The Lander-Green Algorithm in Practice

Biostatistics 666

Lecture 21

Last Lecture: Lander-Green Algorithm

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

- Similar multipoint sib-pair analysis, but with:
  - More general definition for I, the "IBD vector"
  - Probability of genotypes given “IBD vector”
  - Transition probabilities for the “IBD vectors”

Lander-Green Recipe

- 1. List all meiosis in the pedigree
  - There should be *2n* meioses for *n* non-founders
- 2. List all possible IBD patterns
  - Total of 2^{2n} possible patterns defined by setting each meiosis to one of two possible outcomes
- 3. At each marker location, score P(X|I)
  - Evaluate using each possible founder allele graph I

Lander-Green Recipe

- 4. Build transition matrix for moving along chromosome

$$T^{\otimes n+1} = \begin{bmatrix} (1-\theta)T^{\otimes n} & \theta T^{\otimes n} \\ \theta T^{\otimes n} & (1-\theta)T^{\otimes n} \end{bmatrix}$$

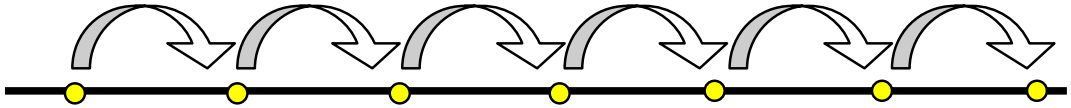
  - Patterned matrix, built from matrices for individual meioses

Lander-Green Recipe

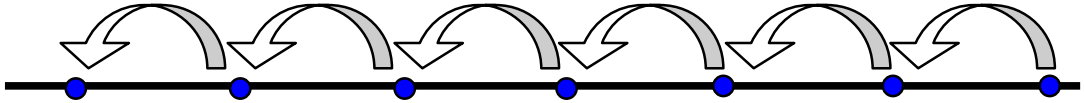
- 5. Run Markov chain
  - Start at first marker, *m*=1
    - Build a vector listing $P(G_{\text{first marker}}|I)$ for each I
  - Move along chromosome
    - Multiply vector by transition matrix
  - Combine with information at the next marker
    - Multiply each component of the vector by $P(G_{\text{current marker}}|I)$
  - Repeat previous two steps until done

Pictorial Representation

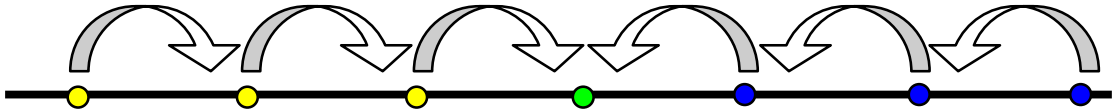
- Forward recurrence



- Backward recurrence



- At an arbitrary location



Today:
Lander-Green Algorithm in practice

- Common applications of the algorithm
  - Non-parametric linkage analysis
  - Parametric linkage analysis
  - Information content calculations
- Refining the Lander-Green algorithm
  - Speeding up transition step
  - Reducing size of inheritance space

Part I: Common Applications

- Non-parametric linkage analysis
- Parametric linkage analysis
- Information content calculation
  - Time permitting!

Nonparametric Linkage Analysis

- *Model-free*
- Does not require specification of a trait model
- Tests for excess IBD sharing among affected individuals

Non-parametric Analysis for Arbitrary Pedigrees

- Must rank general IBD configurations
  - Low scores correspond to no linkage
  - High scores correspond to linkage
- Multiple possible orderings are possible
  - Especially for large pedigrees
- Under linkage, probability for vectors with high scores should increase

Nonparametric Linkage Statistic

- Let *S(I)* be a statistic that ranks IBD vectors
- Then, following Whittemore and Halpern (1995)

$$S(G) = \sum_I S(I) P(I \mid G)$$

$$\mu = \sum_G S(G) P(G)$$

$$\sigma^2 = \sum_G \left[S(G) - \mu \right]^2 P(G)$$

$$Z = \frac{S(G) - \mu}{\sigma} \sim N(0,1)$$

Nonparametric Linkage Statistic

- Original definition not useful for multipoint data…
- Kruglyak et al (1996) proposed:

$$S(G) = \sum_I S(I)P(I \mid G)$$

$$\mu = \sum_I S(I)P(I)$$

$$\sigma^2 = \sum_I \left[S(I) - \mu\right]^2 P(I)$$

$$Z = \frac{S(G) - \mu}{\sigma} \sim N(0,1)$$

The Pairs Statistic

- Sum of IBD sharing for all affected pairs

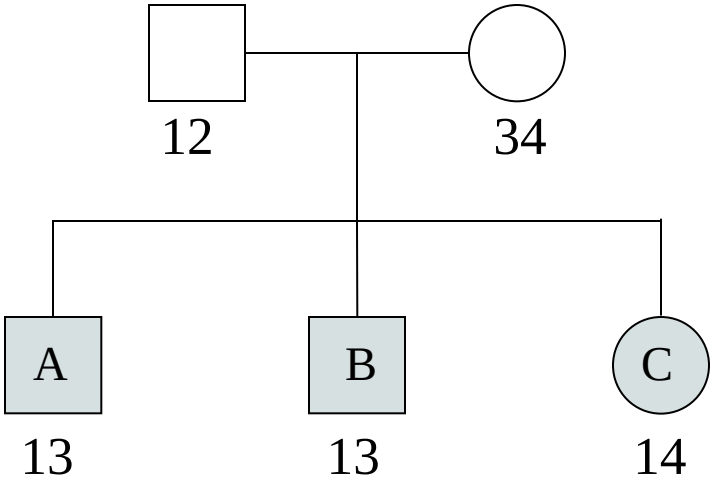
$$S_{pairs}(I) = \sum_{(a,b)\in(\text{affected pairs})} IBD(a,b \mid I)$$

