Bayesian Methods for Microarray Data Analysis

Marco F Ramoni
Harvard Medical School

Paola Sebastiani
Boston University School of Public Health

IMA Workshop
October 2nd, 2003

Introduction

Thesis: Bayesian methods solve your problems:
1. To log or not to log?
2. Normalize/transform: when, why and how?
3. How to handle continuous phenotypes (survival).
4. How many clusters?
5. How to handle temporal experiments?
6. How to move from similarity to control?

Hypotheses:
1. Analyzing 30000 variables with less than 10 observations is not trivial.
2. Predictive accuracy and reproducibility are solid ground for statistical validation (vs anecdotal).
Bayesian Microarray Analysis

Rev. Thomas Bayes

Name: Thomas Bayes.
Death: 1761, Tunbridge, Kent.
Job title: Presbyterian minister.
Publications: 3 (1 posthumous).
Trivia: Unrelated to Bayesians.

Bayesian Microarray Analysis

Bayesian Phenotype

Bayesian: The opposite of Frequentist.
Frequentist: Every (at least most) statistician you know.
Population: 46656 varieties of Bayesians (Good, 1971).
Profile: Hates the cookbook style of classical statistics.
Tools: Total Probability Theorem and Bayes Theorem.
Main Beliefs: Inference is about decisions.
  ✓ Inference is conditional on data (small samples);
  ✓ Inference is updating prior knowledge (even when it is knowledge of not knowing anything);
  ✓ Everything (also an hypothesis) has a probability.
Bayesian Microarray Analysis

**Comparative Experiments**

Healthy cell  Tumor cell

Sample 1  Sample 2  Sample 3  Sample 4

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX -Bio</td>
<td>88</td>
<td>283</td>
<td>309</td>
<td>12</td>
</tr>
<tr>
<td>Affx -alu_</td>
<td>15091</td>
<td>11038</td>
<td>16692</td>
<td>15763</td>
</tr>
<tr>
<td>AFFX -Dap</td>
<td>311</td>
<td>134</td>
<td>378</td>
<td>268</td>
</tr>
<tr>
<td>AFFX -Lys</td>
<td>21</td>
<td>21</td>
<td>67</td>
<td>43</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>215</td>
<td>116</td>
<td>476</td>
<td>155</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>977</td>
<td>433</td>
<td>1474</td>
<td>415</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>14538</td>
<td>615</td>
<td>5669</td>
<td>4850</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>9738</td>
<td>115</td>
<td>3272</td>
<td>2293</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>8529</td>
<td>1518</td>
<td>3668</td>
<td>2569</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>15076</td>
<td>19448</td>
<td>27410</td>
<td>14920</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>11126</td>
<td>13568</td>
<td>16756</td>
<td>11439</td>
</tr>
<tr>
<td>AFFX -HU</td>
<td>17782</td>
<td>18112</td>
<td>23006</td>
<td>17633</td>
</tr>
</tbody>
</table>

Bayesian Microarray Analysis

**Bayesian Solution**

**Challenge:** Compare two conditions and identify genes changed across them using small samples.

**Quantity:** Probability that the ratio between the expression of a gene in condition A and B is > 1.

**Strategy:** Identify the distributions of the expression values, and take into account random changes (no normalization).

**Method:** A good model (distribution) capitalizes as much as possible on the data (small sample sizes).

**Free Goodies:** A parametric method builds predictive models that can be used for statistical validation.
Distributional Assumptions

Microarrays produce data in two distributions:

- **Log-normal**: take the logarithm, data are normal.
- **Gamma**: they remain asymmetrical (exponential).

![Distributional Assumptions Diagram](image)

Bayesian Microarray Analysis

Probabilistic Scoring

- Expression fold change for a gene $g$ between the two conditions A and B: $\theta_g = \frac{E(Y_A)}{E(Y_B)}$.
- Compute the probability: $p(\theta_g > 1 \mid Y_A, Y_B)$ for each gene under the gamma and LN assumption.
- Compute the probability that data for $g$ are generated by gamma $p(\Gamma_g \mid Y_A, Y_B)$ and log-normal $p(LN_g \mid Y_A, Y_B)$.
- Return the weighted average of these quantities: $p_G(\theta_g > 1 \mid \text{data})p(\Gamma_g \mid \text{data}) + p_{LN}(\theta_g > 1 \mid \text{data})p(LN_g \mid \text{data})$
  weights are the odds of gamma versus log-normal.

Bayesian Microarray Analysis
The Effect of Averaging

- Prostate cancer data set: 50 samples from normal tissues; 52 samples from cancer tissues.
- Original .cel files processed with MAS 5.0 and analyzed using model averaging, only gamma, only log-normal.

Posterior probabilities $p(\theta_g > 1 \mid \text{data})$

Prediction

Analysis: Suppose the analysis leads to select a group of genes which are differentially expressed across the two conditions.

Prediction: we may want to classify new samples on the basis of their expression profile $z$ (molecular diagnosis):

$p(\text{class}=i \mid z_1=v_1, z_2=v_2, \ldots, z_n=v_n)$.

Assumptions: gene independent given class and parameters.
Reproducibility

Problem: Simulation not an option.
Data: 50+52 sample of prostate cancer (Singh et al., 2002).
Preprocess/Transform: None.
Reproducibility: Cross sample normalized correlation: (1+cor)/2.
Resampling: 8 groups of 4 data sets each, of increasing sample sizes.
Quantities: Fold change, posterior probability, t- and s2n statistics.

