The structure of genetic data for linkage analysis
Introduction to linkage lod scores

Elizabeth Thompson
University of Washington.

Seattle meeting: Jan 28, 2010.
The Elston-Stewart framework (1971)

Key to variables:
G: genotypes (phased)
Y: phenotypes (observed)

Population

\[ G^{(F)} \] Founders

Transmission

\[ G \] All

Penetrance

\[ Y \] Observed

Model

Latent variables

Observables
Joint computations; arbitrary pedigrees, (1976,1978)

Cannings, Thompson, Skolnick: OCIGGOP:
offspring conditionally independent given genotypes of parents.

Cutset $C$ of individuals divides the pedigree; peeled set may be above some cutset individuals ($C^*$: parents are peeled), and below others ($C^\dagger$: offspring are peeled)

$$R_C(g_c) = \Pr(\text{data peeled}, g_{c^*} | g_{c^\dagger})$$

Equations still the same: population, transmission, penetrance.

Now we have genetic maps!
Transmission involves recombination (male/female) and interference.
Computations become intensive.
Downcoding, and similar ideas, help but do not solve.
The linkage likelihood and lod score

Observed data $\mathbf{Y} = \{Y_i; \ i \text{ observed}\}$, and model $\Gamma$, then

$$L(\Gamma) = \Pr(\mathbf{Y}; \Gamma) = \sum_G \Pr(\mathbf{Y} | G)P(G)$$

$$= \sum_G \left( \prod_{i \in \mathcal{F}} \Pr(G_i) \right) \left( \prod_{i \in \mathcal{N}} \Pr(G_i | G_{M(i)}, G_{F(i)}) \right) \left( \prod_{i \in \mathcal{O}} \Pr(Y_i | G_i) \right)$$

where $\mathcal{F}$, $\mathcal{N}$ and $\mathcal{O}$ denote the sets of founders, non-founders, and observed individuals, respectively.

For trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$

$$\text{lod} = \log_{10} \frac{\Pr(\mathbf{Y}_T, \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T, \mathbf{Y}_M; \Gamma_0)}$$

where $\Gamma_0$ is $\Gamma$ with no $T/M$ linkage

$$= \log_{10} \frac{\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T; \Gamma)\Pr(\mathbf{Y}_M; \Gamma)}$$

$$= \log_{10} \frac{\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T; \Gamma)}$$
• Definition of $S$ for multiple loci, $j, j = 1, \ldots, \ell$,

$$S_{i,j} = 0 \quad \text{if gene at meiosis } i \text{ position } j \text{ is parent's maternal}$$

$$= 1 \quad \text{if gene at meiosis } i \text{ position } j \text{ is parent's paternal.}$$

• Models for meiosis ($S_{i,\bullet}$) for inheritance vectors ($S_{\bullet,j}$):

$$S_{i,\bullet} = \{S_{i,j}; j = 1, \ldots, \ell\}, \quad i = 1, \ldots, m$$

$$S_{\bullet,j} = \{S_{i,j}; i = 1, \ldots, m\}, \quad j = 1, \ldots, \ell$$

where $m$ is the number of meioses in the pedigree, and $\ell$ the number of loci along the chromosome.

• Space of $S$ smaller than of $G$.

and dependence structure of $S$ is simpler (we assume).

• Transmission provides model for each $S_{i,\bullet}$,

and $S_{i,\bullet}$ are independent over $i, i = 1, \ldots, m$. (Mendel’s first law).

$S_{i,j}$ are dependent among loci $j$ on the same chromosome pair
Multiple genetic markers: the GENEHUNTER model (1987)

Meiosis model is primary:
Markov dependence of inheritance gives
\[
\Pr(Y) = \sum_S \Pr(Y \mid S) \Pr(S) = \sum_S \left( \prod_{j=1}^{l} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right) \Pr(S_{\bullet,1}) \left( \prod_{j=2}^{l} P(S_{\bullet,j} \mid S_{\bullet,j-1}) \right)
\]
Population models provide founder allelic types: and hence \(P(Y_{\bullet,j} \mid S_{\bullet,j})\), in principle.
But penetrance models (and hence traits, or marker error) are a computational problem in this framework.
The Lange-Sobel (1991) formulation

For trait data $Y_T$ and marker data $Y_M$

$$\text{lod} = \log_{10} \frac{\Pr(Y_T \mid Y_M; \Gamma)}{\Pr(Y_T; \Gamma)}$$

Now $\Gamma = (\Gamma_M, \Gamma_T)$, with $S_M$ marker location inheritance:

$$\Pr(Y_T \mid Y_M; \Gamma) = \sum_{S_M} \Pr(Y_T \mid S_M; \Gamma_T) \Pr(S_M \mid Y_M; \Gamma_M)$$

$$= E_{\Gamma_M} (\Pr(Y_T \mid S_M, \Gamma_T) \mid Y_M)$$

Marker analysis to provide $S_M$ given $Y_M$ once only; then multiple trait models, and trait-locus locations.

