Computational Methods in Population Genetics

Dirk Metzler

January 25, 2024

Contents

1 Introduction **2**
1.1 Example . . . 2
1.2 How to study . . . 3
1.3 Wright-Fisher model and Kingman's Coalescent . . . 4
1.4 Estimators for θ and Tajima's D . . . 10
1.5 Outline of methods . . . 12
1.5.1 ML with Importance Sampling . . . 12
1.5.2 MCMC for frequentists and Bayesians . . . 13
1.5.3 Approximate Bayesian Computation (ABC) . . . 13

2 Importance sampling for genealogies **14**
2.1 Griffiths und Tavaré . . . 16

3 MCMC with Lamarc (or Migrate) **20**

4 IM, IMa, IMa2 **30**

5 Approximate Bayesian Computation (ABC) **33**
5.1 ABC with local regression correction . . . 35
5.2 MCMC without likelihoods . . . 38
5.3 Sequential / Adaptive ABC . . . 39
5.4 Optimizing sets of summary statistics with PLS . . . 40

6 Jaatha and dadi **43**
6.1 Wild Tomatoes and Jaatha 1.0 . . . 43
6.2 Jaatha 2.0 . . . 51
6.3 Application to genome-wide data . . . 54
6.4 Statistics in jaatha beyond jsfs . . . 54
6.5 Conclusions . . . 55

7 ARG read from left to right and its approximations **55**
7.1 Why ARG read from left to right is not Markovian . . . 55
7.2 SMC, SMC': Markovian approximations of the ARG . . . 56
7.3 MaCS, SCRM: closer approximations of the ARG . . . 56

8 Hidden-Markov model approaches for coalescents with recombination **58**
8.1 The Pairwise Sequential Markov Chain (PSMC) Approach . . . 58
8.2 Excursus: hidden Markov models (HMMs) . . . 60
8.3 Back to PSMC . . . 64
8.4 MSMC and MSMC2 . . . 64

9 The program STRUCTURE **65**
9.1 no admixture, no sampling locations . . . 66
9.2 with admixture . . . 69
9.3 taking sampling locations into account . . . 71
9.4 Faster alternatives to STRUCTURE for large datasets . . . 72
9.4.1 ADMIXTURE . . . 72
9.4.2 fastSTRUCTURE . . . 73
9.4.3 snmf() in the Bioconductor R package LEA . . . 73

10 Simulating and detecting selection **76**
10.1 Ancestral Selection Graphs . . . 77
10.2 Simulating selective sweeps and other kinds of strong selection . . . 78

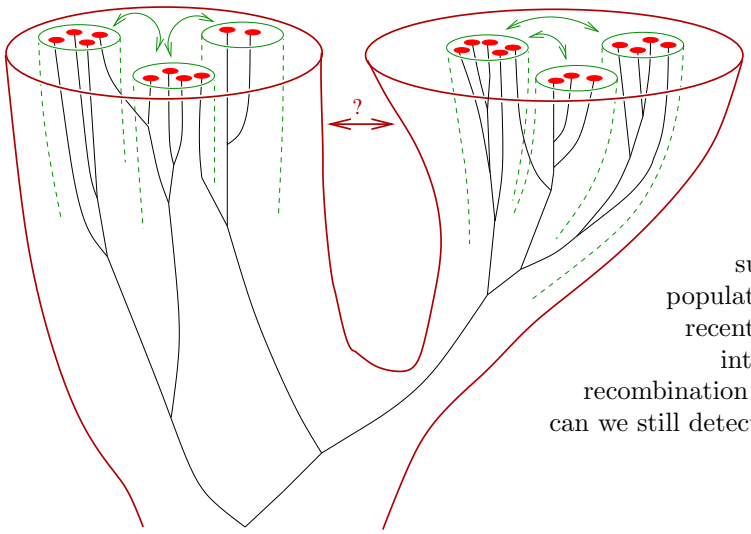
11 Stats for Selection **80**
11.1 Detecting selective sweeps . . . 80
11.1.1 Stats for detecting incomplete sweeps . . . 83
11.1.2 Soft Sweeps . . . 85
11.1.3 Signals of sweeps and demography . . . 85
11.2 Does the gene list make sense? . . . 87
11.3 Balancing selection . . . 88

12 Phasing and PAC **89**
12.1 Li&Stephens' PAC approach . . . 89
12.1.1 Estimating LD and recombination hotspots . . . 89
12.1.2 Population splitting and recombination . . . 94
12.1.3 Diversifying selection and recombination . . . 95
12.1.4 Excursus: Stephens and Donnelly's Importance Sampling . . . 97
12.2 Phasing . . . 101
12.2.1 Excoffier and Slatkin's EM algorithm . . . 102
12.2.2 Excursus: EM algorithm . . . 102
12.2.3 Basic algorithms in PHASE . . . 105
12.2.4 PAC in PHASE . . . 107
12.3 Phasing large genomic datasets . . . 108
12.3.1 fastPHASE . . . 108
12.3.2 Phasing with Beagle software package . . . 110
12.3.3 IMPUTE version 2 . . . 110
12.3.4 MaCH . . . 111
12.3.5 polyHAP . . . 111

1 Introduction

1.1 Example

Complex Demography



substructure
population growth
recent speciation
introgression?
recombination within loci
can we still detect selection?

1.2 How to study

References

[1] R. Nielsen, M. Slatkin (2013) *An Introduction to Population Genetics: Theory and Applications.* Sinauer

[2] J.H. Gillespie (2004) *Population Genetics – A Concise Guide.* 2nd Ed., John Hopkins University Press

[3] J. Felsenstein (2016+) *Theoretical Evolutionary Genetics.* https://felsenst.github.io/pgbook/pgbook.html

[4] J. Wakeley (2008) *Coalescent Theory: An Introduction.* Roberts & Company Publishers

[5] J. Hein, M.H. Schierup, C. Wiuf (2005) *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory.* Oxford University Press

[6] J. Felsenstein (2002) *Inferring Phylogenies*, Chapters 26, 27 and 28. Sinauer Associates

ECTS and work load per week

For *Computational Methods in Evolutionary Biology*, 9 ECTS ≈ 0.6 per week, 18 hours per week:

- 4 hours lecture (each 45 min + break)
- 3 hours exercise sessions
- 6 hours homework (exercises)
- 5 study lecture contents

For *Comp. Meth. Pop. Gen.*, 6 ECTS = 0.75 per week, 22.5 hours per week,

- all as above plus
- 2 hours of practicals and additional exercise session
- 1 hours learn software, apply to data, prepare presentation
- 1.5 more hours to learn algorithms and maths

How to study the content of the lecture

For the case that you are overwhelmed by the contents of this course, and if you don't have a good strategy to study, here is my recommendation:

1. Try to explain the items under "Some of the things you should be able to explain"
2. Discuss these explanations with your fellow students
3. Do this before the next lecture, such that you can ask questions if things don't become clear
4. Do the exercises (at least some of them) in time
5. Study all the rest from the handout, your notes during the lecture, and in books

What will the exam be like

You can bring:

- pocket calculator
- formula sheet, hand-written by yourself

What you need to answer the questions:

- understanding concepts
- be able to apply concepts
- do calculations
- think during the exam
- (not just reproduce facts)

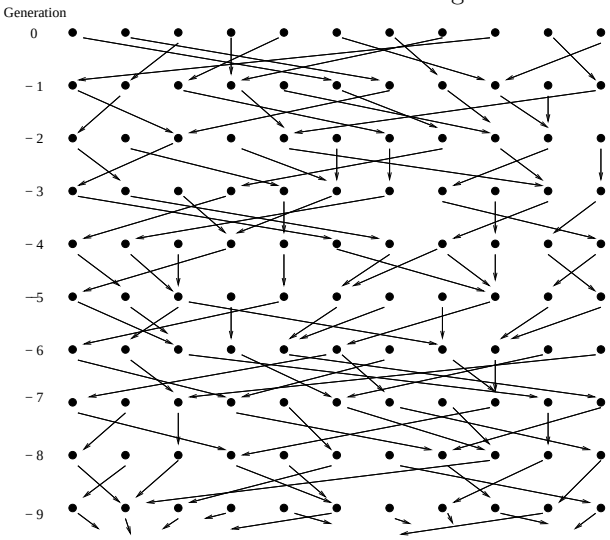
1.3 Wright-Fisher model and Kingman's Coalescent

Basic assumptions of the Wright Fisher model

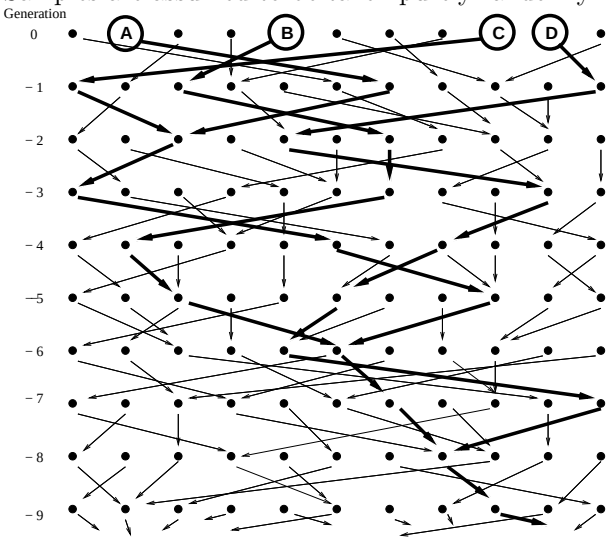
- non-overlapping generations
- constant population size
- panmictic
- neutral (i.e. no selection)
- no recombination
- N diploid individuals $\rightsquigarrow$ population of $2N$ haploid alleles (in case of autosomal DNA)

Wright Fisher model

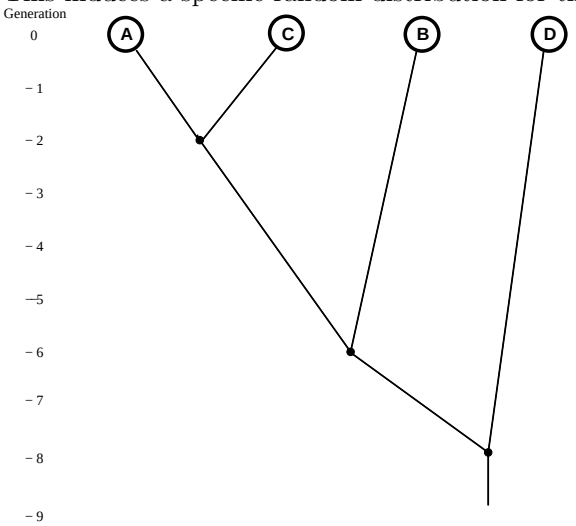
Each allele chooses an ancestor in the generation before.



Samples are assumed to be taken purely randomly from the population.



This induces a specific random distribution for the genealogies of the sampled alleles.



Haploid population of size N_e

Average time until two ancestral lineages coalesce: N_e generations.
Scale time: (1 time unit) = (N_e generations) $\Rightarrow$ pairwise coalescence rate = 1
$\mu :=$ mutation rate per generation

$$\theta := 2N_e \cdot \mu$$

is the expected number of mutations between 2 random individuals
Let $N_e \longrightarrow \infty$

Time until two lineages coalesce

W: time in generations until two lineages coalesce, $X = W/N_e$.

$$\begin{aligned} \Pr(W > k) &= \left(1 - \frac{1}{N_e}\right)^k \approx e^{-k/N_e} \\ \Pr(X > t) &= \Pr(W > t \cdot N_e) \approx e^{-t} \end{aligned}$$

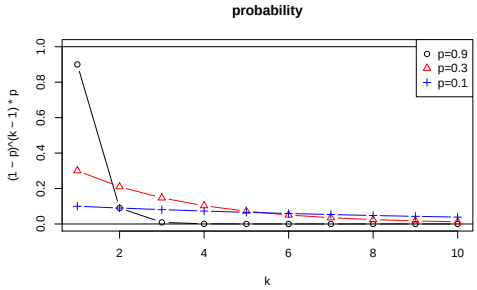
For $N_e \to \infty$ the "$\approx$" becomes "=".Thus

W is geometrically distributed with success probability $\frac{1}{N_e}$.

X is (for $N_e \to \infty$) exponentially distributed with rate 1.

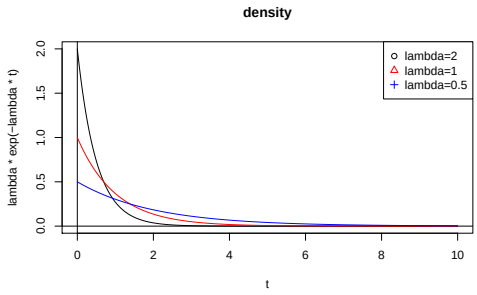
Geometric distribution

$$\begin{aligned} \Pr(W > k) &= (1-p)^k \\ \Pr(W = k) &= (1-p)^{k-1} \cdot p \\ \mathbb{E}W &= 1/p \end{aligned}$$



Exponential distribution

$$\begin{aligned} \Pr(X > t) &= e^{-\lambda t} \\ f(t) &= \lambda \cdot e^{-\lambda t} \\ \mathbb{E}X &= 1/\lambda \end{aligned}$$



If there are k lineages traced back, each of the

$$\binom{k}{2} = \frac{k \cdot (k-1)}{2}$$

pairs of lineages coalesces with rate 1 (per N_e generations).

Going to the limit $N_e \to \infty$ we can neglect that more than two lineages coalesce in the same generation.

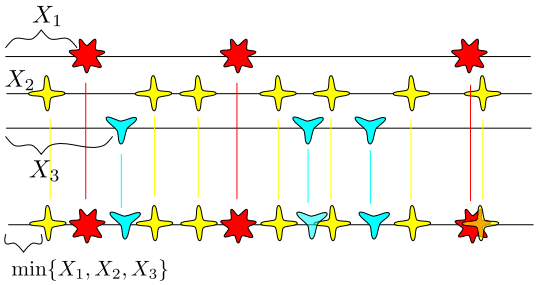
Thus, the total rate of coalescence is $\binom{k}{2}$ and the expectation value for the time back until the first coalescence is

$$\frac{1}{\binom{k}{2}} = \frac{2}{k \cdot (k-1)}$$

time units (of N_e generations).

Minimum of independent exponentials

If $X_1, X_2, \ldots, X_n$ are exponential random variables with rates $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the minimum of the X_i is exponentially distributed with rate $\Lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. (E.g. each λ_i is the rate of certain events, and X_i is the waiting time for the first of these event.)



This implies that for k lineages (and $N_e \to \infty$) the waiting time (in units of N_e generations) is geometrically distributed with rate $\binom{k}{2}$.

(Doob-)Gillespie algorithm

To simulate the process, first simulate the waiting time for the first event according to an exponential distribution with rate Λ.

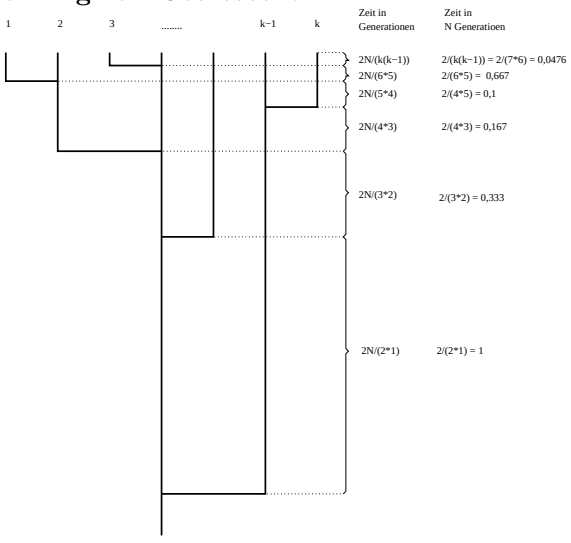
Then make a random decision which type of event occurred, using the probabilities

$$\frac{\lambda_1}{\Lambda}, \frac{\lambda_2}{\Lambda}, \ldots, \frac{\lambda_n}{\Lambda}.$$

Proceed with the next and all other events accordingly.

This is often called **Gillespie's algorithm**. Dan Gillespie applied it in the 1970ies to simulate chemical reactions. The algorithm was, however, known long before, e.g. by Doob in the 1940ies.

The Kingman Coalescent



$$\begin{aligned}
&\mathbb{E}(\text{total length}) \\
&= k \cdot \frac{2}{k \cdot (k-1)} + \cdots + 2 \cdot \frac{2}{2 \cdot 1} \\
&= \sum_{i=2}^{k} i \cdot \frac{2}{i \cdot (i-1)} = \sum_{i=2}^{k} \frac{2}{i-1} \\
&= 2 \cdot \sum_{h=1}^{k-1} 1/h
\end{aligned}$$

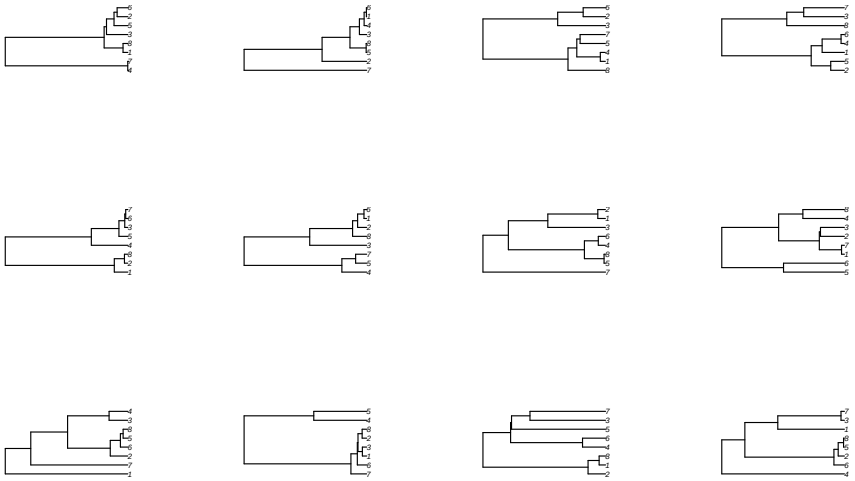
Simulating Kingman's coalescent for k lineages

Apply Doob-Gillespie algorithm.

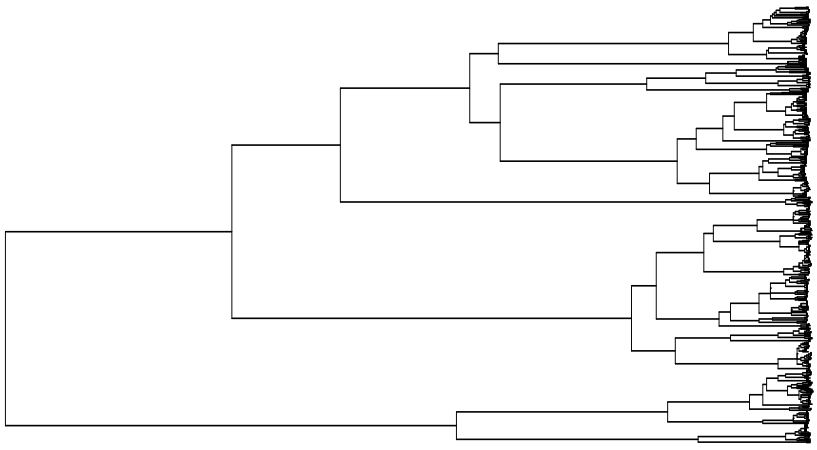
$\binom{k}{2}$ pairs of lineages, each pair coalescing with rate 1.

- Draw the time τ until the first two lineages coalesce from exponential distribution with rate $\binom{k}{2}$.
- choose one of the $\binom{k}{2}$ pairs of lineages (each with the same probability of $1/\binom{k}{2}$) and coalesce them to one lineage
- Draw next time interval from exponential with rate $\binom{k-1}{2}$ and let two of the $k-1$ coalesce, and so on, until the last two lineages coalesce.

typical coalescent trees for $n = 8$:



simulated coalescent tree with $n = 500$:



Time to the most recent ancestor (TMRCA)

$$
\begin{aligned}
\mathbb{E}(\text{TMRCA}) &= \frac{1}{\binom{k}{2}} + \frac{1}{\binom{k-1}{2}} + \cdots + \frac{1}{\binom{3}{2}} + \frac{1}{\binom{2}{2}} \\
&= \frac{2}{k\cdot(k-1)} + \frac{2}{(k-1)\cdot(k-2)} + \cdots + \frac{2}{3\cdot 2} + \frac{2}{2\cdot 1} \\
&= 2\cdot\left(\frac{1}{k\cdot(k-1)} + \frac{1}{(k-1)\cdot(k-2)} + \cdots + \frac{1}{3\cdot 2} + \frac{1}{2}\right) \\
&= 2\cdot\left(\frac{k-(k-1)}{k\cdot(k-1)} + \frac{(k-1)-(k-2)}{(k-1)\cdot(k-2)} + \cdots + \frac{3-2}{3\cdot 2} + \frac{2-1}{2}\right) \\
&= 2\cdot\left(-\frac{1}{k} + \frac{1}{k-1} - \frac{1}{k-1} + \frac{1}{k-2} + \cdots - \frac{1}{3} + \frac{1}{2} - \frac{1}{2} + \frac{1}{1}\right) \\
&= 2\cdot\left(1 - \frac{1}{k}\right)
\end{aligned}
$$

In the diploid case

If N_e is the effective population size of individuals, and the locus under consideration is **diploid**, time is scaled in units of $2N_e$ generations or, in some programs, in $4N_e$ generations. We then define

$$\theta = 4 \cdot N_e \cdot \mu.$$

The expected waiting time until the uncestral lineages of two gene copies coalesce is then $2\cdot N_e$ generations. The expectation values

$$\mathbb{E}(\text{total length}) = 2\cdot\sum_{i=1}^{k-1} 1/i \qquad \text{and} \qquad \mathbb{E}(\text{TMRCA}) = 2\cdot\left(1-\frac{1}{k}\right)$$

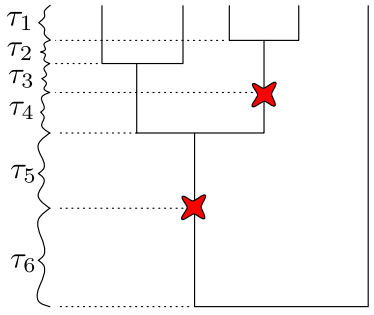
hold in the diploid case in units of $2N_e$ generations.

Simulate coalescent with mutations

k lineages $\Rightarrow$

- each of $\binom{k}{2} = \frac{k\cdot(k-1)}{2}$ pairs of lineages coalesce with rate 1
- each of k lineages mutate with rate $\theta/2$

$\Rightarrow$ time τ_i to next event is exponentially distrib. with rate $\frac{k\cdot(k-1)}{2} + k\cdot\frac{\theta}{2}$



Then, the probabilities are proportional to the rates, that is:

Next event is mutation with probability

$$\frac{k\cdot\frac{\theta}{2}}{\frac{k\cdot(k-1)}{2} + k\cdot\frac{\theta}{2}} = \frac{\theta}{k-1+\theta}$$

and coalescence with probability

$$\frac{k-1}{k-1+\theta}.$$

Some things that you should be able to explain

- Wright-Fisher model for population genetics
- Coalescent model; rate of coalescence of k lineages

- Geometrically distributed coalescent times, approximated by exponential
- Rates of independent processes add up
- the typical genealogy of a sample from W-F population
  - long internal branches
  - very short tip branches for large sample size
  - how to calculate expectation values for
    - total of all branch lengths
    - TMRCA: time to the most recent ancestor
  - how can we let population size go to ∞
- Why time is scaled in const.$\cdot N_e$ generations
- $\theta = 4N_e\mu$ or $\theta = 2N_e\mu$ or...

1.4 Estimators for θ and Tajima's D

Two estimators of θ

θ_π ("Tajima's π" or "nucleotide diversity") Average number of pairwise differences.

θ_W ("Watterson's θ") $= \frac{\text{number of mutations}}{\sum_{i=1}^{k-1} 1/i}$

Both are unbiased estimators of θ, i.e. $\mathbb{E}\theta_W = \mathbb{E}\theta_\pi = \theta$.

Note the difference

parameter:
- hypothetical value in theoretical model;
- example: $\theta = 4N_e\mu$
- non-random (in classical frequentistic stats)
- assumed usually in stats: there is a true value that is unknown

statistic:
- a function of the data (that is, is caluclated from the data)
- examples: θ_W and θ_π,
- are random variables because data is also random due to
  - randomly sampling from natural variation
  - random process
  - measurement error

estimator:
- statistic to estimate the value of a parameter
- example: θ_W and θ_π are estimators for θ

Example: Ward et al. (1991) sampled 360 bp sequences from mtDNA control region of $n = 63$ Nuu Chah Nulth and observed 26 mutations.

$$\theta_W = \frac{26}{\sum_{i=1}^{62} 1/i} = 5.5123$$

This corresponds to 0.0153 Mutations per base and per $2 \cdot N_e$ generations.Assuming a mutation rate $\widehat{\mu} \approx 6.6 \cdot 10^{-6}$ per generation per site this leads to an effective population size of

$$\widehat{N_e} = \frac{\theta_W/360}{2 \cdot \widehat{\mu}} \approx 1150 \text{ females}$$

How precise is this estimation?

$$\mathsf{var}(\theta_W) = \frac{\theta}{\sum_{i=1}^{n} 1/i} + \theta^2 \cdot \frac{\sum_{i=1}^{n} 1/i^2}{\left(\sum_{i=1}^{n} 1/i\right)^2}$$

Theorem 1 *Any unbiased estimator of θ has variance at least*

$$\frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{k+\theta}}.$$

(Here, we assume that the estimation is based on a single locus without recombination).

For the Nuu Chah Nulth data we get:

$$\theta_W = 5.5123$$

$$\sigma_{\theta_W} = 3.42$$

Confidence range? (2σ-rule would leed to negative values...)

Conclusion: N_e could perhaps also be 200 or 3000 females.

How can we improve this estimate? Sample more individuals? How many individuals n would we need to get $\sigma_{\theta_W} = 0.1 \cdot \theta$? From the formula for $\mathsf{var}\theta_W$ follows that we need $n \approx 2 \cdot e^{100/\theta}$. For $\theta = 5$, this is $n \approx 10^9$. For $\theta = 1$, this is $n \approx 10^{43}$. number of water molecules on earth$\approx 10^{47}$ number of seconds since big bang$\approx 4.3 \cdot 10^{17}$

Solution: sample many loci!

References

[Fel06] J. Felsenstein (2006) Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, Or More Loci? *Mol. Biol. Evol.*, **23.3**: 691–700.

How to sample if

- one read is 600 bp long
- costs for developing a new locus is 40$
- costs for collecting a sample is 10 or 0.10$
- costs for a single read is 6$
- you can spend 1000$
- true θ is 1.8 (per locus)

Optimal sampling scheme: $n = 7$ or $n = 8$, respectively, individuals and 11 loci.

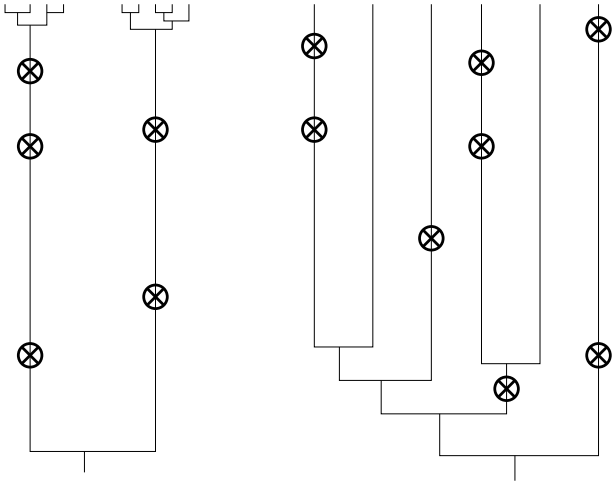
With this sampling scheme we get:

$$\sigma_{\theta_W} \approx 0.2 \cdot \theta \text{ and } \sigma_{\theta_\pi} \approx 0.22 \cdot \theta$$

(all this is based on infinte-sites assumptions)

Tajima's D

$\theta_\pi > \theta_W$: $\theta_\pi < \theta_W$:



$D := \frac{\theta_\pi - \theta_W}{\widehat{\sigma}_{\theta_\pi - \theta_W}}$

substructure?
population
growth?
selection?

Some of the things you should be able to explain

- θ, θ_W, θ_π, Tajima's D
- difference between parameters and statistics and which of the different θs is which
- How do population growth, substructure and selection shape the genealogy of a random sample and how is this reflected in Tajima's D?
- Why does a single locus (or several completely linked loci) not allow for reasonable estimates?
- why are more loci usually better than more or longer sequences?
- Why do we need to know μ to estimate N_e from data and vice versa?

1.5 Outline of methods

1.5.1 ML with Importance Sampling

The Likelihood

$\psi = (\psi_i)_i$ vector of model parameters

D sequence data

$$L_D(\psi) = \Pr_\psi(D) = \sum_{\text{all Genealogies } G} \Pr_\psi(D \mid G) \cdot \Pr_\psi(G).$$

When taking branch lengths into account:

$$L_D(\psi) = \Pr_\psi(D) = \int_{\text{all Genealogies } G} \Pr_\psi(D \mid G) \cdot P_\psi(dG).$$

Importance Sampling

Draw $G_1, \ldots, G_k$ (approx.) i.i.d. with density Q and approximate

$$\int \Pr_\psi(D \mid G)\ P_\psi(dG) \approx \frac{1}{k} \sum_{i=1}^{k} \frac{\Pr_\psi(D \mid G_i) \cdot P_\psi(G_i)}{Q(G_i)}.$$

efficient for ψ with

$$Q(G_i) \approx \text{const.} \cdot \Pr_\psi(D \mid G_i) \cdot P_\psi(G_i)$$

Methods differ in their choice of Q.

Griffiths & Tavaré (1994)

Q: Generate G backwards in time, greedy proportional to coalescence and mutation probabilities. Choose between all allowed events.

Good for infinite sites models, inefficient if back-mutations are allowed.

1.5.2 MCMC for frequentists and Bayesians

Felsenstein, Kuhner, Yamato, Beerli,...

For some initial ψ_0, sample Genealogies G approx. i.i.d. according to $\Pr_{\psi_0}(G \mid D)$ by Metropolis-Hastings MCMC.

Coalescent is a natural prior for G!

Two flavours:

for frequentists: use $G_1, \ldots, G_k$ for Importance Sampling

Optimize approx. Likelihood $\to \psi_1$

Iterate with ψ_0 replaced by ψ_1

for Bayesians: Then sample ψ conditioned on Genealogies and iterate to do Gibbs-sampling from $\Pr(\psi, G \mid D)$.

Problems of full-data methods

- usual runtime for one dataset: several weeks or months
- complex software, development takes years
- most programs not flexible, hard to write extensions

1.5.3 Approximate Bayesian Computation (ABC)

Pritchard et al. (1999)

Approximate Bayesian Computation

1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ
3. repeat until k accepted ψ':
   - Simulate ψ' from prior distribution of ψ
   - Simulate genealogy G according to $\Pr_{\psi'}(G)$.
   - Simulate data and compute values s' of S
   - accept ψ' if $\|s - s'\| \leq \delta$

Only possible if a few summary statistics suffice. We will later discuss refinements and extensions of this approach.

Beaumont, Zhang, Balding (2002)

"[...] the MCMC-based method is consistently superior to the summary-statistics-based methods and highlights that it is well worth making the effort to obtain full-data inferences if possible."

"[...] there are advantages to the use of summary statistics, both in the ease of implementation and in the time to obtain the results [...]"

"Further research is needed to find a more rigorous way for choosing summary statistics, including the use of orthogonalization and 'projection-pursuit' methods"

Some of the things you should be able to explain

Basic ideas of the following approaches in computational population genetics

- maximum likelihood with importance sampling
- full-data Bayesian with MCMC
- Approximate Bayesian Computation (ABC)

2 Importance sampling for genealogies

D: data set of DNA sequences sampled from a population. In case of a structured population sampling locations are known.

Aim: Estimate parameters $\Theta := (\theta_i, M_{ij})_{ij}$.

Maximum-Likelihood (ML) approach: Find the set of parameter values that maximizes the likelihood:

$$\widehat{\Theta} := \arg\max_{\Theta} \mathsf{Pr}_{\Theta}(D)$$

How to compute the likelihood? Let us first consider the case that genealogies without branch lengths are sufficient for the model:

$$L_D(\Theta) = \mathsf{Pr}_{\Theta}(D) = \sum_{\text{all genealogies } G} \mathsf{Pr}_{\Theta}(G) \cdot \mathsf{Pr}_{\Theta}(D \mid G).$$

Problem 1: There are many genealogies G

Idea: sample a smaller number of genealogies G and estimate the sum from their contributions $\mathsf{Pr}_{\Theta}(G) \cdot \mathsf{Pr}_{\Theta}(D \mid G)$.

Problem 2: many G contribute very little to the sum

Importance sampling: Sample preferably G that contribute more and correct for the sampling bias.

Importance Sampling

Assume we need to calculate a sum $\sum_{G \in \mathcal{G}} f(G)$ over a large set $\mathcal{G}$.

Further assume that p is a probability distribution on $\mathcal{G}$, such that

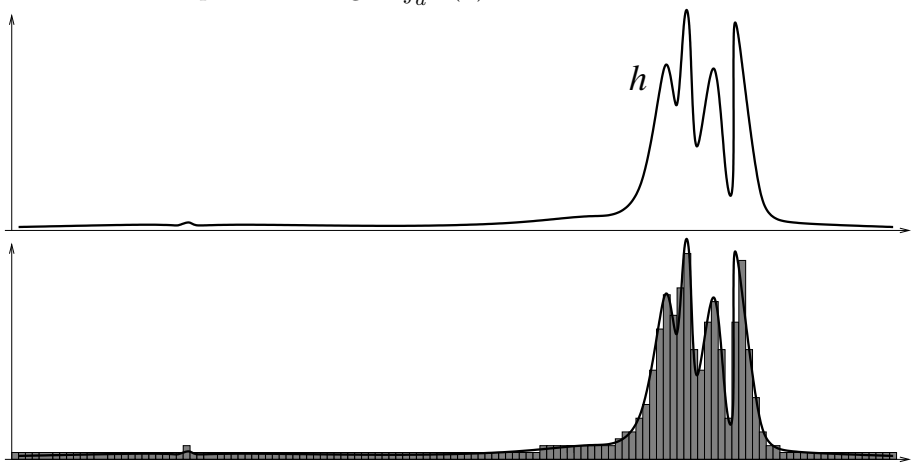
- $p(G)$ is *approximately* proportional to $f(G)$
- Given any $G \in \mathcal{G}$, there is an efficient way to calculate $f(G)/p(G)$
- There is an efficient way to generate random G according to the distribution p.

Then we can generate $G_1, G_2, \ldots, G_n$, all independent of each other according to p and approximate

$$\sum_{G \in \mathcal{G}} f(G) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(G_i)}{p(G_i)}.$$

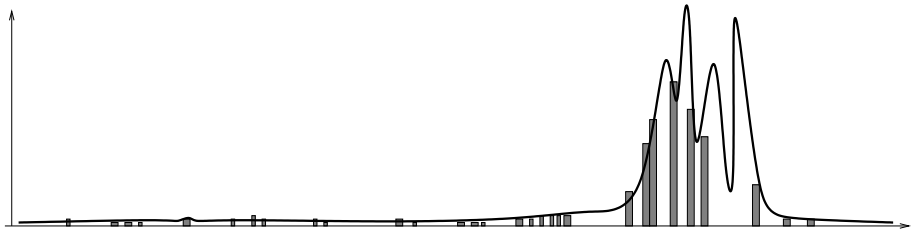
Note: if $p(G)$ is *exactly* proportional to $f(G)$, that is $f(G) = c \cdot p(G)$, then $\sum_G f(G) = c$, which we can caculate as $c = f(G_0)/p(G_0)$ with any G_0 with $f(G_0) \neq 0$.

How can we compute the integral $\int_a^b h(x)dx$ of this function h?



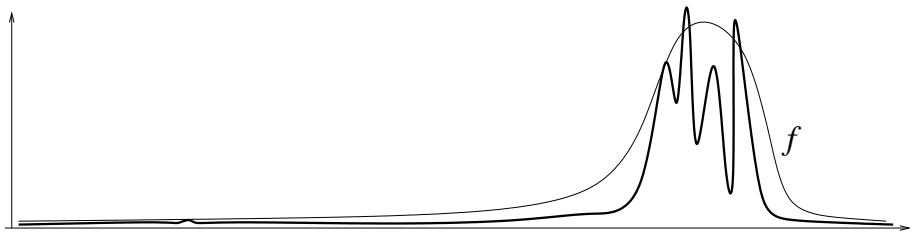
Approximation by a step function: If $x_1, \ldots, x_k$ are the means of the partition intervals and $c = \frac{b-a}{k}$ is their width, then

$$\int_a^b h(x)\ dx \approx \sum_{i=1}^{k} c \cdot h(x_i) = \frac{b-a}{k} \sum_{i=1}^{k} h(x_i).$$

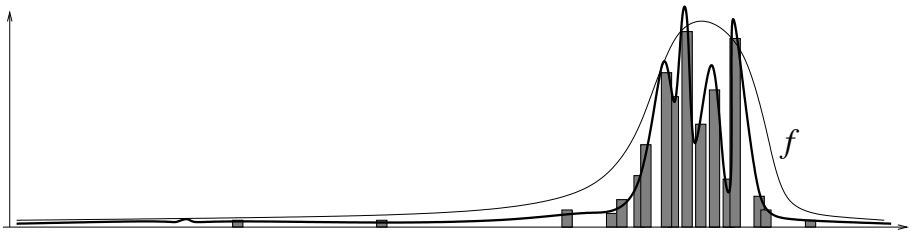


Maybe save some time by just taking a sample of k values $h(x)$.

$$\int_a^b h(x)\ dx \approx \frac{b-a}{k} \sum_{i=1}^{k} h(X_i) = \frac{1}{k} \sum_{i=1}^{k} \frac{h(X_i)}{\frac{1}{b-a}}.$$



Maybe we know a function f that approximates h



We can sample more from the relevant range but we have to correct this by the Importance-Sampling formula:

$$\int h(x)\ dx \approx \frac{1}{k}\sum_{i=1}^{k}\frac{h(X_i)}{q(X_i)}$$

where $X_1, \ldots, X_k$ are independent samples from a distribution whose density q is proportional to f. The closer f is to h, the better the approximation.

Sketch of proof of the IS formula

$$\begin{aligned}\int_a^b h(x)dx &= \int_a^b \frac{h(x)}{q(x)}\cdot q(x)dx \\ &= \mathbb{E}_q\frac{h(X)}{q(X)} \\ &\approx \frac{1}{k}\cdot\sum_{i=1}^{k}\frac{h(X_i)}{q(X_i)},\end{aligned}$$

where $\mathbb{E}_q$ is the expectation value under the assumption that X has probability density q, and $X_1, \ldots, X_k$ are independently sampled with probability density q.

2.1 Griffiths und Tavaré

References

[1] Griffiths and Tavaré (1994) Ancestral Inference in Population Genetics *Statistical Science* 9(3): 307-319.

We roughly follow the rationale and notations of

References

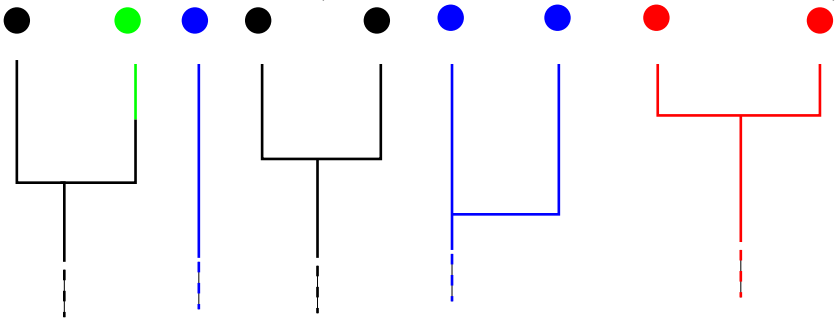
[1] Felsenstein, Kuhner, Yamato, Beerli (1999) Likelihoods on Coalescents: A Monte Carlo Sampling Approach to Inferring Parameters from Population Samples of Molecular Data. In: Seillier-Moiseiwitsch (Ed.): *Statistics in Molecular Biology and Genetics* IMS Lecture Notes-Monograph Series, Vol. 33, pp. 163–185

Start with an initial guess Θ_0. Define the history of a sample to be $H = (H_1, H_2, \ldots, H_\ell)$, where the historical events H_k can be

1. lineages i and j coalesce
2. mutation on lineage i
3. lineage i from island a traces back to island b

and $H_1, H_2, \ldots, H_\ell$ goes from present to past.

For the Importance Sampling procedure, many histories $H^{(1)}, H^{(2)}, \ldots, H^{(M)}$ are generated. For each history $H^{(i)}$ are sampled $H_1^{(i)}, H_2^{(i)}, \ldots$ step by step from the tips to the root of the tree. Given the data, not all events are possible. E.g., lineages cannot coalesce if they are of different allelic type. If the infinite-site mutation model is used (to make the Griffith-Tavaré scheme efficient), not all mutations are



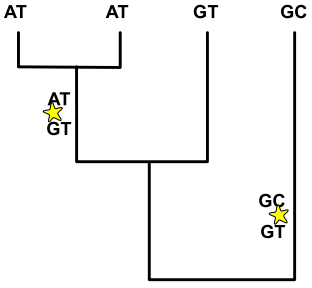
allowed.

- Let $b_{ij}(\theta_0)$ be the rates of the jth event $h = H_j^{(i)}$ in the ith sampled history $H^{(i)}$ and
- let $(a_{ijk}(\theta_0))_k$ be the series of rates of all events that are in accordance with the data and the infinite-sites assumption, and
- $(r_{ij}(\theta))$ the total rate of all possible events, disregarding the data.

Then, the probability to choose h was $b_{ij}(\theta_0)/\sum_k a_{ijk}(\theta_0)$. Thus, $\prod_j b_{ij}(\theta_0)/\sum_k a_{ijk}(\theta_0)$ is the importance-sampling probability $Q_{\theta_0}(H^{(i)})$ of the entire history $H^{(i)}$.According to the importance-sampling formula we get for all θ that are not too far from θ_0:

$$\begin{aligned} L_{(D)}(\theta) &\approx \frac{1}{M}\sum_{i=1}^{M}\left(\prod_j \frac{b_{ij}(\theta)}{r_{ij}(\theta)}\right) \Bigg/ \left(\prod_j \frac{b_{ij}(\theta_0)}{\sum_k a_{ijk}(\theta_0)}\right) \\ &= \frac{1}{M}\sum_{i=1}^{M}\prod_j \frac{b_{ij}(\theta)\cdot\sum_k a_{ijk}(\theta_0)}{r_{ij}(\theta)\cdot b_{ij}(\theta_0)} \end{aligned}$$

For example, let this be the ith sampled history under infinite-sites assumptions:

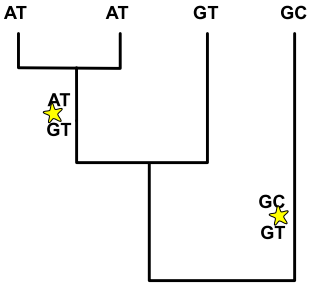


(only segregating sites shown)

step 1: $b_{i,1}(\theta) = 1$, $r_{ij}(\theta) = \binom{4}{2} + 4\cdot\frac{\theta}{2}$
step 2: $b_{i,2}(\theta) = \theta/2$, $r_{ij}(\theta) = \binom{3}{2} + 3\cdot\frac{\theta}{2}$
step 3: $b_{i,3}(\theta) = 1$, $r_{ij}(\theta) = \binom{3}{2} + 3\cdot\frac{\theta}{2}$
step 4: $b_{i,4}(\theta) = \theta/2$, $r_{ij}(\theta) = 1 + 2\cdot\frac{\theta}{2}$
step 5: $b_{i,4}(\theta) = 1$, $r_{ij}(\theta) = 1 + 2\cdot\frac{\theta}{2}$

$$\prod_j \frac{b_{ij}(\theta)}{r_{ij}(\theta)} = \frac{1}{6+2\cdot\theta}\cdot\frac{\theta/2}{3+1.5\cdot\theta}\cdot\frac{1}{3+1.5\cdot\theta}\cdot\frac{\theta/2}{1+\theta}\cdot\frac{1}{1+\theta}$$

For example, let this be the ith sampled history under infinite-sites assumptions:



(only segregating sites shown)

step 1: allowed are coalescence of first two lineages or mutation of 2nd position in sequence 4.

$$b_{i,1}(\theta_0) = a_{i,1,1}(\theta_0) = 1, \qquad a_{i,1,2}(\theta_0) = \theta_0/2$$

step 2: allowed is mutation of position 1 in on lineage 1 or position 2 on lineage on lineage 3

$$b_{i,2}(\theta_0) = a_{i,2,1}(\theta_0) = a_{i,1,2}(\theta_0) = \theta_0/2$$

step 3: allowed are coalescence of lineages 1 and 2 or mutation on lineage 3

$$b_{i,3}(\theta_0) = a_{i,3,1}(\theta_0) = 1, \qquad a_{i,3,2}(\theta_0) = \theta_0/2$$

step 4: position 2 can mutate in lineage 1 or 2

$$b_{i,4}(\theta_0) = a_{i,4,1}(\theta_0) = a_{i,4,2}(\theta_0) = \theta_0/2$$

$$\prod_j \frac{b_{ij}(\theta_0)}{\sum_k a_{ijk}(\theta_0)} = \frac{1}{1+\frac{\theta_0}{2}} \cdot \frac{\theta_0/2}{\frac{\theta_0}{2}+\frac{\theta_0}{2}} \cdot \frac{1}{1+\frac{\theta_0}{2}} \cdot \frac{\theta_0/2}{\frac{\theta_0}{2}+\frac{\theta_0}{2}}$$

- Advantage over MCMC: Histories are sampled really independent of each other.
- Disadvantage: For finite-sites models many different mutation events are allowed in each step, which makes the method very inefficient. Stephens and Donnelly (2000) found a solution for this, which we could discuss maybe later in the semester.

So, importance sampling helps to calculate

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \sum_{\text{all genealogies } G} \mathsf{Pr}_\Theta(G) \cdot \mathsf{Pr}_\Theta(D \mid G).$$

But how to do this if we account for (continuos) time points of events, including branch lengths, and have:

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \int_{\text{all genealogies } G} \mathsf{Pr}_\Theta(D \mid G)\ P_\Theta(G) dG$$

where $P_\Theta(G)$ is the density of the (structured) coalescent distribution at the genealogy G.

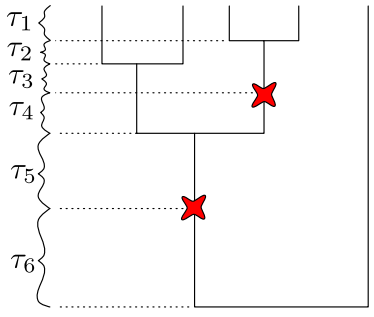
What does this mean?

How to compute $P_\Theta(G)$?

And what is dG?

How to compute $P_\Theta(G)$

Example: Kingman coalescent with mutations as part of G (different than in LAMARC). Let $r_k = \binom{k}{2} + k\frac{\theta}{2}$ be the total rate if k lineages are present.



$$\begin{aligned} P_\Theta(G) \quad = \quad & r_5 \cdot e^{-r_5 \cdot \tau_1} \cdot 1/r_5 \cdot \\ & r_4 \cdot e^{-r_4 \cdot \tau_2} \cdot 1/r_4 \cdot \\ & r_3 \cdot e^{-r_3 \cdot \tau_3} \cdot \theta/2/r_3 \cdot \\ & r_3 \cdot e^{-r_3 \cdot \tau_4} \cdot 1/r_3 \cdot \\ & r_2 \cdot e^{-r_2 \cdot \tau_5} \cdot \theta/2/r_2 \cdot \\ & r_2 \cdot e^{-r_2 \cdot \tau_6} \cdot 1/r_2 \end{aligned}$$

$$= e^{-r_5 \cdot \tau_1 - r_4 \cdot \tau_2 - r_3 \cdot (\tau_3 + \tau_4) - r_2 \cdot (\tau_5 + \tau_6)} \cdot \left(\frac{\theta}{2}\right)^2$$

So what is dG?

Let's first ask: What is the dx in

$$\int_0^1 x^2 dx \qquad ?$$

dx is used in an ambigous way. This is sloppy but intuitive.

It means "a small environment around x", but also the size of this environment.

To explain this we be a little bit less sloppy for a few minutes and write $\mathrm{d}x$ for the environment and dx for its size.

For some large $n \in \mathbb{N}$ and $x \in \mathbb{R}$ we can define $\mathrm{d}x = [x - \frac{1}{2n}, x + \frac{1}{2n}]$.Then, $dx = 1/n$.

We can approximate $\int_0^1 x^2 dx$ by

$$\sum_{x \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}} x^2 \cdot \frac{1}{n} = \sum_{x \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}} x^2 \cdot dx \overset{n \to \infty}{\to} \int_0^1 x^2 dx$$

dx is always meant to be "infinitesimally small", i.e. $dx \to 0$

What is a probability density?

$P(x)$ is the probability density of a random variable X in x if

$$\Pr(X \in \mathrm{d}x) \approx P(x) \cdot dx$$

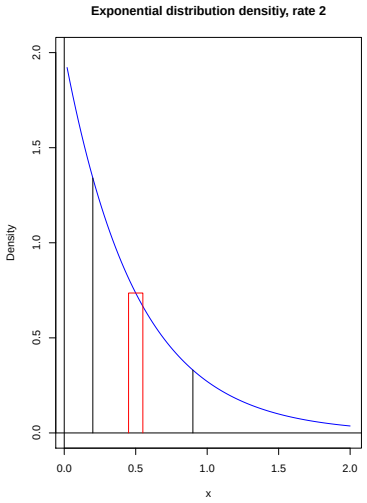
and the "$\approx$" becomes a "$=$" for "infinitesimally small" dx.

This is again sloppy and intuitive. It actually means that

$$\lim_{dx \to 0} \frac{\Pr(X \in \mathrm{d}x)}{dx} = P(x)$$

It then follows that

$$\Pr(X \in [a, b]) = \int_a^b P(x) dx.$$



Examples

The density of the exponential distribution with rate λ at x is

$$\lambda e^{-\lambda x}.$$

The density of the normal distribution with mean value μ and standard deviation σ is

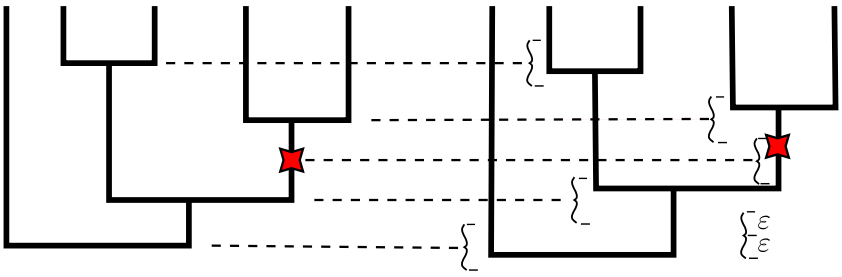
$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Now for dG

Let $\mathrm{d}G$ be a small environment around the genealogy G. This means, $\mathrm{d}G$ consists of all genealogies G' that have the same topology as G and if $\tau_1, \ldots, \tau_n$ are the points in time where coalescent events or migrations of lineages or thelike occurr in G, and $\tau'_1, \ldots, \tau'_n$ are the corresponding points in time for G', then

$$\forall_{k \leq n} |\tau_k - \tau'_k| \leq \varepsilon.$$

Thus, the volume dG of $\mathrm{d}G$ can be defined to be $(2\varepsilon)^n$.



The density $P_\Theta(G)$ is then defined by

$$\mathsf{Pr}_\Theta(G' \in \mathsf{d}G) \approx P_\Theta(G) \cdot dG$$

where $\mathrm{Pr}_\Theta(G' \in \mathsf{d}G)$ is the probability that a genealogy G' that was generated according to the probability distribution of a structured coalecent with parameter values Θ results to be in the environment $\mathrm{d}G$ of G, or, more precisely:

$$\frac{\mathrm{Pr}_\Theta(G' \in \mathsf{d}G)}{dG} \overset{dG \to 0}{\longrightarrow} P_\Theta(G)$$

The equation

$$L_D(\Theta) = \mathsf{Pr}_\Theta(D) = \int_{\text{all genealogies}} \mathsf{Pr}_\Theta(D \mid G)\ P_\Theta(G) dG$$

should now make some more sense to us. But how can we compute it? We use Importance Sampling.

Importance Sampling for computing the likelihood of for a range of parameter values Θ: Generate genealogies $G_1, \ldots, G_k$ (more or less) independently according to a probability density $Q(G_i)$ that **we can calculate**. Then,

$$\begin{aligned} L_D(\Theta) &= \int_{\text{all genealogies } G} \Pr_\Theta(D|G) \cdot P_\Theta(G) dG \\ &\approx \frac{1}{k} \sum_{i=1}^{k} \frac{\mathrm{Pr}_\Theta(D|G_i) \cdot P_\Theta(G_i)}{Q(G_i)}. \end{aligned}$$

Method differ in their choice of Q (e.g. via MCMC, see next section) and will be most efficient if

$$Q(G) \approx \text{const.} \cdot \Pr_\Theta(D|G) \cdot P_\Theta(G).$$

Note: If $Q(G) = c \cdot \mathrm{Pr}_\Theta(D|G) \cdot P_\Theta(G)$ then $L_D(\Theta) = \frac{1}{c}$.

Some of what you should be able to explain

- The formula
$$L_D(\Theta) = \int \mathrm{Pr}_\Theta(D|G) P_\Theta(G) dG$$
and all its terms, including dG
- what is a probability density, in particular $P_\Theta(G)$
- how to compute $P_\Theta(G)$
- importance sampling: idea and formula
- Griffiths-Tavaré method

3 MCMC with Lamarc (or Migrate)

Rate parameters and time scales

For autosomal DNA:

	per generation	per $2N_i$ generations	per $1/\mu$ generations
mutation rate	μ	$\frac{\theta_i}{2} = 2N_i\mu$	1
migration rate of ancestral lineage from i tracing back to j	m_{ij}	$\gamma_{ij} = 2N_i m_{ij}$	$M_{ij} = \frac{m_{ij}}{\mu} = \frac{2\gamma_{ij}}{\theta_i}$
coalescence on island i	$1/(2N_i)$	1	$\frac{1}{2N_i\mu} = \frac{2}{\theta_i}$

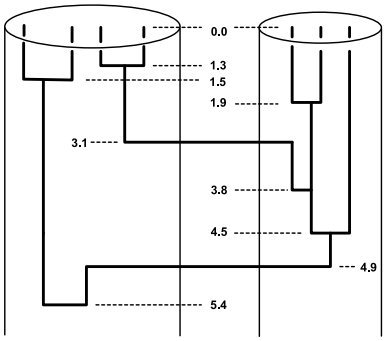
Number of alleles on island i that choose their parent allele on island j:

$$2N_i \cdot m_{ij} = \gamma_{ij}$$

example: two populations, time unit $1/\mu$ generations

total rate when k lineages in pop 1 and ℓ lineages in pop 2:

$$r(k,\ell) = \frac{k\cdot(k-1)}{2}\cdot\frac{2}{\theta_1} + k\cdot M_{1,2} + \frac{\ell\cdot(\ell-1)}{2}\cdot\frac{2}{\theta_2} + \ell\cdot M_{2,1} = \frac{k\cdot(k-1)}{\theta_1} + k\cdot M_{1,2} + \frac{\ell\cdot(\ell-1)}{\theta_2} + \ell\cdot M_{2,1}$$



$$\begin{aligned} p(G) \quad = \quad & r(4,3)\cdot e^{-r(4,3)\cdot 1.3}\cdot\frac{2/\theta_1}{r(4,3)}\cdot \\ & \cdot r(3,3)\cdot e^{-r(3,3)\cdot 0.2}\cdot\frac{2/\theta_1}{r(3,3)}\cdot \\ & \cdot r(2,3)\cdot e^{-r(2,3)\cdot 0.4}\cdot\frac{2/\theta_2}{r(2,3)}\cdot \\ & \cdot r(2,2)\cdot e^{-r(2,2)\cdot 1.2}\cdot\frac{M_{1,2}}{r(2,2)}\cdot \\ & \cdot r(1,3)\cdot e^{-r(1,3)\cdot 0.7}\cdot\frac{2/\theta_2}{r(1,3)}\cdot \end{aligned}$$

$$\cdot r(1,2)\cdot e^{-r(1,2)\cdot 0.7}\cdot\frac{\theta_2/2}{r(1,2)}\cdot r(1,1)\cdot e^{-r(1,1)\cdot 0.4}\cdot\frac{M_{2,1}}{r(1,1)}\cdot r(2,0)\cdot e^{-r(1,1)\cdot 0.5}\cdot\frac{2/\theta_1}{r(2,0)}$$

$$= e^{-r(4,3)\cdot 1.3}\cdot\theta_1\cdot e^{-r(3,3)\cdot 0.2}\cdot\theta_1/2\cdot e^{-r(2,3)\cdot 0.4}\cdot\theta_2/2\cdot e^{-r(2,2)\cdot 1.2}\cdot M_{1,2}\cdot e^{-r(1,3)\cdot 0.7}\cdot\theta_2/2\cdot e^{-r(1,2)\cdot 0.7}\cdot\theta_2/2\cdot e^{-r(1,1)\cdot 0.4}\cdot M_{2,1}\cdot e^{-r(1,1)\cdot 0.5}\cdot\theta_1/2 = e^{-r(4,3)\cdot 1.3 - r(3,3)\cdot 0.2 - r(2,3)\cdot 0.4 - r(2,2)\cdot 1.2 - r(1,3)\cdot 0.7 - r(1,2)\cdot 0.7 - r(1,1)\cdot 0.4 - r(1,1)\cdot 0.5}\cdot\left(\frac{2}{\theta_1}\right)^3\cdot\left(\frac{2}{\theta_2}\right)^3\cdot M_{1,2}\cdot M_{2,1}$$

Combining IS with MCMC

References

[1] M. Kuhner, J. Yamato, J. Felsenstein (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hasings sampling. *Genetics* **140**: 1421–1430

[2] P. Beerli, J. Felsenstein (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *PNAS* **98.8**: 4563–4568

- MIGRATE-N http://popgen.sc.fsu.edu/Migrate/Migrate-n.html
- LAMARC http://evolution.genetics.washington.edu/lamarc/lamarc.html

LAMARC strategy

Begin with initial parameter guess $\Theta_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, M_{12}^{(0)}, M_{12}^{(0)}, M_{23}^{(0)}, \ldots)$, repeat the following steps for $i = 0, 1, 2, \ldots, m-1$

1. Metropolis-Hastings MCMC sampling of genealogies $G_1, G_2, \ldots, G_k$ (approx.) according to the posterior density $p_{\Theta_i}(G|D)$ given the data D. What is Metropolis-Hastings MCMC?

2. importance sampling:
$$\frac{L_D(\Theta)}{L_D(\Theta_i)} \approx \frac{1}{k}\sum_{j=1}^{k} \frac{p_\Theta(G_j)}{p_{\Theta_i}(G_j)} =: F_{\Theta_i}(\Theta)$$
Why is this justified as importance sampling?

3. $\Theta_{i+1} := \arg\max_\Theta F_{\Theta_i}(\Theta)$

and hope that $\Theta_m \approx \widehat{\Theta} = \arg\max_\Theta L_D(\Theta)$

Justification of step 2

$$\begin{aligned}
\frac{L_D(\Theta)}{L_D(\Theta_i)} &\approx \frac{\frac{1}{k}\sum_{j=1}^{k} \frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j|D)}}{\Pr_{\Theta_i}(D)} \qquad \text{(importance sampling)} \\
&= \frac{1}{k}\sum_{j=1}^{k} \frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j|D)\cdot \Pr_{\Theta_i}(D)} \\
&= \frac{1}{k}\sum_{j=1}^{k} \frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{p_{\Theta_i}(G_j, D)} \\
&= \frac{1}{k}\sum_{j=1}^{k} \frac{\Pr_\Theta(D|G_j)\cdot p_\Theta(G_j)}{\Pr_{\Theta_i}(D|G_j)\cdot p_{\Theta_i}(G_j)} = \frac{1}{k}\sum_{j=1}^{k} \frac{p_\Theta(G_j)}{p_{\Theta_i}(G_j)}
\end{aligned}$$

The last equation follows from $\Pr_\Theta(D|G_j) = \Pr_{\Theta_i}(D|G_j)$, which holds since the mutation rate is always 1 and thus the D is independent of Θ when G is given.

Markov-Chain Monte Carlo (MCMC)

MCMC: construct Markov chain $X_0, X_1, X_2, \ldots$ with stationary distribution $\Pr(G \mid D)$ and let it converge.

Markov property:

$$\forall_{i,x}: \quad \Pr(X_{i+1} = x|X_i) = \Pr(X_{i+1} = x|X_i, X_{i-1}, \ldots, X_0)$$

In words: The probabilty for the next state may depend on the current state but not additionally on the past.

"Equilibrium" or "Stationary distribution" p:

$$\forall_{i,x}: \quad p(x) = \sum_y p(y)\cdot \Pr(X_{i+1} = x|X_i = y)$$

In words: If you choose an element of the state space according to p and go one step, the probability to be in x is $p(x)$ not only in the first step but also in the second step and consequently in any further step.When you are once in equilibrium, you'll be forever.

Theorem 2 *If $X_0, X_1, X_2 \ldots$ is a aperiodic, irreducible Markov chain on a finite state space S with equilibrium p, it will converge against the equilibrium p in the following sense:*

$$\forall_{x,y}: \quad \Pr(X_n = x|X_0 = y) \overset{n\to\infty}{\longrightarrow} p(x)$$

Irreducible means:

$$\forall_{x,y}\exists_i\forall_m: \quad \Pr(X_{i+m}=x|X_m=y)>0$$

Aperiodic means:

$$\forall_{x,y,m}: \quad \gcd\left(\{k\in\mathbb{N}|\Pr(X_{k+m}=x|X_m=y)>0\}\right)=1,$$

where gcd means "greatest common divisor".

(let's watch a Tcl/Tk simulation of a Markov chain)

"Equilibrium" or "Stationary distribution" p:

$$\forall_{i,x}: \quad p(x)=\sum_y p(y)\cdot\Pr(X_{i+1}=x|X_i=y)$$

Stronger condition than equilibrium: reversibility (or "detailed balance")

$$p(x)\cdot\Pr(X_{i+1}=y|X_i=x)=p(y)\cdot\Pr(X_{i+1}=x|X_i=y)$$

In words: If you start in equilibrium, and it is reversible, a move from x to y is as probable as a move from y to x.

Alternative explanation: If you watch a movie of the process starting in a reversible equilibrium, the probability of what you see does not change if you watch the movie backwards.

Given the probability distribution $\Pr(.|D)$, how can we construct a Markov chain that converges against it?

One possibility: **Metropolis-Hastings**

Given current state $X_i=x$ propose y with Prob. $Q(x\to y)$

Accept proposal $X_{i+1}:=y$ with probability

$$\min\left\{1,\frac{Q(y\to x)\cdot\Pr(y\mid D)}{Q(x\to y)\cdot\Pr(x\mid D)}\right\}$$

otherwise $X_{i+1}:=X_i$

(All this also works with continuous state space, with some probabilities replaced by densities.)

Why Metropolis-Hastings works

Let's assume that $\frac{Q(y\to x)\cdot\Pr(y\mid D)}{Q(x\to y)\cdot\Pr(x\mid D)}\le 1$. (Otherwise swap x and y in the following argument).Then, if we start in x, the probability $\Pr(x\to y)$ to move to y (i.e. first propose and then accept this) is

$$Q(x\to y)\cdot\frac{Q(y\to x)\cdot\Pr(y\mid D)}{Q(x\to y)\cdot\Pr(x\mid D)}=Q(y\to x)\frac{\Pr(y\mid D)}{\Pr(x\mid D)}$$

If we start in y, the probability $\Pr(y\to x)$ to move to x is

$$Q(y\to x)\cdot 1,$$

since our assumption implies $\frac{Q(x\to y)\cdot\Pr(x\mid D)}{Q(y\to x)\cdot\Pr(y\mid D)}\ge 1$.

This implies that the reversibility condition

$$\Pr(x\mid D)\cdot\Pr(x\to y)=\Pr(y\mid D)\cdot\Pr(y\to x)$$

is fulfilled.This implies that $\Pr(.\mid D)$ is an equilibrium of the Markov chain that we have just constructed, and the latter will converge against it.(let's watch a simulation in R)

Applying Metropolis-Hastings

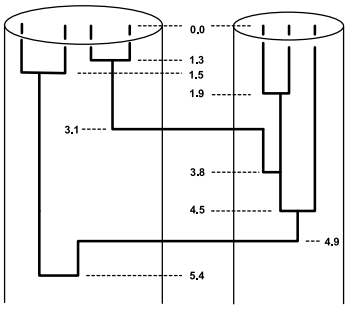
- You are never in equilibrium (your target distribution), but you can get close if you run enough steps.
- You can take more than one sample from the same chain, but you should run enough steps between the sampling steps to make the sampled objects only weakly dependent.
- Your initial state may be "far from equilibrium" (i.e. very improbable). So you should run the chain long enough before you start sampling ("burn-in").

Lamarc's Metropolis-Hastings step

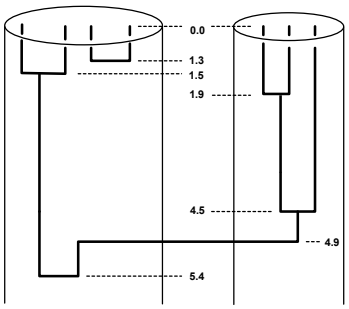
Proposal chain: Remove a randomly picked branch and let the ancestral lineage of the isolated subtree coalesce with the rest accoring to Θ. Total rate of all that can happen to lineage on island 1 or 2 with k other lineages:

$$\lambda_1(k) = \frac{k \cdot 2}{\theta_1} + M_{1,2} \qquad \lambda_2(k) = \frac{k \cdot 2}{\theta_2} + M_{2,1}$$

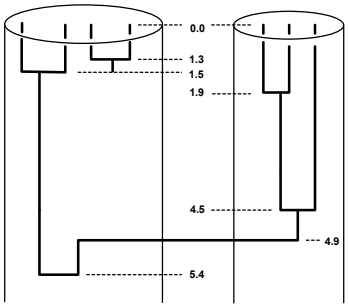
Example:



Step 1: Choose a random branch (each with same prob) and remove it.

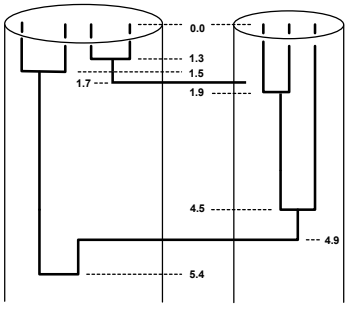


To reconnect genealogy generate missing branch according to coalescent model.



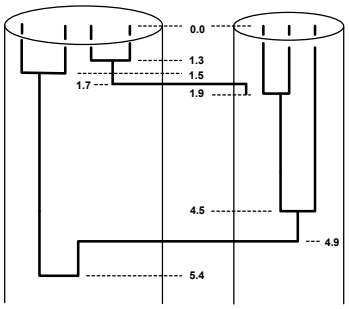
Step 2:

- generate random time T_1 from exp-distribution with rate $\lambda_1(2)$.
- If $1.3 + T_1 < 1.5$, this is the time for the next event (coalescence or migration).
- Otherwise (as shown here) stop at 1.5.



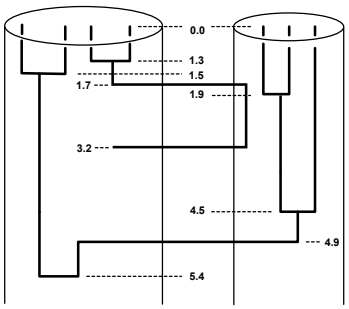
Step 3:

- draw exponential time T_2 with rate $\lambda_1(1)$
- Let's say $T_2 = 0.2$ (rounded)
- As $1.5 + T_2 < 1.9$, an event takes place
- Choose migration (as here) with prob. $M_{1,2}/\lambda_1(1)$ or coalescence with prob. $(2/\theta_1)/\lambda_1(1)$



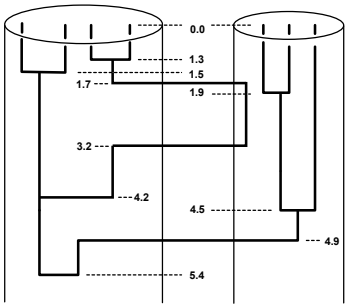
Step 4: continue at time 1.7

- draw exponential time T_3 with rate $\lambda_2(3)$
- if $1.7 + T_3 > 1.9$ continue line down to 1.9
- (we always stop where the rate needs to be updated)



Step 5: continue at 1.9

- draw exponential time T_4 with rate $\lambda_2(2)$
- Let's say $T_4 \approx 1.3$, such that $1.9 + T_4 = 3.2 < 4.5$.
- Choose migration event with prob. $M_{2,1}/\lambda_2(2)$.
- [or coalescence with prob. $(2/\theta_2)/\lambda_2(2)$]



Step 6:

- draw exponential time T_5 with rate $\lambda_1(1)$, let's say $T_5 \approx 1.0$.
- as $3.2 + 1.0 < 4.9$, event takes place, e.g. with prob. $(2/\theta_1)/\lambda_1(1)$ a coalescence

Lamarc's Metropolis-Hastings step

Target distribution density: $p_\Theta(G|D)$, where Θ is the current set of parameter values, G is the genealogy and D is the data.

Proposal chain: Remove a randomly picked branch and let the ancestral lineage of the isolated subtree coalesce with the rest accoring to Θ.

$\Rightarrow$

$$\frac{Q(G' \to G)}{Q(G \to G')} = \frac{p_\Theta(G)}{p_\Theta(G')}$$

$\Rightarrow$ The MH acceptance probability is:

$$\begin{aligned}
\min\left\{1, \frac{Q(G' \to G) \cdot p_\Theta(G'|D)}{Q(G \to G') \cdot p_\Theta(G|D)}\right\} &= \min\left\{1, \frac{p_\Theta(G) \cdot p_\Theta(G', D)/Pr(D)}{p_\Theta(G') \cdot p_\Theta(G, D)/Pr(D)}\right\} \\
&= \min\left\{1, \frac{p_\Theta(G) \cdot \Pr(D|G') \cdot p_\Theta(G')}{p_\Theta(G') \cdot \Pr(D|G) \cdot p_\Theta(G)}\right\} \\
&= \min\left\{1, \frac{\Pr(D|G')}{\Pr(D|G)}\right\}
\end{aligned}$$

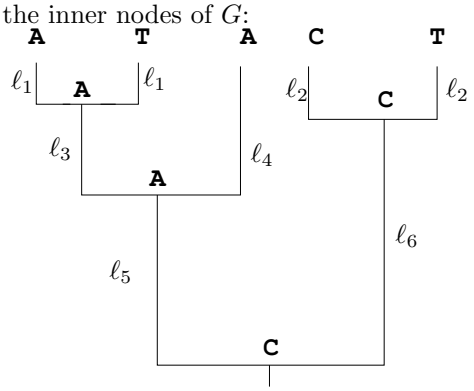
How to compute $\Pr(D|G)$? Felsenstein's pruning!

We assume that all sites evolve independent of each other. $\Rightarrow$

$$\Pr(D|G) = \prod_i \Pr(D_i|G),$$

where D_i is the i-th column in the alignment.

How to compute $\Pr(D_i|G)$? For any nucleotides (or amino acids) x, y let p_x be the frequency of x and $\Pr_{x\to y}(\ell)$ be the probability that a child node has type y, given that the parent node had type x and the branch between the two nodes has length ℓ. Let's first assume that D_i knows the nucleotides at the inner nodes of G:



$$\begin{aligned}
&\Pr(D_i|G) \\
&= p_C \cdot \Pr_{C\to A}(\ell_5) \cdot \Pr_{C\to C}(\ell_6) \cdot \\
&\quad \Pr_{A\to A}(\ell_3) \cdot \Pr_{A\to A}(\ell_4) \cdot \\
&\quad \Pr_{A\to A}(\ell_1) \cdot \Pr_{A\to T}(\ell_1) \cdot \\
&\quad \Pr_{C\to C}(\ell_2) \cdot \Pr_{C\to T}(\ell_2) \cdot
\end{aligned}$$

How to compute or define $\Pr_{x\to y}(\ell)$?

Jukes-Cantor model for DNA evolution

- All nucleotide frequencies are $p_A = p_C = p_G = p_T = 0.25$.
- "mutation events" happen at rate λ and let the site forget its current type and select a new one randomly from {A,C,G,T}. (New one can be the same as old one.)

$\Rightarrow$

$$\Pr_{x\to y}(\ell) = \begin{cases} = & \left(1 - e^{-\lambda\ell}\right) \cdot \frac{1}{4} & \text{if } x \neq y \\ = & e^{-\lambda\ell} + \left(1 - e^{-\lambda\ell}\right) \cdot \frac{1}{4} & \text{if } x = y \end{cases}$$

(More sophisticated sequence evolution models in the phylogenetics part of the lecture.)

Felsenstein's pruning algorithm

How to compute $\Pr(D_i|G)$ if (as usual) the data do only contain the nucleotides for the tips of the tree?

For any node k of the genealogy and any nucleotide (or amino acid) x define $w_k(x)$ to be the probability that, given the nucleotide (or a.a.) in k is x, the tips that stem from k get the nucleotides (or a.a.) given in D_i. Then

$$\Pr(D_i|G) = \sum_{x\in\{A,C,G,T\}} p_x \cdot w_r(x),$$

where r is the root of the genealogy,and for any node k with child nodes i and j and corresponding branch lengths ℓ_i and ℓ_j we get:

$$w_k(x) = \left(\sum_{y\in\{A,C,G,T\}} \Pr_{x\to y}(\ell_i)\cdot w_i(y)\right)\cdot\left(\sum_{z\in\{A,C,G,T\}} \Pr_{x\to y}(\ell_j)\cdot w_j(z)\right)$$

Felsenstein's pruning algorithm

If b is a tip of G, then $w_b(x)$ is 1 if x is the nucleotide at b in D_i, and $w_b(x)$ is 0 otherwise.

With the recursion for$w_k(x)$ given above, we can compute $w_k(x)$ for all x and all k starting with the tips and ending in the root r.

From the $w_r(.)$ we can compute $\Pr(D_i|G)$.

Some of the things you should be able to explain

- Time scaling of $1/\mu$ generations and how to calculate $p(G)$ for the structured coalescent
- convergence of Markov chains against equilibria
  - formula and what it means
  - sufficient conditions for this convergence
  - when is an equilibrium reversible and why are not all equilibria reversible
- MCMC, in particular Metropolis-Hastings
- How Importance Sampling and MCMC is combined in LAMARC
  - proposals for changes in the genealogy
  - why LAMARC's Metropolis-Hastings ratio is so simple
  - Felsenstein's pruning
  - Jukes-Cantor model and how to compute its transition probabilities

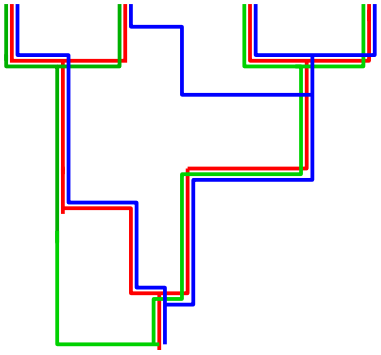
Other MCMC strategy than in LAMARC

References

[1] I. J. Wilson, D. J. Balding (1998) Genealogical inference from microsatellite data. *Genetics* **150**: 499-–510

- assign data to inner nodes
- when choosing new parent node take mutation probs into account
- more intelligent proposals but larger state space
- may be superior for microsatellite data

Ancestral Recombination Graph



When recombination occurs, ancestral lineages for the left and the right part of the sequence split up.
Each site has a tree-shaped ancestry, and these trees have complex stochastic dependencies.
LAMARC can also sample Ancestral Recombination Graphs instead of trees. If k lineages are present, the total recombination rate is $\rho \cdot k$.
As the coalescence rate is

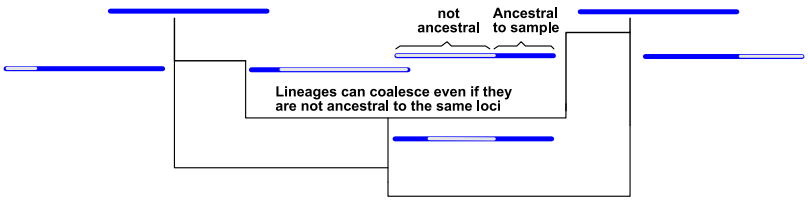
$$\propto k \cdot (k-1),$$

the number of lineages will not go to ∞.
Due to random fluctuations, the number of lineages will sooner or later reach 1 (ultimate ancestor).

LAMARC must define for each recombination event which segments follow which lineage.

