Ecology, 95(5), 2014, pp. 1418–1428

An approximate Bayesian computation approach to parameter estimation in a stochastic stage-structured population model

Katherine Scranton,[1] Jonas Knape, and Perry de Valpine

University of California, Berkeley, Department of Environmental Science, Policy, and Management, 130 Mulford Hall #3114, Berkeley, California 94720-3114 USA

Abstract. Complex population processes may require equally complex models, which can lead to analytically intractable estimation problems. Approximate Bayesian computation (ABC) is a computational tool for parameter estimation in situations where likelihoods cannot be computed. Instead of using likelihoods, ABC methods quantify the similarities between an observed data set and repeated simulations from a model. A practical obstacle to implementing an ABC algorithm is selecting summary statistics and distance metrics that accurately capture the main features of the data. We demonstrate the application of a sequential Monte Carlo ABC sampler (ABC SMC) to parameter estimation of a general stochastic stage-structured population model with ongoing reproduction and heterogeneity in development and mortality. Individual variation in demographic traits has considerable consequences for population dynamics in many systems, but including it in a population model by explicitly allowing stage durations to follow a realistic distribution creates a complex model. We applied the ABC SMC to fit the model to a simulated representative data set with known underlying parameters to evaluate the performance of the algorithm. We also introduced a systematic method for selecting summary statistics and distance metrics, using simulated data and receiver operating characteristic (ROC) curves from classification theory. Evaluations suggest that the approach is promising for model inference in our example of realistic stage-structured population models.

Key words: approximate Bayesian computation; cohort dynamics; individual heterogeneity; life-history variation; sequential Monte Carlo.

Introduction

Models of ecological systems walk a fine line between the complexity of model realism and the simplifications necessary for parameter estimation and model inference. In stage-structured population models, increased realism in the form of variable life-history traits and ongoing reproduction creates very difficult estimation problems. Deriving or approximating the likelihood function may not be possible, necessitating a more flexible statistical framework, such as approximate Bayesian computation (ABC). An ABC approach replaces the likelihood function with general metrics of distance between observed data and repeated simulated data sets. By favoring simulations that are similar to the observed data, the algorithm moves from a prior distribution to the approximate posterior distribution.

The relatively recent development of the ABC framework has been applied to problems in a variety of fields, including ecology (Tavare et al. 1997, Pritchard et al. 1999, Beaumont 2010). ABC methods are used extensively in population genetics to infer demographic parameters from molecular variation in a sample (Csilléry et al. 2010). Parameters of interest include coalescence time (Tavare et al. 1997), crossing-over rates (Padhukasahasram et al. 2006), expansion times and migration rates (Hamilton et al. 2005), and mutation rates (Bazin et al. 2010). Estimation problems in conservation genetics (Rabosky 2009, Lopes and Boessenkool 2010), epidemiology (Blum and François 2008), systems biology (Barnes et al. 2011), human demography (Shriner et al. 2006), and psychology (Turner and Van Zandt 2012) have also been solved with the use of ABC. ABC methods are slowly filtering into the field of ecology and have been used to evaluate models of forest diversity (Jabot and Chave 2009), Lotka-Volterra dynamics (Toni et al. 2009), host–parasite dynamics (Drovandi et al. 2011), and range expansion (Rasmussen and Hamilton 2012).

The ABC framework has the potential to fill a gap in our ability to fit models in population ecology (Beaumont 2010, Hartig et al. 2011). However, a large obstacle to implementing any ABC algorithm is the comparison of an observed and simulated data set, using summary statistics and distance metrics. For those fields in which ABC has been used extensively, summary statistics can be chosen from previous studies, but in novel applications of ABC, the choice of summary statistics and distance metrics is left to the researcher's

Manuscript received 6 June 2013; revised 10 October 2013; accepted 15 October 2013; final version received 20 November 2013. Corresponding Editor: B. D. Inouye

[1] E-mail: katherine.scranton@yale.edu

own knowledge of the system (Marjoram et al. 2003). The most common approach is to create training sets of simulated data for pilot ABC runs with different combinations of distance metrics and summary statistics, and compare the results (Li and Jakobsson 2012). This pilot run approach amounts to trial and error ABC until summary statistics are found that perform adequately, with extremely high computation costs. Thus, for practical application to any population dynamics problems, there is a need for better methods to select summary statistics and distance metrics.

There are a variety of ABC algorithms (Beaumont et al. 2002, Marjoram et al. 2003), including those based on sequential Monte Carlo (SMC), or particle filter, samplers (Cappé et al. 2004, 2007), which we adopt for our population time-series data. Fitting stochastic models of stage-structured populations to cohort data with reproduction is a very difficult estimation problem for which there are no good methods available. Life histories can be highly variable (Benton et al. 2005) but existing models lack individual heterogeneity in life-history processes and assume unrealistic stage durations (Birt et al. 2009, de Valpine 2009). Previous models have been fit to cohort data with independent samples at each time and without reproduction (Read and Ashford 1968, Manly 1990, Hoeting et al. 2003). The temporal nonindependence created by counting an entire population of unmarked individuals through time, as well as ongoing reproduction, make the estimation problem considerably harder. Likelihoods with such data for more realistic models that include individual heterogeneity, cannot be calculated or approximated, eliminating maximum likelihood methods for inference. The more flexible statistical framework of ABC could allow parameter estimation, model inference, and model selection.

We present a novel application of the ABC framework and a systematic method for evaluating potential summary statistics and distance metrics a priori. We fit a stochastic stage-structured model that explicitly includes individual heterogeneity in development and survival and allows for ongoing reproduction. We hold two of the parameters constant, fixing the shape of the distributions of stage durations, as a first step towards a very difficult estimation problem. Although the model is simple, fitting it to nonindependent cohort data has not been successfully done in the past. We use an ABC SMC algorithm with an approximated optimal backwards kernel in the weights and an adaptive threshold schedule for parameter inference. We provide specific descriptions of the population model, the method for selecting summary statistics and distance metrics, and the ABC SMC algorithm. We illustrate the performance of our estimation method with simulated data and provide a discussion of the potential of ABC for population ecology and the obstacles that still remain.

