MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 723–727 doi: 10.1111/j.1755-0998.2010.02832.x

COMPUTER PROGRAM NOTE

msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis

P. PAVLIDIS,* S. LAURENT* and W. STEPHAN
Department of Biology II, Section of Evolutionary Biology, University of Munich, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

Abstract

With the availability of whole-genome sequence data biologists are able to test hypotheses regarding the demography of populations. Furthermore, the advancement of the Approximate Bayesian Computation (ABC) methodology allows the demographic inference to be performed in a simple framework using summary statistics. We present here msABC, a coalescent-based software that facilitates the simulation of multi-locus data, suitable for an ABC analysis. msABC is based on Hudson's ms algorithm, which is used extensively for simulating neutral demographic histories of populations. The flexibility of the original algorithm has been extended so that sample size may vary among loci, missing data can be incorporated in simulations and calculations, and a multitude of summary statistics for single or multiple populations is generated. The source code of msABC is available at http://bio.lmu.de/~pavlidis/msabc or upon request from the authors.

Keywords: Approximate Bayesian Computation, coalescent simulations, demographic inference, population genetics

Received 13 November 2009; revision received 17 December 2009; accepted 23 December 2009

Introduction

Along with the increase in population genomic data sets, an important goal is to understand the relationship between patterns of nucleotide polymorphism in natural populations and their evolutionary history. Statistical methods have been employed to estimate demographic parameters using likelihood approaches (Kuhner 2006; Hey & Nielsen 2007) or analysing summary statistics of the data. Among them, Approximate Bayesian Computation (ABC) benefits from the increase in both available data and computer power (Beaumont *et al.* 2002; Excoffier *et al.* 2005). ABC is applied widely in population genetics studies and usually consists in a two-step procedure. First, simulations are used to sample from the joint distribution of parameters and summary statistics of the simulated data for a given demographic model. Then, a rejection algorithm is applied to retain only values of parameters that generate summary statistics which are similar to the observed values. The retained set of parameter values is then corrected by local linear regression (Beaumont *et al.* 2002) or nonlinear regression (Blum & François 2009) and considered as an approximation of the posterior distribution. Hudson's (2002) ms is a widely used coalescent software that generates neutral polymorphism data for a genomic locus sampled from one or more populations undergoing complex demographic scenarios (including past population size changes, merging of populations and migration). Furthermore, it is computationally efficient for relatively large samples (hundreds or thousands of chromosomes) as well as large genomic segments (tens to a few hundred kilobases).

Here, we propose an extension of ms to facilitate its usage within ABC and, in particular, to perform the sampling procedure. Our aim is to provide a software, named msABC, that (i) draws parameter values from user-specified prior distributions, (ii) allows to choose from a variety of summary statistics, (iii) can be used for multiple unlinked loci and (iv) enables the calculation of summary statistics in cases of incomplete information (i.e. missing data). The randomly drawn parameter values are used to perform coalescent simulations. The simulated data are summarized into a vector of summary statistics and written to a file. This file can then be used to perform the

Correspondence: Pavlos Pavlidis, Fax: +49 89 2180 74104;
E-mail: pavlidis@bio.lmu.de
*These authors have contributed equally

rejection step and the other postsampling adjustments using linear regression (Beaumont *et al.* 2002; Thornton 2009) or nonlinear regression models (Blum & François 2009).

Methods

Generation of data

Currently, ms allows to simulate neutral polymorphism data using a set of constant, user-defined parameter values. Alternatively, employing the tbs option, ms permits some of the parameters to be specified from the standard input. However, even in this case, the parameter values should be generated *a priori*. This may be tedious when many parameters need to be sampled from one or more distributions. msABC enables the user to specify in the command line the desired sampling distributions (uniform, normal, log-normal, gamma) for the parameters of interest. For each simulated data set, new parameter values (e.g. the population mutation parameter θ) are drawn from the specified distribution. In msABC, a data set may consist of multiple independent loci. The sample size is allowed to vary among loci, similarly to the msnsam program (Ross-Ibarra *et al.* 2008), as is often the case in large genome re-sequencing projects. Furthermore, the simulation of missing information is possible.

Calculation of summary statistics

Following the generation of a data set, summary statistics are calculated: (i) estimates of variability such as the Watterson's estimator θ_W (Watterson 1975), or equivalently the number of segregating sites S, the average pairwise differences of sequences θ_π (Tajima 1983), (ii) summaries of the site frequency spectrum such as D (Tajima 1989) and H (Fay & Wu 2000) and (iii) summaries based on linkage disequilibrium (LD), i.e. the average pairwise correlation coefficient ZnS (Kelly 1997). Population differentiation statistics such as F_{ST} (Hudson *et al.* 1992a,b; Slatkin 1993) or pairwise F_{ST} have been implemented for the case of multiple population data sets. Furthermore, fixed differences, shared and private polymorphisms can be calculated between pairs of populations. When data sets are composed of multiple populations, summary statistics i - iii are calculated for each population as well as for the pooled sample. Summary statistics are calculated for each locus, and averages and variances are reported if multiple loci are simulated.

