Recruitment actively underway for
• graduate students
• post-doctoral fellows
• research track faculty
• tenure track faculty

Contributors

- Grier Page
- Jode Edwards
- Prinal Patel
- Tapan Mehta
- Murat Tank
- Jelai Wang
- Moonseong Hoo
- Jose Fernandez
- Gary Gadbury
- Tom Prolla
- Cheol-Koo Lee
- Rick Weindruch
- Jaap Brand
- Christopher Coffey
- Mark Beasley

Collaborators, Consultants, Confidants

- Anthony Ferrante
- Paska Permana
- Angelo Del Parigi
- Antonio Tataranni
- Kei Cheung
- Ken Williams
- Bob Eckel
- Upender Manne
- Tim Garvey
- John Mountz
- Christina Kendzierski
- Bill Grizzle
- Stephen Barnes
- Brad Yoder
- Lisa Gay-Woodford
- William E. Grizzle
- Dale Benos
- Bob Kesterson

Supported by:
- NSF – Plant Genomics (Allison); RCN on Microarray Analysis (Allison); SGER (Edwards)
- NIDDK – CNRU (Allison); Biotechnology Center (Yale; Williams); Short Course on Stat Genet (Allison)
- NIA – P01 & R01 (Weindruch); R01 (Prolla)
- NCI – U54 CNGI (Barnes)
Keeping Up with the Microarray Literature: How Many Can You Read Per Day?

From Mehta, Tanik, & Allison (in preparation).

High Dimensional Biology

What do we mean?

Biological investigations in which we are studying many (e.g., thousands) of variables, estimating many parameters, and making many inferences.

What are some of the issues?

I. Epistemologic Foundations
II. Desired properties of methods.
III. Study design (e.g., ‘power’) with untraditional desired properties.
IV. Computational burden.
V. Combining diverse types of data.
VI. "Pluralitas non est ponenda sine neccesitate"
Epistemological Foundations

• Epistemology is the study of how we come to have and what constitutes knowledge.

• Given a set of statistical procedures judged to be valid, a sound epistemological foundation for biological science comes, in part, from the application of those procedures.

• But how do we derive knowledge about the validity of our statistical methods such that they are also enjoy a solid epistemological foundation?

The problem of error has preoccupied philosophers since the earliest antiquity. According to the subtle remark made by a famous Greek philosopher, the man who makes a mistake is twice ignorant, for he does not know the correct answer, and he does not know that he does not know it.

Borel, Emile
Probability and Certainty
A Perspective

♦ We study:

♦ We wish to obtain knowledge about:

Samples

Data

Populations

Nature

Things Statisticians Do:

Develop Design & Analysis Procedures to Facilitate:

• Measurement - (e.g., produce a variable Y’ that represents Y).

• Prediction - (e.g., ‘impute’ unobserved values of X using observed Y).

• Estimation - (e.g., estimate \( \Delta = \mu_1 - \mu_2 \)).

• Inference - (e.g., conclude whether \( \delta = 0 \)).

• Classification - (e.g., for \( j = 1 \) to \( k \), sort the \( Y_j \) into \( m < k \) groups).
What is Validity?

- It usually (but not necessarily) comes in degrees.
- There are multiple potential criteria.
- The choice among criteria is inherently subjective.
- Whether or the extent to which criteria are met can (should) be objectively documented.

Some Properties of Estimators

Unbiasedness
Mean Square Error
Consistency
Efficiency
Sufficiency
Empirical Bayes (EB) Estimation of Gene-Specific Effects

♦ Why

♦ Who

♦ How

Empirical Bayes (EB) Estimation of Gene-Specific Effects - II

Gene Expression Differences 5 Insulin Sensitive vs. 5 Insulin Resistant Humans
(Data Courtesy Paska Permana - NIDDK).
Suppose we conduct a t-test of the difference between two means and obtain a p-value < .05. Does this mean:

a) There is less than a 5% chance that the results are due to chance.

b) If there really is no difference between the population means, there is less than a 5% chance of obtaining a difference this large or larger.

c) There is a 95% chance that if the study is repeated, the result will be replicated.

d) There is a 95% chance that there is a real difference between the two population means.


---

**Inferential Validity**

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>Truth</th>
<th>Null</th>
<th>Alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>a</td>
<td>b</td>
<td>K-R</td>
</tr>
<tr>
<td>Alt</td>
<td>c</td>
<td>d</td>
<td>R</td>
</tr>
<tr>
<td>K-M</td>
<td>M</td>
<td>K</td>
<td></td>
</tr>
</tbody>
</table>

FWER = P(c>0)

FDR = E[c/(c+d)]

EDR = E[d/(d+b)]

TP = E[d/(c+d)] = 1 – FDR

TN = E[b/(a+b)]
Under the null hypothesis, the distribution of p-values is uniform on the interval [0,1] regardless of the sample size and statistical test used (as long as that test is valid).

Under the alternative hypothesis, the distribution of p-values will tend to cluster closer to zero than to one.

Mixtures of Betas

Any distribution on the interval [0,1] can be modeled as a mixture of \( v + 1 \) separate component distributions where the \( j^{th} \) component is a beta distribution with parameters \( r_j \) and \( s_j \). The beta’s PDF is:

\[
\beta(r, s)(x) = I_{(0,1)}(x) \frac{x^{r-1}(1-x)^{s-1}}{B(r, s)}
\]

When \( r = s = 1 \), the beta distribution is a uniform distribution.

