PopABC: a program to infer historical demographic parameters

Joao S. Lopes1,*, David Balding2 and Mark A. Beaumont1

1School of Biological Sciences, University of Reading, Whiteknights, PO Box 228, Reading RG6 2AJ and 2Department of Epidemiology and Public Health, Imperial College, St Mary’s Campus, Norfolk Place, London W2 1PG, UK

Received on March 17, 2009; revised on August 6, 2009; accepted on August 10, 2009

ABSTRACT

Summary: PopABC is a computer package for inferring the pattern of demographic divergence of closely related populations and species. The software performs coalescent simulation in the framework of approximate Bayesian computation (ABC). PopABC can also be used to perform Bayesian model choice to discriminate between different demographic scenarios. The program can be used either for research or for education and teaching purposes.

Availability and Implementation: Source code and binaries are freely available at http://www.reading.ac.uk/~sar05sal/software.htm. The program was implemented in C and can run on UNIX, MacOSX and Windows operating systems.

Contact: joao.lopes@reading.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

With the development of new techniques in molecular biology, it is possible to scan the entire genomes of multiple individuals within a population in a dense map of SNP markers, microsatellite markers or DNA sequence data. These data can then be used to infer aspects of the history of a population, such as population sizes and migration rates, recombination rates and selection coefficients. The Bayesian statistical paradigm offers an efficient framework for such inferences, because it allows maximal extraction of information from data under the specified model, and background information can be incorporated via prior distributions.

The spread of the use of MCMC (Markov Chain Monte Carlo) methods in the early 1990s made possible accurate statistical inferences from molecular genetic data (Wilson and Balding, 1998). Further increases in the volume of data as well as the ambitions of researchers have quickly exceeded the capabilities of these computer-intensive methods. In the past few years, some developments in computational statistics have arisen that pushed the boundaries of the models that can be analysed at the cost of some approximation (Marjoram and Tavaré, 2006; Pritchard et al., 1999).

These developments typically consist of the characterization of the data by summary statistics and the use of Monte Carlo simulations that avoid the need for an explicit likelihood function. Bayesian versions of these computational methods have come to be known as ABC (approximate Bayesian computation: Beaumont et al., 2002; Marjoram et al., 2003; Sisson et al., 2007). These studies suggest that ABC can compare well with methods based on fully specified likelihood functions.

We present a computer package that enables the user to perform ABC-based inference on problems in population genetics that involve multiple populations, with or without migration events. The model allows multi-locus microsatellites or sequence data to be used either separately or jointly. Recombination events can also be incorporated. Inferences regarding the choice between different branching histories (topologies) of population models can be performed automatically for up to four populations. The developed software is especially aimed for phylogeographic problems with well-defined panmictic populations. The species can have any ploidy, sexual or asexual reproduction and short or long generation times.

The package includes a coalescent simulator, based on routines described in Beaumont (2008). We have tested these routines against theoretical expectations from a coalescent model and by performing comparisons with established coalescent simulators [ms, Hudson, 2002; SIMCOAL, Laval and Excoffier, 2004 (using REJECTOR, Jobin and Mountain, 2008)] (See Supplementary Material).

ABC techniques are becoming more widely used in population genetics and molecular biology, and some general packages are currently available (Cornuet et al., 2008; Hickerson et al., 2007). Indeed, the ABC approach is easily 'pipelined', as discussed in Section 2, and a number of simulation packages have been expanded to allow them to be used for inference [e.g. ms, Hudson, 2002; Serial SimCoal, Anderson et al., 2005 (based on SIMCOAL, Laval and Excoffier, 2004)]. The 'pipelining' approach is extremely flexible, and the preferred option for more computationally adept users. However, there is a strong case for friendlier packages, which has motivated the development of the DIY-ABC package by Cornuet et al. (2008). PopABC complements DIY-ABC in a number of ways. It has the option to use a menu-based console application; the possibility to include recombination events either in linked microsatellites or in sequence data; and the option to add migration events to the population models.

