Guest lecture for course:
Computational Genetics (236608)

Linkage Disequilibrium
Mapping and HaploBlock

Gideon Greenspan
gdg@cs.technion.ac.il
Part 1: LD Mapping

• Basic LD Mapping
  – $\chi^2$-squared test for individual SNPs

• Mapping with Haplotypes
  – Population phenomena

• Haplotyping
  – Clark algorithm
  – EM algorithm
Linkage Disequilibrium

- LD = Another word for ‘correlation’
  - Correlation between markers in a population
- Random recombination destroys correlation
  - Close markers *may* have high LD
  - Above 1 Mb, LD disappears
LD Mapping: The Basics

• Take set of unrelated individuals
  – Ideally from a small, inbred population
• Measure markers at high resolution
  – Single Nucleotide Polymorphisms are ideal
• Test marker–disease correlations
  – Non-parametric disease model
  – Suitable (in theory) for low penetrance
LD Mapping in Action

- 3 1 1
× 0 1 2

- 1 3 1
× 1 1 1
## Chi-Squared Test

### Observed Counts

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>69</td>
<td>236</td>
<td>305</td>
</tr>
<tr>
<td>a</td>
<td>31</td>
<td>264</td>
<td>295</td>
</tr>
<tr>
<td>∑</td>
<td>100</td>
<td>500</td>
<td>600</td>
</tr>
</tbody>
</table>

### Expected Counts

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50.83</td>
<td>254.17</td>
<td>305</td>
</tr>
<tr>
<td>a</td>
<td>49.17</td>
<td>245.83</td>
<td>295</td>
</tr>
<tr>
<td>∑</td>
<td>100</td>
<td>500</td>
<td>600</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum \frac{(o - e)^2}{e} = 15.85 \]

1 degree of freedom
\[ \Rightarrow p-value = 0.0001 \]
SNPs

• Single base pair which exhibits variation
  – Caused by point mutations during meiosis
  – Variation almost always biallelic

• dbSNP contains ~ 4.3×10^6 SNPs
  – Over 1 SNP per 1,000 base pairs
  – About half with minor allele frequency > 20%
  – This number is still growing rapidly!
LD Mapping in Context

Identify chromosome
(10^8 bp)

Linkage analysis
(10^6 ~ 10^7 bp)

Identify genes
(10^5 ~ 10^6 bp)

Resequencing
(10^0 bp)
False Positives

• Causes of spurious LD
  – Population structure
    • Migration and admixture
    • Preferential mating
  – Phenotypic site interaction
    • Disease epistasis

• Key problem: too many SNP tests
  – Bonferroni correction
Haplotypes

Generally, only a few of the $2^{loci}$ possible haplotypes cover >90% of a population, due to bottleneck effects and genetic drift.
Bottleneck Effects

10^6 years

10^5 years
LD Mapping with Haplotypes

• Obtain haplotypes for a genomic region
  – Treat haplotype as correlated allele
• Advantage: fewer tests
  – Reduced false positive rate
• Disadvantage: ignores recombination
  – Different haplotypes could contain target
• Best: consider partial haplotypes…
The Haplotyping Problem

Variable Loci

Maternal Chromosome

Hidden Haplotypes

Paternal Chromosome

Observed Genotypes

A/T  C/G  T/T  A/C  G/T

A  G  T  C  T

T  C  T  A  G
Why is it hard?

- A series of joint measurements containing \( h \) heterozygous loci can be divided \( 2^{h-1} \) ways (we don’t care which is maternal or paternal).
Why is it approachable?

- Many of the haplotypes appear many times.
- Data for many individuals allows inference.

**Solution seems ‘better’ since it uses fewer haplotypes.**
Formalization 1

• Assume all loci biallelic (realistic).
• Individuals numbered $1 \ldots n$
• Loci numbered $1 \ldots l$
• Possible alleles $B = \{0, 1\}$
• Possible haplotypes $H = B^l$
• Possible locus observations $L = \{[B, B]\}$
• Possible genotypes $G = L^l$
• Possible haplotype pairs $D = \{[H, H]\}$
Formalization 2

- Given a true haplotype pair \([h_1, h_2] \in D\), \(G(h_1, h_2) \in G\) is the genotype observed.
- Given an observed genotype \(g \in G\), \(D(g) \subseteq D\) is set of possible haplotype pairs.

