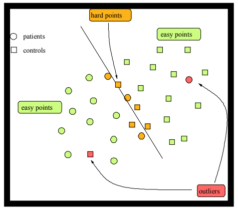
Proteome profiling without selection bias

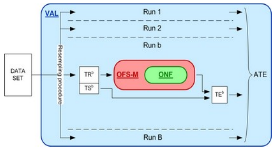
Annalisa Barla, Bettina Irler, Stefano Merler,
Giuseppe Jurman, Silvano Paoli, Cesare Furlanello

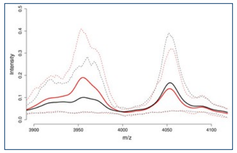
ITC-irst, Trento, Italy

Topics

- **Biomarkers and Predictive classification**
  - Prediction and Bias
  - Biomarker Ranking algorithms with Support Vector Machines (kernel methods)
  - The Complete Validation Platform (BioDCV)
  - Pipeline of Preprocessing Procedures
  - Grid application (EGEE-Biomed VO)
- **Experiments**
  - Cromwell MALDI-TOF simulated data
  - SELDI-TOF Ovarian cancer (NCIFDAProteomics)
  - MALDI-TOF Ovarian cancer (Keck Labs)





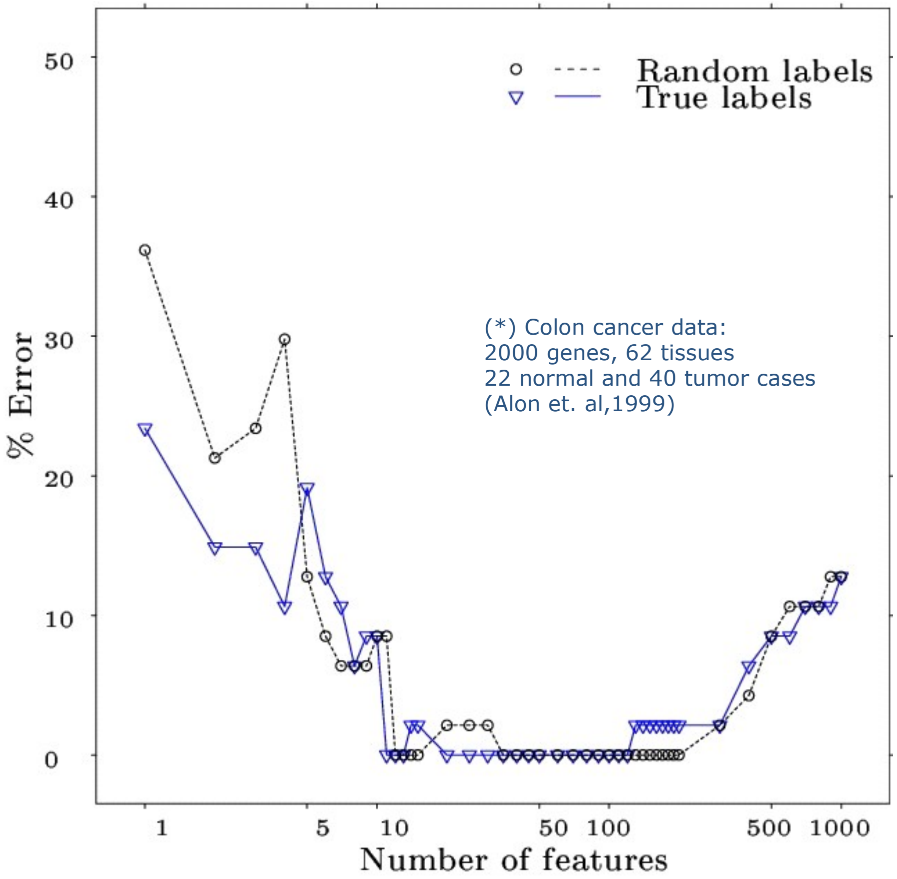


Prediction and Bias

- **Bias: on data preparation, preprocessing (complex!), classification**
  - E Petricoin, A Ardekani, B Hitt, P Levine, B Fusaro, S Steinberg, G Mills, C Simone, D Fishman, E Kohn, and L Liotta. Use of proteomic patterns in serum to identify ovarian cancer. L*ancet*, 359:572-577, 2002.
  - K Baggerly, J Morris, and K Coombes. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777-785, 2004.
- **Controversy: *J Natl Cancer Inst* 2005; 97**
  - K Baggerly, JS Morris, SR Edmonson, KR Coombes. Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer
  - LA Liotta, M Lowenthal, A Mehta, TP Conrads, TD Veenstra, DA Fishman, EFIII Petricoin. Importance of Communication Between Producers and Consumers of Publicly Available Experimental Data
  - DF Ransohoff. Lessons from Controversy: Ovarian Cancer Screening and Serum Proteomics
- DF. Ransohoff. Bias as a threat to the validity of cancer molecular-marker research. *Nature*, 5:142-149, 2005.

The selection bias problem

Pervasive in the first years of microarray classification studies: use CV to evaluate models, pick up best probes, compute again expected error with CV …



(*) Colon cancer data:
2000 genes, 62 tissues
22 normal and 40 tumor cases
(Alon et. al,1999)

METHODOLOGY
- Ambroise & McLachlan, 2002
- Simon et. al 2003
- Furlanello et. al 2003

A zero error (CV) may be obtained with only 8 genes (*).

But when repeating the experiment after a label randomization, a very similar result is reached: 14 genes are sufficient to get a zero error estimate.

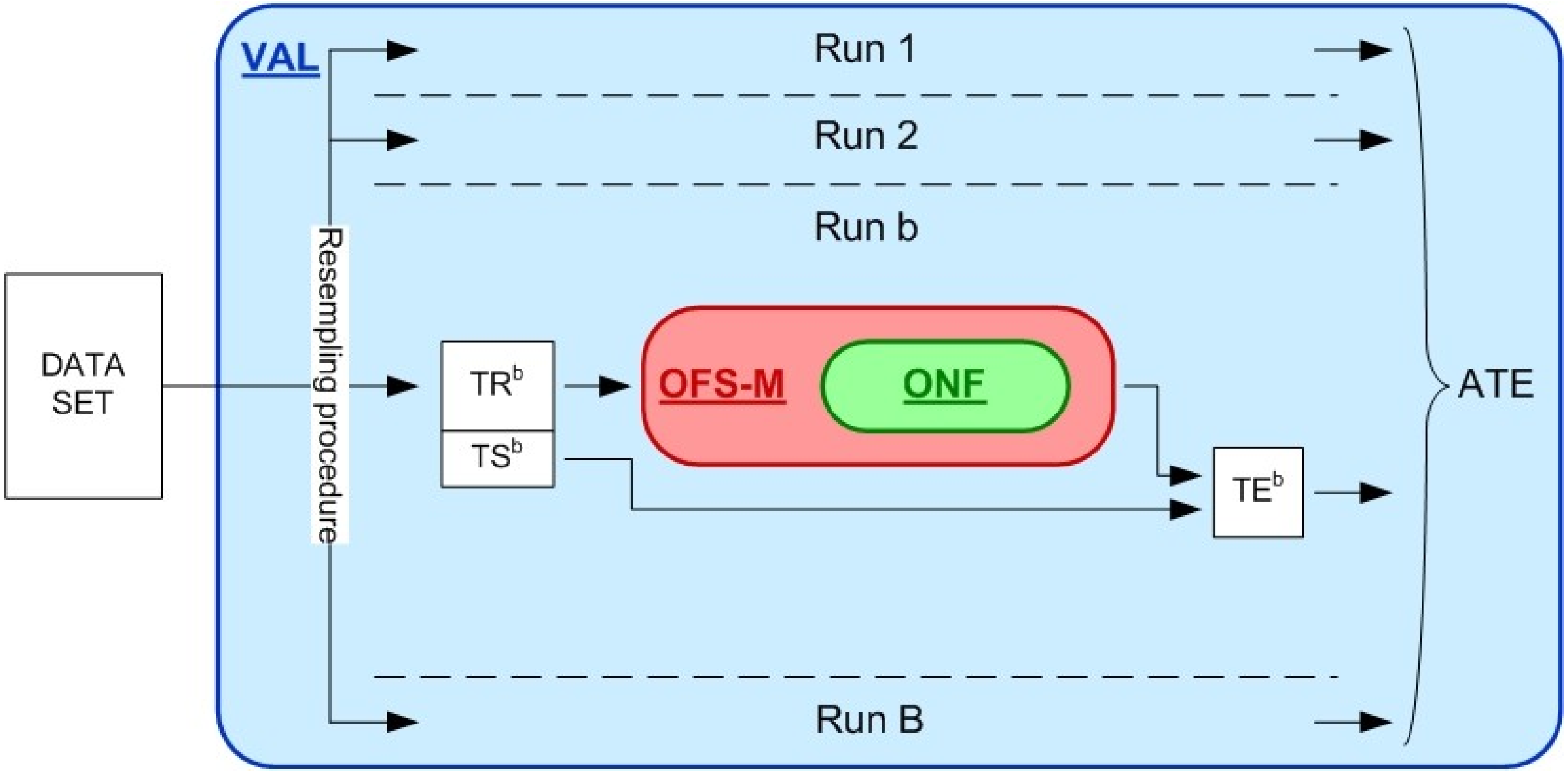
The same effect can be reproduced with no-information datasets !!

(*): similar results of near perfect classification with few genes published in *PNAS*, *Machine Learning*, *Genome Research*, *BMC Bioinformatics*, etc.

COMPLETE VALIDATION

To avoid selection bias (p>>n): *

- externally a stratified random partitioning,
- internally a model selection based on a K-fold cross-validation
- high computational costs due to replicates (10^5 --10^6 models)**



* Ambroise & McLachlan, 2002,
Simon et. al 2003,
Furlanello et. al 2003

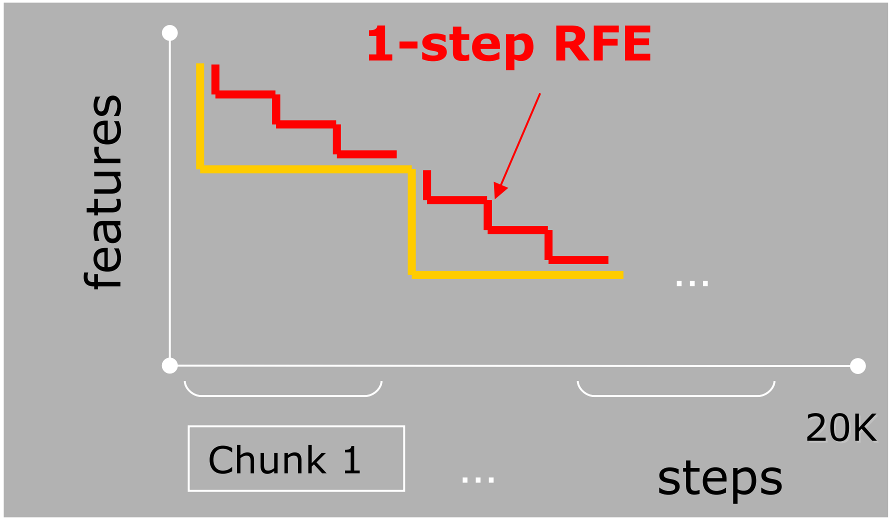
OFS-M: Model tuning and Feature ranking

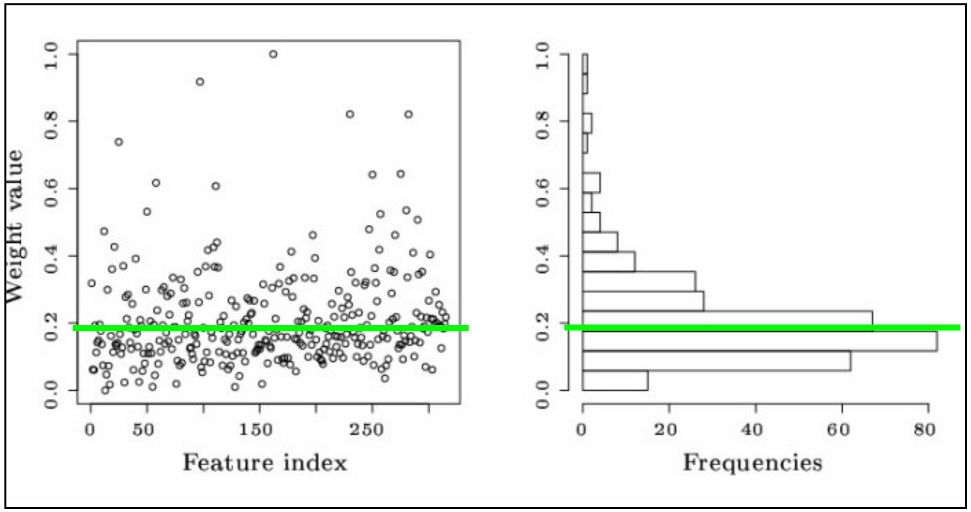
ONF: Optimal gene panel estimator

ATE: Average Test Error

** Binary classification, on a 20 000 genes x 45 cDNA array, 400 loops

E-RFE with SVM Classifiers





- **MODEL: Support Vector Machines (SVM)**

- **RANKING → SELECTION Recursive Feature Elimination (RFE): a stepwise backward selection procedure.**