The log of the likelihood for the collection of \( k \) p-values from a model with \( v + 1 \) components can then be expressed as:

\[
L_{v+1} = \sum_{i=1}^{k} \ln \left( \lambda_{0i}\beta(1,1)(x_i) + \sum_{j=1}^{v} \lambda_{ji}\beta(r_j, s_j)(x_i) \right)
\]

\( x_i \) is the p-value for the \( i^{th} \) test,

\( \lambda_0 \) is the probability that a randomly chosen test from the collection of tests is for a gene for which there is no population difference in gene expression,

\( \lambda_j \) is the probability that a randomly chosen test is for a gene from the \( j^{th} \) component distribution for which there is a true population difference in gene expression.
Allison, David B.

Figure 4. The Weindruch et al. Mouse Cortex Data (Old Ad Lib vs. Old Calorically Restricted)

Estimated parameters: $\lambda_1 = .29; \ r_1 = .78; s_1 = 3.87.$

Null-Simulated Cortex-Real

Posterior (Bayesian) Probabilities

<table>
<thead>
<tr>
<th>Gene</th>
<th>$P_{11}$</th>
<th>$P_{12}$</th>
<th>$P_{22}$</th>
<th>$P_{23}$</th>
<th>$P_{33}$</th>
<th>P-Value ($t$-test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.74</td>
<td>.24</td>
<td>362</td>
<td>.17</td>
<td>.94</td>
<td>3.68</td>
</tr>
<tr>
<td>2</td>
<td>1.74</td>
<td>.04</td>
<td>362</td>
<td>.17</td>
<td>.94</td>
<td>3.68</td>
</tr>
<tr>
<td>3</td>
<td>1.74</td>
<td>.04</td>
<td>362</td>
<td>.17</td>
<td>.94</td>
<td>3.68</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Gene P-Values

<table>
<thead>
<tr>
<th>Gene</th>
<th>P-Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.16982</td>
</tr>
<tr>
<td>2</td>
<td>0.00142</td>
</tr>
<tr>
<td>3</td>
<td>0.48675</td>
</tr>
<tr>
<td>4</td>
<td>0.00027</td>
</tr>
</tbody>
</table>

Posterior Probability

Probability that the gene $i$ is a gene from the population of genes for which the null hypothesis is not true.
A parametric bootstrap approach to sample size estimation for microarray research.


To determine targets of TNF signaling of NF-κB in RASF, replicate samples of RASF transfected with AdIkB.DN (n=3) or control AdTet (n=3).

Steps:
1. Fit a model to a set of observed data.
2. Simulate data from a model with the estimated parameters and observe what happens.
3. Repeat step 2 m times.
4. Summarize results.
A Challenge for Classification: Lack of a Null Hypothesis.

If there is no right answer, how do we know if we got a good one?

Cluster Analysis: Four Caveats

1. It's Not New.


2. It has well-recognized problems.


“...availability of computer calculation methods has led many psychiatrists to using the new data techniques uncritically. Spurious findings may result. The available clustering techniques should be validated, as by applying them to sets of data of known structure. Most present methods may be defective.”

3. It can be very computationally demanding.

4. It may not answer a particular investigator’s questions.

Classification Validity

- Replicability (more obvious how to assess when clustering genes than cases)
- Utility
Epistemologically Valid Frameworks: Induction & Deduction

♦ Deduction: i.e., mathematical proof.
♦ Induction:
  • Simulations
  • Plasmodes
♦ Composite Approaches: Application to multiple real data sets of unknown nature with methods of partially known properties.

A Circular & Epistemologically Invalid Framework

♦ Application to single real data sets of unknown nature.

What Constitutes Proof: What assumptions are being made?

♦ Normality? Non-Parametric: Non-Panacea (Cohen, J.)
♦ Exchangeability?
♦ Independence?
♦ Other? Asymptotic ≠ Exact

Simulated P-value for 42 out of 42
Evaluating Evidence from Simulations

- Number of Trials
- “Nuisance” Variables:
  - Samples size
  - Equality of sample sizes
  - Heterogeneity of variances
  - Distributions
  - Correlation structure
- “Genetic Model”
- Stringency of Criteria

Issues in Simulating

- Computing Power
  - RAM
  - Speed
  - Can importance sampling help?
- Modeling Dependency Structure
  - Get Real?
    - Block Diagonal
    - Autocorrelation
    - Capture Image
  - How to Make ‘Gridable’
Example of dependencies between genes in bio-chemical pathways

(Schwartz, M.W et al, 2000 in Nature)

A Novel $\alpha$-Spending Function for More Powerful Testing in High-Dimensional Biology. Brand et al. (under development)

- Predictive weight $\hat{w}_i$ is the linear combination of differential expression levels of the donor genes which has the maximal correlation with $d_i$ of target gene $i$.

$$\hat{w}_i = \frac{\sum_j |f_{ij}|}{\sigma_y}$$

- Weight $w_i$ is the standardized absolute value of predictive weights such that all weights sum up to number of genes $k$.

Vector of correlations between target gene $i$ and its donor genes

Inverse matrix of correlations between the donor genes for target gene $i$

Observed expression levels of the donor genes of target gene $i$.
An Example of Ongoing Work
A Novel $\alpha$-Spending Function for More Powerful Testing in High-dimensional Biology
Brand et al. (under development)

![Bar graph showing power of alpha-spending relative to Bonferroni]

Increasing Correlation & Fewer True Nulls

Now it Gets Interesting!

<table>
<thead>
<tr>
<th>Gene</th>
<th>Estimated Expression Difference</th>
<th>Estimated Linkage QTL</th>
<th>Protein Characteristic</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Empirical Bayes may be ideally suited.
Allison, David B.

Will Analyze Data for Food