Statistical Issues in Cancer Proteomics

Alexander Ploner

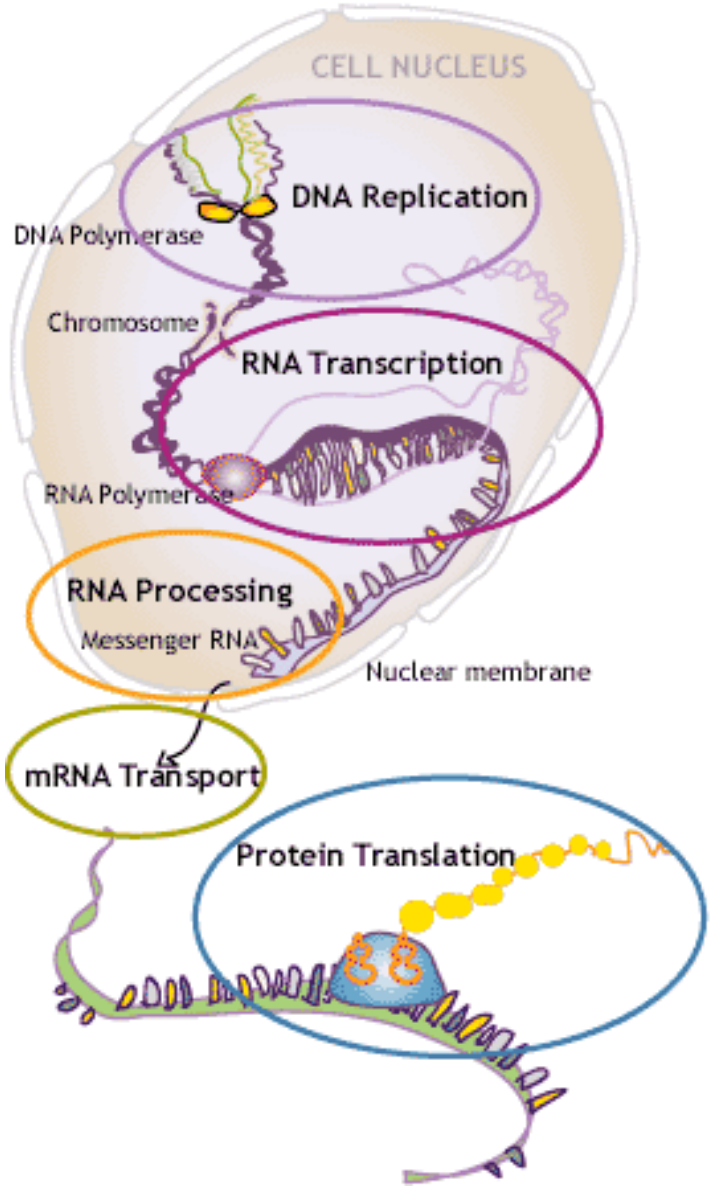
Medical Epidemiology & Biostatistics
Karolinska Institutet

http://www.meb.ki.se/∼aleplo

November 2, 2004

Overview

- Proteomics & Cancer Proteomics
  - Why?
  - Why is it hard?
- Technology & Data
  - How to measure proteins
  - Issues: biological, technical, statistical
- TOF-MS
- Experiences with the analysis of microarray data and how they may translate into working with proteomics data



The Central Dogma

- DNA: What the cell **can** do
- mRNA: What the cell **tries** to do
- Protein: What the cell **does** . . . potentially