  At each step, eliminate the "least interesting variable" and retrain

ACCELERATIONS

- **Parametrics**
  - **Sqrt–RFE**
  - **Decimation-RFE**

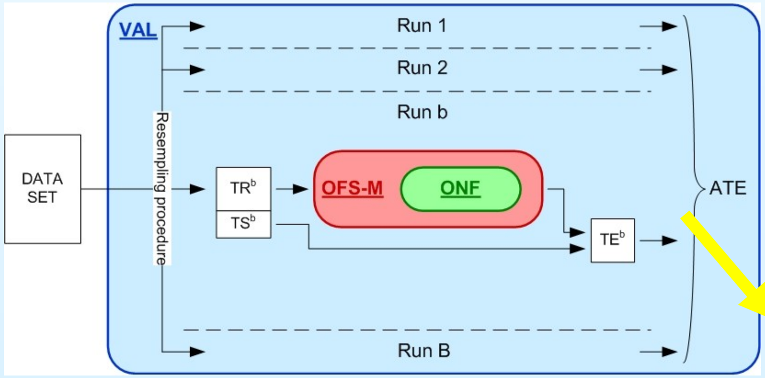
- **Non-Parametrics**
  - **E–RFE: adapting to weight distribution by thresholding the SVM weights at w***

ITC irst
CENTRO PER LA RICERCA
SCIENTIFICA E TECNOLOGICA

High Throughput Computing

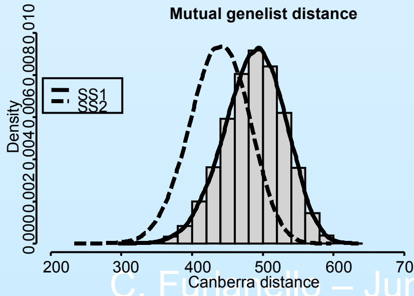
Complete validation on cluster/ Grid

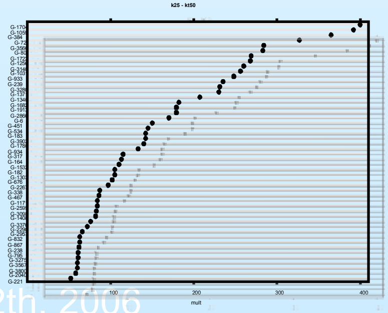
1. **COMPLETE VALIDATION CURES THE SELECTION BIAS**
2. **Computational solutions: Clusters and GRID**
3. **The by-products of complete validation**



Accuracy estimates

Set of lists:
- aggregation
- stability





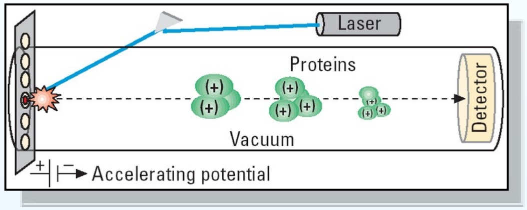
BIODCV
http://biodcv.itc.it
for National Institute of Cancer

PREVIOUS WORK ON MICROARRAY DATA

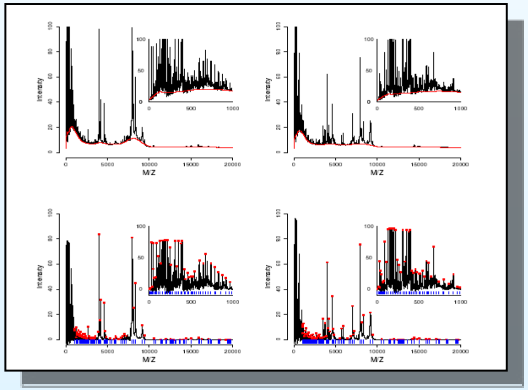
Neural Networks
BMC Bioinformatics
IEEE Trans. Signal Processing
IEEE Trans. Comp. Biology and Bioinformatics
Int. J. of Cancer

Harnessing Bias

Data generation



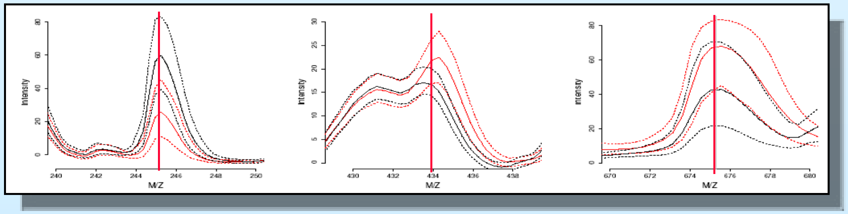
Preprocessing



Complete Validation

BioDCV
biodata distributed complete validation

Biomarkers identification



- **integrate a pipeline for proteomic data preprocessing with the BioDCV complete validation process**

Preprocessing

Single spectrum	***Batch***
Baseline subtraction [1]	
	Normalization (A.U.C.)
Peak Extraction [2]	
	Centroid Identification [2]
	Peak Assignment [2]
	(ms Standardization)