Methods

Stochastic stage-structured population model

We consider a stage-structured population, where survival times and stage durations follow specified distributions, allowing for variation between individuals. Data on such populations commonly result from studies of cohorts: groups of individuals born at the same time and monitored over a fixed time period. Observation is rarely done continuously, so the exact times of stage transitions, births, and deaths are unknown. Instead, the data are interval censored; individuals are observed at set points. Individuals are unmarked, and all are counted at each observation point, so that observations are not temporally independent.

This population is closed to dispersal, and individuals have three life stages: egg, immature, and adult. The duration of the egg stage (t_e) and the duration of the immature stage (t_i) follow Weibull distributions with scale and shape parameters ($1/\lambda_e$, γ_e), ($1/\lambda_i$, γ_i) respectively:

$$f(t_e) = \lambda_e \gamma_e (\lambda_e t_e)^{\gamma_e - 1} \exp\Big(- (\lambda_e t_e)^{\gamma_e} \Big) \quad (1)$$

$$f(t_i) = \lambda_i \gamma_i (\lambda_i t_i)^{\gamma_i - 1} \exp\Big(- (\lambda_i t_i)^{\gamma_i} \Big). \quad (2)$$

Adults experience mortality (m) such that survival times (t_s) follow an exponential distribution

$$f(t_s) = m e^{-m t_s}. \quad (3)$$

Adults reproduce at a constant rate (r) from maturation until death.

In order to simulate a data set, we generate exact stage durations and exact survival times for a cohort of N_0 individuals that found a population and for all of their offspring as they reproduce. We sample the population every two time steps, by counting individuals according to stage, ignoring the exact ages. This yields a data set of the number of individuals in each stage at each sampling time, [0, 2, . . . , 10]. The scenario encompasses two generations, but still makes the assumption of density-independent dynamics realistic.

The full stochastic stage-structured model has six parameters $\{\lambda_e, \gamma_e, \lambda_i, \gamma_i, m, r\}$. For the purposes of investigating an estimation method, we fix the shape parameter of egg-stage duration ($\gamma_e = 6$) and the shape parameter of immature-stage duration ($\gamma_i = 6$). Once we are satisfied as to estimation ability and model performance, future work may extend the use of the method to problems with more parameter dimensions.

Summary statistics and distance metrics

ABC algorithms rely on summary statistics and distance metrics to calculate the distance between an observed and simulated data set. We define a summary statistic, $S(\mathcal{D})$, as a value computed from one data set

that may summarize or capture some information about the observation. We define a distance metric, $\rho(\mathcal{D}, x_i)$ or $\rho(S(\mathcal{D}), S(x_i))$, as a function of two data sets (or two summary statistics) that may capture some information about the difference between observations. Because cohort data has different dimensions (abundance, stage, time), we may wish to include very different classes of summary statistics and distance metrics. Because it will not be possible to combine all summary statistics with all distance metrics from these different classes, we define a distance function as a combination of a summary statistic and a distance metric. Instead of calculating a single distance between two data sets, we incorporate distances from several distance functions into the ABC SMC algorithm, allowing us to make distinctions over the different dimensions of the data. We developed this procedure to select distance functions that discriminate between parameters close to the truth and far from the truth, without resorting to pilot ABC runs with high computation costs (Fig. 1).

As a first step, we compile a large set of candidate summary statistics and distance metrics from the literature, from problems similar to those faced in ABC, and by considering the nature of the data and model at hand (Table 1, full description in Appendix A). We simulate data from different underlying parameters (Fig. 1a) and calculate distances between pairs of data sets using different distance functions (Fig. 1b). In order to select the optimal subset of distance functions, we quantify the performance of each distance function using methods common in classification theory (Hastie et al. 2009).

We treat each distance function as a rule that classifies a pair of data sets as coming from either the same or different underlying parameters according to a threshold value. Any rule that classifies values into two groups (positives and negatives) will make some errors (Hastie et al. 2009). As the threshold changes, the proportion of errors will also change. A perfect classifier would have 100% true positives and 0% false positives over all thresholds, but in practice we expect a tradeoff between the two. The receiver operating characteristic (ROC) curve plots true positives vs. false positives as the threshold varies over the entire range of the data (Hastie et al. 2009). The area under the ROC curve (AUC) summarizes the information so that the perfect classifier would have an AUC of 1.0, and a rule no better than flipping a coin would have an AUC of 0.5 (Hastie et al. 2009).

For each candidate distance function, we would like to ask how well it classifies a pair of data sets as coming from either the same or different underlying parameters. A strong distance function would classify perfectly or with little error (Fig. 1d). A strong distance function would be able to discriminate between two data sets, even when the difference between their underlying parameters was small. A weak distance function would make many errors, even when a pair of data sets were generated by very different underlying parameters (Fig. 1e).

In order to apply classification theory to this problem, we simulate many different data sets from many different underlying parameters. We first select a reasonable, true parameter vector $\boldsymbol{\theta}$. We shift each parameter $\boldsymbol{\theta}_k$ in $\boldsymbol{\theta}$ one at a time, creating a grid of values (called $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$) surrounding $\boldsymbol{\theta}$. For $\boldsymbol{\theta}$ and each value of $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, we simulate 1000 data sets. For each pair of parameter vectors ($\boldsymbol{\theta}$, $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$) and each distance function, we calculate the distances between the 1000 pairs of data sets. We then plot ROC curves and calculate the AUC value for each distance function and each $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ (Fig. 1c). Continuing in this way for each $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, we can compare the distances and AUC values from one distance function, as the real differences between underlying parameter values increase (Fig. 1f, g).

We program the distance functions with original code in R and use the ROCR package in R to calculate AUC values (Sing et al. 2005, R Development Core Team 2010). We select strong distance functions with AUC values that increase to one, with increasing differences between parameter values (Fig. 1f). We make sure we have at least one distance function that performs well for each parameter. We also examine pairwise plots of distances to eliminate highly correlated distance functions.

ABC SMC algorithm

After selecting a subset of J distance functions from our candidate list, we proceed with parameter estimation using the ABC SMC algorithm (Box 1, with notation in Table 2). Instead of moving from the prior distribution to the posterior in one step, an SMC sampler introduces a number of intermediate steps. At the initial step ($s = 0$), we draw a sample of parameters from the prior, $\{\boldsymbol{\theta}_i\}^{(0)}$, which we now call particles. Each particle $\boldsymbol{\theta}_i$ is used to simulate a data set x_i. For each distance function j, the distance between the observed data ($\mathcal{D}$) and the simulated data set (x_i), $\mathbf{d}_{ij} = \rho_j(S(\mathcal{D}), S(x_i))$ is calculated. That value, $\mathbf{d}_{ij}$, is the distance from the observed data set to the ith model simulation using the jth distance function. Each particle then has a corresponding J-dimensional set of distance values, called a distance vector $\mathbf{d}_i$, and a weight w_i. In the initial step, we keep all N particles and assign them equal weight such that $w_i = 1/N$ for $i = 1, \ldots, N$.

