Efficient Parameter Estimation for Human Microsatellite Mutation

Glenn Galvizo, under Dr. Floyd Reed

University of Hawaii at Manoa

December 5, 2018

Overview

1. Introduction
   - Problem Statement
2. Microsatellites
   - DNA Variation: Tandem Repeats
   - Microsatellite Data
   - Mutation Model
3. Methodology
   - Coalescent Simulation
   - ABC MCMC
4. Results / Discussion
   - Likelihood Distribution
   - Future Work
5. Conclusion

Brief overview of modern human history:

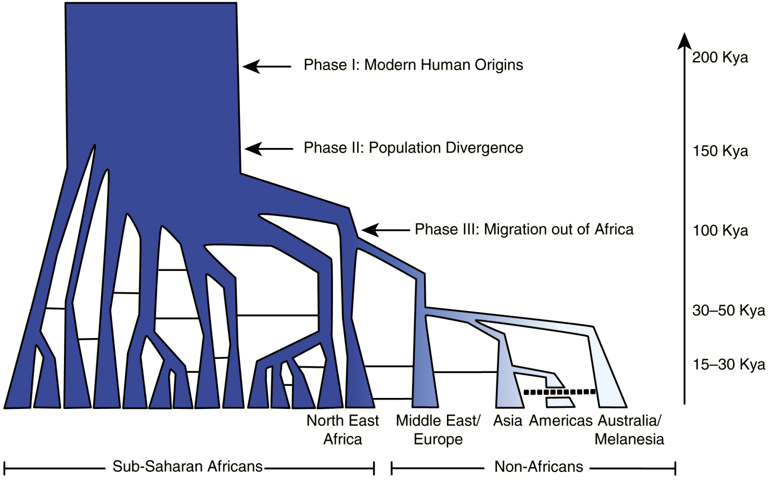


**Image from Campbell & Tishkoff [1].*

What is the goal of this research?

Research Question

Which microsatellite mutation model parameters are the most likely to produce our observed data?

Essential Questions

1. What is a microsatellite?
2. What is the observed data?
3. How do microsatellites mutate? What is the model?
4. How do we simulate evolution?
5. How can we find the best parameters?

What is a microsatellite?

Definition (Microsatellite)

A *microsatellite* is a short sequence in DNA, repeated in tandem.

- Interested in number of repeats.
- Represent variation in humans.
- Infer human history by tracking changes.

. . .AACG**ATATATATATAT**GGCTA . . .

. . .AACG**ATATATATAT**GGCTA . . .

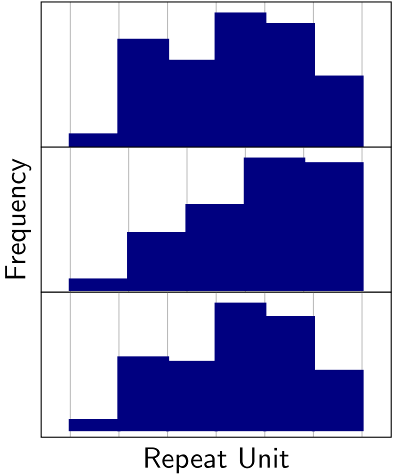
. . .AACG**ATATATAT**GGCTA . . .

. . .AACG**ATATAT**GGCTA . . .

. . .AACG**ATAT**GGCTA . . .

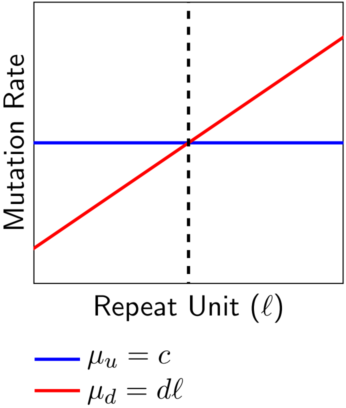
What data are we working with?

- Working with Columbian GATA samples.
- Samples collected from ALFRED (ALlele FREquency Database).
- Interested in frequency of repeat length.



