Statistical Methods for Quantitative Trait Loci (QTL) Mapping

Lectures 4 – Oct 10, 2011
CSE 527 Computational Biology, Fall 2011
Instructor: Su-In Lee
TA: Christopher Miles
Monday & Wednesday  12:00-1:20
Johnson Hall (JHN) 022

Outline

- Learning from data
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori (MAP)
  - Expectation-maximization (EM) algorithm

- Basic concepts
  - Allele, allele frequencies, genotype frequencies
  - Hardy-Weinberg equilibrium

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance (marker regression)
  - Interval mapping (EM)
Continuous Space Revisited...

- Assuming sample \( x_1, x_2, \ldots, x_n \) is from a mixture of parametric distributions,

\[
x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_n
\]

This?

Or this?

A Real Example

- CpG content of human gene promoters

"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417
Mixture of Gaussians

**Parameters $\theta$**

- means
  - $\mu_1$
  - $\mu_2$
- variances
  - $\sigma_1^2$
  - $\sigma_2^2$
- mixing parameters
  - $\tau_1$
  - $\tau_2 = 1 - \tau_1$

P.D.F

$$f(x|\mu_1, \sigma_1^2) f(x|\mu_2, \sigma_2^2)$$

$$p(x) = \tau_1 f(x|\mu_1, \sigma_1^2) + \tau_2 f(x|\mu_2, \sigma_2^2).$$

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2 : x_1, ..., x_n) = p(x_1 : \theta, \mu_1, \sigma_1^2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

---

A What-If Puzzle

**Likelihood**

$$L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2 : x_1, ..., x_n)$$

$$= \prod_{i=1}^{n} \left\{ \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2) \right\}$$

- No closed form solution known for finding $\theta$ maximizing $L$.

- However, what if we knew the hidden data?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$
EM as Chicken vs Egg

- If $z_{ij}$ known, could estimate parameters $\theta$
  - e.g., only points in cluster 2 influence $\mu_2$, $\sigma_2$.

- If parameters $\theta$ known, could estimate $z_{ij}$
  - e.g., if $|x_i - \mu_1|/\sigma_1 << |x_i - \mu_2|/\sigma_2$, then $z_{i1} >> z_{i2}$

- BUT we know neither; (optimistically) iterate:
  - E-step: calculate expected $z_{ij}$, given parameters
  - M-step: do “MLE” for parameters ($\mu, \sigma$), given $E(z_{ij})$
  - Overall, a clever “hill-climbing” strategy

Simple Version: “Classification EM”

- If $z_{ij} < 0.5$, pretend it’s 0; $z_{ij} > 0.5$, pretend it’s 1
  i.e., classify points as component 0 or 1

- Now recalculate $\theta$, assuming that partition

- Then recalculate $z_{ij}$, assuming that $\theta$

- Then recalculate $\theta$, assuming new $z_{ij}$, etc., etc.
EM summary

- Fundamentally an MLE problem

- EM steps
  - E-step: calculate expected z_{ij}, given parameters
  - M-step: do “MLE” for parameters (μ, σ), given E(z_{ij})

- EM is guaranteed to increase likelihood with every E-M iteration, hence will converge.

- But may converge to local, not global, max.

- Nevertheless, widely used, often effective

Outline

- Basic concepts
  - Allele, allele frequencies, genotype frequencies
  - Hardy-Weinberg equilibrium

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance (marker regression)
  - Interval mapping (Expectation Maximization)
Alleles

