Dizzy-Beats

Bioinformatics Advance Access published January 30, 2015

Dizzy-Beats: a Bayesian Evidence analysis tool for systems biology

Stuart Aitken [1,*] Alastair M. Kilpatrick [2,3] and Ozgur E. Akman [4]

[1]MRC Human Genetics Unit, IGMM, University of Edinburgh, Edinburgh EH4 2XU, UK
[2]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
[3]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA
[4]Centre for Systems, Dynamics and Control, College of Engineering, Mathematics & Physical Sciences, University of Exeter EX4 4QF UK

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Model selection and parameter inference are complex problems of long-standing interest in systems biology. Selecting between competing models arises commonly as underlying biochemical mechanisms are often not fully known, hence alternative models must be considered. Parameter inference yields important information on the extent to which the data and the model constrain parameter values.

Results: We report Dizzy-Beats, a graphical Java Bayesian evidence analysis tool implementing nested sampling - an algorithm yielding an estimate of the log of the Bayesian evidence Z and the moments of model parameters, thus addressing two outstanding challenges in systems modelling. A likelihood function based on the L_1-norm is adopted as it is generically applicable to replicated time series data.

Availability: http://sourceforge.net/p/bayesevidence/home/Home/

Contact: s.aitken@ed.ac.uk

1 INTRODUCTION

Bayesian methods provide a sound basis for ranking alternative systems biology models and for characterising the extent to which parameters are constrained by models and data (Kirk *et al.*, 2013). Markov Chain Monte Carlo (MCMC) methods have been applied to model selection (Schmidl *et al.*, 2012) and to parameter inference in systems biology (Hug *et al.*, 2013; Kanodia *et al.*, 2014), but often require considerable algorithmic and conceptual development. Nested sampling promises to ease these complex computational tasks: Recent biological applications include (Aitken and Akman, 2013; Kirk *et al.*, 2013; Pullen and Morris, 2014).

General purpose code for nested sampling is available in R (Skilling, 2006; Aitken and Akman, 2013), and biological applications of the MultiNest tool (Feroz *et al.*, 2013) have been reported (Kirk *et al.*, 2013; Pullen and Morris, 2014). A C-based command-line application implementing nested sampling and providing an SBML interface has recently been released (Johnson *et al.*, 2014), but no graphical tool is currently available. Thus we sought to add nested sampling to the widely-used Dizzy chemical kinetics simulation tool (Ramsey *et al.*, 2005) (over 200 citations as of November 2014). While doing so we also added an optimisation function and SBML 3.1 compatibility. However, as Dizzy's command language has operators that cannot be captured in SBML 3.1, and SBML 3.1 has features not supported by Dizzy, this feature is restricted to the intersection of the modelling languages.

*to whom correspondence should be addressed

2 METHODS

Nested sampling calculates two of the central results of Bayesian inference: the posterior distribution $P(\theta|D, H_i)$ of the parameters θ, and the evidence $P(D|H_i)$, that is, the support for the data D under hypothesis H_i (Skilling, 2006), through a sampling strategy. A selection between two alternative models H_0 and H_1 can be made by calculating the ratio of their posterior probabilities (1), a calculation that can be decomposed into the Bayesian evidence (Z_0 and Z_1) and the prior probability of the respective hypotheses.

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} = \frac{Z_1 P(H_1)}{Z_0 P(H_0)} \quad (1)$$

$$Z = \int L(\theta)\pi(\theta)\,\mathrm{d}\theta \quad (2)$$

The evidence (2) is a scalar quantity that can be viewed as an integral of the likelihood (L) over the elements of mass ($\mathrm{d}X = \pi(\theta)\mathrm{d}\theta$) associated with the prior density $\pi(\theta)$. The prior mass can be accumulated from its elements ($\mathrm{d}X$) in any order. The enclosed prior of likelihood $> \lambda$ can be defined (3), and this allows the evidence to be written as a one-dimensional integral of the (inverse) likelihood $L(X)$ over the unit range (taking the enclosed prior mass X to be the primary variable) (4) (Skilling, 2006).

