The E-M Algorithm in Genetics

Biostatistics 666

Maximum Likelihood Allele Frequencies

- 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:

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

Setting for the E-M 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 E-M Algorithm...

- Problem is simple with complete data …
  - For example, estimating allele frequencies when all genotypes are observed …

- … but available data can be “incomplete”.
  - For example, homozygotes and heterozygotes might be hard to distinguish.

The E-M Algorithm

- Consider a set of starting parameters
- Use these to “estimate” the complete data
- Use estimated 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
  - $p_{rec} = 0.5$

- 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

$$P(+/+;Normal) = \frac{P(+/+,Normal)}{P(Normal)} =$$

$$= \frac{P(+/+,Normal)}{P(+/+,Normal) + P(+/-,Normal)}$$

$$= \frac{P(+/+)}{P(+/+) + P(+/-)}$$

Step 2 (Finally!):

- 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 as necessary …

Round	Estimate	E(+/+)	E(+/-)	E(-/-)	ln L
1	0.50	32.00	64.00	4.00	-14.40240
2	0.36	45.18	50.82	4.00	-9.33657
3	0.29	52.36	43.64	4.00	-8.02405
4	0.26	56.60	39.40	4.00	-7.58067
5	0.24	59.21	36.79	4.00	-7.41213
6	0.22	60.87	35.13	4.00	-7.34396
7	0.22	61.94	34.06	4.00	-7.31540
8	0.21	62.64	33.36	4.00	-7.30317
9	0.21	63.10	32.90	4.00	-7.29787
10	0.20	63.40	32.60	4.00	-7.29555
11	0.20	63.60	32.40	4.00	-7.29453
12	0.20	63.73	32.27	4.00	-7.29408
13	0.20	63.82	32.18	4.00	-7.29388
14	0.20	63.88	32.12	4.00	-7.29379
15	0.20	63.92	32.08	4.00	-7.29375
16	0.20	63.95	32.05	4.00	-7.29374

Alternatives

- Analytical solutions
- Generic maximization strategies
- Calculating second derivates is always a useful complement, why?…

Other Applications of the E-M Algorithm in Genetics

- **Classic example:**
  - ABO blood group

- **Other applications:**
  - Haplotype frequency estimates
  - Inferring population labels
  - Modeling components in mixtures

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

ABO Blood Group II

Phenotype	Antigen		Antibody	
	A	B	A	B
A	+	-	-	+
B	-	+	+	-
O	-	-	+	+
AB	+	+	-	-

There are only 4 possible phenotypes for the ABO blood group.

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…

The iterations give …

Iteration	p_A	p_B	p_O
1	0.300	0.200	0.500
2	.243	.074	.683
3	228	.070	.700
4	.228	.070	.702
5	.228	.070	.702

Alternatives to E-M…

- 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 E-M 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 E-M Algorithm to A Genetic 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

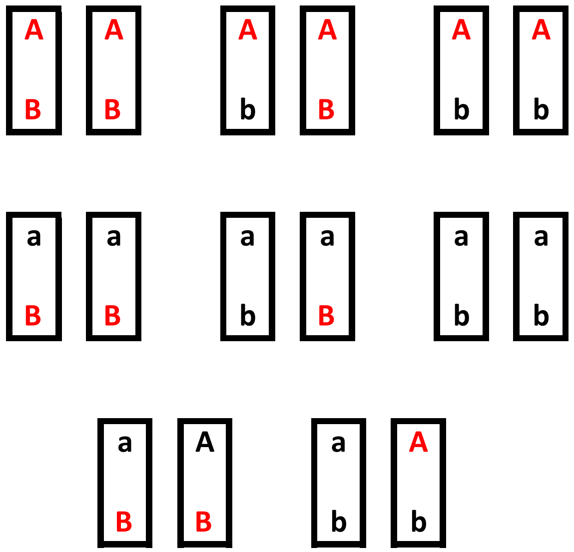
- This was ~20 years before the E-M algorithm was formally outlined in the statistical literature!

Counting for Allele Frequencies

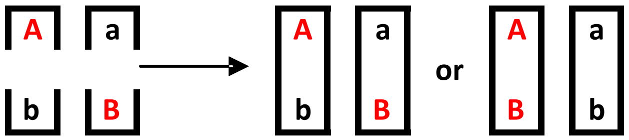
- For co-dominant markers, allele frequency typically carried out in very simple manner:
  - Count number of chromosomes (e.g. 2N)
  - Count number of a alleles (e.g. n_a)
  - Allele frequency is simple proportion (n_a/2N)

- Haplotypes can't always be counted directly
  - Focusing on unambiguous genotypes introduces bias