Excursus: simulation of ARG in ms

- Must define:
  - recombination rate (per locus)
  - number of positions (or unbreakable segments)
- Stop when each nucleotide (or segment) has found its MRCA
- Can be very slow when large genomic regions are simulate
- Other programs such as scrm allow to neglect long-range dependencies



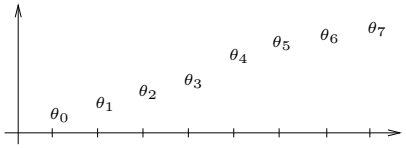
LAMARC Search Strategies

initial chains: several short chains to optimize driving values

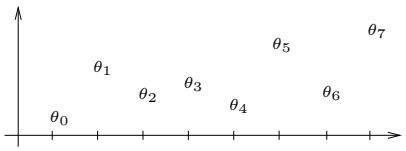
final chain: longer chain to narrow the final interval

burn-in: discard e.g. first 5% of each chain

symptom of too few chains: parameters are still changing directionally



symptom of too short chains: parameters leap wildly from chain to chain



$(\mathsf{MC})^3$=MCMCMC

=Metropolis-Coupled MCMC= MCMC with "heated chains".

If $\beta_i \in (0,1]$ is heat parameter for chain i, then chain i samples from distribution $p^{\beta_i} : x \mapsto p^{\beta_i}(x)\cdot\text{const}$, with $\beta_1 = 1$.

The usual MH acceptance prob. for chain i is

$$\min\left\{1, \frac{p(y)^{\beta_i}}{p(x)^{\beta_i}} \cdot \frac{Q_{y\to x}}{Q_{x\to y}}\right\}.$$

Sometimes a swap between the current state x_i of chain i and the current state x_j of chain j is proposed. The acceptance with probability

$$\min\left\{1, \frac{p(x_j)^{\beta_i}}{p(x_i)^{\beta_i}} \cdot \frac{p(x_i)^{\beta_j}}{p(x_j)^{\beta_j}}\right\}$$

fulfills the requirements of both chaines (check this!).

Bayesian Lamarc

Aim: sample parameter values Θ (and Genealogies) according to the posterior probability distribution $\Pr(\Theta|D)$ (or $\Pr(\Theta, G|D)$) given the data D.

- needs priors $\Pr(\Theta)$ for the parameters
- Gibbs sampling scheme: iterate uptdate of the Θ, given D and G, and update of G, given Θ and D.

Gibbs samping

Assume we want to sample from a joint distribution $\Pr(A = a, B = b)$ of two random variables, and for each pair of possible values (a, b) for (A, B) we have Markov chains with transition probabilities $P^{(A=a)}_{b\to b'}$ and $P^{(B=b)}_{a\to a'}$ that converge against $\Pr(B = b|A = a)$ and $\Pr(A = a|B = b)$.

Then, any Markov chain with transition law

$$P_{(a,b)\to(a',b')} = \begin{cases} \frac{1}{2}P^{(B=b)}_{a\to a} + \frac{1}{2}P^{(A=a)}_{b\to b} & \text{if} \quad a = a' \quad \text{and} \quad b = b' \\ \frac{1}{2}P^{(B=b)}_{a\to a'} & \text{if} \quad a \neq a' \quad \text{and} \quad b = b' \\ \frac{1}{2}P^{(A=a)}_{b\to b'} & \text{if} \quad a = a' \quad \text{and} \quad b \neq b' \\ 0 & \text{else} \end{cases}$$

Priors in Bayesian Lamarc

When new values for Θ are to be proposed,

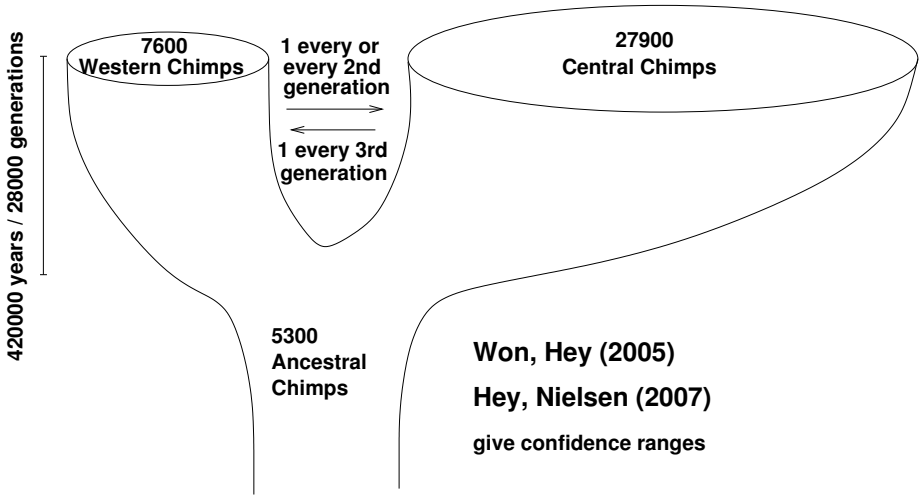
- e.g. the new values of θ and the recombination rate are chosen according to a exponential prior that is uniform on the log scaled interval $[10^{-5}, 10]$and the
- growth rate g is chosen uniformly from $[-500, 1000]$.
- For the MH acceptance step use a U that is uniform on $[0, 1]$ and accept if

$$U < \frac{\Pr(G|\Theta_{\text{proposal}})}{\Pr(G|\Theta_{\text{old}})}$$

Some of the things you should be able to explain

- Difference between LAMARC and Wilson-Balding approach
- Ancestral Recombination Graph (ARG) and its properties
- short and long chains in LAMARC
- how to check convergence in LAMARC
- MCMCMC (heated chains)
- Differences btw. Bayesian and frequentist LAMARC
- Gibbs sampling

4 IM, IMa, IMa2



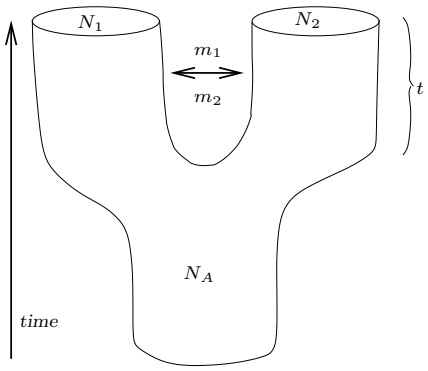
References

[1] Nielsen, R. and J. Wakeley 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**:885-896

[2] Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics* **167**:747-760

[3] Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *PNAS* **104**:2785–2790.

[4] Hey J. 2010. Isolation with Migration Models for More Than Two Populations. *Mol Biol Evol* **27**: 905-20



m_i: proportion of individuals of pop. i that are replaced by immigrants

N_i: effective size of pop. i

N_A: effective size of ancestral population

t: number of generations since the split

μ: mutation rate per generation

Asymptotics and rescaled parameters:

$$\begin{array}{rcl@{\qquad}rcl} N_i & \to & \infty & 2N_i m_i & \to & M_i \\ N_2/N_1 & \to & r & 4N_1\mu & \to & \theta \\ N_A/N_1 & \to & a & t/(2N_1) & \to & \tau \end{array}$$

$$\Theta = (\theta, r, a, \tau, M_1, M_2)$$

IM is an implementation of a Bayesian sampler with flat priors, e.g.

$$\begin{array}{rcl@{\qquad}rcl} M_i & \sim & \mathsf{Unif}([0,10]), & \tau & \sim & \mathsf{Unif}([0,10]) \\ \log(r) & \sim & \mathsf{Unif}([-10,10]), & \log(a) & \sim & \mathsf{Unif}([-10,10]) \end{array}$$

Proposals G^* for genealogy updates like in Lamarc with MH acceptance probability

$$\min\left\{1, \frac{\Pr(D|\Theta_i, G^*)}{\Pr(D|\Theta_i, G_i)}\right\},$$

where G_i is the current genealogy and Θ_i is the current vector of parameter values in MCMC step i.

Proposals for parameter updates: Given the current value λ of some parameter, the new value is proposed from $\mathrm{Unif}[\lambda - \Delta, \lambda + \Delta]$. MH acceptance probability:

$$\min\left\{1, \frac{p(G_i|\Theta^*)}{p(G_i|\Theta_i)}\right\}$$

IM can handle datasets of unlinked loci (but NO intralocus-recombination!).

$D = (D^1, \ldots, D^n)$, D^i: data from locus i. $G = (G^1, \ldots, G^n)$, G^i: genealogy of locus i (including topology, branch lengths, migration times, coalescent times)

$$p(\Theta|D) = \frac{p(\Theta)\cdot \Pr(D|\Theta)}{\Pr(D)} = \frac{p(\Theta)\cdot \prod_i Pr(D_i|\Theta)}{\Pr(D)} = \frac{p(\Theta)}{\Pr(D)} \cdot \prod_{i=1}^{n} \int_{G^i} \Pr(D^i|G^i, \Theta)\cdot p(G^i|\Theta) dG^i$$

additional parameters: locus-specific mutation scalars u_i with constraint $\prod_i u_i = 1$.

Updating $(u_1, \ldots, u_n)$: choose i and j and propose

$$u_i^* = x \cdot u_i \text{ and } u_j^* = u_j/x,$$

where $\log(x) \sim \mathsf{Unif}(-\delta, \delta)$.

In IMa, some MCMC steps are replaced by faster numerical computation. We discuss this first in a 1-population model with sample size m.

- Let τ_k be the time while the number of lineages is k, measured in $1/\mu$ generations.
- $\Rightarrow$ coalescence rate is $2/\theta$

- $\Rightarrow p(G|\Theta) = \left(\frac{2}{\theta}\right)^{m-1} \cdot \exp(-f_m/\theta)$,
- where $f_m := \sum_{i=2}^{m} \tau_i \cdot i \cdot (i-1)$

Assume a flat prior $\theta \sim \mathsf{Unif}(0, \theta_{max})$. This implies (with the substitution $\theta = f_m/x$):

$$\begin{aligned} p(G) &= \int_0^{\theta_{\mathsf{max}}} p(\theta) \cdot p(G|\theta)\ d\theta \ = \ \frac{1}{\theta_{\mathsf{max}}} \int_{\infty}^{f_m/\theta_{\mathsf{max}}} \left(\frac{2x}{f_m}\right)^{m-1} \cdot e^{-x} \cdot \frac{-f_m}{x^2}\ dx \\ &= \frac{2^{m-1}}{\theta_{\mathsf{max}} f_m^{m-2}} \cdot \Gamma(m-2, f_m/\theta_{\mathsf{max}}), \end{aligned}$$

where $\Gamma(a,b) = \int_b^{\infty} x^{a-1}e^{-x}dx$ is the "incomplete Gamma-function".

This implies

$$p(\theta|G) = \frac{p(G|\theta) \cdot p(\theta)}{p(G)} = \frac{(f_m/\theta)^{m-2} \exp(-f_m/\theta)}{\theta \cdot \Gamma(m-2, f_m/\theta_{\mathsf{max}})}$$

Hence, given f_m, the posterior probability can be computed and the expression above gives a smooth curve.

- works in a similar way for models with subpopulations with migration
- for the split time τ a standard MH step is required
- population growth not allowed in IMa (other than IM)
- "branch sliding" proposals for G: move randomly chosen branch a random distance. Current migration events are removed and replaced by a Poisson number of migration events conditioned on odd or even.

Likelihood Ratio Testing with IMa

Let

$$\widehat{\Theta}_0 = \mathsf{arg}\max p(\Theta|D) \text{ in the general model}$$

and

$$\widehat{\Theta}_r = \mathsf{arg}\max p(\Theta|D) \text{ in a restricted model, e.g. without migration.}$$

Since we use uniform priors for all parameters (some log-scaled), we get

$$\frac{p(\Theta_0|D)}{p(\Theta_r|D)} = \frac{\Pr(D|\Theta_0) \cdot p(\Theta_0)}{\Pr(D|\Theta_r) \cdot p(\Theta_r)} = \frac{L_D(\Theta_0)}{L_D(\Theta_r)}$$

Hence, $\widehat{\Lambda} = \log\left(\frac{\widehat{p}(\Theta_0|D)}{\widehat{p}(\Theta_r|D)}\right)$ is an approximation of the log likelihood-ratio and thus, $2\widehat{\Lambda}$ is approximately χ^2_d-distributed under the null hypothesis of the restricted model, where d is the number of additional parameters in the general model.However, this approximation is only appropriate for extremely large datasets. IMa assesses the significance of $\widehat{\Lambda}$ by comparing it to values of $\widehat{\Lambda}$ from simulations based on the null hypothesis (restricted model).

Bayes factors

Other authors use so-called Bayes factors to decide between two models M_1 and M_2:

$$B_{M_1,M_2} = \frac{\Pr(D|M_1)}{\Pr(D|M_2)},$$

where

$$\begin{aligned} \Pr(D|M) &= \int p(D, \Theta|M) d\Theta \\ &= \int \Pr(D|M, \Theta) \cdot p(\Theta|M) d\Theta \\ &\approx \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{Pr(D|\Theta_j, M)}\right)^{-1}, \end{aligned}$$

where $\Theta_1, \ldots, \Theta_m$ are the samples from an MCMC run.

Why harmonic mean estimator for $\Pr(D)$?

Let $\theta_1, \ldots, \theta_m$ be (approx.) independent samples according to $p(\theta|D)$. Then,

$$\begin{aligned} 1 &= \int p(\theta)d\theta \approx \frac{1}{m}\sum_{i=1}^{m}\frac{p(\theta_i)}{p(\theta_i|D)} && \text{(importance sampling)} \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{p(\theta_i)}{\frac{\Pr(D|\theta_i)\cdot p(\theta_i)}{\Pr(D)}} && \text{(Bayes formula)} \\ &= \Pr(D)\cdot\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\Pr(D|\theta_i)}. \end{aligned}$$

$\Rightarrow$

$$\Pr(D) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\Pr(D|\theta_i)}}$$

Advantages of Bayes factors:

- can also support the restricted model while tests can only support the general model by statistically rejecting the restricted one.
- can also compare non-nested models

Problems:

- Prior has influence even for large amount of data
- harmonic mean estimator can have infinite variance (more sophisticated methods exist)
- Tests and Bayesian model selection can lead to opposite results (Lindley's paradox).

Study accuracy of different possibilities to apply IMa if there is recombination within loci (which is usually the case):

References

[1] J.L. Strasburg, L.H. Rieseberg (2010) How robust are "isolation with migration" analyses to violations of the im model? A simulation study. *Mol. Biol. Evol.* **27(2)**:297-310.

Note that meanwhile LAMARC also works with demographic models of populations splits.

Some of the things you should be able to explain

- Model assumptions and time scaling in IM and IMa
- IM/IMa assumptions for paramter variations among loci
- likelihood-ratio testing vs. Bayes factors
- harmonic mean estimator to compute marginal model probability from MCMC samples

5 Approximate Bayesian Computation (ABC)

Problems of full-data methods:

- usual runtime for one dataset: several weeks or months
- complex software, development takes years
- most programs not flexible, hard to write extensions

References

[PSPL+99] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun and M. W. Feldman (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16(12)**:1791–1798

[BZB02] M.A. Beaumont, W. Zhang, D.J. Balding (2002) Approximate Bayesian Computation in Population Genetics. *Genetics* **162**:2025–2035

[MMPT03] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100**:15324–15328

[WCE09] D. Wegmann, C. Leuenberger, L. Excoffier (2009) Efficient approximate Bayesian computation coupled Markov chain Monte Carlo without likelihood. *Genetics* **182**:1207

Pritchard et al. (1999)

- Compute MRCA of human Y chromosome in population models with growth.
- Find strong signal of population expansion in all populations.
- Explanations: recent expansion from a small ancestral population in the last 120,000 years or natural selection on the Y chromosome.
- data: 8 microsatellite loci from 445 humans
- Try various microsatellite mutation models
- Use summary statistics:
  1. mean accross loci in the variance of repeat numbers
  2. mean effective heterozygosity
  3. number of distinct haplotypes

Pritchard et al. (1999)

Approximate Bayesian Computation

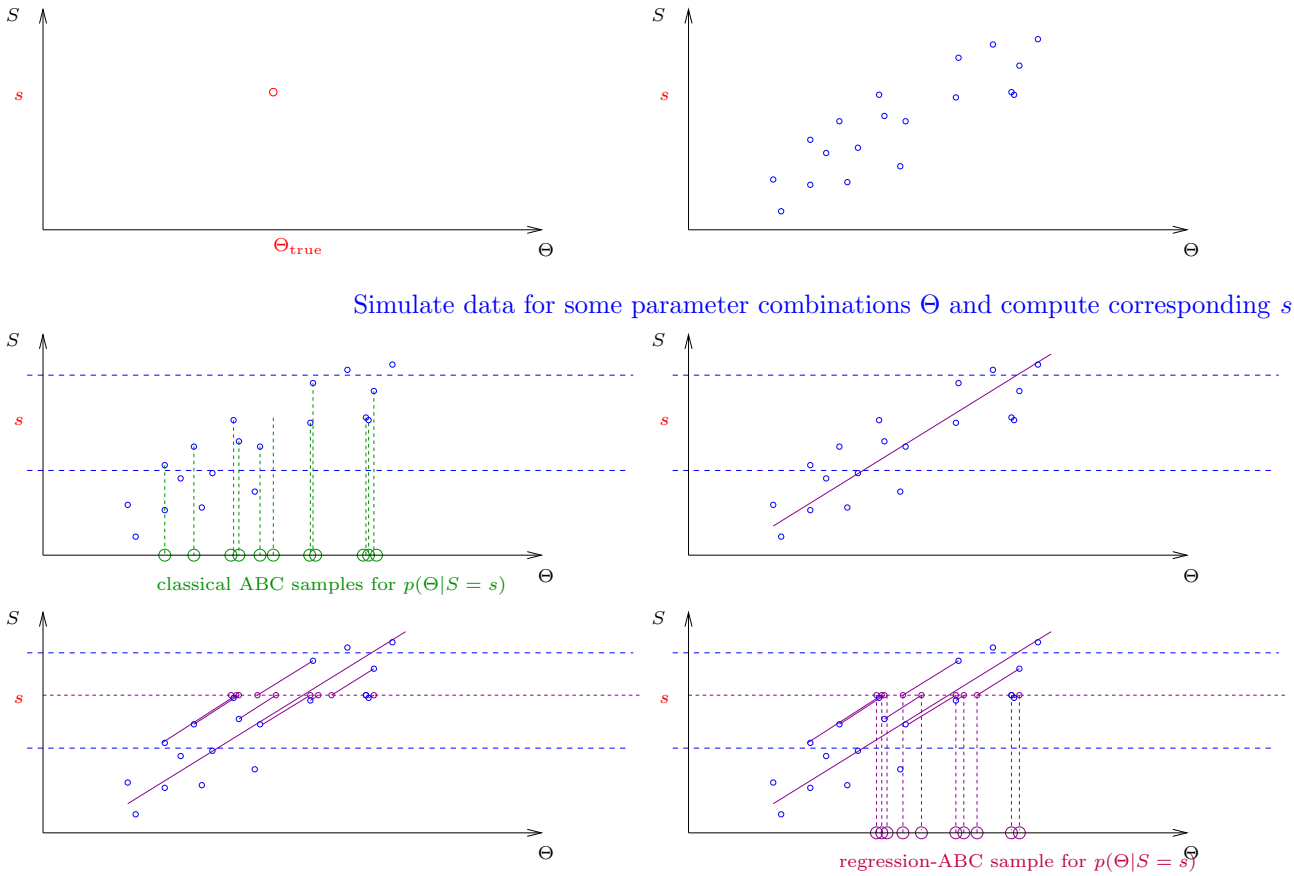
1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ
3. repeat until k accepted parameter combinations Θ':
   (a) Simulate Θ' from prior distribution of Θ
   (b) Simulate genealogy G according to $\Pr_{\Theta'}(G)$.
   (c) Simulate data and compute values s' of S
   (d) accept Θ' if $\|s - s'\| \leq \delta$

Only possible if a few summary statistics suffice. Otherwise acceptance will be rare.

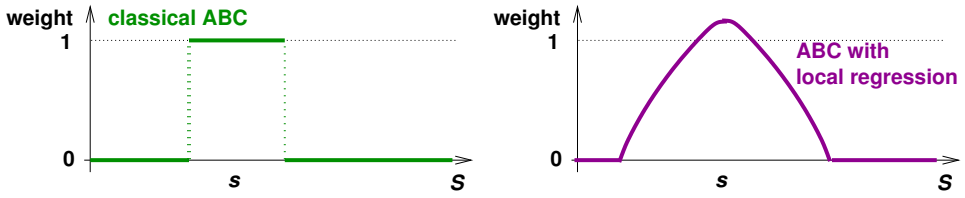
5.1 ABC with local regression correction

Ideas of Beaumont, Zhang, Balding (2002):

- combine ABC with local regression:



- Accept in a wider range but put a smaller weight on s' if $|s - s'|$ is large.

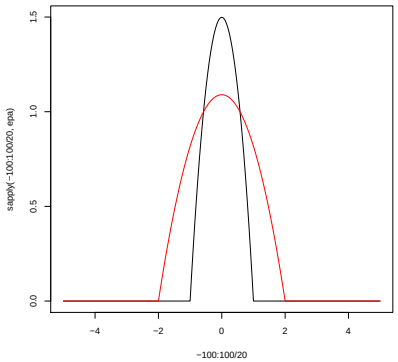


Epanechnikov-Kernel

$$K_\delta(t) = \begin{cases} \frac{3}{4\cdot\delta}\cdot\left(1-\left(\frac{t}{\delta}\right)^2\right) & \text{for} \quad |t| \le \delta \\ 0 & \text{for} \quad |t| > \delta \end{cases}$$

Note that the area under the kernel is 1, as

$$\int_{-\delta}^{\delta} \frac{3}{4\cdot\delta}\cdot\left(1-\left(\frac{x}{\delta}\right)^2\right)\,dx \;=\; 1$$



Epanechnikov-Kernels with

$\delta = 1$ and δ=2

Beaumont, Zhang, Balding (2002)

Simulate pairs $(\Theta^{(i)}, s^{(i)})$ and fit local regression model, i.e. find α and β to minimize

$$\sum_i \left(\Theta^{(i)} - \alpha - (s^{(i)} - s)^T\beta\right)^2 \cdot K_\delta(\|s^{(i)} - s\|),$$

where $||v|| = \sqrt{\sum_i v_i^2}$ (or some other vector norm).

Consider

$$\Theta_*^{(i)} = \Theta^{(i)} - (s^{(i)} - s)^T\widehat{\beta}$$

as random sample from $\Pr(\Theta \mid S = s)$.

Posterior density estimation:

$$\widehat{p}(\Theta_0 \mid S = s) = \frac{\sum_i K_\Delta(\Theta_*^{(i)} - \Theta_0)\cdot K_\delta(\|s - s^{(i)}\|)}{\sum_j K_\delta(\|s - s^{(j)}\|)}$$

where Δ = density estimation bandwidth.

Reminder/Introduction to solution for normal regression

Problem: predict Y from X_1, X_2,...,X_m. Observations:

$$\begin{array}{ccc} Y_1 & , & X_{11}, X_{12}, \ldots, X_{1m} \\ Y_2 & , & X_{21}, X_{22}, \ldots, X_{2m} \\ \vdots & & \vdots \\ Y_n & , & X_{n1}, X_{n1}, \ldots, X_{nm} \end{array}$$

with $$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ni} \end{pmatrix}$$

Model: $Y = a + b_1\cdot X_1 + b_2\cdot X_2 + \cdots + b_m\cdot X_m + \varepsilon$ Equation system for a, b_1, b_2, ..., b_m:

$$\begin{array}{ccccccccccccc} Y_1 & = & a & + & b_1\cdot X_{11} & + & b_2\cdot X_{12} & + & \ldots & + & b_m\cdot X_{1m} & + & \varepsilon_1 \\ Y_2 & = & a & + & b_1\cdot X_{21} & + & b_2\cdot X_{22} & + & \ldots & + & b_m\cdot X_{2m} & + & \varepsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ Y_n & = & a & + & b_1\cdot X_{n1} & + & b_n\cdot X_{n2} & + & \ldots & + & b_m\cdot X_{nm} & + & \varepsilon_n \end{array}$$

Model:

$$\begin{array}{ccccccccccccc} Y_1 & = & a & + & b_1 \cdot X_{11} & + & b_2 \cdot X_{12} & + & \dots & + & b_m \cdot X_{1m} & + & \varepsilon_1 \\ Y_2 & = & a & + & b_1 \cdot X_{21} & + & b_2 \cdot X_{22} & + & \dots & + & b_m \cdot X_{2m} & + & \varepsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ Y_n & = & a & + & b_1 \cdot X_{n1} & + & b_n \cdot X_{n2} & + & \dots & + & b_m \cdot X_{nm} & + & \varepsilon_n \end{array}$$

target variables Y Explanatory variables $X_1, X_2, \dots, X_m$ Parameters to be estimated $a, b_1, \dots, b_m$ Independent normally distributed errors $\varepsilon_1, \dots, \varepsilon_m$ with unknown variance σ^2.

Approach to estimate a and b_i:

minimize residual sum of squares: $f(\beta) := \sum_{i=1}^{n} \left(Y_i - a - \sum_{j=1}^{m} b_j X_{ij} \right)^2$

For this be

$$\beta = \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{und} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nm} \end{pmatrix}$$

and thus $f(\beta) = \langle y - \mathbf{X}\beta, y - \mathbf{X}\beta \rangle = \|\mathbf{y} - \mathbf{X}\beta\|^2$.

This means: find $\widehat{\beta}$, such that $\widehat{y} = \mathbf{X}\widehat{\beta}$ has minimal euclidian distance to y.

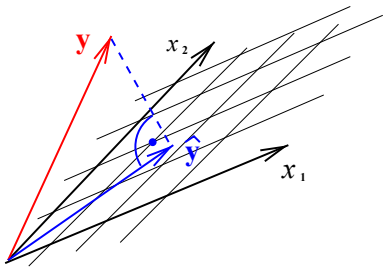
Geometric solution: minimizing $f(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$ means that $\widehat{y} = \mathbf{X}\widehat{\beta}$ must be the projection of $\mathbf{y}$ to the space spanned by the vectors $x_0, x_1, \dots, x_N$ (with $x_0 = (1, \dots, 1)^T$). Therefore, $\mathbf{y} - \mathbf{X}\beta$ must be in a right angle on each x_i,

This implies

$$\forall i \ : \ \langle \mathbf{y} - \mathbf{X}\widehat{\beta}, x_i \rangle = 0,$$

and thus

$$(\mathbf{y} - \mathbf{X}\widehat{\beta})^T \mathbf{X} = (0, \dots, 0),$$



from which we obtain the solution $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Note that the vectors β and $\widehat{\beta}$ above contain also the intercept; the notation changes below.

Solution of the local regression problem

Solution for j-th parameter: $(\widehat{\alpha}, \widehat{\beta_1}, \dots, \widehat{\beta_k}) = \left(X^T W X\right)^{-1} X^T W \Theta^{(j)}$, where

$\Theta^{(j)} = \begin{pmatrix} \Theta_1^{(j)} \\ \Theta_2^{(j)} \\ \vdots \\ \Theta_m^{(j)} \end{pmatrix}$: Values of the j-th parameter from m simulations,

$s = (s^{(1)}, \dots, s^{(k)})$: Vector of summary statistics for observed data,

$s_i = (s_i^{(1)}, \dots, s_i^{(k)})$: Vector of summary statistics from i-th simulation,

$X = \begin{pmatrix} 1 & s_1^{(1)} - s^{(1)} & \cdots & s_1^{(k)} - s^{(k)} \\ 1 & s_2^{(1)} - s^{(1)} & \cdots & s_2^{(k)} - s^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_m^{(1)} - s^{(1)} & \cdots & s_m^{(k)} - s^{(k)} \end{pmatrix}$ and

W is diagonal matrix with diagonal entries $K_\delta(||s_1 - s||), \dots, K_\delta(||s_m - s||)$.

Beaumont, Zhang, Balding (2002)

ABC with local regression

1. Select summary statistics $S = (S_i)_i$ and compute their values $s = (s_i)_i$ for given data set
2. Choose tolerance δ and bandwidth Δ
3. repeat for $i = 1, \ldots, m$:
   (a) Simulate $\Theta^{(i)}$ from prior distribution of Θ
   (b) Simulate genealogy G according to $\Pr_{\Theta^{(i)}}(G)$.
   (c) Simulate data and compute values $s^{(i)}$ of S
4. $(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{m} \left(\Theta_i - \alpha - (s^i - s)^T \beta\right)^2 \cdot K_\delta(||s^i - s||)$
5. $\Theta_*^{(i)} := \Theta^{(i)} - (s^{(i)} - s)^T \widehat{\beta}$
6. Approximate $p(\Theta | S = s)$ by $\frac{\sum_i K_\Delta(\Theta_*^{(i)} - \Theta) \cdot K_\delta(\|s - s^{(i)}\|)}{\sum_j K_\delta(\|s - s^{(j)}\|)}$

Summary statistics used by Beaumont et al. (2002) for microsatellite data:

1. mean accross loci in the variance of repeat numbers
2. mean effective heterozygosity
3. number of distinct haplotypes
4. mean accross loci of kurtosis of repeat numbers
5. variance accross loci of variance of repeat numbers
6. mean accross loci of maximum allele-frequency
7. multivariate kurtosis
8. linkage disequilibrium (LD) measured with Hudson's Δ^2

5.2 MCMC without likelihoods

Marjoram et al. (2003) MCMC without likelihoods

Aim: For given data D with summary statistics $S = s$ sample paramter vectors according to $p(\Theta \mid ||S - s|| \leq \varepsilon)$.

1. If current parameter estimation is Θ', propose Θ^* with probability $Q_{\Theta' \to \Theta^*}$
2. Simulate data D^* according to Θ^* and compute their summary statistics s^*.
3. If $||s^* - s|| > \varepsilon$ reject proposal, else accept with probability
$$\min\left\{1, \frac{p(\Theta^*) \cdot Q_{\Theta^* \to \Theta'}}{p(\Theta') \cdot Q_{\Theta' \to \Theta^*}}\right\}.$$
4. repeat steps 1 to 4.

Application example: Nuu Chah Nulth data, n=63 samples of HVR-I.

Estimate θ and time to the MRCA based on F84 substitution model.

Summary statistics: number of variable sites and number of haplotypes.

Simple approach: when updating parameters, generate entirely new tree. (will usually be rejected $\rightsquigarrow$ inefficient.)

Compromise: keep some information about the tree an modify it slightly for next step:

1. tree topology
2. times of coalescence events
3. number of mutations between two coalescents events

Some of the things you should be able to explain

- Basic rejection ABC
- improvements in local-regression ABC
  - use of Epanechnikov kernel in local regression
  - local regression correction
  - use of Epanechnikov smoothing kernel
  - why this leads to higher accuracy in shorter runtime
- how to combine ABC with Metropolis-Hastings ("MCMC without likelihoods")

5.3 Sequential / Adaptive ABC

Sequential ABC

Basic idea: Run several iterations of ABC, always using the results from the previous run (posterior distribution of parameters) as priors for the new run.

Problem: priors are not allowed to depend on the data. Using posteriors as priors for the analysis of the same data is cheating!

Solution: Make some correction like in importance sampling to make sure that the initial prior is used for the final posterior.

Sequential/Adaptive ABC

References

[1] S.A. Sisson, Y. Fan, M.M. Tanaka (2007) Sequential Monte Carlo without likelihoods *PNAS* **104**: 1760–1765

[2] M.A. Beaumont, J.-M. Cornuet, J.-M. Marin, C.P. Robert (2009) Adaptive approximate Bayesian Computation *Biometrika* **96**: 983–990

[3] S.A. Sisson, Y. Fan, M.M. Tanaka (2009) Correction for Sisson *et al.*, Sequential Monte Carlo without likelihoods *PNAS* **106**

[4] J.-M. Marin, P. Pudlo, C.P. Robert, R.J. Ryder (2012) Approximate Bayesian computational methods *Statistics and Computing* **22**: 1167–1180

Sequential/Adaptive ABC (ABC-PMC)

Proposed by Beaumont *et al.* (2009); PMC="Population Monte Carlo"

Notations for the description of the algorithm:

$\theta_i^{(t)}$ i-the sampled parameter in iteration t (for simplicity assumed one-dimensional in following pseudocode, but can be parameter vector in general.)

S vector of summary statistics

s_o vector of summary statistics for original data

φ densitiy of standard normal distribution $\mathcal{N}(0,1)$ (can also be multivariate).

$p(\theta)$ prior probability density

$p(s|\theta)$ probability density of s given θ

$\delta_1 > \delta_2 > \cdots > \delta_T$ decreasing thresholds

In most applications θ is a vector (a multi-dimensional parameter). Then in the following pseudo-code τ is the matrix such that $\tau^2/2$ (or actually $\tau \cdot T(\tau)/2$, where T is the transpose operation) is a variance matrix (also called covariance matrix).

References

[1] Filippi, S., Barnes, C.P., Cornebise, J. and Stumpf, M.P (2013) On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo *Statistical applications in genetics and molecular biology* **12(1)**: 87–107

ABC-PMC

for $i = 1, \ldots, N$ **do**
  repeat
    Draw $\theta_i^{(1)}$ from prior and simulate $s \sim p(S|\theta_i^{(1)})$
  until $||s, s_0|| < \delta_1$
  $\omega_i^{(1)} := 1/N$
end for
choose matrix τ_1 such that $\tau_1^2/2 =$ empirical variance matrix of the $\theta_i^{(1)}$
for $t = 2, \ldots, T$ **do**
  for $i = 1, \ldots, N$ **do**
    repeat
      Draw θ_i^* from $(\theta_1^{(i-1)}, \ldots, \theta_N^{(i-1)})$ with probability distribution $(\omega_1^{(i-1)}, \ldots, \omega_N^{(i-1)})$
      Draw $\theta_i^{(t)}$ from $\mathcal{N}(\theta_i^*, \tau_{t-1}^2)$ and simulate $s \sim p(S|\theta_i^{(t)})$
    until $||s, s_0|| < \delta_t$
    $\omega_i^{(t)} :\propto p(\theta_i^{(t)}) \Big/ \sum_j \omega_j^{(t-1)} \cdot \varphi\left(\tau_{t-1}^{-1} \cdot \left(\theta_i^{(t)} - \theta_j^{(t-1)}\right)\right)$
  end for
  choose matrix τ_t such that $\tau_t^2/2 =$ empirical variance matrix of the $\theta_i^{(t)}$
end for

($:\propto$ means "set proportional to", such that $\sum_i \omega_i^{(t)} = 1$.)

5.4 Optimizing sets of summary statistics with PLS

Beaumont, Zhang, Balding (2002)

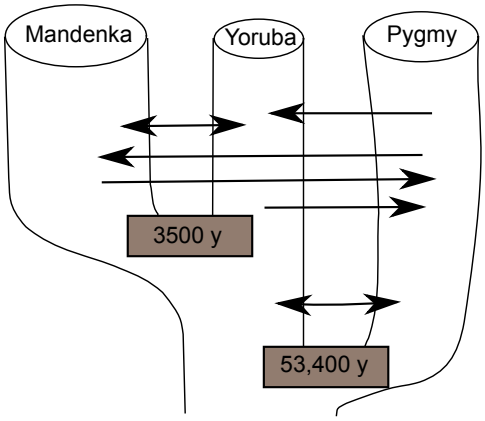
"[...] the MCMC-based method is consistently superior to the summary-statistics-based methods and highlights that it is well worth making the effort to obtain full-data inferences if possible."

[Note that "MCMC-based method" here refers to full-data methods]

"[...] there are advantages to the use of summary statistics, both in the ease of implementation and in the time to obtain the results [...]"

"Further research is needed to find a more rigorous way for choosing summary statistics, including the use of orthogonalization and 'projection-pursuit' methods"

Wegmann et al. (2009)



ABC estimations with microsatellite data.

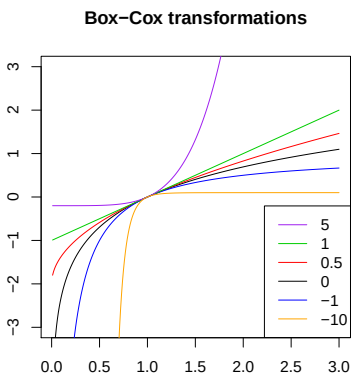
Wegmann et al. (2009)

- combine MCMC-ABC with Beaumont et al.'s regression approach to sample from $p(\Theta|||S-s|| \leq \varepsilon)$.
- apply Box-Cox transformation to each summary statistic with respect to the parameter of interest, based on simulated data
- apply partial least squares (PLS) to find combinations of summary statistics that are informative wrt the parameter of interest
- leave-one-out cross validation to optimize number of PLS components used

Simulation studies show improvements compared to other ABC methods but IMa is still better.

Wegmann et al. "[..] would not recommend using an ABC approach if a full-likelihood method exists [..]".

Box-Cox transformation



$$X^{(\lambda)} = \begin{cases} \frac{(X+c)^{\lambda}-1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(X+c) & \text{for } \lambda = 0 \end{cases}$$

Idea: fit λ and c such that the residuals of the regression model $Y = \alpha + \beta X$ look as normally distributed as possible.

Comparison PCA vs. PLS

Let S be the covariance matrix of the vectors $x_1, \ldots, x_n$ (with $x_i = (x_{i1}, \ldots, x_{im})$) that are normalized, that is $\mu_{x_i} = 0$ and $\sigma_{x_i} = 1$. Then, the principal component directions $v_1, \ldots, v_m$ satisfy:

$$v_j = \arg\max_{\alpha} \left\{ \mathrm{Var}\left(\sum_i x_i \alpha_i\right) \;\middle|\; ||\alpha|| = 1, \forall_{\ell<j}\ v_\ell^T S\alpha = 0 \right\}$$

The PLS directions $\varphi_1, \ldots, \varphi_m$ satisfy:

$$\begin{aligned} \varphi_j &= \arg\max_{\alpha} \left\{ \mathrm{Cor}^2\left(y, \sum_i x_i\alpha_i\right) \mathrm{Var}\left(\sum_i x_i\alpha_i\right) \;\middle|\; ||\alpha|| = 1, \forall_{\ell<j} :\ \varphi_\ell^T S\alpha = 0 \right\} \\ &= \arg\max_{\alpha} \left\{ \mathrm{Cov}^2\left(y, \sum_i x_i\alpha_i\right) \;\middle|\; ||\alpha|| = 1, \forall_{\ell<j} :\ \varphi_\ell^T S\alpha = 0 \right\} \end{aligned}$$

Note that the condition $v_\ell^T S\alpha = 0$ just means that the new vector $\sum_j \alpha_j \cdot x_j$ ist orthogonal on the previous ones $\sum_k v_{\ell,k} x_k$ (for any $\ell < j$).