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\,\mathrm{d}\theta \quad (3)$$

$$Z = \int_0^1 L(X)\,\mathrm{d}X \quad (4)$$

$$L(X(\lambda)) \equiv \lambda$$

Given a sequence of decreasing values $0 < X_m < \ldots X_2 < X_1 < 1$ where the likelihood $L_i = L(X_i)$ can be evaluated, the evidence can be approximated numerically as a weighted sum. Inferences about the posterior can be obtained from the sequence of m discarded points generated by sampling, P. Each point is assigned the weight $p_i = L(\theta_i)w_i/Z$, from which the first and second moments of each parameter in θ can be estimated – for more details see Skilling (2006) and Aitken and Akman (2013). The size of the population of active points (points θ_i within the evolving constraint $L(\theta) > \lambda$) used to sample the parameter space is the only

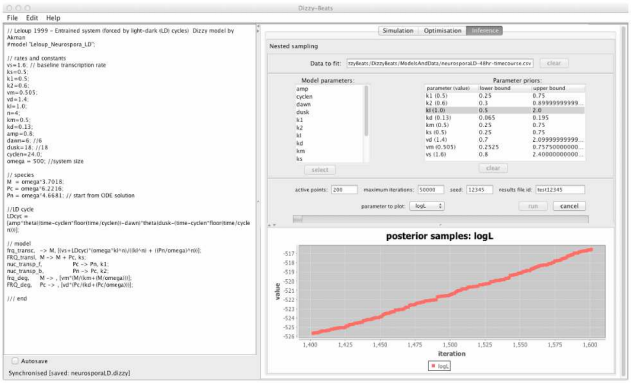

Fig. 1. Dizzy-Beats: an application for parameter inference and model selection.

parameter of the algorithm that the user must specify. For complex likelihood functions with multiple modes, this number may be as high as 10000, and for a single mode as low as 200.

Dizzy-Beats updates the user interface of the original Dizzy program (Ramsey *et al.*, 2005) retaining the text of the model in the left editing panel (see Figure 1) and placing the original simulator choices in the simulation tab on the the right. A histogram plot for visualising the results of stochastic simulations is added to the simulation viewing formats. Two new tabs add optimisation and inference capabilities, and both require a data file to be specified (a simple CSV format is used, with column headers matching the names of species in the model). Using the simulation tab, the user can select parameters, modify their values and see the simulation results plotted over the data. This allows a manual tuning and exploration of the model's fit to the data. Computational optimisation using simulated annealing can be run to explore a larger parameter space. Similarly, the inference tab requires users to select parameters to be included in inference by nested sampling and to input their prior range. A uniform prior is assumed as is typical in nested sampling. A graph of log likelihood, or of the samples of the selected parameters can be viewed as nested sampling progresses to monitor progress. The stopping heuristics of Aitken and Akman (2013) are implemented but the user can in addition specify the maximum number of iterations, and must specify the number of active points. The outputs are a file summarising the results, and a second listing the posterior samples for further analysis.

A likelihood function based on the L_1-norm is used for optimisation and inference – this is defined by eqns. (5) and (6) (Sivia and Skilling, 2006).

$$\epsilon_t = \langle |x_t - \mu_t| \rangle = \int |x_t - \mu_t| p(x) d^N x \tag{5}$$

$$p(x|\{\mu_t, \epsilon_t\}) = \prod_{t=1}^{m} \frac{1}{2\epsilon_t} \exp\left(-\frac{|\tilde{x}_t - \mu_t|}{\epsilon_t}\right) \tag{6}$$

Equation (5) defines the normalising constant ϵ_t as the expected value of the moduli of the differences between the replicate observations at time t and the values predicted by the kinetic model (μ_t). The product of the probabilities of the median observation at time t ($\tilde{x}_t$) defines the likelihood function for a time series x of m samples (eqn. (6)). Maximisation of this likelihood minimises the sum of the moduli of the residuals (rather than their squares) on the basis that the testable information is restricted to the expected value of the modulus of the difference between theory and experiment. Should we know both the mean and variance, maximum entropy considerations would lead instead to the Gaussian distribution (Sivia and Skilling, 2006). Time points where the replicates are most dissimilar contribute least to the likelihood as ϵ_t is larger – as is desirable.

