Estimating Mutation Parameters and Population History from Temporally-Spaced Genome Data

Arman Bilge
abil933@aucklanduni.ac.nz

Computational Evolution Group
The University of Auckland

18 February 2016

Motivation

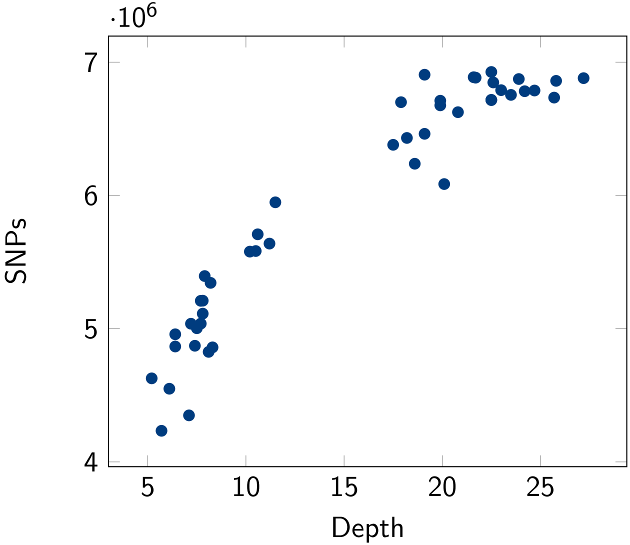
Data

- Really cool (but secret!) dataset
- Feasible to sequence entire genomes
- More recently, even possible to recover ancient genomes
- Temporally-spaced genome data
- Opportunity to do inference previously only possible for viruses
    - mutation rate
    - population size and changes through time

Challenges

- Difficult to phase diploid genomes
- Recombination
- Sequencing error, espescially in ancient genomes
- Cannot use Bayesian phylogenetic methods

Sequencing Error



The Composite Likelihood

Definition (Cox and Reid, 2004)

Consider an n-dimensional vector random variable X with density

$$X \sim f\left(x \mid \boldsymbol{\theta}\right)$$

Suppose that the full n-dimensional density is intractable, but

$$f_i\left(x_i \mid \boldsymbol{\theta}\right) \text{ and } f_{ij}\left(x_i, x_j \mid \boldsymbol{\theta}\right), i \neq j$$

are not. Letting

$$\ell_1(\boldsymbol{\theta}; x) = \sum_i \log f_i\left(x_i \mid \boldsymbol{\theta}\right)$$
$$\ell_2(\boldsymbol{\theta}; x) = \sum_{i \neq j} \log f_{ij}\left(x_i, x_j \mid \boldsymbol{\theta}\right) - an\ell_1\left(\boldsymbol{\theta}; x\right)$$

then ℓ_2 is the **pairwise composite loglikelihood**.

The Composite Likelihood for Serially-Sampled Genomes

Assumptions

- Every site in the alignment has an independent phylogeny
  - No more concerns about recombination and phasing!
- Phylogenies are i.i.d. under the same coalescent process
- Substitution process is identical across sites
- Sites are i.i.d.
- A site is our n-dimensional vector X

The Composite Likelihood for Serially-Sampled Genomes

- $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{N_e}, \boldsymbol{\epsilon})$, substitution parameters, effective population sizes, sequence error probabilities
- Computing $f\left(x \mid \boldsymbol{\theta}\right)$ requires integrating out all phylogenies
- Instead, approximate with a pairwise composite likelihood
- $\ell_1(\boldsymbol{\theta}; x)$ ignored because a single sequence not dependent on $\boldsymbol{\theta}$
- For a pair of genomes, integrate out their coalescence time

$$f_{ij}\left(x_i, x_j \mid \boldsymbol{\theta}\right) = \int_0^\infty p\left(x_i, x_j \mid t, \boldsymbol{\mu}, \boldsymbol{\epsilon}\right) p\left(t \mid \mathbf{N_e}\right) \mathrm{d}t$$

Why use composite likelihoods?

- Frequently used in genetics
- Well-studied in general so plenty of existing theory
- Make the intractable tractable

Theorem (Xu and Reid, 2011)

Under some assumptions, the maximum composite likelihood is consistent under the full model.

