Maximum likelihood inference comparing sequence and allelic markers in a population of variable size: an Importance Sampling approach
Champak Reddy, Francois Rousset, Pierre Pudlo, Raphaël Leblois

To cite this version:

HAL Id: hal-02932322
https://hal.inrae.fr/hal-02932322
Submitted on 7 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Maximum likelihood inference comparing sequence and allelic markers in a population of variable size: an Importance Sampling approach

Champak Beeravolu Reddy1, François Rousset2, Pierre Pudlo1,3, Raphael Leblois1

1 CBGP, INRA, Campus International de Baillarguet CS 30016 34988 Montferrier-sur-Lez cedex
2 ISM, CNRS, Université de Montpellier II - CC 065 34095 MONTPELLIER Cedex 05
3 ISM, Université de Montpellier II - CC 051 34095 MONTPELLIER Cedex 05

Introduction & Objectives

Building on the Griffiths & Tavaré (1994a,b) approach, Stephens & Donnelly (2000) introduced an efficient importance sampling scheme for computing the likelihood of a genetic sample. Their method consists of approximating the conditional probability of the allelic type of an additional gene given those currently in the sample. This technique was further extended by De Iorio & Griffiths (2004a,b) to a subdivided population framework for various mutation models between gene types.

For sequence loci, the results of De Iorio & Griffiths (DIG) represent an improvement in terms of efficiency over those of Bahlo & Griffiths (2000) but this hasn’t been tested on more than a single panmictic population. DIG consider the DNA sequences to be evolving under an infinitely-many Sites Mutation model (SMM). For allelic loci (microsatellites) evolving under a Step-wise Mutation model (SMW), we use the Importance Sampling proposal distributions of De Iorio et al. (2005).

Here, we used results from Griffiths & Tavaré (1994c) in order to implement the case of a single exponentially contracting population (see box: the demographic model) and contrast the performance of estimation from sequence and allelic markers. This approach gave us an insight into the potential complementarity of the considered genetic markers for improving the precision of the demographic and mutational estimates.

Software & Simulations

We have developed software which simulate data in the well known Campus (for allelic data) and the Nexus (for sequence data) formats. These simulations were generated using IBDSim (http://raphaelleblois.free.fr/software) which follows an exact coalescent algorithm. This data was then analysed by the Migraine software (http://kimura.univ-lyon1.fr/~rousset/Migraine.htm) for inferring the demographic and mutational parameters using DIG’s method. Note that the current version of Migraine can also handle more complex demographic models than the one presented here such as Isolation by Distance (IBD) based on DIG’s method.

We choose a simulation strategy which consisted of varying the time since the contraction began (D = 1/T) keeping the other parameters constant. A total of 10 demographic scenarios were thus produced, each of which was repeated 200 times. The initial size of the population was set to 200,000 individuals which progressively diminishes down to 200 individuals. Each sample consists of 100 individuals genotyped at 10 loci.

Maximum likelihood inference comparing sequence and allelic markers in a population of variable size: an Importance Sampling approach

The demographic model:

\[
\begin{aligned}
N_{anc} & = \frac{\theta_{anc}}{2Nm}, \\
D & = \frac{T}{2N}, \\
\theta & = 2Na.
\end{aligned}
\]

<table>
<thead>
<tr>
<th>Past</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>(\theta_{anc} = 2Na)</td>
</tr>
<tr>
<td>(D)</td>
<td>(\frac{T}{2N})</td>
</tr>
</tbody>
</table>

Overall estimation performance

ISM

SMM

Example of Migraine’s output (ISM, D = 1.25)

<table>
<thead>
<tr>
<th>D = 0.025</th>
<th>D = 0.125</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(LRT-pvalue \sim U[0;1]) &lt; 0.01)</td>
<td>(P(LRT-pvalue \sim U[0;1]) &lt; 0.05)</td>
</tr>
<tr>
<td>(P(LRT-pvalue \sim U[0;1]) &lt; 0.05)</td>
<td>(P(LRT-pvalue \sim U[0;1]) &lt; 0.01)</td>
</tr>
</tbody>
</table>

Rel. bias, rel. MSE

1.14, 6.44

Differences between markers

We accounted for the differences in the evolution of the genetic markers as follows: for microsatellites under a SMM, we assumed a mutation rate of \(10^{-5}\) while for DNA sequences under an ISM we choose \(3\times10^{-7}\)/sequence. The simulated data had the following number of haplotypes/alleles (N) and expected heterozygosity (\(H_e\)) for the cases presented here:

ISM

D = 1.25: \(N_e = 2.6, H_e = 0.39\)

D = 0.025: \(N_e = 2.4, H_e = 0.65\)

D = 1.25: \(N_e = 4.4, H_e = 0.58\)

D = 0.025: \(N_e = 13.4, H_e = 0.87\)

An extreme scenario (ISM, D = 0.025)

Very recent and extreme contractions result in flat profile likelihoods (right) or saddle point like situations (left). Here, the surfaces indicate that the markers have little information on the bottleneck and thus the current population size. This results in very wide confidence intervals for the parameters \(D\) and \(N\). Finally large \(N\) values are confounded with the true values of \(N_{anc}\) resulting also in wide confidence intervals.

Results & Conclusions

The bottleneck scenario implemented here:

1. was successfully detected over a wide range of scenarios;\n2. was inferred with reasonable precision for intermediate scenarios. The SMM estimates performed better than their ISM counterparts but this can be understood in terms of heterozygosity and the number of alleles/haplotypes present in the simulated data (see box: Differences between markers);\n3. the time taken in comparison to typical MCMC approaches was very short (~4s for a single repetition);\n4. slightly longer runs improve estimation performance (not shown here).

As the LR statistic approximately follows a chi-square distribution, we calculated the corresponding p-values. The figures on the left plot the empirical cdf of the obtained p-values. This helps us check whether the distribution of the p-values for each parameter is uniform.

References:


Affiliations & Sponsors:

CBR & RL are supported by the BfMD/STORCH project while CBR, FB, PP & RL are supported by the EMILE project, both funded by the French National Research Agency.

Additional References:

Affiliations & Sponsors:

CBR & RL are supported by the BfMD/STORCH project while CBR, FB, PP & RL are supported by the EMILE project, both funded by the French National Research Agency.

Additional References:

Affiliations & Sponsors:

CBR & RL are supported by the BfMD/STORCH project while CBR, FB, PP & RL are supported by the EMILE project, both funded by the French National Research Agency.

Additional References:

Affiliations & Sponsors:

CBR & RL are supported by the BfMD/STORCH project while CBR, FB, PP & RL are supported by the EMILE project, both funded by the French National Research Agency.

Additional References: