Checking Pairwise Relationships

Lecture 21
Biostatistics 666
Today ...

- Multipoint analysis
  - Markov chain for handling many markers

- Other types of relatives
  - Multipoint analysis of other relative pairs
  - Other methods for checking relationships
The final ingredient connects IBD states along the chromosome …
The Likelihood of Marker Data

\[
L = \sum_{I_1} \sum_{I_2} \ldots \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i | I_{i-1}) \prod_{i=1}^{M} P(X_i | I_i)
\]

- General, but slow unless there are only a few markers.
- Combined with Bayes’ Theorem can estimate probability of each IBD state at any marker.
- This is not a linkage test yet!
Example

- Consider two loci separated by $\theta = 0.1$
- Each loci has two alleles, each with frequency $0.50$
- If two siblings are homozygous for the first allele at both loci, what is the probability that IBD = 2 at the first locus?
Solution

| I₁ | I₂ | P(I₁) | P(I₂|I₁) | P(X₁|I₁) | P(X₂|I₂) | P |
|----|----|-------|---------|---------|---------|---|
| 0  | 0  |       |         |         |         |   |
| 0  | 1  |       |         |         |         |   |
| 0  | 2  |       |         |         |         |   |
| 1  | 0  |       |         |         |         |   |
| 1  | 1  |       |         |         |         |   |
| 1  | 2  |       |         |         |         |   |
| 2  | 0  |       |         |         |         |   |
| 2  | 1  |       |         |         |         |   |
| 2  | 2  |       |         |         |         |   |
## Solution

| $I_1$ | $I_2$ | $P(I_1)$ | $P(I_2|I_1)$ | $P(X_1|I_1)$ | $P(X_2|I_2)$ | $P$   |
|-------|-------|----------|---------------|---------------|---------------|-------|
| 0     | 0     | 0.25     | 0.67          | 0.0625        | 0.0625        | 0.00066 |
| 0     | 1     | 0.25     | 0.30          | 0.0625        | 0.125         | 0.00058 |
| 0     | 2     | 0.25     | 0.03          | 0.0625        | 0.25          | 0.00013 |
| 1     | 0     | 0.50     | 0.15          | 0.125         | 0.0625        | 0.00058 |
| 1     | 1     | 0.50     | 0.70          | 0.125         | 0.125         | 0.00551 |
| 1     | 2     | 0.50     | 0.15          | 0.125         | 0.25          | 0.00231 |
| 2     | 0     | 0.25     | 0.03          | 0.25          | 0.0625        | 0.00013 |
| 2     | 1     | 0.25     | 0.30          | 0.25          | 0.125         | 0.00231 |
| 2     | 2     | 0.25     | 0.67          | 0.25          | 0.25          | 0.01051 |
Solution

- Taking into account all available genotype data...
  - $P(I_1 = 2) = 0.57$
  - $P(I_1 = 1) = 0.37$
  - $P(I_1 = 0) = 0.06$

- Considering only one marker, the corresponding probabilities would be 0.44, 0.44 and 0.11.
The Likelihood of Marker Data

\[ L = \sum_{I_1} \sum_{I_2} \ldots \sum_{I_M} P(I_1) \prod_{i=2}^{M} P(I_i \mid I_{i-1}) \prod_{i=1}^{M} P(X_i \mid I_i) \]

- General, but slow unless there are only a few markers.

- How do we speed things up?
A Markov Model

- Re-organize the computation slightly, to avoid evaluating nested sum directly

- Three components:
  - Probability considering a single location
  - Probability including left flanking markers
  - Probability including right flanking markers

- Scale of computation increases linearly with number of markers
The Likelihood of Marker Data

\[ L = \sum_{I_j} P(I_j)P(X_j \mid I_j)P(X_1 \ldots X_{j-1} \mid I_j)P(X_{j+1} \ldots X_M \mid I_j) \]

\[ = \sum_{I_j} P(I_j)P(X_j \mid I_j)L_j(I_j)R_j(I_j) \]

- A different arrangement of the same likelihood
- The nested summations are now hidden inside the \( L_j \) and \( R_j \) functions…
Left-Chain Probabilities

\[ L_m(I_m) = P(X_1, ..., X_{m-1} \mid I_m) \]

\[ = \sum_{I_{m-1}} L_{m-1}(I_{m-1}) P(X_{m-1} \mid I_{m-1}) P(I_{m-1} \mid I_m) \]

\[ L_1(I_1) = 1 \]

- Proceed one marker at a time.
- Computation cost increases linearly with number of markers.
Right-Chain Probabilities

\[ R_m(I_m) = P(X_{m+1}, \ldots, X_M \mid I_m) \]
\[ = \sum_{I_{m+1}} R_{m+1}(I_{m+1}) P(X_{m+1} \mid I_{m+1}) P(I_{m+1} \mid I_m) \]

\[ R_M(I_M) = 1 \]

- Proceed one marker at a time.
- Computation cost increases linearly with number of markers.
Pictorial Representation

- Single Marker
- Left Conditional
- Right Conditional
- Full Likelihood
Extending the MLS Method ...

\[ w_j = P(X_j \mid I_j)P(X_1 \ldots X_{j-1} \mid I_j)P(X_{j+1} \ldots X_M \mid I_j) \]
\[ = P(X_j \mid I_j)L_j(I_j)R_j(I_j) \]

- We just change the definition for the “weights” given to each configuration!
Some Extensions We’ll Discuss

- Modeling error
  - What components might have to change?

