Peak versus AUC to compare SELDI data

Sreelatha Meleth PhD

EDRN Conference

Seattle, WA 2004

In this presentation

- My experience with Proteomics in general SELDI in particular
- Rounding m_z values
- Rationale – AUC
- Three Comparisons – peak versus auc
- Potential uses for AUC
- Conclusions

Proteomics and Early Detection of Cancer

- 2D gels
  - Separation of proteins based on pI and molecular weight 2D = 2 dimensions
  - Advances both in 2D Gel engineering and Image Analysis Software making this valuable technology
  - Statistical issues with 2D–
    - experimental design issues –Sample size, replicates etc.
    - pre-processing & its effect on results of analysis
    - Optimal analysis techniques
- My opinion SELDI + 2D = quicker biomarker discovery

My Reality

- My unit is primarily service provider
- No graduate students, no post docs
- I do not have time to concentrate only on SELDI data and develop novel methods with new language etc. 6 – 10 mths down the road

My Imperative

- My imperative to develop reliable , good methods that can be implemented in SAS
- Yet I must provide investigators with result
- Decided to use known statistical methods tweaked to fit SELDI data better

My experience with SELDI

- Analyzed 4-5 small pilot study data sets
  - 20-30 samples
  - Started more or less blind –applied my experience with 2D data
    - Protocol used – comparison of total protein expression in two groups, normalization, two sample tests, PCA & Discriminant Analysis
  - Developed classifier, identified peaks, anxiously waited to see test data
  - None of the m_z values in training & test matched
    - Close and within error range
  - So developed a SAS program to correct m_z vals

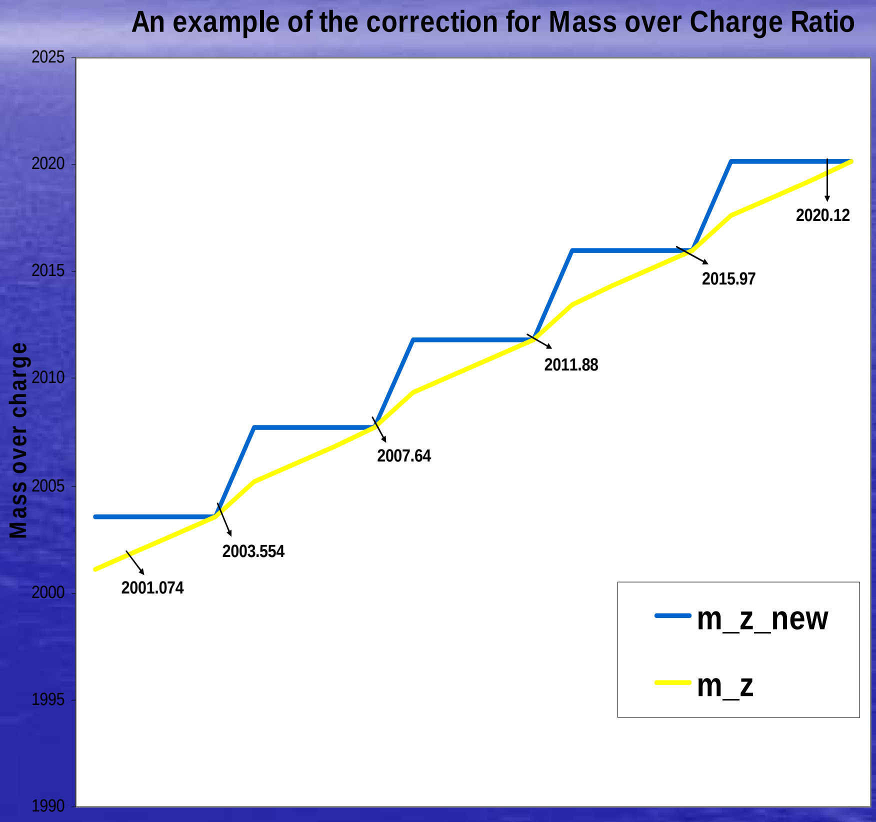
Rounding m_z’ s to reflect error 0.2%

m_z	rndrel	diff	tot	flag	index
2001	4	0	0	0	0
2002	4	0.83	0.83	0	0
2002	4	0.83	1.65	0	0
2003	4	0.83	2.48	0	0
2004	4	0.83	3.31	0	0
2005	4	0.83	0	1	1
2006	4	0.83	0.83	0	1
2007	4	0.83	1.66	0	1
2007	4	0.83	2.48	0	1
2008	4	0.83	3.31	0	1
2009	4	0.83	0	1	2