$$\mu = \sum_{I} S_{pairs}(I) P_{uniform}(I)$$

$$\sigma^2 = \sum_{I} \left(S_{pairs}(I) - \mu\right)^2 P_{uniform}(I)$$

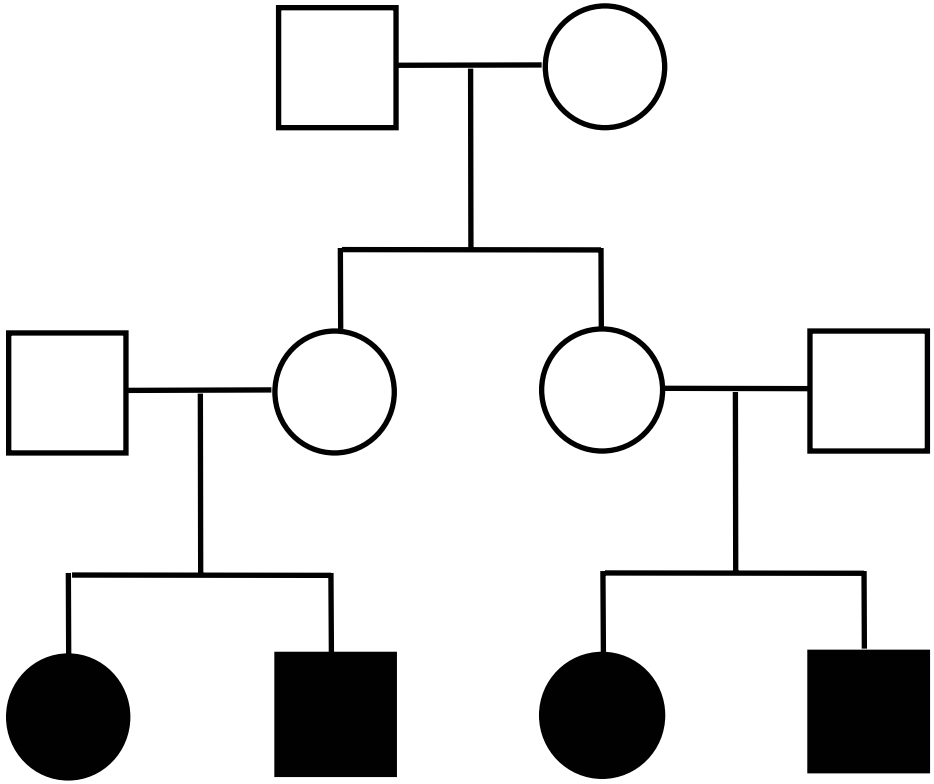
The S_{pairs} Statistic

- Total allele sharing among affected relatives

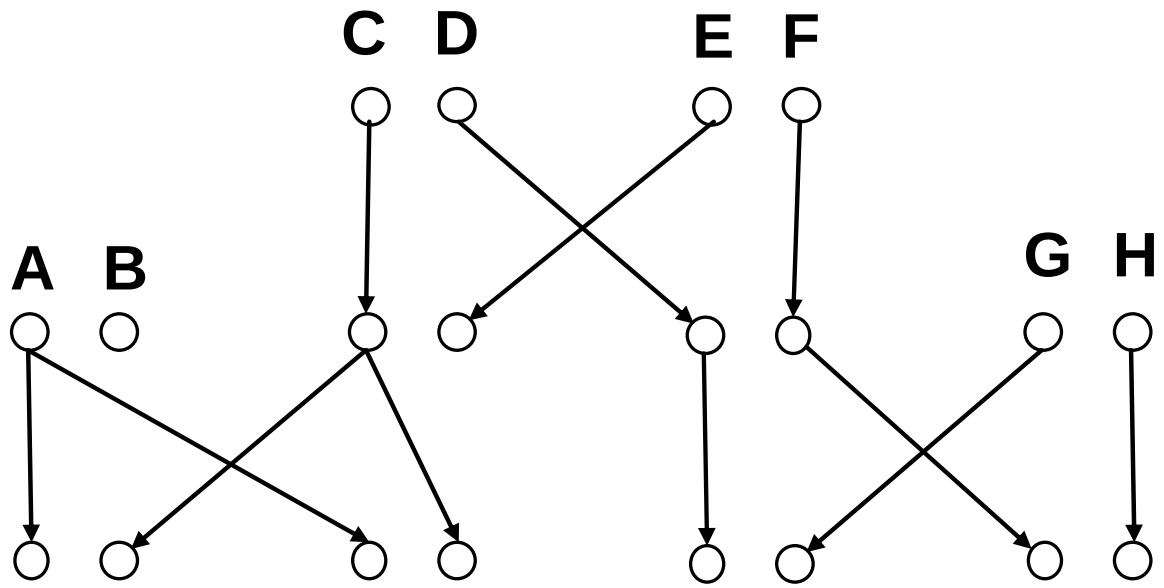


Sibpair:	A-B		A-C		B-C	
S_{Pairs} =	2	+	1	+	1	= 4

Example:
Pedigree with 4 affected individuals



What is $S_{pairs}(I)$ for this Descent Graph?



