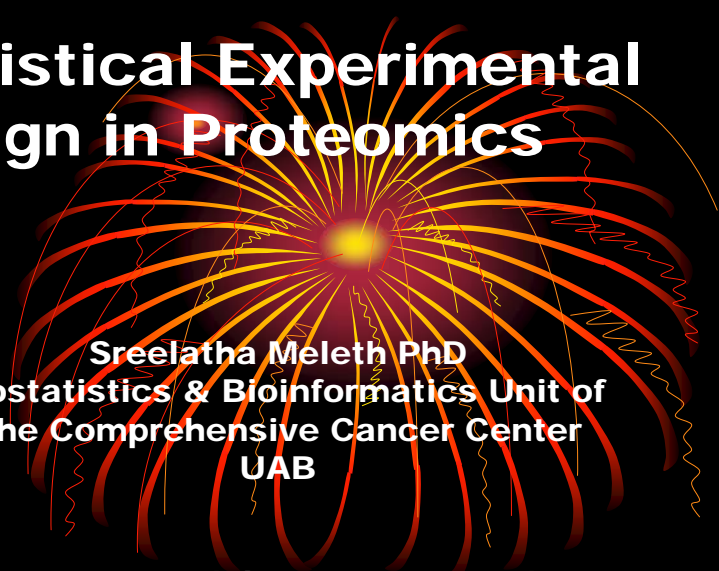



Statistical Experimental Design in Proteomics



Sreelatha Meleth PhD
Biostatistics & Bioinformatics Unit of
The Comprehensive Cancer Center
UAB

University of Alabama in Birmingham
Feb 2007

Proteomics & Biomarker Discovery Promises not kept ?



- To date the contribution of proteomics methods based on mass Spectrometry to the diagnostic armamentarium has been disappointing - Rifai N, Gillette MA, & Carr SA(2006)
- 'Concurrent with the explosion in the number of publications reporting biomarker discovery by profiling technologies such as proteomics and pattern recognition, has been the increase in evidence highlighting the susceptibility of these approaches to analytical & experimental bias' Teahan et al (2006)

Examples of promises- SELDI



- Lancet 2002 - Petricoin et al (2002) - Proteomic Profile for Ovarian Cancer: Se = 100% , Sp = 95% .
- Biometrics (2003) - Yatsui et al : Proteomic Profile for Prostrate Ca : Proposed method nearly perfect.
- SELDI and Cancer & 2006 - Pubmed Search 216 articles.

Examples of Promises - 2D gels



- ❖ Polypeptides associated with HP differentiation of primary lung CA (*Hirano et al 1995*)
- ❖ Protein modifications associated with Heart failure in 3 cellular systems (*Jiang et al 2001*)
- ❖ Identified proteins associated with AD (*Schoenberger et al 2001*)
- ❖ Used to classify human tumors as benign or malignant (*Alaiya et al 2000*)

Promises not kept & Statistics – Is there a connection?



- Expensive Technologies
- Small samples
- Large number of variables – hi-dimensions
- Lack of experimental design
- Particularly – no replication, no randomization

How to use Statistical Techniques Box GEP et al (1978)



- Find out as much as you can about the problem
- Some questions to be asked
 - What is the object of the investigation?
 - I am going to describe your problem. Am I correct?
 - Do you have any data?
 - How were these data collected?
 - In what order?
 - On what days?
 - By whom?

Box et al ...2

- How does the equipment work?
- What does it look like?
- May I see it?
- May I see it work?
- Do you have other data like these?
- How much physical theory is known about this phenomenon?

More from Box..

- Good statistical work seems to result from a genuine interest in practical problems
- Fisher worked closely with experimenters; was one himself
- "To many in the statistical world 'Student' was regarded as a statistical adviser to Guinness's brewery; to others he appeared to be a brewer devoting his spare time to statistics....though there is some truth to both these ideas they miss the central point, which was the intimate connection between his statistical research and the practical problems he was engaged in".

