A Statistical Model of Mapping QTL (Quantitative Trait Loci) for Shape Variability

Guifang Fu

Department of Statistics, Pennsylvania State University

Advisor: Professor Rongling Wu
Outline

- Introduction and Background
- Image Analysis
- QTL Mapping Structure and Statistical Models
- Numerical Results
Outline – Introduction and Background

- Definition and Meaning of Shape
- Our Motivation
- Literature Review and Our Contribution
Outline – Introduction and Background

- Definition and Meaning of Shape
- Our Motivation
- Literature Review and Our Contribution
Outline – Introduction and Background

- Definition and Meaning of Shape
- Our Motivation
- Literature Review and Our Contribution
**Definition of Shape**

**Definition**

*Shape* is all the geometrical information that remains when location, scale and rotation effects are filtered out from an object. (Stegmann, 2002).


**Caption:** Only the boundary of an object determines the shape of this object.
All living creature including humans, animals, plants, flowers, fruits, and even bacteria and tumor and so on, all have their own diversity of shapes. Shape can affect many important biological process, such as fruit yield, metabolism, growth, heartbeat, lifespan, and disease.
Our Motivation

- **The information we have:** the measurements of 106 leaf photos, and marker data.
- **The goal we want:** To detect the genes that control the shape variation of these leaves.
- **The method we want:** QTL Mapping and Image Analysis.
Our Motivation

- **The information we have:** the measurements of 106 leaf photos, and marker data.
- **The goal we want:** To detect the genes that control the shape variation of these leaves.
- **The method we want:** QTL Mapping and Image Analysis.
Our Motivation

- **The information we have:** the measurements of 106 leaf photos, and marker data.
- **The goal we want:** To detect the genes that control the shape variation of these leaves.
- **The method we want:** QTL Mapping and Image Analysis.
Crop scientists have cloned the gene that responsible for controlling the shape of tomatoes. The gene, named **SUN**, is found to play a significant role in the elongation of tomato shape.

**Caption:** Results got by molecular biology experiment from *Esther van der Knaap, Ohio State University*. 
Literature Review: Progress of Quantitative Genetics on Shape

As early as the beginning of this century, the quantitative genetic map on shape started from tomato. However, poor progress is made during the past 100 years, since they all represented shape by simple scale such as length or width.


Currence found that locus $o$ in chromosome 2 controls the relative length of tomato fruit by using length/diameter as phenotype.


Young et al. detected gene $f$ on chromosome 11 controls tomato shape by using locule number to represent the tomato fruit.


Grandillo et al. reported $f$s$8.1$ by using the ratio of longitudinal diameter and equatorial diameter as phenotype.
Literature Review: Progress of Quantitative Genetics on Shape

Fulton et al. detected 16 QTLs by denoting round tomato shape as 1, and elongated tomato shape as 2.


Drawbacks of all above work:

- It is kind of misleading because two totally different shape can have exactly the same length and diameter (or width);
- Only taking length and width will lose the majority of information;
- In reality, most problems are not as easy as "round" or "elongated". It is normally hard to judge the variability by eyes.
Literature Review: Progress of Quantitative Genetics on Shape

Jiang et al. recording 14 measures (lobe numbers, main-lobe length and width, second-lobe length and width et. al) to describe a leaf shape.


**Caption:** This method is much better than simple scale. But it is still not as good as image analysis.
Literature Review: Progress of Quantitative Genetics on Shape

Langlade, N.B. et al. use the **19 fixed points** to represent every shape.