Tail Reproducibility

Badge p>.99 (~1300)  
GeneCluster t>2 (~1300)

Bayesian Microarray Analysis
What Did We Learn?

- Differential experiments usually end up with:
  - A list of genes changed across the two conditions;
  - A “stochastic profile” of each condition.
- These methods are useful to identify diagnostic profiles and prognostic models.
- They are not designed to tell us something about regulatory mechanisms, structures of cellular control.
- The reason: we use the experimental condition as a training signal: i.e. we look only at relations between gene expression and experimental condition.

Sorting and Clustering

Unsupervised: Expression profiles, no training signal.

Method: Sort the expression profiles in a hierarchical using a pair-wise similarity measure (say, correlation) between all the profiles.

Model: Build a single tree merging all sequences. Use the mean of each set of merged sequences as representation of the joint to traverse the tree and proceed until all series are merged.

Clustering: Pick a threshold to break down the single tree into a set of clusters.

Problem: How many clusters? Where do I cut?
Bayesian Solution

**Similarity:** Two profiles are “similar” when they are generated by the same process.

**Example:** EKGs are similar but not identical series generated by the a set of physiological process.

**Clusters:** Cluster profiles on the basis of their similarity is to group profiles generated by the same process.

**Bayesian solution:** The most probable set of generating processes responsible for the observed profiles.

**Strategy:** Compute posterior probability \( p(M|D) \) of each clustering model given the data and take the highest.


---

Bayesian Model Selection

- The set of possible models \( M \) is a stochastic variable with a probability distribution \( p(M) \).
- We want to select the model \( m_i \) with the highest posterior probability given the data \( \Delta \).
- We should search all models (combinations of series) and find the one with highest posterior probability.
- We use Bayes’ theorem:

\[
p(M | \Delta) = \frac{p(\Delta, M)}{p(\Delta)} = \frac{p(\Delta | M) p(M)}{p(\Delta)}
\]
The Clustering Model

- We need to **merge** these separated models in a single global model of the processes.
- We **add a variable** indexing the series and we build a single conditional probability distribution.

\[ y = X\beta + \epsilon \]

Bayesian Microarray Analysis

---

CAGED

- Cluster Analysis of Gene Expression Dynamics.
- Distributed in over 800 copies since August 2002.
- It was the most read PNAS on-line paper in August.
- We use a dataset of 512 genes collected to study the response to fibroblast to serum (Iyer et al. 1999).
**Validation**

* Back in 1999, 238 out of 517 genes were unknown.
* We relabeled the genes according to the current state of the art and less than 20 are left unknown.
* There are 19 repeated genes in the dataset:
  - Original model puts 4 of these in different clusters;
  - We put 1 of these in two different clusters.
* Interestingly enough, if we run the clustering with Markov order 0 (assuming uncorrelated iid data), we get 4 “misplacements” as well, albeit of other genes.

---

**CAGED for Muscle Development**

![Cluster Analysis Diagram](image)

Tomczak et al, FASEB J, to appear
Bayesian Microarray Analysis

Analysis of Cluster 15

Immunofluorescent Staining Caveolin3

Bayesian Microarray Analysis

Decoding Control

**Assumption:** Clustering begs the assumption that genes behaving *similarly* belong to the same process.

**Results:** Clustering breaks down the set of genes into boxes containing genes of the same process.

**Control:** Clustering tells us little about these control structures: it provides boxes, not chains of command.

**Problem:** How do I move from similarity to control?

**Bayesian solution:** To discover chains of command, we need to a new approach: Bayesian networks.
Bayesian Networks

Qualitative: A dependency graph made by:
Node: a variable $X$, with a set of states $\{x_1, \ldots, x_n\}$.
Arc: a dependency of a variable $X$ on its parents $\Pi$.

Quantitative: The distributions of a variable $X$ given each combination of states $\pi_i$ of its parents $\Pi$.

Learning Bayesian Networks

Components: Components to be learned are:
- Probabilistic functions: linking child to parents;
- Dependency structure: the network of links.

Structures: Learning a structure involves:
- Search strategy: to explore the possible structures;
- Scoring metric: to select a dependency (probability).

Challenges: Limitations of current methods:
- Distributions: Only discrete and normal distributions;
- Dependencies: Only linear (for normal distribution).

Solution: Generalized Gamma Networks (GGNs).
**Differential Analysis**

**Example:** Take the genes differentially expressed in the prostate cancer database and learn the structure of each condition (normal vs tumor).

**37605:** collagene is independent of all other nodes, and becomes a child of 914 (oncogene with transcription regulation functions).

Changes in 40282_S_AT determine changes in tumor markers.

Bayesian Microarray Analysis
Leukemia and Multiple Phenotypes

Data: 41 pediatric patients with leukemia, expression measures on 72 genes, 3 clinical phenotypes.

Validation

Cross-validation: A form of predictive validation.
1. For each case, remove it from the database;
2. Learn probability distributions of the network;
3. Predict value on a variable of the removed case.

Oncogene Status:
Coverage: 100% (41); Accuracy: 97.56% (40).
Average Distance: 0.03339.

Survival Status:
Coverage: 97.56% (40); Accuracy: 100% (40).
Average Distance: 0.004146.
Bayesian Microarray Analysis

Answers

To log or not to log?
    Do not log, understand the nature of your data.
Normalize: when, why and how?
    Don’t change your data, may contain information.
How many clusters?
    Ask ‘which clusters’ and get the most probable ones.
How to handle temporal experiments?
    Don’t forget they are dependent observations.
How to move from similarity to control?
    Use arrows and point them in the right direction.
And join the Bayesian cult!