Rounding M_z/ Aligning Spectra

- Since SELDI Reliability = 0.2%
- E.G. , 2000 M-z might represent 1996 or 2004

  We aligned spectra such that SELDI values were rounded up to their maximum possible value



TOF Spectra – rationale for AUC

- Time of Flight Spectra – conversion of time of flight to molecular weights
- *Distribution* of ions around different Mol Wts
- Intuitively it seemed that area (total number of ions) represented a distribution better than the peak (maximum number of ions)
- Decided to examine classifiers using the two metrics

Estimating peaks (local maximums)

- Initially used the idea of maximum value in five / ten adjacent m_z values
- However, once I understood issue of reliability of the m_z values I use the following algorithm
    - Create the m_z_new variable as in previous slide
    - Estimate maximum values at each set of m_z values
    - These local maximums are used in classifier
    - Not strictly peaks, but maximum value at each ‘differentiable’ m_z

Estimate AUC

- Once again the set of m_z values that could represent the same molecular weight were used

- AUC is estimated using a trapezoidal rule

- AUC = (Maxm int + minm int)/ 2
  X
  (Maxm m_z interval – Minm m_z in interval)

Data sets Used

- Data Set 1 – Pilot data :
  - 21 normal serum , 21 HSIL serum
- Data set 2 - Pilot Data :
  - 8 patients with malignant diagnosis, 14 benign
  - Sample used pleural fluid
- Data Set 3 – EVMS prostrate data
  - 80 normal cases, 88 cancer

Building Classifier--1

- Step 1: Identify significantly different peaks / AUC
- Step 2: Used a cross validation type process in Step 2 (Robert Tibshirani – 2003 ASA Meeting SF)
  - In data sets 1 and 2 used a leave one out in disease (normal) using a random process
  - For EVMS data randomly selected 40 cancer and 40 normals
- Step 3: Stepwise Discriminant analysis used to identify potential variables to build classifier – list is stored

Building Classifier - 2

- Step 4: Repeated 500 times DS1, 10000 DS2, 5000 DS

- Step 5: The most frequently occurring m_z's are used in the final discriminant analysis
- Quadratic / linear depending on test of equal covariance matrix

- Data set 1 & 2 –pilot data used only cross validation, EVMS data – used test set to measure quality