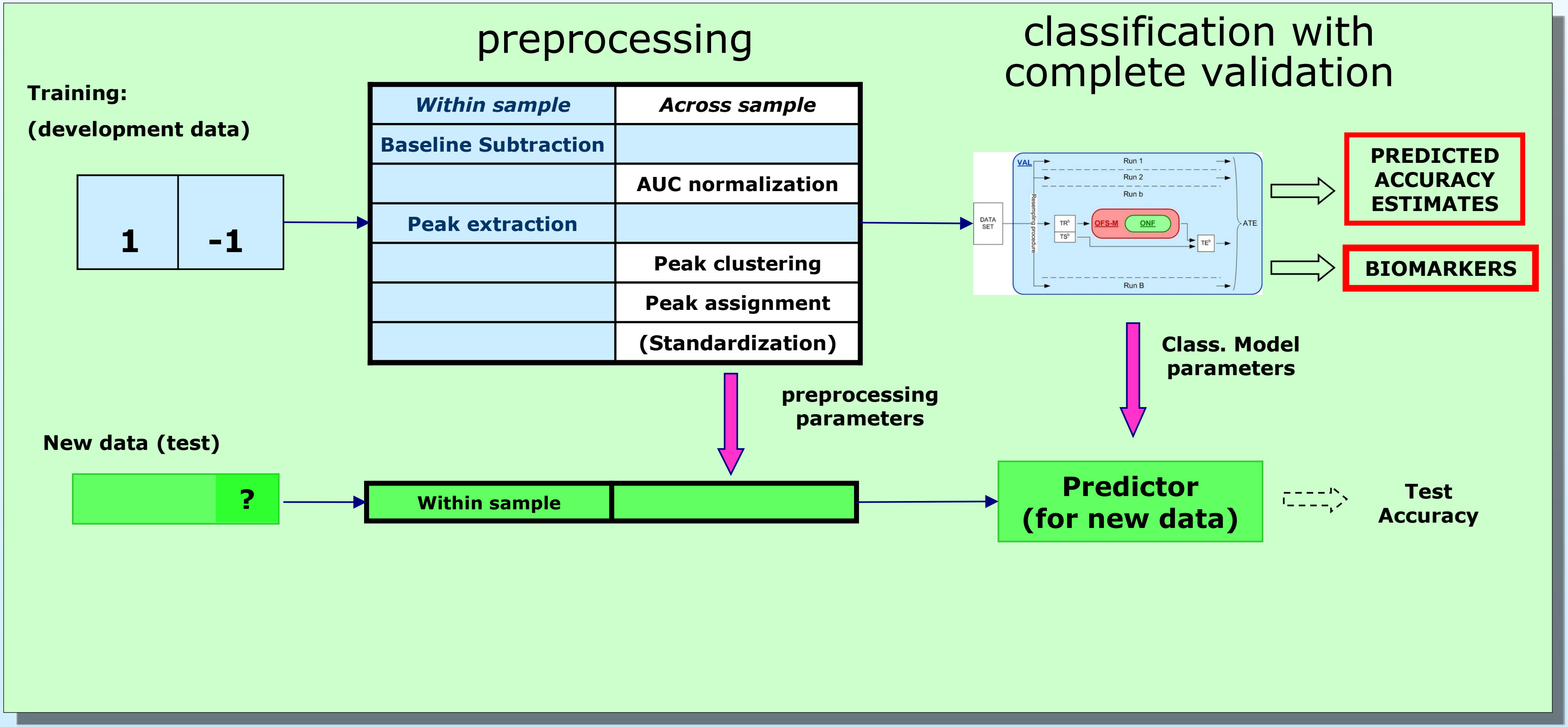
[1] **PROcess: an R package -- lowess for baseline subtraction**
X. Li
A package for processing protein mass spectrometry data.
http://www.bioconductor.org/packages/bioc/1.8/html/PROcess.html

[2] **ppc: another R package – features defined by cluster centroids' location**
R. Tibshirani et al.
Sample classification from protein mass spectrometry, by "peak probability contrasts"
Bioinformatics 20(17):3034-3044, 2004

Not discussed here: calibration, filtering

Complete Validation for Proteomics

- GIVEN mz-ms data: spectra in a standardized mass spectrometry format; a binary label for each spectrum (e.g. +1/-1)
- FIND: Biomarkers valid on novel data & classification error estimates



preprocessing

classification with complete validation

Training:

(development data)

Within sample	Across sample
Baseline Subtraction	
	AUC normalization
Peak extraction	
	Peak clustering
	Peak assignment
	(Standardization)

PREDICTED ACCURACY ESTIMATES

BIOMARKERS

Class. Model parameters

preprocessing parameters

New data (test)

Within sample

Predictor (for new data)

Test Accuracy

Methodology

Goal: study biomarker selection by complete validation setup

PROTEOMIC DATA

TR1 / TS1

TR2 / TS2

⋮

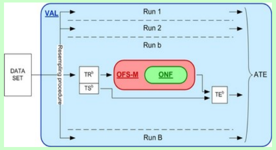
⋮

TR5 / TS5

preprocessing

Within sample	Across sample
Within sample	

classification with complete validation



Predicted Accuracy Estimates

BIOMARKERS

Test Accuracy

Predicted Accuracy Estimates

BIOMARKERS

Test Accuracy

⋮

⋮

Predicted Accuracy Estimates

BIOMARKERS

Test Accuracy

- H=5 splits (80%-20% or 50%-50%)
- B=400 replicates
- Preprocessing
  - AP0: non-batch
  - AP1: batch
- Random labels exps.

Datasets

- ***Simulated MALDI-TOF data (Cromwell's): 4 datasets at increasing levels of noise: $\varepsilon=N(0, \sigma)$ $\sigma=(0,10, 50, 300)$***

tot #	*class 1*	*class -1*	*#m/z (100Da < m/z < 20000Da)*
160	80	80	**17669**

- ***Ovarian 8/7/02 (SELDI-TOF)******

tot #	cancer	control	#m/z (0Da < m/z < 20000Da)
253	162	91	**15153**

http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

- ***Ovarian '05 (MALDI - TOF)***

tot #	cancer	control	#m/z reflectron	**#m/z linear (3450Da < m/z < 28000Da)**
170	93	77	94780	**36890**

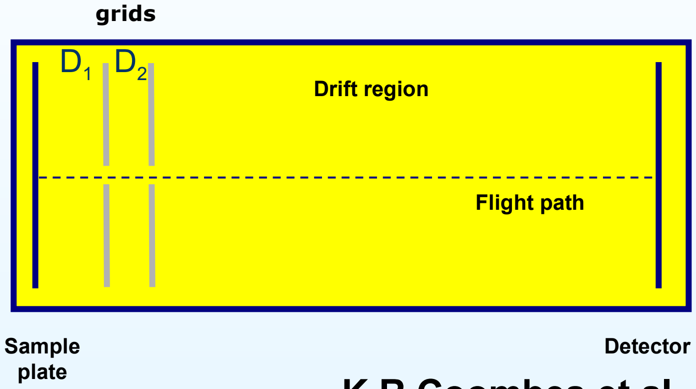
Nat. Ovarian Cancer Early Detection Program Northwestern Univ. Hospital
Micromass M@LDI-L/R , Keck Lab Yale (Wu et al 2005)

http://bioinformatics.med.yale.edu/MSDATA2/

*** Technical and experimental design of this dataset were questioned.**

Simulated Data (I)