Simulations with incomplete information

Often, in Sanger re-sequencing (e.g. Hutter *et al.* 2007), microchip sequencing (e.g. http://www.dpgp.org/) or high-throughput sequencing projects (e.g. http://www.dpgp.org/), a fraction of data contains missing information, i.e. nonidentified nucleotides symbolized as 'N'. Missing data affect the values of summary statistics (e.g. they decrease variability), and therefore may bias the demographic inference. In msABC, one can simulate missing data by specifying the coordinates (position and sequence) of each 'N' in the alignment. In a simulated data set, segregating sites coinciding with the position of 'N' in the alignment are replaced by the missing state. The sample size of each site is then updated and the calculation of summary statistics is adapted. Details and examples are provided in the manual (pg. 12).

Code availability

The source code and documentation of msABC is available at http://bio.lmu.de/~pavlidis/msabc or upon request. msABC has been compiled and run on 32-bit Linux machines with the gcc (version 4.2.4) compiler and on 64-bit Linux machines with the gcc (version 4.1.2) compiler.

Results

The sampling process of an ABC analysis may consist of multiple steps. Parameter values are sampled from the prior distribution to simulate polymorphism data. Coalescent simulation programs such as simcoal2 (Laval & Excoffier 2004) and ms (Hudson 2002) have been used extensively for the data generation. Then, the summary statistics are calculated using the simulation results in appropriate software packages [e.g. the libsequence library (Thornton 2003)]. msABC integrates these steps into one software package that efficiently performs the sampling process of the ABC.

The benefits from this integration are (i) it allows researchers without extensive coding skills to estimate demographic models even for complicated scenarios, when the sample size of loci varies or the data set includes missing information and (ii) computations are considerably faster than combining sequentially the steps of the sampling process mentioned in the previous paragraph (Fig. 1).

Speed measurements

We compared the speed performance of msABC with the combination of the coalescent simulator ms (Hudson 2002) and the libsequence library (Thornton 2003) to calculate summary statistics. msABC out-competes this combination. As illustrated in Fig. 1, msABC (solid line with circles) is compared with the combination ms-libsequence (dashed line with crosses). Hudson's ms (dashed

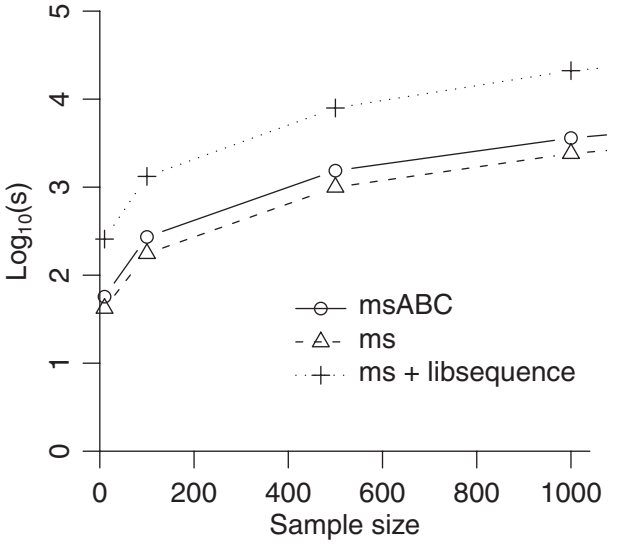

Fig. 1 Speed comparison (in $\log_{10}$ seconds) when the sample size is between 10 and 1000. msABC is about six times faster than the combination of ms with libsequence when the sample size equals 1000.

line with triangles) is used as lower bound for the time, because it simulates data without calculating summary statistics. Simulations refer to a demographic scenario of two populations with gene flow between them ($4Nm$=0.5, where N is the present day effective population size and m is the fraction of each subpopulation made up of new migrants each generation), with a (global) population contraction to $0.3N$ at time 0.01 (backwards in units of $4N$). A genome of 100 independent loci is simulated 1000 times (θ = 10 per locus, ρ = 10 per locus). The set of summary statistics consists of θ_W, θ_π, D, H, F_{ST}, shared and fixed polymorphisms and ZnS. The speed difference is important especially for large (whole-genome) data sets [e.g. the 1001 genome project for *A. thaliana* (Weigel & Mott 2009)], where simulations may require extensive time periods. For example, based on Fig. 1, simulating a genome of 100 independent loci 10^6 times when the sample size is 500 would require about 92 days on a single computer using ms-libsequence. On the other hand, msABC would require 17 days for the same computations.

