

## II. Population Genetics

# The Expectation-Maximization Algorithm

---

### Lecture 16

Instructor: Su-In Lee

## MLE of Allele Frequencies

---

- Find parameter estimates which make observed data most likely
- General approach, as long as tractable likelihood function exists
- Can use all available information
- Provides justification for natural estimators

## Today's Lecture

- The Expectation–Maximization algorithm in Genetics
- Frequency estimates for...
  - Recessive alleles
  - A, B, O alleles
  - Haplotype frequencies

## Setting for the EM Algorithm

- Specific type of incomplete data
  - More possible **categories** (genotypes) than can be **distinguished** (phenotypes)
- For example, consider disease locus with recessive alleles...
  - What are the possible genotypes?
  - What are the possible phenotypes?

## Setting for the EM Algorithm

- Any problem that is simple to solve for complete data ...
  - For example, estimating allele frequencies when all genotypes can be distinguished ...
- ... but the available data is “incomplete” in some way.
  - For example, when disease phenotypes are observed the distinction between homozygotes and heterozygotes is missing.

## The EM Algorithm

- Consider a set of starting parameters
- Use these to “estimate” the complete data
- Use estimates of complete data to update parameters
- Repeat as necessary

## An Example

---

- A random sample of 100 individuals
- 4 express a recessive phenotype
  - Assume the phenotype is controlled by a single gene
- Let's follow E-M algorithm steps ...

## Step 1:

---

- Set starting values for parameters
- For allele frequency estimation...
  - Equal frequencies are a common choice
  - $p_{\text{rec}} = 0.5$
- Useful to repeat process using different starting point

## Step 2:

---

- Estimate “complete data”
- Assign phenotypes to specific genotype categories
- Use Bayes’ Theorem

## Step 2 (continued):

---

- Calculate probability of each genotype among the 96 “normal” individuals

## Step 2:

- At the first iteration, the complete data would be filled in as:
  - 4 individuals with recessive genotype
  - 64 individuals with heterozygous genotype
  - 32 individuals with dominant genotype

## Step 3:

- Estimate allele frequencies by counting...

$$p_{rec} = \frac{N_{het} + 2N_{rec/rec}}{2N}$$

- What would be the estimated allele frequencies?
- Repeat Steps 1,2 & 3 as necessary ...

## Other Applications of the EM Algorithm in Genetics

---

- **Classic example:**
  - ABO blood group
- **Most common application:**
  - Haplotype frequency estimates

## The ABO blood group

---

- Determines compatibility for transfusions
- Controlled by alleles of ABO gene
- 3 alternative alleles  
A, B and O
- 6 possible genotypes,  $n(n + 1) / 2$ 
  - A/A, A/B, A/O, B/B, B/O, O/O

## Genotypes and Phenotypes

| Genotype | Phenotype |
|----------|-----------|
| A/A      | A         |
| A/B      | AB        |
| A/O      | A         |
| B/B      | B         |
| B/O      | B         |
| O/O      | O         |

## ABO Example

- Data of Clarke et al. (1959)
  - *British Med J* 1:603-607
  - Reported excess of gastric ulcers in individuals with blood type O
- $n_A = 186$ ,  $n_B = 38$ ,  $n_{AB} = 36$ ,  $n_O = 284$

## Quick Exercises

---

- Write out the likelihood for these data...
- What are complete data categories?
- Express the complete data “counts” as a function of allele frequency estimates and the observed data...

## Alternatives to EM ...

---

- Analytical solutions are not known for the general case
- Generic maximization strategies could be employed
- Could derive solutions using part of the data...
  - Would this be a good idea?

## The EM Haplotyping Algorithm

- Excoffier and Slatkin (1995)
  - *Mol Biol Evol* **12**:921-927
  - Provide a clear outline of how the algorithm can be applied to genetic data
- Combination of two strategies
  - E-M statistical algorithm for missing data
  - Counting algorithm for allele frequencies

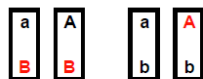
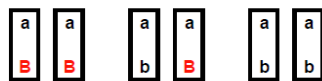
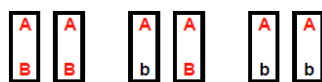
## Original Application of the EM Algorithm to a Genetics Problem

- Ceppellini R, Siniscalco M and Smith CAB (1955) The Estimation of Gene Frequencies in a Random-Mating Population. *Annals of Human Genetics* **20**:97-115
- This was ~20 years before the E-M algorithm was formally outlined in the statistical literature!

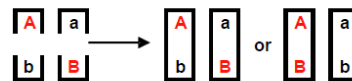
## Counting for Allele Frequencies

- For co-dominant markers, allele frequency typically carried out in very simple manner:
  - Count number of chromosomes (e.g.  $2N$ )
  - Count number of a alleles (e.g.  $n_a$ )
  - Allele frequency is simple proportion ( $n_a/2N$ )
- Haplotypes can't always be counted directly
  - Focusing on unambiguous genotypes introduces bias

## Counting Haplotypes for 2 SNPs

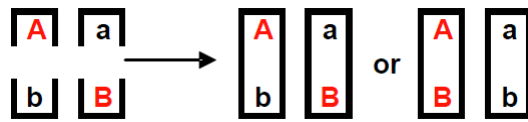


**Unambiguous Genotypes**  
Underlying Haplotype is Known



**Ambiguous Genotype**  
Multiple Underlying Genotypes Possible

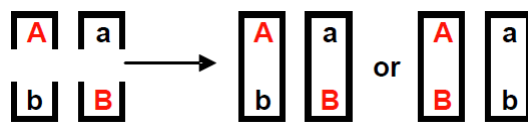
## Probabilistic Interpretation I



For example, if  
 $P_{AB} = 0.3$   
 $P_{ab} = 0.3$   
 $P_{Ab} = 0.3$   
 $P_{aB} = 0.1$

- Probability of first outcome:  
 $- 2P_{Ab}P_{aB} =$
- Probability of second outcome:  
 $- 2P_{AB}P_{ab} =$

## Probabilistic Interpretation II



For example, if  
 $P_{AB} = 0.3$   
 $P_{ab} = 0.3$   
 $P_{Ab} = 0.3$   
 $P_{aB} = 0.1$

- Conditional probability of first outcome:  
 $- 2P_{Ab}P_{aB} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) =$
- Conditional probability of second outcome:  
 $- 2P_{AB}P_{ab} / (2P_{Ab}P_{aB} + 2P_{AB}P_{ab}) =$

## Probabilistic Interpretation II

- 1. “Guesstimate” haplotype frequencies
- 2. Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes
- 3. Estimate frequency of each haplotype by counting
- 4. Repeat steps 2 and 3 until frequencies are stable

## Computational Cost (for SNPs)

- Consider sets of  $m$  unphased genotypes
  - Markers 1.. $m$
- If markers are bi-allelic
  - $2^m$  possible haplotypes
  - $2^{m-1} (2^m + 1)$  possible haplotype pairs
  - $3^m$  distinct observed genotypes
  - $2^{n-1}$  reconstructions for  $n$  heterozygous loci
- For example, if  $m = 10$

## EM Algorithm for Haplotyping

- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
  - Fewer microsatellites
- Fully or partially phased individuals contribute most of the information

## Other Common Applications

- E-M Algorithm also commonly used for:
  - Estimation of recombination fractions
  - Estimating mixture distributions

## Summary

---

- The E-M algorithm in genetics
- Outline the approach
- Examined specific examples

## Acknowledgement

---

- This lecture note is based on material by Profs Goncalo Abecasis (Univ of Michigan).