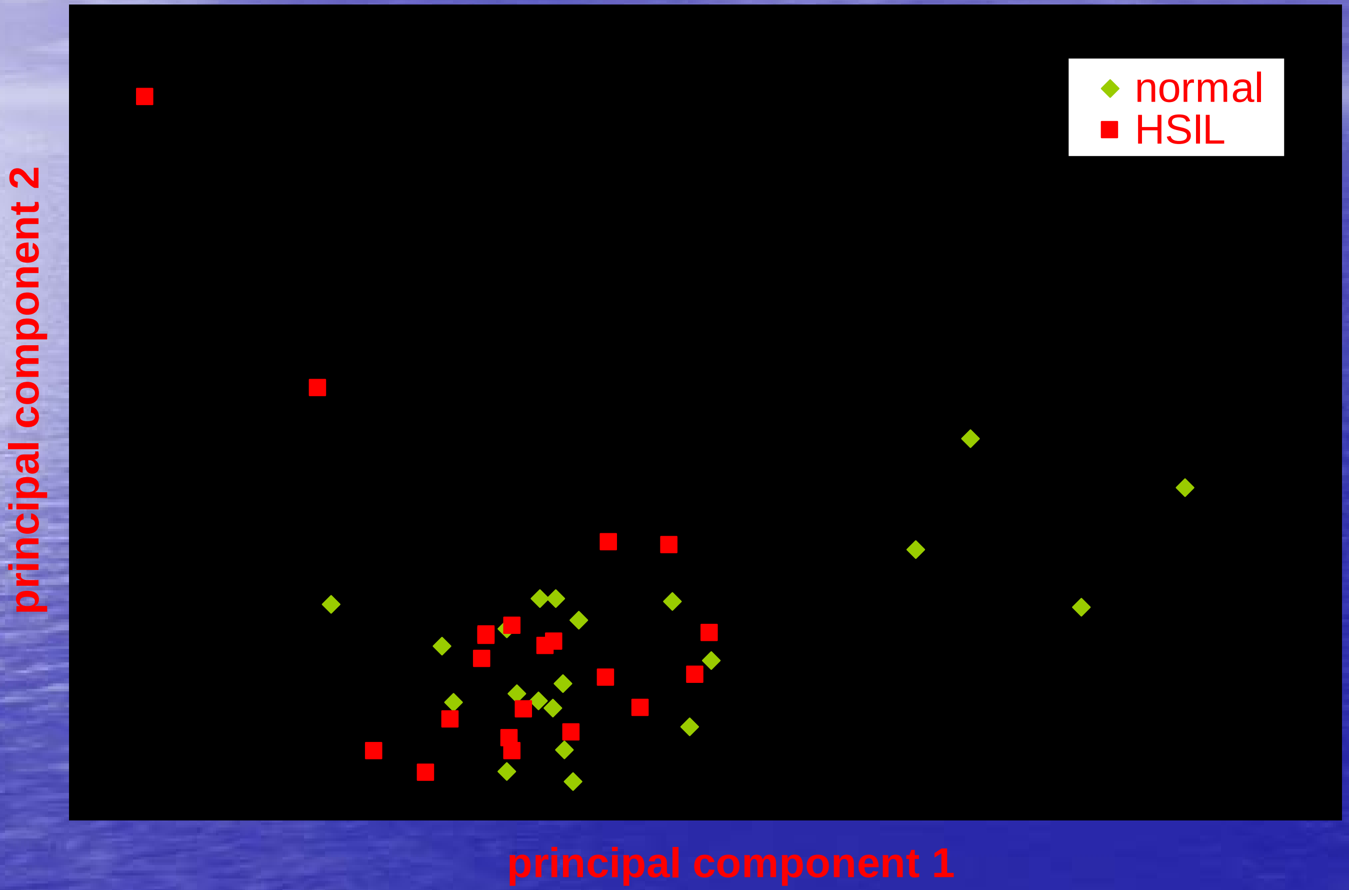
- In DS3 the random training sets chosen before 2 sample tests

Results – Normal versus HSIL - PEAKS

- Total protein expression in two groups – not significantly different p = 0.77
- 13 peaks were significantly different at p=0.05
- Quadratic Discrim Analysis – 6 Peaks (homogeneity test p =0.0001)

  Specificity =76%, Sensitivity=67%
- Caveats:
  - Based on cross validation .
  - Data set too small for test set