The NPL Score

- Non-parametric linkage score

$$Z(I) = \left(S_{pairs} - \mu\right) / \sigma$$

$$Z_{NPL} = \sum_{I} Z(I) P(I \mid G)$$

- Variance will always be ≤ 1 so using standard normal as reference gives conservative test.

Accurately Measuring NPL Evidence for Linkage

- For a single marker…

$$\sigma^2 = \sum_{i \in I^*} \sum_{G} \left(S_{pairs}(G) - \mu \right)^2 P(G|i) P_{uniform}(i)$$

- Estimating variance of statistic over all possible genotype configurations is not practical for multipoint analysis

- One possibility is to evaluate the empirical variance of the statistic over families in the sample… but Kong and Cox (1997) proposed a simple analytical solution

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 is proportional to S(I)
- "Generalization" of the MLS method

Kong and Cox Method

$$P(I|\delta) = P_{uniform}(I)\left(1+\delta\frac{S(I)-\mu}{\sigma}\right)$$

$$L(\delta) = \prod_{families}\sum_{I} P(G|I)P(I|\delta)$$

$$LOD = \log_{10}\frac{L(\hat{\delta})}{L(\delta=0)}$$

Notes: δ should be constrained so $0 < P(I|\delta) < 1$ for all I

: $\sum_{I} P(I|\delta) = 1$ for all δ since $E[\frac{S(I)-\mu}{\sigma}] = 0$

Note: Alternative NPL Statistics

- Any arbitrary statistic can be used
- Vectors with high scores must be more common when linkage exists
- Statistics have been defined that
  - Focus on the most common allele among affecteds
  - Count number of distinct founder alleles among affecteds
  - Evaluate linkage for quantitative traits

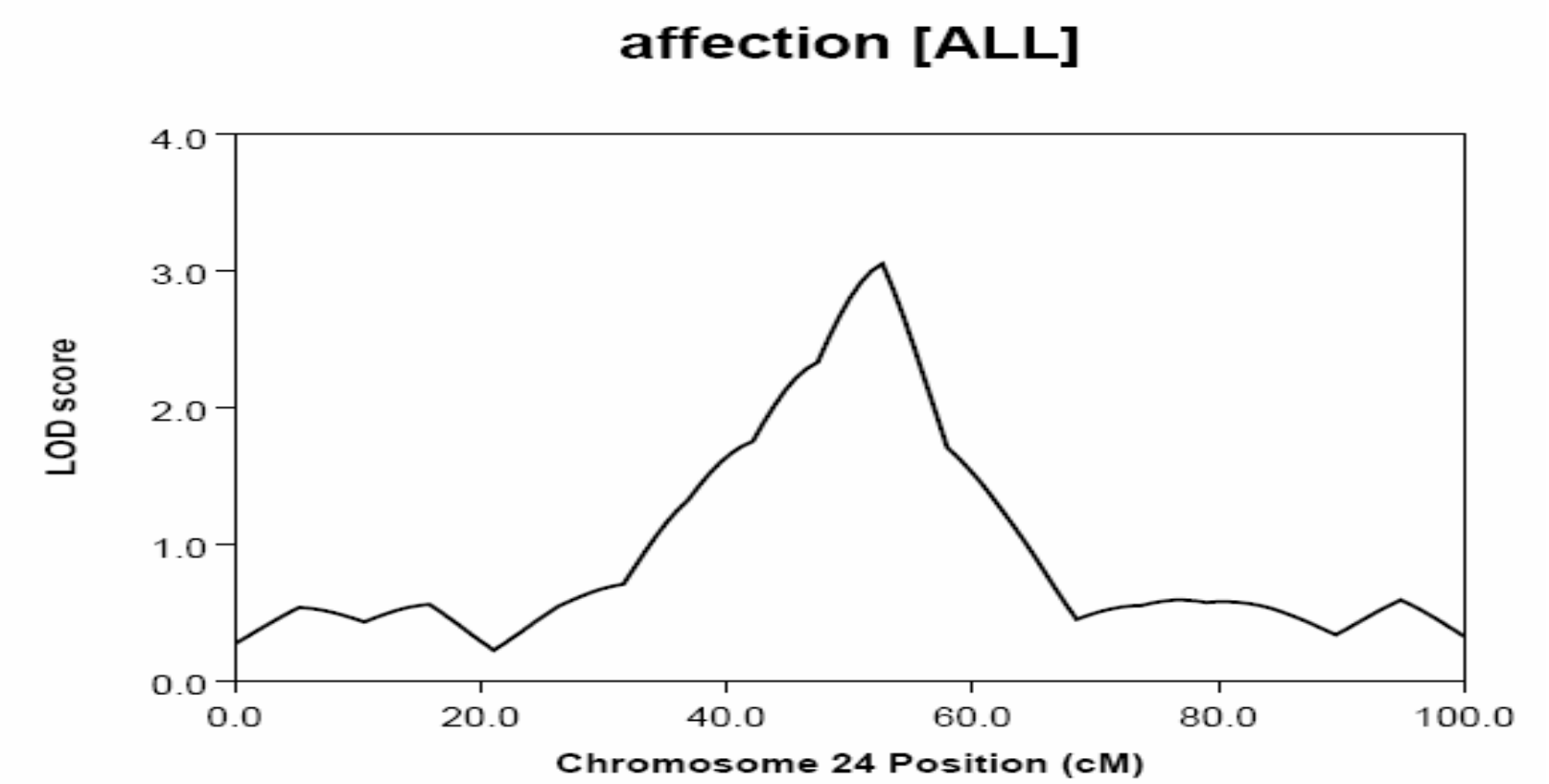
Many Alternative NPL Statistics!

