BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1453–1454
doi:10.1093/bioinformatics/btp162

Systems biology

GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models

Frédéric Y. Bois

Direction des Risques Chroniques, INERIS, Parc ALATA, BP2, F-60550, Verneuil en Halatte, France

Received on January 17, 2009; revised on February 26, 2009; accepted on March 16, 2009

Advance Access publication March 20, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Statistical inference about the parameter values of complex models, such as the ones routinely developed in systems biology, is efficiently performed through Bayesian numerical techniques. In that framework, prior information and multiple levels of uncertainty can be seamlessly integrated. GNU MCSim was precisely developed to achieve those aims, in a general non-linear differential context. Starting with version 5.3.0, GNU MCSim reads in and simulates Systems Biology Markup Language models. Markov chain Monte Carlo simulations can be used to generate samples from the joint posterior distribution of the model parameters, given a dataset and prior distributions. Hierarchical statistical models can be used. Optimal design of experiments can also be investigated.

Availability and Implementation: The GNU GPL source is available at http://savannah.gnu.org/projects/mcsim. A distribution package is at http://www.gnu.org/software/mcsim. GNU MCSim is written in standard C and runs on any platform supporting a C compiler. Supplementary Material is available online at http://www.gnu.org/software/mcsim.

Contact: frederic.bois@ineris.fr

1 INTRODUCTION

Most systems biology models focus on simulating the behavior of pathways or networks. Studying the properties of theoretical systems is a worthy endeavor, but uncovering the characteristics of actual systems, or testing hypotheses, requires a step beyond pure simulation, into *statistical inference*. Such inference is needed to correctly account for most sources of uncertainty that stem from experimentation (dealing with uncertainty about model structure is somewhat more difficult but can be partially tackled through model checking). Inference about the parameter values of complex models, well represented in systems biology by large sets of differential equations, is efficiently performed with Bayesian numerical techniques (Gelman *et al.*, 1996; Rogers *et al.*, 2006), in which prior information and multiple levels of uncertainty can be seamlessly integrated.

GNU MCSim is a numerical simulation and Bayesian statistical inference tool for algebraic or differential equation systems (Bois and Maszle, 1997). Other programs have been created to that end: Matlab is one of the most general and easy to use program, and BioBayes (Vyshemirsky and Girolami, 2008) a noteworthy specialized software. Still, there is room for free GPL software able to efficiently run computer intensive multilevel Bayesian analyses. GNU MCSim was created specifically to perform Monte Carlo analyses in an optimized, and easy to maintain environment.

2 PROGRAM OVERVIEW

GNU MCSim is written in ANSI-standard C language. The source code is freely available under GNU GPL and can be compiled for any system, provided an ANSI C compliant compiler. The code can be freely modified.

The software consists in two parts. A model generator, '*mod*', takes a user defined model written in Systems Biology Markup Language (SBML) or in GNU MCSim native syntax, and translates it into C. A dynamic library of routines is then linked to the C version of the model. The resulting executable can run Monte Carlo simulations under a variety of conditions, using a specified statistical model.

2.1 Defining structural models in native syntax

The native syntax allows you to specify a set of linear or non-linear algebraic equations or ordinary differential equations to solve. The model file in that case is an ASCII text file with several sections, including global declarations of variables and parameters, dynamics' specifications (with derivative calculations), model initialization and output computations.

2.2 Reading SBML models and applying a template

SBML provides a standard for representing biochemical pathway models (Hucka *et al.*, 2003). It can code for models of metabolism, cell-signaling, transcription, etc. SBML is maintained since the mid-2000 by an international community of software developers and users (cf. http://sbml.org).

GNU MCSim reads SBML Level 1 models and most Level 2 models. Some features of Level 2 are not yet recognized by *mod*, such as 'functionDefinition', 'unit' and 'rateRule'). Compartments are ignored unless a model template for circulating species is given.

It is possible to read in several SBML models at once. In that case, they are merged, in the sense that if a chemical species is present in several of them, its rate equation is constructed from the reaction descriptions of each SBML model. Obviously, in that case, the concurrent models should be coherent in their structure and notation.

2.3 Performing simulations and statistical inference

After compilation of a model into a GNU MCSim executable, five types of simulations can be run, as specified in simulation input files.

Simple runs produce time-courses of the model variables. The main integration routine is Lsodes, from the SLAC Fortran library (Gear, 1971).

Monte Carlo simulations perform repeated runs across a randomly sampled region of the model parameter space.

Markov chain Monte Carlo simulations (MCMC) simulate a Markov chain in the model parameter space (Gilks *et al.* 1996). They can provide samples

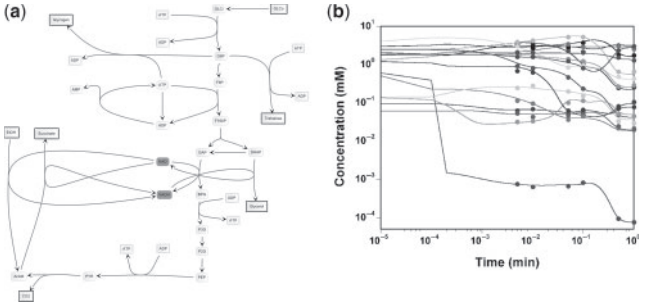

Fig. 1. (**a**) JDesigner representation of the yeast glycolysis model of Pritchard and Kell (2002). (**b**) A simulation of the time course of all model species (note the logarithmic *x*-axis.) The dots mark the values taken as simulated data for the Bayesian calibration exercise.

