Improving the accuracy of demographic and clock model comparison while accommodating phylogenetic uncertainty

Guy Baele

Evolutionary and Computational Virology Section,
KU Leuven, Belgium

June 19th, 2012

1. Introduction
2. HIV-1 data - Results
3. Demographic simulations
4. HSV data - Results
5. S. aureus - Results
6. Conclusions

Goal of model selection: accurately calculate Bayes factors

- given prior belief that the models are equally likely: expresses how much more (less) likely one model is compared to the other
- ratio of marginal likelihoods
- takes into account differences in dimensions, so higher dimensional models are not automatically preferred
- the aim of model selection is not necessarily to find the true model that generated the data but to select a model that best balances simplicity with flexibility and biological realism in capturing the key features of the data (Steel 2005)

Calculating reliable (log) Bayes factors:

- depends on reliable calculation of the marginal likelihoods
- the harmonic mean estimator (HME) remains widely used because of its computational efficiency (and ease of implementation)
- but the HME is often severely biased, overestimating the true marginal likelihood (Lartillot & Philippe 2006, Xie et al. 2011) and tends to prefer models with higher dimensions
- in some (simple) cases even the arithmetic mean estimator (AME) outperforms the HME

Compare marginal likelihood estimation using different methods:

- the harmonic mean estimator (HME; Raftery & Lewis, 1994)
- the stabilized harmonic mean estimator (sHME; Suchard & Redelings, 2005)
- a posterior simulation-based analogue of Akaikes information criterion (AIC) through Markov chain Monte Carlo (MCMC) (AICM; Raftery & Newton, 2007)
- path sampling (PS; Lartillot & Philippe, 2006)
- stepping-stone sampling (SSS; Xie et al., 2011)

Applied to demographic and relaxed molecular clock comparison (using BEAST)

The AICM is defined as:

$$AICM = 2s_l^2 - 2\bar{l}$$

where $\bar{l}$ and s_l^2 are the sample mean and variance of the posterior log likelihoods (Raftery et al., 2007)

HME, sHME and AICM properties:

- only require samples from the posterior
- reasonably fast
- quick convergence towards the marginal likelihood

PS and SS/SSS properties:

- require samples from a series of power posteriors, along a path between prior and posterior:

$$q_\beta(\theta) = p(Y \mid \theta, M)^\beta p(\theta \mid M),$$

- PS and SS differ in the way the collected samples are used to estimate the log marginal likelihood
- reduces to the posterior when $\beta = 1$
- reduces to the prior when $\beta = 0$
- slow / computationally demanding
- slower convergence than HME/AICM

Placing more computational effort near $\beta = 0$ leads to a substantial increase in the efficiency of the estimator

Stepping-stone sampling convergence compared to path sampling and harmonic mean for a green plant 10-taxa dataset

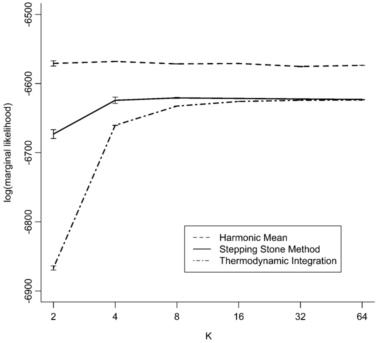


FIGURE 5. Log marginal likelihood for three estimation methods as a function of the number of β intervals, *K*, for the green plant Ribulose Bisphosphate Carboxylase/Oxygenase large subunit (*rbc*L) example. β values are evenly spaced quantiles from a Beta(0.3,1.0) distribution. Error bars represent ±1 standard error based on 30 independent MCMC analyses.