Statistical Experimental Design



- Measuring variability and attributing variability to different sources is a major part of statistical analysis
- Statistical Experimental design – aims to estimate, isolate or neutralize the variability
- Uses - Replication, Randomization & Blocking

Replication



- Biological replicates-sample size-power to detect between group variance
- Technical (same sample) replicates- helps estimate within group variance
- In techniques such as 2D gel, & micro-array technical replicates also help as a quality control measure
- E.g. Are protein spots seen in all replicates of a sample?

Blocking

- Blocking - Create blocks of observations that have very similar variance
- Have every treatment group represented in each block
- e.g., Processing a 2D gel extraneous variability caused by day of processing and / technician involved
- Technicians, day will both be used as blocking factors



Randomization

- After getting a good understanding of process, and variables decide
 - Which variables to block for
 - Which variables are uncontrollable
- Uncontrollable variables neutralized by randomizing across those variables



2D Gel Experience

- Opposite from the SELDI
- Initially small data sets, reliable data.
- Useful biological results
 - 2D advantage – ID of proteins – easier to confirm biological plausibility
- As experiments became more complicated
- Data quality suffered

2D GELS

- 2D gels made by distributing proteins of a biological material in 2 dimensions
- X-axis- isoelectric point (pI)
- Y-axis- molecular weight
- Phosphorylation - Change in location along the X-axis
- Change in location along Y-axis not usual
- Objective - find changes in proteins-expression /location

Issues with statistical analysis of 2D gels .1



- Although 2D gels have been around for a long time statisticians have not been very interested in the analysis of 2D gel data
- Partly because a lot of for years was done by the protein biologists, who depended only on visual analysis, and often based his/her results on the visual inspection of one sample from each group of interest.
- New and improved gel creation as well as improvement in imaging technology has opened the door to larger more sophisticated experiments
- However, much like micro array, most of us have plunged headlong into it with no time to really understand the technology and its limitations and strengths

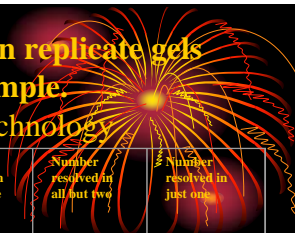
Reproducibility of a 2D Gel



- Variability between GELS - Big problem
- Voss & Haberi (2000) - 49 gels -from 25 patients with Chronic Lymphatic Leukemia
- Pair wise matching - 90 - 95% of all spots matched
- Pairs compared to master - matching efficiency 89%
- Only 7 spots matched in all gels


The number of protein spots resolved in replicate gels of the same biological sample

Meleth et al 2005 – BMC Biotechnology



Sample No (Treatment group)	No. of Replicates	Total Proteins resolved (occurring in at least one replicate) (%)	Number resolved in all replicates (%)	Number resolved in all but one replicate (%)	Number resolved in all but two (%)	Number resolved in just one (%)
6 (GSE)	3	4064 (100)	309 (8)	563 (14)	3192 (78)	-
7 (GSE)	4	546 (100)	169 (31)	130 (26)	97 (18)	150 (27)
8 (GSE)	4	954 (100)	186 (19)	120 (12)	105(11)	543 (57)
9 (GSE)	2	904 (100)	342 (39)	562 (62)	-	-
10 (GSE)	2	396 (100)	229 (58)	167 (42)	-	-
22 (CONT)	4	924 (100)	234 (25)	89 (10)	109(12)	492 (53)
23 (CONT)	4	950 (100)	161 (17)	151 (16)	102 (11)	536 (59)
24 (CONT)	2	879 (100)	312 (35)	567 (65)	-	-
25 (CONT)	3	957 (100)	272 (28)	117 (12)	568 (59)	-
26 (CONT)	2	432 (100)	183 (42)	249 (58)	-	-

