Design for a three-stage proteomic study

Irene Zeng, Thomas Lumley and Kathy Ruggiero, Department of Statistics, University of Auckland.

Background

- In the majority of current reported proteomic studies, laboratory selection and clinical validation of protein markers are two separate processes in study design.
- The US National Cancer Institute advocated a systematic design framework
  - stage I laboratory discovery
  - stage II targeted verification
  - stage III clinical validation in large, multicenter study

- We are developing methods to plan these designs
  - control of false positives and false negatives, under cost constraints for the entire design
  - ways to use biological information on groups of proteins to help in selection

	Individual proteins are selected separately	Group of proteins are selected from a biological pathway
Protein levels are independent	Individual t-tests at each stage [results below]	Hotelling's T^2 or tests based on ranking p-values [we know roughly how to do this]
Protein levels are correlated within pathways	Individual tests taking account of correlation [needs research]	Hotelling's T^2 or tests based on ranking p-values, taking account of correlation [needs research]

Unbiased discovery

- Proteins are identified from discovery stage. The selection of protein candidates are determined by pathways, and statistical criteria
- Sample size: 10's
- Analytical platform : High throughput Liquid Chromatographic Mass Spectrometry LC-MSMS with iTRAQ label

Targeted verification

- Filter out false positives due to chance or to platform artifacts.
- Sample size: 100's
- Analytical platform : MRM Mass Spectrometry

Clinical validation

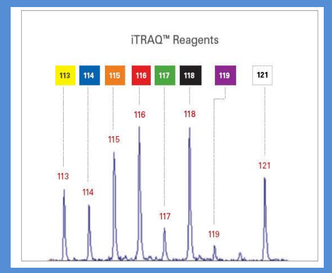
- Final validation: are differences in this protein clinically useful to measure.
- Sample size: 1000's
- Analytical platform : Immunoassay

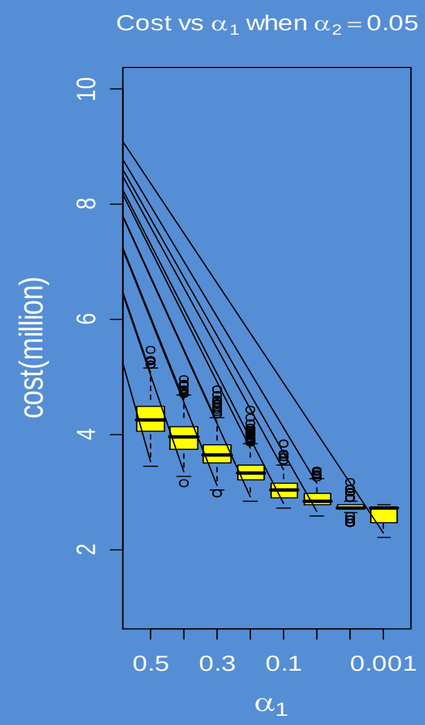
What is iTRAQ?

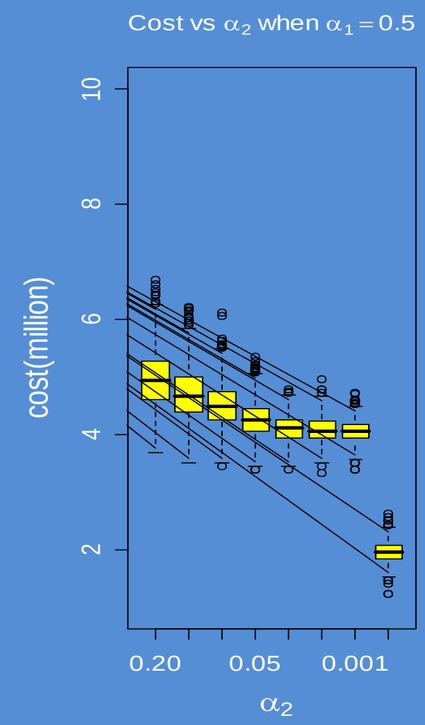
A popular approach is to use high throughput mass spectrometry with labeling technology such as iTRAQ. iTRAQ is a set of labeling reagents that allow multiple samples to be processed together and still produce distinguishable results

What is MRM?

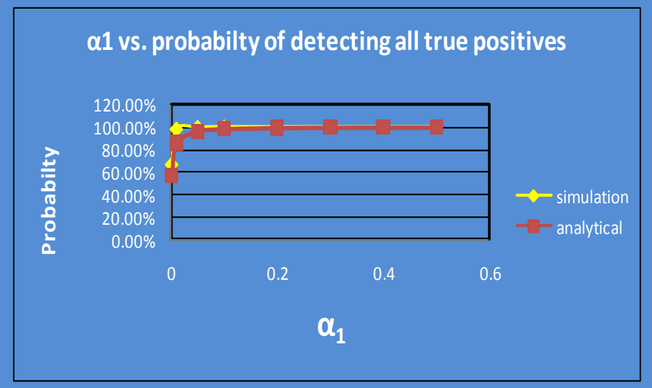
Multiple Reaction Monitoring Mass Spectroscopy is used to measure known proteins. When the sample proteins are fragmented, only fragments that correspond to the target proteins are measured.

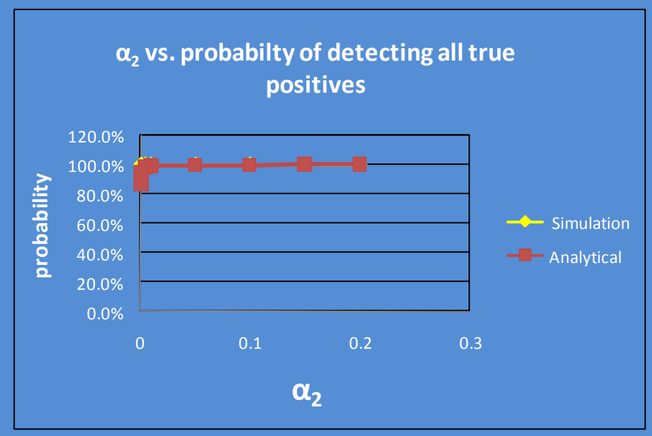






*All box plots are generated from 1000 simulations with different thresholds at stage I and stage II; the last box plot of the second graph had smaller sample size at stage III.
α1 represents the significant level at stage I and α2 represents the significant level at stage II.





Conclusions

- Stage I threshold has large impact on cost and power
- The stage II significance threshold has little impact on the results.
- Stage III sample size affects power, but 1000 may be too large
- Since stage II is partly a technical verification, future simulations should include technical artifacts
- The simulation code agrees well with the analytical code. Analytical results may not be available for more complex scenarios.