- HIV-1 data: 162 taxa, 997 bp
- "The inability to strongly reject the model with a constant population size prior is counterintuitive because it is clear that the HIV-1 population size has increased notably. We speculate that this finding might be due to the simplest model providing a good fit to a relatively short, information-poor alignment, in comparison with more parameterized models.' (Worobey et al., 2008)'

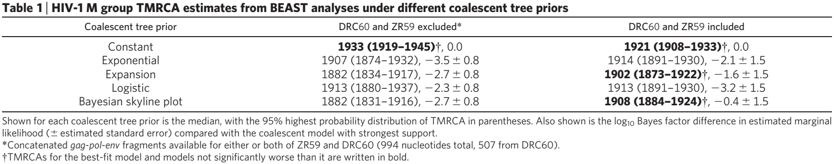


Table 1 | HIV-1 M group TMRCA estimates from BEAST analyses under different coalescent tree priors

Coalescent tree prior	DRC60 and ZR59 excluded*	DRC60 and ZR59 included
Constant	**1933 (1919–1945)**†, 0.0	**1921 (1908–1933)**†, 0.0
Exponential	1907 (1874–1932), −3.5 ± 0.8	1914 (1891–1930), −2.1 ± 1.5
Expansion	1882 (1834–1917), −2.7 ± 0.8	**1902 (1873–1922)**†, −1.6 ± 1.5
Logistic	1913 (1880–1937), −2.3 ± 0.8	1913 (1891–1930), −3.2 ± 1.5
Bayesian skyline plot	1882 (1831–1916), −2.7 ± 0.8	**1908 (1884–1924)**†, −0.4 ± 1.5

Shown for each coalescent tree prior is the median, with the 95% highest probability distribution of TMRCA in parentheses. Also shown is the $\log_{10}$ Bayes factor difference in estimated marginal likelihood (± estimated standard error) compared with the coalescent model with strongest support.

*Concatenated *gag-pol-env* fragments available for either or both of ZR59 and DRC60 (994 nucleotides total, 507 from DRC60).

†TMRCAs for the best-fit model and models not significantly worse than it are written in bold.

Analysis of the HIV-1 data set using HME, AICM, PS and SS

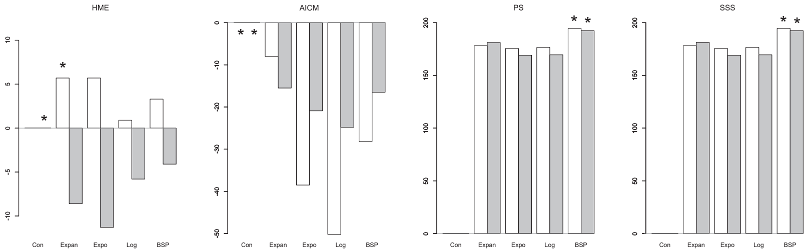


Figure: Differences in log marginal likelihood estimates and AICM for two independent fittings (first fitting shown in white, second in gray) of the HIV dataset using the HME, AICM, PS and SS. For each estimator, the constant population size model (Con) was used as the reference model.

Simulation settings:

- consider the sampling dates of 60 sequences that represent the diversity in the original HIV-1 dataset
- simulate dated-tip genealogies under two simple demographic models: a constant population size and an exponentially growing population size through time
- increasing growth rates under the exponential growth model: 0.01, 0.025, 0.05, and 0.10 per year
- rescale to fit a reasonable TMRCA $\sim N(1910, 10)$
- simulate 1000 nucleotides per sequence
- calculate marginal likelihood for constant and exponential population sizes using HME, AICM, PS and SSS

Coalescent prior	Growth rate	HME	AICM	PS	SS	log BF HME	ΔAICM	log BF PS	log BF SS
Constant	-	48	59	72	72	0.61	0.57	1.76	1.76
Exponential	0.010	50	45	57	57	0.28	0.20	-0.81	-0.80
Exponential	0.025	59	73	92	92	-1.33	-1.36	-6.81	-6.81
Exponential	0.050	80	99	100	100	-4.43	-4.34	-12.54	-12.54
Exponential	0.100	78	100	100	100	-7.75	-7.66	-18.24	-18.24

