Multipoint Analysis for Sibling Pairs

Biostatistics 666

Previously ...

- Linkage analysis with pairs of individuals
  - “Non-parametric” IBS Methods
  - “Maximum Likelihood” IBD Based Method
  - Possible triangle constraint
- Examined one marker at a time

IBS Based Linkage Test

$$\chi^2_{2df} = \sum_i \frac{\left[N_{IBS=i} - E(N_{IBS=i})\right]^2}{E(N_{IBS=i})}$$

- Expect counts calculated using:
  - Allele frequencies for marker
  - Relationship for affected individuals

Likelihood for a Single ASP

$$L_i = \sum_{j=0}^{2} P(IBD = j \mid ASP) P(Genotypes \mid IBD = j) = \sum_{j=0}^{2} z_j w_{ij}$$

Risch (1990) defines

$$w_{ij} = P(Genotypes_i \mid IBD = j)$$

We only need proportionate w_{ij}

MLS Linkage Test

$$L(z_0, z_1, z_2) = \prod_i \sum_j z_j w_{ij}$$

$$LOD = \log_{10} \prod_i \frac{z_0 w_{i0} + z_1 w_{i1} + z_2 w_{i2}}{\frac{1}{4} w_{i0} + \frac{1}{2} w_{i1} + \frac{1}{4} w_{i2}}$$

The MLS statistic is the LOD evaluated at the MLEs of z_0, z_1, z_2

Important Limitation

- A major limitation of all these models is that they consider only one marker at a time
- This may not allow us to extract all available information about IBD…

Today ...

- Refresher on IBD probabilities
- Intuition behind multipoint calculations
- Framework for multipoint calculations
- Using a Markov Chain to speed analyses

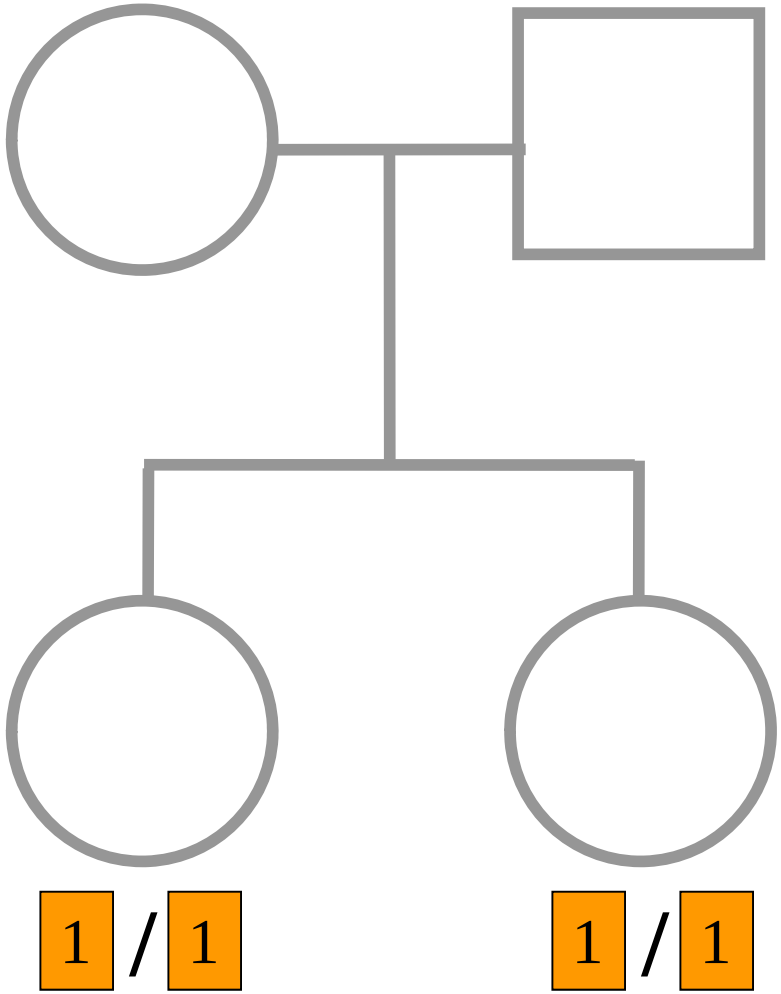
IBD Probabilities

- Number of alleles identical by descent
- For sibling pairs, must be:
  - 0
  - 1
  - 2
- Not always determined by marker data

