Statistical Analysis of proteomic data

Grier P Page Ph.D.
Associate Professor
Section on Statistical Genetics
Department of Biostatistics
School of Public Health

Gpage@uab.edu
4-4930
Ryals 317D
Keeping Up with the Microarray Literature: How Many Can You Read Per Day?

![Microarray Articles in PubMed Graph]

From Mehta, Tanik, & Allison.

A Perspective on Statistics

- We study: Samples
- We wish to obtain knowledge about: Populations
- Data 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).

---

**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?
What is High Dimensional Biology?

- High Dimensional Biology – is a broad topic covering biological systems where the number of variables is very large.
- Topics that often fall in HDB are microarray, proteomics, linkage, and genomics.
- HDB is also highly collaborative both ‘wet’ and ‘dry’ lab people.
What Do All These Topics Have in Common?

Lots and Lots and Lots of Numbers !!!

If you have numbers what do you do?

- Statistics (and Design) !
- Or as most of you think Statistics Ugh!

- Most of the statistics used in HDB are identical to statistical methods that have been used for years.
- The thought process that goes into design is also similar to those that have been used for years.
Design

• Design is the art of designing an experiment in such a way that the question that is being asked can be easily and unambiguously answered.
• The experimental hypothesis drives the design.

Statistics

• Methods for making inferences about a population as a whole by taking a sample.
• Statistics and design work in harmony with the biology, while design and statistical may be the cause of alterations in experiments, the biology is the *sine qua non*.
What are Statistics and Design?

• The goal of experimental design and statistical analysis is to allow an investigator to answer the question that they would like to ask correctly and efficiently.

• Often statisticians are a reality check. If you can’t explain your experiment to a statistician will it make sense in a publication?

From T. Speed
Quality Issues - I

- Known sources of non-biological error (not exhaustive) that must be addressed
  - Technician
  - Chip lot
  - Reagent/gel lot
  - Printer tip
  - Time of printing
  - Date
  - Fluidics well/ Scanner/ position on scanner
  - Order of scanning
  - Location
  - Cage/ Field position
  - Far and away the largest issue is labeling
  - Software!

First Steps

- Look at your data/images do they look like they are supposed to?
- Then conduct semi-statistical analyses.
Which one is good or bad?

Distribution of the geography index (GEODEX) for individual chips

Image of chip 12

Image of the right corner on the bottom.
Top 2.5% of ratios red, bottom 2.5% of ratios green

Cluster Analysis of GG/BG Study

From Susan Hilsenbeck with permission
UMSA Analysis

Time of printing effects

Green channel intensities (log₂G). Printing over 4.5 days.
The previous slide depicts a slide from this print run.
From T. Speed/H Yang
Quality Issues – II

• How to address these issues
  – Make the experiment as uniform as possible
    • Agree on exactly what defines the tissue to be used, use same technician, same chip lot, same reagents (always buy a little too much), same scanner, do sample extraction, labeling and hybridization on one day if possible, establish quality control
  – Randomize when uniformity is not possible
    • Don’t do all of condition 1 on day 1 and condition 2 on day 2
    • Randomize the time a chips sits waiting to be scanned
    • Randomize animal cage/plant field position
• Microarrays generate such a huge volume of data that is is possible to detect these issues, I suspect that northerns, Southerns, RT-PCR, westerns, and more have similar problems.

Elements of Statistics

• Power – the probability of detecting something if it is there. Usually a function of sample size and size of difference to be detected
• Image Analysis
• Quality Control- normalization/transformation
• Normalization
• Statistical Analysis
  – Class discrimination
  – Class prediction
  – Class differentiation
• Annotation
• Bioinformatics issue
Image Analysis

- How do you go from an image to a number?
Which Size Circle?

Parts of other Proteins

Image Analysis

Inside the boundary is spot (foreground), outside is not.

From Helen Kim

From T. Speed
Measurement Properties of Microarrays


According to Biomed Central over 2000 accessions in first 6 weeks
Experimental design

8 rats
24 chips
15923 genes
Quality Control/Normalization

- Not all gels, chips, sequencing runs, etc are perfect
- Some are so bad they should be dropped
- Other can be fixed
  - Identify problem values/ areas
  - Fix them – adjustments and normalization
Types of Normalization

- Changes in mean
  - Add or multiple every value on a chip to take to some consistent value across chips.
  - Log transform
  - Quantile

- Changes in position
  - Lowess
  - Variance adjustments

Mean Normalization
Intensity correction only

Before  After
Composite normalization

Before and after composite normalization

From T. Speed

Statistical Analysis

- Statistical Analysis
  - Class discrimination
  - Class prediction
  - Class differentiation
Class Discovery

• Data visualization
• Cluster analysis
  – Clustering
  – Self organizing maps
• Multidimensional scaling
• Similarity searching

Cluster Analysis

• There are a large number of clustering algorithms.
  – Hierarchical
  – Non-hierarchical
  – Different weights
  – All will give different answers.
  – None are statistical tests