Drawbacks

- Langlade et al. roughly connected 19 fixed points for all different leaves to represent their boundaries. It is not a powerful way, especially for those leaves with complicated and non-smooth outlines. But we represent a shape very accurately, including tiny corners.
Outline – Image Analysis

- Four Steps to turn images into numbers
  - Data Structure
  - Dimension Reduction
Outline – Image Analysis

- Four Steps to turn images into numbers
- Data Structure
- Dimension Reduction
Outline – Image Analysis

- Four Steps to turn images into numbers
- Data Structure
- Dimension Reduction
Outline – Image Analysis

- Four Steps to turn images into numbers
- Data Structure
- Dimension Reduction
After image analysis, a leaf photo will be transformed into the data that we can manipulate with.
Each shape is represented by a $360 \times 1$ vector. Actually the dimension is decreased from $900 \times 600 \times 3$ to $360 \times 1$. 
Definition

PCA (Principal Components Analysis) is a dimension reduction method that projects the high dimensional data to the subspace that best account for the variance of the original pattern. That is to say, PCA summarizes original variables (might correlated) to a smaller set of principle variables (uncorrelated) such that PC 1 has the highest variance, PC 2 has the next highest variance, ......, and these PCs are orthogonal to each other.

From http://www.plantbiology.siu.edu/PLB444/PCA.ppt.
Outline – QTL Mapping Structure and Statistical Models

- Genetic Design
- Statistical Models
Outline – QTL Mapping Structure and Statistical Models

• Genetic Design
• Statistical Models
Genetic Design

One Gene: A/a

Two Genes: A/a and B/b

Genetic Design: LD (Linkage Disequilibrium)

- Denote the true gene by A/a. Denote the marker by B/b. Marker can be observed but the true gene is unknown.

- Assume that the frequency of $A$ is $q$, then the frequency of $a$ is $1 - q$. It forms three genotypes $AA, Aa, aa$. Notation: $j = 0(aa), 1(Aa), 2(AA)$

- Assume that the frequency of $B$ is $p$, then the frequency of $b$ is $1 - p$. It forms three genotypes $BB, Bb, bb$;

- Assume that the gene and marker are correlated with linkage disequilibrium $D$, and they form four genotypes $AB, aB, Ab, bb$, with frequencies respectively

$$
\begin{align*}
    p_{11} &= pq + D, \\
    p_{10} &= p(1-q) - D, \\
    p_{01} &= (1-p)q - D, \\
    p_{00} &= (1-p)(1-q) + D.
\end{align*}
$$
Genetic Design: How to compute $\pi_{j|i}$?

Let $\pi_{j|i}$ denotes the conditional probabilities of the gene has genotype $j$ when a marker genotype of the $ith$ individual is given.

### Table: Joint genotype frequencies at the marker and QTL

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>$p_{11}^2$</td>
<td>$2p_{11}p_{10}$</td>
<td>$p_{10}^2$</td>
</tr>
<tr>
<td>Bb</td>
<td>$2p_{11}p_{01}$</td>
<td>$2p_{11}p_{00} + 2p_{10}p_{01}$</td>
<td>$2p_{10}p_{00}$</td>
</tr>
<tr>
<td>bb</td>
<td>$p_{01}^2$</td>
<td>$2p_{01}p_{00}$</td>
<td>$p_{00}^2$</td>
</tr>
</tbody>
</table>

Note: This joint frequency table need to be derived to conditional frequency. $\pi_{j|i}$ is a function of $p$, $q$, $D$. 
Mixture Normal Model

Subscripts

- \( i = 1, \ldots, n \) — individuals
- \( j = 0(\text{aa}), 1(Aa), 2(AA) \) — three genotype
- \( Y_i \) — The phenotypic value of i-th individual

Model

\[
Y_i \sim f(Y_i|\Omega) = \sum_{j=0}^{2} \pi_j f_j(Y_i|\mu_j, \sigma^2),
\]

\[
f_j(Y_i|\mu_j, \sigma^2) \sim N(\mu_j, \sigma^2).
\]

