Advanced Statistical Methods for the Analysis of Gene Expression and Proteomics

Lecture 10 – Bayesian multiple comparison; ArrayCGH analysis

Yuan Ji

yuanji@mdanderson.org

Department of Bioinformatics and Computational Biology
University of Texas M.D. Anderson Cancer Center
Bayesian Multiple Testing Based on Test Statistics

Yuan Ji

January 28, 2008
1. The problem of multiple testing

2. The proposed approach
   1. A hierarchical modeling approach for multiple testing
   2. An illustrative example – $F$-tests
   3. A model assessment tool

3. Application

4. A brief discussion
A Bayesian framework

- Suppose a sequence of $m$ null hypotheses $H_{0i}$ is tested against a corresponding sequence of alternative $H_{1i}$ for $i = 1, \ldots, m$.

- A Bayesian procedure for this problem:
  - Construct a latent indicator $J_i = 0$ if $H_{0i}$ is true and $J_i = 1$ if $H_{1i}$ is true.
  - Compute the marginal posterior probability $\Pr(J_i = 1 | \text{data})$ based on some appropriate models.
  - Adjust for multiplicity using the marginal posterior probabilities.
For test $i$, observed data $y_i$. A Bayesian hierarchical model consists of

- Probability distribution $p(y_i | J_i = k) = p_k(y_i; \theta_k)$, $k = 0, 1$.
- The likelihood function:
  $$p_0(y_i; \theta_0)^{1 - J_i} p_1(y_i; \theta_1)^{J_i}.$$ 
- Priors for $\theta_k$ is $f_k(\theta_k)$; prior $Pr(J_i = 1) = \pi$.
- Hyperpriors for the parameters in the priors (e.g., $\pi$).

Compute

$$r_i = Pr(J_i = 1 | y_1, \ldots, y_m)$$

the marginal posterior probability that $H_{1i}$ is true.
- Probabilities $r_i$ adjust for multiplicities automatically as long as
  - $\Pr(J_i = 1) > 0$ for all $i = 1, \ldots, m$;
  - $\pi \sim p(\pi)$, rather than fixed.
  - Ref. Scott and Berger (2003); Müller et al. (2006)

- Optimal decision (Müller et al., 2004) is

$$I(r_i > t),$$

...to reject all the null hypotheses with $r_i > t$ for some fixed value $t$.

- Choice of $t$ depends on choice of loss functions.
Construction of appropriate Bayesian models can be difficult. (e.g., construction of priors for $\theta_k$).

Values of posterior probabilities $r_i$ are often sensitive to the prior densities.

MCMC computation can be intensive, especially for high-dimensional data (e.g., genomics/proteomics data).
Hierarchical model based on test statistics

- Johnson (2005) proposed computing posterior probabilities $r_i$ based on test statistics.
- Main idea:
  - Base the models on the sampling distributions of test statistics.
  - The null distributions are often completely specified – no need for prior specification.
  - The alternative distributions of test statistics can often be described with a parsimonious parametrization.
Therefore,

- Models under the null \( p_0(y_i) \) are free of parameters.

- Models under the alternative \( p_1(y_i; \theta_1) \) depend on few parameters (often just one).

- \( \Pr(J_i = 1|y_1, \ldots, y_m; \theta_1) \) has a closed-form solution – easy to sample.
Let $f_i$ be the test statistic (e.g., $\chi^2$, $F$, $t$ or $z$-statistic) for null $H_{0i}$ vs. $H_{1i}$;

- Likelihood $p(f_i|J_i, \tau) = p_0(f_i)^{1-J_i} p_1(f_i|\tau)^{J_i}$;
- Prior of $J_i \sim \text{Bin}(1, \pi)$;
  - Hyperprior of $\pi \sim \text{Beta}(p_0, (1 - p_0))$, where $p_0$ is fixed.
- Prior of $1/\tau \sim \text{Gamma}(1, 2)$;
MCMC algorithm for \( \{\pi, \tau, J_1, \ldots, J_m\} \)

- **Full conditional**

\[
\Pr(J_i = 1|f_1, \ldots, f_m, \tau, \pi) = \frac{p_1(f_i|\tau)\pi}{p_1(f_i|\tau)\pi + p_0(f_i)(1 - \pi)}
\]

- \( \pi|J_1, \ldots, J_m \sim \text{Beta}(p_0 + \sum J_i, (1 - p_0) + m - \sum J_i) \).