TABLE I. Example 1: Outbred Sib Pair and First Cousin

Configuration (sib, sib, cousin)	Null prob.	$S_{pairs} - \mu_0$	$S_{all} - \mu_0$	$S_{-\#alleles} - \mu_0$	$S_{everyone} - \mu_0$	$S_{-\#geno} - \mu_0$	$S_{fewest} - \mu_0$
c_1 1 2 3 4 5 6	.125	–1.5	–.41	–1.375	–.125	–.25	–.0625
c_2 1 2 3 4 1 5	.125	–.5	–.16	–.375	–.125	–.25	–.0625
c_3 1 2 1 3 4 5	.3125	–.5	–.16	–.375	–.125	–.25	–.0625
c_4 1 2 1 3 2 4	.125	.5	.09	.625	–.125	–.25	–.0625
c_5 1 2 1 2 3 4	.1875	.5	.09	.625	–.125	.75	–.0625
c_6 1 2 1 3 1 4	.0625	1.5	.59	.625	.875	–.25	–.0625
c_7 1 2 1 2 2 3	.0625	2.5	.84	1.625	.875	.75	.9375

McPeek (1999) *Genetic Epidemiology* **16:**225–249

Non-Parametric Linkage Curve



Non-Parametric Linkage Scan

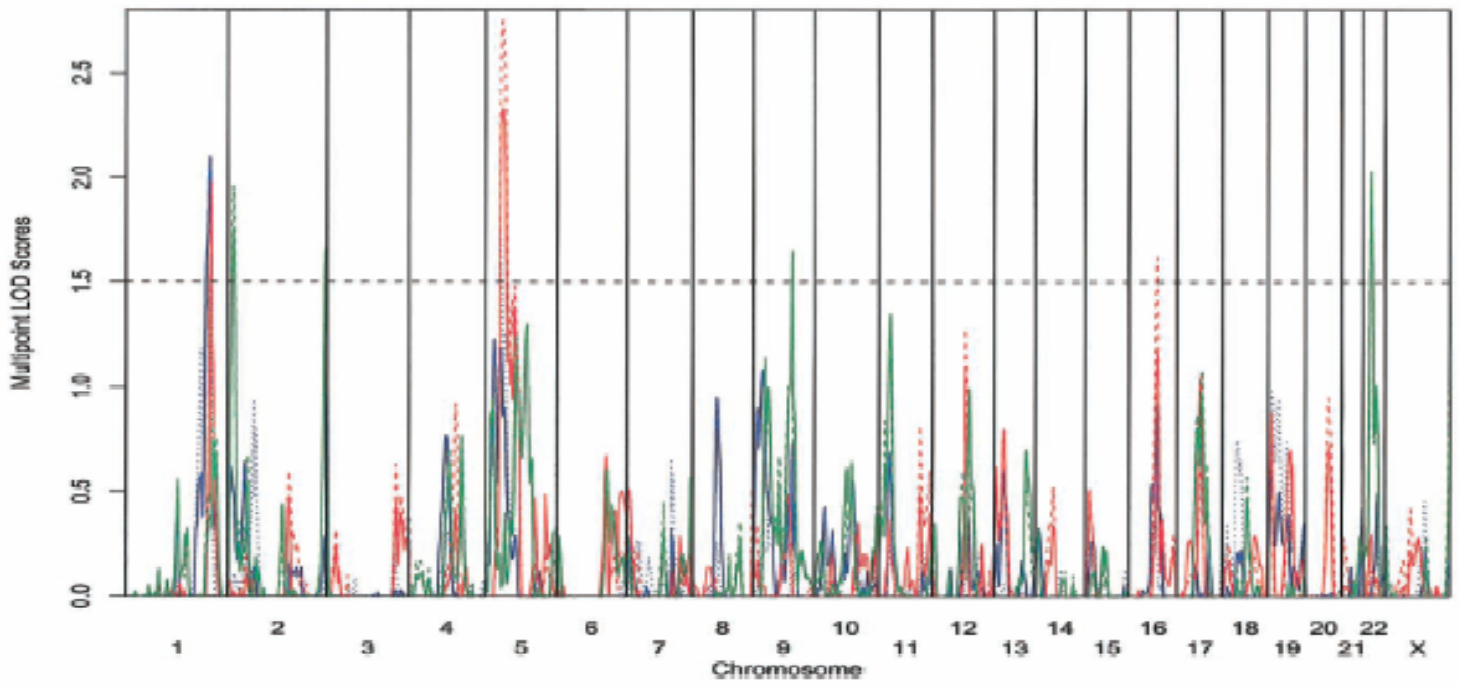


Figure 2 Multipoint nonparametric linkage analysis at equally spaced locations throughout the genome by use of the NPL_{ALL} statistics (Whittemore and Halpern 1994). LOD scores (Kong and Cox 1997) for six disease subtypes are plotted at 1-cM intervals. The solid lines denote the strict trait definition, and the dotted lines denote the broad trait definition. The blue line indicates AMD, the red line indicates GA, and the green line indicates CNV.

