Part II. Genetics:  
5. Haplotype Reconstruction II

Lecture 11 – Feb 10, 2015  
CSE 427 Computational Biology  
Instructor: Su-In Lee  
TAs: Scott Lundberg  
TTh 12:00-1:20 @ MGH 238

Outline

- Problem statement  
  - Haplotype reconstruction
- Statistical methods for haplotype reconstruction  
  - Parsimony algorithm  
  - Expectation-maximization algorithm

Typical genotype data

- Two alleles for each individual  
  - Chromosome origin for each allele is unknown

Typical genotype data

- Two alleles for each individual  
  - Chromosome origin for each allele is unknown

- Multiple haplotype pairs can fit observed genotype
What if there are no relatives?

- Rely on linkage disequilibrium (LD)
  - LD: non-random association of variants at different sites in the genome

- Assume that population consists of small number of distinct haplotypes

Haplotype reconstruction

- Also called, phasing, haplotype inference or haplotyping

Data

- Genotypes on $N$ tag SNPs from $M$ individuals

Goals

- Haplotype reconstruction for individuals
- Frequency estimation of all possible haplotypes

Outline

- Background
  - Haplotype, haplotype frequency
  - Limitations of a general approach for GWAS

- Problem statement
  - Haplotype reconstruction

- Statistical methods for haplotype reconstruction
  - Parsimony algorithm
  - Expectation-maximization algorithm

Clark’s haplotyping algorithm


- One of the first haplotyping algorithms
  - Computationally efficient
  - Very fast and widely used in 1990’s
  - More accurate methods are now available
Clark’s haplotyping algorithm
- Find unambiguous individuals
  - Initialize a list of known haplotypes
- Unambiguous individuals
  - Homozygous at every locus (e.g. {TT} {AA} {CC})
    Haplotypes: TAC
  - Heterozygous at just one locus (e.g. {TT} {AA} {CG})
    Haplotypes: TAC or TAG

Unambiguous vs. ambiguous
- Haplotypes for 2 SNPs (alleles: A/a, B/b)
- Unambiguous Genotypes
  - Multiple Underlying Genotypes Possible
  - Underlying Haplotype is Known

Clark’s haplotyping algorithm
- Find unambiguous individuals
  - Initialize a list of known haplotypes
- Resolve ambiguous individuals
  - If possible, use two haplotypes from the list
  - Otherwise, use one known haplotype and augment list
- If unphased individuals remain
  - Assign phase randomly to one individual
  - Augment haplotype list and continue from previous step

Parsimonious phasing - example
- Notation (more compact representation)
  - 0/1: homozygous at each locus (00,11)
  - h: heterozygous at each locus (01)
  - Ambiguous Genotype

| 1 0 1 0 0 h | 1 0 1 0 0 0 |
| 0 0 1 1 0 0 |
| 0 1 0 1 1 0 |

| 1 0 1 0 0 0 |
| 1 0 1 0 0 1 |
| 0 0 1 1 0 0 |
| 0 1 0 1 1 0 |
Parsimony algorithm (Clark 1990)

- **Pros**
  - Very fast
  - Can deal with very long sequences

- **Cons**
  - No homozygotes or single SNP heterozygotes in the data
  - Some haplotypes may remain unresolved
  - Outcome depends on order in which lists are transversed
  - Naïve, not very accurate (no modeling)

Outline

- **Background**
  - Haplotype, haplotype frequency
  - Limitations of a general approach for GWAS

- **Problem statement**
  - Haplotype reconstruction

- **Statistical methods for haplotype reconstruction**
  - Parsimony algorithm
  - Expectation-maximization algorithm

The EM haplotyping algorithm


  - Why EM for haplotyping?
    - EM is a method for estimating parameters with hidden variables.
    - Hidden variables: haplotype state of each individual
    - Parameters: haplotype frequencies

  
  ![Haplotype frequencies](image)

  **Individual /**
  **Haplotype state (hidden variable)**

  $z = 0$

  $z = 1$

  **Haplotype frequencies (parameters)**

  $P_{AB}$, $P_{aB}$, $P_{AB}$, $P_{aB}$

  Assume that we know haplotype frequencies

  - **For example, if**
    - $P_{AB} = 0.3$
    - $P_{aB} = 0.3$
    - $P_{AB} = 0.3$
    - $P_{aB} = 0.1$

  - **Probability of first outcome:**
    - $2P_{AB}P_{aB} = 0.06$

  - **Probability of second outcome:**
    - $2P_{AB}P_{aB} = 0.18$
Conditional probabilities

For example, if
\[ P_{AB} = 0.3 \]
\[ P_{ab} = 0.3 \]
\[ P_{Ab} = 0.3 \]
\[ P_{aB} = 0.1 \]