3 DISCUSSION

Dizzy-Beats is a graphical application for simulating and optimising systems models based on an established simulator (Ramsey *et al.*, 2005) and its simple textual model syntax, to which we have added SBML 3.1 import/export functionality. Uniquely, Dizzy-Beats provides model comparison and parameter inference functions through the nested sampling algorithm in a graphical application. Comparable functions are implemented in BioBayes (Vyshemirsky and Girolami, 2008), however, users must edit the XML representation of the model should they wish to make modifications. SYSBIONS (Johnson *et al.*, 2014) implements nested sampling but all interaction is via the command-line. The use of a likelihood based on the L_1-norm derived from biological replicate data makes fewer assumptions than a Gaussian error model (Vyshemirsky and Girolami, 2008; Johnson *et al.*, 2014), and is less computationally complex than a transitional likelihood function derived from reaction propensities (Heron *et al.*, 2007; Aitken and Akman, 2013).

ACKNOWLEDGEMENT

Funding: This work was funded by BBSRC grant BB/I023461/1 (Bayesian evidence analysis tools for systems biology; S.A. and O.E.A.).

REFERENCES

Aitken, S. and Akman, O. E. (2013). Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. *BMC Systems Biology*, **7**(1), 72.

Feroz, F., Hobson, M. P., Cameron, E., and Pettitt, A. N. (2013). Importance Nested Sampling and the MultiNest Algorithm. *ArXiv e-prints: 1306.2144*.

Heron, E. A., Finkenstadt, B., and Rand, D. A. (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, **23**(19), 2596–2603.

Hug, S., Raue, A., Hasenauer, J., Bachmann, J., Klingmüller, U., Timmer, J., and Theis, F. (2013). High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. *Mathematical Biosciences*, **246**(2), 293 – 304.

Johnson, R., Kirk, P., and Stumpf, M. P. (2014). SYSBIONS: nested sampling for systems biology. *Bioinformatics doi: 10.1093/bioinformatics/btu675*.

Kanodia, J., Chai, D., Vollmer, J., Kim, J., Raue, A., Finn, G., and Schoeberl, B. (2014). Deciphering the mechanism behind Fibroblast Growth Factor (FGF) induced biphasic signal-response profiles. *Cell Communication and Signaling*, **12**(1), 34.

Kirk, P., Thorne, T., and Stumpf, M. P. (2013). Model selection in systems and synthetic biology. *Current Opinion in Biotechnology*, **24**(4), 767–774. Nanobiotechnology Systems biology.

Pullen, N. and Morris, R. J. (2014). Bayesian model comparison and parameter inference in systems biology using nested sampling. *PLoS ONE*, **9**(2), e88419.

Ramsey, S., Orrell, D., and Bolouri, H. (2005). Dizzy: Stochastic simulation of large-scale genetic regulatory networks. *Journal of Bioinformatics and Computational Biology*, **03**(02), 415–436.

Schmidl, D., Hug, S., Li, W., Greiter, M., and Theis, F. (2012). Bayesian model selection validates a biokinetic model for zirconium processing in humans. *BMC Systems Biology*, **6**(1), 95.

Sivia, D. and Skilling, J. (2006). *Data analysis: A Bayesian tutorial.* OUP.

Skilling, J. (2006). Nested sampling for general Bayesian computations. *Bayesian Analysis*, **1**(4), 833–860.

Vyshemirsky, V. and Girolami, M. (2008). BioBayes: A software package for Bayesian inference in systems biology. *Bioinformatics*, **24**(17), 1933–1934.