- Modeling other types of relatives
  - What components might have to change?

- Modeling larger pedigrees
Efficient computational framework for multipoint analysis of sibling pairs
Coming up ...

- Checking of relationships for pairs of individuals
  - Maximum likelihood based strategy

- Extend multipoint algorithm to different types of relatives
Verifying relationships is crucial

- Genetic analyses require relationships to be specified

- Misspecifying relationships can lead to tests of inappropriate size
  - Inflate Type I error
  - Decrease power
Results

Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.
Strategy:

- Information we have:
  - $G$ – observed genotypes at each marker
  - $p$ – allele frequencies at each marker
  - $\theta$ - recombination fraction between consecutive markers

- $P(X|R)$ for each possible relationship $R$
  - unrelated, half-sib, sib-pairs, MZ twins
Likelihood

- Sum over IBD states at each location

\[ L = \sum_{I_1} \ldots \sum_{I_m} P(I_1) \prod_{i=2}^{m} P(I_i \mid I_{i-1}) \prod_{i=1}^{m} P(X_i \mid I_i) \]

- Set of possible I changes with R
Notation

- $R$  
  Hypothesized Relationship
- $I_k = (I_{km}, I_{kf})$  
  Allele sharing at locus $k$
- $X_k$  
  Genotype pair at locus $k$

- $\alpha_k(j \mid R) = P(X_1, X_2, ..., X_{k-1}, I_k = j \mid R)$  
  Joint probability of data at first $k-1$ markers and IBD vector $I_k = j$ at marker $k$
Details on I

- Possible inheritance patterns
  - (0,0) – no sharing
  - (1,0) – share maternal allele
  - (0,1) – share paternal allele
  - (1,1) – share both alleles

- For convenience, separate IBD=1 into maternal and paternal sharing states
Algorithm for Likelihood Calculation

\[
\alpha_1(j \mid R) = P(I_1 = j \mid R)
\]

\[
\alpha_{k+1}(j \mid R) = \sum_i \alpha_k(i \mid R) P(X_k \mid I_k = i) t_k(i, j)
\]

\[
L = \sum_j \alpha_M(j \mid R) P(X_M \mid I_M = j)
\]
Relationship between I and R

- Probability of $I_1=(0,0), (1,0), (0,1)$ and $(1,1)$:
  - MZ Twins: $(0, 0, 0, 1)$
  - Unrelated: ?
  - Parent-Offspring: ?
  - Full sibs: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
  - Maternal half sibs: $(\frac{1}{2}, \frac{1}{2}, 0, 0)$
  - Paternal half sibs: ?
### P(X|I) for pairs of individuals

| GENOTYPE | P(X_1,X_2|I) for I = |
|----------|---------------------|
|          | (0,0)  | (0,1) or (1,0) | (1,1) |
| ii       | p_i^4     | p_i^3         | p_i^2 |
| ii       | 2p_i^3p_j | p_i^2p_j     | 0     |
| ii       | p_i^2p_j^2 | 0             | 0     |
| ii       | 2p_i^2p_jp_k | 0          | 0     |
| ii       | 4p_i^2p_j^2 | p_i^2p_j(p_i+p_j) | 2p_i^2p_j |
| ij       | 4p_i^2p_jp_k | p_i^2p_jp_k | 0     |
| ij       | 4p_i^2p_jp_k | 0             | 0     |
## Transition Matrix (Full Sibs)

\[
\begin{bmatrix}
(0,0) & (1,0) & (0,1) & (1,1) \\
(0,0) & (1-\psi)^2 & (1-\psi)\psi & \psi(1-\psi) & \psi^2 \\
(1,0) & (1-\psi)\psi & (1-\psi)^2 & \psi^2 & \psi(1-\psi) \\
(0,1) & (1-\psi)\psi & \psi^2 & (1-\psi)^2 & (1-\psi)\psi \\
(1,1) & \psi^2 & (1-\psi)\psi & (1-\psi)\psi & (1-\psi)^2 \\
\end{bmatrix}
\]

\[\psi = 2\theta(1-\theta)\]

\[r(i, j) = |i_1 - j_1| + |i_2 - j_2|\]

\[t(i, j) = \psi^{r(i,j)} (1-\psi)^{2-r(i,j)}\]
Transition Matrix  
(Maternal Half Sibs)

<table>
<thead>
<tr>
<th></th>
<th>(0,0)</th>
<th>(1,0)</th>
<th>(0,1)</th>
<th>(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>(1 − ψ)</td>
<td>ψ</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0)</td>
<td>ψ</td>
<td>(1 − ψ)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