- Conditional probability of first outcome:
  \[ \frac{2P_{Ab}P_{ab}}{2P_{Ab}P_{ab} + 2P_{AB}P_{ab}} = 0.25 \]

- Conditional probability of second outcome:
  \[ \frac{2P_{AB}P_{ab}}{2P_{Ab}P_{ab} + 2P_{AB}P_{ab}} = 0.75 \]

Assume that we know the haplotype state of each individual

- Computing haplotype frequencies is straightforward

Notation

- Hidden variables \( Z \)
  \[ Z_{ij} = 1 \] if individual \( i \) has haplotype pairs \( j \)
  \[ 0 \] otherwise

- Parameters \( \theta \)
  - Haplotype frequencies: \( p_{Ab}, p_{aB}, p_{AB}, p_{ab} \)

EM as Chicken vs Egg

- IF \( Z \) is known, could estimate parameters \( \theta \)
  - e.g., frequencies of haplotypes
**EM as Chicken vs Egg**

- IF parameters $\theta$ known, could estimate $z_{ij}$
  - Compute conditional probabilities

**Simple version: “hard” EM**

- If $z_{ij} < 0.5$, pretend it’s 0; $z_{ij} > 0.5$, pretend it’s 1
  i.e., classify points as component 0 or 1
- Now recalculate $\theta$, assuming that partition
- Then recalculate $z_{ij}$, assuming that $\theta$
- Then recalculate $\theta$, assuming new $z_{ij}$, etc., etc.
In “soft” EM, $z_{ij}$’s are continuous-valued

**EM as Chicken vs Egg**

- IF $\mathbf{Z}$ known, could estimate parameters $\theta$
  - Frequencies of haplotypes
- IF parameters $\theta$ known, could estimate $z_{ij}$
  - Compute conditional probabilities

Convergence provable? YES

- BUT we know neither; (optimistically) iterate:
  - E-step: calculate expected $z_{ij}$, given parameters
  - M-step: Estimate parameters given $E(z_{ij})$
- Overall, a clever “hill-climbing” strategy

**Phasing By EM**

- EM: Method for parameter estimation with hidden variables

- Inferring haplotype state of each individual
- Parameters (haplotype frequencies)
- Hidden variables (haplotype states of individuals)
- Frequency estimation
- Find expected values

**Estimating haplotype frequencies**
EM 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

Phasing by EM

Data:

1 0 h 1
h 0 1 h
1 h 1 1

Z11 = ½
Z12 = ½
Z21 = ½
Z22 = ½
Z31 = ½
Z32 = ½

Phasing by EM

Data:

1 0 h 1
h 0 1 h
1 h 1 1

Z11 = ½
Z12 = ½
Z21 = ½
Z22 = ½
Z31 = ½
Z32 = ½

Phasing by EM

Data:

1 0 h 1
h 0 1 h
1 h 1 1

Z11 = ½
Z12 = ½
Z21 = ½
Z22 = ½
Z31 = ½
Z32 = ½

Z11 + Z12 = 1
Phasing by EM

**Data:**

<table>
<thead>
<tr>
<th>1 0 h 1</th>
<th>h 0 0 1</th>
<th>1 h 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1</td>
<td>0 0 0 1</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 1 1</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>1 0 0 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

** Frequencies**

<table>
<thead>
<tr>
<th>0 0 1 0</th>
<th>1/12</th>
<th>0 0 1 1</th>
<th>1/12</th>
<th>0 1 0 1</th>
<th>1/12</th>
<th>0 1 1 1</th>
<th>1/12</th>
<th>1 0 0 1</th>
<th>1/12</th>
<th>1 0 1 1</th>
<th>1/12</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0</td>
<td>0 0 1 1</td>
<td>0 1 0 1</td>
<td>0 1 1 1</td>
<td>1 0 0 1</td>
<td>1 0 1 1</td>
<td>1 1 0 1</td>
<td>1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Haplotypes**

<table>
<thead>
<tr>
<th>Z_{11}</th>
<th>Z_{12}</th>
<th>Z_{21}</th>
<th>Z_{22}</th>
<th>Z_{31}</th>
<th>Z_{32}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 4</td>
<td>0 6</td>
<td>0 7 5</td>
<td>0 2 5</td>
<td>0 6 7</td>
<td>0 4 7</td>
</tr>
</tbody>
</table>

**Expectation**

**Phasing by EM**

**Data:**

<table>
<thead>
<tr>
<th>1 0 h 1</th>
<th>h 0 0 1</th>
<th>1 h 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1</td>
<td>0 0 0 1</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 1 1</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>1 0 0 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

