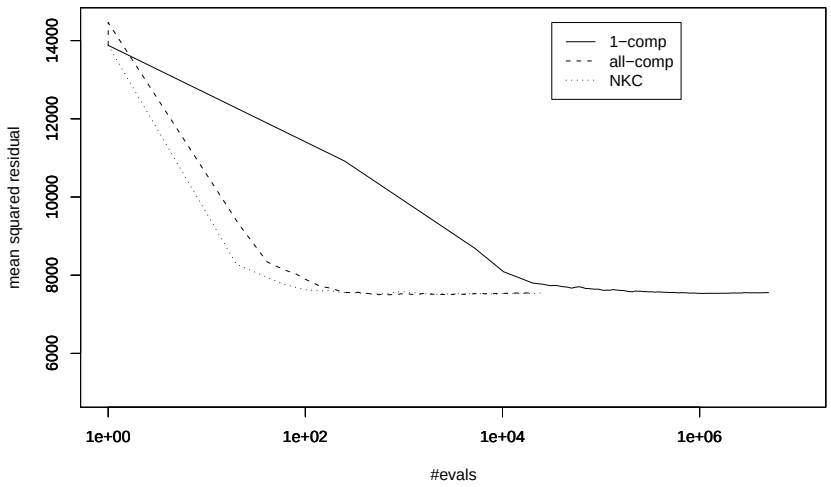


Figure 8: Mean squared residuals vs. number of likelihood evaluations for the 5-reaction model

the number of likelihood evaluations for the three algorithms. The relatively long convergence time for the 1-component Metropolis algorithm is a result of the high correlations between some of the parameters (Figure 6). With the 1-component Metropolis algorithm, one parameter is updated for each evaluation of the likelihood method, i.e., the algorithm searches the parameter space by moving parallel to the axes and thus it can sample a density with a diagonal axis of symmetry very slowly. This is not a problem with the all-component Metropolis and NKC algorithms since they update all the parameters at each iteration.

The actual convergence times on a 3.2GHz P4 machine with 1GB of memory are 1–2 hours for the 1-component algorithm and about 1 minute for the all-components and Normal Kernel Coupler algorithms.

5 Discussion

`$Id: discussion.Snw 1039 2006-12-06 19:20:46Z warnes $`

We are interested in assessing the usefulness of Markov chain Monte Carlo (MCMC) methods for the fitting of models of metabolic pathways. Fitting metabolic pathway models can be difficult because pathways are described by sets of reaction rate equations with overlapping sets of parameters. We have found that the MCMC algorithms handle this situation well, and produce reasonable joint probability densities for the model parameters. This output can be used for the estimation of confidence intervals for the parameters and the detection of correlations and multimodality. Thus MCMC compares favorably with maximum likelyhood methods that produce point estimates of the parameters and nonlinear regression methods that find approximations of the parameters and their variances.

MCMC methods and Bayesian statistics are particularly useful for modeling networks of biological reactions. These networks typically are modeled by large numbers of parameters and frequentist methods require at least as many observations as there are parameters to fit a model. In contrast, Bayesian methods incorporate our prior knowledge of the system and use the experimental data to refine the estimates (Figure 5). Thus the model fitting procedure described here lends itself to iterative experimentation where the experimental results, even if they consit of a single datum, can be used to update the prior for the next experiment.

The models used here have the form of the Hill function, $\frac{x^n}{\theta^n + x^n}$, with an exponent of 1. This form was chosen because the functions exhibit two of the characteristics of enzyme-catalyzed reactions: linearity at low concentrations of substrate and saturability. This, of course, is also the form of the Michaelis-Menten equation. These functions have the reactant concentrations as the independent variables since they are quantities that are relatively easy to measure. A drawback with these models is that they describe irreversible reactions whereas most enzymatic reactions are reversible. We have tried using a model of reversible reactions, the Haldane equation, but it does not fit the data from a perturbation equation very well. It can be used with MCMC simulation for multiple steady states, a situation we will continue to examine.