To see this, note that from $\mu_{x_k} = 0 = \mu_{x_j}$ follows

$$S_{(k,j)} = \mathrm{Cov}(x_k, x_j) = \frac{1}{m-1} \sum_i (x_{ki} - \mu_{x_k}) \cdot (x_{ji} - \mu_{x_j}) = \frac{\sum_i x_{ki} x_{ji}}{m-1}$$

and thus

$$v_\ell^T S\alpha = \sum_{k,j} v_{\ell,k} \frac{\sum_i x_{ki} x_{ji}}{m-1} \cdot \alpha_j = \frac{1}{m-1} \left\langle \sum_k v_{\ell,k} x_k \ ,\ \sum_k \alpha_j x_j \right\rangle.$$

(Remember that the scalar product $\langle v, w\rangle = \sum_i v_i w_i$ of two vectors v and w has the geometric interpretation $\langle v, w\rangle = ||v|| \cdot ||w|| \cdot \cos(\gamma)$, where γ is the angle between the vectors. Thus, $\langle v, w\rangle = 0$ holds if and only if v and w are orthogonal on each other.)

The scalar product will also be useful on the next slide, on which the algorithm to compute PLS is shown.

The slope of a regression line with response variable y and explanatory variable x (both of length m) can be expressed as

$$b = \mathrm{Cov}(x, y)/\sigma_x^2$$

and the intercept is $a = \mu_y - b \cdot \mu_x$.

If y is centered and x is normalized such that $\mu_x = \mu_y = 0$ and $\sigma_x = 1$, we obtain the regression line

$$\begin{aligned} y &= a + bx \ = \ 0 + \frac{\mathrm{Cov}(x,y)}{1} x \ = \ \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{m-1} x \\ &= \frac{\sum x_i y_i}{m-1} x \ = \ \frac{\langle x, y\rangle}{m-1} x. \end{aligned}$$

partial least squares (PLS)

Aim: find combinations of explanatory variables $x_1, \ldots, x_m$ that have highest covariance with variable y.

let y be centered and x_j be normalized, i.e. $\mu_y = 0$, $\mu_{x_j} = 0$, $\sigma_{x_j} = 1$.

1. ($(m-1)$-fold of) univariate regression coefficient: $\varphi_j := \langle x_j, y\rangle := \sum_i x_{ji} y_i \qquad \Rightarrow y \approx \frac{1}{m-1} \cdot \varphi_j \cdot x_j$
2. first partial least squares direction: $z_1 := \sum_j \varphi_j \cdot x_j$

3. first regression coefficient: $\delta := \frac{\langle z_1, y\rangle}{\langle z_1, z_1\rangle} \qquad \Rightarrow y \approx \delta \cdot z_1$

4. now orthogonalize $x_1, x_2, \ldots, x_m$ with respect to z_1: $x_j^{(2)} := x_j - \frac{\langle z_1, x_j\rangle}{\langle z_1, z_1\rangle} \cdot z_1$

5. and compute the residuals: $y^{(2)} := y - \delta \cdot z_1$

repeat 1-5 with x_j and y replaced by $x_j^{(2)}$ and $y^{(2)}$. $\rightsquigarrow z_2, x_j^{(3)}, y^{(3)}$

iterate to get $z_1, z_2, \ldots, z_m$.

PLS for multiple response variables

Wegmann *et al.* (2009) PLS for multiple response variables (here: summary statistics), implemented in the command `plsr` in the R package `pls`.

Several possible generalizations of PLS exist for multiple response variables $y_1, \ldots, y_q$, e.g. SIMPLS: For all $i = 1, \ldots, m$ let φ_i be the vector α, for which $z_i := x_1\alpha_1 + \cdots + x_n\alpha_n$ maximizes

$$\sum_{j=1}^{q} \mathrm{Cov}^2(z_i, y_j)$$

subject to the conditions that $||\alpha|| = 1$ and that $\forall_{k<i} : \langle z_i, z_k\rangle = 0$.

References

[BS06] A.-L. Boulesteix, K. Strimmer (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data *Briefings in Bioinformatics* **8.1**: 32–44

Some of the things you should be able to explain

- Why can't we just iterate ABC on a data set, always using posteriors from previous run as prior?
- How this is corrected in ABC-PMC
- What is PLS and how can it be applied to improve sets of summary statistics, e.g. for ABC?

6 Jaatha and dadi

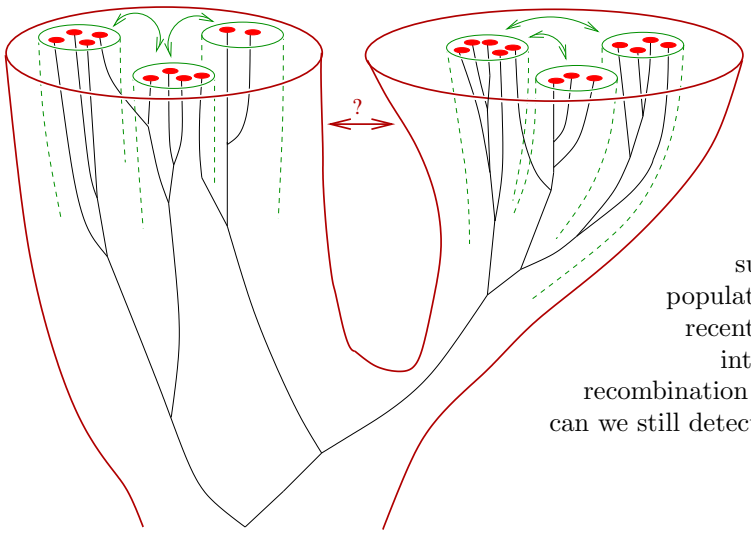
6.1 Wild Tomatoes and Jaatha 1.0

Tomato Data

Solanum peruvianum, Canta, Peru

Solanum chilense, Moquegua, Peru

Complex Demography

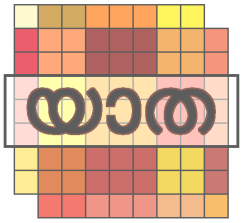


substructure
population growth
recent speciation
introgression?
recombination within loci
can we still detect selection?

Jaatha

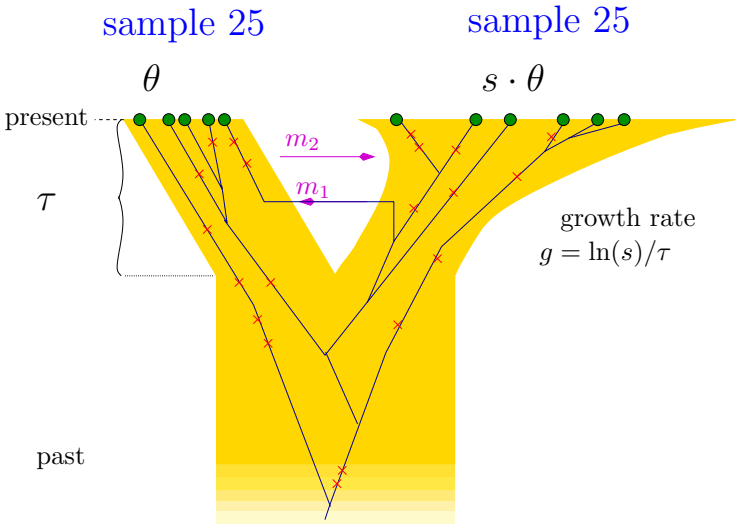
JSFS associated approximation of the ancestry

Malayalam word for “past”.

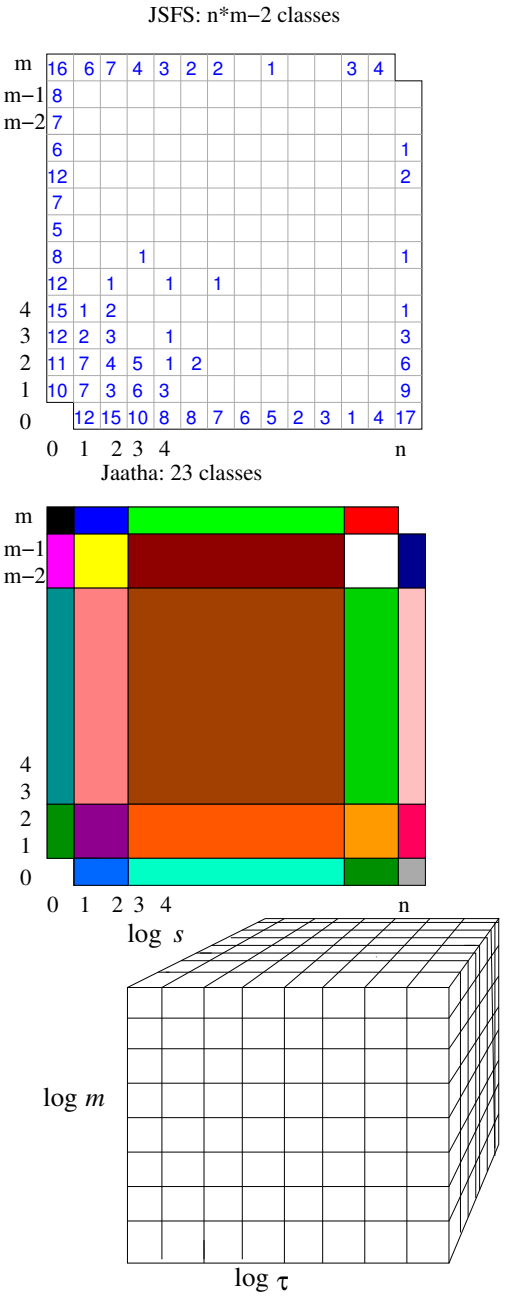


Strategy: Compare data to data that has been simulated with various combinations of parameter values.

Demographic Model



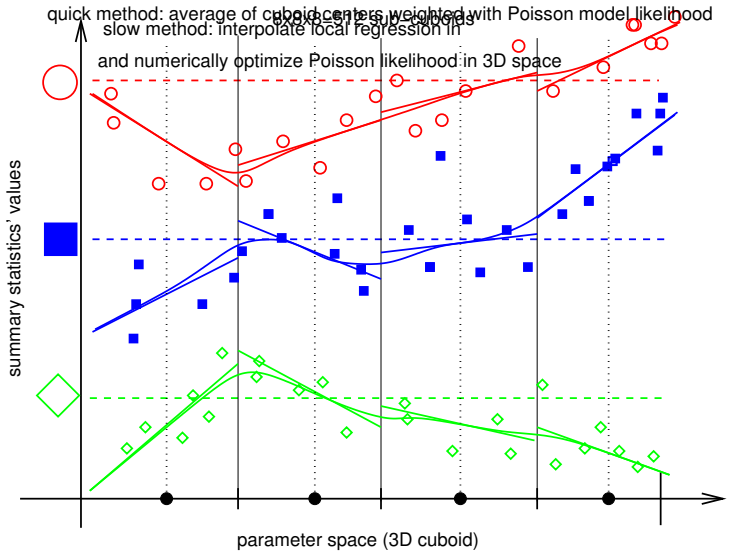
Summarizing the JSFS



JSFS =
Joint Site Frequency
Spectrum *Comparison of summaries:*
A. Tellier, P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. Rose, T. Städler, W. Stephan, D. Metzler (2011) Estimating parameters of speciation models based on refined summaries of the joint site frequency spectrum. *PLoS ONE* 6(5): e18155.
doi:10.1371/journal.pone.0018155

How do the 23 summary statistics depend on the parameters?
Linearize on each of the 8x8x8 cuboids.

Simple methods in continuous parameter space



How to get from $\mathbb{E}$ to Likelihoods?

Actually not likelihoods but **composite** likelihoods, that is: approximations that neglect certain dependencies.

Composite Likelihood Approach $\Rightarrow$ 23 summary statistics are treated as independent and Poisson distributed. (As if all polymorphisms were independent of each other.)

This means, if $s_1, s_2, \ldots, s_{23}$ are the observed summary statistics and $\lambda_1, \lambda_2, \ldots, \lambda_{23}$ their expectations, the composite likelihood is

$$\frac{\lambda_1^{s_1} \cdot e^{-\lambda_1}}{s_1!} \cdot \frac{\lambda_2^{s_2} \cdot e^{-\lambda_2}}{s_2!} \cdots \frac{\lambda_{23}^{s_{23}} \cdot e^{-\lambda_{23}}}{s_{23}!}$$

Runtime (Jaatha version 1; limited to 4 parameters

Given model with 4 parameters and sample sizes for two populations, simulate data and fit local linear models. 3-5 days

Analyse dataset with quick method: 1-3 seconds

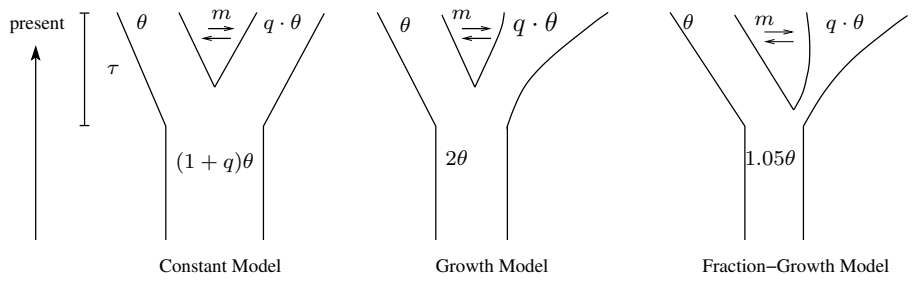
Analyse dataset with slow method: 15 minutes

Compromise "J-Med": <15 seconds

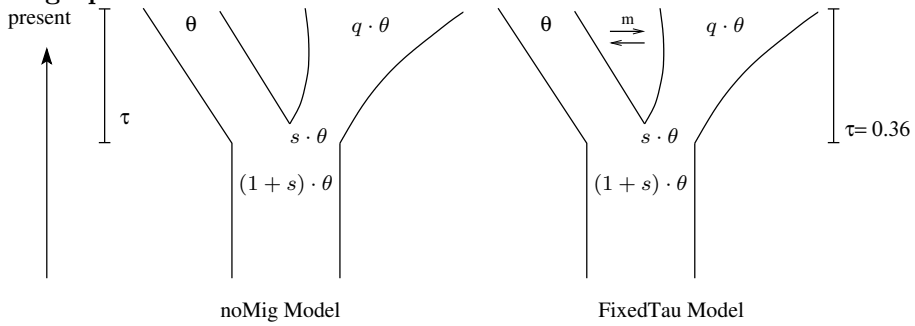
References

[1] L. Naduvilezhath, L. Rose, D. Metzler (2011) Jaatha: A Fast Composite Likelihood Approach to Estimate Demographic parameters. *Molecular Ecology*, **20(13)**: 2709–2723

Demographic Models



Demographic Models

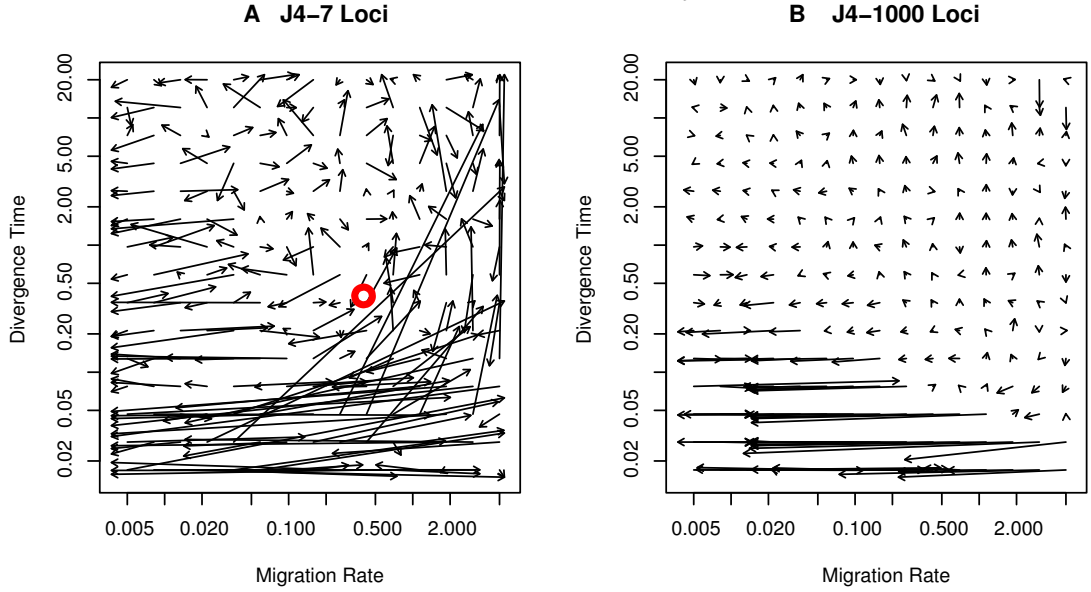


Tomato data: parameter estimations and bootstrap confidence intervals

7 loci, varying from 0.8 to 1.9 kb in size, sampled 23 individuals (i.e. 46 sequences) per species

Parameter	*Constant*	*Growth*	*Fraction-Growth*	*noMig*	*fixedTau*
$\widehat{\theta}_1$	9.41 (7.14-12.59)	10.30 (8.29-13.02)	12.56 (9.61-16.38)	13.34 (10.29-17.35)	12.22 (9.37-15.09)
$\widehat{q}$	1.83 (1.23-2.69)	4.24 (2.58-6.95)	4.29 (2.71-6.38)	8.67 (5.34-15.00)	4.94 (3.28-7.85)
$\widehat{m}$	0.36 (0.06-4.89)	0.36 (0.09-2.34)	0.73 (0.39-1.27)	*0*	0.55 (0.22-1.03)
$\widehat{\tau}$	0.41 (0.05-1.82)	0.37 (0.11-0.93)	0.79 (0.37-1.63)	0.14 (0.10-0.23)	*0.36*
$\widehat{s}$	$\widehat{q}$	*1*	*0.05*	0.44 (0.18-0.98)	0.33 (0.11-1.10)
log-likelihood	-189.51	-119.70	-101.58	-133.06	-93.96

Growth model: Tomato estimates vs. simulation study



parametric-bootstrap confidence intervals

- Simulate 200 datasets according to inferred parameter values
- Infer parameters for each simulated dataset
- For any parameter θ let $\widehat{\theta}$ be the original estimation and $\theta^*_{0.025}$ and $\theta^*_{0.975}$ the 0.025- and 0.975-quantiles of the estimates for the bootstrap datasets
- parametric-bootstrap confidence interval:

$$[2\widehat{\theta} - \theta^*_{0.975}, 2\widehat{\theta} - \theta^*_{0.025}]$$

- based on the idea that $\mathcal{L}(\theta - \widehat{\theta}) \approx \mathcal{L}(\widehat{\theta} - \theta^*)$, where θ^* is the estimate from some bootstrap simulation
- (some refinements exist, e.g. BC_a bootstrap intervals)

Simulation-based significance testing

- Simulated data according to NoMig model with ML parameter values
- Estimated parameters from simulated data with other models
- Only few ($\approx$ 5 out of 1000) estimated migration rates were as high as for tomato data
- log likelihood-ratios "Mig-NoMig" < 25 (most < 0) for simulated data, > 30 for tomato data with models "fraction-growth" and "growth"

∂**a**∂**i**

fastsimcoal2 and ∂a∂i

References

[1] Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., Sousa, V. C. (2021) fastsimcoal2: demographic inference under complex evolutionary scenarios Bioinformatics **37**:4882-4885.

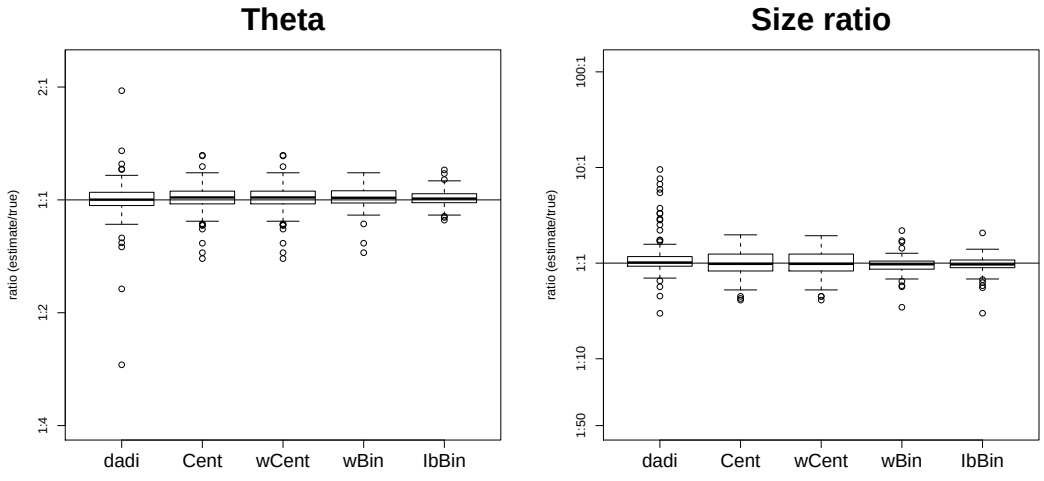
- Very fast simulator
- built-in jsfs based composite-likelihood method

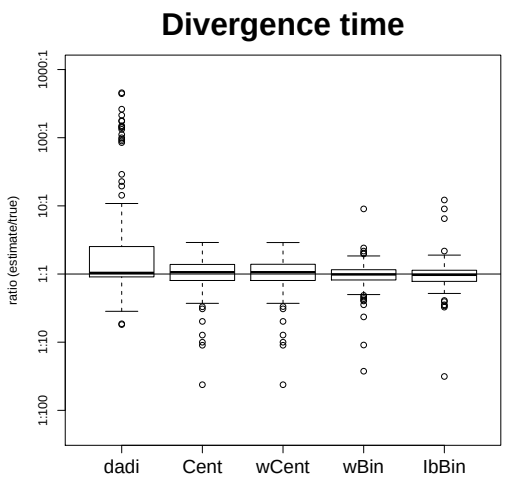
References

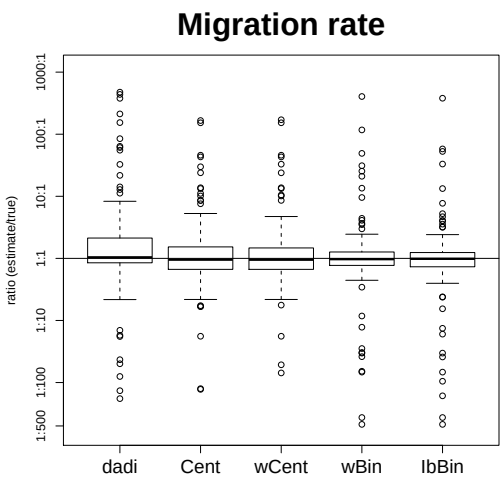
[1] R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, C.D. Bustamante (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data *PLoS Genetics*

- also a composite likelihood approach
- computes expected JSFS by diffusion approximation
- uses full JSFS

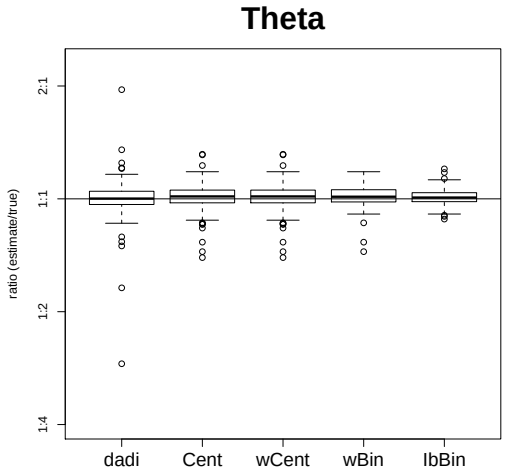
100 datasets with 100 loci

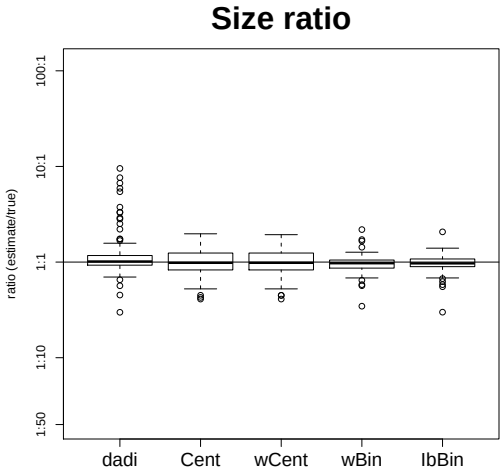


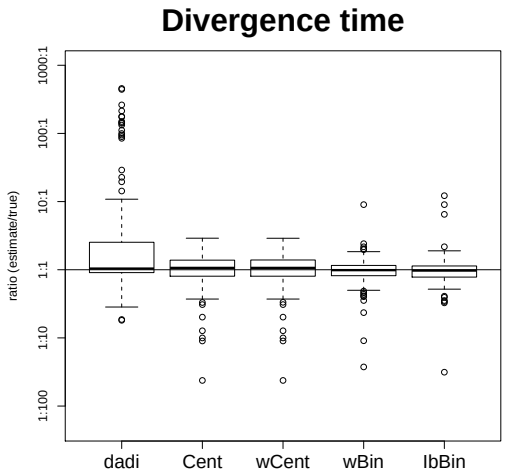


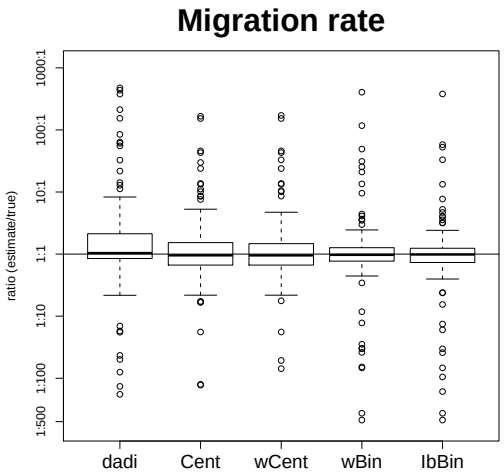


100 datasets with 7 loci





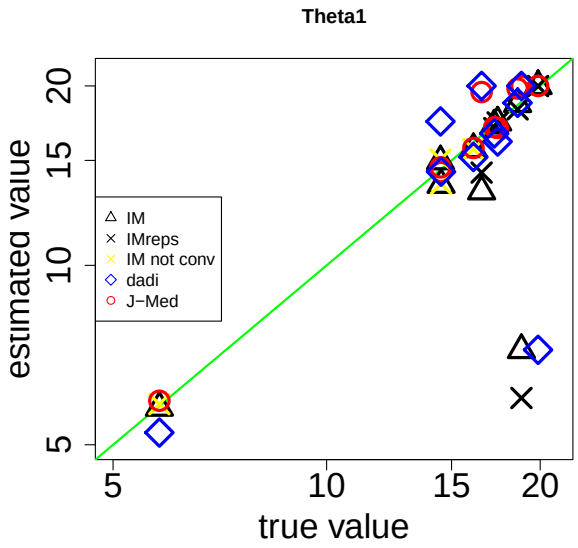


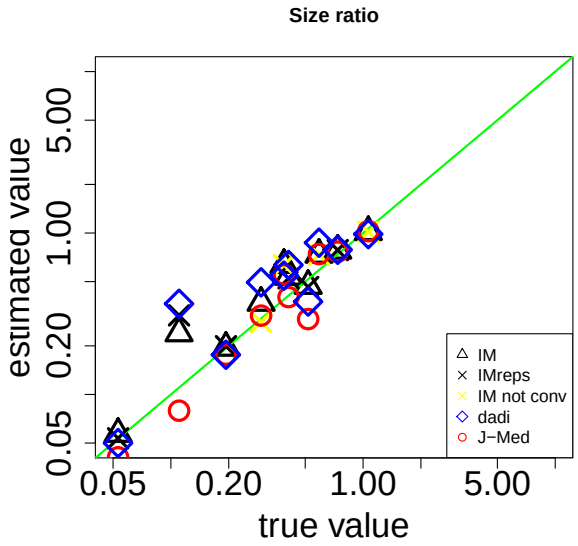


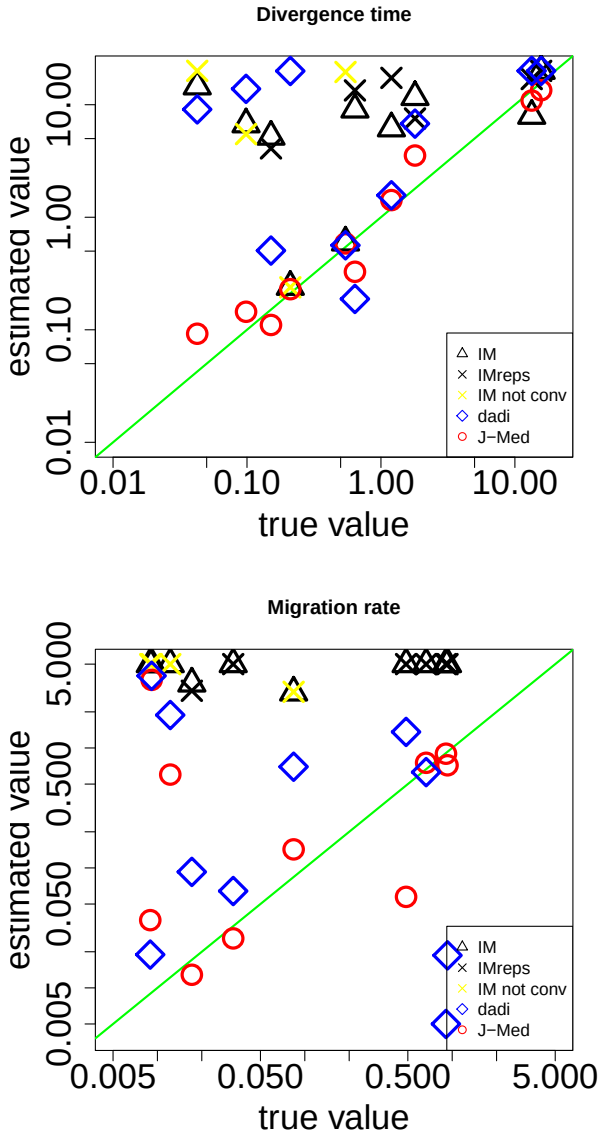
Jaatha vs. ∂a∂i vs. IM

Simulation Study

- Growth model, equally-sized founder populations
- 100 loci, no recombination within loci
- $\theta \in [5, 20]$ (per locus)
- size ratio $q \in [0.05, 20]$
- divergence time $\tau \in [0.01, 20]$
- migration rate $m \in [0.05, 5]$
- IM runs for 10 datasets, stopped after 5 weeks







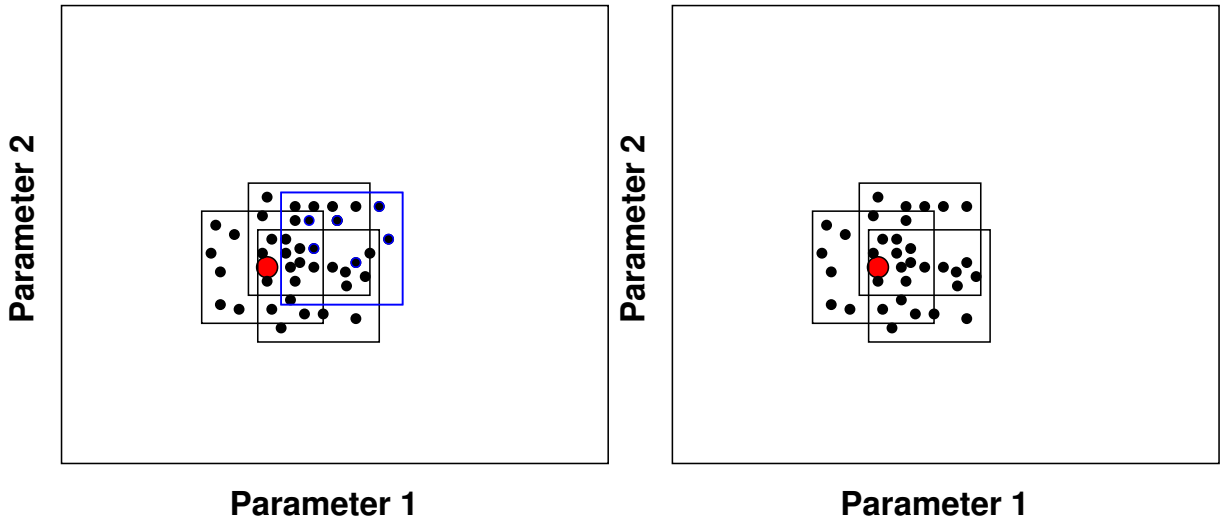
6.2 Jaatha 2.0

Jaatha 2.0

- R package
- Also for more than 4 parameters
- Training data are simulated when needed

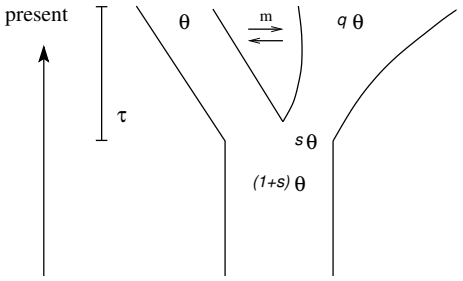
References

[1] Lisha A. Mathew, Paul R. Staab, Laura E. Rose, Dirk Metzler (2013) Why to account for finite sites in population genetic studies and how to do this with Jaatha 2.0. *Ecology and Evolution*



Simulations with 7 or 200 loci

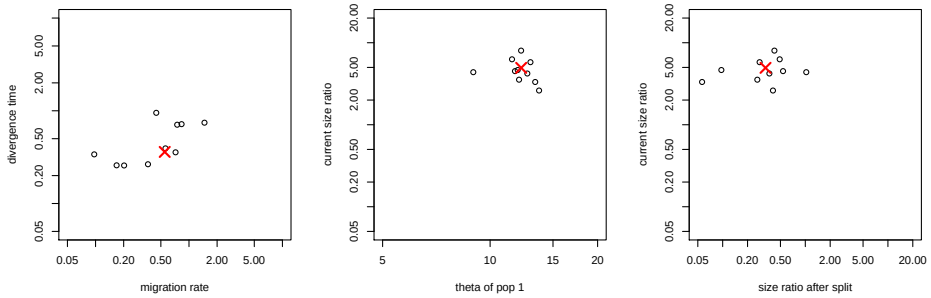
First experiment with infinite-sites model, demographic parameters inspired by tomato data:



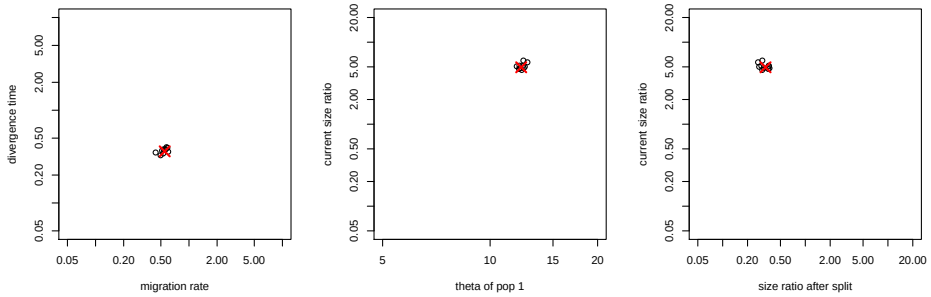
$\theta = 12.22$ (per locus)
$\tau = 0.36$
$m = 0.55$
$q = 4.94$
$s = 0.33$

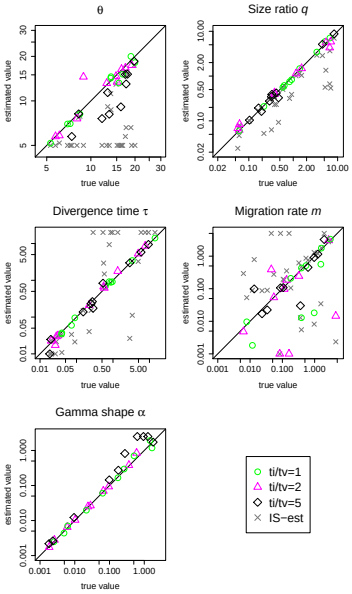
Recombination rate between 5 and 20 per locus; 25 sampled sequences per population and locus

Simulation Results with 7 loci

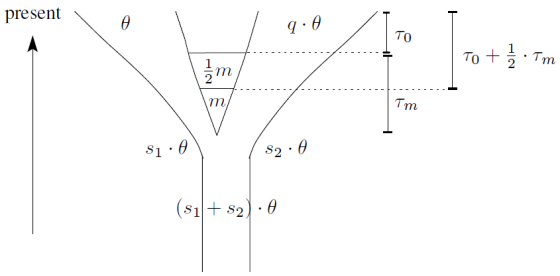


Simulation Results with 200 loci



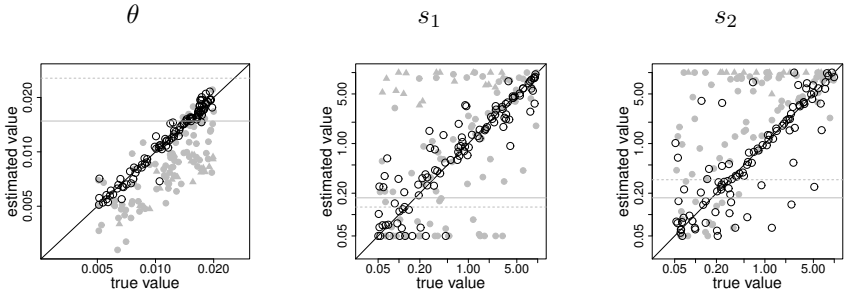


- 100 loci
- θ, q, τ, m, α estimated simultaneously
- ti/tv ratio fixed and assumed to be known
- 7 extra summary stats for double hits and separate counts of ti and tv within and between pops



7 demographic parameters, only 7 loci

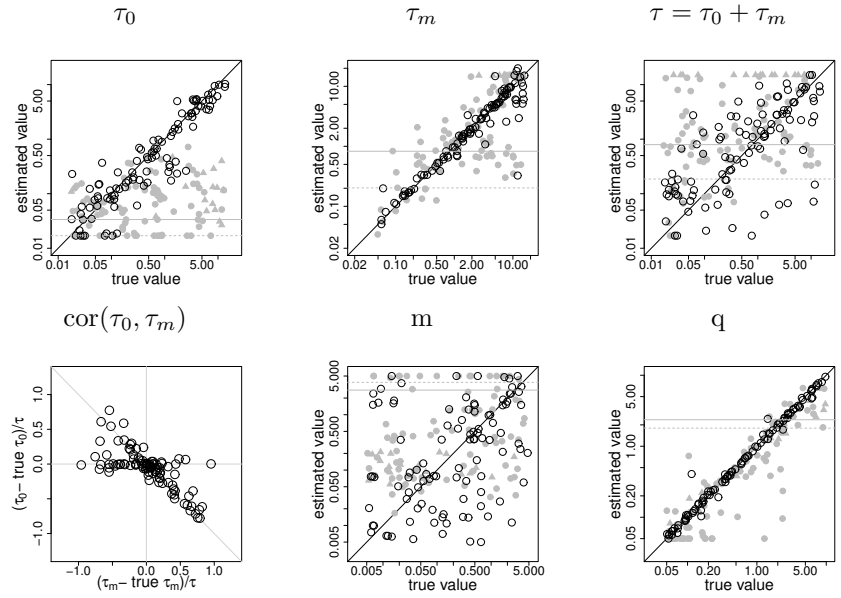
- τ_0 very small, suggesting ongoing gene flow
- estimation quite imprecise
- infinite-sites estimation substantially different from finite-sites estimations
- However, gene flow is significant (simulation-based composite-likelihood ratio test)



dashed line: infinite-sites

estimate grey line: finite-sites estimate grey dots: simulation with 7 loci cirles: simulation with 200 loci

triangles: τ_m estimated ≥ 15



6.3 Application to genome-wide data

NGS data of *Arabidopsis thaliana*

- 1.1 million SNPs (after filtering out ambiguous)
- 12 individuals from Spain, 12 from Italy, 5 from Novosibirsk (outgroup)

Model assumptions

- split of southern European populations, constant migration, constant sizes
- Finite-sites, estimate ti/tv
- separately for first or second codon position or UTR (FS), third codon position (Th), and non-coding (NC).