Cromwell: a proteomic MALDI-TOF simulation engine



K.R.Coombes et al.
Understanding the characteristics
of mass spectrometry data
through the use of simulation
Cancer Informatics 2005:1(1) 41-52

Configuration

A. Default parameters:
- Voltage between plates (20 KV)
- Length of drift tube (L=1 m)
- Distance between charged grids (8 mm)
- Standard deviation on initial particles' velocity (50)

B. Defined Parameters:
- Peak sites (chosen from a real dataset)
- Peak intensity (max no. of a set of particles)
- Standard deviation on noise over intensity

Software: v 2.0 in R, from S-Plus code
http://bioinformatics.mdanderson.org/cromwell.html

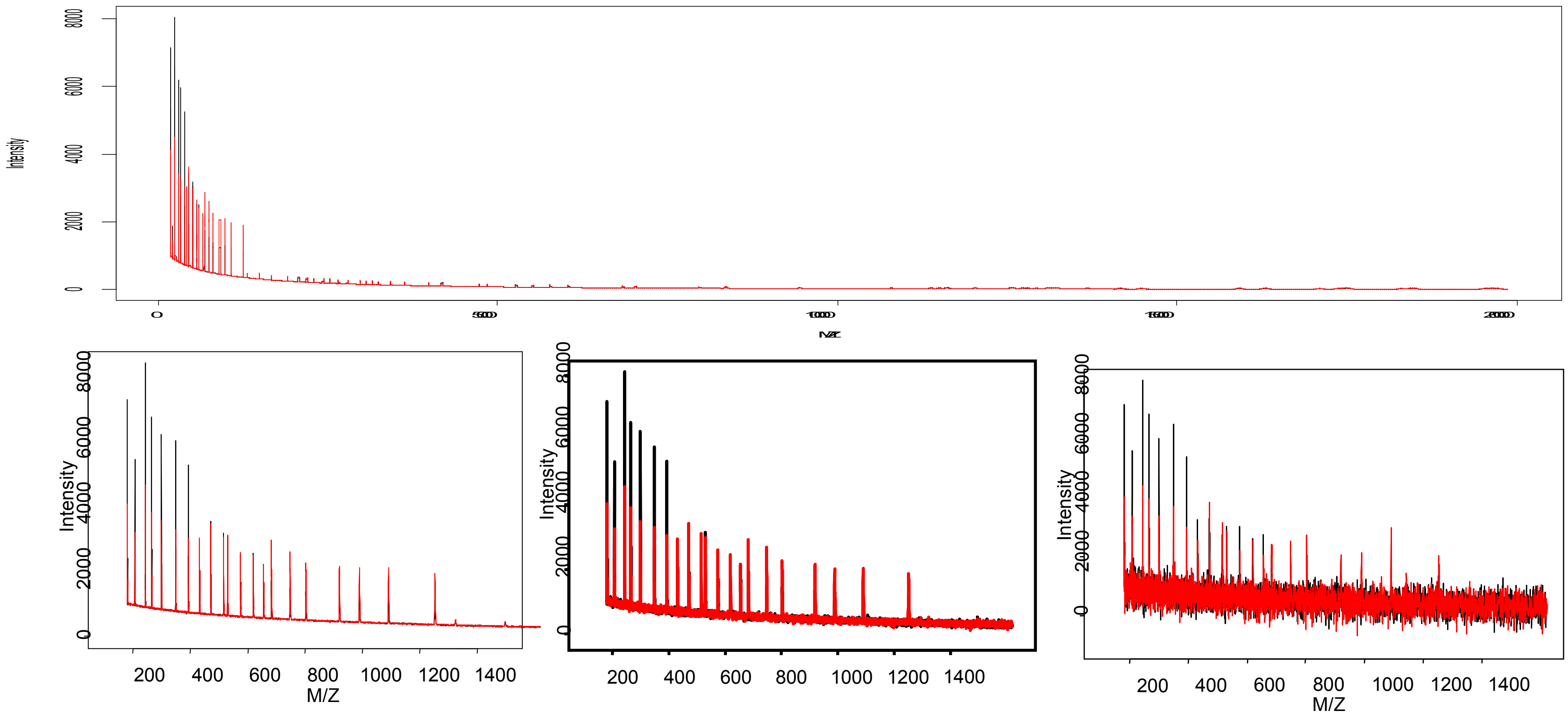
Hypotheses

- Different peak intensity at a panel of m/z locations discriminates the two classes
- A "band" structure

Simulated Data (II)

class	Peak Intensity [**Number of Peaks**]			
	B1	B2	B3	B4
1	10000 [**7**]	7000 [**7**]	5000 [**7**]	1000 [**60**]
-1	5000 [**7**]	7000 [**7**]	10000 [**7**]	1000 [**60**]

Design: the 2 classes are discriminated by peak intensities in bands B1 and B3, but no discriminations in B2 and B4



ITC irst
CENTRO PER LA RICERCA SCIENTIFICA E TECNOLOGICA

Results: Simulated Data

Note: the 4 synthetic MALDI-TOF datasets were built each with a total of 14 discriminant peaks, but our preprocessing procedure detected only 13 of them since the first one is located too close to the inf of spectrum border.