From Nature
From

K-Means Clustering

Source Unknown
Stability Results

Findings:
- Low stability (~55%) achieved for all four clustering algorithms even at the elevated sample sizes of n=50.
Class Prediction

• Discriminate Analysis
  – Build a predictive model for future data based upon previous data.
  – Each new sample is assigned the probability that it will fall into one the classes.
• Assign new samples to one of several groups
  – e.g. is a new tumor adenoma or squamous cell carcinoma

Class Prediction II

• Methods
  – Genes are selected with a discriminant function. Many exist, all very similar.
• Methods
  – Modeling building and test dataset are needed.
  – Often a form of cross validation such as 10 fold or leave one out are used.
From Nature
**Classification:**

**Train, Validation, Test split**

- **Results Known**
  - Training set
  - Validation set
  - Predictions
  - $+$
  - $-$

**Model Builder**

- **Evaluate**
  - Final Model
  - Final Test Set
  - Final Evaluation

---

**Cross-validation**

- *Cross-validation* avoids overlapping test sets
  - First step: data is split into $k$ subsets of equal size
  - Second step: each subset in turn is used for testing and the remainder for training
- This is called *$k$-fold cross-validation*
- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate

---

From Gregory Piatetsky-Shapiro
Cross-validation example:

— Break up data into groups of the same size
— Hold aside one group for testing and use the rest to build model

Test

Repeat

More on cross-validation

• Standard method for evaluation: stratified ten-fold cross-validation
• Why ten? Extensive experiments have shown that this is the best choice to get an accurate estimate
• Stratification reduces the estimate’s variance
• Even better: repeated stratified cross-validation
  – E.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)
Leave-One-Out cross-validation

- Leave-One-Out: a particular form of cross-validation:
  - Set number of folds to number of training instances
  - I.e., for $n$ training instances, build classifier $n$ times
- Makes best use of the data
- Involves no random subsampling
- Very computationally expensive
  - (exception: NN)

From Gregory Piatetsky-Shapiro

Class Differentiation

- Supervised Analysis
- What genes are most different between two or more groups
What is ‘Significant’

• There is no such thing as significant.
• P-values are a continuum of evidence
• RA Fisher thought a p < 0.05 merited further study.
• Has been abused toward significant.

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.

“There are other experiments, however, which cannot easily be repeated very often; in such cases it is sometimes necessary to judge the certainty of the results from a very small sample, which itself affords the only indication of the variability.”

-- Student (1908)
What should I use for 2-group testing?

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.
These are indications of bad data of lack of information.

Adjustedments for multiple testing

- All HDB studies involve many tests
- Given the definition of a p-value an adjustment is needed when many tests are conducted.
Testing Defined

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

\[ FDR = E\left(\frac{c}{c + d}\right) \]

\(c = \) type 1 error (alpha) – false positive
\(b = \) type 2 error (beta) – false negative

FDR - False Discovery Rate

- When many hypotheses are tested the sample size required for a Bonferroni corrected \( p < 0.05 \) were prohibitive in most contexts.
- Some attempts were made for intermediate adjustments
  - Lander and Botstein (1989) for linkage data
- Benjamini and Hochberg 1995 pulled together several streams of research on adjusting for multiple testing.
  - Developed method for setting an adjusted p-value that controlled for type I error
  - Like many statistical methods it has been ‘extended’ and abuse to a FDR estimating procedure
- Methods were developed for epidemiology and genetic studies, but were adapted for HDB studies
Family Wise Error Rate vs. False Discovery Rate

• Traditional FWER
  – Bonferroni $\alpha^* = \frac{\alpha}{n}$
  – Sidak $(1-(1-a)^n)$
    • Very conservative
    • Minimize False discovery rates
    • Assume independence

• False Discovery Rate
  – Designed to estimate the rate of error

Power and Sample Size

• This is where microarray experiments get the most criticism.
• Experiments performed without replication
• Impression that arrays much more expensive than they are now
• Belief that microarrays are not liable to the same experimental error that experiments are
• There also has not been a good way to calculate sample size
Power

- All power and sample size calculations require and estimate of population variability
- For microarrays we use a pilot project
- Based upon the posterior probability that a gene is differentially expressed it test statistic may be increased as a function of proposed increase in sample size
Data Interpretation

• The most time consuming portion of a HDB experiment is the interpretation
• Many databases and resources exist
  – Dr. Loraine talked about these in great detail
**a posteriori vs. a Priori data interpretation**

- Many people get the data and then stare at it and tell a story based on their subjective observations about the data.
- *a posteriori* observations are highly biased.
- *a priori* observations require knowledge of pathway, gene family, etc. There can be a large number of classes.