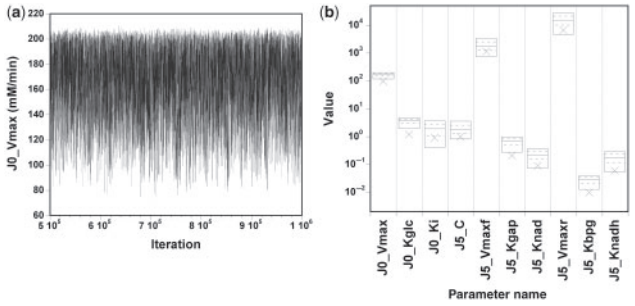

Fig. 2. (**a**) Trajectory of three Markov chains for parameter J0_Vmax at convergence over the last half million runs. (**b**) Percentile plot (5th, 25th, 50th, 75th and 95th percentiles) of the posterior sample obtained for each parameter, overlaid with their 'true' value (crosses).

from a Bayesian joint posterior distribution of the parameters, given a structural model (e.g. an SBML model), a statistical model (specified in a simulation file), prior parameter distributions and data for which a likelihood function can be computed. Fast Metropolis (vector) sampling is available, as well as Metropolis within Gibbs sampling (one component at a time). GNU MCSim handles multilevel (e.g. random effects and mixed effects) statistical models (Gelman *et al.*, 1996), particularly useful when several datasets are at hand or when variance structures are complex.

Set-points simulations solve the model for a series of specified parameter sets, placed on a regular grid, for example, or generated from previous Monte Carlo or MCMC simulations.

The optimal design procedure optimizes the number and location of observation times for experimental conditions, minimizing the variance of a set of parameters or outputs, given a structural model, a statistical model and prior parameter distributions (Amzal *et al.*, 2006; Bois *et al.*, 1999).

2.4 Application examples

The yeast glycolysis model (YeastGlycolysisJDClean.xml) of Pritchard and Kell (2002), provided with JDesigner 2.1c and the Systems Biology Workbench 2.7.8 (http://sbw.kgi.edu/index.htm), was used. Figure 1 shows the model (17 state variables and 95 parameters) and the results of time-course simulations over a minute. Those simulation results are identical to those obtained with the Jarnac simulator of SBW 2.7.8. To generate a synthetic dataset, Gaussian error [10% coefficient of variation (CV)] was added to points along each species trajectories at 7 times (including time 0.) Those points, used as 'data' in the subsequent analyses, are also plotted on Figure 1.

Posterior distributions were obtained by MCMC sampling for 10 model parameters (see details in the Supplementary Material). Independent uniform priors and lognormal likelihood functions were used. Figure 2 shows results obtained with three concurrent chains, in two steps each (10 000 runs with component sampling, 1 million runs with vector sampling). Vector sampling was more efficient, given the very strong posterior covariance between some model parameters, for example, between J0_Vmax and J0_kglc. Convergence was achieved, with all R diagnostics (Gelman *et al.*, 1996) = 1, over the last 5×10^5 iterations. Each chain took ~80 min on a i686 machine clocked at 3.6 GHz. Keeping one sample for every 50 in the last 5×10^5 of the three chains led to a joint posterior sample of 30 000 values for each parameter, whose empirical percentiles are shown Figure 2. Due to strong posterior covariances, the posterior samples tend to be higher than the true parameter values used to generate the simulated data. An application of design optimization to fix that problem is given in the Supplementary Material, together with an example of multi-level modeling.

3 CONCLUSIONS

GNU MCSim can efficiently perform Bayesian inference and optimal design computations for systems biology models. Multi-level models can also be used. Despite a few limitations in reading in SBML Level 2, which will be removed in a coming version, GNU MCSim can import most SBML models and allows users to take advantage of its speed and flexibility. Its code is open and can be freely tailored to specific needs.

Funding: European Commission Sixth Framework Program, Priority 6 (Global change and ecosystems), project 2-FUN (contract #036976); French Ministry of Sustainable Development Program (BCRD 2004 DRC05).

Conflict of Interest: none declared.

REFERENCES

Amzal,B. *et al.* (2006) Bayesian optimal design via interacting MCMC. *J. Am. Stat. Assoc.* **101**, 773–785.

Bois,F.Y. and Maszle,D. (1997) MCSim: a simulation program. Available at http://www.jstatsoft.org/v02/i09 (last accessed date April 2, 2009).

Bois,F.Y. *et al.* (1999) Optimal design for a study of butadiene toxicokinetics in humans. *Toxicol. Sci.*, **49**, 213–224.

Gear,C.W. (1971) The automatic integration of ordinary differential equations. *Commun. ACM*, **14**, 176–179.

Gelman,A. *et al.* (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Stat. Assoc.*, **91**, 1400–1412.

Gilks,W.R. *et al.* (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Pritchard.L and Kell,D.B. (2002) Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. *Eur. J. Biochem.*, **269**, 3894–3904.

Rogers,S. *et al.* (2006) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8** (**Suppl. 2**), S2.

Vyshemirsky,V. and Girolami,M. (2008) BioBayes: a software package for Bayesian inference in systems biology. *Bioinformatics*, **24**, 1933–1934.