Scan for Age-Related Macular Degeneration Subtypes

Parametric Linkage Analysis

- *D* disease status information (affected/normal)
- *I* inheritance vector (meiosis outcomes)

- Calculate *P(D|I)* based on…

- Trait locus allele frequencies
  - *p* and *q*
- Penetrances for each genotype
  - f_{11}, f_{12}, f_{22}

Parametric Linkage Analysis

$$P(D|I)=\sum_{a_1}\cdots\sum_{a_{2f}}\prod_{i}P(a_i)\prod_{j}P(D_j|\mathbf{a},I)$$

- Sum over possible allele states for each founder
- Once *P(D|I)* is available, we can either:
  - "Plug" disease status into calculation with other markers
  - Calculate *P(I|D)* and use it to replace *P(I)* in the likelihood

Likelihood Ratio Test

- Evaluate evidence for linkage as…

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

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

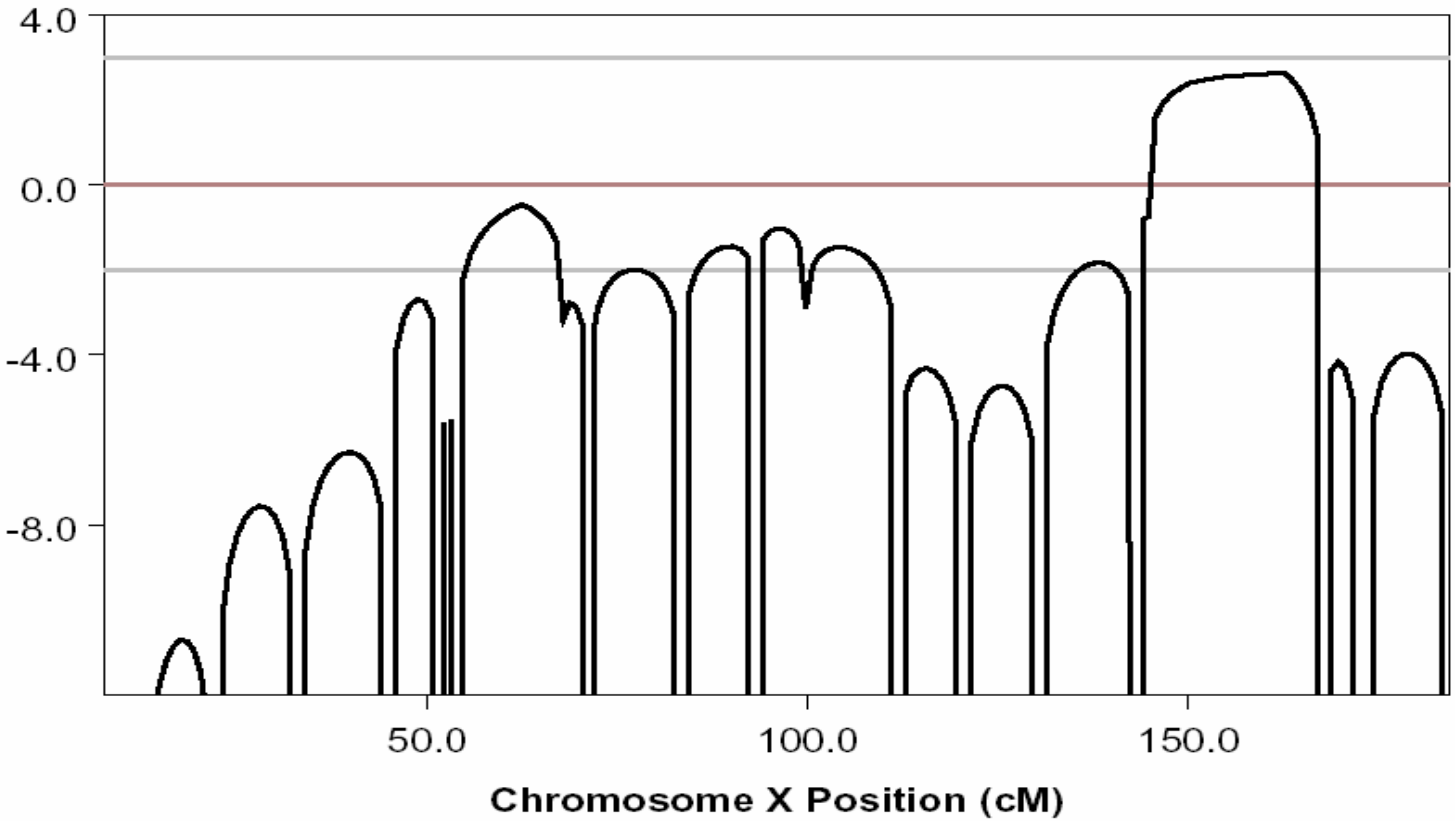
Allowing for uncertainty…

- Weighted sum over possible meiotic outcomes…

$$LR = \sum_{i \in I^*} LR(i) P(i|G)$$

$$= \frac{\sum_{i \in I^*} P(D|i) P(i|G)}{\sum_{i \in I^*} P(D|i) P_{uniform}(i)}$$

Multipoint LOD Score Plot (For An X-Linked Blindness)



Information Content

- When evaluating a scan it is useful to assess the informativeness of the available genotype data…
- … was the available genotype data sufficient to elucidate inheritance patterns along the genome?

