Basics of modern population genetics and inference of model parameters

Peter Beerli
Department of Scientific Computing
Florida State University, Tallahassee
Basics of modern population genetics and inference of model parameters

Peter Beerli
Department of Scientific Computing
Florida State University, Tallahassee
Questions of interest

- How many humpback whales lived in the North-Atlantic 100,000 years ago?
- What is the original host of the SARS virus?
- What are the seasonal dynamics of influenza?
- Are Florida panthers different from Texan cougars?
- How and when did human populate this world?
Species
Population models

Species

Population

Mutation

introduces variability
Population models

Species
Population
Migration
Mutation
introduces variability
Population models

Population

Mutation

Migration

Species

Loss of variability
[genetic drift]
Population models

\[ \Theta_i = a N_e \mu = a \times \text{the effective population size} \times \text{the mutation rate} \]

\[ M = \frac{m}{\mu} = \text{immigration rate over the mutation rate for all migration routes} \]

\[ a \text{ is a multiplier dependent on the ploidy and transmission mode} \]
Population models

\[ \Theta_i = aN_e \mu = a \times \text{the effective population size} \times \text{the mutation rate} \]

\[ M_{ij} = \frac{m_{ij}}{\mu} = \text{immigration rate from population } i \text{ to } j \text{ over the mutation rate} \]

\[ aN_e^{(j)} m_{ij} = \Theta_j M_{ij} \]
Population size $= f(\text{Alleles, Mutation, Migration, population size in last generation})$

$$N_t = f(X, \mu, m, N_{t-1})$$

Simply looking only at a single population this is

$$N_t = f(X, \mu, N_{t-1})$$
Population models
Population models
Population models

past present

©2010 Peter Beerli
Population models

Fisher, Wright
Population models

past

present
Population models

past present

©2010 Peter Beerli
Population models

past

present

©2010 Peter Beerli
Population models

past present
Population models

past present
Most commonly used and often wrong?

F-statistic

Simulated data using a model with two populations

<table>
<thead>
<tr>
<th>True values</th>
<th>Expected values</th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.0</td>
<td>1.14±0.77</td>
</tr>
<tr>
<td>0.01</td>
<td>5.5</td>
<td>7.80±22.20</td>
</tr>
<tr>
<td>0.05</td>
<td>5.5</td>
<td>11.46±18.54</td>
</tr>
</tbody>
</table>

Most commonly used and often wrong?

Simulated data using a model with two populations

<table>
<thead>
<tr>
<th>True values</th>
<th>Expected values</th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.0</td>
<td>1.14±0.77</td>
</tr>
<tr>
<td>0.01</td>
<td>5.5</td>
<td>7.80±22.20</td>
</tr>
<tr>
<td>0.05</td>
<td>5.5</td>
<td>11.46±18.54</td>
</tr>
</tbody>
</table>

Population models
Population models

Genealogy
Population models

Genealogy

past present

©2010 Peter Beerli
Population models

Genealogy

past

present

©2010 Peter Beerli
Coalescence theory

Present

Past

Kingman
The distribution of time intervals $u_k$ follows an exponential distribution with

$$E(u_k) = \frac{\Theta}{k(k-1)}$$

The probability distribution is given by

$$p(G \mid \Theta, n) = \prod_{k=2}^{n} \exp\left(-u_k \frac{k(k-1)}{\Theta} \right) \frac{2}{\Theta}$$
Population size and Time to coalescence

Large populations need more time to coalesce to a single lineage than small populations.
All genealogies were simulated with the same population size $N_e = 10,000$
Variability of the coalescent process

MRCA = most recent common ancestor (last node in the genealogy)
Coalescence applied

For a Bayesian inference we want to calculate the probability of the model parameters given the data $p(\text{model}|X)$.