The first ($s = 0$) $\{\boldsymbol{\theta}_i\}^{(0)}$ set of particles now has a corresponding set of simulated data sets $\{x_i\}^{(0)}$, distance vectors $\{\mathbf{d}_i\}^{(0)}$, and weights $\{w_i\}^{(0)}$. For each of the next steps ($s = 1, \ldots, S$), we resample particles from the previous set $\{\boldsymbol{\theta}_i\}^{(s-1)}$ by weights $\{w_i\}^{(s-1)}$. We perturb each resampled particle, using a transition kernel. For our model, we chose a multivariate Gaussian kernel with means equal to the parameter values and a diagonal covariance matrix. The variance for each parameter in $\boldsymbol{\theta}$ is reset at each step to twice the sample variance of

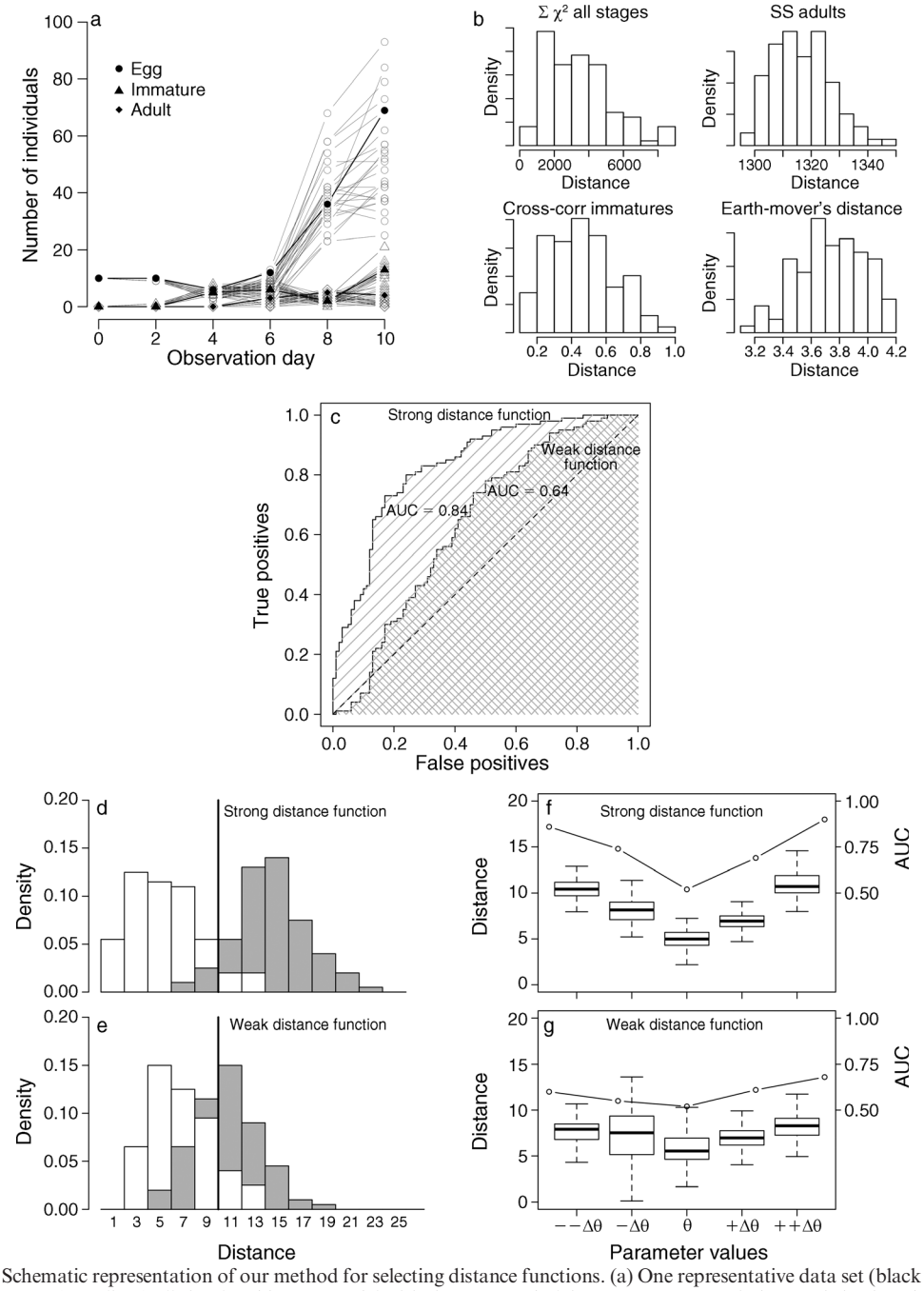

FIG. 1. Schematic representation of our method for selecting distance functions. (a) One representative data set (black line) and 25 replicate data sets (gray lines) all simulated by our model with the same underlying parameters. Each data set is broken down into the numbers of individuals in each stage at each time. (b) Distances calculated between the representative data set and 100 simulated data sets using four distance functions: the sum of χ^2 differences in counts of individuals in each stage at each time, the sum of squared differences (SS) in the number of adults at each time, the cross-correlation function (cross-corr; with zero lag) between the counts of immatures at each time, and the sum of earth mover's distance between the distributions at each time of the relative number of individuals in each stage. (c) Receiver operating characteristic (ROC) curves for two distance functions: one strong (light shading) and one weak (dark shading), calculated using the distances given in panels d and e, respectively. (d) The distribution of distances calculated by a strong distance function for pairs of data sets generated by the same underlying parameters (white) and different underlying parameters (gray). The solid black line shows a threshold that might be used to discriminate between the two groups. (e) The distribution of distances calculated by a weak distance function for pairs of data sets generated by the same underlying parameters (white) and different underlying parameters (gray). The solid black line shows a threshold that might be used to discriminate between the two groups. (f) Distances and area under the ROC curve (AUC) values for a strong distance function for five different parameter combinations. Each AUC value is calculated by comparing data sets generated by $\boldsymbol{\theta}$ to data sets generated by $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$, where $\Delta\boldsymbol{\theta}$ is a shift in the underlying parameter $\boldsymbol{\theta}$. At the center $\Delta\boldsymbol{\theta} = 0$, $\Delta\boldsymbol{\theta}$ increases to the right and decreases to the left. (g) Distances and AUC values for a weak distance function for five different parameter combinations. Each AUC value is calculated by comparing data sets generated by true parameter vector $\boldsymbol{\theta}$ to data sets generated by $\boldsymbol{\theta} + \Delta\boldsymbol{\theta}$. At the center $\Delta\boldsymbol{\theta} = 0$, $\Delta\boldsymbol{\theta}$ increases to the right and decreases to the left.