Genotype Data Informativeness

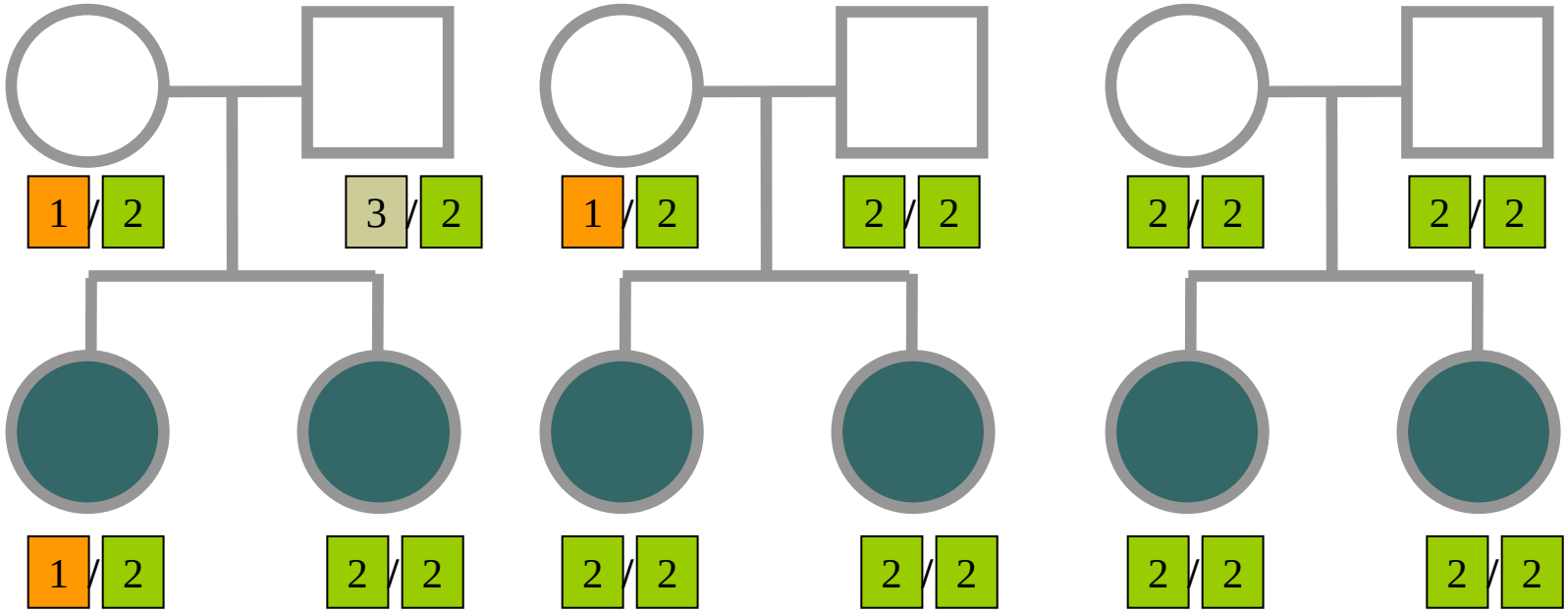
- Based on the Shannon entropy measure:

$$E = -\sum P(I = i \mid X) \log_2 P(I = i \mid X)$$

$$I = 1 - \frac{E}{E_0}$$

- Ranges between 0 and 1.
- Randomness in distribution of conditional probabilities.

