A new framework for reconstructing epidemic dynamics from virus sequences: Model validation and application to the HIV CRF07-BC epidemic in China

Art FY Poon[1,2,3]

TUPEA056

[1]BC Centre for Excellence in HIV/AIDS, Vancouver, Canada; [2]Department of Medicine, University of British Columbia, Vancouver, Canada; [3]Faculty of Health Sciences, Simon Fraser University, Burnaby, Canada

CONTACT: Art Poon
BC Centre for Excellence in HIV/AIDS
680-1081 Burrard St.
Vancouver, BC V6Z 1Y6

apoon@cfenet.ubc.ca

BACKGROUND

Many RNA viruses, like HIV, evolve so rapidly that genetic differences accumulate between related infections within weeks.

Consequently, the epidemic spread of the virus leaves a distinct imprint on the genetic diversity of infections. This is often visualized in the shape of the virus **phylogeny** - a tree that models how infections are related through common ancestors.

Phylodynamics is the study of how to infer epidemiological parameters from phylogenetic tree shapes.

For example, BEAST is a popular software package that reconstructs the dynamics of population size from the spacing between ancestral nodes (the coalescence times) in trees.

These software require the ability to calculate the exact likelihood of a data set for a given model. This requirement limits us to relatively simple epidemic models that make unrealistic assumptions.

Approximate Bayesian computation (ABC) is a radical departure from conventional inference methods in that it no longer requires the exact likelihood. We fit a model by using it to simulate data sets under different parameters, until it yields simulations that resemble the observed data.

OBJECTIVES

- To develop a new simulation-based framework for phylodynamic inference.
- To validate this framework against BEAST2 on simulated data sets.
- To apply the framework to study the HIV CRF07_BC epidemic in China.

METHODS

In the context of phylodynamic inference, ABC has two requirements:
(1) the ability to simulate trees from the model; and
(2) some measure of the similarity between the simulated trees and the observed tree.

Trees can be simulated from an epidemic model using both forward-time and reverse-time (coalescent) methods. MASTER is a versatile program for forward-time simulation of trees that is distributed with BEAST2. To facilitate comparison with BEAST2, I used this program to simulate trees for ABC.

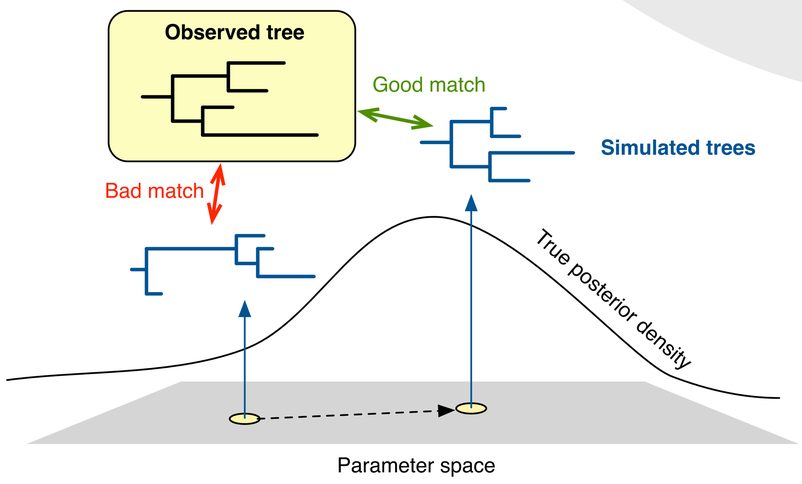
To compare trees, I used a kernel-based similarity measure developed in previous work (**Poon et al, 2014**). A kernel function maps complex objects into a more mathematically-convenient space. The intuition behind this particular kernel function is that it breaks down each tree into fragments, and then counts the number of times each kind of fragment appears in both trees.

I implemented the kernel function for ABC (kernel-ABC) in a Markov chain Monte Carlo (MCMC) framework. This generates a sample of model parameters by a random walk over the posterior distribution. I used Metropolis-Hastings sampling with full dimensional updates from the proposal distribution.

All methods and data can be obtained at http://github.com/ArtPoon/kamphir

Figure 1 - Schematic of kernel-ABC method.

A random walk is more likely to accept a proposed change in model parameters if the resulting model generates tree simulations that resemble the observed tree more closely. The distribution of parameters visited by the walk eventually approximates the posterior density.



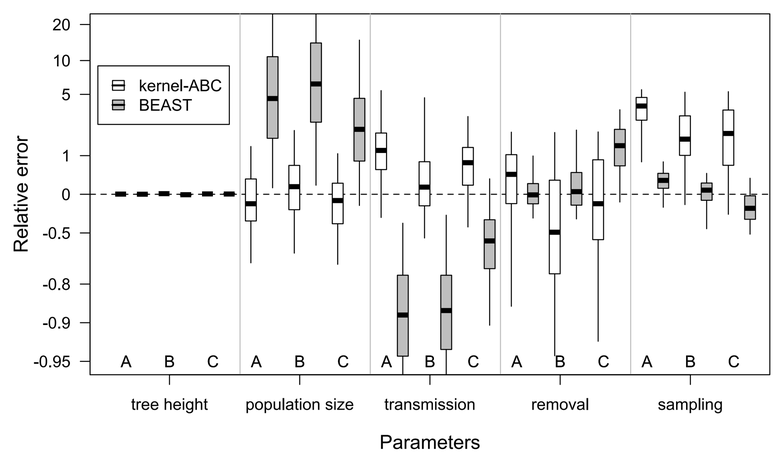
RESULTS

Transmission trees were simulated under a birth-death susceptible-infected-recovered (BDSIR; **Kühnert et al. 2014**) model under 3 epidemic scenarios (A-C).