** Frequencies**

<table>
<thead>
<tr>
<th>0 0 1 0</th>
<th>1/12</th>
<th>0 0 1 1</th>
<th>1/12</th>
<th>0 1 0 1</th>
<th>1/12</th>
<th>0 1 1 1</th>
<th>1/12</th>
<th>1 0 0 1</th>
<th>1/12</th>
<th>1 0 1 1</th>
<th>1/12</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0</td>
<td>0 0 1 1</td>
<td>0 1 0 1</td>
<td>0 1 1 1</td>
<td>1 0 0 1</td>
<td>1 0 1 1</td>
<td>1 1 0 1</td>
<td>1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Haplotypes**

<table>
<thead>
<tr>
<th>Z_{11}</th>
<th>Z_{12}</th>
<th>Z_{21}</th>
<th>Z_{22}</th>
<th>Z_{31}</th>
<th>Z_{32}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 4</td>
<td>0 6</td>
<td>0 7 5</td>
<td>0 2 5</td>
<td>0 6 7</td>
<td>0 4 7</td>
</tr>
</tbody>
</table>

**Expectation**

**Phasing by EM**

**Data:**

<table>
<thead>
<tr>
<th>1 0 h 1</th>
<th>h 0 0 1</th>
<th>1 h 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1</td>
<td>0 0 0 1</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 1 1</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>1 0 0 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

** Frequencies**

<table>
<thead>
<tr>
<th>0 0 1 0</th>
<th>1/6</th>
<th>0 0 1 1</th>
<th>0 1 0 1</th>
<th>0 1 1 1</th>
<th>1 0 0 1</th>
<th>1 0 1 1</th>
<th>1 1 0 1</th>
<th>1 1 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0</td>
<td>0 0 1 1</td>
<td>0 1 0 1</td>
<td>0 1 1 1</td>
<td>1 0 0 1</td>
<td>1 0 1 1</td>
<td>1 1 0 1</td>
<td>1 1 1 1</td>
<td></td>
</tr>
</tbody>
</table>

**Haplotypes**

<table>
<thead>
<tr>
<th>Z_{11}</th>
<th>Z_{12}</th>
<th>Z_{21}</th>
<th>Z_{22}</th>
<th>Z_{31}</th>
<th>Z_{32}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0 1 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

**Expectation**

**Phasing by EM**

**Data:**

<table>
<thead>
<tr>
<th>1 0 h 1</th>
<th>h 0 0 1</th>
<th>1 h 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 1</td>
<td>0 0 0 1</td>
<td>1 0 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>0 0 1 1</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>1 0 1 1</td>
<td>1 0 0 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

** Frequencies**

<table>
<thead>
<tr>
<th>0 0 1 0</th>
<th>1/6</th>
<th>0 0 1 1</th>
<th>0 1 0 1</th>
<th>0 1 1 1</th>
<th>1 0 0 1</th>
<th>1 0 1 1</th>
<th>1 1 0 1</th>
<th>1 1 1 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 1 0</td>
<td>0 0 1 1</td>
<td>0 1 0 1</td>
<td>0 1 1 1</td>
<td>1 0 0 1</td>
<td>1 0 1 1</td>
<td>1 1 0 1</td>
<td>1 1 1 1</td>
<td></td>
</tr>
</tbody>
</table>

**Haplotypes**

<table>
<thead>
<tr>
<th>Z_{11}</th>
<th>Z_{12}</th>
<th>Z_{21}</th>
<th>Z_{22}</th>
<th>Z_{31}</th>
<th>Z_{32}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0 1 1</td>
<td>1 1 1 1</td>
</tr>
</tbody>
</table>

**Expectation**

**Computational cost (for SNPs)**

- Consider sets of $m$ unphased genotypes
  - Tag SNPs 1..$m$
  - For example, if $m=10$, $2m = 20$
- 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 $m=10$, $2^{m-1} = 1024$
EM Algorithm For Haplotyping

- Cost grows rapidly with number of markers
- Typically appropriate for < 25 SNPs
  - Fewer microsatellites
- More accurate than Clark’s method
- Fully or partially phased individuals contribute most of the information

Enhancements to EM

- List only haplotypes present in sample
- Gradually expand subset of markers under consideration, eliminating haplotypes with low estimated frequency from consideration at each stage
  - SNPHAP, Clayton (2001)
  - HAPLOTYPER, Qin et al (2002)

Divide-And-Conquer Approximation

- Number of potential haplotypes increases exponentially (Qin et al. 2002)
  - Number of observed haplotypes does not
- Approximation
  - Successively divide marker set
  - Locally phase each segment through EM
  - Prune haplotype list as segments are ligated
  - Merge by phasing vectors of haplotype pairs

  1 0 0 1 0 1 0 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 0 1 1 1 1 1 1 0

  Computation order: \( \sim m \log m \)
  - Exact EM is order \( \sim 2^m \)

Summary

- Background
  - Haplotype, haplotype frequency
  - Limitations of a general approach for GWAS
- Problem statement
  - Haplotype reconstruction
- Statistical methods for haplotype reconstruction
  - Parsimony algorithm
  - Expectation-maximization algorithm