

Approximate Bayesian Computations Done Exactly: Towards a Thousand Human Genomes

Principal Investigator:

Dr. Raazesh Sainudiin,

Department of Mathematics and Statistics, University of Canterbury, NZ

Associate Investigators:

Professor Pierre Del Moral,

Head of INRIA Research Team ALEA, Université Bordeaux I, Talence, France

Dr. Amandine Véber,

Department of Mathematics and its Applications, École Polytechnique, Palaiseau, France

Dr. Kevin Thornton,

Department of Ecology and Evolutionary Biology, University of California, Irvine, USA

January 28, 2011

Abstract

Currently, 1000 whole human genomes are being sequenced. It is becoming exceedingly difficult to extract critical information from such extensive population-level genomic data reliably to solve pressing biomedical problems. Several approximate methods are being used without sound statistical justification to extract information from such complex data. This project aims to develop novel theoretically justified approximate methods for robust population-level genomic inference.

5A. ABSTRACT OF RESEARCH PROPOSAL

The 1000 Genomes project¹ is yielding massive amounts of data that document genetic diversity within human populations. Such population-level genomic data carry vital information about the genetic basis of disease, the history of the population and the evolutionary process itself. The sheer volume and complex nature of these data makes it exceedingly difficult to extract critical information to solve pressing biomedical problems. A new and fast *likelihood-free* inference algorithm, known as Approximate Bayesian Computation²⁻⁶, has recently gained popularity to bridge the ever-widening gap between massive data and the computational capability to analyse it. Such approximate estimators use simulations of sample datasets from the model to extract partial information about the parameters from various *summary statistics* of the data, without making explicit use of the likelihood function. These methods have been proven to perform well for simple additive models.⁷ But population genomic models such as the *coalescent*⁸ (a parametric family of Markov processes with a combinatorially dependent unobserved space of ancestral histories) are far more complex. Whether the summary statistics used by the algorithm for such genomic data are *approximately sufficient* (i.e. contain enough information about the parameters) for reliable inference is not known. Furthermore, the performance measures of these estimators for genomic data are assessed exclusively by simulations. Therefore, sound statistical justification of these likelihood-free approximate estimators for population genomic inference is urgently needed.

Recent analytic work by PI Sainudiin and AI Thornton⁹ has produced the first counterexamples that demonstrate the statistical inconsistency of several likelihood-free approximate estimators. These estimators rely on linear combinations of a popular summary statistic called the site frequency spectrum (SFS). To address these issues, the investigators exhaustively integrated the appropriate resolution of the unobserved space of ancestral histories, conditioned on the summary statistics, using the theory of lumped¹⁰, controlled¹¹ and coalescent⁸ Markov chains, and algebraic statistical Markov bases.¹² This novel approach is an improvement on existing approximate estimators, as it is proven to be approximately sufficient and significantly increases the power of classical genome scans. Currently, this exhaustive method works for samples of size 10 or less for simple models of ancestral histories. The goal of the proposed research is to advance this novel approximate estimator to (1) large sample sizes, (2) more realistic models (3) statistics that are more informative than the SFS and (4) real data analysis.

(1) We will forge *sequential importance samplers* developed by AI Del Moral¹³ (a class of evolutionary algorithms based on genealogical and interacting particle systems¹⁴) through a new refining family of lumped controlled coalescents⁹ in order to analyse large samples of size 1000. Our basic strategy is to find an increasingly informative family of observed statistics for which the likelihood depends on a corresponding family of unobserved statistics described by lumped Markov chains of the coalescent. This will allow us to propagate the information from the coarsest to the finest statistic via interacting particle systems in the parameter space, coupled with elements from the appropriate lumped resolution of the unobserved space. (2) We will extend our samplers to test hypotheses under AI Veber's more realistic model¹⁵ that incorporates a continuous spatial structure into the coalescent process. We will do this by generalising the underlying binary coalescent to Λ -coalescent¹⁶ (which allows for more than two lineages to coalesce) and then producing the lumped Markov chains needed for importance sampling over the sequence of SFS statistics. (3) We will also consistently extract more information in the data than that provided by the SFS and thereby develop an urgently needed theory of approximately sufficient samplers. (4) Finally, we will apply our methods to infer the demography and structure of two fruit fly species from high resolution whole-genome data from AI Thornton's laboratory and to the 1000 human genomes currently available at low resolution.

The expertise of the investigators in statistical genetics,⁹ evolutionary algorithms,^{13;14} coalescent processes¹⁵ and molecular population genomics¹⁷⁻¹⁹ is ideally suited to address these research goals and will strengthen approximate methods for population genomic inference.

5B. REFERENCES

References

- [1] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.
- [2] G Weiss and A von Haeseler. Inference of population history using a likelihood approach. *Genetics*, 149:1539–1546, 1998.
- [3] M Beaumont, W Zhang, and DJ Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.
- [4] SA Sisson, Y Fan, and MM Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104:1760–1765, 2007.
- [5] MA Beaumont, J-M Cornuet, J-M Marin, and CP Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.
- [6] G Bertorelle, A Benazzo, and S Mona. ABC as a exible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology*, 19:2609–2625, 2010.
- [7] RD Wilkinson. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. pages 1–13 (In Submission) <http://www.maths.nottingham.ac.uk/personal/pmzrdw/Papers/ABCisExact.pdf>.
- [8] JFC Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.
- [9] R Sainudiin, K Thornton, J Harlow, J Booth, M Stillman, R Yoshida, RC Griffiths, G McVean, and P Donnelly. Experiments with the site frequency spectrum. *Bulletin of Mathematical Biology: Special Issue in Algebraic Biology*, pages 1–48 (In Press). See Indian Statistical Institute Technical Report, isibang/ms/2010/8, September 16, 2010. http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/AlgBioISIBC_ms_2010_8.pdf, 2011.
- [10] CJ Burke and M Rosenblatt. A Markovian function of a Markov chain. *The Annals of Mathematical Statistics*, 29(4):pp. 1112–1122, 1958.
- [11] M Duflo. *Random Iterative Models*. Springer, Berlin, 1997.
- [12] P Diaconis and B Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26:363–397, 1998.
- [13] P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B in Statistical Methodology*, 68(3):411–436, 2006.
- [14] P Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications. Probability and its Applications*. Springer, New York, 2004.
- [15] N Barton, A Etheridge, and A Véber. A new model for evolution in a spatial continuum. *Electronic Journal of Probability*, 15:162–216, 2010.
- [16] J Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27:1870–1902, 1999.
- [17] R Sainudiin, A Clark, and R Durrett. Simple models of genomic variation in human SNP density. *BMC Genomics*, 8:146, 2007.
- [18] JD Jensen, KR Thornton, and P Andolfatto. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *drosophila*. *PLoS Genet*, 4(9):e1000198, 09 2008.
- [19] MK Burke, JP Dunham, P Shahrestani, KR Thornton, MR Rose, and AD Long. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*, 467:587–590, 2010.