- Sample \( \tau \), e.g., using random-walk Metropolis-Hastings.
Suppose

\[ y_i|\beta_i, \sigma_i^2 \sim N_n(X_i\beta_i, \sigma_i^2 I). \]

For testing the validity of linear constraint \( H_{0i} : Q'\beta_i = \xi \), the classical \( F \) statistic \( f_i \) is the ratio of average sums of squares.

- \( p_0(f_i) \) is a central \( F \) distribution;
- Suppose alternative \( H_{1i} \) assumes that

\[ \beta_i \sim N(\beta_i^*, \tau \sigma_i^2 (X'_iX_i)^{-1}) \]

where \( \beta_i^* \) is a value satisfying \( H_{0i} \),
- then \( p_1(f_i|\tau) \sim (1 + \tau)p_0(f_i) \).
Decision rules

- Posterior probability $r_i = \Pr(J_i = 1 | f_1, \ldots, f_m)$ is computed using the MCMC sample.

- Reject $H_{0i}$ if $r_i > t$ for some value of $t$ (more discussion later)
A quantile-quantile plot is proposed to check the model fitting.

- Suppose \( \{\tau^1, \ldots, \tau^B\} \) is the MCMC sample.
- Randomly draw \( \tau^s \).
- Obtain the corresponding posterior sample \( \{J^s_1, \ldots, J^s_m\} \) from the \( s^{th} \) iteration of the MCMC.
- Assign the test statistics \( f_i \) to the null group if \( J^s_i = 0 \), and to the alternative group if \( J^s_i = 1 \).
Plot the sample quantiles of $f_i$ in the null group against the theoretical quantiles based on the distribution $p_0(f_i)$;

Plot the sample quantiles of $f_i$ in the alternative group against the theoretical quantiles based on the distribution $p_1(f_i|\tau^k)$;

Compare the curves with the 45 degree line.

This procedure only works for quantities of which the sampling distributions are free of parameters – such as the $F$-statistics (its distribution only depends on two degrees of freedom).
Consider one-sample t-tests $H_{0i} : \mu_i = 0$, $i = 1, \ldots, m$.

Observed data for test $i$ are samples $\{y_{i1}, \ldots, y_{in}\}$. The $F-$statistic $f_i$ is the square of the one-sample $t-$statistic.

- We generated $m = 1000$ tests.
- Sample sizes per test $n = 11$.
- Under $H_{0i}$, $f_i \sim F_{1,10}$ and under alternative $f_i \sim (1 + \tau)F_{1,10}$.

Simulation scheme consists of sampling $\tau$, $\pi$, $J_i|\pi$, and $f_i|J_i, \pi$ (in this order), from their true distributions under the proposed model.
Simulation 2

- Sample $y_{i1}, \ldots, y_{in} \sim iid \ N(3, 1)$ for $i = 1, \ldots, 100$;
- Sample $y_{i1}, \ldots, y_{in} \sim iid \ N(0, 1)$ for $i = 101, \ldots, 1000$;
- $H_{0i} : \mu_i = 0$
- Compute

$$ t_i = \frac{\bar{y}_i}{\hat{\sigma}/\sqrt{n}} $$

where $\bar{y}_i$ is the sample mean and $\hat{\sigma}$ is the sample standard deviation.

After applying the proposed method,
An siRNA screening experiment conducted by Gordon Mills and his lab.