Refresher ...
Bayes Theorem for IBD Probabilities

$$
\begin{aligned}
P(IBD = i \mid X) &= \frac{\mathrm{P}(\mathrm{IBD} = i, X)}{P(X)} \\
&= \frac{P(IBD = i)P(X \mid IBD = i)}{P(X)} \\
&= \frac{P(IBD = i)P(X \mid IBD = i)}{\sum_j P(IBD = j)P(X \mid IBD = j)}
\end{aligned}
$$

Worked Example



$$p_1 = 0.5$$

$$P(X \mid IBD=0) = p_1^4 = 1/16$$

$$P(X \mid IBD=1) = p_1^3 = 1/8$$

$$P(X \mid IBD=2) = p_1^2 = 1/4$$

$$P(X) = 1/4\, p_1^4 + 1/2\, p_1^3 + 1/4\, p_1^2 = 9/64$$

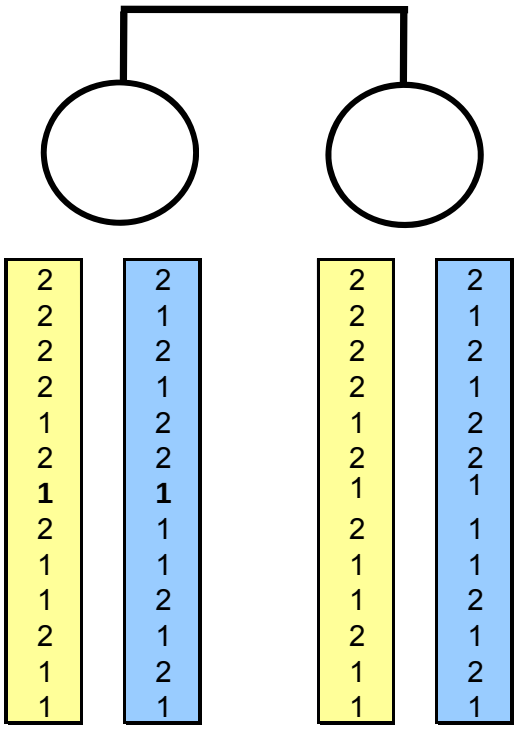
$$P(IBD=0 \mid X) = \frac{1/4\, p_1^4}{P(X)} = 1/9$$

$$P(IBD=1 \mid X) = \frac{1/2\, p_1^3}{P(X)} = 4/9$$

$$P(IBD=2 \mid X) = \frac{1/4\, p_1^2}{P(X)} = 4/9$$

Intuition For Multipoint Analysis

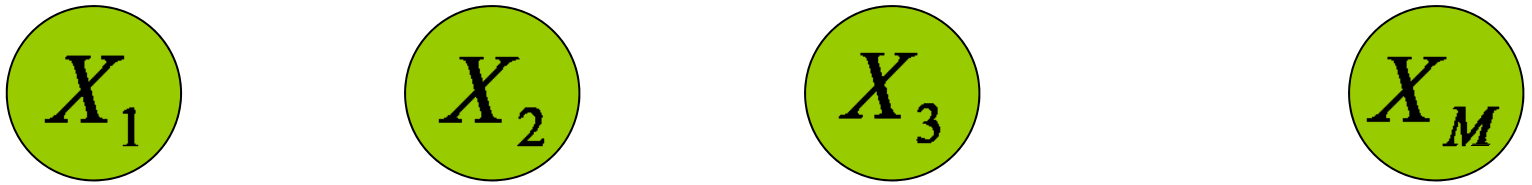
- IBD changes infrequently along the chromosome
- Neighboring markers can help resolve ambiguities about IBD sharing
- In the Risch approach, they might ensure that only one **w** is *effectively* non-zero



The Ingredients ...