ψ = 2θ(1 − θ)  \quad r(i, j) = |i_1 − j_1|  
\quad t(i, j) = ψ^{r(i, j)}(1 − ψ)^{1−r(i, j)}
### Transition Matrix (Paternal Half Sibs)

$$
\begin{bmatrix}
(0,0) & (1,0) & (0,1) & (1,1) \\
(0,0) & (1-\psi) & 0 & \psi & 0 \\
(1,0) & 0 & 0 & 0 & 0 \\
(0,1) & \psi & 0 & (1-\psi) & 0 \\
(1,1) & 0 & 0 & 0 & 0 \\
\end{bmatrix}
$$

\[
\psi = 2\theta(1-\theta) \\
r(i, j) = |i_2 - j_2| \\
t(i, j) = \psi^{r(i,j)}(1-\psi)^{1-r(i,j)}
\]
### Transition Matrix (Unrelated)

<table>
<thead>
<tr>
<th></th>
<th>(0,0)</th>
<th>(1,0)</th>
<th>(0,1)</th>
<th>(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Transition Matrix (MZ twins)

<table>
<thead>
<tr>
<th></th>
<th>(0,0)</th>
<th>(1,0)</th>
<th>(0,1)</th>
<th>(1,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(0,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(1,1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Example 1

- Consider genotypes for one marker
- \( X = (1/1, 1/1) \)
- Assume \( p_1 = .2, .5 \) or .8

- Calculate \( P(X|R) \) for each relationship
  - MZ twin, Full Sibs, Half-Sibs, Unrelated
Example II

- Consider genotypes for 2 markers
  - $X_1 = (1/1, 2/2)$
  - $X_2 = (1/1, 2/2)$
- Assume $p_1=p_2=\frac{1}{2}$
- Assume
  - $\theta = 0.0528, \psi = 0.10$
  - $\theta = 0.5000, \psi = 0.50$
- Calculate $P(X|R)$ for each relationship
Simulations (θ=.1, M=50)

<table>
<thead>
<tr>
<th>True R</th>
<th>Inferred R</th>
<th>Full Sibs</th>
<th>Half Sibs</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Sibs</td>
<td>0.914</td>
<td>0.085</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>Half Sibs</td>
<td>0.044</td>
<td>0.872</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>Unrelated</td>
<td>&lt;0.001</td>
<td>0.059</td>
<td>0.941</td>
<td></td>
</tr>
</tbody>
</table>
## Simulations (\(\theta = .2, \ M = 50\))

<table>
<thead>
<tr>
<th>True R</th>
<th>Full Sibs</th>
<th>Half Sibs</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Sibs</td>
<td>0.948</td>
<td>0.052</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Half Sibs</td>
<td>0.038</td>
<td>0.899</td>
<td>0.064</td>
</tr>
<tr>
<td>Unrelated</td>
<td>&lt;0.001</td>
<td>0.062</td>
<td>0.938</td>
</tr>
</tbody>
</table>
Simulations ($\theta = .1$, $M=400$)

<table>
<thead>
<tr>
<th>True R</th>
<th>Full Sibs</th>
<th>Half Sibs</th>
<th>Unrelated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Sibs</td>
<td>1.000</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Half Sibs</td>
<td>&lt;.001</td>
<td>1.000</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Unrelated</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Bayesian Approach

- Given some prior information on the expected frequency of each relative pair...
  - Alternative to simply maximizing $P(X|R=r)$

$$P(R = r \mid X) = \frac{Prior(R)P(X \mid R = r)}{\sum_{R} Prior(R)P(X \mid R)}$$
More distant relationships

Figure 1  Autosomal transition probabilities for grandparent-grandchild (GG), half-sib (HS), and avuncular (AV) pairs. $P(I_{k+1} = 1 | I_k = 0) = P(I_{k+1} = 0 | I_k = 1)$ is shown. Note that $P(I_{k+1} = 0 | I_k = 0) = 1 - P(I_{k+1} = 1 | I_k = 0)$ and $P(I_{k+1} = 1 | I_k = 1) = 1 - P(I_{k+1} = 0 | I_k = 1)$. 
Problem ...

- Consider some genome scan data
  - 380 microsatellite markers
- Consider some pair of individuals
  - Putative siblings
- Observed Sharing
  - Identical for 379/380 genotype pairs
- \( L(G|R=MZ\ Twins) = 0 \)
  - \( L(G|R=\text{Any other}) > 0 \)
Solution: Allow for Genotyping Errors

- A small proportion of errors could lead to misclassification
  - Allow for possibly erroneous genotypes
- \( \epsilon \) – error rate parameter

\[
P(X_i | I_i) = \sum_{G_i} P(X_i | G_i, \epsilon) P(G_i | I_i)
\]

\[
= (1 - \epsilon)^2 P(G_i | I_i) + [1 - (1 - \epsilon)^2] P(X_{i1})P(X_{i2})
\]
Conclusions

- Likelihood approach provides reliable manner to infer relationships

- Can incorporate multiple linked markers
  - Some distant relationships can only be discerned by likelihood approach
Today

- Checking of Relationships for Pairs of Individuals
- Multipoint algorithm for calculating likelihoods for genotype data
Recommended Reading


- Optional
  - Epstein et al. (2000), *AJHG* 67:1219-31