- A kinase library of about 900 siRNA’s are screened for their silencing properties.
- A functional silencing siRNA significantly reduced cell viability (measured as a continuous variable).
- Using 96-well plates, the library is screened with 30 plates in triplicates.
- $F$-statistics $f_i$ are computed for all 900 siRNA’s with degrees of freedom $(1, 4)$. 
Yuan Ji
Bayesian Multiple Testing Based on Test Statistics
Results using the Bayesian procedure

We applied the proposed method for the 900 $F$-statistics $f_i$.

- Assume $\pi \sim \text{Beta}(0.5, 0.5)$.
- Assume $1/\tau \sim \text{Gamma}(1, 2)$.
- Under null, $f_i \sim F(1, 4)$.
- Under alternative, $f_i \sim (1 + \tau)F(1, 4)$. 

Yuan Ji

Bayesian Multiple Testing Based on Test Statistics
Posterior probability and FDR

Bayesian Multiple Testing Based on Test Statistics
Decision rules

The optimal decision takes the form (Müller et al., 2004)

\[ d_i = I(r_i \leq t) \]

- If the "goal" (loss function) is to minimize FNR subject to \( \text{FDR} \leq \alpha \), then \( t \) equals the largest \( r_i \) such that the corresponding posterior expected FDR (by rejecting all the \( r_j \leq r_i \)) is \( \leq \alpha \).

- In the above plot, draw a horizontal line at \( y \)-axis = 0.2. Draw a vertical line at the intersection between the horizontal line and the dotted curve. The intersection between the vertical line and the solid curve is the optimal \( t \) value in \( d_i \).
Model assessment

Alternative

Null

Sample quantiles

Model quantiles

Sample quantiles

Model quantiles
Khodarev et al. (2005) studied the association between progression of Barrett’s Metaplasia to Adenocarcinoma and gene expression levels. Three conditions are examined:

- Normal esophageal epithelium
- Premalignant Barrett’s metaplasia,
- Esophageal adenocarcinoma

For each condition, $n = 8$ Affymetrix U133A arrays were produced from 8 different patients with the same condition. After normalization using dChip (Li and Wong, 2001), we obtained $m = 16384$ genes, each with 24 measurements.
For each gene, we performed a one-way ANOVA using the three conditions as a factor. We obtained $m = 16384$ $F$–statistics with degrees of freedom $(2, 21)$. Therefore,

- $p_0(f_i)$ follows $F_{2,21}$
- $p_1(f_i|\tau)$ follows $(1 + \tau)F_{2,21}$.

We applied the proposed method and computed

$$r_i = \Pr(J_i = 1|f_1, \ldots, f_m)$$

for each gene $i$. 
Posterior probability and FDR

---

**Yuan Ji**
Bayesian Multiple Testing Based on Test Statistics
We let $\pi \sim \text{Beta}(0.5, 0.5)$. 
Conclusions

- The proposed model simplifies the process of specifying prior distributions for unknown parameters, which can be tricky.
- Only one parameter needs to be sampled using M-H; others are sampled directly from Bernoulli distributions.
- Information across all the tests is used in the decision making for each single test – through the common parameter $\tau$.
- We provide a simple model-assessment tool to check the model fitting.
- Additional research is needed to explore more general assumptions under the alternative when model does not fit.
ArrayCGH Analysis

- Introduction of the technology
- Two existing methods for analysis
- Bayesian parametric/nonparametric modeling (ongoing)
http://creativecommons.org/licenses/by-sa/2.0/

Attribution-ShareAlike 2.0

You are free:

• to copy, distribute, display, and perform the work
• to make derivative works
• to make commercial use of the work

Under the following conditions:

BY: Attribution. You must give the original author credit.

