Extending Approximate Bayesian Computation with Supervised Machine Learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest

Francois-David Collin[1], Ghislain Durif[1], Louis Raynal[1], Eric Lombaert[1], Mathieu Gautier[2], Renaud Vitalis[1], Jean-Michel Marin[3], and Arnaud Estoup[4]

[1]Affiliation not available
[2]INRA
[3]Institut Montpelliérain Alexander Grothendieck
[4]INRAE

July 15, 2020

Abstract

Simulation-based methods such as Approximate Bayesian Computation (ABC) are well adapted to the analysis of complex scenarios of populations and species genetic history. In this context, supervised machine learning (SML) methods provide attractive statistical solutions to conduct efficient inferences about scenario choice and parameter estimation. The Random Forest methodology (RF) is a powerful ensemble of SML algorithms used for classification or regression problems. RF allows conducting inferences at a low computational cost, without preliminary selection of the relevant components of the ABC summary statistics, and bypassing the derivation of ABC tolerance levels. We have implemented a set of RF algorithms to process inferences using simulated datasets generated from an extended version of the population genetic simulator implemented in DIYABC v2.1.0. The resulting computer package, named DIYABC Random Forest v1.0, integrates two functionalities into a user-friendly interface: the simulation under custom evolutionary scenarios of different types of molecular data (microsatellites, DNA sequences or SNPs) and RF treatments including statistical tools to evaluate the power and accuracy of inferences. We illustrate the functionalities of DIYABC Random Forest v1.0 for both scenario choice and parameter estimation through the analysis of two example datasets corresponding to pool-sequencing and individual-sequencing SNP datasets. Because of the properties inherent to the implemented RF methods and the large feature vector (including various summary statistics and their linear combinations) available for SNP data, DIYABC Random Forest v1.0 can efficiently contribute to the analysis of large SNP datasets to make inferences about complex population genetic histories.

Hosted file

`Collin et al diyabcrf MER 10-07-2020 main text submitted.pdf` available at https://authorea.com/users/343084/articles/469772-extending-approximate-bayesian-computation-with-supervised-machine-learning-to-infer-demographic-history-from-genetic-polymorphisms-using-diyabc-random-forest