- Alternative forms of a particular sequence

- Each allele has a frequency, which is the proportion of chromosomes of that type in the population

\[
\begin{align*}
\text{C, G and \text{--} are alleles} \\
\ldots\text{ACTCGTTGGCCTTAATTCGCCGGGACTCGTTGGCCTTAATTCGCCG} \ldots \\
\ldots\text{ACTCGTTGGCCTTAATTCGCCGGGACTCGTTGGCCTTAATTCGCCG} \ldots \\
\ldots\text{ACCCCGTAAGGCCTTAATTCGCCGGGACCCTAGGCCTTAATTCGCCG} \ldots \\
\ldots\text{ACCCGGTAAGGCCTTAATTCGCCGGGACCCTAGGCCTTAATTCGCCG} \ldots \\
\ldots\text{ACCCGGTAAGGCCTTAATTCGCCGGGACCCTAGGCCTTAATTCGCCG} \ldots
\end{align*}
\]

- Single nucleotide polymorphism (SNP)

- Allele frequencies for C, G, --

Allele frequency notations

- For two alleles
  - Usually labeled \( p \) and \( q = 1 - p \)
  - e.g. \( p \) = frequency of C, \( q \) = frequency of G

- For more than 2 alleles
  - Usually labeled \( p_A, p_B, p_C \ldots \)
  - ... subscripts A, B and C indicate allele names
    \[ p_C \quad p_A \quad p_\text{--} \]
Genotype

- The pair of alleles carried by an individual
  - If there are \( n \) alternative alleles ...
  - ... there will be \( \frac{n(n+1)}{2} \) possible genotypes
  - In most cases, there are 3 possible genotypes

- **Homozygotes**
  - The two alleles are in the same state
  - (e.g. CC, GG, AA)

- **Heterozygotes**
  - The two alleles are different
  - (e.g. CG, AC)

Genotype frequencies

- Since alleles occur in pairs, these are a useful descriptor of genetic data.

- However, in any non-trivial study we might have a lot of frequencies to estimate.

\[ p_{AA}, p_{AB}, p_{AC}, ..., p_{BB}, p_{BC}, ..., p_{CC} \]
The simple part

- Genotype frequencies lead to allele frequencies.

- For example, for two alleles:
  - \( p_A = p_{AA} + \frac{1}{2} p_{AB} \)
  - \( p_B = p_{BB} + \frac{1}{2} p_{AB} \)

- However, the reverse is also possible!

Hardy-Weinberg Equilibrium

- Relationship described in 1908
  - Hardy, British mathematician
  - Weinberg, German physician

- Shows \( n \) allele frequencies determine \( \frac{n(n+1)}{2} \) genotype frequencies
  - Large populations

- Random union of the two gametes produced by two individuals
Random Mating: Mating Type Frequencies

- Denoting the genotype frequency of $AA$ by $p_{ijr}$

<table>
<thead>
<tr>
<th>Mating Frequency</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1A_1$</td>
<td>$p_{11}^2$</td>
</tr>
<tr>
<td>$A_1A_1A_2$</td>
<td>$2p_{11}p_{12}$</td>
</tr>
<tr>
<td>$A_1A_2A_2$</td>
<td>$2p_{11}p_{22}$</td>
</tr>
<tr>
<td>$A_1A_2A_2$</td>
<td>$p_{12}^2$</td>
</tr>
<tr>
<td>$A_2A_2A_2$</td>
<td>$2p_{12}p_{22}$</td>
</tr>
<tr>
<td>Total</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Mendelian Segregation: Offspring Genotype Frequencies

<table>
<thead>
<tr>
<th>Mating Frequency</th>
<th>Offspring</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_1A_1$</td>
<td>$A_1A_1$ $A_1A_2$ $A_2A_2$</td>
</tr>
<tr>
<td>$p_{11}^2$</td>
<td>1 0 0</td>
</tr>
<tr>
<td>$2p_{11}p_{12}$</td>
<td>0.5 0.5 0</td>
</tr>
<tr>
<td>$2p_{11}p_{22}$</td>
<td>0 1 0</td>
</tr>
<tr>
<td>$p_{12}^2$</td>
<td>0.25 0.5 0.25</td>
</tr>
<tr>
<td>$2p_{12}p_{22}$</td>
<td>0 0.5 0.5</td>
</tr>
<tr>
<td>$p_{22}^2$</td>
<td>0 0 1</td>
</tr>
</tbody>
</table>
Required Assumptions

- Diploid (2 sets of DNA sequences), sexual organism
- Autosomal locus
- Large population
- Random mating
- Equal genotype frequencies among sexes
- Absence of natural selection

Conclusion: Hardy-Weinberg Equilibrium

- Allele frequencies and genotype ratios in a randomly-breeding population remain constant from generation to generation.