Source: `http://nobelprize.org/medicine/educational/dna/`

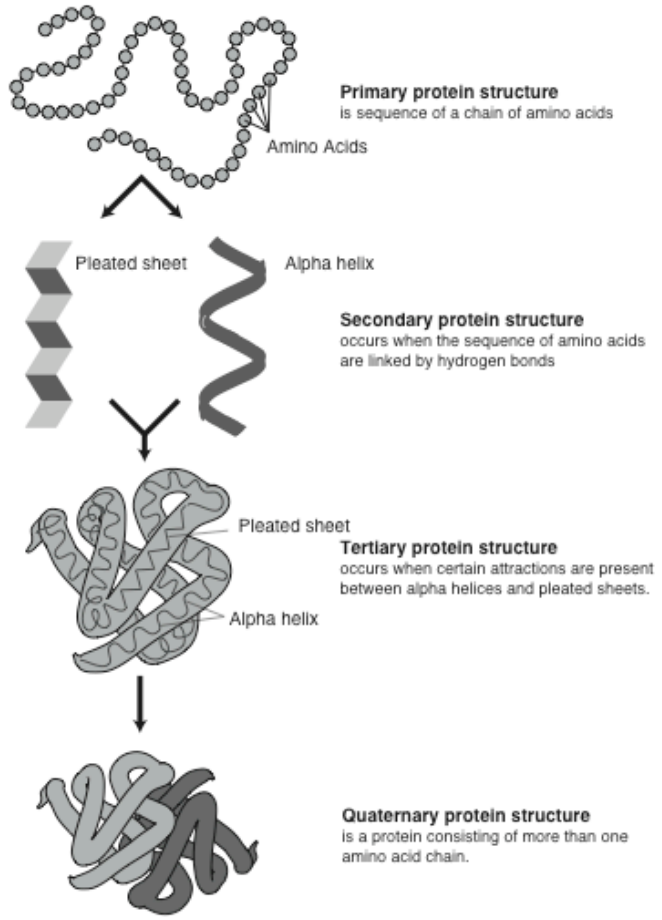
Proteins

Functions Enzymes, hormones, growth factors, transcription factors, membrane receptors, transporters, antibodies, . . .

Modifications Phosphorylation of proteins can regulate their biological activity

Splice variants Often expressed sub-sequences of a gene (exons) can be selectively arranged to yield different proteins

Structure



The Proteome

- The collection of proteins found in a specific $cell\ type$ in a particular $environment$ OR any useful sub- or superset thereof (human, mitochondrial p.)

- Dynamic range from 1 copy/cell to 10^6 copies/cell

- The human genome of ca. 30000 genes codes for many more proteins:
  - alternative splicing (2 to 5 fold increase)
  - post-translational modifications (2 to 5 fold increase)
  - alternative initiation, RNA editing, ribosomal frameshifting, protein splicing, protein aging, . . .

'Genes were easy ' – Cancer Proteomics

Of general biological interest:

- Protein inventory of reference
- Cell signaling
- Characterization of active proteins
- Functional protein interactions
- . . .

Clinical perspectives:

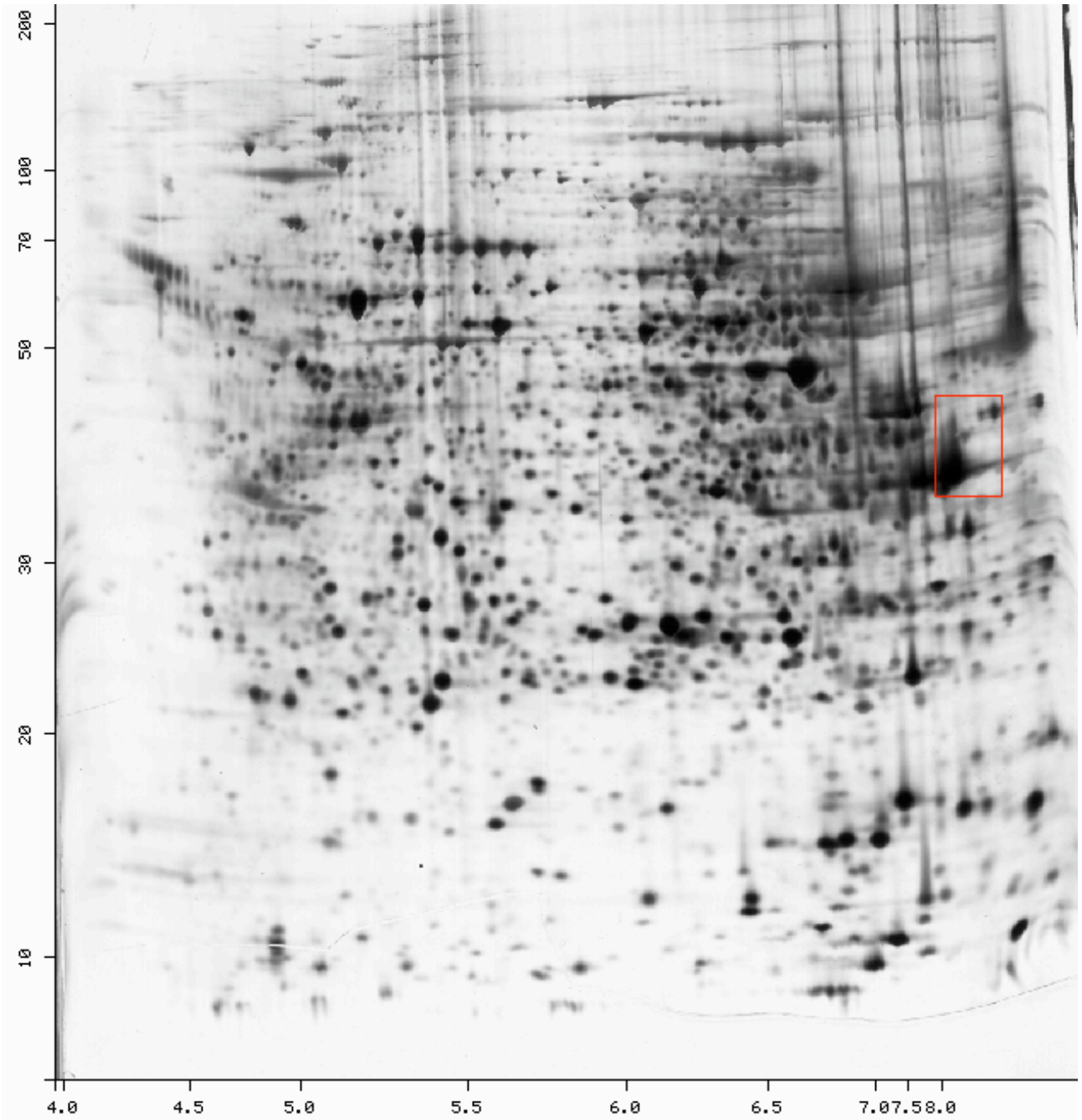
- Biomarker discovery for
  - early diagnostic
  - treatment response
- Identification of drug targets

Technologies

- 2d gel electrophoresis
- SELDI/MALDI mass spectrometry
- Protein Chips
  - the next big thing
  - analogous to mRNA chips
- Pre-fractionation
  - separate proteins by physical/chemical properties
  - high-throughput possible

2d Gel Electrophoresis

- Proteins in a sample are separated by isoelectric point and molecular mass within a polyacrylamide gel & stained $\Rightarrow$ protein expression map
- Technically known since the 70s, high-throughput in combination with
  - scanning of expression maps,
  - digital quantification of spot sizes,
  - protein identification for interesting spots (through MS)
- Resolution of several 1000 proteins per map
- Active development [8, 6]:
  - improved digital processing
  - combinations with pre-fractionation, mass spectrometry
  - two-color systems


200
100
70
50
30
20
10
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

Issues with 2D Gels

- Biological:
  - biased towards high-abundance proteins
  - no suitable for high mass, high hydrophobia, low copy number
- Technical:
  - poor linear response
  - spot recognition is hard
  - spot matching between gels is hard
- Practical: quantification requires expertise & time
- Statistical: little systematic treatment

