Parametric Linkage Analysis

Biostatistics 666
Lecture 25
Last Lecture

- Elston Stewart Algorithm
  - Can handle large pedigrees
  - Proceeds one nuclear family at a time
  - Limited to small numbers of markers

- Calculates conditional probabilities for sections of the pedigree
Today

- Refresher on non-parametric analysis

- Parametric linkage analysis
  - Evaluates a specific trait model
    - Disease allele frequencies
    - Probability of disease for each genotype

- Successfully mapped >1,000 rare disorders
Data for a Linkage Study

- Pedigree
  - Set of individuals of known relationship

- Observed marker genotypes
  - SNPs, VNTRs, microsatellites

- Phenotype data for individuals
Linkage Analysis

- Aims to relate sharing of specific chromosomal regions to phenotypic similarity

- Parametric methods define explicit relationship between phenotypic and genetic similarity

- Non-parametric methods test for increased sharing among affected individuals
Non-parametric Linkage Analysis

- No specific model for disease locus
- Evaluates whether segregation at specific locations is "not-random"
- Specifically, the objective is to show increased IBD sharing among sets of affected individuals
Allele Sharing Analysis

- Reject random sharing at a particular region
- Less powerful than classic methods
  - When disease model is known
- More robust than classic methods
  - When disease model is unknown
Historical References

- Penrose (1953) suggested comparing IBD distributions for affected siblings.
  - Possible for highly informative markers (eg. HLA)

- Risch (1990) described the MLS method for evaluating the evidence for linkage in affected sibling pair data.

- Soon after, large-scale microsatellite genotyping became possible and geneticists attempted to tackle more complex diseases...
Non-parametric Analysis for Arbitrary Pedigrees

- Must rank general IBD configurations which include sets of more than 2 affected individuals
  - Low ranks correspond to no linkage
  - High ranks correspond to linkage

- Multiple possible orderings are possible
  - Especially for large pedigrees

- Interesting regions are those where IBD configurations with higher rank are more common
Non-Parametric Linkage Scores

- Introduced by Whittemore and Halpern (1994)

- The two most commonly used ones are:
  - *Pairs* statistic
    - Total number of alleles shared IBD between pairs of affected individuals in a pedigree
  - *All* statistic
    - Favors sharing of a single allele by a large number of affected individuals.
Kong and Cox Method

- A probability distribution for IBD states
  - Under the null and alternative

- Null
  - All IBD states are equally likely

- Alternative
  - Increase (or decrease) in probability of each state is modeled as a function of sharing scores

- "Generalization" of the MLS method
Typical Plot for NPL Along Chromosome
Parametric Linkage Analysis

- Requires a model for the disease
  - Frequency of disease allele(s)
  - Penetrance for each genotype

- Typically employed for single gene disorders and Mendelian forms of complex disorders
Typical Interesting Pedigree
How it works...

- Uses disease model to infer segregation of disease alleles through pedigree

- Estimate the recombination fraction between disease locus and a genetic marker of known location
  - Summarize evidence for linkage (co-segregation) in a LOD score
Two Point Analysis

- Proceeds one marker at a time
- Estimates the recombination fraction between each marker and disease locus
- Trait model used to infer segregation of disease alleles
Two Point LOD Score

\[ LOD = \log_{10} \frac{L(\theta \leq \frac{1}{2})}{L(\theta = \frac{1}{2})} \]

- Compares two quantities:
  - Probability of the data assuming disease locus is unlinked
  - Probability of the data assuming disease locus is at a specific location

- Usually, LOD scores are tabulated either for:
  - A specific grid of possible locations
  - A unique location for each marker, chosen to maximize the LOD score
Example: Track Segregation of Disease Allele

In this case, the model might specify an autosomal locus, with a very rare dominant allele with full penetrance.
Example:
Test Segregation of Marker Alleles

- Which allele is on the disease haplotype?
- How often does it co-segregate with disease?
- How often is it separated from the disease allele?
Example:
Test Segregation of Marker Alleles

- Disease allele originally in chromosome with marker allele 3.
- Disease allele and marker allele 3 co-segregate in 5 meioses
  - No recombination.
- Disease allele and marker allele 3 are separated in 1 meiosis.
  - Recombinant.
The next step

- Likelihood as a function of $\theta$?
- LOD score as function of $\theta$?
- Maximum likelihood estimate of $\theta$?
  - Corresponds to MLOD, maximum LOD score
Exemplar LOD Score Calculation
(Phase Known Data)

\[ L(\theta) = (1 - \theta)^{\text{non-recombinants}} \theta^{\text{recombinants}} \]
\[ = (1 - \theta)^5 \theta \]

Likelihood

\[ LOD = \log_{10} \frac{L(\hat{\theta})}{L(\theta = \frac{1}{2})} = \]
\[ = \log_{10} \frac{L(\frac{1}{6})}{L(\frac{1}{2})} \]
\[ = 0.63 \]

Compares two likelihoods, one with \( \theta = \frac{1}{2} \) and another with \( \theta \leq \frac{1}{2} \).