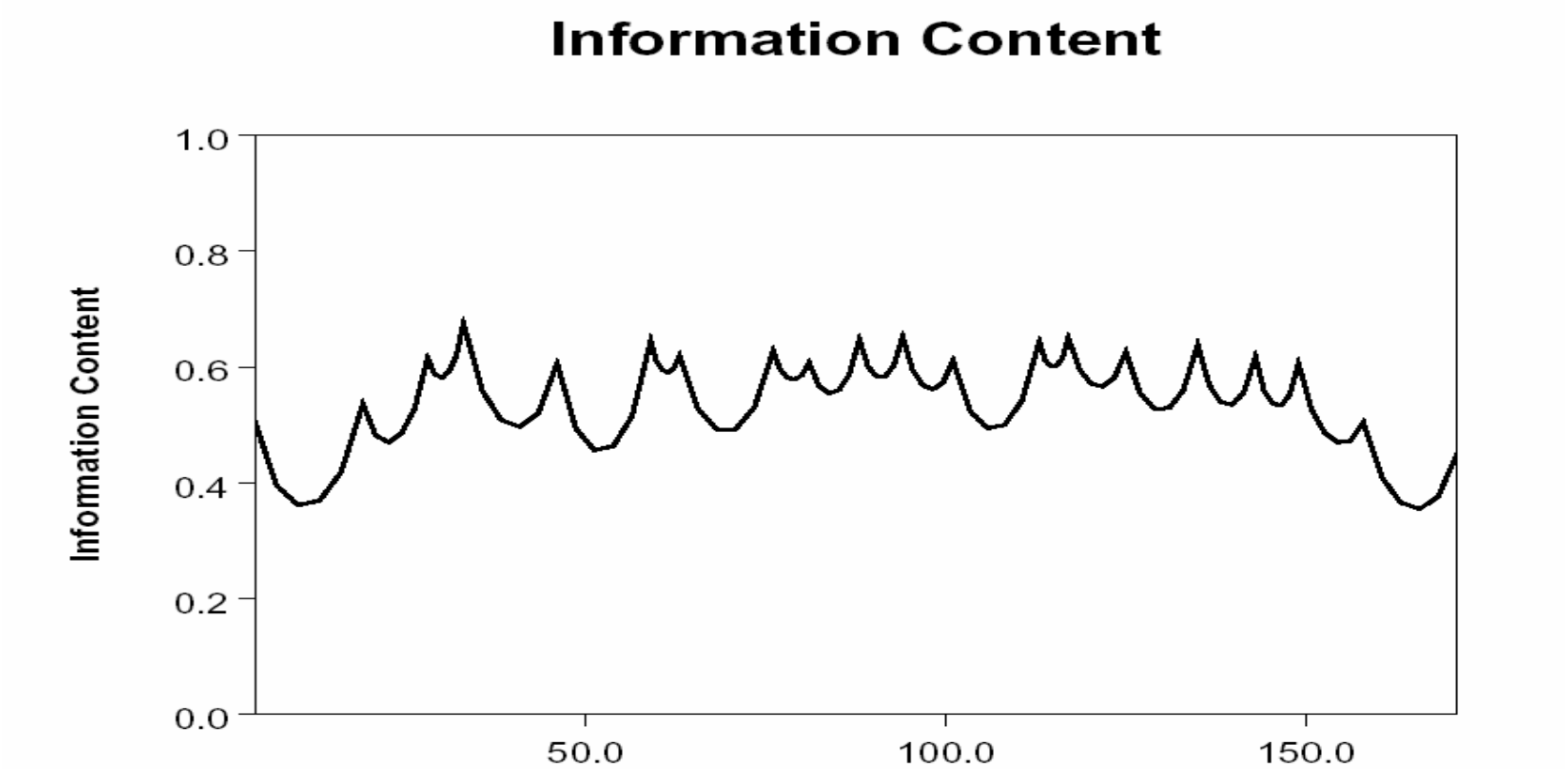
Some Exemplar Entropies



Information = 1 **Information = 0.5** **Information = 0**

(with 4 inheritance vectors)

Example of Multipoint Information Content



More on Information Content...

- The theoretical maximum is 1.0
  - All probability concentrated on one inheritance vector
- The practical maximum is lower
  - It will depend on which individuals are genotyped
- Useful in a comparative manner
  - Identifies regions where study conclusions are less certain

References on Lander-Green Based Linkage Analysis

- Kruglyak, Daly, Reeve-Daly, Lander (1996) *Am J Hum Genet* **58:**1347-63

- Whittemore and Halpern (1994) *Biometrics* **50:**109-117

Part II:
Revving Up the Markov Chain

- The ingredients we described last week are all we need to implement the Lander-Green algorithm …

- … however, a naïve implementation would be quite slow and most implementations use various refinements.

Markov Chain Calculations

$$P(X_1,...,X_m \mid I_m) = \sum_{I_{m-1}} P(X_1,...,X_{m-1} \mid I_{m-1}) P(I_m \mid I_{m-1}) P(X_m \mid I_m)$$

- $P(X_1,\ldots,X_m|I_m)$ for each I_m define a vector
- $P(I_m|I_{m-1})$ for each pair I_{m-1} , I_m defines a matrix
- $P(X_m|I_m)$ for each I_m define another vector

As Matrix Operations ...

$$\left[P(X_{1..m-1}|I_{m-1}=00),...,P(X_{1..m-1}|I_{m-1}=11)\right]\begin{bmatrix} P(I_m=00|I_{m-1}=00) & \dots & P(I_m=11|I_{m-1}=00) \\ \dots & \dots & \dots \\ P(I_m=00|I_{m-1}=11) & \dots & P(I_m=11|I_{m-1}=11) \end{bmatrix} \circ \begin{bmatrix} P(X_m|I_m=00) \\ \dots \\ P(X_m|I_m=11) \end{bmatrix}$$

- Given all the ingredients are available, how complex is this operation?

First Refinement ...