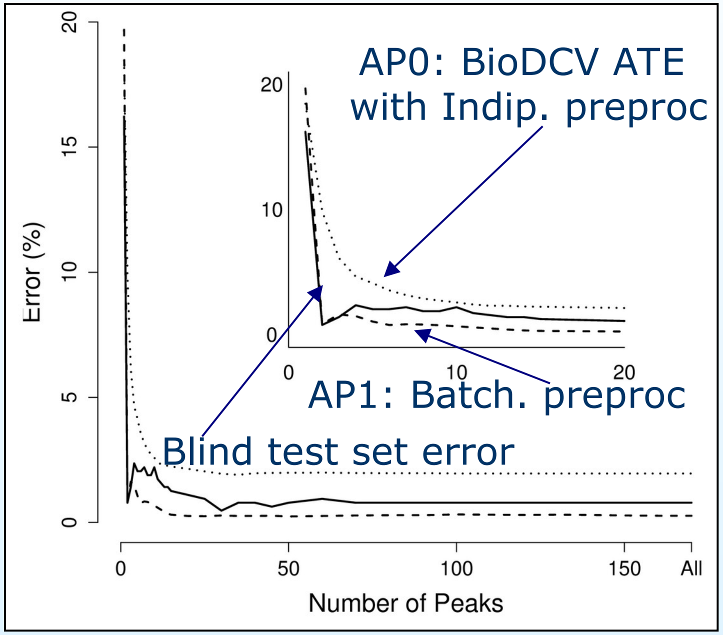
PREPROCESSING PIPELINE RESULTS

σ=0	81 peaks detected
σ=10	
σ=50	- Two extra non-valid peaks identified in the preprocessing phase (due to noise) - After the BioDCV procedure one was rejected
σ=300	

COMPLETE VALIDATION RESULTS

- **The actual 13 discriminant peaks were found among the most significant features extracted**
- **A list stability indicator showed that the number of relevant variables over all run is exactly 13**

Results: SELDI-TOF Ovarian 8/7/02



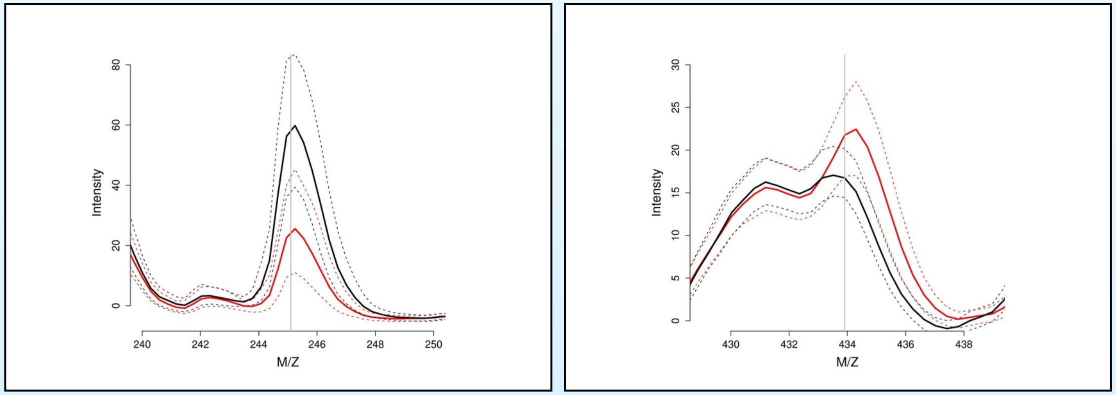
AP0: BioDCV ATE with Indip. preproc

AP1: Batch. preproc

Blind test set error

AVG Error on blind test set (5 features): ~3%
Random labels:
ATE on blind test=41.1% (CI 34.6, 56.8)
No Info = 36%

Biomarker analysis



The first and the second most relevant peaks for BioDCV classification in all the sublists of the dataset confirm previous studies (and their concerns)

K. Baggerly et al.
Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.
Bioinformatics, 20(5):777-785,2004.

W. Zhu et al.
Detection of cancer-specific markers amid massive mass spectral data.
PNAS 100(25):14666-14671, December 2003.

ITC irst
CENTRO PER LA RICERCA SCIENTIFICA E TECNOLOGICA

Results: MALDI-TOF Keck Lab

AVG Error (AP0 mode) test set (14 features): 32.5% (CI 32.1,32.7)
AVG Error (AP0 mode) test set (all features): 24.5%

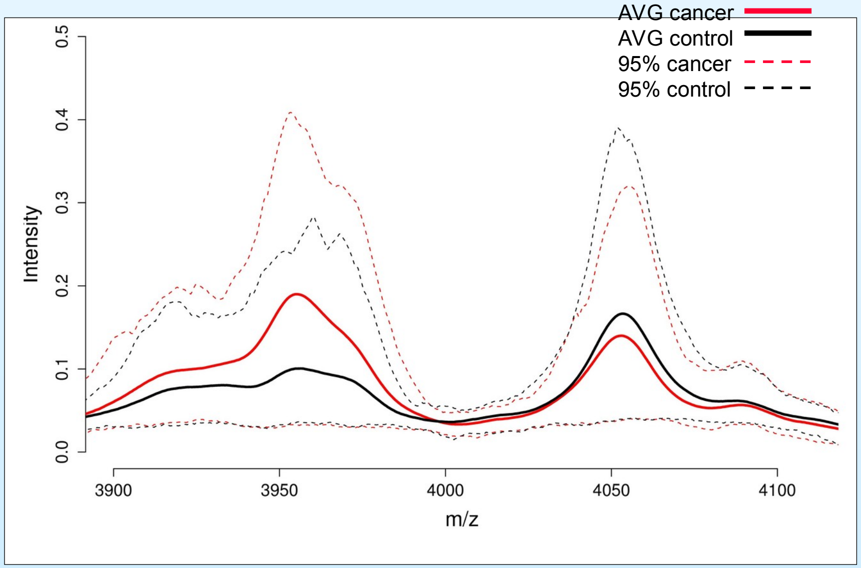
AVG Error (AP1 mode) test set (14 features): 25.7%

Random labels: ATE on blind test=49.1%
No Info=45.3%

Results compliant with:

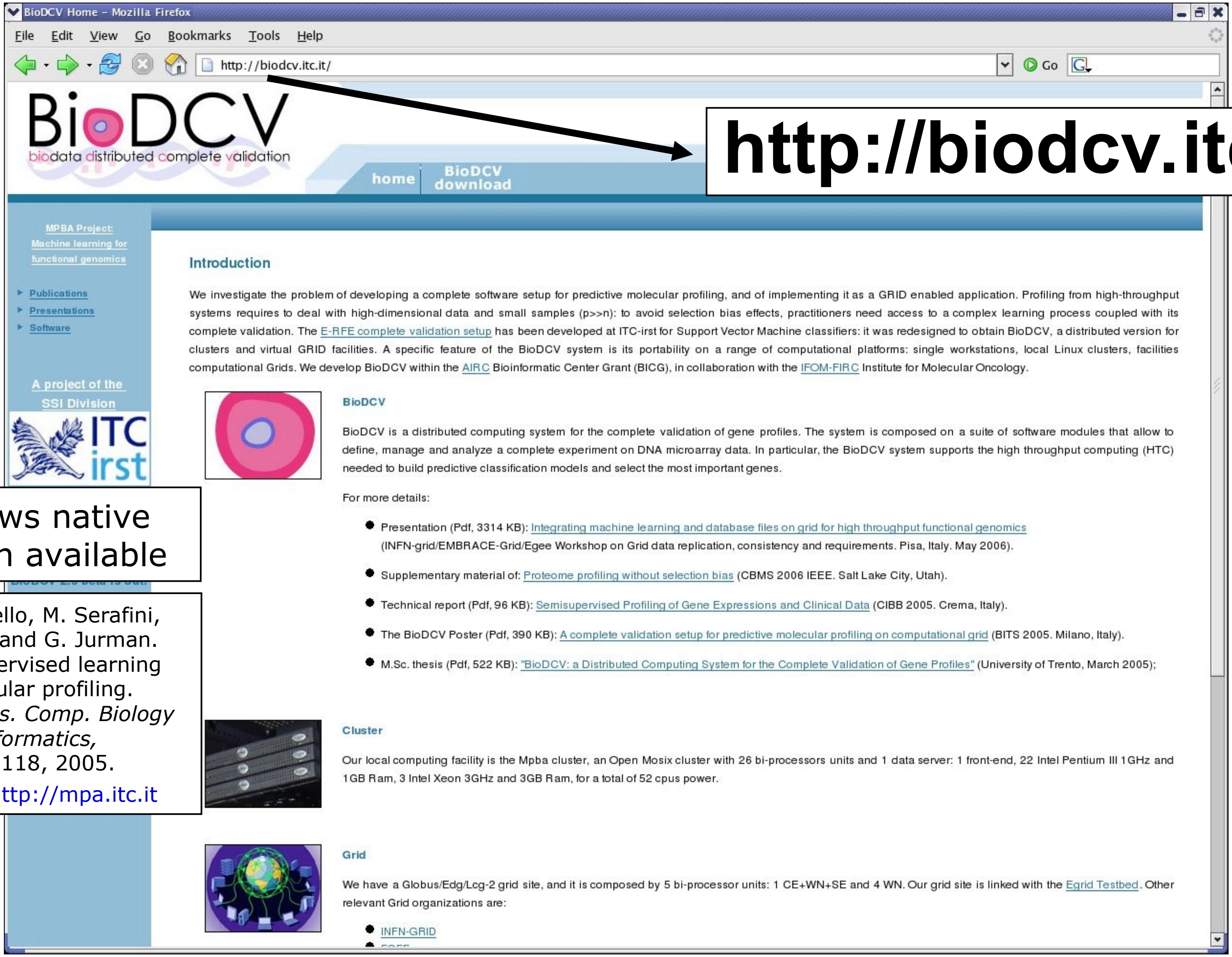
Baolin Wu et al.
Ovarian cancer classification based on mass spectrometry analysis of sera.
Cancer Informatics, 2005.

Biomarker analysis:



The first and the fifth most relevant peaks in the Keck Lab dataset

BioDCV SubVersion Repository


BioDCV Home - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://biodcv.itc.it/
Go
BioDCV
biodata distributed complete validation
home
BioDCV download
MPBA Project: Machine learning for functional genomics
Publications
Presentations
Software
A project of the SSI Division
ITC irst
BioDCV 2.0 beta is out!

Introduction

We investigate the problem of developing a complete software setup for predictive molecular profiling, and of implementing it as a GRID enabled application. Profiling from high-throughput systems requires to deal with high-dimensional data and small samples (p>>n): to avoid selection bias effects, practitioners need access to a complex learning process coupled with its complete validation. The E-RFE complete validation setup has been developed at ITC-irst for Support Vector Machine classifiers: it was redesigned to obtain BioDCV, a distributed version for clusters and virtual GRID facilities. A specific feature of the BioDCV system is its portability on a range of computational platforms: single workstations, local Linux clusters, facilities computational Grids. We develop BioDCV within the AIRC Bioinformatic Center Grant (BICG), in collaboration with the IFOM-FIRC Institute for Molecular Oncology.

BioDCV

BioDCV is a distributed computing system for the complete validation of gene profiles. The system is composed on a suite of software modules that allow to define, manage and analyze a complete experiment on DNA microarray data. In particular, the BioDCV system supports the high throughput computing (HTC) needed to build predictive classification models and select the most important genes.

For more details:

- Presentation (Pdf, 3314 KB): Integrating machine learning and database files on grid for high throughput functional genomics (INFN-grid/EMBRACE-Grid/Egee Workshop on Grid data replication, consistency and requirements. Pisa, Italy. May 2006).
- Supplementary material of: Proteome profiling without selection bias (CBMS 2006 IEEE. Salt Lake City, Utah).
- Technical report (Pdf, 96 KB): Semisupervised Profiling of Gene Expressions and Clinical Data (CIBB 2005. Crema, Italy).
- The BioDCV Poster (Pdf, 390 KB): A complete validation setup for predictive molecular profiling on computational grid (BITS 2005. Milano, Italy).
- M.Sc. thesis (Pdf, 522 KB): "BioDCV: a Distributed Computing System for the Complete Validation of Gene Profiles" (University of Trento, March 2005);

Cluster

Our local computing facility is the Mpba cluster, an Open Mosix cluster with 26 bi-processors units and 1 data server: 1 front-end, 22 Intel Pentium III 1GHz and 1GB Ram, 3 Intel Xeon 3GHz and 3GB Ram, for a total of 52 cpus power.