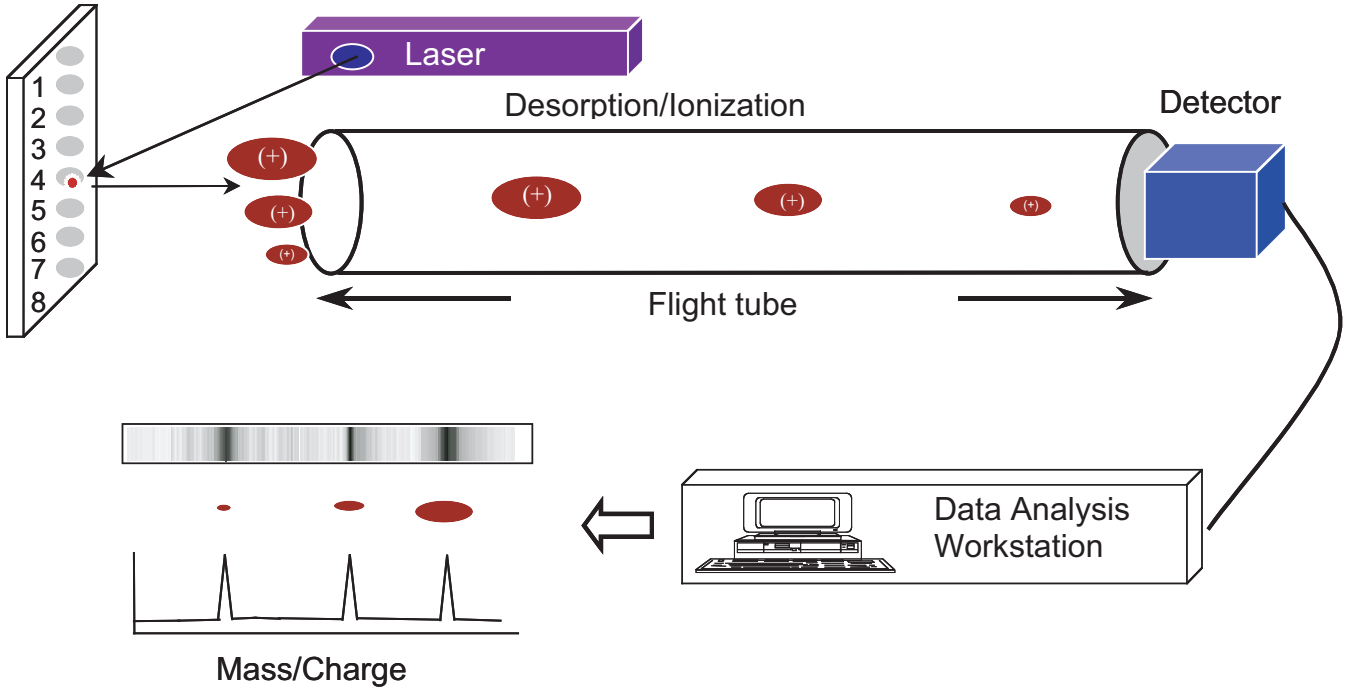
2D Gel Data

- Observations may be heavily censored and/or missing
- Often useful structure:
  - percentage of unobserved depends on expression level of observed
  - observations log nicely
  - mean/variance relationship
- However, mostly ad-hoc methods, e.g.:
  - Breast cancer samples with $< 10\%$ unobserved: impute by iteratively regressing on principal components
  - Colorectal cancer samples with 50% unobserved: use missingness to seperate sporadic and familial cases

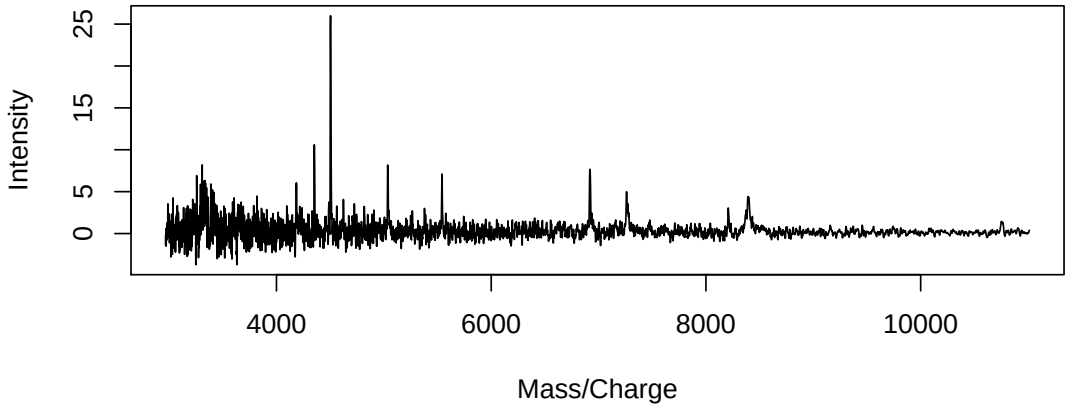
Mass Spectrometry

- MALDI-TOF MS: Matrix-assisted laser desorption/ionisation-time of flight mass spectrometry
  - sample proteins are embedded in energy-absorbent matrix molecules
  - laser stimulates the matrix to transfer energy to proteins
    $\Rightarrow$ proteins get ionized
  - an electric field hurls the ions down the flight tube
  - a detector counts impacts in regular time intervals:
    * count $\approx$ number of proteins
    * time of flight $\approx$ mass/charge of proteins