- Genotype frequencies are function of allele frequencies.
  - Equilibrium reached in one generation
  - Independent of initial genotype frequencies
  - Random mating, etc. required

- Conform to binomial expansion.
  - \((p_1 + p_2)^2 = p_1^2 + 2p_1p_2 + p_2^2\)
Outline

- Basic concepts
  - Allele, allele frequencies, genotype frequencies
  - Hardy-Weinberg Equilibrium

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance (marker regression)
  - Interval mapping

Quantitative Trait Locus (QTL)

- Definition of QTLs
  - The genomic regions that contribute to variation in a quantitative phenotype (e.g. blood pressure)

- Mapping QTLs
  - Finding QTLs from data

- Experimental animals
  - Backcross experiment (only 2 genotypes for all genes)
  - F2 intercross experiment
Backcross experiment

- Inbred strains
  - Homozygous genomes
- Advantage
  - Only two genotypes
- Disadvantage
  - Relatively less genetic diversity

F2 intercross experiment

Karl Broman, Review of statistical methods for QTL mapping in experimental crosses
Trait distributions: a classical view

QTL mapping

- **Data**
  - Phenotypes: \( y_i \) = trait value for mouse \( i \)
  - Genotypes: \( x_{ik} = 1/0 \) (i.e. AB/AA) of mouse \( i \) at marker \( k \) (backcross)
  - Genetic map: Locations of genetic markers

- **Goals**
  - Identify the genomic regions (QTLs) contributing to variation in the phenotype.
  - Identify at least one QTL.
  - Form confidence interval for QTL location.
  - Estimate QTL effects.
The simplest method: ANOVA

- "Analysis of variance": assumes the presence of single QTL
- For each marker: Split mice into groups according to their genotypes at each marker.
- Do a t-test/F-statistic
- Repeat for each typed marker
- t-test/F-statistic will tell us whether there is sufficient evidence to believe that measurements from one condition (i.e. genotype) is significantly different from another.
- LOD score ("Logarithm of the odds favoring linkage")
  \[ \log_{10} \text{likelihood ratio}, \text{ comparing single-QTL model to the "no QTL anywhere" model.} \]

Advantages
- Simple.
- Easily incorporate covariates (e.g. environmental factors, sex, etc).
- Easily extended to more complex models.

Disadvantages
- Must exclude individuals with missing genotype data.
- Imperfect information about QTL location.
- Suffers in low density scans.
- Only considers one QTL at a time (assumes the presence of a single QTL).
Interval mapping [Lander and Botstein, 1989]

- Consider any one position in the genome as the location for a putative QTL.
- For a particular mouse, let $z = 1/0$ if (unobserved) genotype at QTL is AB/AA.
- Calculate $P(z = 1 \mid$ marker data).
  - Need only consider nearby genotyped markers.
  - May allow for the presence of genotypic errors.
- Given genotype at the QTL, phenotype is distributed as $N(\mu + \Delta z, \sigma^2)$.
- Given marker data, phenotype follows a mixture of normal distributions.

**IM: the mixture model**

- Let’s say that the mice with QTL genotype AA have average phenotype $\mu_A$ while the mice with QTL genotype AB have average phenotype $\mu_B$.
- The QTL has effect $\Delta = \mu_B - \mu_A$.
- What are unknowns?
  - $\mu_A$ and $\mu_B$
  - Genotype of QTL
References

- Prof Goncalo Abecasis (Univ of Michigan)'s lecture note
- Broman, K.W., Review of statistical methods for QTL mapping in experimental crosses