How do microsatellites mutate?

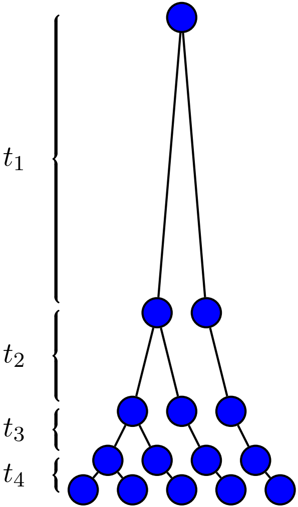
- *Single Step*: Mutate up one, down one, or not at all [5].
- *Proportional:* Mutation rate dependent on length [2]
- *Focal Bias*: Mutate toward some length [3].
- μ_u = upward mutation rate
  μ_d = downward mutation rate.



How do we simulate evolution?

Answer: We construct a evolutionary tree (coalescent)!

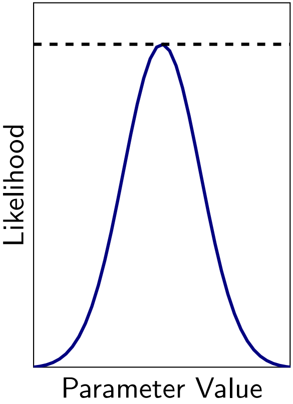
1. Given sample size n, mutation parameters c, d.
2. Construct random tree with n leaves and common ancestor.
3. Mutate children from an ancestor length until leaves are reached.



How can we find the best parameters?

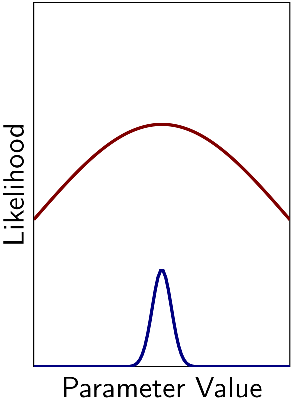
- *Problem*: Which model parameters are the most likely to generate our observed data?
    1. How do we compute this likelihood?
    2. How can we maximize this likelihood?

- *Solution*: ABC – MCMC
  (Approximate Bayesian Computation – Markov Chain Monte Carlo)



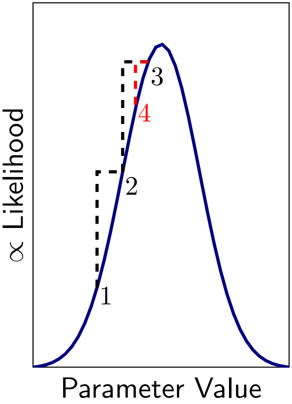
How do we compute likelihood?

- *Naive Approach*: Count number of exact matches.
- *Problem*: Frequency of exact matches is low.
- *Solution*: Count approximate matches instead!
    1. Compute distance between generated and observed samples.
    2. Count number of generated samples where distance is below some threshold.
    3. Results in wider and flatter distribution (red vs. blue) [4].

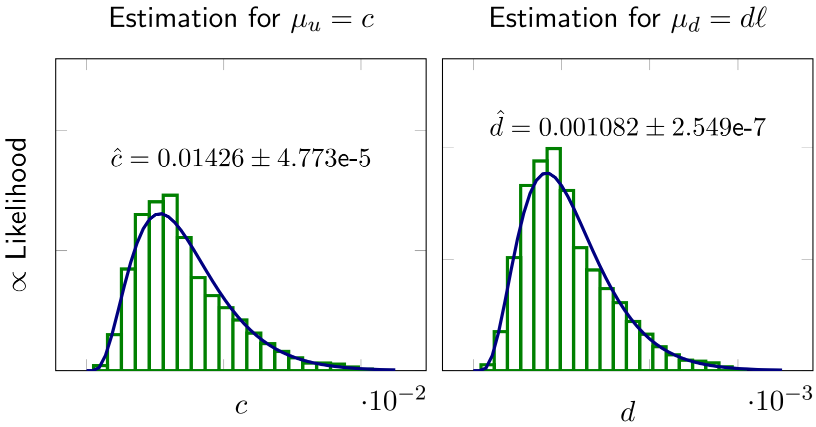


How do we maximize likelihood?