	τ	m	α	θ_{site}
complete data set	0.16	3.45	2.87	$3.54 \cdot 10^{-3}$
1st or 2nd codon pos or UTR	0.12	2.81	4.83	$2.73 \cdot 10^{-3}$
3rd codon pos	0.19	3.31	1.53	$3.70 \cdot 10^{-3}$
non-coding	0.18	3.33	2.26	$4.31 \cdot 10^{-3}$

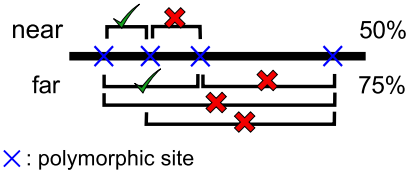
Parameter estimates for *A. thaliana* using FSM. Jaatha's estimates using the HKY model for the mutation rate θ, time τ of the split of both demes, the subsequent migration rate m between populations, and the rate heterogeneity parameter α. The parameter τ is scaled in $2N_e$ generations, m is twice the number of immigrants to each deme per generation, and θ is $2N_e$ times the mutation rate per base.

Significance of population structure: for 100 simulated panmictic populations τ was always estimated smaller.

6.4 Statistics in jaatha beyond jsfs

Summary statisics with linkage

Idea: Group loci by how often the 4-gamete condition is violated



near\far	0-33	33-66	66-100
0-33			
33-66			+1
66-100			

⇝ Novel summary statisitics based on spectrum of loci

6.5 Conclusions

Conclusions

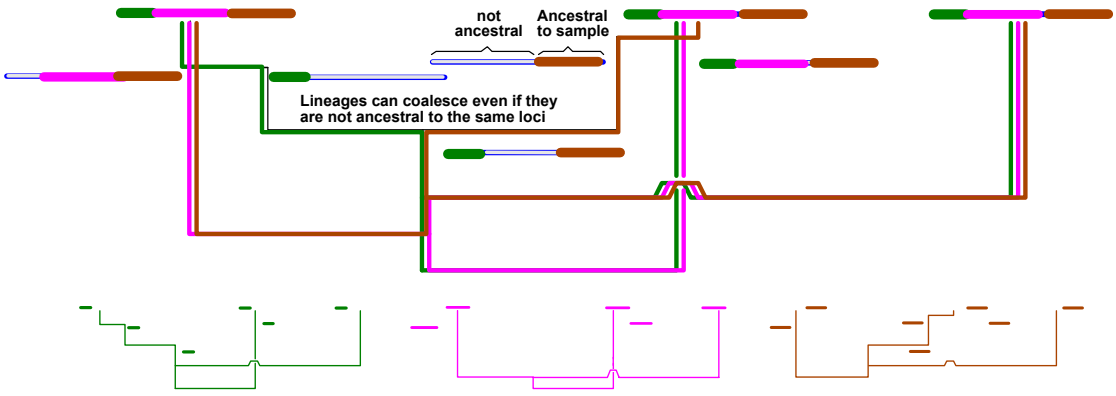
- intra-locus recombination difficult to handle rigorosly but allows for composite-likelihood approximations
- more loci needed for getting reasonable estimates
- small datasets require different methods and different sets of summary statistics than large datasets
- not always appropriate to use time-consuming methods for small datasets
- very large datasets can also be analysed with simple methods if not too many parameters to be estimated
- improving choice of summary statistics or smooth estimators for JSFS may be more important than numerics

Some of the things you should be able to explain

- JSFS
- Composite-likelihood approach and Poisson approximation
- Why is the JSFS further summarized in jaatha or ABC?
- bootstrap confidence interval and simulation-based p-values

7 ARG read from left to right and its approximations

7.1 Why ARG read from left to right is not Markovian



7.2 SMC, SMC': Markovian approximations of the ARG

References

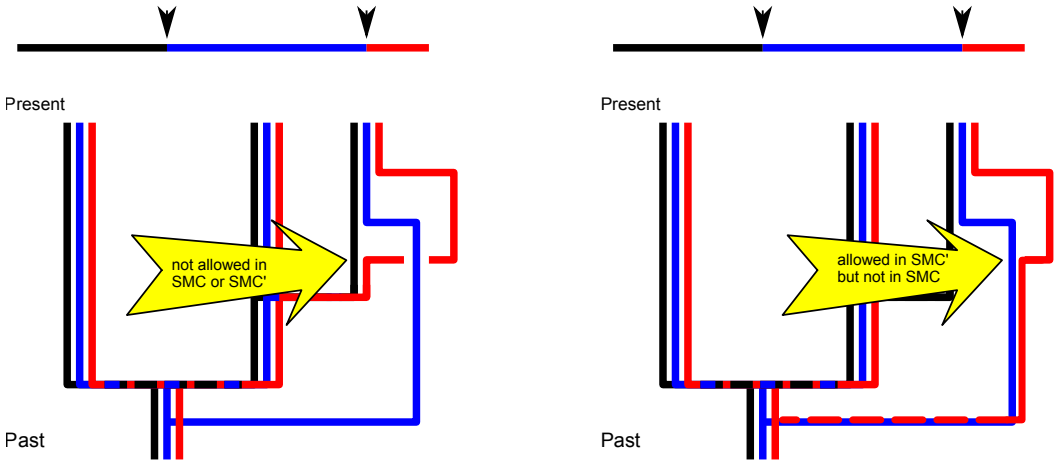
[WH99] Wiuf C, Hein J. (1999) Recombination as a point process along sequences. *Theor. Popul. Biol.* 55: 248–259

[MVC05] McVean, GA, Cardin, NJ (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* 360, 1459:1387–93. Introduce the **sequentially Markov coalescent** (SMC)

[MW06] Marjoram, P, Wall, JD (2006) Fast "coalescent" simulation. *BMC Genet* 7:16 Introduce SMC'

[EF11] Excoffier, L, Foll, M (2011) fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27(9):1332–4. implementation of SMC' combined with ABC

[EM+21] Excofffier, L, Marchi, N, Marques, DA, Matthey-Doret, R, Gouy, A, Sousa, VC (2021) fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* 37(24):4882–4885

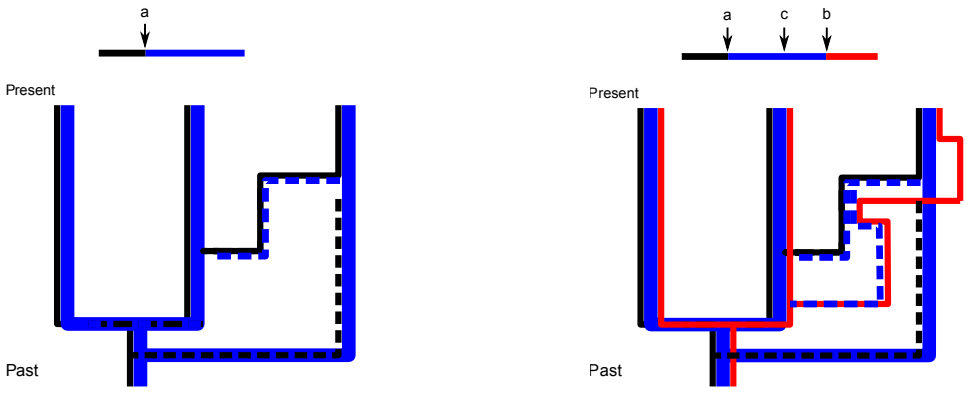


7.3 MaCS, SCRM: closer approximations of the ARG

References

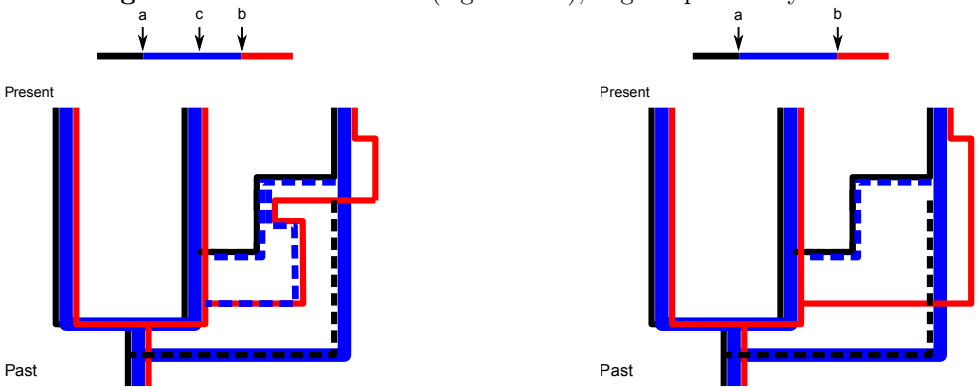
[CMW09] Chen, GK, Marjoram, P, Wall, JD (2009) Fast and flexible simulation of DNA sequence data. *Genome Res* 19(1):136–42 introduce MaCS ("Markovian Coalescent Simulator")

[SZML15] Staab, PR, Zhu, S, Metzler, D, Lunter, G (2015) scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* 31(10):1680–2



Main idea of SCRM and MaCS

If distance $b - a$ is **greater than threshold** (e.g. 100 kb), neglect possibility of coalescence.

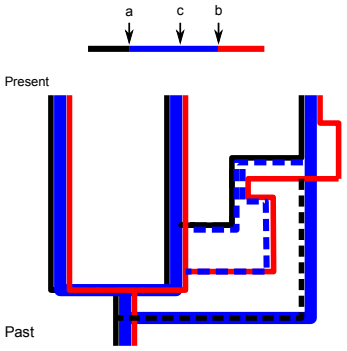


Rationale: recombination will anyway quickly separate the lines.

Main difference between SCRM and MaCS

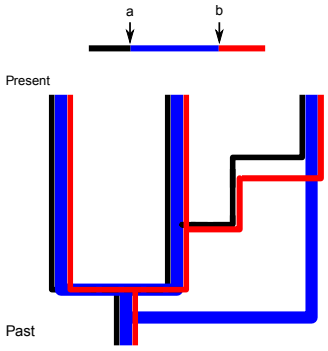
If distance $b - a$ is **smaller than threshold**:

SCRM:

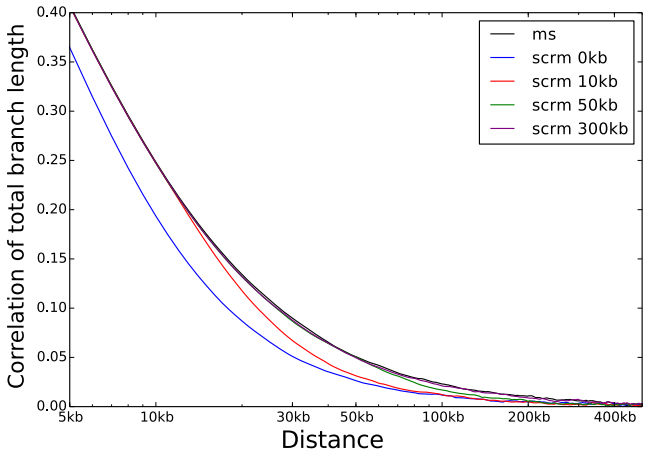


Simulate recombination in non-ancestral segments **when needed**.

MaCS (at least original version):

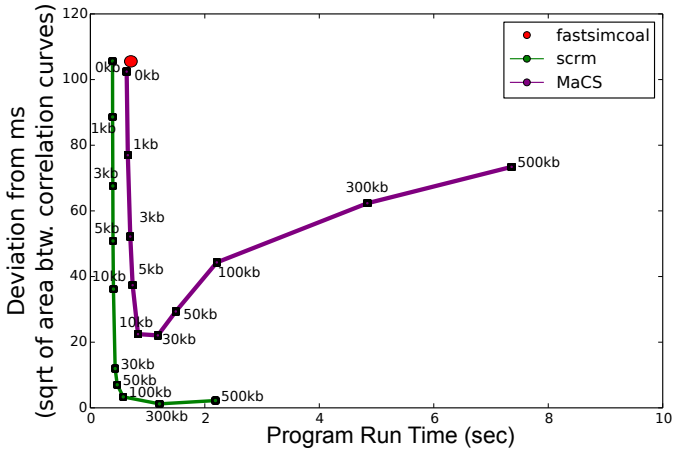


Neglect recombination in non-ancestral segments.



Simulation in Staab et al. (2015)

- 20 sequences
- sequence length 1000 kb
- $\rho = 4000$ for the entire region (that is 4 / kb)



Simulation in Staab et al. (2015)

- fastsimcoal implementation of SMC'
- newer versions of fastsimcoal (fastsimcoal2) or MaCS may of course give different results
- not clear how much accuracy is actually needed for ABC

8 Hidden-Markov model approaches for coalescents with recombination

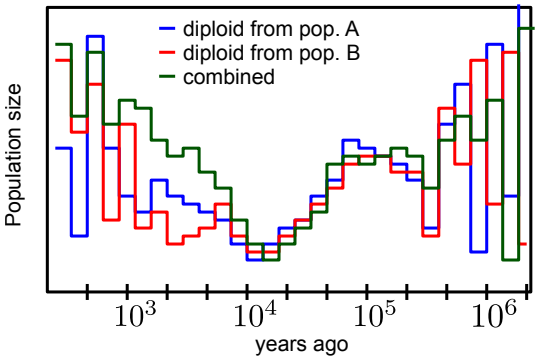
8.1 The Pairwise Sequential Markov Chain (PSMC) Approach

References

[LD11] Li, H, Durbin, R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475(7357):493–6

- More loci and few samples better than vice versa, so how about sample size 2 but whole-genome?
- Even a single diploid genome contains a lot information on demographic history of the whole population
- (implicitly) estimate coalescence times of many regions
- N_e at a time in history is inverse of frequency of coalecence at that time.
- Markov model for how pairwise coalescence time varies along the genome

What PSMC results may look like



For actual results see `https://www.nature.com/articles/nature10231`

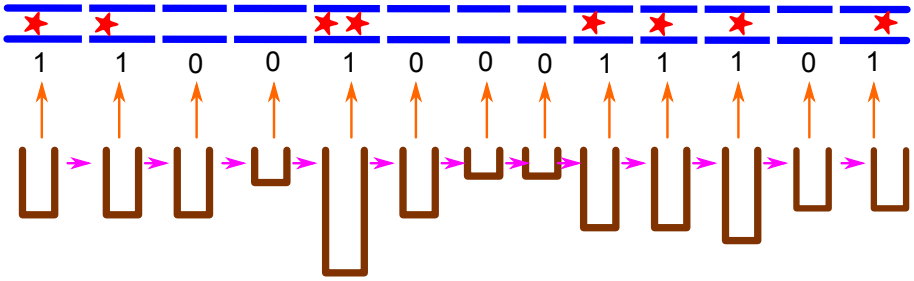
Li and Durbin's findings for human populations

- Data: diploid genomes of 1 Chinese male, 1 Korean male, 3 Europeans, 2 Yoruba (West-Africa)
- Reconstruct within-population N_e histories from diploid genomes of individuals and between-population from X chromosomes of male individuals from the two populations.
- Chinese and European ancestry: very similar population sizes before 10-20 kya, including a bottleneck 10-60 kya.
- gene flow between all populations 20-40 kya.
- seems like West-Africans and non-Africans decended from a homogeneous population at most 100 kya.

Pairwise Sequential Markov Chain (PSMC)
Split pair of input sequences into bins of 100 bp

0: homozygous

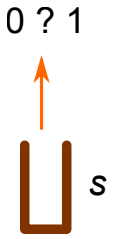
1: heterozygous



- Hidden Markov chain of coalecent tree lengths according to SMC
- Coalescent trees emit observed 0-1 sequence.

Let $\theta = 4N_0\mu$ and $\rho = 4N_0 r$, where μ and ρ are per 100 bp.

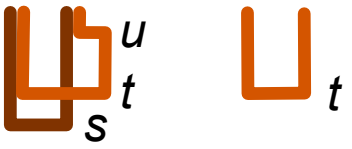
Emission probabilities for 2-sequence coalescent of length s:



0: $e^{-\theta s}$

1: $1 - e^{-\theta s}$

SMC transition rate from tree length s to tree length t
First for $t < s$:



For the simple case of constant $N_e = N_0$:

$$\left(1 - e^{-\rho t}\right) \frac{1}{t} \int_0^t 1 \cdot e^{-(t-u)\cdot 1} \mathrm{d}u \;=\; \left(1 - e^{-\rho t}\right) \frac{1 - e^{-t}}{t}$$

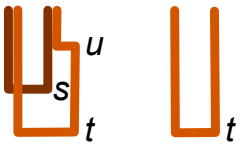
For $N_e(t) = N_0 \cdot \lambda(t)$:

$$\left(1 - e^{-\rho t}\right) \frac{1}{t} \int_0^t \frac{1}{\lambda(t)} \cdot e^{-\int_u^t \frac{1}{\lambda(v)} \mathrm{d}v} \mathrm{d}u$$

Note that PSMC uses SMC (not SMC') and thus neglects cases like .

SMC transition rate from tree length s to tree length t

Now $t > s$ and $N_e(t) = N_0 \cdot \lambda(t)$:



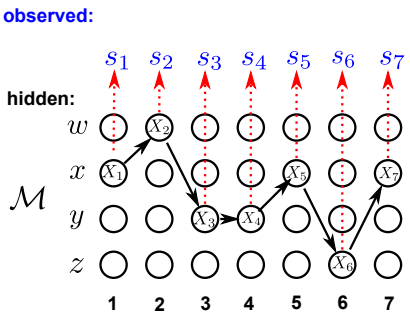
$$\left(1 - e^{-\rho s}\right) \frac{1}{s} \int_0^s \frac{1}{\lambda(t)} \cdot e^{-\int_u^t \frac{1}{\lambda(v)} \mathrm{d}v} \mathrm{d}u$$

In the case of $t = s$ account for the probability $e^{-\rho s}$ that no recombination takes place.

8.2 Excursus: hidden Markov models (HMMs)

Hidden Markov Models (HMMs) in general

$X_1, X_2, \ldots, X_n$ hidden (=unobserved) Markov chain on finite state space $\mathcal{M}$. $S_1, S_2, \ldots, S_n$ series of "signals" with observations $s_1, s_2, \ldots, s_n$.



$P_{x \to y}$: transition probability $\Pr(X_i = y \mid X_{i-1} = x)$

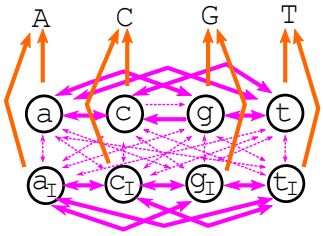
$e_x(s)$: emission probability $\Pr(S_i = s \mid X_i = x)$ (S_i depends only on X_i)

HMM examples

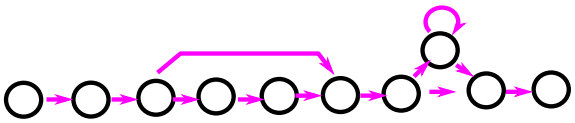
application	hidden	observed
PSMC	Coalscent times	homozygous or heterozygous
CpG island detection	in island or not	nucleotides
Structure	which cluster	alleles
PAC	closest relative haplotype	alleles
protein profileHMM	sequence position	amino acid
alignment pairHMM	alignment	pair of sequences
nanopore sequencing	tuple of nucleotides	electrostatic signal

HMM application examples

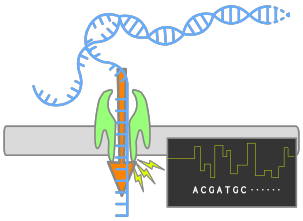
CpG island detection



profileHMM of a protein family



nanopore sequencing



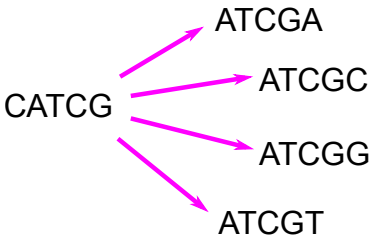


image source:

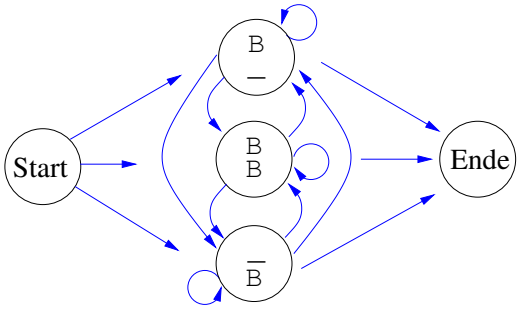
https://commons.wikimedia.org/wiki/File:202001_nanopore_sequencing.svg DataBase Center for Life Science (DBCLS)

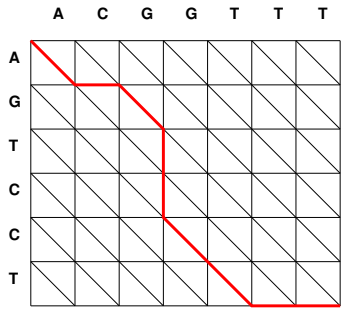
Creative Commons License

https://doi.org/10.7875/togopic.2020.01

pairHMM for sequence alignment

```
ACG--GTTT
A-GTCCT--
```





Classical HMM algorithms

forward algorithm: Calculate Likelihood of parameters, that is, for given parameter values the probability $\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n)$ of the observed sequence $s_1, \ldots, s_n$, summing over all possible chains $x_1, \ldots, x_n$.

backward algorithm: Calculates the same as forward algorithm in another way.

forward–backward algorithm Calculates for each i and x, y of the Markov chain state space

$$\Pr(X_i = x \mid S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) \text{ and}$$

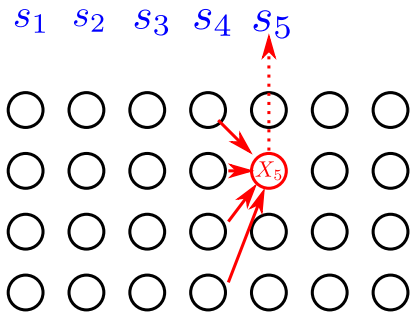
$$\Pr(X_i = x, X_{i+1} = y \mid S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n)$$

Baum–Welch algorithm: Fit HMM parameter to given observation $s_1, s_2, \ldots, s_n$.

Viterbi algorithm: Find most probable chain of hidden states $x_1, x_2, \ldots, x_n$ for given observation $s_1, s_2, \ldots, s_n$ (and given parameter values).

HMM forward algorithm

$$f_i(x) = \Pr(S_1 = s_1, S_2 = s_2, \ldots, S_i = s_i, X_i = x)$$



$$f_i(x) = \sum_{y \in \mathcal{M}} f_{i-1}(y) \cdot P_{y \to x} \cdot e_x(s_i)$$

Time complexity: $O(|\mathcal{M}|^2 \cdot n)$.
(if all transitions $x \to y$ for $x, y \in \mathcal{M}$ allowed)

$$\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = \sum_{x \in \mathcal{M}} f_n(x)$$

Forward algorithm with vector–matrix operations

With $f_i = (f_i(x), f_i(y), \ldots, f_i(k))$

$$P = \begin{pmatrix} P_{x \to x} & P_{x \to y} & \ldots & P_{x \to k} \\ P_{y \to x} & P_{y \to y} & \ldots & P_{y \to k} \\ \vdots & & \ldots & \vdots \\ P_{k \to x} & P_{k \to y} & \ldots & P_{k \to k} \end{pmatrix}, \qquad e(a) = (e_x(a), e_y(a), \ldots, e_k(a)))$$

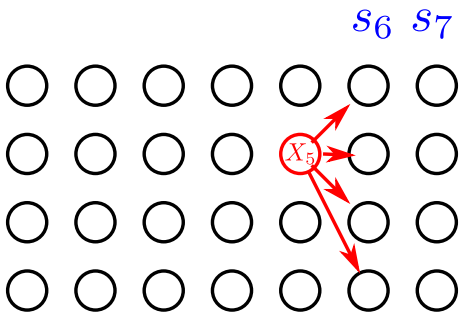
the forward recursion takes the form

$$f_i = (f_{i-1} \cdot P) \circ e(s_i),$$

where $\circ$ is the entrywise Hadamard product.

HMM backward algorithm

$$b_i(x) = \Pr(S_{i+1} = s_{i+1}, S_{i+2} = s_{i+2}, \ldots, S_n = s_n \mid X_i = x)$$



$$\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = \sum_{x \in \mathcal{M}} p_x \cdot e_x(s_1) \cdot b_1(x)$$

$$b_i(x) = \sum_{y \in \mathcal{M}} P_{x \to y} \cdot e_y(s_{i+1}) \cdot b_{i+1}(y)$$

combination of forward and backward variables

From $\Pr(S_1 = s_1, \ldots, S_n = s_n \mid X_i = x) = \Pr(S_1 = s_1, \ldots, S_i = s_i \mid X_i = x) \cdot \Pr(S_{i+1} = s_{i+1}, \ldots, S_n = s_n \mid X_i = x)$ follows

$$\begin{aligned} \Pr(X_i = x \mid S_1 = s_1, \ldots, S_n = s_n) &= \frac{\Pr(X_i = x, S_1 = s_1, \ldots, S_n = s_n)}{\Pr(S_1 = s_1, \ldots, S_n = s_n)} \\ &= \frac{f_i(x) \cdot b_i(x)}{\Pr(S_1 = s_1, \ldots, S_n = s_n)} \end{aligned}$$

and

$$\Pr(X_{i-1} = x, X_i = y \mid S_1 = s_1, \ldots, S_n = s_n) = \frac{f_{i-1}(x) \cdot P_{x \to y} \cdot e_y(s_i) \cdot b_i(y)}{\Pr(S_1 = s_1, \ldots, S_n = s_n)}.$$

This allows us to calculate for given $s_1, \ldots, s_n$ the expected numbers of emissions of any s from any $x \in \mathcal{M}$ and the numbers of transitions $x \to y$ for all $x, y \in \mathcal{M}$, averaged over all hidden chains weighted with their posterior probabilities.

These expectation values are used in the Baum–Welch algorithm, an EM algorithm to estimate transition and emission probabilities (from emissions only; chains are still hidden).

Baum–Welch algorithm

Expectation–Maximization algorithm to estimate HMM parameters from observed data.

Begin with initial HMM transition and emission probability values and repeat the following steps until convergence:

E step:

- With the current HMM parameters apply forward–backward algo to get

$$\Pr(X_i = x, X_{i+1} = y | s_1, \ldots, s_n) \quad \text{and} \quad \Pr(X_i = x | s_1, \ldots, s_n)$$

  for each step i and all MC states x and y.

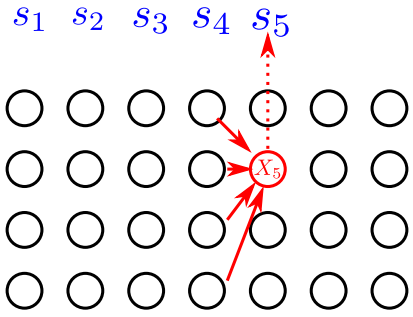
- With this calculate the expectation values for the fractions of transitions to y from x and the distribution of emissions from x, e.g.:

$$\sum_i \Pr(X_i = x, X_{i+1} = y | s_1, \ldots, s_n) \Big/ \sum_i \Pr(X_i = x | s_1, \ldots, s_n)$$

M step: Fit the transition probabilities and emission probabilities to the expected relative frequencies and repeat E-step and then M-step with the updated parameters.

Viterbi algorithm

$$v_i(x) = \max_{(x_1, \ldots, x_{i-1}) \in \mathcal{M}^{i-1}} \Pr(S_1 = s_1, S_2 = s_2, \ldots, S_i = s_i, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x)$$



Allows us to trace back the most probable $x_1, x_2, \ldots, x_n$ for given $s_1, s_2, \ldots, s_n$.

$$v_i(x) = \max_{y \in \mathcal{M}} v_{i-1}(y) \cdot P_{y \to x} \cdot e_x(s_i)$$

8.3 Back to PSMC

How HMM algorithms are applied in PSMC

- time is binned with boundaries $0.1 \cdot e^{\frac{i}{k}\log(1+10T_{\max})} - 0.1$ for $i = 0, 1, \ldots, k$
- thus, each run of the forward, backward has $O(k^2 \cdot n)$ runtime
- parameter estimation with Baum–Welch, combined with Powell's numerical optimization algorithm

Possible Problem for PSMC approach

- used data is distribution of nucleotide diversity across genomic regions
- but this distribution may depend also on variation in mutation rates
- Li and Durbin (2011) carry out simulation study
  - mutation rate variation according to human–macaque alignment
  - result is that these rate variations alone would not explain PSMC results
- More simulation studies for robustness in supplement to Li and Durbin (2011)

8.4 MSMC and MSMC2

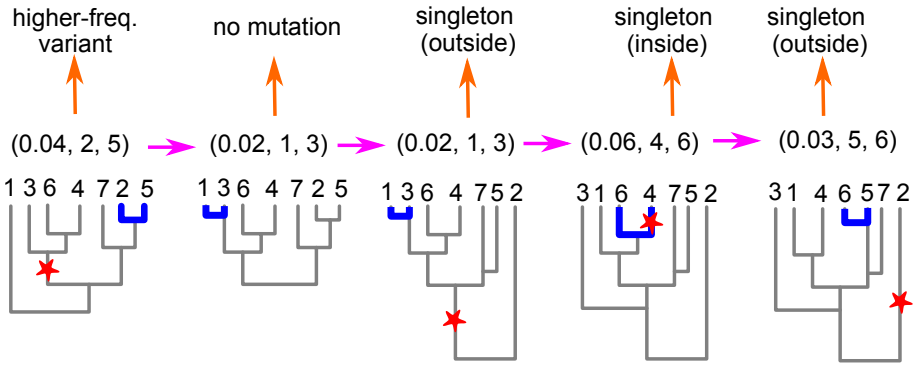
MSMC

References

[SD14] Schiffels, S, Durbin, R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8): 919–25

- input multiple sequences
- Hidden Markov chain states refer to most recent coalescent events
- Aim to get better resolution for more recent demographic history
- Calculations based on SMC' (instead of SMC)

Hidden state in MSMC: time and lineages of most recent coalescence
Emissions:



MSMC neglects:

- all dependencies that are neglected in SMC'
- all other information contained in coalescent outside pair (grey lines)
- emissions depend not only on chain state

MSMC2

References

[SW20] Schiffels S., Wang K. (2020) MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In: Dutheil J.Y. (eds) *Statistical Population Genomics.* Methods in Molecular Biology, vol 2090. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0199-0_7

- MSMC for practical purposes limited to 8 haplotypes
- Biases in MSMC due to approximations in the HMM emission rates
- MSMC2: composite likelihood approach using pairwise coalescents

Some of what you should be able to explain

- Differences between SMC, SMC', MaCS, SCRM and the ARG
- PSMC
  - model assumptions and HMM approach
  - how to draw conclusions
- HMM algorithms
  - forward
  - forward–backward
  - Baum–Welch
  - Viterbi
- basic ideas of MSMC and MSMC2

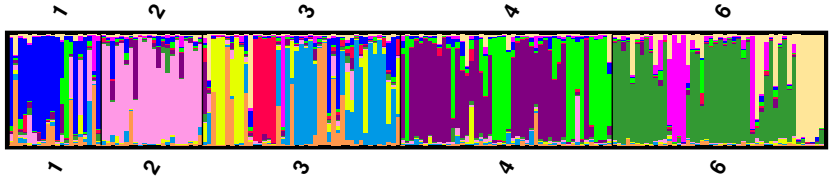
9 The program STRUCTURE

examples

References

[1] M. Linnenbrink, J. Wang, E.A. Hardouin, S. Künzel, D. Metzler, J.F. Baines (2013) The role of biogeography in shaping diversity of the intestinal microbiota in house mice Molecular Ecology 22(7): 1904–1916. (have a look at Fig 1)

[2] B.M. vonHoldt et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids Genome Research 21(8): 1294–1305. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3149496/figure/F4/



References

[PSD00] Pritchard, Stephens, Donnelly (2000) Inference of Population Structure Using Multilocus Genotype Data *Genetics* **155**: 945–959

[FSP03] Falush, Stephens, Pritchard (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* **164**: 1567–1587

[FSP07] Falush, Stephens, Pritchard (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes*

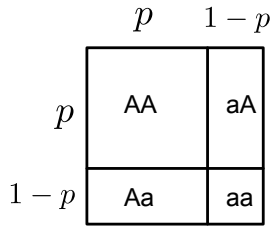
[HFSP09] Hubisz, Falush, Stephens, Pritchard (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resources* **9**: 1322–1332

9.1 no admixture, no sampling locations

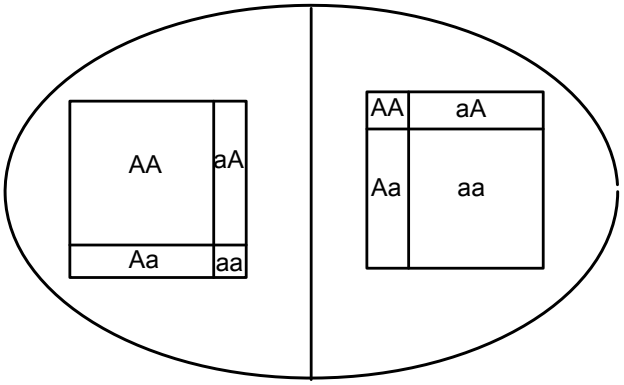
Hardy-Weinberg

A: p a: $1-p$

AA p^2 Aa $2p\cdot(1-p)$ aa $(1-p)^2$



Wahlund effect

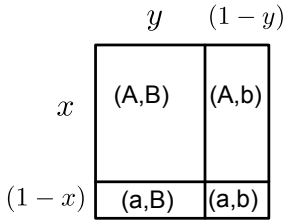


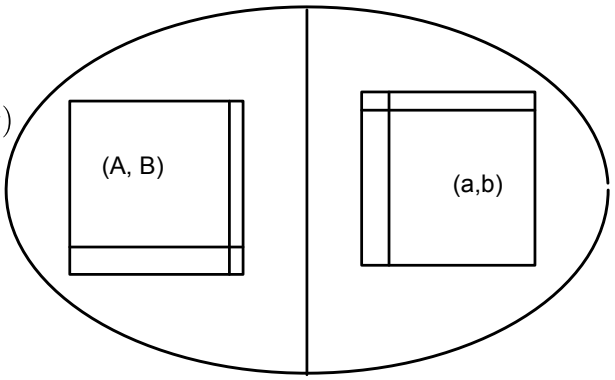
Linkage disequilibrium due to population structure

A: x a: $(1-x)$ B: y b: $(1-y)$

(A, B): $x\cdot y$ (A, b): $x\cdot(1-y)$

(a, B): $(1-x)\cdot y$ (a, b): $(1-x)\cdot(1-y)$





Structure: A program for model-based clustering of genotypes (Microsatellites, SNPS, AFLPs, ...)

N diploid individuals, L loci, K (sub)populations

unknown which individuals belong to which population, even if sampling locations are known, i.e. subpopulations may not correspond to sampling locations.

known is the genotype of individual each i at locus ℓ:

$$X = (x_\ell^{(i,1)}, x_\ell^{(i,2)})_{i\le N,\ell\le L}$$

unknown are the populations from which individual i originates:

$$Z = (z^{(i)})_{i\le N}$$

and the frequencies of allele j at locus ℓ in population k:

$$P = (p_{k\ell j})_{k \le K, \ell \le L, j \le J_\ell}$$

Assumption 1: each population is in Hardy-Weinberg equilibrium

Assumption 2: linkage equilibrium between loci

Bayesian approach: approximate sample from

$$\Pr(Z, P \mid X) \propto \Pr(Z) \cdot \Pr(P) \cdot \Pr(X \mid Z, P)$$

Priors for origin population of individual i:

$$\Pr(z^{(i)} = k) = 1/K$$

Dirichlet prior for allele frequencies in each population:

$$p_{k\ell} \sim \mathcal{D}(\lambda_1, \lambda_2, \ldots, \lambda_{J_\ell}) \text{ with } \lambda_1 = \lambda_2 = \cdots = \lambda_{J_\ell} = 1$$

(uniform distribution on all distributions)

$\Pr(X|Z, P)$:

$$\Pr(x_\ell^{(i,a)} = j \mid Z, P) = p_{z^{(i)}\ell j}$$

Dirichlet distribution

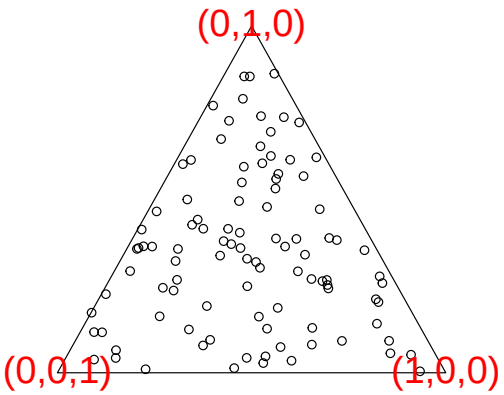
If $Y \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ then

$$\Pr(Y = (y_1, \ldots, y_k)) = c(\alpha) \cdot \prod_{i=1}^{k} y_i^{\alpha_i - 1}$$

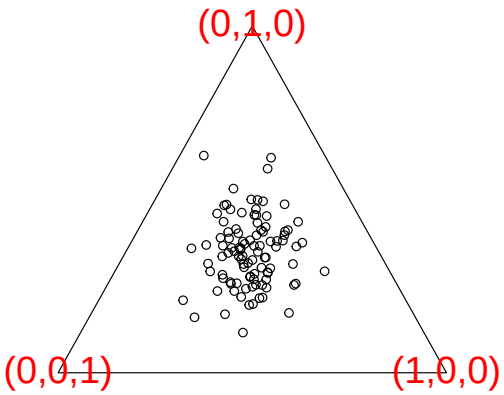
if all $y_i \ge 0$ and $\sum_i y_i = 1$, else 0.