In this case, the maximum possible LOD score would be \( \sim 1.8 \), for a marker with zero recombinants.
Table of Two Point LOD Scores

<table>
<thead>
<tr>
<th>Recombination Fraction (θ)</th>
<th>MLE</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.167</th>
</tr>
</thead>
<tbody>
<tr>
<td>L(θ)</td>
<td></td>
<td>0.000</td>
<td>0.059</td>
<td>0.066</td>
<td>0.050</td>
<td>0.031</td>
<td>0.016</td>
<td>0.067</td>
</tr>
<tr>
<td>LOD(θ)</td>
<td>-∞</td>
<td>0.577</td>
<td>0.623</td>
<td>0.509</td>
<td>0.299</td>
<td>0.000</td>
<td>0.632</td>
<td></td>
</tr>
</tbody>
</table>

- A row could be added for each marker
- Typically, $LOD = -\infty$ when $\theta = 0$... why?
- Marker closest to disease locus has highest LOD
  - And relatively small $\theta$
Example: A Less Informative Pedigree

- Which allele is on the disease haplotype?
- How often does it co-segregate with disease?
- How often is it separated from the disease allele?
Two possibilities: the disease allele might original occur in a chromosome with marker allele 3 or 1.

Either 5 non-recombinants and 1 recombinant…

… or 5 recombinants and 1 non-recombinant.
The next step

- Likelihood as a function of $\theta$?
- LOD score as function of $\theta$?
- Maximum likelihood estimate of $\theta$?
  - Corresponds to MLOD, maximum LOD score
Exemplar LOD Score Calculation (Phase Ambiguity for One Individual)

\[ L(\theta) = \frac{1}{2} (1 - \theta)^{\text{phase1\_nr}} \theta^{\text{phase1\_r}} + \frac{1}{2} (1 - \theta)^{\text{phase2\_nr}} \theta^{\text{phase2\_r}} \]

\[ = \frac{1}{2} (1 - \theta)^5 \theta + \frac{1}{2} (1 - \theta)\theta^5 \]

\[ LOD = \log_{10} \frac{L(\hat{\theta})}{L(\theta = \frac{1}{2})} = \]

\[ = \log_{10} \frac{L(0.168)}{L(\frac{1}{2})} \]

\[ = 0.33180 \]

In this case, the maximum possible LOD score would be \( \sim 1.5 \), for a marker where one of the two phases results in zero recombinants.
### Table of Two Point LOD Scores

<table>
<thead>
<tr>
<th>Recombination Fraction ((\Theta))</th>
<th>MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.030</td>
</tr>
<tr>
<td>0.2</td>
<td>0.033</td>
</tr>
<tr>
<td>0.3</td>
<td>0.026</td>
</tr>
<tr>
<td>0.4</td>
<td>0.019</td>
</tr>
<tr>
<td>0.5</td>
<td>0.016</td>
</tr>
<tr>
<td>0.156</td>
<td>0.033</td>
</tr>
</tbody>
</table>

- A row could be added for each marker
- Again, LOD = \(-\infty\) when \(\theta = 0\)...
- Marker closest to disease locus has highest LOD
  - And relatively small \(\theta\)
In practical settings ...

- LOD score calculations require evaluating and summing over many alternatives …
  - Segregation patterns for disease locus
  - Phases for ambiguous individuals
  - Alternatives for missing genotypes

- Penetrances may vary between individuals or groups of individuals

- Nowadays, LOD calculations are usually left to specialized computer programs…
  - Using either the Elston-Stewart or Lander-Green
An Uninformative Pedigree

For some markers, it may not be possible to count recombinants at all...
Most markers will not be informative for all the meiosis of interest

LOD scores will vary with θ and marker informativeness

Multipoint analysis uses a map of genetic markers to reconstruct inheritance along a chromosome.
  - Calculate LOD score by comparing each possible location to an unlinked locus.
Multipoint LOD Score Plot (For X-Linked Type of Blindness)
Parametric Linkage Analysis, Elston Stewart Algorithm

- Include disease locus in haplotype and haplo-genotype when evaluating likelihood

- Trait model affects all three components:
  - Probabilities for founder alleles
  - Probabilities of offspring given parents
  - Probabilities of observed data given genotypes
Parametric Linkage Analysis, Lander Green Algorithm

\[ P(X \mid I) = \sum_{a_1} \cdots \sum_{a_{2f}} \prod_i P(a_i) \prod_j P(X_j \mid a, I) \]

- Sum over all allele states for each founder
- For the last factor, notice that founder allele states and inheritance vector determine every genotype
Lander Green Algorithm

- One way to calculate LOD scores would be to “plug in” the trait locus at different positions and compare likelihoods …

- But this should be done carefully, to avoid unnecessary calculations…
Likelihood Ratio Test, Fully Informative Data

- Evaluate evidence for linkage as…

\[
LR(I) = \frac{P(X \mid I_{\text{observed}})}{\sum_{i \in I^*} P(X \mid i) P_{\text{uniform}}(i)}
\]

- Is a particular set of meiotic outcomes likely for a given trait model?
Allowing for uncertainty...

- Weighted sum over possible meiotic outcomes...

\[
LR = \sum_{i \in I^*} LR(i) P(i \mid G)
\]

\[
= \frac{\sum_{i \in I^*} P(X \mid i) P(i \mid G)}{\sum_{i \in I^*} P(X \mid i) P_{uniform}(i)}
\]
Concepts for Today ...

- Parametric linkage analysis
- Two-point analysis
- Multipoint analysis
- Mapping of Mendelian traits
Polio Symposium on Tuesday

- Rackham Auditorium, 9:30
  - No class!

- Polio used to affect 1:5,000 children

- Award in memory of Thomas Francis, Jr., of the University of Michigan School of Public Health, who directed field trials for the vaccine