TABLE 1. Candidate summary statistics and distance functions.

Summary statistic	Distance metric
Individuals in each stage at each time; total individuals at each time	sum of χ^2
Eggs at each time; new eggs at each time; immatures at each time; adults at each time; individuals in each stage at each time; total individuals at each time; relative stage class distribution; transformed stage class distribution	sum of squared differences
Relative stage class distribution	sum of Kullback-Leibler divergence; sum of squared Kullback-Leibler divergence; sum of Bhattacharyya distance; sum of Hellinger distances; sum of squared Hellinger distances; sum of earth mover's distance
Eggs at each time; immatures at each time; adults at each time; total individuals at each time	cross-correlation function
Individuals in each stage at each time	ln $\Pr[X = x_i]$ where $X \sim$ Poisson($\mathcal{D}$)

parameter values at the previous step (Beaumont et al. 2009). Because the perturbation is local, the kernel is reset at each step, and the variance ls large compared to the particles at the previous step, we believe this choice of transition kernel is robust to correlations between parameters. The perturbed particle is used to simulate a new data set x_i, and a distance vector, $\mathbf{d}_{ij}$, is calculated using the distance functions, $\{\rho_j\}$.

Unlike the initial step, where all particles were kept, we keep only those particles that pass a threshold test. Thresholds are specific to the distance function so that we need not worry about scale or weighting of distance functions. This ensures that accepted particles produce data sets that are similar to the observation across all dimensions of the data. The jth distance function has a corresponding threshold ε_j, set to the value that would have included the particles with the smallest q_s percentage of distances in step $s-1$ (Drovandi and Pettitt 2011). For any particle $\boldsymbol{\theta}_i$, we require every $\mathbf{d}_{ij}$ in $j = 1, \ldots, J$, to be less than some threshold, ε_j. This creates a rectangular acceptance region for a two-dimensional distance vector as in Pritchard et al. (1999). We continue sampling and perturbing particles until we have N new particles that generate data sets that pass our threshold test. We then calculate the new weights $\{w_i\}^{(s)}$ using a particle weighting scheme after Beaumont et al. (2009) and move on to the next step $(s+1)$.

We fit the stage-structured population model to simulated data with the ABC SMC algorithm. The ABC SMC algorithm was programmed in R (R Development Core Team 2010). The population model was programmed and compiled in C and called from R

Box 1. Approximate Bayesian Computation (ABC) Sequential Monte Carlo (SMC) Algorithm Pseudo-code.

1) At $s = 0$ initialize
Repeat N times:
  1.1 Draw a particle $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta})$
  1.2 Simulate data $x_i \sim \mathcal{M}(\boldsymbol{\theta}_i)$
  1.3 Calculate the distance vector $\mathbf{d}_i$ for the set of J distance functions $\{\rho_j\}$: for $j = 1, \ldots, J$: $\mathbf{d}_{ij} = \rho_j [S_j(\mathcal{D}), S_j(x_i)]$
  1.4 Set weight $w_i = 1/N$
2) For each step $s = 1, \ldots, S$:
  2.1 For each distance function reset the threshold: for $j = 1, \ldots, J$: $\varepsilon_j = q_s^{th}$ quantile of $\{d_i\}_j^{(s-1)}$
  2.2 Repeat until N particles are accepted
    2.3.1 Draw $\boldsymbol{\theta}^* \sim \{\boldsymbol{\theta}_i\}^{(s-1)}$ with probabilities $\{w_i\}^{(s-1)}$
    2.3.2 Perturb the particle: $\boldsymbol{\theta}_i \sim \mathcal{K}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)$; if $\pi(\boldsymbol{\theta}_i) = 0$ return to 2.3.1
    2.3.3 Simulate data $x_i \sim \mathcal{M}(\boldsymbol{\theta}_i)$
    2.3.4 Calculate the distance vector $\mathbf{d}_i$ for the set of J distance functions $\{\rho_j\}$: for $j = 1, \ldots, J$: $\mathbf{d}_{ij} = \rho_j [S_j(\mathcal{D}), S_j(x_i)]$
    2.3.5 Reject particles with distances greater than the threshold: for $j = 1, \ldots, J$: if $\mathbf{d}_{ij} \geq \varepsilon_j$, return to 2.3.1 otherwise, accept $\boldsymbol{\theta}_i$ into $\{\boldsymbol{\theta}_i\}^{(s)}$
    2.3.6 Calculate the weight for particle $\boldsymbol{\theta}_i$:

$$w_i = \frac{\pi(\theta_i)}{\sum_{l=1}^{N}\left(w_l^{(s-1)}\prod_{k=1}^{K}\varphi\left[\tau_k^{-1}(\theta_{ik}^{(s)} - \theta_{lk}^{(s-1)})\right]\right)}$$ where $\varphi = \mathcal{N}(0,1)$, the standard normal distribution

  2.4 Normalize the weights

TABLE 2. Approximate Bayesian computation sequential Monte Carlo (ABC SMC) notation.