- Problem input: \((g_1, \ldots, g_n)\) where \(g_i \in G\)
- Problem output: \((d_1, \ldots, d_n)\) where \(d_i \in D(g_i)\)
Clark’s Algorithm

1. Initialize set $S$ to $\{\}$.
2. For genotypes $g_i$ with a single possibility $[h_1,h_2]$ assign $d_i=[h_1,h_2]$ and add $h_1,h_2$ to $S$.
3. For genotypes $g_i$ with a possibility containing a member $h_1 \in S$ and another haplotype $h_2$, assign $d_i=[h_1,h_2]$ and add $h_2$ to $S$.
4. Repeat step 3 until all haplotypes are assigned or we add nothing new to $S$.
5. Assign any remaining $d_i$ arbitrarily.
Clark: Rerun (same input)

- A/T C/G T/T A/C G/T
- A/A C/G T/T C/C T/T
- A/A C/G C/T A/C T/T
- A/T C/G C/C A/A G/T
- A/T C/C C/T A/A G/T

5 haplotypes used

- TCTAG
- ACTCT
- ACCAT
- ACCAT
- TCTAG
Clark: Comments

- Implementation is very fast, $O(ln^2)$
- Total failure if no starting point.
- Blind haplotyping of ‘orphans’ at end.
- Arbitrary selections based on input order.
  - Try multiple orderings, select best results.
- Or formulate choices as integer program
  - Solve approximately by linear relaxation.
Part 2: HaploBlock

- Haplotype blocks
- Statistical model
- Model inference
- Model criterion
- Applications
  - Haplotyping
  - Block-based LD mapping
Recombination Hotspots
Haplotype Blocks

1. GAACCTGC  ATTTCGACTGC  CCAGTAGGC
2. ACGTACA  GATGAGCTGC  CCAGTAGGC
...
99. ACGTACA  AACCGAGGT  TGTACTAA
100. GAACCTGC  GATGAGCTGC  TGTGCCTAA

Recombination hotspot separates blocks
Few block variants due to bottlenecks, drift
Mutation hotspot
Bayesian Network Model

\[ \text{Pr}(C = c) \text{ is frequency of haplotype } c \]

Values of variable \( C \) are 1\ldots q

denoting index of block’s haplotype

\[ \text{Pr}(a_j \mid c) \text{ is deterministic} \]

Values of variable \( A_j \) are A,C,G,T,–
denoting allele at site \( j \) of haplotype.
Example: \( A_1 A_2 A_3 = CTA \) for \( C = 2 \)

\[ \text{Pr}(h_j \mid a_j) \text{ is cumulative mutation rate} \]

Values of variable \( H_j \) are A,C,G,T,–
denoting allele at site \( j \) observed
after possible haplotype mutations
Bayesian Network Model

\[
\Pr(c, a_1, a_2, a_3, h_1, h_2, h_3) = \\
\Pr(c) \times \\
\Pr(a_1 \mid c) \times \\
\Pr(a_2 \mid c) \times \\
\Pr(a_3 \mid c) \times \\
\Pr(h_1 \mid a_1) \times \\
\Pr(h_2 \mid a_2) \times \\
\Pr(h_3 \mid a_3)
\]
Bayesian Network Model

- **haplotype block**
- **recombination hotspot**
Data Likelihood

• For haplotypes $H$, likelihood is:

$$\Pr(H) = \prod_{h \in H} \left[ \sum_{c_1} \cdots \sum_{c_b} \sum_{a_1} \cdots \sum_{a_l} \left[ \Pr(c_1) \prod_{k=2}^{b} \Pr(c_k \mid c_{k-1}) \prod_{k=1}^{b} \prod_{j=s_k}^{e_k} \Pr(a_j \mid c_k) \Pr(h_j \mid a_j) \right] \right]$$