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

• For any reuse or distribution, you must make clear to others the license terms of this work.
• Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer: ©
A normal human genome has 2 copies of DNA
Copy number alterations (CNA) can lead to disease

- CNAs can lead to adverse expression changes of affected genes
- CNAs are a hallmark of tumor genomes
- CNAs are diagnostic of mental retardation
- Recurrent CNAs in individuals with common phenotype represent molecular markers of disease
- Task: find recurrent CNAs for diagnostics, gene-disease association, disease susceptibility

Bayani et al, Cancer Research 2002

Nature 437, 1084-1086
• Array hybridization - similar to cDNA array studies:
  * Test DNA sample - Unknown DNA copy number
  * Reference DNA sample - DNA copy number of 2
  * Label, mix, hybridize to BAC, cDNA, or oligonucleotide targets/probes spotted on a glass array
  * Scan

• Array analysis - resulting data are normalized log test over reference intensities for genomic targets
Measuring CNAs with array comparative genomic hybridization (aCGH)

Array comparative genomic hybridization (aCGH)

\[ y_t = \log_2 \left( \frac{\text{sample}_t}{\text{ref}} \right) \]

- \( \log_2 4/2 = 1.00 \)
- \( \log_2 3/2 = 0.58 \)
- \( \log_2 2/2 = 0.00 \)
- \( \log_2 1/2 = -1 \)

Gain
Neutral
Loss

Indexed by chromosomal location

Intensity of sample probe \( t \)
Intensity of reference
Examples of CNAs acquired in a lymphoma patient
DNA Copy Number Example Data
DNA Copy Number

Analytic Goals

Things that might be done with this data are:

1. Identification of genes and regions that often have abnormal copy number

2. Association of copy number with clinical data

3. Classification

4. Clustering of genes or samples

Our methods are focused on 1, but may help for 2 and 3
DNA Copy Number

Previous Work

• Hodgson et al. (2001) fit a three-component Gaussian mixture model

• Bourdon et al. (2002) transformed data to be Gaussian and identified outliers

• Pollack et al. (2002) smoothed data and determined cut-offs using normal data and FDR

• Olshen and Venkatraman and Fridlyand et al. explored hidden Markov models
Let $Z_1, Z_2, \ldots, Z_K$ be the data.

If $Z_1, \ldots, Z_\nu \sim F_0$ and $Z_{\nu+1}, \ldots, Z_K \sim F_1$, then $\nu$ is a change-point.

For our data, a change-point would correspond to where the DNA copy number has changed. There may be multiple changes within a chromosome.
Suppose the data are $Z_1, \ldots, Z_K$. For $k : 0 < k < K$,

$$T_k = \frac{|\bar{Z}_k - \bar{Z}_{K-k}|}{\sigma \sqrt{(K - k)^{-1} + k^{-1}}}$$

where $\bar{Z}_k = \frac{\sum_{i=1}^{k} Z_i}{k}$ and $\bar{Z}_{K-k} = \frac{\sum_{i=k+1}^{K} Z_i}{K}$.

To test for a change-point, the test statistic is

$$T^* = \max_{0 < k < K} |T_k|.$$
DNA Copy Number

A Problem
We find the maximum of

$$T_{k_1,k_2} = \frac{|\bar{Z}_{k_2-k_1} - \bar{Z}_{k_1,k_2}|}{\sigma \sqrt{(k_2 - k_1)^{-1} + (k_1 + K - k_2)^{-1}}},$$

for $1 \leq k_1 < k_2 \leq K$, where

$$\bar{Z}_{k_2-k_1} = \frac{\sum_{i=k_1+1}^{k_2} Z_i}{(k_2 - k_1)}.$$

and

$$\bar{Z}_{k_1,k_2} = \frac{(\sum_{i=1}^{k_1} Z_i + \sum_{i=k_2+1}^{K} Z_i)}{(K - k_2 + k_1)}.$$

The reference distribution is found by permutation.
• Repeatedly split the data into either 2 or 3 segments. If the split is ternary, the two end segments are grouped together when determining the maximum statistic.