Symbol	Value	Definition
S	10	number of steps
s		indicates current step
N	10 000	number of particles at each step
i		indicates current particle
$\boldsymbol{\theta}$		particle (model parameter vector)
$\boldsymbol{\theta}^*$		intermediate particle
K	4	number of parameters
k		indicates current parameter
$\pi(\boldsymbol{\theta})$	Uniform	joint prior distribution of parameters
$\boldsymbol{\theta}_{ik}$		kth parameter value in the ith particle
$\{\boldsymbol{\theta}_i\}^{(s)}$		set of all particles at step s
$\{\boldsymbol{\theta}_i\}_k^{(s)}$		set of the kth parameter in all particles at step s
τ_k^2		variance of perturbation kernel for the kth parameter
$\mathcal{M}(\boldsymbol{\theta})$		population model
x		simulated data set
$\mathcal{D}$		observed data set
$S()$		summary statistic that may be used in a distance function
$\mathbf{d}$		vector of distances
$\rho(\cdot)$		distance function
J	4	number of distance functions
j		indicates current distance function
$\{\mathbf{d}_i\}^{(s)}$		set of all distance vectors at step s
$\{\mathbf{d}_i\}_j^{(s)}$		set of the jth distance in all distance vectors at step s
$\{\rho_j(\cdot)\}$		set of all distance functions
ε		threshold
q		fraction denoting the quantile
w		weight
$\{w_i\}^{(s)}$		set of all weights at step s
$\varphi(\cdot)$	$\mathcal{N}(0,1)$	standard normal distribution

to increase efficiency. We initiated a population with a cohort of 10 newly laid eggs. We used parameter values $\{1/\lambda_e = 4,\ \gamma_e = 6,\ 1/\lambda_i = 3,\ \gamma_i = 6,\ m = 0.6,\ r = 5\}$ to simulate 1000 data sets and chose one observed data set as representative of the group. The values in this observed data set were near the median of counts of individuals in each stage at each time to ensure that we were not evaluating the method on unlikely data. The number of particles (N) in each step was 10 000. Uninformative uniform priors were used for each parameter, with limits based on biological realism for a small, quickly developing, arthropod species.

In setting the threshold schedule q_s and the number of steps, s, we required a threshold schedule that declines steadily and maintains a large effective sample size (ESS), $\text{ESS}(\{w_i\}^{(s)}) = [\sum_{i=1}^{N}(w_i^{(s)})^2]^{-1}$. Under equal particle weights, ESS = N. ESS values decrease as particles differentially contribute to the next step. However, we also considered the distribution of distances that arises from pairs of data sets produced by the same underlying parameters. We investigated different threshold schedules, paying attention to the specific threshold values they produced compared to the variation in distances we would expect from the same underlying parameters, and the q_s schedule {0.9, 0.8, 0.7, 0.6, 0.6, 0.6, 0.5, 0.4, 0.4, 0.4} was chosen. The algorithm proceeded for 10 steps, after which there was no apparent change in the distribution of accepted particles and thresholds became so strict as to reject particles that have distances within the expected distribution of distances.

There are many aspects of ABC SMC performance that are beyond the scope of this assessment; we limit our assessment to illustrating the results of the model-fitting and the performance of distance functions. Although the true posterior distribution of our parameter values is unknown, we compare the posterior distribution to the underlying parameter values of our observed data set. We show the convergence of the intermediate distributions to the estimated posterior by comparing the distributions and by tracking the mean, median, and variance over all iterations.

RESULTS

Summary statistics and distance metrics

We varied the parameters in $\boldsymbol{\theta}$ one at a time and evaluated a plot of AUC and distance values over the range of $\Delta\boldsymbol{\theta}$ for each of 21 candidate distance functions. Our assessment procedure resulted in four distance metrics that met our criteria: (1) the sum of chi-square differences in counts of individuals in each stage at each observation, (2) the sum of squared differences in the number of adults at each observation, (3) the cross-correlation function (with zero time lag) between the observed and simulated number of immatures over all observations, and (4) the sum of earth mover's distances between the distribution of individuals between stages at each observation (Fig. 2).