- SELDI-TOF MS: surface-enhaced laser desorption/ionisation-time of flight mass spectrometry
  - uses specially treated protein chips that bind a specific class of protein


Laser
Desorption/Ionization
Detector
1
2
3
4
5
6
7
8
(+)
Flight tube
Data Analysis
Workstation
Mass/Charge

Source: [12]


Intensity
0
5
15
25
4000
6000
8000
10000
Mass/Charge

Workflow

Sample collection $\Rightarrow$ Run samples (1/many chips)

$\Downarrow$

Normalization/Calibration

$\Downarrow$

Peak selection $\Leftarrow$ Peak detection

$\Downarrow$

Protein identification

This is really about:

Statistical Challenges in Biomarker Discovery for Cancer Using Time-of-Flight Mass Spectrometry

Biomarkers

- Proteins that are *somehow* associated with incidence, treatment response, or outcome of the disease

- Current biomarkers:
  - e.g. PSA for prostate cancer
  - usually based on disease mechanism and/or antibody screening
  - weak (low sensitivity/specificity)

- In future hopefully:
  - systematic exploration of possible markers
  - reliable panels of markers

Biomarkers for Early Cancer Detection

- This is HARD when aimed at population screening:
  - detection before clinical onset
  - very high specificity required
  - fairly non-invasive method of sample collection
- Somewhat easier for screening high-risk groups
- Suitable biological material:
  - urine (bladder c)
  - nipple aspirate fluid (breast c)
  - **serum** – hopefully general
    $\Rightarrow$ Human Plasma Proteome Project (at HUPO)
- [4] suggest a five-step framework for systematic discovery

Proteomics, Ovarian Cancer, and A New Paradigm?

Petricoin et al., 2002 [7]:

- Used a genetic algorithm (combined peak selection/detection) to establish a predictive pattern of peaks for 50 cases and 50 controls
- Validation on 116 additional samples:
  - 100% sensitivity
  - 95% specificity
- '. . . a new diagnostic paradigm'

Not yet . . .

- Ad Petricoin et al.: [9, 1]
  - greatest difference between cases and controls outside of measurement range ($< 500kD$)
  - sample processing causes blatant changes
  - separating features not reproducible across experiments
  - suspect mass calibration
- A general review of SELDI profiling papers finds [3]
  - discrepancies between studies,
  - failure to reproduce known biomarkers,
  - re-discovery of old biomarkers discarded for lack of sensitivity and specificity

Workflow for Biomarkers/SELDI

(Hypothesis) $\Rightarrow$ Experimental Design $\Rightarrow$ Sample Collection

$\Downarrow$

Peak Detection $\Leftarrow$ Normalization $\Leftarrow$ Quality Control I

$\Downarrow$

Peak Selection $\Rightarrow$ (Cross-) Validation $\Rightarrow$ Protein Identification

$\Downarrow$

Quality Control II

Experimental Design

- Identification of potential confounders in sample collection and processing
- Protein chip blocking structure
- Technical replication
- Pre-fractionation

Normalization

- Remove techhnical variation unrelated to biological sample properties
- Something has to remain constant between samples
- Mostly ad-hoc:
  - total ion current,
  - area under curve,
  - baseline

Peak Detection

- aka feature selection, aka variable selection
- Mostly ad hoc:
  - spectrum-by-spectrum,
  - based on local extremes & smoothing
- Suggestion K. Coombes: smooth mean spectra via wavelet decomposition & do peak detection there
- Project at MEB:
  - split spectra into small windows (ca. 5 consecutive observations)
  - compute a smoothed F-statistic with spectrum as grouping factor for each window
  - identify discriminatory peaks via false discovery rate

