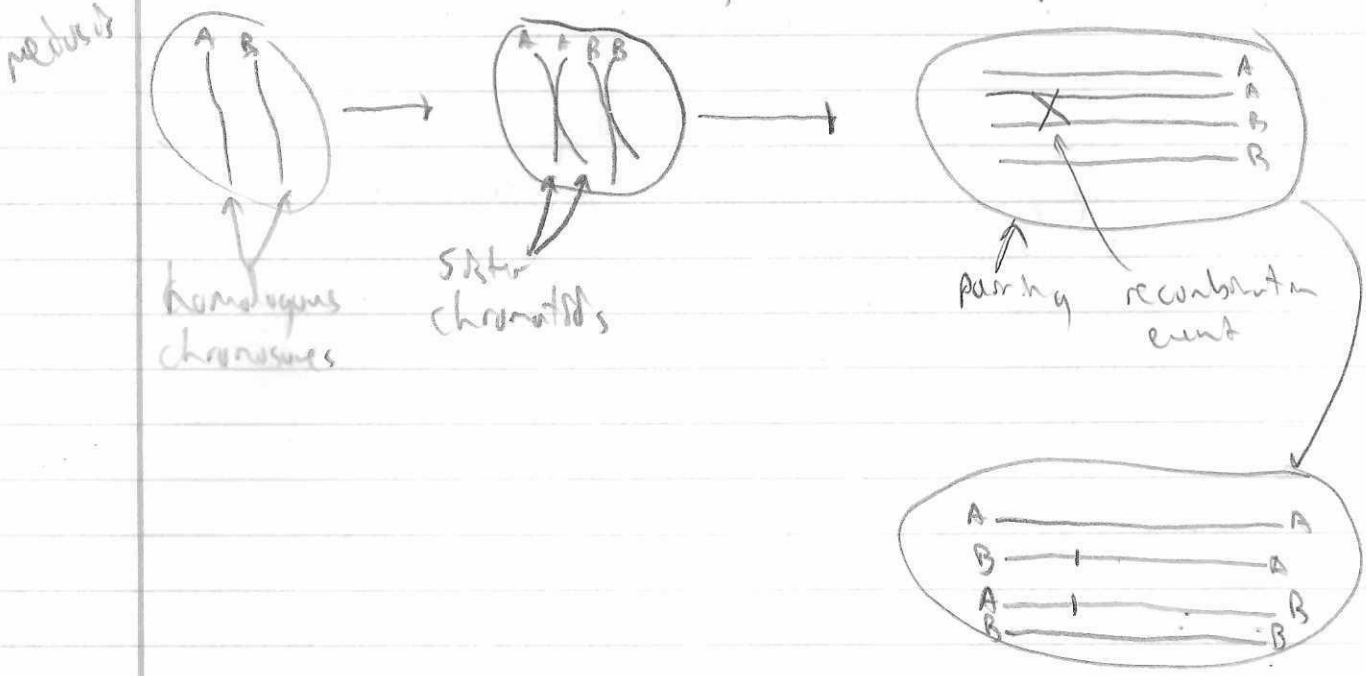
Li & Stephens (2003)

- Recombination: In meiosis, chromosomes pair and recombine



- Recombination makes it so that loci are not perfectly dependent.
- Dependence between Loci = "LD"
- This article relates LD to recombination ρ through a model
- Let $h_k \in \{0,1\}^s$ be the haplotype for individual k at s loci.