Three algorithms from the Hydra library were used. Two of the algorithms, the component-wise Metropolis and the all-components Metropolis algorithms, are conceptually simple and execute fairly quickly when there is not

too much structure in the joint density. The third algorithm, the Normal Kernel Coupler, is more computationally intensive but it can sample from complex densities more efficiently than the other two algorithms.

All three algorithms produced very similar results but the execution times varied considerably depending on the size of the model and the algorithm. The all-component Metropolis and NKC algorithms converged quickly but the 1-component algorithm was quite slow to converge. This is undoubtably due to the high correlations between some parameters which is evident in Figure 6. The 1-component algorithm has trouble sampling from this density because it is restricted to movements parallel to the axes. This is less of a problem for the all-components and NKC algorithms since they can move diagonally.

Each of the algorithms produced essentially identical results. The mean squared residuals do not vary much, nor do the calculated values for R^2_{adj}. These last values are indicative of a useful fit of the models to the data. These results also demonstrate the usefulness of MCMC simulation. The derived probability densities can be used to find good estimates for the parameters as well as reveal the correlations between the parameters. The result is a well-fitted model with realistic confidence intervals (Figure 7).

In this paper we have developed models of linear sequences of irreversible biochemical reactions and shown that several Markov chain methods can be used to generate useful fits of the models to simulated data. We also have shown that the results so obtained are unique in that they provide probability density estimates as opposed to simple point estimates for the parameters. The probability densities are useful for inference since they make evident the correlations between parameters. Lastly, we outlined areas for further research. Thus, this paper provides a starting point for the future development of tools for modeling complex biochemical pathways.

References

[1] Oates, P.J., 2002, Polyol pathway and diabetic peripheral neuropathy, *Int. Rev. Neurobiol.*, **50**, 325–392.

[2] Brownlee, M., 2001, Biochemistry and molecular biology of diabetic complications, *Nature*, **414**, 813–820.

[3] Nishikawa, T., Edelstein, D., Du, X. L., Yamagishi, S., Matsumura, Y.,Kaneda, Y., Yorek, M.A., Beebe, D., Oates, P.J., Hammes, H-P., Giardino, I., and Brownlee, M., 2000, Normalizing mitochondrial superoxide production blocks three pathways of hyperglycaemic damage, *Nature*, **404**, 787–790.

[4] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A., and Teller, E., 1953, Equation of state calculations by fast computing machines, *J. Chem. Physics*, **21**, 1087–1092.

[5] Gilkes, W.R., Richardson, S., and Spiegelhalter, D.J. (Ed.), *Markov Chain Monte Carlo in Practice* (Boca Raton, FL: Chapman & Hall/CRC)

[6] Gillespie, D.T., 1977, Exact Stochastic Simulation of Coupled Chemical Reactions, *J. Phys. Chem.*, **81**, 2340–2361.

[7] Loew, L., *Virtual Cell Modeling and Simulation Environment*, http://vcell.org (accessed 29 Sep 06)

[8] Arkin, A., Ross, J., and McAdams, H.H., 1998, Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells, *Genetics*, **149**, 1633–1648.

[9] R Development Team, *R: A Language and Environment for Statistical Computing*, http://www.r-project.org (accessed 7 Nov 06)

[10] Michaelis, L. and Menten, M.L., 1913, Die kinetic der invertinwirkung, *Biochem. Zeit.*, **49**, 333–369.

[11] Briggs, G.E. and Haldane, J.B.S., 1925, A note on the kinetics of enzyme action, *Biochem. J.*, **19**, 338–339.

[12] Hastings, W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97–109.

[13] Warnes, G.R., Hydra MCMC Library, http://www.sourceforge.net/projects/hydra-mcmc

[14] Warnes, G.R., 2000, The Normal Kernel Coupler: An adaptive Markov chain Monte Carlo method for efficiently sampling from multi-modal distributions, thesis, University of Washington.

[15] Byrd, R.H., Lu, P., Nocedal, J., and Zhu, C., 1995, A limited memory algorithm for bound constrained optimization, *SIAM J. Sci. Comput.*, **16**, 1190–1208.