End of Intro part
Part 2: Lod scores using multiple dense markers

- MORGAN: lm_markers and lm_multiple
- MCMC approaches to sampling $S_M$ given $Y_M$
- Computing on the $ibd$ graph for markers
- Computing on the $ibd$ graph for traits
- New framework for lod scores using dense markers — the MORGAN gl_auto program and the $ibd$ graph library
Introduction to MORGAN

- MORGAN online tutorial: MORGAN does many kinds of analysis.
- With many loci, we cannot use Elston-Stewart approach.
- With large pedigrees, we cannot use HMM approach.
- Exact computation fails.

- Lodscore programs lm_markers and lm_multiple (up to 2008)
  - use the Lange-Sobel framework
  - sample $S_M$ given $Y_M$ by MCMC
    (or i.i.d on small pedigree components)
MCMC sampling of $S_M$ given $Y_M$

- For each sampled $S_M$, each hypothesized trait location, each trait penetrance model desired,
  - compute $P(Y_T \mid S_M)$
  - estimate $P(Y_T \mid Y_M)$ by mean $P(Y_T \mid S_M)$.
  - and hence the lod score.

Block-Gibbs MCMC Samplers:

L-sampler: resample $S_{\bullet, j}$ given $Y$ and $S_{\bullet, j'}, j \neq j'$

M- (or MM-) sampler: resample \( \{S_{i, \bullet}; i \in I^*\} \) given $Y$ and \( \{S_{i', \bullet}; i' \notin I^*\} \)

LM-Sampler: computational complexity, and mixing performance

L-sampler: requires peeling over the pedigree to resample $S_{\bullet,j}$.

$$\Pr(Y_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1}) = \sum_{S_{\bullet,j}} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1})$$

and to compute $\Pr(Y_T \mid S_M) = \Pr(Y_T S_{\bullet,j-1}, S_{\bullet,j+1})$

M-sampler: requires peeling along the chromosome (Baum, HMM, or fwds-backward algorithm)

$$\Pr(\{S_{i,\bullet}; i \in I^*\} \mid Y, \{S_{i',\bullet}; i' \notin I^*\})$$

L-sampler mixes poorly for tight linkage
M-sampler mixes poorly on extended pedigrees.
L-sampler is irreducible (theoretically).
Together they can do better.
S defines the Identity-by-descent (ibd) graph

- Nodes are distinct genome labels.
- Edges are observed individuals.

Only *ibd* matters, not founder origins: the *ibd* graph is an equivalence class of S w.r.t data probabilities at a locus.
Given \emph{ibd}-graph; probabilities for marker data.
Sobel & Lange, 1996; Kruglyak et al. (1996)

- If allelic types are independent, over loci and over genomes: each node $g$ has type $a_k$ independently with prob $q_k$. $\Pr(\mathcal{A}) = \prod_g q(\mathcal{A}(g)) = \prod_k q_k^{n(k)}$ where $n(k)$ is number of nodes $g$ with type $a_k$.

- $\Pr(Y_{\bullet,j} \mid S) = \sum \Pr(\mathcal{A}_j)$: sum over all $\mathcal{A}_j$ consistent with $Y_{\bullet,j}$.

\begin{itemize}
  \item Suppose $A$, $B$, $J$ are all $a_1a_4$, $G$ is $a_1a_6$, $D$ is $a_4a_6$, $E$ is $a_4a_2$, $C$ is $a_2a_2$, $F$ is $a_3a_6$, $H$ is $a_2a_3$, and $L$ is $a_1a_3$.
  \item Then 2 is $a_1$; 9, 13 are $a_4$; 4 is $a_6$; 6 is $a_2$; 15 is $a_3$, and 17 is $a_1$. The probability is $q_1^2 q_2 q_3 q_4^2 q_6$.
  \item There are always 2, 1, or 0 possible $\mathcal{A}_j$.
  \item Probabilities multiply over disjoint components.
  \item Ignore “missing” nodes: contribute a factor 1.
\end{itemize}
Given $S$: data probabilities on the $ibd$-graph

- The distinct nodes $g_i$ have independent allelic types; say type $A(g_i)$ with probabilities $q(A(g_i))$.
- The allelic types of the two nodes of an observed individual $n$ determine his genotype, and hence probabilities of his observed data $Y_n$: $\Pr(Y_n|A_{n,1}, A_{n,2})$ independently for each $n$.

$$\Pr(Y|ibd) = \sum_{A(g)} \left( \prod_n \Pr(Y_n|A_{n,1}, A_{n,2}) \right) \left( \prod_i q(A(g_i)) \right)$$

- Disjoint components of $ibd$-graph are independent.
- Only genome nodes present in observed individuals matter.