E.g. if $s = 3$, the possible values of h_k:

$h_k \in$ (0,0,0) (0,1,1)
(0,0,1) (1,0,1)
(0,1,0) (1,1,0)
(1,0,0) (1,1,1)

$1 = A, \quad 0 = a$

- We want a likelihood for $h_1, h_2, \ldots, h_n$ given the recombination rate ρ.
- By definition of conditional probability,

$$P_\rho(h_1, h_2, \ldots, h_n | \rho) = P_\rho(h_1|\rho)\, P_\rho(h_2|h_1, \rho) \cdots P_\rho(h_n | h_1, h_2, \ldots, h_{n-1}, \rho)$$

- Idea: Find approximation s.t.

$$P_\rho(h_1, \ldots, h_n|\rho) \approx \hat{\pi}(h_1|\rho)\, \hat{\pi}(h_2|h_1, \rho) \cdots \hat{\pi}(h_n | h_1, \ldots, h_{n-1}, \rho)$$

- Ewen's sampling formula approximation ($\theta = 4N\mu$)

h_k w.p. $\frac{k}{k+\theta}$ will equal one of $(h_1, \ldots, h_{k-1})$

↕ choose which one uniformly

w.p. $1 - \frac{k}{k+\theta}$, will be something completely different

Issues: ① Does not indicate that # differences should be small
② Does not account for recombination / LD

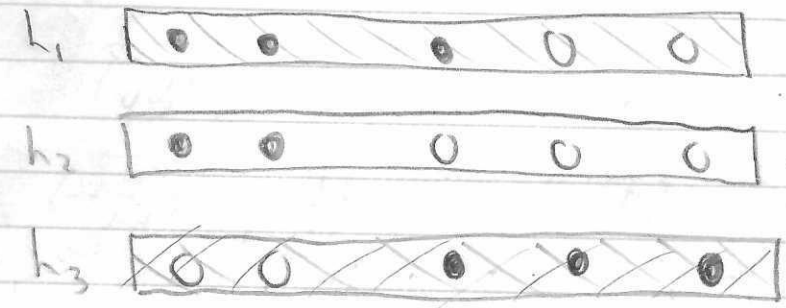
- Li and Stephen's $\hat{\pi}$ (Appendix A)

Haplotypes $h_1, h_2, \ldots, h_n$ (so $n/2$ diploid individuals)

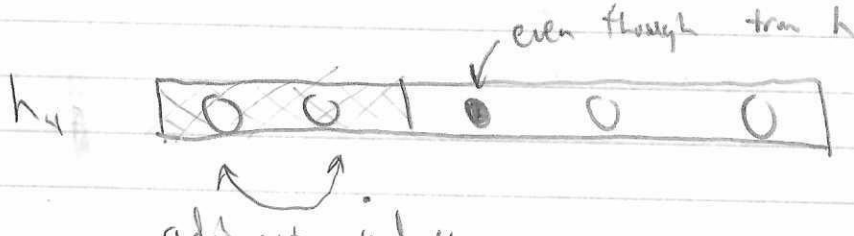
$\hat{\pi}(h_1) = \frac{1}{2^s}$ (all haplotypes are equally likely)

$\pi(h_{k+1} \mid h_1, \ldots, h_k)$

↳ imperfect mosaic of the other k haplotypes



h_4 is sometimes from h_1, h_2, or h_3. But it may or may not have copied those values.



$\rho \uparrow \Rightarrow$ less likely adjacent pairs are from same haplotype

$\rho \downarrow \Rightarrow$ more likely

- Let X_j = which haplotype h_{k+1} comes from at site j

$$X_j \in \{1, 2, \ldots, k\}$$

E.g. in example $X = (3, 3, 2, 2, 2, 2)$

- Model using a markov chain:

$$P_r(X_1 = x) = \frac{1}{k} \text{ for } x \in \{1, \ldots, k\}$$

$$P_r(X_{j+1} = x' \mid X_j = x) = \begin{cases} \exp\{-\rho d_j / k\} + (1 - \exp\{-\rho d_j / k\}) \frac{1}{k} & \text{if } x' = \\ (1 - \exp\{-\rho d_j / k\}) \frac{1}{k} & \text{otherwise} \end{cases}$$

d_j = distance between adjacent loci (known)

If $\rho = 0$ (no recombination)

$$= \begin{cases} 1 & x' = x \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow$ Never switches.