Parameters

- \( u_j, j = 0, 1, 2 \) are the expectation of each genotype.
- \( \sigma^2 \) is the variance assumed to be common to all different genotype groups.
- \( \pi = (\pi_0|i, \pi_1|i, \pi_2|i)' \) are the mixture proportions.
The likelihood of \( n \) progeny can be represented by

\[
L(\Omega) = \prod_{i=1}^{n} \left[ \sum_{j=0}^{2} \pi_{j|i} f_{j}(Y_{i}|\mu_{j}, \sigma^{2}) \right].
\]

**What is unknown:** \( p, q, D, \mu_{0}, \mu_{1}, \mu_{2}, \sigma^{2} \), true gene genotypes \( AA, Aa, aa \).

**What is known:**
Phenotypic data \( Y \), marker genotypes \( BB, Bb, bb \)
**EM Algorithm**

**E step:**

\[
\Pi_{ij} = \frac{\pi_j(i) f_j(Y_i)}{\sum_{j'=1}^{3} \pi_{j'}(i) f_{j'}(Y_i)}.
\]

**M step:**

\[
\mu_j = \frac{\sum_{i=1}^{N} (\Pi_{ij} * Y_i)}{\sum_{i=1}^{N} \Pi_{ij}},
\]

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left[ \Pi_{i1}(Y_i - \mu_1)^2 + \Pi_{i2}(Y_i - \mu_2)^2 + \Pi_{i3}(Y_i - \mu_3)^2 \right],
\]

\[
\hat{p}_{11} = \frac{1}{2N} \left[ \sum_{i=1}^{N_1} (2\Pi_{i1} + \Pi_{i2}) + \sum_{i=1}^{N_2} (\Pi_{i1} + \theta \Pi_{i2}) \right],
\]
**EM Algorithm**

\[
\hat{p}_{10} = \frac{1}{2N} \left[ \sum_{i=1}^{N_1} (\Pi_{i2} + 2\Pi_{i3}) + \sum_{i=1}^{N_2} (\Pi_{i3} + (1 - \theta)\Pi_{i2}) \right],
\]

\[
\hat{p}_{01} = \frac{1}{2N} \left[ \sum_{i=1}^{N_3} (2\Pi_{i1} + \Pi_{i2}) + \sum_{i=1}^{N_2} (\Pi_{i1} + (1 - \theta)\Pi_{i2}) \right],
\]

\[
\hat{p}_{00} = \frac{1}{2N} \left[ \sum_{i=1}^{N_3} (\Pi_{i2} + 2\Pi_{i1}) + \sum_{i=1}^{N_2} (\Pi_{i3} + \theta\Pi_{i2}) \right],
\]

where \( \theta = p_{11}p_{00} / (p_{11}p_{00} + p_{10}p_{01}) \).

Keep iterating until all estimates converge.
Outline – Numerical Results

- Hypothesis Test
- Data Fitting Graphs
Outline – Numerical Results

- Hypothesis Test
- Data Fitting Graphs
Hypothesis Tests

Testing Gene Existence for Controlling Poplar Leaf Shape Variation

\[ H_0 : \mu_0 = \mu_1 = \mu_2 \text{ vs. } H_1 : \text{Not } H_0 \]

Reject if

\[ LOD = -2[\log L(\hat{\Omega}) - \log L(\tilde{\Omega})] > \chi^2_{df=2} \left( \frac{0.05}{16} \right) = 11.54. \]

Testing Linkage Disequilibrium between the Gene and Marker

\[ H_0 : D = 0 \text{ vs. } H_1 : \text{Not } H_0 \]

Reject if

\[ \chi^*_2 = \frac{2nD^2}{(p(1-p)q(1-q))} > \chi^2_{df=1} \left( \frac{0.05}{16} \right) = 8.73. \]
Significant Result for Global Case

Significant Markers

- Marker 2, with LOD 46.04, and $\chi^2 = 60.96$
- Marker 5, 8, 12, 13
- Marker 9, 16
- Marker 1, 6
You can notice the bending, elongating, and widening trend near the middle and bottom of the leaf blade.
Thank you!