**Coalescent** to describe the population genetic processes.

**Mutation model** to describe the change of genetic material over time.

We calculate the Posterior distribution $p(\Theta|X)$ using Bayes’ rule

$$p(\Theta|X) = \frac{p(\Theta)p(X|\Theta)}{p(X)}$$

where $p(X|\Theta)$ is the likelihood of the parameters.
Likelihood calculation

\[ p(X|\Theta) = \int_G p(G|\Theta)p(X|G)dG \]

- \( p(G|\Theta) \) is the probability of a genealogy given parameters.
- \( p(X|G) \) is the probability of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.
<table>
<thead>
<tr>
<th>Tips</th>
<th>Labeled histories</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>180</td>
</tr>
<tr>
<td>6</td>
<td>2700</td>
</tr>
<tr>
<td>7</td>
<td>56700</td>
</tr>
<tr>
<td>8</td>
<td>1587600</td>
</tr>
<tr>
<td>9</td>
<td>57153600</td>
</tr>
<tr>
<td>10</td>
<td>2571912000</td>
</tr>
<tr>
<td>15</td>
<td>695805766896240000</td>
</tr>
<tr>
<td>20</td>
<td>5644809895887305913369600000000</td>
</tr>
<tr>
<td>30</td>
<td>4368466613103069512464680019862076389144064000000000000000</td>
</tr>
<tr>
<td>40</td>
<td>302733382994800735654630336455145720004293943205386250170788872192000000000000</td>
</tr>
<tr>
<td>50</td>
<td>$3.28632 \times 10^{112}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.37416 \times 10^{284}$</td>
</tr>
</tbody>
</table>

\[
p(X|\Theta) = \int_G p(G|\Theta)p(X|G)dG
\]

The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.
**Irreducibility**: the Markov chain must be able to reach all interesting parts of the distribution.

**Recurrence**: all interesting parts must be reached (in principle) infinitely often if the chain is run infinitely long.

**Convergence**: the sample mean must converge to the expectation.

In reality only *Irreducibility* is crucial and if all states can be reached often (recurrence) then the convergence criteria follows.
Gene flow among populations
Bayesian inference

Bayesian inference

Bayesian inference

Bayesian inference

Bayesian inference

Bayesian inference

Bayesian inference


©2010 Peter Beerli
Obvious migration pattern

Beerli, P. (2009) How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use?
The model used to do the inference specifies the condition: a model that estimates only a single parameter therefore assumes that the parameter is constant over time. Many different models are possible.

The more parameters, the longer the MCMC needs to run to approximate the likelihood with some accuracy.

In Bayesian context the prior distributions for the parameters need considerable attention.

Reports of parameter estimates without discussion of run conditions, and support intervals are not all that useful.
Researchers from the frequency-based camp claim that the coalescence-based methods are working on an evolutionary time-scale and therefore are not really usable in a conservation genetics or management context.

There is some truth to this claim because the time scale for the genealogies is in generations and with large populations such genealogies are deep, but ...
Average of parameters over long time

- True value
- MIGRATE estimate
- Support interval
- Harmonic mean

Coalescent-based methods

---

©2010 Peter Beerli
Basic method “summary”

- Probabilistic modeling using the coalescent allows the separation of mutation model and population model, this allows the estimation of parameters of fairly complex models. Typically more complex than frequency-based methods.

- Parameter estimates are long-term averages but because more information is available near the present, estimates are weighted towards more recent values when the true parameters vary over time.
Bedford et al. (submitted) Global migrational dynamics underlie evolution and persistence of human influenza A (H3N2).
Beerli and Palczewski (in prep.) Sequence partitioning reveals recombination hotspots.
Thanks to the organization committee for the invitation to speak to you.

Research is supported by the joint NSF/NIGMS Mathematical Biology program with NIH grant R01 GM 078985.

download MIGRATE-N from http://popgen.sc.fsu.edu
Thanks to the organization committee for the invitation to speak to you.

Research is supported by the joint NSF/NIGMS Mathematical Biology program with NIH grant R01 GM 078985.

download MIGRATE-N from http://popgen.sc.fsu.edu

That was it folks: questions, comments?