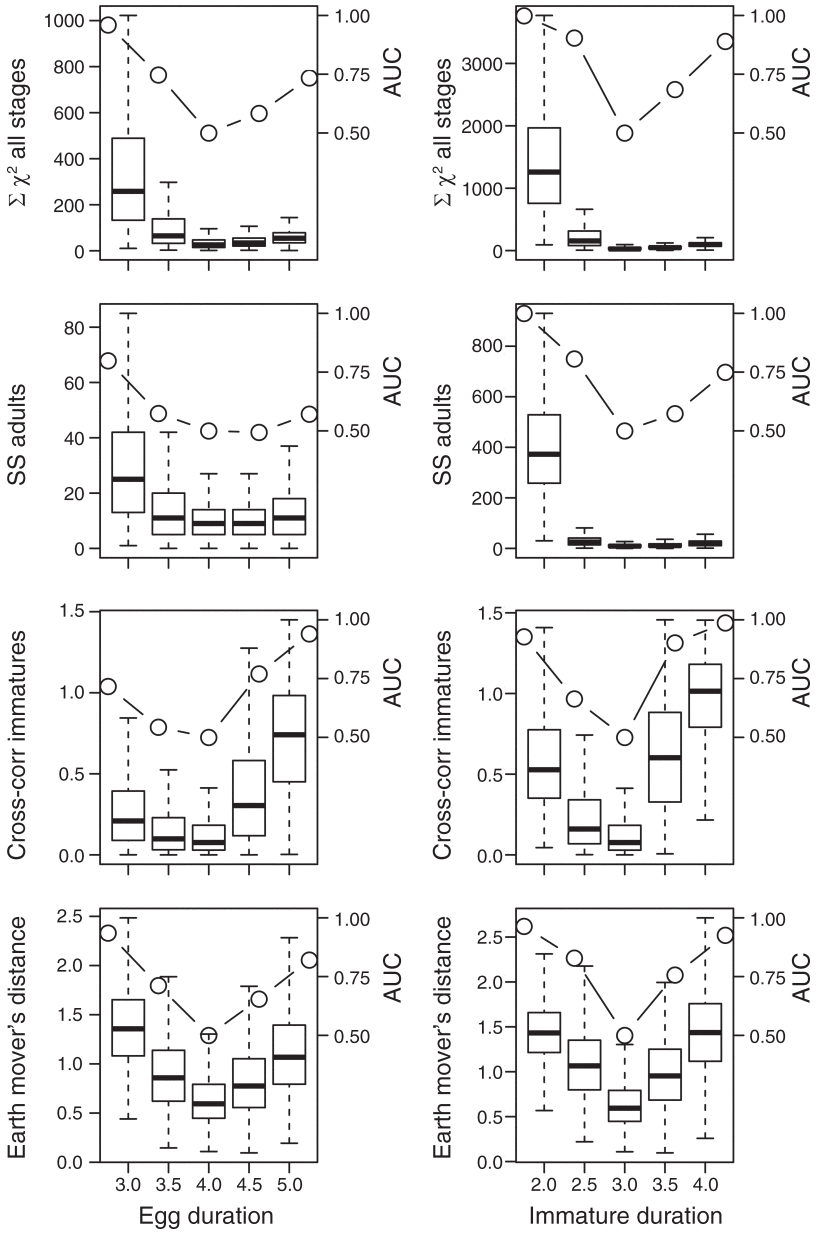

FIG. 2. Distance values and AUC statistics used to evaluate the ability of each distance function to track changes in each parameter. Each column of graphs tracks changes in one parameter while the other three are held constant, from left to right: the scale parameter of egg stage duration, the scale parameter of immature stage duration, mortality, and fecundity. Each row of plots represents a single distance function. The first row uses the sum of χ^2 differences in counts of individuals in each stage at each time. The second row uses the sum of squared differences in the number of adults at each time. The third row uses the cross-correlation function (with zero lag) between the counts of immatures at each time. The fourth row uses the earth mover's distance between the distributions at each time of the relative number of individuals in each stage. See Appendix A for further explanation of metrics. In each plot, the center values are the true parameters $\boldsymbol{\theta}$. Underlying parameters increase to the right and decrease to the left. Distance scale is given on the left axis and AUC scale is on right axis. Open circle show the AUC statistic of the ROC curve for each group of pairs of data sets. As expected, AUC for two groups with the same underlying parameters is the coin-flipping level of 0.5.

Parameter estimation

The ABC SMC algorithm yielded posterior distributions for each parameter (Fig. 3). The estimated posteriors were unimodal, and 95% credible intervals contained the true value of each parameter. The variance differed between parameters, as is normal for a model where some parameters are better informed by the data than others.

Algorithm convergence with the selected distance functions performed well. Over the 10 resampling steps, the intermediate distributions shifted from the prior to a

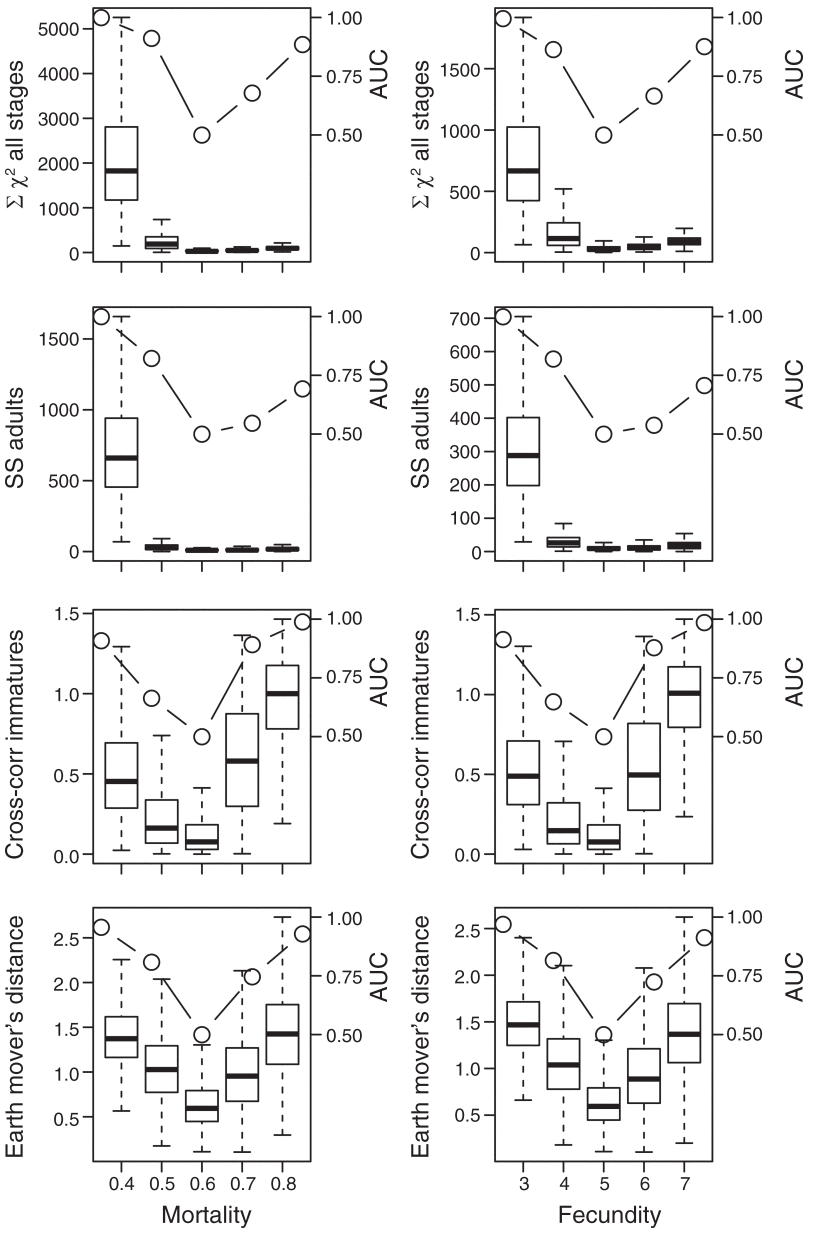

FIG. 2. Continued.

steady posterior (Appendix B). The mean and median of the parameter values shifted from the prior to steady values, with little change over the last five steps of the algorithm (Appendix B). The variances for egg and immature scale also stabilized over the SMC iterations, while those for mortality and fecundity have nearly stabilized but show small downward trends even after 10 steps. The egg scale and immature scale variances were also smaller than those for fecundity and mortality. ESS values decreased slightly from 10 000 to approximately 8500, but did so smoothly, indicating an acceptable threshold schedule.

