Parametric Linkage Analysis

- Oldest (Newton Morton, 1950’s) and most statistically complete (full likelihood model) analysis method.

- “Parametric” – (1) the input must include parameters defining how we think the genotypes at the trait locus influence the trait phenotype, a.k.a. the mode or model of inheritance, or the penetrance values; (2) the output includes an estimate of a parameter describing the degree of linkage, i.e., location, not just the degree of significance.
Overview

• Parametric Linkage (and NPL) is a pedigree-based analysis, i.e., all results are due to relationships within the pedigrees, no analysis is done between pedigrees. However, one can combine results from many pedigrees to get an overall result.

• If the region of interest is smaller than 1 – 2 Mb, then there will be very few recombinations in this region within any one pedigree, no matter how many markers are added. Since Parametric Linkage only uses results within pedigrees, it will be less useful in this small region.
Association (including Haplotype Analysis) is a population-based method, i.e., results take into account data across all pedigrees. Association analysis can give significant results in small regions. However, these significant results can usually only be found over smaller distances.
More Overview

- There are Many Programs but Few Methods. Don’t just try every program you can get your hands on.

- If only one program gives you good results, be suspicious. If results are not robust to small perturbations in your data (including the model), be suspicious.

- In fact, always be suspicious. Don’t treat the programs as impenetrable black boxes. Know their assumptions and shortcomings.
Complex Traits Require a Penetrance Function (a.k.a. Mode of Inheritance)

- The input parameter in Parametric Linkage is the model of how the genotype at the trait locus influences the trait phenotype.

- The model is specified as a penetrance function: \( Pr(\text{phenotype} \mid \text{genotype}) \)

- For example,

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Normal</th>
<th>Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1/2</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2/2</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Penetrance Functions

- Another example,

<table>
<thead>
<tr>
<th></th>
<th>1/1</th>
<th>1/2</th>
<th>2/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>affected</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

- More generally,

<table>
<thead>
<tr>
<th></th>
<th>1/1</th>
<th>1/2</th>
<th>2/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.99</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>affected</td>
<td>0.01</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

- You can also have liability classes, for example, decade of life, smoking, value at second trait, etc.

- Several conditions can increase the phenocopy rate; but the value here should still probably be small (but if using a multipoint analysis, then it should definitely be positive).
General Pedigree Likelihood Function

\[ L = \sum_{G_1} \cdots \sum_{G_n} \prod_j \text{Prior}(G_j) \prod_{\{m,k,l\}} \text{Trans}(G_m | G_k G_l) \prod_i \text{Pen}(X_i | G_i) \]

- \( G_p \) runs through all possible multilocus-genotypes at person \( p \)
- \( X_i \) is the phenotype (observed genotype) at person \( i \)
- \( j \) runs over all founders; \( i \) runs over all people; \( \{m,k,l\} \) runs over all parent offspring triples
LOD Scores

• Given data consisting of trait phenotypes and genotypes at a marker locus, one can form the likelihood ratio, 
\[
\text{Likelihood(Data | Trait at position } d) \\
\text{Likelihood(Data | Trait unlinked)}
\]

• This is the odds that the trait locus is at position \( d \). The logarithm (base 10) of the ODds is the LOD score. The log allows summing over independent pedigrees.

• One can either search for the location \( d \) that maximizes the LOD or calculate the LODs at a predefined set of locations (e.g., 0.01, 0.05, 0.1, ..., 0.5).
Location Scores

- Exactly analogous to single point analysis, we again want to form the likelihood ratio test,
  \[ \frac{\text{Likelihood(Data | Trait at position } d)}{\text{Likelihood(Data | Trait unlinked)}} \]

- Here the marker map is considered fixed in place and the putative positions for the Trait are on a grid between the markers.

- The log of this LR again allows for summing over independent pedigrees and is equivalent to a multipoint LOD.
Multipoint Analysis

- Parametric Linkage can be used with many markers and large pedigrees, using the method of Location Scores

- Multipoint has the advantage of using all your data simultaneously; it can turn uninformative markers, informative

- As with all multipoint analyses, it is crucial that the map order be correct. The marker positions may be approximate but their order must be right. Choose markers accordingly.
Locus Heterogeneity

- When more than one trait locus is suspected, the $\Pr(\text{any pedigree is segregating a disease gene linked to the current position, } \theta )$ is called $\alpha$

- Moreover, one can simultaneously obtain MLE for $\theta$ and $\alpha$ in an unbiased fashion using Parametric Linkage Analysis

- In addition, at any $\theta$ one can obtain a Heterogeneous LOD (HLOD) which is maximized over all $\alpha$. This is the score one should normally use for complex traits.
Multipoint versus Two Point Analysis

• Two Point (aka Single Point) analysis is more flexible in allowing the data to fit the model, since $\theta$ is not constrained by neighboring markers. That is, for models of inheritance that are not precise (and none are), the maximum likelihood estimate for $\theta$ will be less accurate but the LOD score will still be accurate.

