Available online at www.sciencedirect.com

ScienceDirect

Mathematics and Computers in Simulation 103 (2014) 33–38

MATHEMATICS AND COMPUTERS IN SIMULATION

www.elsevier.com/locate/matcom

Original Article

Quasi-Monte Carlo method in population genetics parameter estimation

Hongmei Chi [a,*], Peter Beerli [b]

[a] *Department of Computer and Information Sciences, Florida A& M University, Tallahassee, FL 32307-5100, United States*
[b] *Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120, United States*

Received 21 June 2007; received in revised form 20 August 2013; accepted 24 February 2014
Available online 25 March 2014

Abstract

The computations of likelihood or posterior distribution of parameters of complex population genetics models are common tasks in computational biology. The numerical results of these approaches are often found by Monte Carlo simulations. Much of the recent work of Monte Carlo approaches to population genetics problems has used pseudorandom sequences. This paper explores alternatives to these standard pseudorandom numbers and considers the use of uniform random sequences, more specifically, uniformly distributed sequences (quasi-random numbers) to calculate the likelihood. We demonstrate by examples that quasi-Monte Carlo can be a viable alternative to the Monte Carlo methods in population genetics. Analysis of a simple two-population problem in this paper shows that parallel quasi-Monte Carlo methods achieve the same or better parameter estimates as standard Monte Carlo and have the potential to converge faster and so reduce the computational burden.

Keywords: Quasi-Monte Carlo; Phylogenetics; Population genetics; Coalescence theory; Completely uniformly distributed sequences

1. Introduction

Population genetic models often use parameters, such as migration rates, recombination rate and population sizes. Scientists use these models to understand the evolutionary history of humans (for example [28]), and pathogens such as the malaria parasites [13,22], HIV, SARS at the molecular level [1], among others. Population genetics [30] describes how the genes present in a population of parents are redistributed among a population of offspring. Mathematical population genetics [14] helps us to understand the epidemic history of a virus. Many biologists are pursuing research questions that explore the evolutionary significance of genotypic diversity. Early exploration of genetic diversity in the seventies used mainly gene frequencies to study the diversity among populations and species. Great progress was achieved in the estimating of parameters of population genetics models after the introduction of the coalescence theory in 1982 by Sir JFC Kingman [16]. The coalescent traces possible interactions of individuals, from today into the past, describing genealogical relationships among individuals. Several likelihood and Bayesian methods use coalescent theory to estimate population genetics parameters from genetic marker data from sampled individuals. These

* Corresponding author. Tel.: +1 850 412 7355.
E-mail addresses: hchi@cis.famu.edu, chi20082008@gmail.com (H. Chi), beerli@fsu.edu (P. Beerli).

http://dx.doi.org/10.1016/j.matcom.2014.02.005
0378-4754/

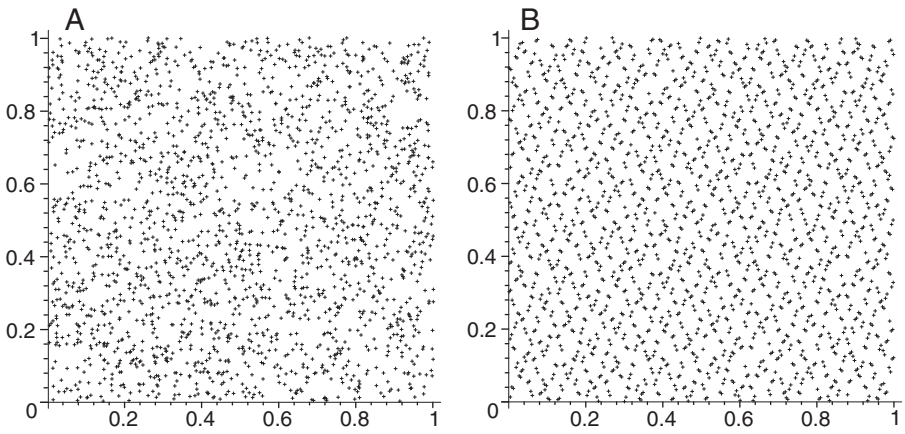

Fig. 1. Comparison of pseudorandom numbers and quasirandom numbers in two dimensions. (A) 2000 pseudorandom numbers (linear congruental generator); (B) 2000 quasirandom numbers (Soboĺ sequence).

methods integrate over all possible relationships (genealogies) of the sampled individuals. The number of genealogies for a typical data set is so large, however, that only approximations based on Markov chain Monte Carlo (MCMC) techniques are possible. Despite these computational problems, these new methods are more accurate and often more flexible than frequency-based methods for a wide range of population models.

The advances in population genetics are an impressive demonstration that approximate methods, such as MCMC, are very general and can help many researchers to estimate parameters of complex models taking into account all aspects of the data at hand. Biologists are still asking for better models to analyze data that comes from whole genomes rather than a single small region (locus) on a chromosome. In human genetics the amount of data has now surpassed the usability of the above mentioned MCMC approaches because the analysis of more than hundred thousands of loci takes more time than researchers are willing to wait. The large number of loci and the wish to investigate more complex models forces researchers to run MCMC approaches on parallel machines.

One major advantage of Monte Carlo methods is that they are usually very easy to parallelize. This is, in principal, also true of quasi-Monte Carlo (QMC) methods. Parallel computations using QMC require a source of quasirandom sequences, which are distributed among the individual parallel processes [10]. In these environments, a large QMC problem can be broken up into many small subproblems. These subproblems are then scheduled on the parallel, distributed, or grid-based environment. To our knowledge, all Monte Carlo approaches in population genetics and phylogenetics are based on pseudorandom number generators. Parallel computation using pseudorandom numbers has been embedded into MIGRATE [3] and as shown promising results [2]. Very recently, it has been recognized that the convergence rate of Monte Carlo approaches based on pseudorandom numbers is slow and that an important improvement of the convergence rate (and thus of speed) can be achieved by using quasi-Monte Carlo methods [23]. Many problems in population genetics can be viewed as high-dimensional numerical integration problems. Thus, QMC is applicable to problems in population genetics. We will explore the application of QMC to population genetics problems by using quasirandom numbers. This quasi-MCMC approach is novel in population genetics or phylogenetics inference. A similar approach has been used in an application to online retailing [12].

The paper will explore benefits of quasi-random numbers for complex population genetics models. The interaction with genealogy searches is of special interest because such complexity has not yet been explored using QMC.

2. Quasi-Monte Carlo methods

Monte Carlo methods are based on the simulation of stochastic processes whose expected values are equal to computationally interesting quantities. Despite the universality of Monte Carlo methods, a serious drawback is their slow convergence, which is based on the $O(N^{-1/2})$ behavior of the size of statistical sampling errors. Quasirandom numbers are constructed to be as evenly distributed as is mathematically possible. The difference between pseudorandom and quasirandom numbers can be easily seen in Fig. 1. Pseudorandom numbers tend to be clumpy while quasirandom numbers [5] are more uniformly distributed.