If $\theta = \infty$

$$= \begin{cases} \frac{1}{k} & \text{if } x' = x \\ \frac{1}{k} & \text{otherwise} \end{cases}$$

- No dependence between adjacent loci.
- Model mutation via imperfect copying

Copy is exact w.p. $\frac{k}{k+\theta}$

Copy is imperfect w.p. $1 - \frac{k}{k+\theta}$

} via Ewen's sampling

If imperfect, then 0 w.p. $\frac{1}{2}$ and 1 w.p. $\frac{1}{2}$

$$P_r\left(h_{n+1,j} = a \mid X_j = x, h_{1,j}, h_{n}\right) = \begin{cases} \frac{k}{k+\theta} + \frac{1}{2}\frac{\theta}{k+\theta} & \text{if } h_{x_j} = a \\ \frac{1}{2}\frac{\theta}{k+\theta} & \text{if } h_{x_j} \neq a \end{cases}$$

Same

different

We $\theta = \left(\sum_{m=1}^{n-1} \frac{1}{m}\right)^{-1}$

HW:

1.) What is length of h_i?

(5)

2.) What are possible values of h_{ij}?

(0 or 1)

3.) $n=2$, what is $\hat{\theta}$?

$\theta = \left(\sum_{m=1}^{1} \frac{1}{n}\right)^{-1} = (1)$

4.) $n=2$, $Pr(h_{2j}=0 \mid h_{1j}=0)$?

$= \frac{k}{k+\theta} + \frac{1}{2}\frac{\theta}{1+\theta}$

$= \frac{1}{1+1} + \frac{1}{2}\frac{1}{1+1} = \frac{1}{2} + \frac{1}{4} = \left(\frac{3}{4}\right)$

5.) $n=2$, $Pr[h_{2j}=0 \mid h_{1j}=1]$?

$= \frac{1}{2}\frac{\theta}{k+\theta} = \frac{1}{2}\frac{1}{1+1} = \left(\frac{1}{4}\right)$

6.) Suppose $\theta=\infty$, what is $Pr(X_{n+1}=1)$?

Did this in the notes. $\left(\frac{1}{k}\right)$

Given observed haplotypes $h_1, \ldots, h_n$, very easy to compute likelihood and maximize it over ρ

↳ Markov chains make this efficient

- Forward/backward algorithm for computing for estimating ρ
- Forward part can calculate likelihood
- Viterbi algorithm can give us hidden states

↳ easy and computationally fast (but 5 lectures to learn them)

- Why we $\theta = \left(\sum_{m=1}^{n-1} \frac{1}{m}\right)^{-1}$?

$\theta \sum_{m=1}^{n-1} \frac{1}{m}$ is expected # mutations on a tree with n alleles

↳ (we did this in Gillespie)

Set Expected number to be 1 at each size (since we know it occurred)

$$\Rightarrow \theta \approx \frac{1}{\sum_{m=1}^{n-1} \frac{1}{m}}$$

Likelihood depends on order $h_1, h_2, \ldots, h_n$

↳ so average ≈ 20 random orderings of haplotypes

- Why that form for transition?

Suppose $d \in (0, \infty)$, what is distribution of # jumps?

does not jump up: $\exp\{-\rho d / h\}$ ← true for all d

jumps (possibly to same haplotype) up: $1 - \exp\{-\rho d_i / h\}$

Let D = waiting time

$$f(D = d) \propto \exp\left\{-d \frac{\rho}{h}\right\}$$

↑ kernel of exponential

$$E[D] = \frac{h}{\rho}$$

If $\rho \uparrow$, then waiting time is small
$\rho \downarrow$, then waiting time is large

If $h \uparrow$, then waiting time is large
$h \downarrow$, then waiting time is small

If $h \uparrow$, then more likely to have a close relative that is not similar in sample, so stay on that relative longer.

Phylogenetics suggests $E(D) \approx \frac{\rho}{h+\rho} \approx \frac{\rho}{h}$ for small ρ