Example of parameter estimation

We infer the parameters of a simple demographic model characterized by two diverging populations with recombination, to illustrate the usage of msABC. The model consists of three parameters: the population mutation parameter θ, which is identical in the present and ancestral populations; the time τ at which the two populations diverged; and the population recombination rate ρ. We used msABC to sample parameter values from uniform priors U(0; 10) and U(0; 1) for θ and τ, respectively, to simulate polymorphism data set under this demographic model and to summarize these data sets into a series of summary statistics. In all simulations, ρ = 20 and the simulated data sets consist of 50 loci of 500 bp with sample size n = 12. The output of msABC allows to investigate the relations between the parameters θ and τ and summary statistics of the simulated data. Figure 2a illustrates the relation between τ and the amount of differentiation between the two populations as it is measured by F_{ST}. Strong correlations between parameters and summary statistics indicate that summary statistics can be used to infer values of demographic parameters (Fig. 2a). Posterior distributions of parameters based on observed summary statistics, the joint distribution of parameters and simulated summary statistics can be computed from the output of msABC and the rejection/regression analysis (Beaumont *et al.* 2002; Excoffier *et al.* 2005; Thornton 2009). To illustrate this estimation procedure, we simulated a data set by setting θ = 5 and τ = 0.5 and re-estimated the values of θ and τ using 10^6 simulated data sets. Posterior distributions were

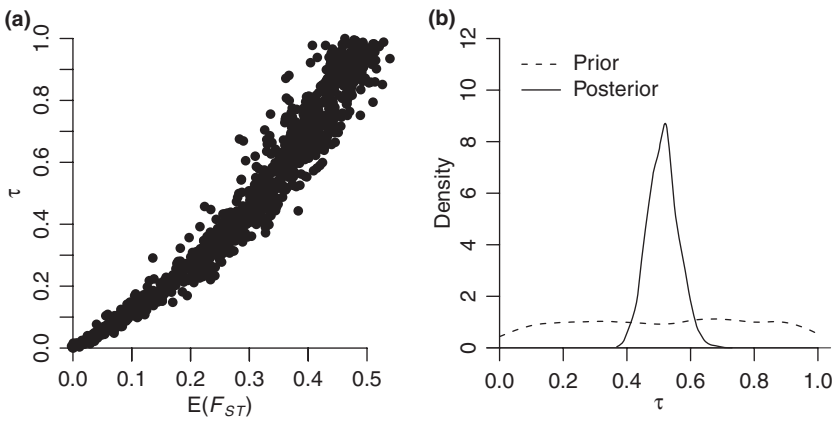

Fig. 2 Results obtained from msApproximate Bayesian Computation (msABC). a) Examining the relationship between the parameter τ and the summary statistic $E(F_{ST})$. b) Posterior distribution of the parameter τ, obtained after applying the output of msABC in algorithms that perform the rejection and regression steps of ABC analysis.

estimated by summarizing the data into the mean and the variances of the number of segregating sites, *D*, *ZnS* and F_{ST}. Figure 2b illustrates the posterior distribution of the parameter τ.

Discussion

msABC facilitates the sampling process of an ABC analysis. The command line is similar to the command line of ms, thus shortening the learning curve for a user who is familiar with ms.

Although msABC can be used to simulate single loci, most demographic analyses in molecular population genetics are characterized by large data sets composed of several chromosomal fragments scattered along the genomes (Nordborg *et al.* 2005; Ometto *et al.* 2005; Hutter *et al.* 2007). msABC can simulate multi-locus data sets, where each fragment is characterized by its own length, sample size, recombination rate and mutation rate. msABC provides a collection of commonly used summary statistics that allow to quantify levels of polymorphisms, LD, population differentiation and the shape of the frequency spectrum of derived mutations. The complete list of available summary statistics can be found in the user's manual (see http://bio.lmu.de/~pavlidis/msabc).

Furthermore, msABC extends the flexibility of Hudson's ms by allowing variable sample size among fragments and missing data simulation. It allows to analyse data sets that contain missing data by simulating them and then calculating summary statistics. This may be important in demographic inference of large data sets which typically consist of a large amount of incomplete information (e.g. http://www.dpgp.org/).

The speed performance can be important for large data sets. Assuming that simulating data of tens or hundreds of kb (with typical values of recombination rates) for a sample that consists of hundreds or thousands of individuals may require months of computational time, an improvement of five to six times shortens considerably the time of the inference project. This is especially true if the project is carried out on personal computers instead of cluster machines (Fig. 1).