PCA HSIL versus NORMAL Peaks

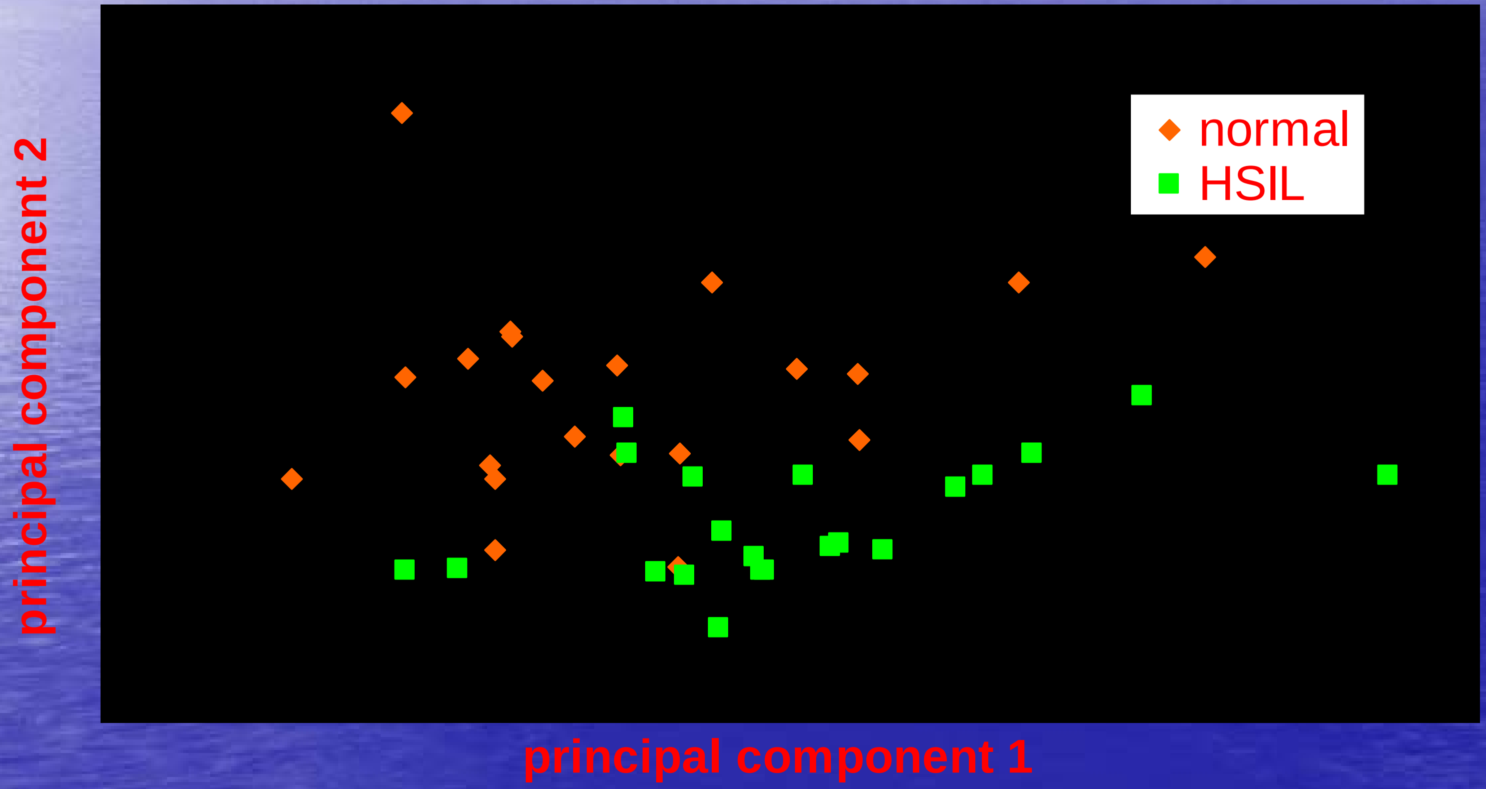


Results – Normal versus HSIL - AUC

- 33 AUC were significantly different at p = 0.05
- Quadratic Discrim Analysis – AUC (homogeneity test p =0.03) – 6 aucs

  Specificity =100%, Sensitivity=67%
- Caveats:
  - Based on cross validation .
  - Data set too small for test set

PCA HSIL versus Normal AUC



Results –Pleural Fluid Ca vs benign Peaks

- Total protein expression cancer significantly higher than benign p = 0.0044
- 84 m_z values significant at p=0.0002
- Quadratic Discrim Analysis – AUC (homogeneity test p =0.0001) – 4 peaks