But we can calculate this efficiently using a suitable elimination ordering!
Data Criterion

• Maximum Likelihood leads to over-fitting
  – No hotspots, no mutations, many ancestors
  – Need to consider model complexity
  – Min $\text{DL}(H,M) = \text{DL}(M) - \log_2 \text{Pr}(H|M)$

• $\text{DL}(M)$ considers variable elements only
  – Ancestor block sequences
  – Markov chain parameters
Model Search

4
C_1

3
C_2

4
C_1

3
C_2

5
C_1

Add/remove

Nudge

Ancestors
Model for Haplotyping

• Learn model directly from genotypes

• Haplotype pair: choose most likely under model
## Haplotyping Results

<table>
<thead>
<tr>
<th>Site pairwise error rate</th>
<th>C21a</th>
<th>C21b</th>
<th>C21c</th>
<th>C21d</th>
<th>C21e</th>
<th>ACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clark</td>
<td>.0548</td>
<td>.0251</td>
<td>.0280</td>
<td>.0329</td>
<td>.0234</td>
<td>.0381</td>
</tr>
<tr>
<td>Hierarchical EM</td>
<td>.0095</td>
<td>.0042</td>
<td>.0009</td>
<td>.0047</td>
<td>.0083</td>
<td>.0152</td>
</tr>
<tr>
<td>HAPLOTYPER</td>
<td>.0224</td>
<td>failed</td>
<td>.0204</td>
<td>.0077</td>
<td>failed</td>
<td>.0102</td>
</tr>
<tr>
<td>PHASE</td>
<td>.0669</td>
<td>.0403</td>
<td>.0655</td>
<td>.0262</td>
<td>.0183</td>
<td>.0419</td>
</tr>
<tr>
<td>HaploBlock</td>
<td>.0047</td>
<td>.0020</td>
<td>.0005</td>
<td>.0014</td>
<td>.0048</td>
<td>.0098</td>
</tr>
<tr>
<td><strong>Improvement factor</strong></td>
<td>2x</td>
<td>2x</td>
<td>2x</td>
<td>3x</td>
<td>2x</td>
<td>=</td>
</tr>
</tbody>
</table>

C21x data: 20 haplotypes, 100 SNPs over ≤ 35kb, Patil et al. (2001)  
ACE data: 22 haplotypes, 52 SNPs over 24kb, Rieder et al. (1999)  

*Average shown for 10 random pairings of true haplotypes*
Model for LD Mapping

- Learn model from marker data
- Mapping: try making phenotype dependent on each block
<table>
<thead>
<tr>
<th>Resequencing required</th>
<th>5q31 haplos</th>
<th>5q31 genos</th>
<th>Chr 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLADE</td>
<td>144 kb</td>
<td>–</td>
<td>107 kb</td>
</tr>
<tr>
<td>No Blocks</td>
<td>131 kb</td>
<td>105 kb</td>
<td>33 kb</td>
</tr>
<tr>
<td>HaploBlock</td>
<td>40 kb</td>
<td>37 kb</td>
<td>24 kb</td>
</tr>
<tr>
<td>Improvement factor</td>
<td>3x</td>
<td>3x</td>
<td>1.4x</td>
</tr>
</tbody>
</table>

5q31 data: 258 haplotypes, 98 SNPs over 464kb, Daly et al. (2001)
Chr 21 data: 20 haplotypes, 5 sets of 200 SNPs, Patil et al. (2001)

Average shown for 5 random selections of target SNP
HaploBlock: Comments

• Our model boils down to an HMM
  – Calculations have linear complexity
  – Forward/backward probability caching

• Better to infer multiple models
  – Prevent getting stuck in local minima
  – Account for uncertainty of block identification
  – Use Gibbs-style iterations on hotspots
  – Take ‘average’ result over set of models