• We are performing binary segmentation as if the data were in a circle. Thus we call this procedure _circular binary segmentation_ (CBS).
• If ternary split, check first and third segment by test of binary split with middle segment

• Outliers are smoothed before segmentation

• Overlapping windows are used if more than 500 – 1000 markers; our method is now a hybrid of binary segmentation and sequential methods (Page 1957)
DNA Copy Number Validation

• The data we have examined are discussed in Snijders et al. (2001).

• The arrays have 2460 BACs, mapped, spotted in triplicate.

• There are 15 fibroblast cell lines. They have been identified as containing monosomies or trisomies, on either parts or all of 1-2 chromosomes.
DNA Copy Number

An Example
DNA Copy Number

An Example

\[ \log(T/R) \]
DNA Copy Number

Another Example

![Graphs showing DNA copy number analysis](image-url)
DNA Copy Number

A Final Example
Suppose there are $N$ change-points after CBS.

1. Find the best set of change-points $n = 1, \ldots, N$, choosing only among those previously identified.

2. Choose the smallest $n$ such that $SS_n/SS_N - 1 < c$, where $c$ is some pre-specified constant (such as 0.05) ($SS_n$ equivalent to the error sum of square in one way ANOVA; If $n$ change-points, $n + 1$ groups)
DNA Copy Number A Simulation

- Start with a step function from a CBS fit to chromosome 11 from a real breast cancer study
- Add Gaussian noise with SD estimated from same data
- Apply CBS
- Repeat process 100 times
There were 6 change-points in the step function.

<table>
<thead>
<tr>
<th>Change-points</th>
<th>α level for permutation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>6</td>
<td>18 (97)</td>
</tr>
<tr>
<td>7</td>
<td>16 (2)</td>
</tr>
<tr>
<td>8</td>
<td>26 (1)</td>
</tr>
<tr>
<td>9</td>
<td>19 (0)</td>
</tr>
<tr>
<td>10-14</td>
<td>21 (0)</td>
</tr>
</tbody>
</table>
• **Specimens**: 12 breast cancer tissues (thanks to Mike Wigler, Rob Lucito, and members of their labs)

• **Array**: 9820 oligonucleotides of length 70

• **Analysis**: Chromosome 17, which contains ERBB2/HER2NEU.

Increased HER2neu protein is associated with 30% of breast cancers, and correlates with poor prognosis and aggressive tumor growth. The drug Herceptin has been shown to slow progression and increase tumor shrinkage.
DNA Copy Number

Chromosome 17

Specimen 1

Specimen 2

Specimen 3

Specimen 4

Specimen 5

Specimen 6

Specimen 7

Specimen 8

Specimen 9

Specimen 10

Specimen 11

Specimen 12
A recurrent CNA in 5 lung cancer cell lines

MYC oncogene
**Goal: Detecting recurrent CNAs from multiple aCGH samples**

**Goal:** find CNAs that are common to a set of samples to determine which CNAs are contributing to disease

- Shared CNAs define a molecular characterization, or *profile* of the particular phenotype

**Challenges:**
- Not all samples will exhibit CNAs in the same place
- Only a small proportion of locations will show recognizable patterns
- Frequency of occurrence in those locations is variable

**Contributions:**
- Studying and comparing joint models to discover recurrent CNAs
Signals missed by AF

Borrow statistical strength across samples
Models for recurrent CNAs

\[ M_t, Z_t \in \{L, N, G\} \]

\begin{align*}
\text{Alteration frequency (AF):} & \quad \text{Discretize each sample separately and summarize} \\
\text{✓ Use HMM on each sample} & \quad \text{✗ Premature thresholding} \\
\text{✗ Cannot ‘share’ information} & \\
\text{Factored likelihood HMM (FL-HMM):} & \quad \text{Use a ‘joint’ emission model to infer } M \text{ from the raw data} \\
\text{✓ No discretization} & \quad \text{✗ One sample can dominate} \\
\text{Buffered FL-HMM (BFL-HMM):} & \quad \text{Simultaneously infer } M \text{ and } Z \\
\text{✓ ‘Buffer’ the master} & \quad \text{✗ Ambiguous locations?}
\end{align*}
Considering ambiguous regions

Many clones will not have consensus
What should the master do?