Alternative ways to obtain summary statistics from simulated data could be implemented by replicating ms commands with different parameters. In the best case, this would require extensive scripting for calculating the priors and summary statistics. However, when missing data are included in the data set or the sample sizes of loci vary, it would not be possible to perform simulations that match the observed data.

msABC can be used to examine the relationship between parameters and summary statistics (Fig. 2a). This helps to inspect the usability of certain summary statistics in estimating parameters. Summary statistics that are related monotonically to target parameters are expected to be useful for estimating them. Additionally, msABC can be used to obtain the null distributions of a multitude of summary statistics under demographic models.

msABC outputs samples from the joint distribution of parameters and summary statistics under a given demographic model. A follow-up step in the analysis (rejection) retains the closest points to the observed data. The parameter values that have been used to generate those simulations comprise an approximation of the true posterior distribution of the parameters of interest. An improvement of this approximation has been proposed by Beaumont *et al.* (2002) that corrects for the fact that the accepted simulations never match precisely the observed data (linear regression). A more sophisticated approach has been suggested by Blum & François (2009). msABC does not perfom the rejection and regression steps. Algorithms needed to perform these postsampling steps have been implemented elsewhere [e.g. abcReg by K. Thornton (http://www.molpopgen.org/software/abcreg) or nonlinear regression models by Blum & François (2009) (http://membres-timc.imag.fr/Michael.Blum/my_publications.html)].

A critical point in ABC refers to the model choice (Pritchard *et al.* 1999; Fagundes *et al.* 2007). Typically, different demographic scenarios are simulated, and the scenario with the highest relative posterior probability is then used (Fagundes *et al.* 2007; François *et al.* 2008). However, this model does not necessarily provide a good fit to the observed data, because it simply indicates the best model among the tested models (Ratmann *et al.* 2009). Therefore, once the best model and its parameters have been inferred, it is necessary to investigate whether simulations under this model are able to predict the observations (predictive simulations).

Finally, in ABC, the set of summary statistics may be crucial. It has been shown that uninformative summary statistics add noise to the distance between simulations and observations (Joyce & Marjoram 2008), thus they should be avoided. Therefore, the smallest set of summary statistics that captures the information carried by the data set should be used. The choice of summary statistics is an active area of research. Joyce & Marjoram (2008) suggested a scheme for scoring statistics according to whether they improve the inference. Alternatively, Wegmann *et al.* (2009) proposed partial least square regression (Boulesteix & Strimmer 2007) to reduce the dimensionality. In Table 1, we suggest which summary statistics should be used to infer certain demographic parameters. However, because the information provided by statistics may vary between demographic scenarios, investigating the relationship between them and

Table 1 Demographic parameters and population genetics summary statistics that can be used for the inference of parameter values. Summary statistics are described in the section Calculation of summary statistics

Demographic parameters	Summary statistics
θ (population mutation rate)	θ_W (or S), θ_π
ρ (population recombination rate)	*ZnS*
Time of population size expansion	θ_W (or S), θ_π, *D*
Time of population size contraction	θ_W (or S), θ_π, *D*, *ZnS*
Magnitude of population size change	θ_W (or S), θ_π, *D*, *ZnS*
Migration rate (island model)	F_{ST}
τ (time of divergence between two populations)	Pairwise F_{ST}

demographic parameters under various demographic scenarios of interest is necessary.

Acknowledgements

We thank Dirk Metzler, Stephan Hutter and Aurelien Tellier (LMU Munich) for useful discussions. This work is supported by grants from the Volkswagen-Foundation (I/82770) to PP and by the DFG (Ste 325/5-3 and 325/12-1) to WS.

References

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Blum M, François O (2009) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*. Available from http://dx.doi.org/10.1007/s11222-009-9116-0.

Boulesteix A, Strimmer K (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44.

Excoffier L, Estoup A, Cornuet J (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Fagundes NJR, Ray N, Beaumont M *et al.* (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17614–17619.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

François O, Blum MG, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, **18**, 337–338.

Hudson R, Boos D, Kaplan N (1992a) A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution*, **9**, 138–151.

Hudson RR, Slatkin M, Maddison WP (1992b) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide single nucleotide polymorphism data. *Genetics*, **177**, 469–480.

Joyce P, Marjoram P (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7**, Article26.

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kuhner MK (2006) LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics (Oxford, England)*, **22**, 768–770.

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics (Oxford, England)*, **20**, 2485–2487.

Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution*, **22**, 2119–2130.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

Ratmann O, Andrieu C, Wiuf C, Richardson S (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10576–10581.

Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.

Slatkin M (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, **47**, 264–279.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics (Oxford, England)*, **19**, 2325–2327.

Thornton K (2009) Automating approximate Bayesian computation by local linear regression. *BMC Genetics*, **10**, 35.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Weigel D, Mott R (2009) The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biology*, **10**, 107.