- Much easier than computing on the pedigree structure, where summation over unobserved types is required. (Thompson, Heath, 1997, 1999, 2005).
Given *ibd*-graph: computations with penetrance

\[
\text{Pr}(Y|\text{ibd}) = \\
\sum_{A_6} q(A_6) \text{Pr}(Y_C|A_6) \left( \sum_{A_{15}} q(A_{15}) \text{Pr}(Y_H|A_{15}, A_6) \\
\sum_{A_{17}} q(A_{17}) \text{Pr}(Y_L|A_{15}, A_{17}) \left( \sum_{A_4} q(A_4) \text{Pr}(Y_F|A_{15}, A_4) \\
\sum_{A_{13}} q(A_{13}) \text{Pr}(Y_E|A_6, A_{13}) \text{Pr}(Y_D|A_4, A_{13}) \right) \\
\sum_{A_2} q(A_2) \text{Pr}(Y_B|A_2, A_{13}) \right) \text{Pr}(Y_J|A_2, A_{13}) \text{Pr}(Y_G|A_2, A_4) \\
\left( \sum_{A_9} q(A_9) \text{Pr}(Y_A|A_2, A_9) \right)
\]

16
Markers are (too) numerous; traits are complex

• Now we have all components of the lod score:

\[
\Pr(Y) = \sum_{S} \Pr(Y | S) \Pr(S)
\]

\[
= \sum_{S} \left( \prod_{j=1}^{l} \Pr(Y_{\cdot,j} | S_{\cdot,j}) \right) \Pr(S_{\cdot,1}) \left( \prod_{j=2}^{l} P(S_{\cdot,j} | S_{\cdot,j-1}) \right)
\]

for markers (with error) and single-locus traits.

• But SNP markers are numerous; if \( l \) is large, even an algorithm linear in \( l \) has problems.

• Even with dense SNP data \( S \) is quite uncertain, but the \textit{ibd}-graph in each chromosome region is often well determined.

• Traits are complex (oligogenic): \textit{ibd}-graphs are sufficiently simple, that computation of phenotype probabilities is feasible even where the (trait) data depend of \textit{ibd}-graphs jointly at several loci.
Separating traits and markers

• Given marker data $Y_M$, sample $S_M$ (jointly) or compute $S_{\bullet,j}$ (marginally) for markers, once and once only: sample of $D_M$.

• At (dense) marker locations $D_T$ for trait locus (or loci), is as for markers.

• Use trait model to compute $P(Y_T \mid D_T)$, which is the contribution to the lod score for this $S_M$.

• Compute probabilities of trait data, given $S_M$, for multiple trait locations/models/.... hence lod scores.
gl_auto output and ibd-graph equivalence

- MORGAN program gl_auto outputs multiple realizations of the genome labels of all individuals across a chromosome in a compact format: only changes in genome label are recorded, and these are few.

- $D_M$ is constant over multiple markers
- Many different $S_M$ give same (unlabeled) $D_M$
- Many different realizations of $S_M$ give same (unlabeled) $D_M$ – and even more, the same $D_T$. 
ibd-graph equivalence and the \textit{ibd} graph library

- A library of C-functions has been written by Hoyt Koepke

1) hash coding is used to uniquely identify graphs,
2) The (unlabeled) nodes of a graph have an identity only through the (labeled) edges that connect them, and
3) any feature of the graph has a marker-range over which it exists.

- An M-Table is a hash table augmented to hold additional information.
The M-table allows for efficient insertion, querying, equality testing, and set operations on the collection of \textit{ibd}-graphs.

- This reduction can be an order of magnitude or more, and is already being used in real-data examples.
Conclusion

- With multiple markers, marker computations should be done once only.

- With dense markers, marker inheritance should be imputed/stored by chromosome segment, not marker by marker.

- Multiple realizations of marker-based \textit{ibd} over the chromosome can then be used in trait analyses.

\textit{bullet} Software to recognize equivalence of \textit{ibd}-graphs can greatly increase efficiency.