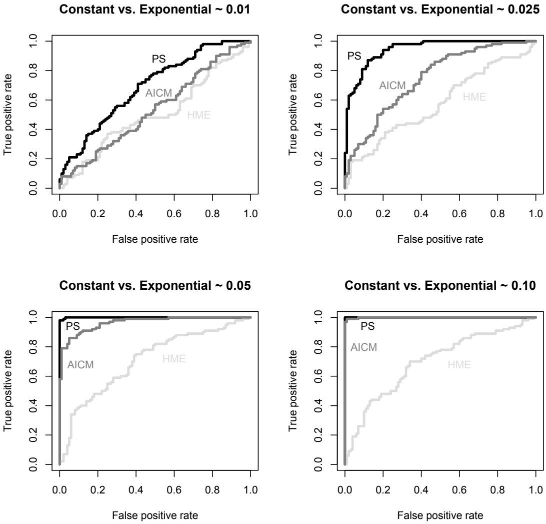
Conclusions:

- an exponential demographic prior with a growth rate of 0.01 is a difficult case for each estimator
- HME is unable to reach an accuracy higher than 80%
- AICM outperforms HME in all but one case
- PS/SS outperform HME and AICM in all cases

These simulation results

- were obtained using log BF of 0 as cut-off for binary classification of models
- to assess the discriminatory power of the HME, AICM, and PS/SS across a range of cutoffs, we plot the true positive rate as a function of the false positive rate
- determine the ROC (receiver operating characteristic) curves
- evaluate BF distributions that compare the fit of both coalescent models on data simulated under constant population size and a particular growth rate

ROC curves for the different demographic simulation scenarios



- HSV-1 data: 84 taxa, 1.135 bp, sampling range: 1981-2008 (Firth et al., 2010)
- compare strict clock (SC) and uncorrelated relaxed clock models with an underlying exponential distribution (UCED) and lognormal distribution (UCLD)
- with and without sampling dates
- of the different clock models available, the HME often prefers the UCED

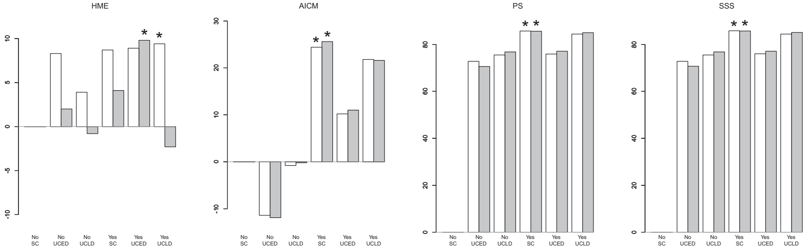


Figure: Differences in log marginal likelihood estimates for two independent fittings (first fitting shown in white, second in gray) for the HSV dataset using HME, AICM, PS and SS using a strict clock (SC), an uncorrelated relaxed clock with an exponential distribution (UCED) and an uncorrelated relaxed clock with a lognormal distribution (UCLD). The data was analyzed excluding the sampling dates (No) and including the sampling dates (Yes). We used the strict clock model excluding the sampling dates as the reference model.

- Staphylococcus aureus data (Gray et al., 2011)
- full data set: 63 taxa, 4.310 bp
- intergenic data set: 63 taxa, 962 bp
- synonymous data set: 63 taxa, 1.055 bp
- molecular clock models: SC and UCLD
- demographic models: constant population size and Bayesian skyline plot
- original analysis: three independent fittings were combined to obtain sufficient independent samples from the posterior (because of mixing issues and improper priors)

Table: Marginal likelihood estimates for two independent fittings for the HA-MRSA ST239 dataset using the HME, AICM, PS and SS (with the overall ranking of the models shown in parentheses for each estimator) after specifying proper priors. Consistent results across all three data partitions could only be obtained when using proper priors and PS/SS.