Counting Haplotypes for 2 SNPs

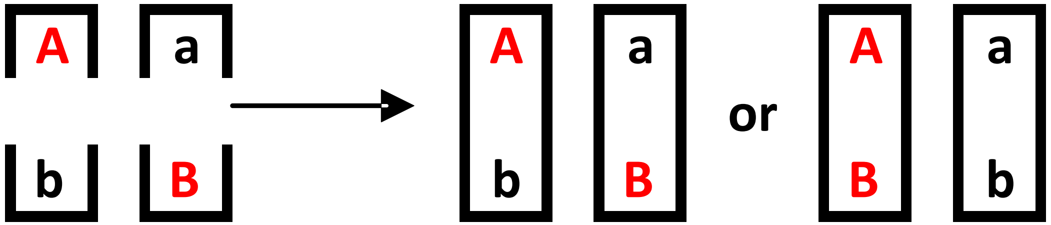


Unambigous Genotypes
Underlying Haplotype is Known



Ambigous Genotype
Multiple Underlying Genotypes Possible

Probabilistic Interpretation



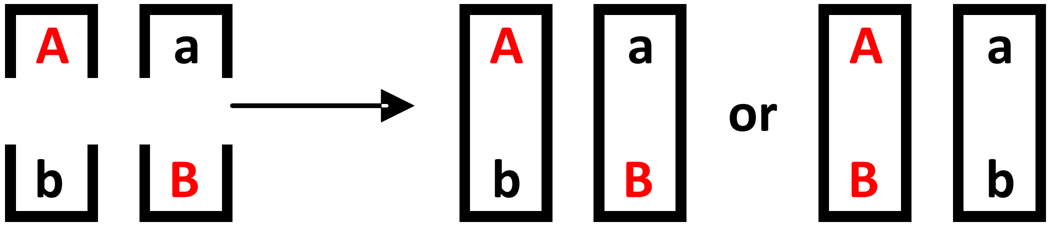
Probability of first outcome:

$2\ P_{Ab}\ P_{aB}$

Probability of second outcome:

$2\ P_{AB}\ P_{ab}$

Probabilistic Interpretation



For example, if:

$P_{AB} = 0.3$
$P_{ab} = 0.3$
$P_{Ab} = 0.3$
$P_{aB} = 0.1$

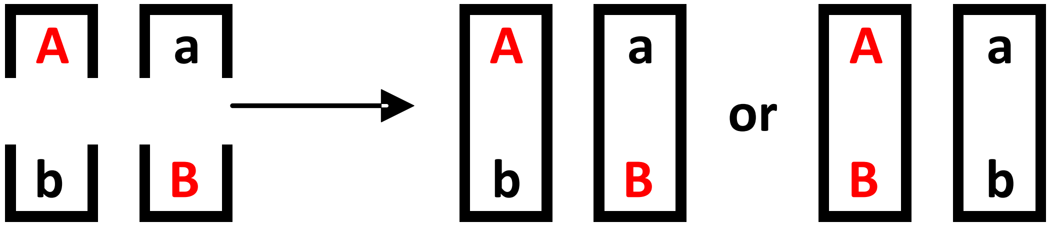
Probability of first outcome:

$$2\, P_{Ab}\, P_{aB} = 0.06$$

Probability of second outcome:

$$2\, P_{AB}\, P_{ab} = 0.18$$

Probabilistic Interpretation II



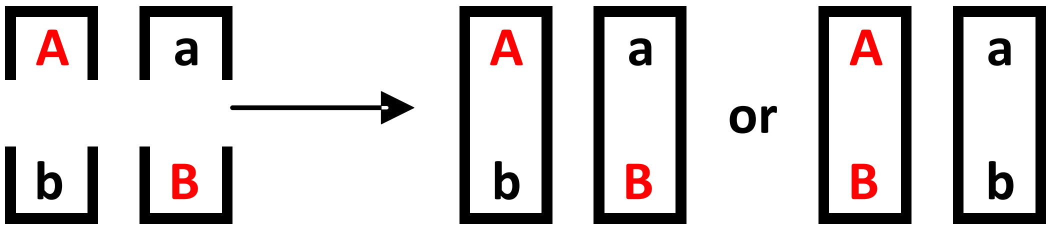
Conditional probability of first outcome:

$$2\ P_{Ab}\ P_{aB} / (2\ P_{Ab}\ P_{aB} + 2\ P_{AB}\ P_{ab})$$

Conditional probability of second outcome:

$$2\ P_{AB}\ P_{ab} / (2\ P_{Ab}\ P_{aB} + 2\ P_{AB}\ P_{ab})$$

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:

$$2\, P_{Ab}\, P_{aB} \,/\, (2\, P_{Ab}\, P_{aB} + 2\, P_{AB}\, P_{ab}) = 0.25$$

Conditional probability of second outcome:

$$2\, P_{AB}\, P_{ab} \,/\, (2\, P_{Ab}\, P_{aB} + 2\, P_{AB}\, P_{ab}) = 0.75$$

Basic E-M Algorithm For Haplotyping

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

Basic E-M Algorithm for Haplotyping

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

Other Common Applications

- E-M Algorithm also commonly used for:
  - Estimating recombination fractions
  - Defining genotype intensity clusters
  - Finding sub-populations and their allele frequencies

Today:

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

Next Lecture …

- E-M algorithm for Haplotyping
- Historical Alternatives
- Recent Enhancements and Alternatives
- Hypothesis testing

Recommended Reading

- Excoffier and Slatkin (1995)
  Mol Biol Evol **12:**921-927

- Introduces the E-M algorithm in the context of haplotyping