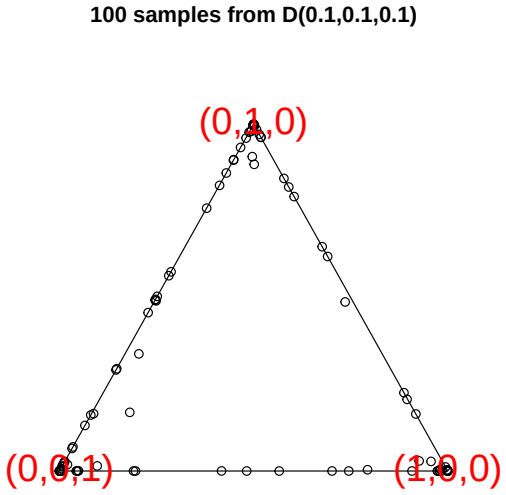
$$\mathbb{E}(Y) = \frac{(\alpha_1, \ldots, \alpha_k)}{\sum_i \alpha_i}$$

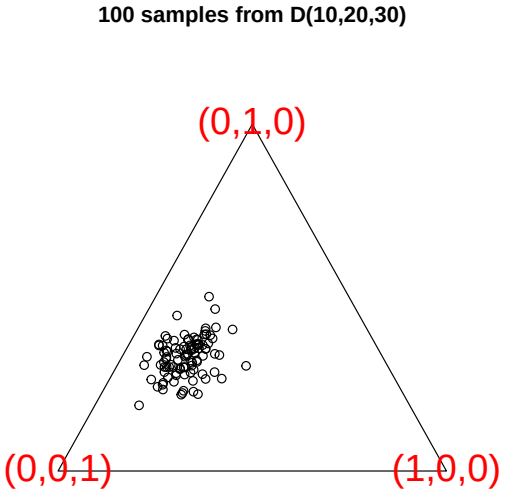
100 samples from D(1,1,1)

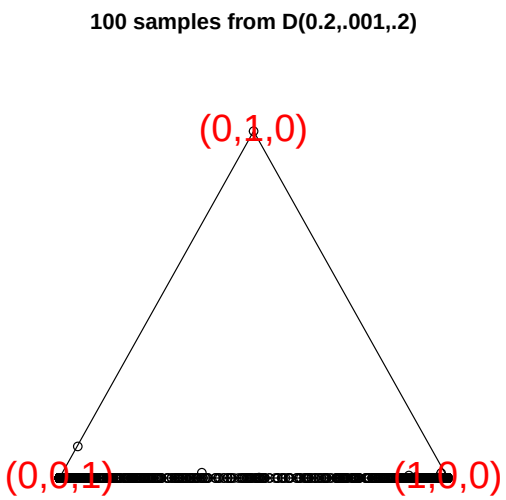


100 samples from D(10,10,10)









Important property of Dirichlet distributions

Let $N = (n_1, \ldots, n_K)$ multinomially distributed with (unknown) probabilities $P = (p_1, \ldots, p_K)$, i.e.

$$\Pr(N = (n_1, \ldots, n_k)) = \frac{(n_1 + n_2 + \cdots + n_k)!}{n_1! \cdot n_2! \cdots n_k!} \prod_{i=1}^{k} p_i^{n_i}.$$

If the prior distribution of P is $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$, then the posterior distribution of P given $N = (n_1, \ldots, n_k)$ is

$$\mathcal{D}(\lambda_1 + n_1, \ldots, \lambda_k + n_k).$$

(Exercise!)

MCMC method for sampling from $\Pr(Z, P|X)$: Start with $Z^{(0)}$ (e.g. sampled from prior) and iterate 2 steps for $m = 1, 2, 3, \ldots$:

1. Sample $P^{(m)}$ from $\Pr(P|X, Z^{(m-1)})$

   $$p_{k\ell\cdot}|X, Z \sim \mathcal{D}(\lambda_1 + n_{k\ell 1}, \ldots, \lambda_{J_\ell} + n_{k\ell J_\ell}),$$

   where $n_{k\ell j} = \#\left\{(i,a)|x_\ell^{(i,a)} = j \text{ and } z^{(i)} = k\right\}$. (using the important property of the Dirichlet distribution.)

2. Sample $Z^{(m)}$ from $\Pr(Z|X, P^{(m)})$

   $$\Pr(z^{(i)} = k|X, P) = \frac{\Pr(x^{(i)}|P, z^{(i)} = k)}{\sum_{k'=1}^{K} \Pr(x^{(i)}|P, z^{(i)} = k')},$$

   using $\Pr(x^{(i)}|P, z^{(i)} = k) = \prod_{\ell=1}^{L} p_{k\ell x_\ell^{(i,1)}} \cdot p_{k\ell x_\ell^{(i,2)}}$.

9.2 with admixture

admixture: present individuals stem from k populations that were admixed recently.

Q : $\left(q_k^{(i)}\right)_{i\le N, k\le K}$ = proportion of individual i's genome originating from population k

Z : $\left(z_\ell^{(i,a)}\right)$ = population of origin of allele copy $x_\ell^{(i,a)}$

$$\Pr\left(x_\ell^{(i,a)} = j \middle| Z, P\right) = p_{z_\ell^{(i,a)},\ell,j}, \qquad \Pr\left(z_\ell^{(i,a)} = k \middle| Q\right) = q_k^{(i)}$$

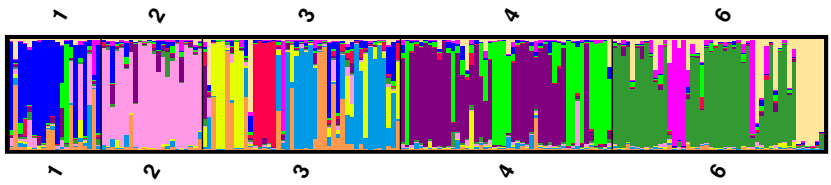
Prior on Q:

$$q^{(i)} = \left(q_1^{(i)}, \ldots, q_k^{(i)}\right) \sim \mathcal{D}(\alpha, \ldots, \alpha),$$

where α is also random with prior $\alpha \sim \text{unif}([0, \alpha_{\max}])$.

Note:

$$\begin{aligned} \alpha = 0 &\Leftrightarrow \text{no admixture} \\ \alpha \to \infty &\Leftrightarrow \text{all completely admixed} \end{aligned}$$



[.5cm] **Interpretation of bars**

without admixture: probabilities of subpopulations to be the origin of individual

with admixture: relative contributions of subpopulations to the genome of the individual

MCMC for case of admixture

Start with initial $P^{(0)}$, $Q^{(0)}$, $Z^{(0)}$ and $\alpha^{(0)}$ and iterate for $m = 1, 2, \ldots$:

1. Sample $P^{(m)}$ and $Q^{(m)}$ from $\Pr(P, Q|X, Z^{(m-1)})$:

   update $p_{k,\ell,j}$ based on the number of ℓ copies of type j that come from population k

   $$n_{klj} = \left\{(i,a)|x_\ell^{(i,a)} = j \text{ and } z_\ell^{(i,a)} = k\right\}$$

   and sample $q^{(i)}|X, Z$ according to

   $$\mathcal{D}\left(\alpha + \#\left\{(\ell,a) : z_\ell^{(i,a)} = 1\right\}, \ldots, \alpha + \#\left\{(\ell,a) : z_\ell^{(i,a)} = K\right\}\right)$$

2. Sample $Z^{(m)}$ from $\Pr(Z|X, P^{(m)}, Q^{(m)})$ according to:

$$\Pr\left(z_\ell^{(i,a)} = k \middle| X, P\right) = \frac{q_k^{(i)} \cdot p_{k\ell x_\ell^{(i,a)}}}{\sum_{h=1}^K q_h^{(i)} \cdot p_{h\ell x_\ell^{(i,a)}}}$$

3. Metroplis Hastings step $\alpha^{(m-1)} \rightsquigarrow \alpha^{(m)}$:
   propose $\alpha' \sim \mathcal{N}(\alpha, \text{some } \sigma^2)$, reject immediately if $\alpha' < 0$, else perform MH step.

Inference for Z, P, Q from MCMC samples
for example for Q it seems obvious to estimate

$$\mathbb{E}(q_i|X) \approx \frac{1}{M} \sum_{m=1}^M q_i^{(m)},$$

but the theoretical posterior mean is

$$\mathbb{E}(q_i|X) = \left(\frac{1}{K}, \ldots, \frac{1}{K}\right)$$

due to symmetries in the model (numbering of populations exchangeable).

$\rightsquigarrow$ use modes of $\left(q_i^{(1)}, \ldots, q_i^{(M)}\right)_i$ instead of means or use Noah Rosenberg's software CLUMPP to evaluate STRUCTURE output.

Inference for the number K of populations

$$\Pr(K|X) \propto \Pr(X|K) \cdot \Pr(K)$$

can be approximated using the harmonic mean estimator

$$\Pr(X|K) \approx M \Bigg/ \sum_{i=1}^M \frac{1}{\Pr\left(X \middle| K, Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)}\right)},$$

but the harmonic mean estimator is know to be imprecise.

Instead, we hope that $-2 \log L(\widehat{Z, P, Q}, \alpha|X)$ is approximately normally distributed and estimate

$$\Pr(X|K) \approx e^{-\widehat{\mu}/2 - \widehat{\sigma}^2/8}$$

with $\widehat{\mu} = \frac{1}{M} \sum_{i=1}^M -2 \log \Pr\left(X|Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)}\right)$
and $\widehat{\sigma}^2 = \frac{1}{M} \sum_{i=1}^M \left(-2 \log \Pr\left(X|Z^{(i)}, P^{(i)}, Q^{(i)}, \alpha^{(i)}\right) - \widehat{\mu}\right)^2$
Pritchard et al. write about this approximation:

"*In fact the assumption underlying* [this] *are dubious at best, and we do not claim (or believe) that our procedure provides a quantitatively accurate estimate of the posterior distribution of K. We see it merely as an ad hoc guide to which models are most consistent to the data, with the main justification being that it seems to give reasonable answers in practice.*"

and:

"*The inferred value of K may not always have a clear biological interpretation.*"

and about the multiple-modes problem:

"[The] *Gibbs-sampler did not manage to move between two modes in any of the runs*"

Data examples

Bird example: Without using informations on sampling locations, STRUCTURE gave clear clusters corresponding to sampling locations, up to a few exceptions. Neighbor-Joining results did not show clear clusters when labels were removed.

`http://www.genetics.org/content/155/2/945/F4.large.jpg`
`http://www.genetics.org/content/155/2/945/F3.expansion.html`

Human data: Found $K \geq 2$ corresponding to African and European origin of samples. Evidence for $K > 2$ may indicate substructure.

Some of the things you should be able to explain

- Wahlund effect and its analogon for linkage
- Dirichlet distribution and how they are used in Structure
- How Dirichlet priors lead to Dirichlet posteriors
- MCMC method in structure
- meaning of Structure barplots with and without admixture

9.3 taking sampling locations into account

First attempt: populations correspond to sampling locations with a few migrants in the last few generations.

$g(i)$**:** sampling location of individual i

ν**:** probability that i is immigrant or offspring of an immigrant in the last G generations, where G is not too large.

$\Rightarrow q^{(i)}_{g(i)} = 1$ with probability $1 - \nu$ and for $t \leq G$:
$q^{(i)}_{g(i)} = 1 - 2^{-t}$ and $q^{(i)}_{j} = 2^{-t}$ with probability $\frac{2^t \nu}{(k-1)\sum_{T=0}^{G} 2^T}$ (neglecting the possibility of more than one migranting ancestor in the last G generations.)

in MCMC: sampling of $q^{(i)}$ is conditioned P, X and Z.
Falush et al. (2003) allow for LD between loci. Advantages:

1. detection of admixture further back into past
2. inference of population of origin of chromosomal regions
3. more accurate estimate of statistical uncertainty when linked loci are used

Sources of LD:

mixture LD: variation in ancestry among sampled individuals (Prichard et al.)

admixture LD: correlation of ancestry along each chromosome causes additional LD between linked markers (Falush et al.)

background LD: within population decaying on a much shorter scale, e.g. tens of kb in humans. (not yet in STRUCTURE)

Approach of Falush et al. (2003):

- breakpoints occur as Poisson process at rate r
- uniform prior on $\log(r)$
- use HMM to sample from conditional distribution of Z

- data allowed to be unphased

more options: correlated allele frequencies between populations according to star-shaped phylogeny of populations with drift rates $F_1, \ldots, F_K$ and ancestral allele frequency distribution $p_A \sim \mathcal{D}(\lambda_1, \ldots, \lambda_{J_\ell})$.

$$p_{k\ell\cdot}|p_A \sim \mathcal{D}\left(p_{A\ell 1}\frac{1-F_1}{F_1}, \ldots, p_{A\ell K}\frac{1-F_K}{F_K}\right)$$

(be careful with this model!)
Approach of Hubisz et al. (2009): Allow uncertainty in the information about sampling location

$$\begin{aligned}
r &\sim \text{unif}([0, r_{\max}]) \qquad \text{(informativeness of sampling location)} \\
q^{(i)} &\sim \mathcal{D}\left(\alpha_{h_1}, \ldots, \alpha_{h_K}\right), \quad \text{if individual } i \text{ comes from location } h \\
\alpha_{h_k} &\sim \Gamma\left(r \cdot \alpha_k^{\text{glob}}, 1/r\right), \quad \text{(which entails that the mean is } \alpha_k^{\text{glob}}\text{)} \\
\alpha_k^{\text{glob}} &\sim \text{unif}(0, \alpha_{\max})
\end{aligned}$$

Hubisz et al.: "*However, we would still encourage users to run the original models as well, and to check that substantial differences between the results from the new and the old models seem biologically sensible.*"

When STRUCTURE has problems

- number of clusters not well-defined when allele frequencies vary slowly accross the landscape
- inbreeding or relatedness between individuals

  In this case, the software INSTRUCT may help, cf.

References

[GWB07] H. Gao, S. Williamson, S.D. Bustamante (2007) An MCMC Approach for Joint Inference of Population Structure and Inbreeding Rates from Multi-Locus Genotype Data. *Genetics (online)*

9.4 Faster alternatives to STRUCTURE for large datasets

9.4.1 ADMIXTURE

ADMIXTURE

- is based on the same modeling approach as STRUCTURE
- (faster) ML optimization instead of Bayesian sampling

References

[GWB07] D.H. Alexander, J. Novembre, K. Lange (2009) Fast Model-Based Estimation of Ancestry in Unrelated Individuals *Genome Res.* **19**: 1655–1664

optimization strategy: similar to Newton's method

Problem: Huge Hesse matrix (2nd derivatives) as there are many parameters.

Q**:** q_{ik} is proportion of individual i genome coming from population k

F**:** $f_{k\ell}$ is the frequency of allele 1 of locus ℓ in population k (assuming two alleles per locus).

$\Rightarrow$ Many second derivatives

$$\frac{\partial^2}{\partial q_{ik}\partial f_{k\ell}}.$$

Also the constraints $0 \leq f_{kj} \leq 1$, $q_{ik} > 0$, $\sum_k q_{ik} = 1$ make optimization a bit tricky.

ADMIXTURE uses **Block Relaxation Algorithm**

- like Newton method uses first two derivatives
- To optimize $L(Q, F)$ iterate
  - update Q for fixed F
  - update F for fixed Q
- no mixed 2nd derivatives $\frac{\partial^2 L}{\partial q_{ik}\partial f_{k'\ell}}$ needed
- need $\frac{\partial^2 L}{\partial q_{ik}\partial q_{i'k'}}$ only if $i = i'$.
- need $\frac{\partial^2 L}{\partial f_{k\ell}\partial f_{k'\ell'}}$ only if $\ell = \ell'$.
- optimization problems are convex.

9.4.2 fastSTRUCTURE

fastSTRUCTURE

References

[1] A. Raj, M. Stephens, J.K. Pritchard (2013) Variational Inference of Population Structure in Large SNP Datasets *preprint available on bioRχiv*

Variational Distributions: tractable family of distributions to approximate posterior.

Variational Bayesian Inference: Instead of sampling from posterior, optimize parameters of variational distributions

Kullback-Leibler Divergence: (=relative entropy)

$$D_{KL}(q||p) = \mathbb{E}_q \log \frac{q(X)}{p(X)} = \int_x q(x) \log \frac{q(x)}{p(x)} dx$$

Approach: Find variational distribution q that minimizes $D_{KL}(q||p)$ to posterior p.

Unrealistic assumption to make variational distributions q tractable:

Z, P, and Q are independent.

Their joint variational distribution density is the product of multinomial probabilities for Z, Dirichlet densities for Q, and beta densities for P.

The parameters of these distributions are optimized.

Also here, the optimization of the parameters of one of distribution, keeping the others fixed, is a convex optimization problem.

9.4.3 snmf() in the Bioconductor R package LEA

The LEA package

`http://membres-timc.imag.fr/Olivier.Francois/LEA/index.htm`

Installation:

```
source("https://bioconductor.org/biocLite.R")
biocLite("LEA")
```

References

[1] E. Frichot, F. Mathieu, T. Trouillon, G. Bouchard, O. François (2014) Fast and efficient estimation of individual ancestry coefficients *Genetics* **196** (4): 973–983

[2] E. Frichot, S.D. Schoville, G. Bouchard, O. François (2013) Testing for associations between loci and environmental gradients using latent factor mixed models *Molecular Biology and Evolution* **30** (7): 1687–1699

(We focus on the sNMF method of Frichon et al., 2014)
Ideas behind sNMF:

- By **not** assuming Hardy-Weinberg equilibrium
  - we may lose a bit of information but
  - make the method more robust against inbreeding
  - and make the method computationally less demanding.
- Combine efficient solutions from numerics / statistics
  - linear regression / least squares:

    Minimize $S(q) = \sum_j (y_j - \sum_k q_k V_{kj})^2$ by $\widehat{q} = (V^T V)^{-1} V^T y$
  - regularization / ridge regression::

    In linear regression avoid overfitting by extreme coefficients by an additional penalty. That is, minimize
    $$S(q) = \sum_j (y_j - \sum_k q_k V_{kj})^2 + \alpha \cdot \sum_k q_k^2$$
    for some regularion penalty $\alpha > 0$.
  - convex optimization:

    minimize a convex function f on a convex definition range $\mathcal{D} \subseteq \mathbb{R}^n$.

Convex Optimization

$\mathcal{D}$ is convex if $\forall x \in \mathcal{D}, y \in \mathcal{D}, \lambda \in [0,1] : \lambda \cdot x + (1-\lambda) \cdot y \in \mathcal{D}$.

$f : \mathcal{D} \to \mathbb{R}$ is convex if $\forall x \in \mathcal{D}, y \in \mathcal{D}, \lambda \in [0,1]$:

$$f(\lambda \cdot x + (1-\lambda) \cdot y) \leq \lambda \cdot f(x) + (1-\lambda) \cdot f(y).$$

Nice: All local minima of convex functions on convex sets $\mathcal{D}$ are global minima and form a convex set.

Example for convex optimization: Minimize

$$S(q) = \sum_j (y_j - \sum_k q_k V_{kj})^2 + \alpha \cdot \sum_k q_k^2$$

(where $\alpha > 0$) with the side condition that $\forall k : q_k > 0$.

This can be solved efficiently by nonnegative matrix factorization (NMF), see

References

[1] J. Kim, H. Park (2011) Fast nonnegative matrix factorization: an active-set-like method and comparisons. *SIAM J. Sci. Comput.* **33**:3261-3281

How these methods are combined in Frichot et al.'s sNMF.

Notations:

$j \in \{0, 1, 2\}$ genotype at diploid locus; 1 means heterozygous

$x_{i\ell}(j) \in \{0, 1\}$ 1 if individual i has genotype j at locus ℓ.

q_{ik} fraction of i's genome coming from subpopulation k

$g_{k\ell}(j)$ frequency of genotype j at locus ℓ in population k

Aim: Find $q = (q_{ik})_{ik}$ and $g = (g_{k\ell}(j))_{k\ell j}$ that minimize

$$LS(q, g) = \sum_{i,\ell,j} \left(x_{i\ell}(j) - \sum_k q_{ik} g_{k\ell}(j) \right)^2$$

Choose some $\alpha > 0$, then start with initial (q, g) and iterate the following steps 1 and 2 until you observe convergence:

1. keeping the current q fixed, update g:
   (a) minimize (like in linear regression):
   $$LS_1(g) = LS(q, g) = \sum_{i,\ell,j} \left(x_{i\ell}(j) - \sum_k q_{ik} g_{k\ell}(j) \right)^2$$
   (b) set all negative $g_{k\ell}(j)$ to 0
   (c) $\forall k, \ell$: normalize $(g_{k\ell}(0), g_{k\ell}(1), g_{k\ell}(2)$ such that $\sum_j g_{k\ell}(j) = 1$.
2. keeping the current g fixed, update q:
   (a) Apply NMF to numerically minimize
   $$LS_2(q) = \sum_{j\ell i} \left(x_\ell(j) - \sum_k q_{ik} g_{k\ell}(j) \right)^2 + \sum_i \alpha \cdot \left(\sum_k q_{ik} \right)^2$$
   with the boundary condition $\forall j, k : q_{jk} \geq 0$.
   (b) normalize q such that $\forall i : \sum_k q_{ik} = 1$.

Some of the things you should be able to explain

- Different types of LD and how they are covered in STRUCTURE
- Limitations in STRUCTURE
- alternative tools and what is different in them
- LEA:
  - why not assuming Hardy-Weinberg
  - ridge regression
  - convex optimization

10 The genetic footprints of selection: simulation and detection

We will now consider the effect of selection and adaptation on genealogies. We will discuss how these effects can be simulated, because

1. this is a way to specify the theoretical model and
2. if we know how to simulate data, we can apply ABC and similar methods for statistical inference of model parameters.

Possible scenarios for the case of positive selection (directional selection):

1. A beneficial mutation appears once, spreads in the population and will eventually be fixed
2. there is a balance between mutation and selection
3. selection pressure changes in time and a certain allele is favored for a while and increases in frequency during that time

Other forms of selection:

background selection , also called negative selection

balancing selection can lead to maintance of two alleles over a long period of time

diversifying selection : new types appear by mutation and have an advantage until the reach a certain frequency

etc.

Basic model of positive selection: Each individual i in a population of size N has a fitness w_i, which is the expected number of kids. The N surviving offspring of the next generation are sampled from the kids of all individuals. Thus, the expected number of surviving offspring of i is $w_i / \sum_{j=1}^{N} w_j$. Let's

assume a simple scenario: haploid population with one type A of fitness $1 + s$ and one type a of fitness 1.

Moran model

Assume a population of $2N$ gametes. An alternative to the Wright-Fisher model is the Model of

Moran(1958): Each gamete has a rate of 1 to generate one offspring and replace one randomly chosen gamete. For $N \to \infty$ and time scaled in units of N (not $2N$!) generations, the genalogy of a sample

from the Moran model converges to the standard Kingman coalescent. Add selection to Moran model:

Type A produces offspring at rate 1 and type a at rate $(1 - s)$. This approximates the diploid case with

fitness 1 of AA, $1 - s$ of Aa, and $(1 - s)^2 \approx 1 - 2s$ of aa. (Note that capital letter A does not indicate dominance.)

Transition rates of number of allele A gametes in Moran model with selection:

$$i \to i + 1 \quad \text{at rate} \quad \frac{(2N - i) \cdot i}{2N}$$

$$i \to i - 1 \quad \text{at rate} \quad \frac{(2N - i) \cdot i}{2N} \cdot (1 - s)$$

More facts about Moran model with selection:

Fixation probability: If we start with i gametes of type A, the fixation probability of A is

$$\frac{1-(1-s)^i}{1-(1-s)^{2N}} \approx \frac{1-e^{-is}}{1-e^{-2Ns}}$$

Fixation time: Assume that type A starts with one gamete. Conditioned on the fixation of A, the expectation value of the fixation time is in the limit of large populations is asymptotically

$$\frac{2}{s}\log N$$

generations.

For proofs see:

References

[D08] R. Durrett (2008) *Probability Models for DNA Sequence Evolution* 2nd Ed., Springer

In mathematical population genetics two cases are considered:

weak selection: as $N \to \infty$, $s \to 0$ such that $Ns \to \widetilde{s} < \infty$.

strong selection: as $N \to \infty$, s is stays constant, with the consequence that the fixation time of the advantageous allele is 0 on the time scale of N generations.

10.1 Ancestral Selection Graphs

References

[NK97] C. Neuhauser, S.M. Krone (1997) Ancestral processes with selection *Theor. Pop- Biol.* **51**:210–237

[KN97] S.M. Krone, C. Neuhauser (1997) The genealogy of samples in models with selection. *Genetics* **145**:519–534

Ancestral Selection Graph (ASG)

Weak selection: $2Ns \to \widetilde{s}$ as population size $2N \to \infty$. Each pair of ancestral lineages coalesces at rate 1. At rate $\theta/2$ lineage of type a mutates into A and vice versa. Each lineage x is hit by "arrow" at rate $\widetilde{s}$. Arrow was shot by random indivdual y from population. If (further in the past) x was of type a and y of type A, replace (in future direction) x by type A. To find out whether this applies, trace lineages back into past. At latest when all lineages are coalesced, types of all lineages are determined.

This happens almost surely after finite time because number j of lineages to trace back jumps to $j+1$ at rate $\widetilde{s} \cdot j$ only but jumps to $j-1$ at rate $j \cdot (j-1)/2$.

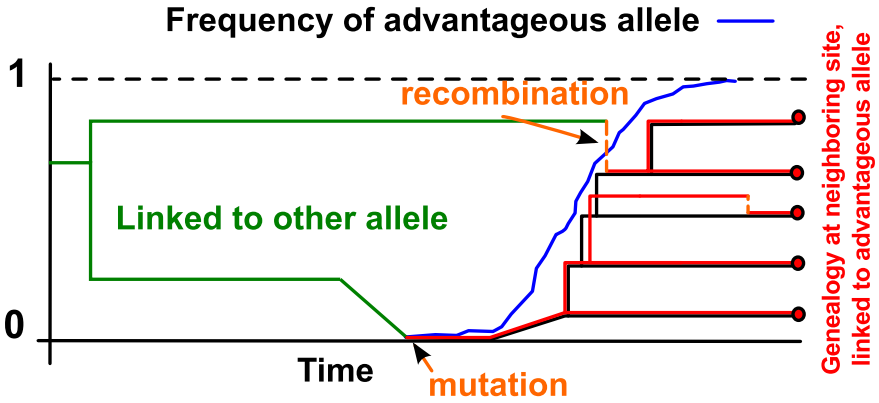
ASG for frequency-dependent selection with advantage of rare alleles

References

[N99] C. Neuhauser (1999) The ancestral graph and gene genealogy under frequency-dependent selection. *Theoretical Population Biology* **56**:203–214

When lineage x is hit by "replacement arrow" from y, it shoots a "check arrow" to some random individual z from the population. It copies the type of y if and only if the type of z is different than that of y. Thus, lineages of x, y, and z have to be traced back. But again all lineages will coalesce in finite time because rate of adding lineages is linear whereas rate of coalescence is quadratic in number of lineages. In structured population one can also assume that the arrows are shot locally to model that selection depends on local frequencies.

10.2 Simulating selective sweeps and other kinds of strong selection



In the case of strong selection, coalescent based simulation faces the problem that in time span of length 0 advantageous type becomes fixed, and backward in time all lineages would coalesce at that time. In case of the ancestral recombination graph (ARG) there is even no time for recombination. However, if we model a locus that is far from selected locus, recombination rate is also high and in mathematical models we could also let it go to ∞ such that lineages can escape the selective sweep. In the simulation programm MSMS, however, the approach is to first simulate the locus under selection for finite N, and then the ARG around (or next to) it conditioned on the frequency trajectory of the selected allele.

References

[EH10] G. Ewing, J. Hermisson (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**(16): 2064–2065

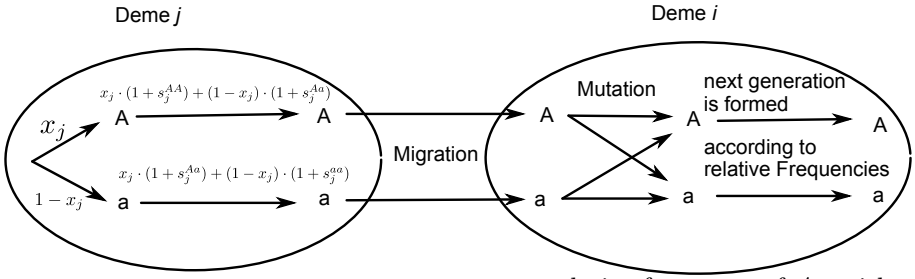
Strategy: First simulate the site under selection forward in time with discrete generations and finite N. Then generate the ARG backwards in time conditioned on the simulated development of allele frequencies. Allows to specify when selection starts and when it ends (or some condition, e.g. fixation).

MSMS model for selected locus

Works for fitness function of haploids or diploids.

Selection can depend on deme.

Fitness of genotypes aa, aA and AA on deme i:
$(1 + s_i^{aa})$, $(1 + s_i^{aA})$, and $(1 + s_i^{AA})$
Also balancing selection possible by choosing $s_i^{aa} < s_i^{aA} > s_i^{AA}$



m_{ij} : fraction of island j immigrants on island i

$m_{ii} := 1 - \sum_{j \neq i} m_{ij}$.

x_i : relative frequency of A on island i

μ : mutation rate $a \to A$

ν : mutation rate $A \to a$

$$\begin{aligned} \eta_i^A &:= \sum_j m_{ij} x_j \left(1 + (1 - x_j) s^{aA} + x_j s^{AA}\right) \\ \eta_i^a &:= \sum_j m_{ij} (1 - x_j) \left(1 + x_j s^{aA} + (1 - x_j) s^{aa}\right) \end{aligned}$$

$$x_i' := \frac{(1 - \nu)\eta_i^A + \mu \eta_i^a}{\eta_i^A + \eta_i^a}$$

Then, the number of copies of A on island i in the next generation is drawn from binomial distribution with parameters $(2N_i, x_i')$. For ARG backwards simulations, continuous time is assumed. For this, each generation from the forward simulation is replaced by suitable time span.

MSMS simulation strategy for ARG

Simulate migration, coalescence, mutation and recombination. As far as possible use that the time to the first event (back in time) of several possible events with exponential waiting time is also exponentially distributed with rate being the sum of the single rates. But take into account that rates vary in time due to population size changes and changes of allele frequencies at locus under selection. Lineages can coalesce only if they have the same type A or a at the selected locus.

E.g. if there are k lineages of type A on island i, the total coalescence rate of these lineages is

$$\frac{k \cdot (k - 1)}{4 N_i x_i}$$

as long as type A has frequency x_i on island i.

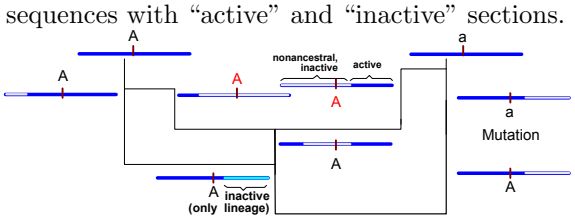
Migration and Mutation rate must be corrected for allele frequencies simulated afore. If m_{ij} is the fraction of island i inhabitants that immigrated from island j, then the fraction of immigrants from j of type A is $x_j m_{ij}$. Thus, the fraction of immigrants among the type A inhabitants of island i is $x_j m_{ij} / x_i$.

Similarly, lineages of type A are traced back to be mutated from lineages of type a on island i at rate $\mu \cdot (1 - x_i)/x_i$.

Simulating the ARG, given the locus under selection

During the backwards simulation, all lineages have sequences with "active" and "inactive" sections.

Active means:
a mutation in such a region would lead to
a polymorphic site in the sampled sequences.



- If recombination happens, non-ancestral material is deactivated. If a region is active in only one lineage, it becomes inactive.
- When a recombination happens in an inactive range, its exact location is usually irrelevant.
- Exception: If the locus under selection is in the inactive region, it matters whether recombination happened left or right of it.

We must alway keep track of the type a or A of each lineage at the selected locus, even if this locus is outside the range for which we simulate the ARG. In particular, when recombination leads to a split

of a lineage, one lineage keeps the locus under selection (which one is clear from the position of the recombination point). The type of the other lineage is of type A or a with probabilities x_i or $1 - x_i$, respectively.

There have been plans to make it possible in msms to simulate quite generals scenarios with more than one locus under selection but it is not clear whether development will go on, see https://github.com/delt0r/msms.

Alternative: use a forward simulater, e.g. SLiM

Some of the things you should be able to explain

- Moran model with selection
- Fixation probability and fixation time with selection
- mathematicians' distiction between weak and strong selection
- Ancestral selection graph and its extension for frequency-dependent selection
- Simulation strategy in MSMS

11 Statistics for detecting genomic footprints of selection

Classical genetic signal of selection: dN/dS.

We will now discuss statistics to detect genomic signals of various kinds of selection processes. These statistics are just examples; many more statistics can be found in the literature and in several software packages.

11.1 Detecting selective sweeps

Kimura (1971) has shown for standard neutral model that the expected number of mutated sites with frequency of derived allele in $[p, p + dp]$ is

$$\phi_0(p)dp = \frac{\theta}{p}dp.$$

Fay and Wu (2000) have generalized this to mutations neighboring a sweep with selection strength s, recombination rate r between the sites, initial frequency of ε, and $C := 1 - \varepsilon^{r/s}$. In this case, ϕ_0 is replaced by (approximately)

$$\phi_1(p) = \left(\frac{\theta}{p} - \frac{\theta}{C}\right) \cdot I_{0<p<C} + \left(\frac{\theta}{C}\right) \cdot I_{(1-C)<p<1},$$

where $I_A = 1$ if A is fulfilled and $I_A = 0$ otherwise.

$\Rightarrow$ Probability of finding k derived alleles in sample of n:

$$P_{n,k} = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \phi(p) dp$$

References

[1] M. Kimura (1971) Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology* **2**: 174–208

[2] J. Fay, C.-I. Wu (2000) Hitchhiking under positive Darwinian selection *Genetics* **155**: 1405–1413

[3] Y. Kim, W. Stephan (2002) Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome *Genetics* **166**: 765–777

[4] R. Nielsen, S. Williamson, Y. Kim, M.J. Hubisz, A.G. Clark, C. Bustamante (2005) Genomic scans for selective sweeps using SNP data. *Genomic Research* **15**: 1566–1575

Kim and Stephan (2002) propose composite-likelihood ratio statistic. Here, "composite" means that stochastic dependencies (linkage) among neutral sites are neglected, and just the product of the probabilities $P_{n,k}$ for all sites is used.

Nielsen *et al.* (2005) propose a variant of this, which

- is not restricted to one null model; instead average site-frequency spectrum of full chromosome is used.
- provides correction for ascertainment bias

$$L(\Theta) \propto \Pr(D_i|\Theta, A_i) = \Pr(D_i|\Theta) \cdot \frac{\Pr(A_i|D_i, \Theta)}{\Pr(A_i|\Theta)}$$

where D_i is the data at SNP i, and A_i is the condition why i is considered.

Software: SweepFinder2 `http://degiorgiogroup.fau.edu/sf2.html`

References

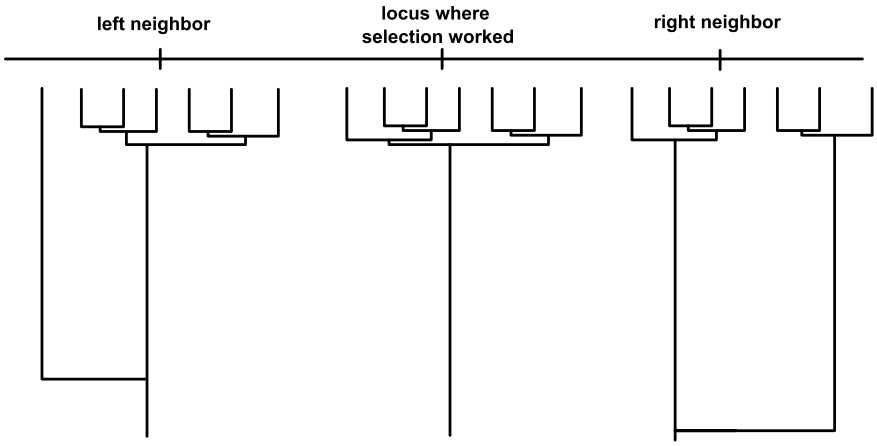
[1] Y. Kim and R. Nielsen (2004) Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics* **167**:1513–1524

propose statistic ω based on LD (linkage disequilibrium) pattern.

Given two loci with alleles A/a and B/b with frequencies $\pi_A, \pi_B, \pi_a, \pi_b$ and frequency π_{AB} of haplotype AB, one possible measure of LD is

$$r^2 = \frac{(\pi_{AB} - \pi_A \cdot \pi_B)^2}{\pi_A \cdot \pi_B \cdot \pi_a \cdot \pi_b}$$

(= squared correlation of indicator functions of A and B)



Kim and Nielsen's ω

Given S segregating sites, split them into the set L of the first ℓ sites and the other $S - \ell$ sites R. Then compute

$$\omega_\ell = \frac{\left(\sum_{i,j\in L} r_{ij}^2 + \sum_{i,j\in R} r_{ij}^2\right) \Big/ \left(\binom{\ell}{2} + \binom{S-\ell}{2}\right)}{\left(\sum_{i\in L, j\in R} r_{ij}^2\right) / (\ell \cdot (S-\ell))}$$

and set

$$\omega = \max_\ell \omega_\ell.$$

To assess significance of any evidence for selection indicated by such statistics, we need to use a null model that accounts for population structure and demography.

XP-CLR

XP-CLR: cross-population composite-likelihood ratio test

Detects sweeps in one of two closely related species (or populations).

References

[1] H. Chen, N. Patterson, D. Reich (2010) Population differentiation as a test for selective sweeps. *Genome Research* **20**:393–402

XP-CLR

Assume sweep at locus A in population 1. Locus B is linked to locus A.

s strength of selection in sweep

r recombination rate between the two loci

p_2 allele frequency at locus B in population 2

w parameters of demographic model such as split time, genetic drift rates...