2 MODELS

2.1 Statistical model

Although there are a number of different flavours of ABC (Blum and Francois, 2008; Sisson et al., 2007), PopABC follows the standard rejection/regression approach, which has been frequently

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
Y-or X-linked data, can be used separately or jointly. (Blum and Francois, 2008). Approaches such as the use of Sequential Monte Carlo sampling have the advantage of allowing the user to apply more recently developed methods like linear regression to correct the simulated points (Beaumont, 2008).

In addition, the output files have been designed so that other programs can be used to estimate the posterior probability for the priors chosen for the parameters in the respective models. Note that Bayesian model choice can be highly dependent on the priors chosen for the parameters in the respective models. We would like to thank E. Bazin, L. Chikhi, J.-M. Cornuet, K. Dawson, V. Sousa, and D. Welch for helpful advice and discussions. We acknowledge three anonymous reviewers and editor for very useful comments which highly contributed to raise the quality of the software.

Funding: Engineering and Physical Sciences Research Council; Fundacao Ciencia e Tecnologia.

Conflict of Interest: None declared.

REFERENCES

J.S.Lopes et al.

tested against full-likelihood alternatives (Excoffier et al., 2005; Hickerson et al., 2006; Sousa et al., 2009; Tallmon et al., 2004). The base algorithm (Pritchard et al., 1999) can be summarized as:

1. sample parameters, \( \Phi \), from the priors: \( \Phi \sim p(\Phi) \);
2. simulate data, \( D \), given \( \Phi \); \( D \sim p(D | \Phi) \);
3. summarize \( D \) with a set of chosen summary statistics to obtain \( S \); go to step 1 until \( N \) sample points from the joint distribution \( p(S, \Phi) \) have been created;
4. accept the points whose \( S \) is within a distance \( \lambda \) from \( s^* \), the real data summarized by the same set of summary statistics, \( |S_i − s^*| < \lambda \).

PopABC implements these four base steps of the ABC algorithm. In addition, the output files have been designed so that other subsequent steps such as the use of a weighting scheme and a local linear regression to correct the simulated points (Beaumont et al., 2002) can be later applied by the use of simple R scripts (available in Beaumont’s webpage http://www.rubic.rdg.ac.uk/~mah/). Although this latter step still involves an element of pipelining, it has the advantage of allowing the user to apply more recently developed approaches such as the use of Sequential Monte Carlo sampling (Beaumont et al., 2009; Sisson et al., 2007) and non-linear regression (effective population size; time of splitting events between sister populations, and also an equilibrium estimator of the number of migrants). For DNA sequences: the mean pairwise difference; the number of segregating sites; the number and frequency of private segregating sites; the number of different haplotypes; an index of haplotype diversity; the mean and SD of the mutation frequency spectrum; and an FST-based estimator of the number of migrants. Summary statistics are computed for each population, and then for all pairs of populations combined as described in Beaumont (2008).

3 ASSUMPTIONS
While assuming ‘Isolation with Migration’ the following should be verified (Nielsen and Wakeley, 2001): there should not be other populations that are more closely related to the sampled populations than they are to each other; and there should not be unsampled populations exchanging genes with the studied populations or their own ancestors.

Other assumptions come from the coalescent method employed: the variation within the DNA data has to be neutral; free recombination between DNA sequence loci is assumed; and mutations follow the mutation model considered.

ACKNOWLEDGEMENTS
We would like to thank E. Bazin, L. Chikhi, J.-M. Cornuet, K. Dawson, V. Sousa, and D. Welch for helpful advice and discussions. We acknowledge three anonymous reviewers and editor for very useful comments which highly contributed to raise the quality of the software.

Funding: Engineering and Physical Sciences Research Council; Fundacao Ciencia e Tecnologia.

Conflict of Interest: none declared.

REFERENCES