- *Problem*: Cannot iterate through all possible likelihoods.
- *Solution*: Use MCMC!
  - Randomly samples from $\propto$ likelihood distribution [4]
  - Spends longer time in regions of high likelihood.
  - Fit frequency to curve, maximize this curve.



What are our results?



Estimation for $\mu_u = c$

Estimation for $\mu_d = d\ell$

**Preliminary results given above.*

What do we do with this?

Mutation Model:

- Use more samples from different locations.
- Verify and test our parameters against different data.
- Run more and longer MCMCs.

Demographic Models:

- Estimate time, admixture, population size of Africa split.
- Integrate Neanderthal, Denisovan populations.
- Answer, "Who did we come from?"

Conclusion

- Microsatellite $=$ a short sequence in DNA repeated in tandem.
- Microsatellites mutate $\pm 1, 0$ repeat lengths, toward focal bias.
- Likely parameters (c, d) were found with ABC-MCMC.
- Future work $=$ more samples & MCMC, different demographics models.

Acknowledgments

Amazing People:

- Dr. Floyd Reed
- Reed Lab
- Undergraduate Showcase
- UHM Mathematical Biology Committee
- The Audience

References & Questions :-)

Michael C. Campbell and Sarah A. Tishkoff.
The Evolution of Human Genetic and Phenotypic Variation in Africa.
Current Biology, 20(4):R166–R173, February 2010.

H. Ellegren.
Heterogeneous mutation processes in human microsatellite DNA sequences.
Nature Genetics, 24(4):400–402, April 2000.

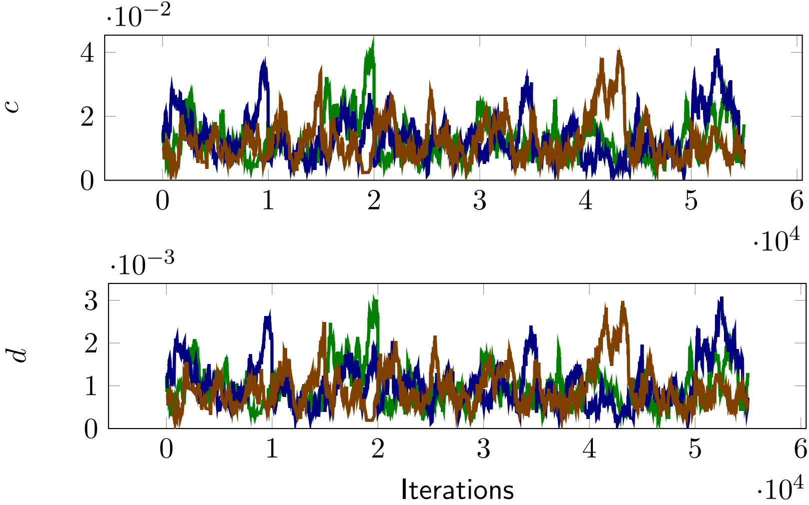
J. C. Garza, M. Slatkin, and N. B. Freimer.
Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size.
Molecular Biology and Evolution, 12(4):594–603, July 1995.

Jarno Lintusaari, Michael U. Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander.
Fundamentals and Recent Developments in Approximate Bayesian Computation.
Systematic Biology, 66(1):e66–e82, January 2017.

Tomoko Ohta and Motoo Kimura.
A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*.
Genetics Research, 89(5-6):367–370, December 2007.

Extra Slides

Assessing MCMC Convergence (Trace Plots)


$\cdot 10^{-2}$
c
$\cdot 10^{4}$
$\cdot 10^{-3}$
d
Iterations
$\cdot 10^{4}$