$f(p_1|p_2, w, s, r)$ approximate probability density of allele frequency at locus B in population 1, conditioned on r, s, p_2 and w, combining several theoretical results, e.g. by Durrett and Schweinsberg (2004)

XP-CLR

Data set consists of sites i that are linked to selected site with r_i and have allele frequency $p_{i,2}$ in population 2 and m_i/n in population 1 (where n is the sample size in population 1).

composite likelihood approach: neglect linkage among sites i

$$CL(r,w,s) = \prod_{i=1}^{k} \int_0^1 f(p_{1,i}|p_{i,2},w,s,r_i) \cdot \binom{n}{m_i} \cdot p_{1,i}^{m_i} \cdot (1-p_{1,i})^{n-m_i} \; dp_{1,i}$$

The statistic is then:

$$2 \cdot \left(\max_{s,r,w} \log CL(r,w,s) - \max_{w} \log CL(r,w,0) \right)$$

11.1.1 Stats for detecting incomplete sweeps

Extended Haplotype Homozygosity

References

[1] P.C. Sabeti, D. Reich,..., E.S. Lander (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837

$EHH(x)$ Among all pairs of sampled sequences the fraction of those that are identical in the range from x cM downstream to x cM upstream of the candidate site.

iHS for incomplete sweeps

References

[1] B.F. Voight, S. Kudaravalli, X. Wen, J.K. Pritchard (2006) A map of recent positive selection in the human genome. *PLoS Biology* **4**: e72

Scale genomic location x in units of $4N_e r$ (for autosomes), where r is the recombination rate and use EHH with linear interpolation between SNPs. Let a and b be the first points up- and downstream of the candidate site where $EHH(x) \leq 0.05$. Then define:

$$iHH = \int_a^b EHH(x) \; dx$$

Calculate iHH_A and iHH_D separately for ancestral and derived allele. Then define:

$$\begin{aligned} uiHS &= \ln \frac{iHH_A}{iHH_D} \qquad \text{(unscaled iHS)} \\ iHS &= \frac{uiHS - \mu_p}{\sigma_p} \end{aligned}$$

In

$$iHS = \frac{uiHS - \mu_p}{\sigma_p}$$

μ_p and σ_p are mean and standard deviation of $uiHS$ estimated from SNPs in which the derived allele has the same distribution as in the candidate site.

Software: http://hgdp.uchicago.edu/Software/ or R package rehh

References

[1] Y. Field,..., J.K. Pritchard (2016) Detection of human adaptation during the past 2000 years. *Science* **354**: 760–764

idea of Single Density Score (SDS): if rare allele gets under recent positive selection, its tip branch lenghts are shorter than those for ancestral allele.

Software: https://github.com/yairf/SDS

1. Assume there are n diploid individuals in the sample of unphased data.
2. "singleton" is a derived allele that appears in only one the $2n$ sequences of diploid individuals
3. for each diploid individual i we consider the two sites that are the nearest to the candidate in upstream and downstream direction, respectivly, among all singletons that are heterozygous in i. Let d_i be the distance between these two sites.
4. from the distances d_i of i with derived or ancestral alleles first estimate the average tip branch lenghts $\widehat{t}_d$ and $\widehat{t}_a$ for both groups
5. $$SDS^* = \log \frac{\widehat{t}_d}{\widehat{t}_a}$$
6. SDS is normalized SDS^* with μ and σ estimated from SNPs of similar frequency as candidate site

estimating $\widehat{t}_d$ and $\widehat{t}_a$ from $(d_i)_i$

- Consider one individual i, assume that the two tip branches leading to its alleles have lengths $t_{i,1}$ and $t_{i,2}$.
- let U and D be the distances upstream and downstream to the first singletons that are herterozygous in i, such that $d_i = U + D$.
- If we neglect recombination and assume a mutation rate μ, then U and D are independent and approximately exponentially distributed with rate $(t_{i,1} + t_{i,2}) \cdot \mu$.
- Thus, conditioned on $t_{i,1} + t_{i,2}$, $d_i = U + D$ is approx. Γ-distributed with shape parameter 2 and rate parameter $(t_{i,1} + t_{i,2}) \cdot \mu$.
- Also the distribution of $t_{i,j}$ for all i with ancestral type at the candidate site and that for all i with the derived type are assumed to be Γ distributions.
- Field et al. found this fitting well with simulation studies
- Γ distribution densities are relatively convenient for likelihood optimization
- Using an empirical branch length distributions can, to some extent, account for sequencing errors

References

[1] H.M.T. Vy, Y. Kim (2015) A Composite-Likelihood Method for Detecting Incomplete Selective Sweep from Population Genomic Data. *Genetics* **200**: 633-649

propose another statistic for detecting incomplete sweeps and compare it to iHS and nS_L.

11.1.2 Soft Sweeps

References

[1] A. Ferrer-Admetlla, M. Liang, T. Korneliussen, R. Nielsen (2014) On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Mol. Biol. Evol.* **31**: 1275–1291

Many sweeps detected in European *D. melanogaster* are due to out-of-Africa adaptations.

In humans, population size and short time span make it improbable that out-of-Africa adaptations were due to new mutations.

If selection acts on standing variation, that is, the advantageous allele existed before selection began, the sweep signal ("soft sweep") is less clear. Reason: several haplotypes that carry the advantageous allele and thus increase in frequency may differ in neighboring neutral SNPs.

Methods based on site-frequency spectra (like SweepFinder) or LD based statistics (like ω) may be appropriate to detect sweeps only if they are hard.

Ferrer-Admetlla *et al.* (2014) propose a haplotype-based statistic nS_L

n : number of sampled haplotypes (=sequences)

S_n : number of segregating sites

H : $n \times S_n$ matrix, where H_{ik} is indicator function that haplotype i carries the derived allele at site k

p_k : position of segregating site k in units of recombination distance

$H_{i,p_k:p_\ell}$: row vector for segregating sites between p_k and p_ℓ (may be empty).

$L_{ij}(x)$: $\max\{r-\ell \ : \ p_\ell < x < p_r, H_{i,p_\ell:p_r} = H_{j,p_\ell:p_r}\}$

$SL_D(k)$: $\frac{2\cdot\sum_{i<j} L_{ij}(p_k)\cdot H_{ik}\cdot H_{jk}}{(\sum_i H_{ik})\cdot((\sum_i H_{ik})-1)}$

$SL_A(k)$: $\frac{2\cdot\sum_{i<j} L_{ij}(p_k)\cdot(1-H_{ik})\cdot(1-H_{jk})}{(n-\sum_i H_{ik})\cdot(n-1-\sum_i H_{ik})}$

Note that $SL_D(k)$ is the average lengths of ranges that are identical by state around the derived allele and $SL_A(k)$ is the same for the ancestral allele.

$$nS_L(k) := \ln\left(\frac{SL_A(k)}{SL_D(k)}\right)$$

Variant: replace $nS_L(k)$ by standardized version

$$\frac{nS_L(k) - \mathbb{E}\left(nS_L(k)\right)}{\sigma\left(nS_L(k)\right)},$$

where $\mathbb{E}$ and standard deviation σ are conditioned on $\sum_i H_{ik}$, the number of sequences that have the derived allele at k. That is, estimate mean and sd from other loci with the same number of derived alleles in the sample.

11.1.3 Signals of sweeps and demography

Several studies compare different sweep statistics, for example:

References

[1] A. Vatsiou, E. Bazin, O.E. Gaggiotti (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular Ecology* **25**: 89–103

compare iHS, XP-CLR, nS_L and several other statistics for detecting sweeps in different demographic scenarios.

Vatsiou et al. (2016)

Scenarios of population structure/demography

- Four islands, migration between all
- Four islands, stepping-stone
- One population splits into two, later each of them splits into two, resulting in hierarchical island structure
- Stepping-stone with 52 populations and hard sweep in half of the populations of mutant that is detrimental in other half

In first three scenarios: hard or soft sweep in one of the four populations.

Vatsiou et al. (2016)

Strategies for setting thresholds ("p values" !?) for sweep signals:

in main text: 5% loci with highest values of the statistics.

limitation: this will always exist, also if there was no sweep signature by selection

in online supplement: use simulations of neutral loci according to same population structure/demography model to set thresholds

limitation: true population structure/demography usually unknown. Furthermore two different approaches are possible:

- threshold is value such that only 5% of the simulation show loci with such high (or higher) values
- threshold is value such that 5% of the loci from all simulation reach this (or higher) values. Such thresholds are *not* p values for the null hypothesis "everything neutral".

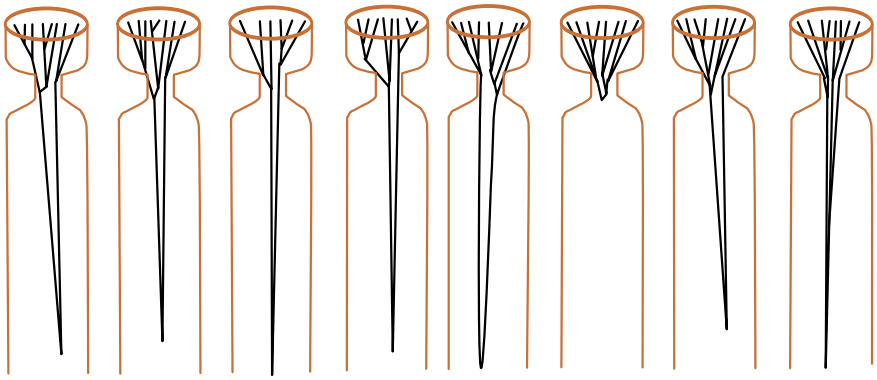
Vatsiou et al. (2016)

Main results:

- XL-PCR and hapFLK are best for performing soft sweeps
- all single statistics had problems when migration rates were higher or for weaker selection or incomplete sweeps
- improvements by combining statistics, e.g. XP-CLR with iHS or nS_L.

Problem: When population demography is unknown, how can we decide which patterns are significant evidence of selective sweeps?

naive idea: demography is the same for all loci outlier must be under selection



Bottlenecks can create outliers! (and branch length distribution is long-tailed anyways)

Possible approach

1. Infer population structure and demography from genomic data of putatively neutral loci
2. Use ω and/or SweepFinder statistic to search for evidence of selection in candidate loci (or genomic scans)
3. Simulate many datasets according to inferred demographic model, each neutral and with same amount of data as in candidate loci (or genomic scan). Account for ascertainment (*e.g.* by rejection sampling).
4. For each simulated neutral dataset i compute **maximum** value s_i of statistic.
5. Choose a threshold s for the statistic, such that only *e.g.* 5% of the s_i are larger.
6. Consider only loci as significant if their statistic value exceeds threshold s.

References

[1] P. Pavlidis, J.D. Jensen, W. Stephan (2010) Searching for Footprints of Positive Selection in Whole-Genome SNP Data from Nonequilibrium Populations. *Genetics* **185**: 907–922

propose and explore with simulated data the following procedure and apply it to *Drosophila melanogaster* data, where demography is known from previous studies:

1. Simulate loci according to known population demography with and without selection.
2. Use this simulated data to train a Support Vector Machine (SVM), which is a computational method for discriminant analysis (aka supervised learning). Input data are (slightly modified) ω and SweepFinder and other statistics based on combining the two, *e.g.* the distance of the positions where the two methods would locate the site under selection.
3. The trained SVM is then applied to descriminate between loci that have been affected by a selective sweep from neutral loci.

Still a problem: How to infer demography if we do not know which loci were really neutral?

Possible approach: Simultaneously infer population demography and fraction of loci affected by sweeps (and other kinds of selection).

11.2 Does the gene list make sense?

Once lists of genes that show evidence of selection have been identified, it is common practice to argue that the gene list makes sense by showing that certain GO categories are significantly overrepresented.

Pavlidis *et al.* (2012) explain why this may not make much sense.

References

[1] P. Pavlidis, J.D. Jensen, W. Stephan, A. Stamatakis (2012) A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Mol. Biol. Evol.* **29**: 3237–3248

Pavlidis *et al.* (2012) simulate completely neutral sequences mimicking the structure of *D. melanogaster* X chromosomes and scanned for sweeps.

False positives for sweeps tended to be clustered on the chromosomes.

As genes of related function can also be clustered on the chromosome, some GO categories were significantly overrepresented.

11.3 Balancing selection

Genetic signatures of balancing selection:

- enrichment of intermediate gene frequencies
- trans-specific polymorphisms
- increased frequency of polymorphic sites

References

[1] M. DeGiorgio, K.E. Lohmueller, R. Nielsen (2014) A model-based approach for identifying signatures of balancing selection in genetic data. *PLOS Genetics* **10**: e1004561

- Composite-likelihood ratio test for balancing selection
- Based on modelling the effect on linked neutral loci
- Software BALLET (*BAL*ancing selection *Lik*E*lihood* *T*est)
- Define and examine two different statistics T_1 and T_2
- Apply method to human data and detect previously found loci but also new ones

S : locus under strong balancing selection with two alleles A_1, A_2 with maintained frequencies x and $1-x$.

$\rho_i = 2Nr_i$, where N is the population size and r_i is the per-generation recombination rate btw. S and locus i.

data : n genomes from population, 1 from outgroup (*e.g.* chimp for humans)

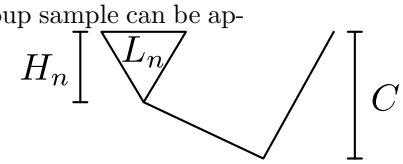
C : estimated genome-wide divergence time btw. in- and outgroup

$L_n(x, \rho_i)$ expectation value for total length of all branches of an ingroup genealogy for samples at site i, given n, x and ρ.

$H_n(x, \rho_i)$ Expected time since the root of the ingroup genealogy at site i, given n, x and ρ

The probability that a site that is segregating is polymorphic in within the ingroup sample can be approximated by

$$p_{n,\rho,x} \approx \frac{L_n(x,\rho)}{2C - H_n(x,\rho) + L_n(x,\rho)}$$



$H_n(x,\rho)$ and $L_n(x,\rho)$ can be computed by solving linear equation systems with variables $L_{k,n-k}$ and $H_{k,n-k}$, which are (for fixed x and ρ) the expected total ingroup genealogy length and height given a sample with k A_1-linked and $n-k$ A_2-linkes lineages.

With the total rate

$$\lambda_{k,n-k} = \frac{\binom{k}{2}}{x} + \frac{\binom{n-k}{2}}{1-x} + \frac{k\cdot(\theta_2+\rho_i x)\cdot(1-x)}{x} + \frac{(n-k)\cdot(\theta_1+\rho_i(1-x))\cdot x}{(1-x)}$$

we obtain, for example:

$$\begin{aligned} L_{k,n-k} &= \frac{n}{\lambda_{k,n-k}} + \frac{\binom{k}{2}\cdot L_{k-1,n-k}}{x\cdot\lambda_{k,n-k}} + \frac{\binom{n-k}{2}\cdot L_{k,n-k-1}}{(1-x)\cdot\lambda_{k,n-k}} \\ &\quad + \frac{k\cdot(\theta_2+\rho_i\cdot x)\cdot(1-x)\cdot L_{k-1,n-k+1}}{x\cdot\lambda_{k,n-k}} \\ &\quad + \frac{(n-k)\cdot(\theta_1+\rho_i\cdot(1-x))\cdot x\cdot L_{k+1,n-k-1}}{(1-x)\cdot\lambda_{k,n-k}} \end{aligned}$$

After numerically solving the equations systems for L_n and H_n and computing $p_{n,\rho,x}$ (approximately), the composite likelihood for all can be computed:

$$\mathcal{L}_1 = \max_x \prod_{i \in M} (1 - p_{n_i,\rho_i,x}) \cdot \prod_{i \in U} p_{n_i,\rho_i,x},$$

where M is the set of segregating sites that are monomorphic within the ingroup and U is the set of sites that are polymorphic in the ingroup.

For the neutral null model, we set

$$\mathcal{L}_0 = \prod_{i \in M} (1 - \tilde{p}_{n_i}) \cdot \prod_{i \in U} \tilde{p}_{n_i},$$

where $\tilde{p}_{n_i}$ is the proportion of loci that are in U among all segregating sites with the same sample size n_i.

$$T_1 := 2 \cdot \ln(\mathcal{L}_1 / \mathcal{L}_0)$$

The statistic T_2 is a refinement of T_1. For T_2 it is not only considered whether a site is polymorphic in the ingroup, but also how many of the sampled sequences show the derived allele.

In simulations T_2 works better than T_1 but T_1 may be advantageous when pooled sequencing is used.

Violations of Hardy-Weinberg equilibrium could also indicate balancing selection, but DeGiorgio *et al.* propose to filter out loci with strong violations of Hardy-Weinberg equilibrium because this may indicate mistakes in bioinformatic preprocessing (aligment/mapping/assembly).

Some of the things you should be able to explain

- Different types of sweeps
- Several statistics to detect various kinds of selection and the ideas behind these statistics
- problem to distinguish genomic footprints of selection from those of demography and why also bootlenecks can create outliers
- why overrepresentation analysis of GO can be misleading

12 Phasing genotypes and other applications of Li&Stephens' PAC approach

12.1 Li&Stephens' PAC approach

12.1.1 Estimating LD and recombination hotspots

Problems of models to estimate local recombination rates:

LAMARC etc. (ARG-based): not feasible for larger parts of the genome

Summary-statistics-based: lose too much information

some composite-likelihood methods: Hudson (2001), **Fearnhead, Donnelly (2002)**, McVean (2002) assume fixed recombination rate along the genome

References

[1] P. Fearnhead, P. Donnelly (2001) Estimating Recombination Rates From Population Genetic Data *Genetics* **159**: 1299–1318

Aim: Approximate the joint likelihood surface for the recombination rate and the mutation rate.

Model assumption: panmictic population, constant size N

$\theta = 4N\mu$

μ Mutation rate per generation and chromosome

$\rho = 4Nr$

r Recombination rate per generation and chromosome

Two different mutation modes:

- infinite-sites model
- at each site finitely many types with transition matrix $P_{\alpha\beta}$

$\mathcal{G}$ set of all ancestral histories (containing all mutations) that are consistent with the data D, such that $\forall_{G\in\mathcal{G}} \Pr(D|G) = 1$

Importance Sampling: If $G_1, \ldots, G_m$ are sampled independently according to some density q with $\mathcal{G} \subseteq \text{supp}(q)$, then

$$L(\rho, \theta) \approx \int_{\mathcal{G}} P(G|\theta, \rho) dG \approx \frac{1}{M} \sum_{i=1}^{M} \frac{P(G_i|\rho, \theta)}{q(G_i)}$$

What is a good proposal distribution q?

idea: Extend method of Stephens, Donnelly (2000) by recombination

H set of already sampled haplotypes

α potential type of $j+1$st sampled haplotype

$p(\alpha|H)$ will, like in Stephens, Donnelly (2000), be approximated to be used in importance sampling scheme

important: To use approximation q in importance sampling it must be possible to **sample** according to $q(.|H)$ and to **compute** $q(\alpha|H)$ for given α.

We specify $q(\alpha|H)$ by showing how to sample from it:

initialization: Let $x_1, \ldots, x_s$ be the segregating sites in the j chromosomes in H.

recombination: For $i = 1, \ldots, s-1$ there is a recombination event in the middle between x_i and x_{i+1} in α with probability

$$a_i := \frac{(x_{i+1} - x_i)\rho}{(x_{i+1} - x_i)\rho + j}.$$

Let k be the number of recombinations and $r = \{r_1, \ldots, r_{k+1}\}$ the resulting fragments.

imputation: For nonancestral sites in H impute types according to their frequency at that site in H.

mutations: Each r_i is simulated (independently of any r_j) according to Stephens, Donnelly's (2000) approximation $\widehat{\pi}$ for sequence data.

To compute $q(\alpha|H)$ for the correction in the Importance Sampling formula, we need to sum over all possible combinations of recombinations, imputations, and mutations that would lead to α.

This is done by dynamic programming: compute iteratively

$q_i(\alpha)$ probability that simulated type will coincide with α at first i loci.

$q_i(\alpha|\ell, t)$ as above, but conditioned that ith locus is a mutated copy of the ith locus in H_ℓ with Poisson(θt) mutations.

For this, the following approximation is used:

$$q_i(\alpha) \approx \sum_{m=1}^{k} \sum_{b=1}^{j} w_m q_i(\alpha|b, t_m/j)/j,$$

where $w_1, \ldots, w_k$ and $t_1, \ldots, t_k$ are the weights and points from the Gauß quadrature $\int_0^\infty e^{-t} f(t)dt \approx \sum_{m=1}^{k} w_m f(t_m)$.

To compute $q_i(\alpha|\ell, t)$ from previously computed $q_{i-1}(\alpha|\ell, t)$ and $q_{i-1}(\alpha)$ first compute the transition matrix for time t:

$$Q(t) = \exp\left(\theta t(P - I)/s\right),$$

where P as before is the transition matrix given that a mutation happens.

If H_ℓ is ancestral at locus i and has type β there, set

$$R := Q_{\beta,\alpha_i}(t),$$

and otherwise

$$R := (\pi_i Q(t))_{\alpha_i},$$

where π_i is the vector of proportions of types in H at position i. Then:

$$q_i(\alpha|\ell, t) = [(1 - a_{i-1}) \cdot q_{i-1}(\alpha|\ell, t) + a_{i-1} \cdot q_{i-1}(\alpha)] \cdot R.$$

Using these regression formulas in a dynamic-programming approach, $q(\alpha|H)$ can be computed and used to compute the proposal probability.

In the Importance Sampling step, the proposal probability is compared to the original ARG probability (ARG=ancestral recombination graph) and corrected accordingly to approximate the likelihood function for ρ and θ in the ARG model.

Li and Stephens' PAC approach, in contrast, replace the ARG model by a simpler model.

Li & Stephens' approach to analyze patterns of LD

References

[LS03] Na Li, Matthew Stephens (2003) Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data *Genetics* **165**

ideas:

- relate LD directly to underlying recombination process
- Sometimes, block-like LD structure is reported. True or artifact of LD mapping? Allow for both.
- consider all loci simultaneously, not pairwise
- should be compuationally tractable even for complete chromosomes

Li & Stephens' PAC approach

$h_1, h_2, \ldots, h_n$**:** haplotypes sampled from panmictic population with constant size and random mating

ρ**:** recombination parameter (may be a vector if recombination rate varies within the region of interest)

Product of Approximate Conditionals (PAC)

$$\Pr(h_1, \ldots, h_n|\rho) = \Pr(h_1) \cdot \Pr(h_2|h_1, \rho) \cdot \ldots \cdot \Pr(h_n|h_1, \ldots, h_{n-1}, \rho)$$

approximate $\Pr(h_k|h_1, \ldots, h_{k-1}, \rho)$ by simpler $q(h_k|h_1, \ldots, h_{k-1}, \rho)$.

Properties of $\Pr(h_k|h_1, \ldots, h_{k-1}, \rho)$

1. h_k is more likely to match another haplotype if the latter is frequent among $h_1, h_2, \ldots, h_{k-1}$
2. the probability of seeing a novel haplotype decreases as k increases
3. the probability of seeing a novel haplotype increases with $\theta = 4N_e\mu$.
4. if a new haplotype does not exactly match any previous one, it will differ from one of those only by a small number of mutations.
5. effect of recombination: the next haplotype will be composed by segments which are similar to segments in previously sampled haplotypes. These segments tend to be longer if recombination rates are low.

Assume the sampled haplotypes $h_1, h_2, \ldots, h_n$ are typed at S biallelic loci (e.g. SNPs).

$$q(h_1) = \left(\frac{1}{2}\right)^S$$

For the definition of $q(h_{k+1}|h_1, h_2, \ldots, h_k)$ let $X_i := j$ if at the i-th locus, the closest relative of h_{k+1} among $h_1, \ldots, h_k$ is h_j.

d_i distance between loci i and $i+1$

c_i recombination rate between loci i and $i+1$ per site and per generation

$\rho_i = 4N_e c_i$

The simplifying assuption is then that $X_1, X_2, \ldots, X_S$ is a Markov chain on $\{1, \ldots, k\}$ with $Pr(X_1 = j) = 1/k$ and

$$\Pr(X_{i+1} = j|X_i = \ell) = \begin{cases} (1 - e^{-\rho_i d_i/k})/k & \text{if} \quad j \neq \ell \\ e^{-\rho_i d_i/k} + (1 - e^{-\rho_i d_i/k})/k & \text{if} \quad j = \ell \end{cases}$$

Mutations

For SNP data we assume that each locus is hit by one mutation, such that

$$\widetilde{\theta} := 1 \Big/ \sum_{m=1}^{n-1} \frac{1}{m}$$

is assumed to be the corrected rate of mutations per SNP site. Note that this does not exclude double hits (just some bias if double hits are frequent.)

Then, with probability $\frac{k}{k+\widetilde{\theta}}$ the copy has the same type as the original

and with probability $\frac{\widetilde{\theta}}{(k+\widetilde{\theta})}$ the haplotype has the other of the two possible alleles.

Compute $q(h_{k+1}|h_1, \ldots, h_k)$ by HMM forward algo:

$h_{k+1,\leq j} := (h_{k+1,1}, \ldots, h_{k+1,j}) :=$ types of the first j sites in h_{k+1}

$\alpha_j(x) := \Pr(h_{k+1,\leq j}, X_j = x | h_1, \ldots, h_k)$

(note that with mutations any $X_1, \ldots, X_S$ can emit h_k.)
Then,

$$q(h_{k+1}|h_1, \ldots, h_k) = \sum_{x=1}^{k} \alpha_S(x).$$

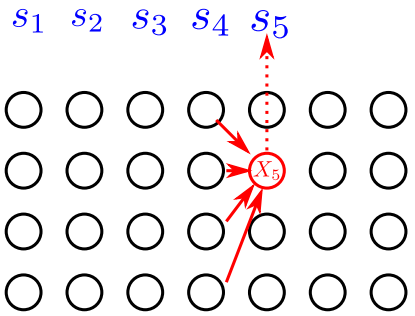
"dynamic programming": we can compute all $\alpha_j(x)$ by the recursion

$$\begin{aligned}
\alpha_{j+1}(x) &= \Pr(h_{k+1,j+1}|X_{j+1} = x, h_1, \ldots, h_k) \cdot \sum_{x'=1}^{k} \alpha_j(x') \cdot \Pr(X_{j+1} = x | X_j = x') \\
&= \Pr(h_{k+1,j+1}|X_{j+1} = x, h_{x,j+1}) \cdot \\
&\qquad \left(e^{-\rho_j d_j/k} \cdot \alpha_j(x) + \left(1 - e^{-\rho_j d_j/k}\right) \cdot \frac{1}{k} \sum_{x'=1}^{k} \alpha_j(x') \right)
\end{aligned}$$

Remember the usual HMM algorithms!

HMM forward algorithm

$f_i(x) = \Pr(S_1 = s_1, S_2 = s_2, \ldots, S_i = s_i, X_i = x)$



$$f_i(x) = \sum_{y \in \mathcal{M}} f_{i-1}(y) \cdot P_{y \to x} \cdot e_x(s_i)$$

$$\Pr(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n) = \sum_{x \in \mathcal{M}} f_n(x)$$

Li & Stephens' bias correction

Simulations show that estimations of ρ based on q are biased.

For bias-correction replace ρ_j in the computation of $\Pr(X_{j+1} = x' | X_j = x)$ by

$$\rho_j \cdot e^{a + b \log_{10} \rho_j},$$

where a and b are fitted to simulated data, taking the numbers of haplotypes and segregating sites into account.

Models for ρ considered by Li and Stephens

1. constant ρ
2. single-hotspot model
3. all recombination rates $\rho_1, \rho_2, \ldots, \rho_{S-1}$ may differ

Software by Matthew Stephens using PAC: Hotspotter, PHASE

References

[YS16] Yun S. Song (2016) Na Li and Matthew Stephens on Modeling Linkage Disequilibrium *Genetics* **203**(3): 1005–1006 `https://doi.org/10.1534/genetics.116.191817`

12.1.2 Population splitting and recombination

References

[DPC09] D. Davison, J.K. Pritchard, G. Coop (2009) An approximate likelihood for genetic data under a model with recombination and population splitting. *Theoretical Population Biology* **75**:331-345

- two populations split G generations ago, $F = G/(2N)$.
- no ongoing geneflow
- for simplicity: assume that both populations and the ancestral population have size N
- Copying occurs in daughter population ($S = d$) and in ancestral population ($S = a$)

Again, the PAC approach is used and we have to approximate the probability of a haplotype $h_{k_1+k_2+1}$ given already sampled haplotype $h_1, \ldots, h_{k_1+k_2}$, of which k_1 were sampled on island 1 and k_2 on island 2.

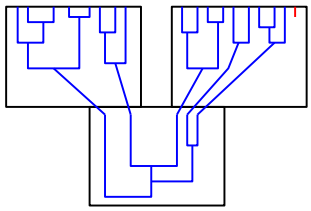
Let $z_i \in \{1, 2\}$ indicate the island where h_i was sampled, $z_* := z_{k_1+k_2+1}$, and X_ℓ indicates the h_i that is the closest relative to $h_{k_1+k_2+1}$ at site ℓ.

What we need to specify as model assumptions:

1. Probability of hidden copying states (S_ℓ, X_ℓ) at a single site ℓ.
2. Probability of new allelic state conditioned on the state of the copied allele and the level S_ℓ.
3. Transition probabilities between the hidden copying state at adjacent states (in case of linked loci).

Point 1: $\Pr(X_\ell = i | S_\ell = d)$**, unlinked case**

In the case of unlinked sites, we obtain



$$\begin{aligned}
\Pr(X_\ell = i | S_\ell = d) &= \begin{cases} \frac{1}{k_{z_*}} & \text{if } z_* = z_i \\ 0 & \text{else} \end{cases} \\
\Pr(X_\ell = i | S_\ell = a) &= \mathbb{E}\left(\frac{J_{z_i}}{J_1 + J_2}\right) \cdot \frac{1}{k_{z_i}},
\end{aligned}$$

Where k_{z_*} is the no. of lineages sampled from pop. z_* so far and J_i is the number of ancestral lineages that enter the ancestral pop. from pop. i. To compute the expectation first compute for all $j_i < k_i$ the probabilities that k_i lineages coalesce down to j_i lineages in G generations. These values are also needed to compute $\Pr(S_\ell = d)$.

Point 2: Mutation probability; unlinked case

Simplification: For time T_{coal} of coalescence use expectation

$$t_s = \mathbb{E}(T_{coal} | S, k_1, k_2, F)$$

Then:

$$\begin{aligned}
u(h_{k_1+k_2+1} | h_i, s) &= \Pr(h_{k_1+k_2+1} | S_\ell = s, X_\ell = i, k_1, k_2, F) \\
&= \begin{cases} 1 - e^{-\theta t_s} & \text{if} \quad h_{k_1+k_2+1} \neq h_i \\ e^{-\theta t_s} & \text{if} \quad h_{k_1+k_2+1} = h_i \end{cases},
\end{aligned}$$

where θ is corrected for using only polymorphic sites as in Li&Stephens.

Thus, we approximate $\Pr(h_{k_1+k_2+1} | h_1, \ldots, h_{k_1+k_2})$ by

$$\sum_{s \in \{a,d\}} p(S = s) \cdot \sum_{i=1}^{k_1+k_2} u(h_{k_1+k_2+1} | h_i, s) \cdot p(X = i | S = s)$$

Point 3: now for the case of (loosly) linked loci

- Now $(S_1, X_1), \ldots, (S_L, X_L)$ are not independent.
- Simplify applying Markov model, such that HMM algorithms are applicable
- Computing transition probabilities

$$p(S_{\ell+1} = s', X_{\ell+1} = i' | S_\ell = s, X_\ell = i)$$
$$= p(S_{\ell+1} = s' | S_\ell = s) \cdot p(X_{\ell+1} = i' | S_{\ell+1} = s') + \delta_{ii'}\delta_{ss'} \cdot p(NR | S_\ell),$$

  where NR stands for "no recombination" is tricky, several simplifying approximation are applied, e.g.:

- if A is the event that a recombination happens on the new lineage in the daughter population, then

$$p(S_{\ell+1} = d | S_\ell = a) = p(A) \cdot p(S = d | A) \quad (1)$$
$$\approx \left(1 - e^{-\rho_\ell F}\right) \cdot p(S = d) \quad (2)$$

- Simulation study
  - works well in case of unlinked loci
  - estimates of F biased for version with linked loci
  - not well understood where this bias comes from, but suggest a bias correction
- Discuss how method could be extended to models with gene flow

12.1.3 Diversifying selection and recombination

References

[WM06] D.J. Wilson, G. McVean (2006) Estimating diversifying selection and functional constraints in the presence of recombination *Genetics* **172**:1411–1425

Apply Bayesian variant of PAC to infer from population genetic data which regions are under diversifying selection and which are under purifying selection.

$\omega = d_N/d_S$ fraction of rates of nonsynonymous vs. synonymous mutations

diversifying selection corrsponds to large ω and

purifying selection to small ω.

Software: omegaMap `http://www.danielwilson.me.uk/omegaMap.html`

NY98 Codon mutation model

Nielsen and Yang (1998)

Mutation rate $q_{ij} = \pi_j \cdot \mu_{ij}$, where i and j are codons, $i \neq j$ or one of then is an insertion/deletion (indel), π_j is the frequency of j and

$$\mu_{ij} = \begin{cases} 1 & \text{for synonymous transversion} \\ \kappa & \text{for synonymous transition} \\ \omega & \text{for nonsynonymous transversion} \\ \omega\kappa & \text{for nonsynonymous transition} \\ \omega\phi & \text{if exactly one of } i \text{ and } j \text{ is an indel} \\ 0 & \text{otherwise} \end{cases}$$

(original NY98 is without indels)

Model for ω along the gene:

There are B transition points $s_1, \ldots, s_B$, such that $\forall_j : \omega$ is constantly ω_j between s_j and s_{j+1}.

p_ω Probabilty of transition point between two codons.

ω_j are independent of each other and have prior $\exp(\lambda)$

Similar model for change of recombination rate ρ; independent of ω configuration.

H haplotypes sampled from population

Θ model parameters, including all ω_j and ρ_j and change points.

$MCMC$ sample parameter values according to

$$P(\Theta|H) \propto P(H|\Theta) \cdot P(\Theta)$$

PAC with HMM forward algorithm is applied to approximate $P(H|\Theta)$ via PAC approximations of

$$p(H_{k+1}|H_1, H_2, \ldots, H_k, \Theta)$$

MCMC moves

To propose e.g. μ' as a replacement of current μ, choose $U \sim \text{unif}(-1, 1)$, und set $\mu' = \mu \exp(U)$. Accept or reject with MH step.

Same for κ, ω_j, ρ_j.

MCMC step to shift block change point.

Reversible Jump steps to update blocks:

- split block
- merge block

$\lambda, \phi, p_\omega, p_\rho$ are specified by user to specify prior.

Application example

79 alleles of *porB* locus of *Neisseria meningitidis*

permutation test shows significant correlation of LD and distance between sites. $\Rightarrow$ Phylogenetic methods not appropriate.

Found four sections in the gene where high ω values are probable, indicating diversifying selection, whereas almost everywhere else, $\omega < 1$, indicating purifying selection.

Indeed, regions with large ω are loops exposed to immune system of host, such that diversifying selection is plausible, and other regions form beta sheet barrel, which explains functional constraints. (nice picture in paper!)

Model without recombination leads to different results, but

Model Criticism via posterior predictive P-value shows that model without recombination fits the data poorly. This means, if D is some statistic of the data and $D_{H'}$ is the statistic for a dataset H' simulated under the PAC model used for inference, then the

posterior predictive P-value is:

$$p = \int P(D_{H'} \geq D_H|\Theta, H)P(\Theta|H)d\Theta \approx \frac{1}{M}\sum_{i=1}^{M} I(D_{H'_i} \geq D_H)$$

Wilson and McVean perform simulation studies for several conditions to assess how well their method works.

This is very important for heuristic approaches like PAC because it is otherwise not clear how accurate these methods are even if sampling from some approximate posterior approximate confidence intervals are computed.

Some of what you should be able to explain

- Basic approach of Li&Stephen's PAC
- Simplifying assumptions in PAC
- How PAC fits into HMM methodology
- How PAC can be applied to infer
  - local recombination rates
  - population demography
  - sections in genes under different kinds of selection

12.1.4 Excursus: Stephens and Donnelly's Importance Sampling

References

[SD00] M. Stephens, P. Donnelly (2000) Inference in molecular population genetics *J. R. Statist. Soc. B* **62(4)**:605–655

improved Griffiths und Tavaré's Importance sampling scheme for the case of the classical (unstructured) coalescent.

A history $\mathcal{H}$ is a sequence $(H_{-m}, H_{-(m-1)}, \ldots, H_0)$, where

H_{-i} is the unorded list of types of the uncestral lineages, i events (mutations and coalescent events) before present. Thus,

H_0 is the sampled data.

$H_{i-1} = H_i - \alpha + \beta$ stands for a mutation from α to β (back into past; note that $i < 0$), and

$H_{i-1} = H_i - \alpha$ for a coalescence of two lineages of type α.

$P_{\alpha\beta}$ probability that an α that is hit by a mutation becomes a β.

n_α number of lineages in H_{i-1} of type α,

$n = \sum_\alpha n_\alpha$,

$$\Pr_\theta(H_i \mid H_{i-1}) = \begin{cases} \frac{n_\alpha}{n} \cdot \frac{\theta}{n-1+\theta} P_{\alpha\beta} & \text{if} \quad H_i = H_{i-1} - \alpha + \beta \\ \frac{n_\alpha}{n} \cdot \frac{n-1}{n-1+\theta} & \text{if} \quad H_i = H_{i-1} + \alpha \\ 0 & \text{otherwise} \end{cases}$$

$\pi_\theta(.)$: distribution of genotype vector A_n in a sample of size n. Note that A_n is an ordered list.
n_α: number of α in H_0

$$\pi_\theta(A_n \mid \mathcal{H}) = \pi_\theta(A_n \mid H_0) = \begin{cases} \left(\prod_\alpha n_\alpha!\right)/n! & \text{if } H_0 \text{ compatible with} A_n \\ 0 & \text{otherwise} \end{cases}$$

If histories $\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(M)}$ are generated independently according to the proposal distribution $Q_{\theta_0}(.)$, the importance sampling formula implies:

$$L(\theta) \approx \frac{1}{M} \sum_{i=1}^{M} \pi_\theta(A_n \mid \mathcal{H}^{(i)}) \cdot \frac{P_\theta(\mathcal{H}^{(i)})}{Q_{\theta_0}(\mathcal{H}^{(i)})}$$

E.g. with the proposal distribution Q_θ^{GT} of Griffiths and Tavaré, for given H_0 the histories $H_{-1}, H_{-2}, \ldots$ are generated by a Markov chain with $q_\theta(H_{i-1} \mid H_i) \propto p_\theta(H_i \mid H_{i-1})$.