Virus sequences were simulated on these trees and analyzed with both kernel-ABC and the serial BDSIR method in the BEAST2 phylodynamics module.

Phylogenies were reconstructed from these data by maximum likelihood (RAxML) and then rooted and time-scaled by root-to-tip regression.

Figure 2 - Method comparison

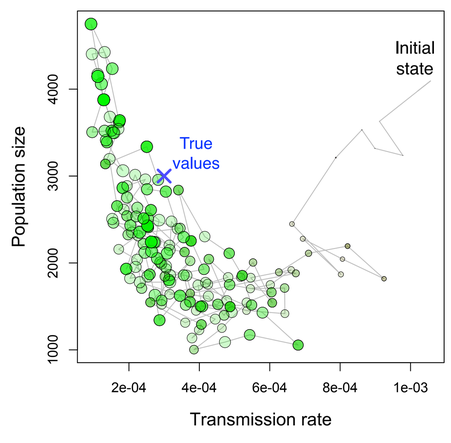


Accuracy was quantified by relative error (E/A – 1), where E is the estimated value and A is the actual value.

BEAST2 tended to overestimate the population size (N) by about 5-fold and underestimate the transmission rate (β) by 10-fold. However, it was highly effective at estimating the lineage death parameters (rates of removal, γ, and sampling, ϕ).

kernel-ABC performed relatively well on lineage birth parameters (N and β) but did not yield reliable estimates of sampling rate, ϕ.

Figure 3 - Example kernel-ABC run on scenario B



The simulated tree under scenario B had 300 tips. The true parameter values are indicated by a blue 'X'.

Each point represents a step in the chain sample. Point sizes were scaled to the kernel score.

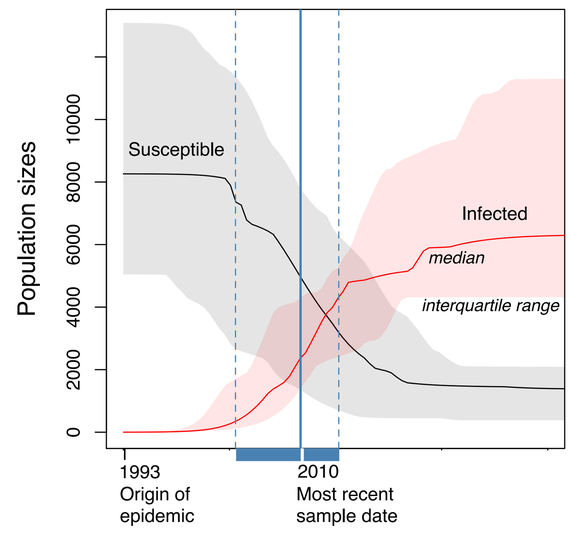
This illustrates the known problem of confounding between N and β for the BDSIR model.

APPLICATION TO HIV DATA

HIV CRF07_BC is the predominant variant among injection drug users in northwestern China (Xinjiang). It was first identified in 1997. The origin of the epidemic has previously been mapped to about 1993 (95% highest posterior density, HPD: 1991-1995; **Tee et al. 2008**).

I obtained all published 07BC env gp120 sequences isolated in China with known years of collection (1997-2010), excluded repeated samples from the same individual (n=314) and analyzed these data with both kernel-ABC and BEAST2.

Figure 4 - Inferred epidemic dynamics of CRF07 in China



Consistent with model validation experiments, BEAST2 estimates of the BDSIR lineage-birth parameters N and β were roughly 5-fold greater and 10-fold less than estimates from kernel-ABC.

Scaling the reconstructed epidemic trajectories from simulation time to real time, the analysis suggests that this epidemic was still undergoing exponential growth by the most recent sample date.

CONCLUDING REMARKS

- Simulation experiments with a simple model indicate that kernel-ABC can be as good (and sometimes better) than the current "gold standard" methods.
- The chief advantage of kernel-ABC is that it can potentially be used to fit a much broader range of epidemic models, since all that is needed is the ability to simulate trees.
- Ongoing work will extend the kernel-ABC method to incorporate non-genetic data (treatment status, risk factors) and simulate network structures.

Kühnert, D et al. 2014. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. J Roy Soc Interface 11: 20131106.
Poon, AFY et al. 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. PLOS ONE 8: e78122.
Tee, KK et al. 2008. Temporal and spatial dynamics of human immunodeficiency virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. J Virol 82: 9206-9215.

This work was supported by grants from the Canadian Institutes for Health Research (CIHR HOP-111406 to AFYP). AFYP is supported by a CIHR New Investigator Award and by a Career Investigator Scholar Award from a partnership between the Michael Smith Foundation for Health Research, St. Paul's Hospital Foundation, and the Providence Health Care Research Institute.