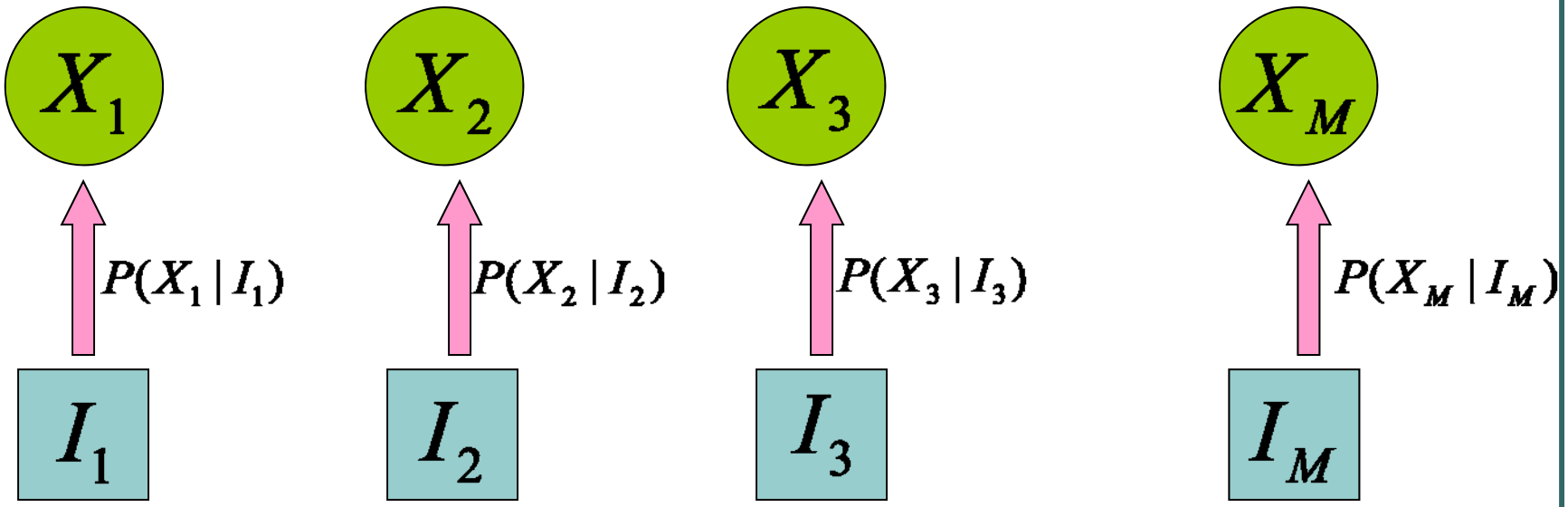
- Probability of observed genotypes at each marker conditional on IBD state
- Probability of changes in IBD state along chromosome
- Hidden Markov Model

Ingredients



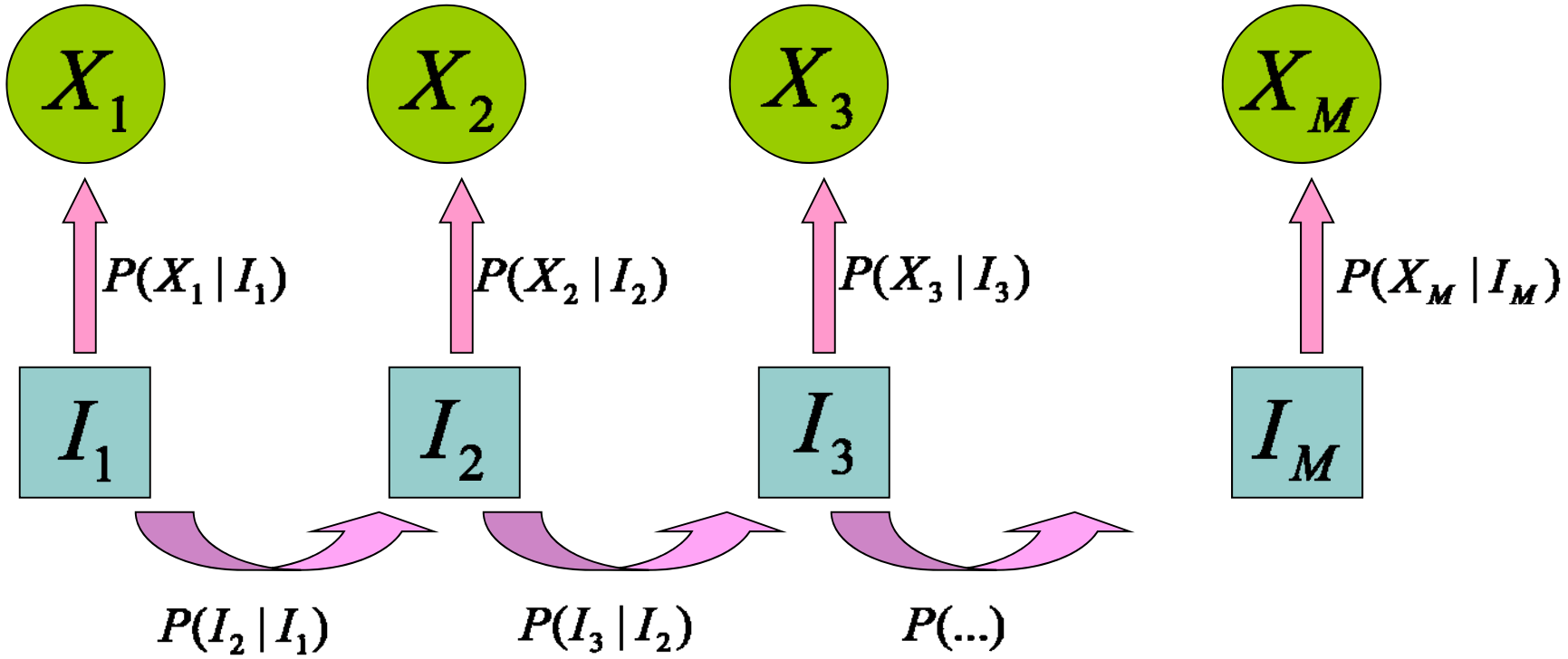
One ingredient will be the observed genotypes at each marker …