Hierarchical HMM (H-HMM):
Allow $M_t$ to ‘opt-out’ of L,N,G states

- Explicitly model ambiguity
- Leads to more accurate and interpretable output
- Exert an adjustable level of control over $Z$s
The $Z^s$ chains are coupled which makes inference hard.

Conditioned on $M$, $Z^s$ are independent and can be updated in parallel.

Use Blocked Gibbs MCMC sampling
- Compute $p(M|Z, \theta, Y)$
- See paper for details

Running time is $O(TSN)$
- Number of data points and $N$ MCMC samples
- Considerably slower than other models due to $Z$ chains
Synthetic data experiment

Data based on real data from 8 MCL cell lines and insert recurrent CNAs

Alter:

- Width
- Height
- Frequency
- Spike in random CNAs
- Repeat 3 times

- ~100 data sets, ~600 probes each

- ROC analysis measuring accuracy in predicting recurrent CNAs
H-HMM is quantitatively better than other models
Qualitatively, the H-HMM is sparse, yet accurate

18 Non small cell lung cancer Adenocarcinoma cell lines

![Graph showing chromosomal activity](image-url)
Summary

- Developed and compared 3 new statistical models to detect recurrent CNAs in array CGH data
  - Infer a profile representing canonical locations without first discretizing the data

- H-HMM was quantitatively and qualitatively better than other models and the standard approach

- Evaluating results in large scale study and refining the model to detect low-frequency CNAs

Consider a chromosome with \( n + 1 \) probes, indexed by \( i = 0, 1, \ldots, n \).

Define a partition of \( \mathcal{A} \) as \( P(\mathcal{A}) \equiv \{ c_0 = 0, c_1, c_2, \ldots, c_K = n \} \), in which \( c_k \in [0, n] \) are change points that break \( \mathcal{A} \) into \( K \) segments.

Denote a segment by \( \Delta_k = (c_k, c_{k+1}] \). Then \( \mathcal{A} = \bigcup_{k=0}^{K-1} \Delta_k \).

The goal is to find an appropriate partition and estimate the copy number state, denoted as \( \mu_k \), for each segment of the chromosome defined by the partition.
Denote $Y_{ij}$ the $\log_2$ ratio of the copy numbers of probe $i$ in array $j = 1, \ldots, J$.

Suppose that probe $j$ is in segment $k$, i.e., $i \in \Delta_k$.

The model for $Y_{ij}$ is given by

$$Y_{ij} \mid P(A), \mu_k, \sigma_k^2 \sim N(\mu_k, \sigma_k^2), \quad \text{if } i \in \Delta_k.$$ 

which can be rewritten as

$$Y_{ij} \mid P(A), \mu_k, \sigma_k^2 \sim N\left(\sum_{k=0}^{K-1} \mu_k I(c_k \leq i < c_{k+1}), \sum_{k=0}^{K-1} \sigma_k^2 I(c_k \leq i < c_{k+1})\right)$$

where $I(\cdot)$ is the indicator function. The full likelihood function is given by
Like = \prod_{j=1}^{J} \prod_{i=1}^{n} \left( 2\pi \sum_{k=0}^{K-1} \sigma_k^2 I(c_k \leq i < c_{k+1}) \right)^{-1/2} \times \exp \left[ -\frac{\left( Y_{ij} - \sum_{k=0}^{K-1} \mu_k I(c_k \leq i < c_{k+1}) \right)^2}{2 \sum_{k=0}^{K-1} \sigma_k^2 I(c_k \leq i < c_{k+1})} \right]. \quad (2)
Priors

- Prior for the partition $P(A)$.
- Prior for the copy number state $\mu_k$ – parametric or nonparametric.
- Prior for other parameters.