Let $\mathcal{M}$ be the class of proposal distributions, for which $H_{-1}, H_{-2}, \ldots$ is Markovian with start in H_0 and

$$\begin{aligned} \text{supp}\{q_\theta(. \mid H_i)\} &:= \{H_{i-1} \ : \ q_\theta(H_{i-1} \mid H_i) > 0\} \\ &= \{H_{i-1} \ : \ p_\theta(H_i \mid H_{i-1}) > 0\}. \end{aligned}$$

Optimal would be $Q^*_\theta(\mathcal{H}) = P_\theta(\mathcal{H} \mid A_n)$, because

$$\pi_\theta(A_n \mid \mathcal{H}) \frac{P_\theta(\mathcal{H})}{Q^*_\theta(\mathcal{H})} = \frac{P_\theta(\mathcal{H} \cap A_n)}{P_\theta(\mathcal{H} \mid A_n)} = \pi_\theta(A_n) = L(\theta)$$

Theorem 3 *Let $\pi_\theta(\alpha \mid A_n) = \frac{\pi_\theta((A_n, \alpha))}{\pi_\theta(A_n)}$ be the conditioned probability that the $n+1$-st allele sampled from the population is of type α, given that the first n types are given by A_n. The optimal proposal distribution Q^*_θ belongs to $\mathcal{M}$ and is defined by*

$$q^*_\theta(H_{i-1} \mid H_i) = \begin{cases} \frac{\theta \cdot n_\alpha}{n \cdot (n-1+\theta)} \frac{\pi(\beta \mid H_i - \alpha)}{\pi(\alpha \mid H_i - \alpha)} P_{\beta\alpha} & \text{für} \quad H_{i-1} = H_i - \alpha + \beta \\ \\ \frac{n_\alpha \cdot (n_\alpha - 1)}{n \cdot (n-1+\theta)} \frac{1}{\pi(\alpha \mid H_i - \alpha)} & \text{für} \quad H_{i-1} = H_i - \alpha \end{cases}$$

Proof

Consider the case $H_{i-1} = H_i - \alpha + \beta$

Let $a_k(t)$ be the type of lineage k at time t. Assume $\delta > 0$ and let Y_m be the event that in the last δ time units a mutation from $a_k(t-\delta) = \beta$ to $a_k(t) = \alpha$ occurred.

We obtain:

$$\begin{aligned} &\Pr\{Y_m \cap A_k(t-\delta) = (\alpha_1, \ldots, \alpha_{k-1}, \beta) \mid A_k(t) = (\alpha_1, \ldots, \alpha_{k-1}, \alpha)\} \\ &\qquad = \frac{\pi(\alpha_1, \ldots, \alpha_{k-1}, \beta) \cdot \delta \cdot \theta \cdot P_{\beta\alpha}/2}{\pi(\alpha_1, \ldots, \alpha_{k-1}, \alpha)} + o(\delta) \\ &\qquad = \delta \cdot \theta \cdot \frac{\pi(\beta \mid A_k - \alpha)}{2\pi(\alpha \mid A_k - \alpha)} \cdot P_{\beta\alpha} + o(\delta) \end{aligned}$$

This implies the result if we let δ go to 0, multiply by n_α (as instead of α_k any other α could be affected, and H_i is unordered) and divide by the total rate.

The proof for $H_{i-1} = H_i - \alpha$ is analogous. □

But:

In general, $\pi(\alpha \mid A_n)$ are hard to compute and we cannot use Q^*_θ.

Ansatz: If $\pi(\alpha \mid A_n)$ cannot be calculated, approximate it and use the approximations in the formulas in the theorem.

Definition 1

$$\widehat{\pi}(\beta \mid A_n) := \sum_{\alpha \in E} \sum_{m=0}^{\infty} \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta}\right)^m \cdot \frac{n}{n+\theta} (P^m)_{\alpha\beta}.$$

This probability distribution can be approximated as follows: Choose a purely randomly individual from A_n and mutate it according to P geometrically often with parameter $\frac{\theta}{n+\theta}$.

properties of $\widehat{\pi}$:

(a) For parent-independent mutation: $\widehat{\pi}(. \mid A_n) = \pi(. \mid A_n)$.

(b) For reversible P with $n = 1$: $\widehat{\pi}(. \mid A_n) = \pi(. \mid A_n)$.

(c) The distribution $\widehat{\pi}(. \mid A_n)$ fulfills

$$\widehat{\pi}(\beta \mid A_n) = \sum_\alpha \frac{n_\alpha}{n} M^{(n)}_{\alpha\beta} \qquad (*)$$

for suitable $M^{(n)}$. Thus, it can be simulated by drawing a random lineage and draw the type according to a distribution that depends only on n and on the type of the drawn lineage. (In the case of $\widehat{\pi}$ holds $M^{(n)} = (1-\lambda_n)(I - \lambda_n P)^{-1}$ with $\lambda_n = \frac{\theta}{n+\theta}$.)

more properties of $\widehat{\pi}$:

(d) $\widehat{\pi}$ is the only distribution that fulfills $(*)$ and (b) and

$$\widehat{\pi}(\beta \mid A_n) = \sum_\alpha \widehat{\pi}(\alpha \mid A_n) \cdot \widehat{\pi}(\beta \mid (A_n, \alpha)) \qquad (**)$$

This means: Given the first n sampled alleles, the $n+1$st has the same distribution as the $n+2$nd.

(e) $\widehat{\pi}(. \mid A_n)$ is the stationary distribution of a Markov chain with transition matrix $T_{\alpha\beta} = \frac{\theta}{n+\theta} P_{\alpha\beta} + \frac{n_\alpha}{n+\theta}$

Proofs

(a) For parent-independent mutation $P_{\alpha\beta} = P_\beta$ holds $P = P^m$ and thus:

$$\pi(\beta \mid A_n) = \frac{n_\beta + \theta P_\beta}{n+\theta} = \widehat{\pi}(\beta \mid A_n)$$

(b) Let X and Y be the types of the leaves, R the type of the root, m_1 the number of mutations between R und X and m_2 that of the mutation between R and Y. Then:

$$\begin{aligned} \Pr(Y = \beta \mid R = \gamma) &= (P^{m_2})_{\gamma\beta} \\ \Pr(R = \beta \mid X = \alpha) &= (P^{m_1})_{\alpha\beta} \\ \Pr(Y = \beta \mid X = \alpha) &= (P^{m_1+m_2})_{\alpha\beta} \end{aligned}$$

The total number of mutations between X and Y is geometrically distributed with parameter $\frac{\theta}{1+\theta}$

(c)

$$\begin{aligned} \widehat{\pi}(\beta \mid A_n) &= \sum_\alpha \sum_{m=0}^\infty \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta}\right)^m \frac{n}{n+\theta} \left(P^m\right)_{\alpha\beta} \\ &= \sum_\alpha \sum_{m=0}^\infty \frac{n_\alpha}{n} (1-\lambda_n) \left[(\lambda_n P)^m\right]_{\alpha\beta} \\ &= \sum_\alpha \frac{n_\alpha}{n} (1-\lambda_n) \left[(I - \lambda_n P)^{-1}\right]_{\alpha\beta} \end{aligned}$$

The last equation follows from the geometric sum formula of matrices.
$\sum_{m=0}^\infty M^m = (I - M)^{-1}$.

proof of (d)

Let $\widetilde{\pi}(\beta \mid A_n) = \sum_\alpha \frac{n_\alpha}{n} M^{(n)}_{\alpha\beta}$ for some $M^{(\cdot)}_{\cdot\cdot}$ fulfilling $(**)$:

$$\begin{aligned}
\left(\frac{n_\alpha}{n}, \frac{n_\beta}{n}, \ldots, \frac{n_\gamma}{n}\right) \cdot M^{(n)}_{\alpha\beta} &= \widetilde{\pi}(\beta \mid A_n) \\
&= \sum_\alpha \widetilde{\pi}(\alpha \mid A_n)\widetilde{\pi}(\beta \mid (A_n, \alpha)) \\
&= \sum_\alpha \sum_\gamma \frac{n_\gamma}{n} M^{(n)}_{\gamma\alpha} \cdot \sum_\xi \frac{n_\xi + \delta_{\alpha\xi}}{n+1} M^{(n+1)}_{\xi\beta} \\
&= \sum_\gamma \sum_\xi \left(\frac{n_\gamma}{n} \frac{n_\xi}{(n+1)} M^{(n+1)}_{\xi\beta} + \frac{n_\gamma}{n \cdot (n+1)} M^{(n)}_{\gamma\xi} M^{(n+1)}_{\xi\beta} \right) \\
&\qquad \text{(because } \forall\gamma : \sum_\alpha M^{(n)}_{\gamma\alpha} = 1) \\
&= \frac{1}{n+1} \left[\left(\frac{n_\alpha}{n}, \ldots, \frac{n_\gamma}{n}\right) \cdot \left(n M^{(n+1)} + M^{(n)} M^{(n+1)} \right) \right]_\beta
\end{aligned}$$

proof of (d) continued

As this holds for all vectors $\frac{n_\alpha}{n}, \ldots, \frac{n_\gamma}{n}$, we conclude:

$$(n+1)M^{(n)} = n \cdot M^{(n+1)} + M^{(n)} \cdot M^{(n+1)}$$

From this recursion and the initial value $M^{(1)} = (1-\lambda_1)(I - \lambda_1 P)^{-1}$ set by (b), follows $M^{(n)} = (1-\lambda_n)(I-\lambda_n P)^{-1}$. This implies $\widetilde{\pi} = \widehat{\pi}$.

□

For the more complicated proof of (e), see Stephens und Donnelly (2000).

We define the proposal distribution Q^{SD}_θ with $\widehat{q}$ like Q^*_θ with q by replacing π by the approximation $\widehat{\pi}$.

Theorem 4

$$\sum_H \widehat{q}_\theta(H \mid H_i) = 1$$

and $\widehat{q}_\theta(. \mid H_i)$ can be simulated as follows:

1. *Choose a purely random $\alpha \in H_i$.*
2. *For all β compute $\widehat{\pi}(\beta \mid H_i - \alpha)$*
3. $$H_{i-1} := \begin{cases} H_i - \alpha + \beta & \textit{with probability} \quad \propto \theta\widehat{\pi}(\beta \mid H_i - \alpha) \cdot P_{\beta\alpha} \\ H_i - \alpha & \textit{with probability} \quad \propto n_\alpha - 1 \end{cases}$$

Thus, $\widehat{\pi}(\beta \mid H_i - \alpha)$ must be computed only for a few pairs (α, β). First sample α and then decide whether it mutated to beta β or coalesces with another α. It is efficient to compute $\widehat{\pi}(\beta \mid H_i - \alpha)$ and to simulte Q^{SD}_θ.

Proof:

The probability that a mutation of a type α is involved, is

$$p_m(\alpha) = \frac{1}{n(n-1+\theta)} \sum_\beta \frac{\theta}{2} n_\alpha \frac{\widehat{\pi}(\beta \mid H_i - \alpha)}{\widehat{\pi}(\alpha \mid H_i - \alpha)} P_{\beta\alpha}.$$

The probability that two lineages of type α coalesce, is:

$$p_c(\alpha) = \frac{n_\alpha(n_\alpha - 1)}{n(n-1+\theta)} \cdot \frac{1}{\widehat{\pi}(\alpha \mid H_i - \alpha)}$$

This implies $p_m(\alpha) + p_c(\alpha) = 1$. □

What to do with sequence data???

For nucleotide (or protein) sequences of length ℓ there are 4^ℓ (or 20^ℓ) different possible genotypes $\alpha = (\alpha_1, \ldots, \alpha_\ell)$, and the transition matrix $(P_{\alpha\beta})_{\alpha\beta}$ could be very large.

$\theta/2$: mutation rate per site.

For $\widehat{\pi}(. \mid A_n)$ draw a geometrically distributed number m of mutations with parameter $\frac{\ell\theta}{n+\ell\theta}$ and spread them randomly on the sites.

Equivalent: draw exp(1)-distributed time t and then for each site i a Poisson$(t\theta/n)$ distributed number m_i of mutations. This implies

$$\widehat{\pi}(\beta \mid A_n) = \sum_{\alpha \in A_n)} \frac{n_\alpha}{n} \int \exp(-t) F^{(\theta,t,n)}_{\alpha_1\beta_1} \cdots F^{(\theta,t,n)}_{\alpha_\ell\beta_\ell} \quad dt$$

with

$$F^{(\theta,t,n)}_{\alpha_i\beta_i} = \sum_{m=0}^{\infty} \frac{(\theta t/n)^m}{m!} \exp(-\theta t/n)(P^m)_{\alpha_i\beta_i}.$$

Stephens and Donnelly suggest to approximate the integral with Gauß quadrature (siehe Press et al. (1992)) to obtain

$$\widehat{\pi}(\beta \mid A_n) = \sum_{\alpha \in A_n)} \sum_{i=1}^{s} \frac{n_\alpha}{n} w_i F^{(\theta,t_i,n)}_{\alpha_1\beta_1} \cdots F^{(\theta,t_i,n)}_{\alpha_\ell\beta_\ell}$$

for certain s, w_i and t_i. The $F^{(\theta,t,n)}_{\alpha_i\beta_i} = \sum_{m=0}^{\infty} \ldots$ can be approximated by finite sums.

12.2 Phasing

Why phasing?

Many sequence datasets from diploid (polyploid) organisms are *unphased*. For example, it is known that some individual has an A and a T at one locus, and a G and a C at another locus on the same chromosome, but not wheter the A is on the same haplotype (chromosome copy) as the C or as the G.

—[A,T]——[C,G]— $\Rightarrow$ —A——C— / —T——G— or —A——G— / —T——C— ?

Estimating this ("phasing the data") can be important, e.g. because Linkage Disequilibrium (LD) is informative about

- population structure
- epistatis
- selective sweeps
- whether a gene locus is associated with a trait of interest or just phsically linked to a relevant locus

Clark's Algorithm

- parsimonious approach to minimize the total number of haplotype classes observed in the sample
- greedy algorithm
- starting with individuals that are homozygous at all loci or at all up to one

$$[0,0],[0,1],[0,0],[1,1] \Rightarrow \begin{matrix} \text{–0—1—0—1–} \\ \text{–0—0—0—1–} \end{matrix}$$

- successively searches individuals that can be phased such that one or both haplotypes is identical to already inferred one.

$$[0,1],[0,1],[0,1],[1,1] \Rightarrow \begin{matrix} \text{–0—1—0—1–} \\ \text{–1—0—1—1–} \end{matrix}$$

- final result depends on input order

12.2.1 Excoffier and Slatkin's EM algorithm

References

[ES95] L. Excoffier, M. Slatkin (1995) Maximum-Likelihood estimation of molecular hyplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12(5)**: 921–927

"phenotype": multilocus genotype with unknown phase, e.g.

genotype= $\begin{matrix}\text{–0—1—0—1–}\\ \text{–0—0—1—1–}\end{matrix}$ $\Rightarrow$ phenotype=$[0,0],[0,1],[0,1],[1,1]$ (unordered pair) (unordered pairs)

P_i: phenotype probability
n_i: absolute frequency of phenotype i in sample, $n = \sum_i n_i$

$$\Pr(\text{sample}|P_1, P_2, \ldots, P_m) = \frac{n!}{n_1! \cdot n_2! \cdots n_m!} \cdot P_1^{n_1} \cdot P_2^{n_2} \cdots P_m^{n_m}$$

one aim:
estimate population frequencies $p_1, p_2, \ldots$ of haplotype classes $h_1, h_2, \ldots$.

Expectation Maximization (EM) algorithm
Iterate E step and M step:

E step Use current estimates of $p_1, p_2, \ldots$ to compute expected frequencies $f_{k\ell}$ of all genotypes (k, ℓ) (with $k \leq \ell$) **in the sample**, given the sampled phenotypes. For this, let $I_{k\ell,j}$ be the indicator function that $[k, \ell]$ leads to phenotype j (i.e. $I_{kl,j} = 1$ in this case and 0 otherwise), and $\delta_{k\ell}$ be the indicator function of $k = \ell$. Then

$$P_j = \sum_k \sum_\ell p_k p_l \cdot I_{k\ell,j}$$

and

$$f_{k\ell} = \sum_j I_{k\ell,j} \cdot \frac{n_j}{n} \cdot \frac{p_k p_\ell}{P_j} \cdot 2 \cdot \left(\frac{1}{2}\right)^{\delta_{k\ell}}.$$

M step Use expected genotype frequencies in sample to estimate haplotype class probabilities p_i (=frequencies in population).

$$p_i = f_{ii} + \frac{1}{2} \cdot \left(\sum_{k=1}^{i-1} f_{ki} + \sum_{k=i+1}^{\ldots} f_{ik} \right)$$

Excoffier and Slatkin use Fisher Information to estimate variance of the estimators, and use estimated p_i to infer haplotypes.

12.2.2 Excursus: EM algorithm

EM algorithm in general

References

[DLR77] Dempster, A.P., Laird. N.M., Rubin, D.B. (1977) Maximum-Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39 (1)**: 1–38

X observed data

U unobserved data

θ parameter to be estimated

ℓ log likelihood

$$\ell(\theta; x, u) = \log P_\theta(X = x, U = u)$$

P Probability or probability density

$$\begin{array}{rcl}
\ell(\theta; u \mid x) & := & \log P_\theta(U = u \mid X = x) \\
\Rightarrow \quad \ell(\theta; x, u) & = & \ell(\theta; u \mid x) + \ell(\theta; x) \\
U \text{ unobserved} & \Rightarrow & \ell(\theta; x, U) \text{ is a random variable} \\
Q_{\theta'}(\theta) & := & \mathbb{E}_{\theta'}\left(\ell(\theta; x, U)|x\right) \\
& = & \displaystyle\sum_u P_{\theta'}(u|x) \cdot \ell(\theta; x, u) \\
& & \left(\text{or } \displaystyle\int P_{\theta'}(u|x) \cdot \ell(\theta; x, u)\ du\right) \\
C_{\theta'}(\theta) & = & \mathbb{E}_{\theta'}\left(\ell(\theta; U|x)|x\right) \\
\ell(\theta) & := & \ell(\theta; x) \ = \ Q_{\theta'}(\theta) - C_{\theta'}(\theta)
\end{array}$$

To estimate θ iterate the following steps:

E step with current estimate θ' compute the function $Q_{\theta'} : \theta \mapsto Q_{\theta'}(\theta)$

M step

$$\theta^{new} := \arg\max_\theta Q_{\theta'}(\theta)$$

Iterate E step with θ' replaced by θ^{new}.

Note that from

$$\ell(\theta) \ = \ Q_{\theta'}(\theta) - C_{\theta'}(\theta)$$

follows that

$$Q_{\theta'}(\theta^{new}) \geq Q_{\theta'}(\theta')$$

implies

$$\begin{array}{rcl}
\ell(\theta^{new}) - \ell(\theta') & \geq & C_{\theta'}(\theta') - C_{\theta'}(\theta^{new}) \\
& = & \displaystyle\int P_{\theta'}(U = u|x) \cdot \log \frac{P_{\theta'}(U = u|x)}{P_{\theta^{new}}(U = u|x)} du \ \geq \ 0
\end{array}$$

Note that the integral is a Kullback-Leibler Divergence, which is according to Gibbs' inequality always ≥ 0, and $= 0$ only if the two distributions are equal.

$$\sum_i p_i \cdot \log \frac{p_i}{q_i} = -\sum_i p_i \cdot \log \frac{q_i}{p_i} \geq -\sum_i p_i \cdot \left(\frac{q_i}{p_i} - 1\right) = -\sum_i q_i + \sum_i p_i = -1 + 1 = 0$$

Therefore, an EM step will never decrease the likelihood (which is not true, e.g., for Newton optimization steps).

Why is Excoffier and Slatkin's EM algorithm a special case of this?

x "phenotypes"

U genotypes

θ haplotype frequencies $p_1, p_2, \ldots$

$$\mathbb{E}_{\theta'}(\ell(\theta; x, U)|x) = ?$$

If the phenotypes x are in accordance with the genotypes u, then

$$\begin{aligned}\ell(\theta; x, u) &= \log(P_\theta(u)) \\ &= \log\left(\frac{n!}{n_1! \cdots n_m!}\right) + \sum_i \log\left(2p_{u(i,1)}p_{u(i,2)} \cdot \left(\frac{1}{2}\right)^{\delta_{u(i,1),u(i,2)}}\right)\end{aligned}$$

$$\begin{aligned}Q_{\theta'}(\theta) &= \mathbb{E}_{\theta'}(\ell(\theta; x, U)|x) \\ &= \log\left(\frac{n!}{n_1! \cdots n_m!}\right) + \sum_i \mathbb{E}_{\theta'}\left(\log\left(2p_{U(i,1)}p_{U(i,2)} \cdot \left(\frac{1}{2}\right)^{\delta_{U(i,1),U(i,2)}}\right)\middle| x_i\right)\end{aligned}$$

The conditional expectation in the last line is, by definition,

$$\sum_{[k,\ell]\in\mathcal{P}_i} P_{\theta'}(U = [k,\ell]|x_i) \cdot \log\left(2p_k p_\ell \cdot \left(\frac{1}{2}\right)^{\delta_{k\ell}}\right)$$

Where the sum is taken over the set $\mathcal{P}_i$ of all (unordered) haplotype pairs $[k,\ell]$ that are in accordance with phenotype i.

With

$$P_{\theta'}(U = [k,l]|x_i) = \frac{2p'_k p'_\ell \cdot 0.5^{\delta_{k,\ell}}}{\sum_{[k',\ell']\in\mathcal{P}_i} 2p'_{k'}p'_{\ell'} \cdot 0.5^{\delta_{k',\ell'}}},$$

where $\theta' = (p'_1, p'_2, ...)$.

Putting it all together and rearranging the sums, we obtain

$$Q_{\theta'}(\theta) = \sum_k \sum_{i=1}^n \frac{p'_k \sum_\ell I_{k\ell,i} p'_\ell}{\sum_{k'} \sum_{\ell'} I_{k'\ell',i} p'_{k'} p'_{\ell'}} \log p_k + \text{const.}$$

where const is a term that does not depend on any p_k. Thus, $Q_{\theta'}(\theta)$ is optimized by setting

$$p_k \propto \sum_{i=1}^n \frac{\sum_\ell I_{k\ell,i} p'_k p'_l}{\sum_{k'} \sum_{\ell'} I_{k'\ell',i} p'_{k'} p'_{\ell'}}$$

because, in general, the distribution $p_1, p_2, \ldots$ that maximizes $\sum_i n_i \log p_i$ is $p_i = n_i / \sum n_j$. This follows from the information inequality

$$\sum p_i \log p_i > \sum p_i \log q_i,$$

which is equivalent to

$$\sum p_i \log \frac{p_i}{q_i} > 0$$

(always assuming that distributions p and q are not equal).

Note that, indeed,

$$p_k \propto \sum_{i=1}^n \frac{\sum_\ell I_{kl,i} p'_k p'_l}{\sum_{k'} \sum_{\ell'} I_{k'l',i} p'_{k'} p'_{l'}}$$

is the same as the M step in Excoffier and Slatkin's EM algorithm.

12.2.3 Basic algorithms in PHASE

References

[SSD01] M. Stephens, Smith, P. Donnelly (2001) A New Statistical Method for Haplotype Reconstruction from Population Data *The American Journal of Human Genetics* **68(4)**

$G = (G_1, \ldots, G_n)$ observed genotypes of n individuals

$H_i = (h_{i1}, h_{i2})$ unknown (unordered) haplotypes of individual i

Gibbs samping with target distribution $\Pr(H|G)$: Start with initial guess for H and iterate the following steps.

- choose individual i purely randomly from all ambiguous individuals
- sample updated H_i from $\Pr(H_i|G, H_{-i})$, where H_{-i} is H without H_i.

Problem: $\Pr(H_i|G, H_{-i})$ depends on genetic and demographic models, e.g. on priors of haplotype frequencies

$$\Pr(H_i|G, H_{-i}) \propto \Pr(H_i|H_{-i}) \propto \Pr(h_{i1}|H_{-i}) \cdot \Pr(h_{i2}|H_{-i}, h_{i1})$$

$\Pr(h_{i1}|H_{-i})$ is only easy in parent-independent mutation model, which is usually unrealistic.

Stephens, Smith and Donnelly (2001) discuss two possible approximations, a "naive" one and their preferred one.

The naive Gibbs sampler

assumes parent-independent mutation

$$\Pr(h|H) = \frac{r_h + \theta v_h}{r + \theta}$$

r_h number of haplotypes of type h in H

r total number of haplotypes in H

v_h in case of mutation this is the probability that it leads to h

θ population-scaled mutation rate

If individual i has k heterozygous loci, 2^{k-1} different haplotypes h are possible. If this may be to many, just set $v_h = 1/M$, where M is the number of possible haplotypes.

The naive algorithm

1. pick individual i uniformly, let k be its number of heterogeneous loci; let $\{h_1, \ldots, h_m\}$ be the other individuals' haplotypes.
2. **for** $j = 1, \ldots, m$ **do**
   if H_i could be (h_j, h') **then**
   if h' is some $h_k \in \{h_1, \ldots, h_m\}$ **then**
   $p_j = \left(r_j + \frac{\theta}{M}\right)\left(r_k + \frac{\theta}{M}\right) - \left(\frac{\theta}{M}\right)^2$
   else
   $p_j = r_j \frac{\theta}{M}$
   end if
   else
   $p_j = 0$
   end if
   end for
3. With prob $\frac{2^k\left(\frac{\theta}{M}\right)^2}{\sum_j p_j + 2^k\left(\frac{\theta}{M}\right)^2}$ reconstruct H_i completely at random.
   Else: Choose $H_i = (h_j, h')$ with probability $p_j / \sum_k p_k$.

The basic standard algorithm in PHASE

The basic standard algorithm in PHASE uses an approximation proposed by Stephens and Donnelly (2000):

$$\Pr(h|H) \approx \sum_{\alpha} \sum_{s=0}^{\infty} \frac{r_\alpha}{r} \left(\frac{\theta}{r+\theta}\right)^s \frac{r}{r+\theta} (P^s)_{\alpha h}$$

r_α number of haplotypes of type α in H

r total number of haplotypes in H

θ population-scaled mutation rate (see next slide)

P transition matrix between types, given a mutation happens

s number of mutations

The basic standard algorithm in PHASE

- For polymorphic sites, assume only one mutation happened and set

$$\theta = \frac{1}{\log 2n}$$

- Problem: For sequence data, P will be huge. In this case use Gaussian quadrature (see also next section).
- For microsatellite data assume a stepwise mutation model with 50 alleles and set

$$\theta_j = \frac{1}{2}\left(\frac{1}{1+\mathcal{H}} - 1\right),$$

where $\mathcal{H}$ is the observed heterozygousity at that locus.

The basic standard algorithm in PHASE

Start with initial phasing and iterate the following steps

1. Choose individual i uniformly
2. Select subset S of (e.g. 5) ambiguous loci i
3. Phase the loci in S in individual i conditioned on the current phase of all other loci and of all loci in the other individuals.

References

[SD03] M. Stephens, P. Donnelly (2003) A comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data *Am. J. Hum. Genet.* **73**: 1162–1169

introduce a few improvements:

- in each step the genomes of all individuals are subdivided into blocks of the same number of loci (6,7, or 8, with probs. 0.3, 0.3, 0.4). Then, a block is chosen for all individuals in random order and the loci of this block are updated (conditioned on all other individual and on all other loci in the focal individual).
- in a certain fraction of individuals, it is allowed that only one haplotype is a copy of another haplotype in the data. This fraction is reduced down to 0 during the MCMC procedure.
- After the blockwise MCMC, haplotype frequencies are estimated for each block and blocks are iteratively ligated with adjacent blocks into larger blocks.

12.2.4 PAC in PHASE

References

[SS05] Matthew Stephens, Paul Scheet (2005) Accounting for Decay in Linkage Disequilibrium and Missing-Data Imputation *Am. J. Hum. Genet.* **76**:449–462

- Use information about order and distance between marker positions
- recombination rates may vary and are estimated
- applicable also when LD is "blocklike"
- imputation of missing data

n number of individuals

L number of loci

$G = (G_1, \ldots, G_n)$ genotypes; observed up to missing loci

$H = (H_1, \ldots, H_n)$ haplotypes; to be reconstructed

$H_{-i} = (H_1, \ldots, H_{i-1}, H_{i+1}, \ldots, H_n)$

$\rho = (\rho_1, \ldots, \rho_{L-1})$ recombination rates between loci

$$\rho_\ell = \frac{4N_e c_\ell}{d_\ell}$$

c_ℓ recombination probability per generation between loci ℓ and $\ell + 1$; to be estimated

d_ℓ known distance between loci ℓ and $\ell + 1$

strategy of PAC approach in PHASE

Start with initial H and ρ the following steps many times:

1. for each i update H_i by sampling from $\Pr(H_i | G_i, H_{-i}, \rho)$
2. propose change of ρ and accept or reject with Metropolis-Hastings (MH) step
3. update ordering ν of individuals with MH step for order-dependent PAC probabilities

Needed in these steps:

$$\Pr(H_i = (h, h') | H_{-i}, \rho) \propto (2 - \delta_{hh'}) \cdot p(h' | H_{-i}, \rho) \cdot p(h | H_{-i}, h', \rho)$$

(where $p(h|H_{-i}, h', \rho)$ will be approximated by $p(h|H_{-i}, \rho)$; note that $2 - \delta_{hh'} = 2 \cdot \left(\frac{1}{2}\right)^{\delta_{hh'}}$)

Simplifying assumption in the computation of $p(h|h_1, \ldots, h_k, \rho)$:

- when a locus in h is copied from some h_i only two possible coalescence times are allowed ($t_1 = 0.586/k$, $t_2 = 3.414/k$) and taken with probabilities $w_1 = 0.854$ and $w_2 = 0.146$. This is a Gauß quadrature approximation of the exponential distribution.
- if X_ℓ is the allele from which locus ℓ in h is copied and T_ℓ the corresponding coalescence time, then $(X_1, T_1), (X_2, T_2), \ldots$ is a Markov chain with $\Pr(X_1 = x, T_1 = t_r) = w_r/k$ and transition probabilities $\Pr(X_{\ell+1} = x', T_{\ell+1} = t_{r'} | X_\ell = x, T_\ell = t_r) =$

$$(1 - e^{-\rho_l d_l / k}) \cdot w_r/k + \delta_{xx'} \delta_{rr'} e^{-\rho_l d_l / k}.$$

Thus, HMM algorithms can be applied again.

- $$\Pr(h_{k+1} = a | X_\ell = x, T_\ell = t, h_1, \ldots, h_k, \rho) = \sum_m \frac{(\theta t)^m}{m!} e^{-\theta t} (P^m)_{h_{x,\ell} a}$$

Proposals for step 1 are halplotypes that are composed by blocks as described in Stephens and Donnelly (2003), leading to a list of promising haplotypes compatible with G_i.

For these haplotypes probabilities are computed with forward algorithm and one of them is chosen randomly according to the computed probabilities.

For imputation of missing types, probabilities are computed with forward-backward algorithm and types are sampled accordingly.

performance studies
haplotype inference:

dataset1 40 X chromosomes from unrelated males, paired into 20 pseudo-individuals, 8 regions of 87–327 kb and 45–165 segregating sites

dataset2 autosomal data from 129 children with known phase (as parents were also genotyped),

result PHASE with recombination PAC model best overall and for most regions.

imputing missing data:

data 50 genes sequenced for 24 humans of African descent and 23 of European descent, 15–230 segregating sites per gene.

simulation remove 5% of the data (in addition to 4.6% that was actually missing), either single alleles or the genotypes.

result PHASE with recombination PAC model always best

References

[CB+04] Crawford, Bhagale, Li, Hellenthal, Rieder, Nickerson, Stephens (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome *nature genetics*

[CB+04] Myers, Freeman, Auton, Donnelly, McVean (2008) A common sequence motif associated with recombination hot spots and genome instability in humans *nature genetics*

12.3 Phasing large genomic datasets

12.3.1 fastPHASE

References

[SS06] P. Scheet, M. Stephens (2006) A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase *Am. J. Hum. Genet.* **78**:629–644

If n haplotyps are to be reconstructed (that is, from $n/2$ sampled individual) at M marker positions, than the complexity of HMM algorithms in PHASE is $\mathcal{O}(n^2 M)$.

fastPHASE reduces this to $\mathcal{O}(nM)$.

Instead of sampling sections of haplotypes from copies of other haplotypes, all section are sampled from K clusters, similar to STRUCTURE.

First consider clustering method of haplotypes, where cluster just is a set of closely related haplotypes.

Then extend for phasing with in Hardy-Weinberg Equilibrium (HWE) within clusters.
First: consider local clustering method for haplotypes.

Given $h = (h_1, \ldots, h_n)$ haplotypes with M biallelic $\{0, 1\}$ (can be relaxed) marker positions.

$z_{im} \in \{1, \ldots, K\}$ origin of h_{im}, (marker position m in h_i). $z_i = (z_i, \ldots, z_{iM})$ modeled as a Markov chain with

$$p(z_{i1} = k) = \alpha_{k1}$$

and $p_m(k \to k') :=$

$$p(z_{im} = k' | z_{i(m-1)} = k, \alpha, r) = \left(1 - e^{-r_m d_m}\right) \cdot \alpha_{k'm} + \delta_{kk'} \cdot e^{-r_m d_m},$$

where d_m is the distance between markers $m-1$ and m, and the recombination parateters $r = (r_2, \ldots, r_M)$ as well as $\alpha = (\alpha_{km})$ are to be estimated.

Now for the emission probabilities:

$$p(h_i | z_i, q) = \prod_{m=1}^{M} p(h_{im} | z_{im}, q) = \prod_{m=1}^{M} q_{z_{im}m}^{h_{im}} \cdot (1 - q_{z_{im}m})^{1 - h_{im}},$$

where q_{km} is the frequency of allele 1 at marker m in cluster k.

Again, we obtain an HMM, parameter estimation can be done with EM, assigments of haplotype sections to culsters e.g. with Viterbi-Algorithm or Bayesian sampling tracing back contributions in the forward algorithm.

Now assume that unphased genotypes $g = (g_1, \ldots, g_n)$ are given, $g_{im} \in \{0, 1, 2\}$ is the genotype at marker m in individual i. Now assume HWE in each cluster. Let $\widetilde{z}_{im}$ be the unordered pair of clustes of

origin of g_{im}:

$$p(\widetilde{z}_{i1} = \{k_1, k_2\}) = (2 - \delta_{k_1 k_2})\, \alpha_{k_1} \alpha_{k_2}$$

and assume that $\widetilde{z}_i = (\widetilde{z}_{i1}, \ldots, \widetilde{z}_{iM})$ is a Markov chain with transition probabilities $p_m(\{k_1, k_2\} \to \{k_1', k_2'\}) =$

$$p_m(k_1 \to k_1') \cdot p_m(k_2 \to k_2') + \left(1 - \delta_{k_1 k_2} \delta_{k_1' k_2'}\right) \cdot p_m(k_1 \to k_2') \cdot p_m(k_2 \to k_1').$$

Emission probabilities:

$$p(g_{im} | \widetilde{z}_{im} = \{k_1, k_2\}, q) = \begin{cases} (1 - q_{k_1 m})(1 - q_{k_2 m}) & \text{if } g_{im} = 0 \\ q_{k_2 m}(1 - q_{k_1 m}) + q_{k_1 m}(1 - q_{k_2 m}) & \text{if } g_{im} = 1 \\ q_{k_1 m} q_{k_2 m} & \text{if } g_{im} = 2 \end{cases}$$

Difference to approach of Falush *et al.* (2003) implemented in STRUCTURE: Here, α varies between marker positions but not between individuals. In Falush *et al.* it is vice versa (for the parameter there called q). That is, here, α controls the frequency of the common haplotypes, not the contribution of the different clusters to an individual's genome.

HWE assumption is violated if the population is substructured. Applications of fastPHASE for data imputation or phasing may be robust against such violations. Moreover, extension of model is possible, assuming that individuals are sampled from known subpopulations and the parameters r and α vary between the subpopulations.

Parameter estimation with EM: Found that 20 independent starts with 25 iterations each is enough.

K =?: How many clusters to choose? Cross validation: Mask 15% of the genotypes, impute the genotyes with fastPHASE with various K between 4 and 12. Choose the K for which the genotypes a correct as often as possible (was $K = 8$ for data used in Scheet and Stephens, 2006). But also suggest to run with various K and compare results rather than relying on a single value of K.

12.3.2 Phasing with Beagle software package

References

[BB07] S.R. Browning and B.L. Browning (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering *Am. J. Hum. Genet.* **81**:1084–1097

For an initial guess of the haplotypes first construct a directed acyclic graph (DAG), in which

- each edge has a level m corresponding to marker position m
- and a lable corresponding to an allele occuring at marker position m,
- and for each of the occuring haplotypes $h_i = (h_{i1}, \ldots, h_{iM})$ there is path from the start node to the end node such that the lables of the edges are $h_{i1}, \ldots, h_{iM}$.
- (If a node has ingoing edges of level m than all its outgoin edges are of level $m+1$, and vice versa.)

Method of graph construction described in paper by S.R. Browning (2006)

Then construct an HMM whose possible states at step m are ordered pairs of DAG edges of the same level m. Transition probabilities:

$$P((e_1, e_2) \to (e_3, e_4)) = P(e_1 \to e_3) \cdot P(e_2 \to e_4),$$

where

$$P(e_i \to e_j) = \frac{\sharp\{\text{haplotypes whose path contains } e_i \text{ and } e_j\}}{\sharp\{\text{haplotypes whose path contains } e_i\}}$$

Emission probability: 1 if genotype at marker position m is compatible with labels of edges belonging to state, otherwise 0.

Now use "diploid HMM" to sample for each individual several haplotypes (in Beagle software 4 haplotype pairs per individual). Pool these haplotypes to construct DAG for next iteration. For

sampling use forward algorithm restricted on states corrsponding to states of the focal individual, and random tracebacks, with probabilities always proportional to current state. In last iterations use Viterbi

paths instead of random paths.

12.3.3 IMPUTE version 2

References

[HDM09] B.N. Howie, P. Donnelly, J. Marchini (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies *PLoS Genetics* **5**(6)

Similar to PAC-approach in PHASE, but is made for situation when reference haplotype data is available. Reference haplotypes and unphased genotypes are used together in phasing update step. For

runtime efficiency untyped SNPs are imputed in haplotype HMM framwork rather than in diploid HMM.

For accelaration restrict set of possible haplotypes in each iteration to those that are similar to existing ones.

12.3.4 MaCH

References

[LW+10] Y. Li, C.J. Willer, J. Ding, P. Scheet, G.R. Abscasis (2010) MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes *Genet. Epidemiol.* **24(8)**: 816–834

Similar to fastPHASE, but uses a larger number of haplotype templates instead of few haplotype groupings. In emission probabilities use error parameter ε_j that can depend on position j and covers sequencing error, gene conversion,...

12.3.5 polyHAP

References

[SW+08] S.Y. Su, J. White, D.J. Balding, L. J. M. Coin (2008) Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions *BMC Bioinformatics* **9**:513

Similar HMM approach like fastPHASE but for polyploid data. Computationally very demanding because states are unordered lists stating how many alleles have been sampled from how many clusters. Thus, many possible transitions between hidden states are considered.