Speeding up the transitions in the Markov Chain

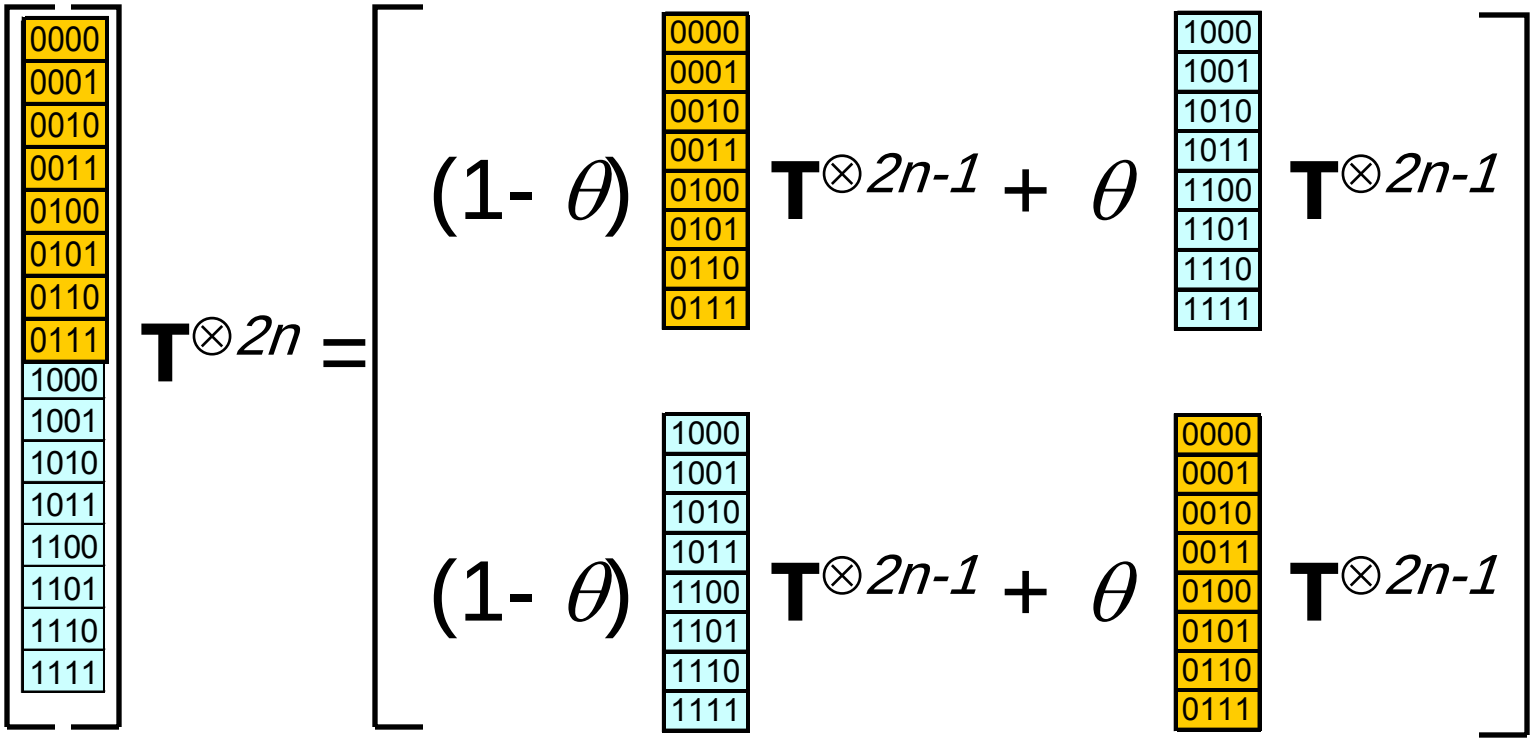
Divide and Conquer Algorithm
(Idury and Elston 1997)

Fast Fourier Transforms
(Kruglyak and Lander 1998)

Matrix Multiplication Bottleneck

- At each location we track 2^{2n} IBD patterns
- To move along genome we consider
  - $2^{2n} * 2^{2n}$ transition probabilities
- How much computation is required in a nuclear family with 5 offspring?

Elston-Idury Algorithm



Replace one matrix multiplication with 4 smaller ones …

Operations required ...

- Multiplication by full transition matrix:

- Multiplication by smaller transition matrix
  - Per matrix
  - Two of these operations needed
  - Multiplication by (1-θ) and θ

Elston-Idury Algorithm

- Matrix multiplication is an expensive operation
- Replaces multiplication by a matrix with $2^{2n}*2^{2n}$ elements with …
- .. multiplication by 2 matrices each with $2^{2n-1}*2^{2n-1}$ elements and 3 * 2^{2n} additions and multiplications
- Can be applied recursively!
  - Overall cost becomes 3 * 2n * 2^{2n} instead of 2^{2n} * 2^{2n}

Second Refinement ...

Reducing number of inheritance vectors

IBD Space

- Default recipe is inefficient
- Check resulting number of IBD states for:
  - Sibling pair
  - Half-sibling pair
  - Uncle nephew pair
- Many ad-hoc solutions …
- … but a general strategy for reducing IBD space?

Improvements: Reducing the inheritance space

- Kruglyak et al (1996)
  - Founder symmetry
- Gudbjartsson et al (2000)
  - Founder couple symmetry
- Abecasis et al (2001)
  - Arbitrary symmetries depending on genotypes
- Approaches to avoid consideration of inheritance vectors that always produce equivalent founder allele graphs

Founder Symmetry

- Allele ordering for founders is unknowable
  - Grand-paternal allele?
  - Grand-maternal allele?
- Arbitrarily fix outcome of meiosis for one offspring
- Inheritance space becomes 2^{2n-f}

Founder Couple Symmetry

- Maternal / paternal origin for ungenotyped couples is unknowable
  - Except if male-female recombination rates differ
- Arbitrarily fix outcome of meiosis for one grandchild
- Inheritance space becomes 2^{2n-f-c}

Example Application of Inheritance Vector Symmetries

- Assume that allele frequency $p_1 = 0.1$
- Consider a first cousin pair sharing genotype 1/1
- Try the following:
  - Enumerate reduced set of inheritance vectors
  - Calculate probability for each one
  - Calculate probability that the pair is IBD=0
  - Calculate probability that the parents are IBD=1

Reduced Inheritance Spaces

- Greatly speed up calculations
- Each state examined considered now represents collection inheritance vectors
  - Vectors in the collection are indistinguishable
- Requires changes to transition matrices

Next Week

- Methods for the analysis of large pedigrees