• Multipoint has more places that data error can enter the problem; but it can also help you find these errors.
How much power does linkage analysis have?

Power to detect linkage depends strongly on the magnitude of the contribution that the disease locus makes to the genetic variation of the trait.

\[ \text{Obs}_\text{Geno} \rightarrow \text{Obs}_\text{Pheno} \]

\[ \text{G}_\text{Markers} \rightarrow \text{G}_\text{Locus} \]

\[ \text{G}_\text{Locus} \rightarrow \text{Mode of inheritance} \]

\[ \text{Pheno} \rightarrow \text{Env} \]

\[ \text{Tested correlation} \]

Figure from Joe Terwilliger
What factors influence power?

- Pedigree size and structure
- Sample size
- Marker informativeness
- Distance of marker from disease locus
- Phenocopy rate
- Genetic heterogeneity
- Magnitude of genetic effect
Design Issues

• Pre-specify a few trait models, e.g., reduced penetrance dominant and recessive, and an additive model. Don’t need to be exactly right

• Simplify problem with either a highly specific phenotype or an isolated population

• After positive result, test for model robustness

• Replication is vital
Assumptions

• Hardy-Weinberg Equilibrium

• Linkage Equilibrium!

• Random Mating

• No Chiasma Interference (Haldane map function)

• No Epistasis, i.e., no interaction between alleles at different loci
Weaknesses of PL

• Bilinearity sensitivity

• Explicitly Model-based (although other methods are often implicitly model-based)

• Analysis only within pedigrees, not across pedigrees

• Originally designed for simple traits, so it is difficult to include multiple trait loci simultaneously

• Localized region will very rarely be less than 1 Mb wide, more often around 5 Mb
Strengths of PL

• Not sensitive to allelic heterogeneity nor population history

• No candidate genes to guess

• Well developed 400 and 800 marker sets for genome scan: cheap and potentially few mistypings

• Can detect a signal up to 20 cM away

• P-value (LOD score) is accurate given the data (including the model); for other methods estimated p-values may not be accurate, either too conservative or anti-conservative
Overview of current general-pedigree linkage analysis programs

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Programs</th>
<th>Solution</th>
<th>Size Restriction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elston-Stewart</td>
<td>FASTLINK</td>
<td>exact</td>
<td>varies: ~8 loci, less with loops</td>
</tr>
<tr>
<td></td>
<td>LINKAGE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mendel</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vitesse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lander-Green</td>
<td>Allegro</td>
<td>exact</td>
<td>~16 people (2n − f ≤ 16)</td>
</tr>
<tr>
<td></td>
<td>GeneHunter</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mendel</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Merlin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markov chain Monte Carlo</td>
<td>Loki</td>
<td>estimate</td>
<td>much larger (&gt; 1000 individuals,</td>
</tr>
<tr>
<td></td>
<td>SimWalk</td>
<td></td>
<td>&gt; 1000 loci)</td>
</tr>
</tbody>
</table>

| Algorithm                  | Increase in computational time with increase in: |  |
|----------------------------|-------------------------------------------------|  |
|                            | people                                          | markers | missing data                  |
| Elston-Stewart             | linear                                          | exponential | severe                      |
| Lander-Green               | exponential                                     | linear   | modest                       |
| Markov chain Monte Carlo   | linear                                          | linear   | mild                         |

02 November 2003 Mendel Short Course @ IGES
Mendel Keywords for Location Score Analysis

- **NUMBER_OF_MARKERS INCLUDED** sets the number of surrounding markers used in the calculations for each marker interval.

- **TRAVEL** determines if a **SEARCH** is made for best location or calculations are over a **GRID** of positions. (The grid is influenced by the keywords: **INTERIOR_POINTS**, **GRID_INCREMENT**, & **STANDARD_GRID**.)

- **ESTIMATE_LINKED_PROPORTION** determines if the $\alpha$ value will be estimated. (**LINKED_PROPORTION** sets the fixed or initial $\alpha$ value.) **PROBABILITY_PEDIGREE_LINKED** determines if posterior probabilities are calculated for each pedigree.