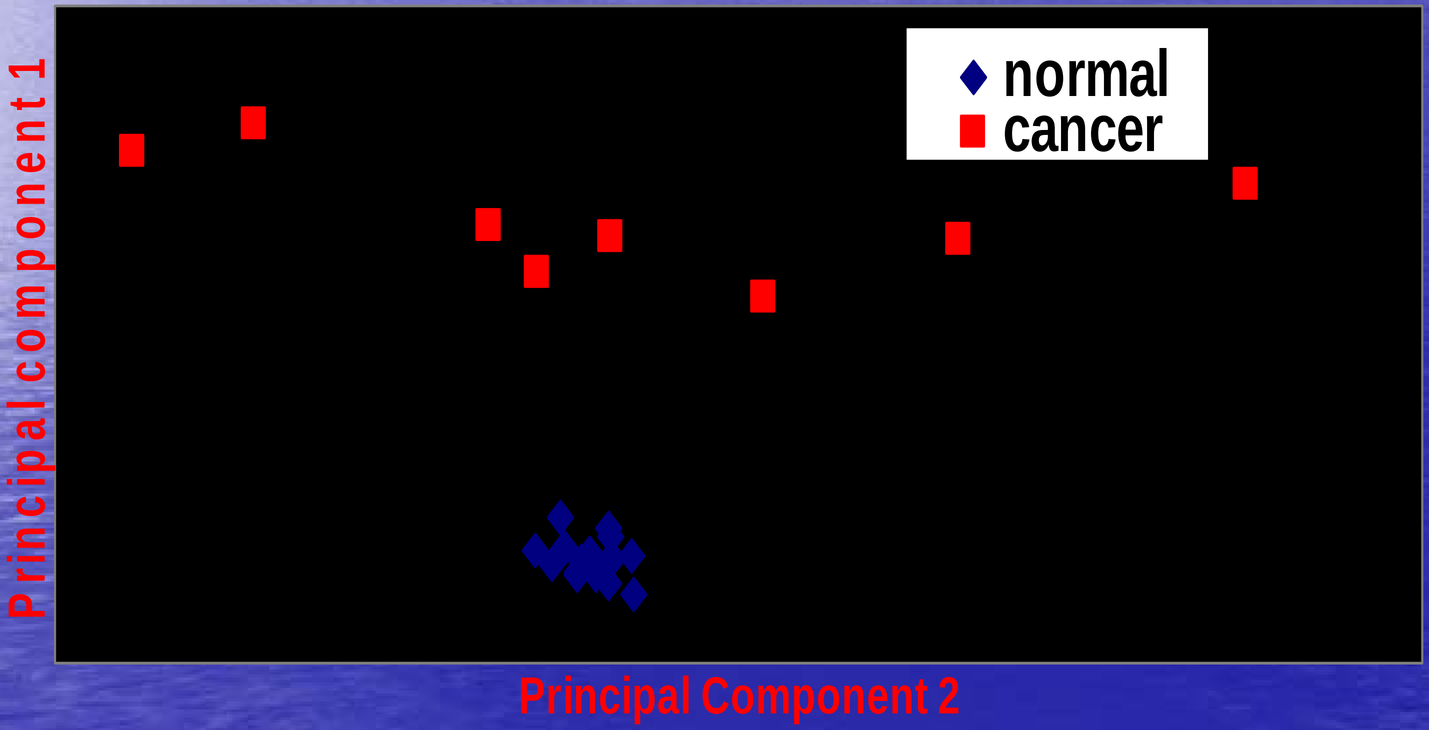
  Specificity =100%, Sensitivity=62.5%
- Caveats:
  - Based on cross validation .
  - Data set too small for test set

Results – Body Cavity Fluid Mets versus none - AUC

- 39 AUC were significantly different at p = 0.0002
- Quadratic Discrim Analysis – AUC (homogeneity test p =0.0001) – 5 aucs

  Specificity =100%, Sensitivity=100%
- Caveats:
  - Based on cross validation .
  - Data set too small for test set

PCA – Mets versus none - Peaks



Results – EVMS Ca versus Normal Peaks

- Total protein expression cancer significantly higher than benign p = 0.0044
- 220 m_z values significant at p=0.0001
- Quadratic Discrim Analysis – AUC (homogeneity test p =0.0001) – 7 peaks

  Specificity =90%, Sensitivity=95%
- PCA – good separation
  - Based on test set.

Results – EVMS Ca versus Normal AUC

- 220 m_z values significant at p=0.0001
- Quadratic Discrim Analysis – AUC (homogeneity test p =0.0001) -7 aucs

  Specificity =90%, Sensitivity=85%
- PCA separates well
  - Based on test set.

Conclusions

- It is possible to use ‘everyday’ regular SAS programs to develop reasonable classifiers
- Different data sets may require different metrics to get optimal classifier
- Too early to confirm but these analyses suggest that for data sets with smaller differences AUC might be a more sensitive feature