DISCUSSION

Our study has shown how an ABC SMC algorithm can be applied to a difficult estimation problem in population ecology. Fitting the stochastic population model to stage-structured cohort data yielded posterior distributions centered on or near to true underlying parameter values. ABC provides an appealing alternative to likelihood-based inference in situations where the likelihood is unknown or too computationally expensive to approximate. The successful application of the ABC framework to our problem depended on the increased

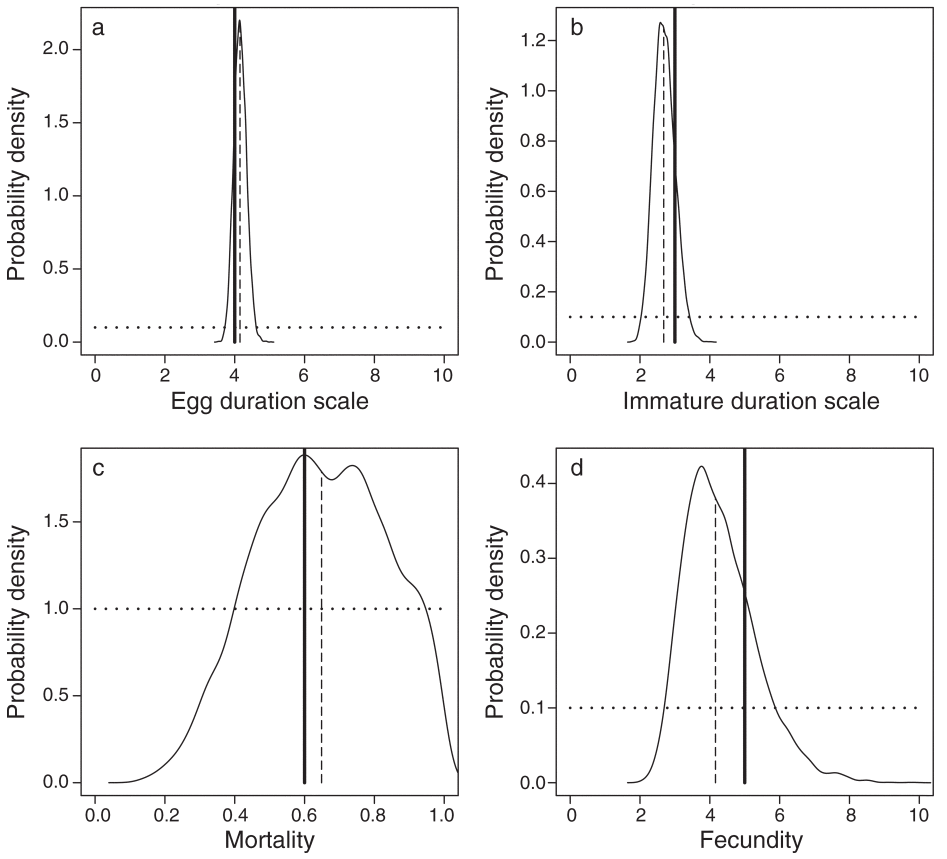

FIG. 3. Estimated posterior distributions (solid curves) for (a) the scale of egg stage duration, (b) the scale of immature stage duration, (c) mortality, and (d) fecundity. Heavy lines indicate the true parameter values that were used to simulate the observed data set. Dashed lines show the posterior median and dotted lines show uniform prior distributions.

efficiency of the SMC algorithm and on the careful assessment of summary statistics and distance metrics.

The specific SMC sampler used in this study is one of many algorithms that increases the efficiency of a simple rejection algorithm. These algorithms improve on the large inefficiencies in a rejection algorithm caused by the disparity between the target posterior and an uninformative prior. The SMC sampler, in particular, produces uncorrelated samples and does not become stuck in areas of low probability as an ABC MCMC chain might. The algorithm eliminates particles that do not represent the posterior in favor of those that do, causing the intermediate distributions to move more quickly from the prior to the posterior. The SMC algorithm also allows us to observe the intermediate distributions of parameter estimates, yielding information on the convergence.

Another advantage of an ABC SMC algorithm is the fact that it is independent of the specific simulation model. The algorithm can be applied to different problems, as long as a new simulation model, appropriate proposal distribution, and new distance functions are developed. The flexibility of the ABC SMC algorithm makes it likely that it will perform well across different systems with varying model assumptions. One limitation of any ABC algorithm is the computation cost involved simulating the large number of data sets needed. This cost cannot be avoided, but it can be minimized by efficient code and in the choice of programming language.

Any application of an ABC method will be challenged by the choice of summary statistics and distance metrics. The approach we have identified here avoids the intense computation time necessary for pilot ABC SMC runs with different subsets of summary statistics and distance metrics. Systematically investigating a large set of metrics and choosing one subset for use in the algorithm allows us to spend the computation time in the actual algorithm. We are also limited to estimating the posterior distribution of parameters given the choice of summary statistics and distance metrics. Our choice of algorithm may improve this estimated posterior, but it is the choice of strong and sufficient summary statistics and distance metrics that allows us to approach the true (unknown) posterior of the parameters.

Our assessment may be limited in its application to real data, as we were able to evaluate distance functions in the appropriate neighborhood of the true parameter

values. In practice, we may not be able to readily identify this range from the parameter space of the prior distributions. Identifying proper parameter ranges that would inform the choice of distance metrics and summary statistics will be problem specific, and may be more difficult for certain parameters or certain observed data sets. This limitation may be overcome by conducting simulation studies similar to the one presented here. Showing that certain distance metrics and summary statistics perform well for a particular class of data would lend support for the use of those distance functions in real world applications.

The ABC framework is a particularly good fit for ecologists, who commonly represent their systems with straightforward and intuitive stochastic models. Several software packages are available for estimation problems specific to population genetics (Cornuet et al. 2008) and systems biology (Liepe et al. 2010). More general software packages exist that implement different algorithms (Wegmann et al. 2010, Csilléry et al. 2012), including the sequential Monte Carlo sampler (Jabot et al. 2013). These general software packages still rely on the user to program the model, summary statistics, and distance metrics. The ABC framework can easily incorporate many extensions, such as correlations between parameters (Drovandi and Pettitt 2011), and model uncertainty (Ratmann et al. 2009). Model selection can also potentially be done in the ABC framework (Grelaud et al. 2009, Toni et al. 2009, François and Laval 2011), but see Robert et al. (2011) for the limitations to ABC model selection. Future applications of ABC methods will allow population ecologists to fit more realistic models to data, estimate parameters and make inferences about highly influential, variable processes.

Acknowledgments

We thank the NSF for funding, all of the members of the de Valpine lab for helpful discussions, and Nick Mills and Menelaos Stavrinides for their expertise in the arthropod system that motivated our questions.

