

SHREEDHARAN SRIRAM

Scribe (03/16/09)

Last Class: (Lecture 14)

- $P = 4Nr$ $r \rightarrow$ recombination rate: important parameter to get the dist. of n_{AB} .
- Log likelihood $l(4Nr) = \sum_{i,j} l(n_i, n_j, n_{i,j} / 4Nr_{i,j})$
Goal: Get the MLE of $r_{i,j}$.

This class: LECTURE 15

- For m individuals, for n sites on their genome
 - Estimate allele frequencies and haplotype frequencies
 - Methods:
 1. Maximum likelihood estimate
 2. E.M. algorithm (in the case of missing data)
 - Challenges :
 1. Dealing with related individuals
 2. Dont know which of the maternal or paternal chromosomes have that genotype
- For e.g. consider three biallelic sites
A G \rightarrow chr1 1000th position
T C \rightarrow chr1 1001th position
A C \rightarrow chr1 1002th position
There are many different cases of haplotypes that could be possible:
Such as: A G or A G
 T C C T
 A C A C
Total number of cases = $2^3/2$
- Haplotype inference problem: Given the genotype dataset, infer the haplotype of the individual.
- ALLELE FREQUENCY INFERENCE:
Objective: To infer the characteristics of the population (in this case, the allele freqs) from a set of sequences from unrelated individuals by computing the MLE of the allele frequencies.

Consider n sequences:

Let X of these have type 'a' \rightarrow Let its freq be p

Therefore, $n-X$ have type 'A' \rightarrow its freq is $(1-p)$

Likelihood function: $L(p|n,X) = C \cdot p^X (1-p)^{n-X}$ where $C \rightarrow nCX$

Log Likelihood: $\log C + X \log p + (n-X) \log(1-p)$

Take the derivative w.r.t. p and set it to 0, we get the MLE of $p = X/n$.

Note: In the case of ' μ ' inference:

Consider n chromosomes and S sites that vary across the sequences

$\mu \rightarrow$ mutation rate

Task: infer μ given n & S

Likelihood function: $L(\mu | D) = P(D | \mu)$
 $= P_n(S | \mu) \rightarrow P(\text{n sequences have S mutations given } \mu)$

Following similar steps as before, we can get an MLE of μ that maximizes the likelihood function.

Getting back to allele freq inference:

Case I: unrelated individuals (as above)

- The likelihood function $L(p | X, n)$ can be written as $L(p | X_1, X_2, \dots, X_n)$ if we consider a specific order.
- $L(p | X_1 \dots X_n) = \prod_i p^{X_i} (1-p)^{1-X_i}$
- Taking the derivative of the log likelihood function and setting it to 0, we get the MLE for $p = X/n$

Case II: Inferring the allele frequencies from the genotype frequencies

Consider a locus A with two possible alleles A_1 and A_2

Possible genotypes are:

$A_1A_1: n_{11}:p_{11}$

$A_1A_2: n_{12}:p_{12}$

$A_2A_2: n_{22}:p_{22}$

Allele frequencies are:

$A_1: p_1 = n_1/2n$

$A_2: (1-p_1) = n_2/2n$

$A_1:n_1 = 2n_{11} + n_{12}$

$A_2:n_2 = 2n_{22} + n_{12}$

$L(p_1 | n_{11}, n_{12}, n_{22}) = p_1^{(2n_{11} + n_{12})} \cdot (1-p_1)^{(2n_{22} + n_{12})}$

Taking the log, and then the derivative w.r.t. p , the MLE of $p = (2n_{11} + n_{12}) / (2n_{11} + n_{12} + 2n_{22} + n_{12})$

CasIII: Parent and Offspring pairs

Say we have genotypes of a parent and a child

Parent	child		
	A1A1	A1A2	A2A2
A1A1	a1	a2	0
A1A2	a3	a4	a5
A2A2	0	a6	a7

Assume the relation between the genotype freq and the allele freq follows HWE.

Parent	X	Spouse	P (child being A1A1)
A1A1		A1A1 : p_1^2	1
		A1A2 : $2p_1p_2$	0.5
		A2A2 : p_2^2	0

$$P(\text{parent and the child being A1A1}) = \frac{p_1^2(p_1^2 \cdot 1 + 2p_1p_2 \cdot 0.5)}{p_1^3(p_1+p_2)}$$

Similarly, one can calculate the probabilities shown in the table on slide 22.

The next step is to define the likelihood function which is a product of all the probabilities of the various cases, each raised to the power of its counts shown in the table above.

$$\begin{aligned} \log L &= a_1 \log p_1^3 + a_2 \log(p_1^2 p_2) + \dots \\ &= B \log p_1 + C \log(1-p_1) \end{aligned}$$

(the values B and C are shown on slide 23)

Taking the derivative w.r.t. p_1 , and setting it to 0, we get the MLE of $p_1 = B/(B + C)$

Note: This is not the same as the natural estimator for p_1 , and this abnormality could be explained by the fact that the individuals in this case are related and hence the MLE is more accurate than the natural estimator in such a case.