**Global/Meta Analytical Tests of Pathways**

**Premise:** We can learn something additional and/or test with more power if we consider the fact that genes may exist within ‘families.’ Several Tests –
- Fisher’s meta analytical tests – combine the individual p-values from n genes \( \sim \chi^2_{(2n-2)} \)
- Vote Counting methods
  - Onto-express
  - GSEA
- Normalize all the data to Z scores and compare the expression levels
- Issues even under \( H_0 \), if genes in a pathway are correlated there will be an increase in type I error
- Address FEWR vs FDR per group
Gene Family-Based Hypothesis Testing: 
*What people say they are testing vs what they are testing.*

**Which Null?**

1. None of the genes in family c are differentially expressed. 
2. The proportion of genes in family c that are differentially expressed is equal to the proportion of genes in the remainder of the genome that are differentially expressed. 
3. The correlation matrix among the expression levels of the genes in family c is an identity matrix. 
4. The correlation matrix among the expression levels of the genes in family c is the same across experimental conditions. 
5. The intersection of #1 and #3.

Mootha et al (2003). “We introduce an analytical strategy, Gene Set Enrichment Analysis, designed to detect modest but coordinate changes in the expression of groups of functionally related genes.”

This implies that the null of interest is #1, but the test appears to be the intersection of #2 and #3.

---

**Global/Meta Analysis**

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Function Name</th>
<th>Total</th>
<th>P-Value</th>
<th>FDR</th>
<th>Bonferroni</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>inflammatory response</td>
<td>71</td>
<td>1.11E-16</td>
<td>4.72E-14</td>
<td>4.72E-14</td>
</tr>
<tr>
<td></td>
<td>immune response</td>
<td>95</td>
<td>8.44E-15</td>
<td>1.79E-12</td>
<td>3.59E-12</td>
</tr>
<tr>
<td></td>
<td>epidermal differentiation</td>
<td>38</td>
<td>1.65E-11</td>
<td>2.34E-09</td>
<td>7.02E-09</td>
</tr>
<tr>
<td></td>
<td>cell-cell signaling</td>
<td>100</td>
<td>3.14E-10</td>
<td>3.14E-08</td>
<td>1.57E-07</td>
</tr>
<tr>
<td></td>
<td>cell adhesion</td>
<td>77</td>
<td>5.72E-09</td>
<td>4.86E-07</td>
<td>2.43E-06</td>
</tr>
<tr>
<td></td>
<td>chemotaxis</td>
<td>43</td>
<td>8.73E-09</td>
<td>6.18E-07</td>
<td>3.51E-06</td>
</tr>
<tr>
<td></td>
<td>cellular defense response</td>
<td>40</td>
<td>1.74E-08</td>
<td>1.06E-06</td>
<td>7.39E-06</td>
</tr>
<tr>
<td></td>
<td>development</td>
<td>80</td>
<td>1.44E-08</td>
<td>1.31E-06</td>
<td>1.46E-06</td>
</tr>
<tr>
<td></td>
<td>antimicrobial humoral response</td>
<td>45</td>
<td>9.90E-08</td>
<td>4.68E-06</td>
<td>4.21E-05</td>
</tr>
<tr>
<td></td>
<td>response to viruses</td>
<td>18</td>
<td>7.16E-07</td>
<td>3.04E-05</td>
<td>3.04E-04</td>
</tr>
<tr>
<td></td>
<td>cell surface receptor linked signal transduction</td>
<td>54</td>
<td>2.99E-06</td>
<td>1.97E-04</td>
<td>1.44E-03</td>
</tr>
<tr>
<td></td>
<td>cell motility</td>
<td>47</td>
<td>3.55E-06</td>
<td>1.26E-04</td>
<td>1.51E-03</td>
</tr>
<tr>
<td></td>
<td>cell proliferation</td>
<td>79</td>
<td>1.81E-05</td>
<td>5.90E-04</td>
<td>7.67E-03</td>
</tr>
<tr>
<td></td>
<td>protein biosynthesis</td>
<td>6</td>
<td>1.81E-05</td>
<td>5.49E-04</td>
<td>7.69E-03</td>
</tr>
<tr>
<td></td>
<td>skeletal development</td>
<td>36</td>
<td>2.39E-03</td>
<td>7.34E-04</td>
<td>1.10E-02</td>
</tr>
</tbody>
</table>
Use of FDR for Union-Intersection tests

• Traditional
  – The ‘min’ test.
  – Low power
  – Not of definitive size
  – Ignores information (i.e., the p-value for min test is largest p-value for \( H_0 \in H_0 \) regardless of the value of any other p-values).

• Informational based approaches
  – All p-values are not equal
  – A variety of ways to weight
  – Let’s consider FDR or PTP – these are equal across datasets
  – Can conduct simple product of FDR.
Venn Diagram of 100 gene more Significantly associated with known CESA genes: ICE Intersection of Coexpression

Primary

Secondary
Bioinformatics Issues

- HDB studies generate a huge amount of information.
- Storage and handling of the data can be difficult.
- Data standards are developing (MIAME for microarrays), proteomics just beginning.