Effect of Pre-Processing


- 
- In a paper published earlier last year (Meleth et al 2005), we have demonstrated that simple changes in a statistical protocol such
 - as a different normalization formulae
 - log transformation versus no transformation
 - Different methods of missing spot intensity imputation
 - Averaging across tech reps versus treating them as independent samples(maybe justifiable given poor reproducibility)
 - Produces different lists of significant spots.
 - Large differences (spot present / absent in all) constant – however subtle differences important

Initial Reaction to these differences




- Need data with known proteins, known quantities
- Our usual solution of simulated data difficult to obtain
- Design experiments to identify optimal statistical methods?
- Important solution but...
- As always it is more complicated

Sources of data variability - Do we really know?




- *Teahan et al (2006) report that*
 - Blood samples being collected onto ice vs in absence of ice
 - Over a series of serum
 - Clot contact times
 - The stability of NMR samples over time
 - Effect of freezing on the metabolic profile
- *All caused slight alterations to the NMR profile that could produce a systematic bias.*
- *Statistical Experimental Design can help one reduce or mitigate the effects of different sources of variability - only if we have a clear idea of process involved in creating the data set*

PI / Statistician interaction




- A number of different designs- CRBD, Latin squares, Split-plots
- Choice depends on close consultation with PI, lab personnel
- Is this design practical?
- You need to say 'yes it is', or 'no it is not'
- Good idea to let statistician to see process in lab

Importance of interaction



- Experimental design has to be adapted to technology and laboratory procedures
- Case in point
 - Recent observation of DIGE gel creation
 - Process takes 4 days at a minimum
 - Monday / Tuesday start 2nd→Dim done – Thursday /Friday
 - Later start freezing of 1st dimension gels

Importance of interaction - 2



- Initial reaction – randomize so that all groups have equal probability of being early in the week or later in the week - so that any variability caused by freezing after randomization is spread across all groups, technical and biological replicates
- Later – Is it possible to avoid this variability all together? Is it really an uncontrollable source of variation?

Benefits of Interaction - Customizing Experimental Design to suit specific technology



In 2D experiment

- Image analysis – a crucial part in the data collection
- Image Analysis is very dependent on matching of gels first within a group and then across a group
- Random assignment of treatment groups across days, technicians etc, might convert variability into random error
- However, it might also increase the variability between gels with the same treatment group to unacceptable levels.
- The need to match gels → more important to reduce variability that to distribute it as random error across the group

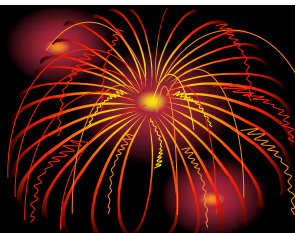
PS: We do not know the effect of freezing after the first dimension on spot location / intensity – need to study

Statistician /PI interaction - 3



- Allow for a learning period.. For both statisticians and you
- E.G: Randomization in a 2D gel experiment
- As above, randomization is used to convert uncontrollable sources of variation, into random error that is distributed with equal probability across all groups of interest
- Advantage – this uncontrollable variability does not affect one group disproportionately and will not be treated as an artifactual treatment effect

Statistical Analysis



- Important to use a statistician
- Most software provides two sample t-tests/ ANOVA
- Both tests above assume equality of variance and normal distribution of samples
- No provision in software to transform data or assess adherence to assumptions
- Meleth et al (2005) demonstrated that different techniques of normalization, transformation, missing data imputation alter conclusions drawn.
- Karp et al demonstrate that treating replicates as independent versus, nested alters the list of significant proteins

In summary

- Before new technologies are implemented in the search for new biomarkers it is important for the Biologist and Statistician to understand the controllable and uncontrollable sources of variation in the process
- Frequent interaction between PI, Statistician and lab personnel important to design experiments that are both practical and scientifically sound
- Data from Proteomics experiments should be sent to a statistician for analysis in order to ensure the validity of the results.

Questions?

Thank you!