Literature Cited

Barnes, C. P., D. Silk, and M. P. H. Stumpf. 2011. Bayesian design strategies for synthetic biology. Interface Focus 1:895–908.

Bazin, E., K. J. Dawson, and M. A. Beaumont. 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. Genetics 185: 587–602.

Beaumont, M. A. 2010. Approximate Bayesian computation in evolution and ecology. Pages 379–406 *in* D. J. Futuyma, H. B. Shafer, and D. Simberloff, editors. Annual review of ecology, evolution, and systematics. Volume 41. Annual Reviews, Palo Alto, California, USA.

Beaumont, M. A., J. Cornuet, J. Marin, and C. P. Robert. 2009. Adaptive approximate Bayesian computation. Biometrika 96:983–990.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Benton, T. G., A. P. Beckerman, and R. A. Desharnais. 2005. Population dynamics in a noisy world: lessons from a mite experimental system. Pages 143–181 *in* Population dynamics and laboratory ecology. Volume 37. Academic Press, Waltham, Massachusetts, USA.

Birt, A., R. M. Feldman, D. M. Cairns, R. N. Coulson, M. Tchakerian, W. Xi, and J. M. Guldin. 2009. Stage-structured matrix models for organisms with nongeometric development times. Ecology 90:57–68.

Blum, M. G. B., and O. François. 2008. Nonlinear regression models for approximate Bayesian computation. Statistics and Computing 20:63–73.

Cappé, O., S. Godsill, and E. Moulines. 2007. An overview of existing methods and recent advances in sequential Monte Carlo. Proceedings of the IEEE 95:899–924.

Cappé, O., A. Guillin, J. M. Marin, and C. P. Robert. 2004. Population Monte Carlo. Journal of Computational and Graphical Statistics 13:907–929.

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin, D. J. Balding, T. Guillemaud, and A. Estoup. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics 24:2713–2719.

Csilléry, K., M. G. B. Blum, O. E. Gaggiotti, and O. François. 2010. Approximate Bayesian computation (ABC) in practice. Trends in Ecology and Evolution 25:410–418.

Csilléry, K., O. François, and M. G. B. Blum. 2012. abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution 3:475–479.

de Valpine, P. 2009. Stochastic development in biologically structured population models. Ecology 90:2889–2901.

Drovandi, C. C., and A. N. Pettitt. 2011. Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. Biometrics 67:225–233.

Drovandi, C. C., A. N. Pettitt, and M. J. Faddy. 2011. Approximate Bayesian computation using indirect inference. Journal of the Royal Statistical Society: Series C (Applied Statistics) 60:317–337.

François, O., and G. Laval. 2011. Deviance information criteria for model selection in approximate Bayesian computation. Statistical Applications in Genetics and Molecular Biology 10:33.

Grelaud, A., C. P. Robert, J. Marin, F. Rodolphe, and J. Taly. 2009. ABC likelihood-free methods for model choice in Gibbs random fields. Bayesian Analysis 4:317–335.

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont, and L. Excoffier. 2005. Bayesian estimation of recent migration rates after a spatial expansion. Genetics 170:409–417.

Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth. 2011. Statistical inference for stochastic simulation models—theory and application. Ecology Letters 14:816–827.

Hastie, T. J., R. J. Tibshirani, and J. J. H. Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, New York, USA.

Hoeting, J. A., R. L. Tweedie, and C. S. Olver. 2003. Transform estimation of parameters for stage-frequency data. Journal of the American Statistical Association 98:503–514.

Jabot, F., and J. Chave. 2009. Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. Ecology Letters 12:239–248.

Jabot, F., T. Faure, and N. Dumoulin. 2013. EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. Methods in Ecology and Evolution 4:684–687.

Li, S., and M. Jakobsson. 2012. Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. BMC Genetics 13:22.

Liepe, J., C. Barnes, E. Cule, K. Erguler, P. Kirk, T. Toni, and M. P. H. Stumpf. 2010. ABC-SysBio—approximate Bayesian

computation in Python with GPU support. Bioinformatics 26:1797–1799.
Lopes, J. S., and S. Boessenkool. 2010. The use of approximate Bayesian computation in conservation genetics and its application in a case study on yellow-eyed penguins. Conservation Genetics 11:421–433.
Manly, B. F. J. 1990. Stage-structured populations: sampling, analysis, and simulation. Chapman and Hall, New York, New York, USA.
Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare. 2003. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences USA 100:15324–15328.
Padhukasahasram, B., J. D. Wall, P. Marjoram, and M. Nordborg. 2006. Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. Genetics 174:1517–1528.
Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16:1791–1798.
R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org
Rabosky, D. L. 2009. Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. Systematic Biology 58:629–640.
Rasmussen, R., and G. Hamilton. 2012. An approximate Bayesian computation approach for estimating parameters of complex environmental processes in a cellular automata. Environmental Modelling and Software 29:1–10.
Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson. 2009. Model criticism based on likelihood-free inference, with an application to protein network evolution. Proceedings of the National Academy of Sciences USA 106:10576–10581.
Read, K., and J. R. Ashford. 1968. A system of models for life cycle of a biological organism. Biometrika 55:211–221.
Robert, C. P., J.-M. Cornuet, J.-M. Marin, and N. S. Pillai. 2011. Lack of confidence in approximate Bayesian computation model choice. Proceedings of the National Academy of Sciences USA 108:15112–15117.
Shriner, D., Y. Liu, D. C. Nickle, and J. I. Mullins. 2006. Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution 60:1165–1176.
Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21:3940–3941.
Tavare, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times from DNA sequence data. Genetics 145:505–518.
Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface 6:187–202.
Turner, B. M., and T. Van Zandt. 2012. A tutorial on approximate Bayesian computation. Journal of Mathematical Psychology 56:69–85.
Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier. 2010. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. BMC Bioinformatics 11:116.

Supplemental Material

Appendix A

Summary statistics and distance metrics (*Ecological Archives* E095-122-A1).

Appendix B

Algorithm convergence (*Ecological Archives* E095-122-A2).

Supplement

C++ code and R script files for summary statistics, distance metrics, population simulation, and sequential Monte Carlo algorithm used for parameter estimation of a stochastic stage-structured population model (*Ecological Archives* E095-122-S1).