Peak selection

- aka discrimination, aka biomarker (pattern) idenfication
- Often a univariate pre-selection step
- Conventional statistical methods have been employed:
  - linear discrimination,
  - quadratic disrimination,
  - logistic regression with variable selection
- Numerous machine learning methods have been employed:
  - classification trees,
  - neural networks,
  - support vector machines,

  - k nearest neighbors,
  - bagging & boosting of all of the above

- Interestingly enough:
  - conventional statistical tools usually do not do much worse than the machine learning approaches,
  - modern statistical tools adapted for microarray data have **not** been used (reduced rank discrimination, empirical Bayes methods)

Cross-Validation

- Classification on the original data is biased towards the original data *specifically* for high-dimensional data

- Validation: apply the prediction to a new data set, optimally.
  - split into test & trainings set,
  - fit the model to the trainings set,
  - predict the test set for an unbiased error estimate

- Less wasteful: average error rates over repeated random partitions (training/test) of the data (in extremis: leave-one-out)

- It still needs stressing [10]: any informative pre-selection of peaks must also be cross-validated.

- A useful baseline: crossvalidate for a randomly shuffled classification.
- All this comes much more natural to machine learning than to statistics

Prediction & Quality Control II

- After establishment of a panel of biomarkers, status of new samples will be predicted
- New samples need to be assessed whether they can be reasonably be assumed to come from the original population
- [2] suggest:
  - establish a reference sample and its properties
  - spot the reference sample along with the new samples
  - use principal components & Mahalanobis distance to identify aberrant chips

Microarrays and Proteomics

On our way to a full Omic Analysis?

(Genome, Proteome, Transcriptome, Metabolome . . .)

Depends:

- Only if we connect all available biological data can we hope to model & understand biological activity on the molecular level in a comprehensive way.
  $\Rightarrow$ Doh

- All omics data is exchangeable & can be fed into the same black box.
  $\Rightarrow$ I don't think so . . .

Some useful reading

- Useful review issues:
  - Proteomics: Nature Insight Overview, Vol. 422, 2003
  - Processing: Proteomics 3, 2003 (First Annual Proteomics Data Mining Conference)
- The ovarian cancer study with follow-up: [7, 9, 1, 3]
- Applications & comparisons: [12, 11, 10]
- Machine learning for statisticians: [5]

References

[1] K.A. Baggerly, J.S. Morris, and K.R. Coombes. Reproducibility of seldi-tof protein patterns in serum: comapring datasets from different experiments. *Bioinformatics*, 20(5):777–785, 2004.

[2] K.R. Coombes, H.A. Fritsche Jr., C. Clarke, J.-N. Chen, K.A. Baggerly, J.S. Morris, L.-C. Xiao, M.C. Hung, and H.M. Kuerer. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.*, 49(10):1615–1623, 2003.

[3] E.P. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Molecular & Cellular Proteomics*, 3:367–378, 2004.

[4] R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, and L. Hartwell. The case for early detection. *Nat. Rev. Cancer*, 3, 2003.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[6] T. Kondo, M. Seike, Y. Mori, K. Fuji, T. Yamada, and S. Hirohashi. Application of sensitive flourescent dyes in linkage of laser microdissection and two-dimensional gel electrophoresis as a cancer proteomic study tool. *Proteomics*, 3:1758–1766, 2003.

[7] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577, 2002.

[8] T. Rabilloud. Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains. *Proteomics*, 2:3–10, 2002.

[9] J.M. Sorace and M. Zhan. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bionformatics*, 4:24, 2003.

[10] M. Wagner, D.N. Naik, A. Pothen, S. Kasukurti, R.R. Devineni, B.-L. Adam, O.J. Semmes, and G.L. Wright Jr. Computational protein biomarker prediction: a case study for prostate cancer. *BMC Bionformatics*, 5(26), 2004.

[11] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparisons of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

[12] Y. Yasui, M. Pepe, M.L. Thompson, B.-L. Adam, G.L. Wright Jr., Y. Qu, J.D. Potter, M. Winget, M. Thornquist, and Z. Feng. A data-analytic strategy for protein biomarker discovery: profiling high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3):449–463, 2003.