			Fitting 1				Fitting 2			
Data	Clock	Coalescent	HME	AICM	PS	SS	HME	AICM	PS	SS
Full	SC	Constant	-28420.5 (4)	56865.6 (4)	-28738.2 (4)	-28735.9 (4)	-28418.2 (3)	56865.2 (4)	-28735.5 (4)	-28734.2 (4)
Full	SC	BSP	-28419.4 (3)	56860.7 (3)	-28724.9 (3)	-28723.2 (3)	-28420.8 (4)	56860.2 (3)	-28723.8 (3)	-28722.3 (3)
Full	UCLD	Constant	-28304.2 (2)	56681.8 (2)	-28641.1 (2)	-28638.3 (2)	-28308.2 (2)	56682.4 (2)	-28647.5 (2)	-28644.2 (2)
Full	UCLD	BSP	-28304.1 (1)	56679.6 (1)	-28635.6 (1)	-28631.9 (1)	-28304.4 (1)	56680.1 (1)	-28631.8 (1)	-28628.2 (1)
Intergenic	SC	Constant	-6493.7 (4)	13016.8 (2)	-6749.5 (4)	-6749.3 (4)	-6495.9 (4)	13016.7 (2)	-6750.0 (4)	-6749.6 (4)
Intergenic	SC	BSP	-6489.4 (3)	13001.4 (1)	-6740.0 (3)	-6739.7 (3)	-6488.9 (3)	13001.4 (1)	-6742.3 (3)	-6742.0 (3)
Intergenic	UCLD	Constant	-6479.9 (1)	13037.7 (3)	-6730.1 (2)	-6729.4 (2)	-6481.9 (1)	13038.3 (3)	-6725.2 (2)	-6724.8 (2)
Intergenic	UCLD	BSP	-6480.6 (2)	13048.2 (4)	-6716.7 (1)	-6716.1 (1)	-6482.0 (2)	13043.7 (4)	-6717.1 (1)	-6716.5 (1)
Synonymous	SC	Constant	-6563.9 (4)	13149.7 (4)	-6816.3 (4)	-6815.8 (4)	-6561.9 (4)	13149.2 (4)	-6816.8 (4)	-6816.3 (4)
Synonymous	SC	BSP	-6556.1 (3)	13133.1 (2)	-6806.4 (3)	-6806.0 (3)	-6558.5 (3)	13133.8 (2)	-6806.6 (3)	-6806.1 (3)
Synonymous	UCLD	Constant	-6541.7 (2)	13138.7 (3)	-6787.4 (2)	-6786.7 (2)	-6538.6 (2)	13138.5 (3)	-6786.8 (2)	-6786.1 (2)
Synonymous	UCLD	BSP	-6533.6 (1)	13122.8 (1)	-6780.8 (1)	-6780.3 (1)	-6536.1 (1)	13123.9 (1)	-6780.9 (1)	-6780.0 (1)

Main conclusions:

- PS/SS outperform AICM and HME in a series of demographic simulations (and AICM outperforms HME)
- PS/SS are the most consistent across different runs for empirical data sets, followed by the AICM
- HME clearly the worst of the four methods compared
- hence, the increased computational demands / additional implementation for PS/SS are worth it
- the different estimators incorporate phylogenetic uncertainty
- be careful to provide proper priors for your parameters!

Availability:

- AICM, PS/SS are now available in BEAST (as of release 1.7) through XML specification
- AICM is available in Tracer (only in repository)
- general implementation: allows to calculate marginal likelihoods for any model that can be fitted in BEAST
- e.g. demographic and molecular relaxed clock models, models of sequence evolution, trait evolution and phylogeographic models, . . .

Current work:

- Li and Drummond (2011) developed a Bayesian model averaging (BMA) approach for relaxed molecular clock models
- Wai Lok Sibon Li & Alexei J. Drummond (2012) *Model Averaging and Bayes Factor Calculation of Relaxed Molecular Clocks in Bayesian Phylogenetics*. Mol. Biol. Evol. 29(2):751-761.
- using estimates of the posterior probability of each model, Li and Drummond (2011) examine the performance of identifying the maximum a posteriori (MAP) model under BMA as a model selection criterion
- compare the performance of HME, sHME, AICM, PS/SS and BMA for relaxed molecular clocks

Acknowledgements:

- Philippe Lemey
- Trevor Bedford
- Andrew Rambaut
- Marc Suchard
- Alexander Alekseyenko

Funding: European Union Seventh Framework Programme (FP7/2007-2013), ERC Grant agreement no. 260864

G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard and A. V. Alekseyenko (2012) *Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty.* Mol. Biol. Evol. *(in press).*