Probability for a Pair of Nucleotides

- Closed-form (but fairly complex) expression for f_{ij} under HKY and skyline coalescent
- Under Jukes-Cantor with a constant population size, we have

$$
\begin{aligned}
p\left(i,j \mid \mu, N_e\right) &= \int_0^\infty \frac{1}{4} p_{ij}\left(\left(\tau + 2t\right)\mu\right) p\left(t \mid N_e\right) \mathrm{d}t \\
&= \begin{cases} \frac{1}{4}\left(\frac{1}{4} + \frac{9\exp\left(-\frac{4}{3}\mu\tau\right)}{64\mu N_e + 12}\right) & \text{if } i = j \\ \frac{1}{4}\left(\frac{1}{4} - \frac{3\exp\left(-\frac{4}{3}\mu\tau\right)}{64\mu N_e + 12}\right) & \text{if } i \neq j \end{cases}
\end{aligned}
$$

- Sequencing error considered by integrating over uncertainties

The Complete Composite Loglikelihood

- Just a big summation over all sites and all pairs

$$\mathcal{L}\left(\boldsymbol{\theta};\mathcal{A}\right)=\sum_{k=1}^{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log\left(f_{ij}\left(\mathcal{A}_{ik},\mathcal{A}_{jk}\mid\boldsymbol{\theta}\right)\right)$$

- To consider multiple categories of sites, substitution parameters $\boldsymbol{\mu}$ for each category are independent but $\mathbf{N_e}$ and ϵ are linked

An Efficient Implementation

Data Structure

- Preprocessing is $O\left(mn^2\right)$, where $m =$ number of sites
- For every pair of genomes, count every substitution $i \to j$
- Calculation per category is $O\left(n^2\right)$
- Preprocessing/calculation are parallelised

Maximum Likelihood Optimisation

- L-BFGS-B algorithm (quasi-Newton method)
- Uses automatically-calculated gradient
- Single tuning parameter affects number of iterations necessary before convergence

Assessing the Estimates

Confidence Intervals

- Distribution is much narrower than it should be
- Using non-parametric bootstrapping across sites
- Better to use block-bootstrap b/c of dependence across sites

Date-Randomisation Test

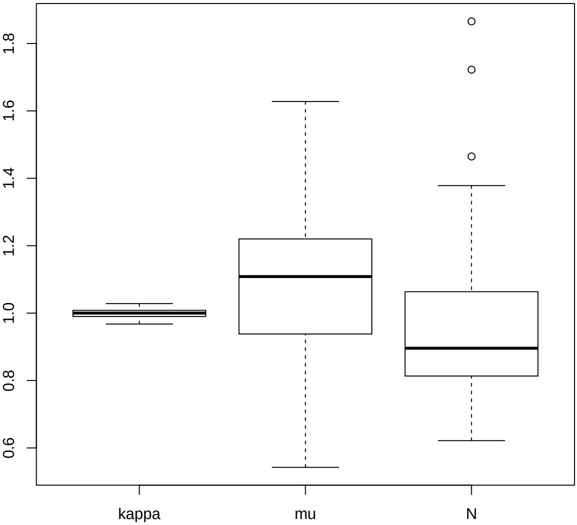
- Is there signal for the mutation rate in our data?
- Randomly reassign times to taxa and reestimate rate
- Verify that estimates fall outside of true confidence interval
- Even stronger: verify no overlap of intervals

Irreproducible Simulation Results

Set Up

- 46 diploid taxa, with 22 ancient individuals as old as 50k years
- Mutation rate is $\mu = 10^{-9}$
- HKY model with $kappa = 5$
- Constant-size population with $N_e = 3 \times 10^6$
- 10^6 independent phylogenies for 10^6 sites
- Attempt to estimate parameters from simulated data

Irreproducible Simulation Results



Concluding Remarks

Summary

- An efficient method for inference from serially-sampled genomes
- Applying to an exciting dataset
- Looks promising but needs more validation

Future Work

- Fix-up implementation and reign in numerical instability
- Successfully estimate all parameters
- Prepare some convincing simulation studies
- Consider structured population in model
- Find faster alternatives to bootstrapping
- Can we show consistency theoretically?

Acknowledgements

Thanks to

- Tanja Stadler and Alexei Drummond
- Vladimir Minin
- Matthew Parks, Sankar Subramanian, and David Lambert
- New Zealand eScience Infrastructure

References

RH Byrd et al. (1995). *SIAM J Sci Comput* 16.5. `doi:10.1137/0916069`

DR Cox, N Reid (2004). *Biometrika* 91.3. `doi:10.1093/biomet/91.3.729`

S Duchêne et al. (2015). *Mol Biol Evol* 32.7. `doi:10.1093/molbev/msv056`

F Larribe, P Fearnhead (2011). *Statistica Sinica* 21.12.

X Xu, N Reid (2011). *J Statist Plann Inf*. `doi:10.1016/j.jspi.2011.03.026`