Ingredients



Another ingredient will be the possible IBD states at each marker …

Ingredients



The final ingredient connects IBD states along the chromosome …

The Likelihood of Marker Data

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

- General formulation, allows for any number of markers.
- Combined with Bayes' Theorem can estimate probability of each IBD state at any marker.
- This is not a linkage test yet!

$P(X_m \mid I_m)$

		IBD		
Sib	**CoSib**	**0**	**1**	**2**
(a,b)	(c,d)	$4p_ap_bp_cp_d$	0	0
(a,a)	(b,c)	$2p_a^2p_bp_c$	0	0
(a,a)	(b,b)	$p_a^2p_b^2$	0	0
(a,b)	(a,c)	$4p_a^2p_bp_c$	$p_ap_bp_c$	0
(a,a)	(a,b)	$2p_a^3p_b$	$p_a^2p_b$	0
(a,b)	(a,b)	$4p_a^2p_b^2$	$(p_ap_b^2+p_a^2p_b)$	$2p_ap_b$
(a,a)	(a,a)	p_a^4	p_a^3	p_a^2
Prior Probability		¼	½	¼

Note: Assuming unordered genotypes

Question:
What to do about missing data?

- What happens when some genotype data is unavailable?

$P(I_{m+1} \mid I_m)$

- Depends on recombination fraction θ
  - This is a measure of distance between two loci
  - Probability grand-parental origin of alleles changes between loci
- Naturally, leads to probability of change in IBD:

$$\psi = 2\theta(1-\theta)$$

$P(I_{m+1} \mid I_m)$

		IBD State at m + 1		
		0	1	2
IBD state at marker m	0	$(1-\psi)^2$	$2\psi(1-\psi)$	ψ^2
	1	$\psi(1-\psi)$	$(1-\psi)^2+\psi^2$	$\psi(1-\psi)$
	2	ψ^2	$2\psi(1-\psi)$	$(1-\psi)^2$

$$\psi = 2\theta(1-\theta)$$

Example

- Consider two loci separated by $\theta = 0.1$
- Each loci has two alleles, each with frequency .50
- If two siblings are homozygous for the first allele at both loci, what is the probability that IBD = 2 at the first locus?

The Likelihood of Marker Data

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

- 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

$$L_m(I_m) = P(X_1,\ldots,X_{m-1} \mid I_m)$$
$$= \sum_{I_{m-1}} L_{m-1}(I_{m-1}) P(X_{m-1} \mid I_{m-1}) P(I_{m-1} \mid I_m)$$
$$L_1(I_1) = 1$$

- 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.

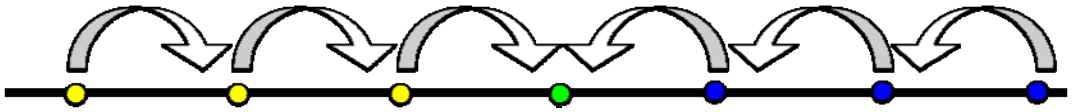
Pictorial Representation

- 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
  - What components might have to change?

Today

- Efficient computational framework for multipoint analysis of sibling pairs