Grid

We have a Globus/Edg/Lcg-2 grid site, and it is composed by 5 bi-processor units: 1 CE+WN+SE and 4 WN. Our grid site is linked with the Egrid Testbed. Other relevant Grid organizations are:

- INFN-GRID

http://biodcv.itc.it

- Windows native version available

- C. Furlanello, M. Serafini, S. Merler and G. Jurman. Semi-supervised learning for molecular profiling. *IEEE Trans. Comp. Biology and Bioinformatics,* 2(2):110-118, 2005.
- More on http://mpa.itc.it

Grid Computing for Proteomics

1. IEEE CBMS 2006: series of experiments on proteomics data
    - standard complete validation analysis
    - random labels analysis
2. A strict deadline for the final version

Solution:

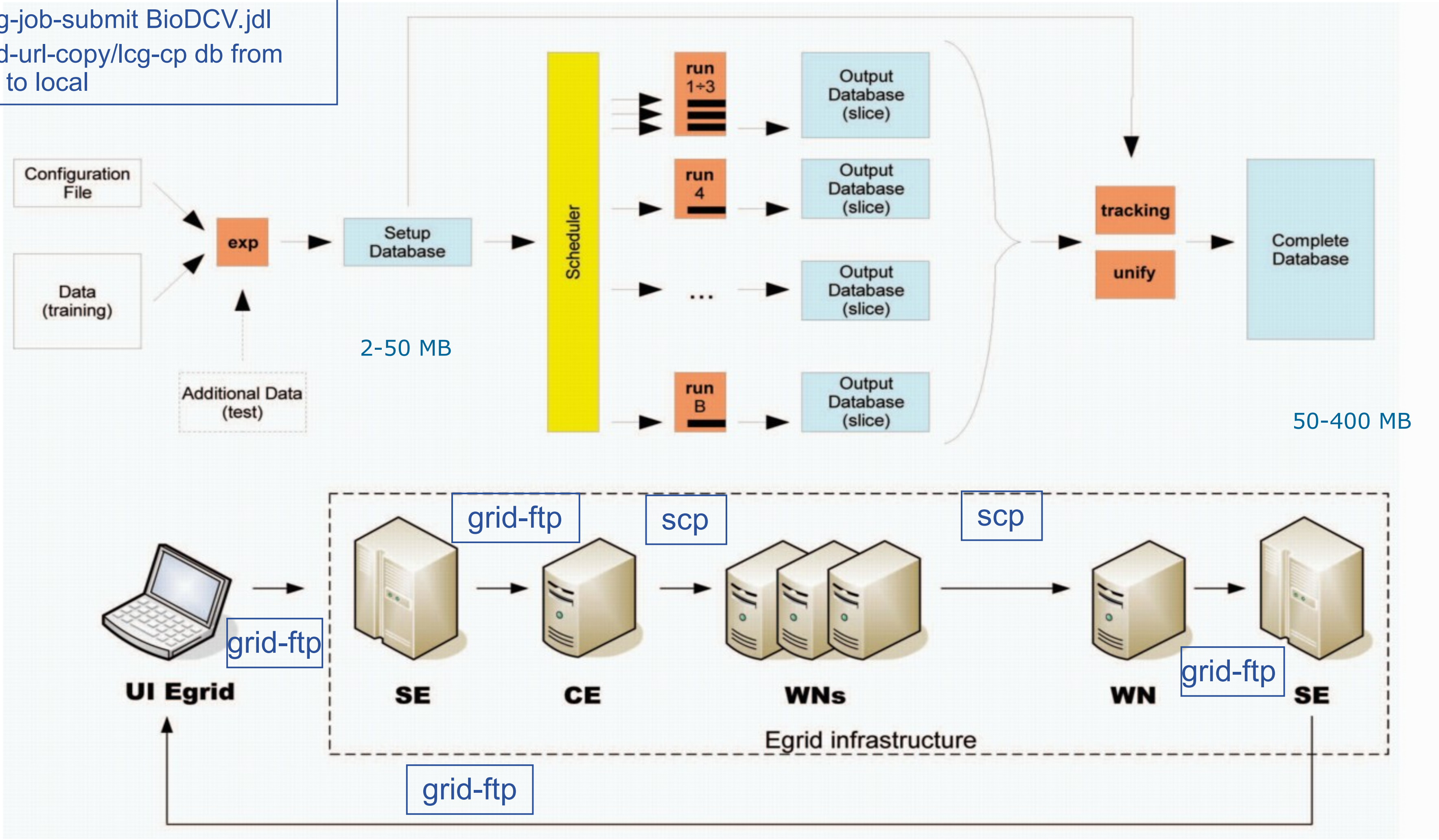
- We used the EGEE Biomed grid infrastructure
- 20 cpus/job, for a total of 100+120 jobs
- BioDCV jobs were run on many Biomed Sites in all Europe

Grid implementation

Commands:

1. grid-url-copy/lcg-cp db from local to SE
2. edg-job-submit BioDCV.jdl
3. grid-url-copy/lcg-cp db from SE to local

The BioDCV system



BioDCV jobs on CEs in Europe

BioDCV jobs was run on these Biomed’s CEs in Europe:

- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-3hr
- Destination: cluster.pnpi.nw.ru:2119/jobmanager-pbs-biomed
- Destination: helmsley.dur.scotgrid.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: ce101.grid.ucy.ac.cy:2119/jobmanager-lcgpbs-biomed
- Destination: grid-ce.ii.edu.mk:2119/jobmanager-lcgpbs-biomed
- Destination: ce01.kallisto.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-3hr
- Destination: lcgce01.gridpp.rl.ac.uk:2119/jobmanager-lcgpbs-bioL
- Destination: ce01.ariagni.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: lcg-ce.its.uiowa.edu:2119/jobmanager-lcgpbs-biomed
- Destination: lcgce01.gridpp.rl.ac.uk:2119/jobmanager-lcgpbs-bioL
- Destination: ce01.grid.acad.bg:2119/jobmanager-lcgpbs-biomed
- Destination: mu6.matrix.sara.nl:2119/jobmanager-pbs-short
- Destination: cluster.pnpi.nw.ru:2119/jobmanager-pbs-biomed
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-3hr
- Destination: gridba2.ba.infn.it:2119/jobmanager-lcgpbs-long
- Destination: ce1.pp.rhul.ac.uk:2119/jobmanager-pbs-biomedgrid
- Destination: cluster.pnpi.nw.ru:2119/jobmanager-pbs-biomed
- Destination: fal-pygrid-18.lancs.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: grid012.ct.infn.it:2119/jobmanager-lcglsf-short
- Destination: ce1.pp.rhul.ac.uk:2119/jobmanager-pbs-biomedgrid
- Destination: ce01.grid.acad.bg:2119/jobmanager-lcgpbs-biomed
- Destination: grid-ce.ii.edu.mk:2119/jobmanager-lcgpbs-biomed
- Destination: grid0.fe.infn.it:2119/jobmanager-lcgpbs-grid
- Destination: grid012.ct.infn.it:2119/jobmanager-lcglsf-short
- Destination: ce01.ariagni.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: grid0.fe.infn.it:2119/jobmanager-lcgpbs-grid
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-12hr
- Destination: epgce1.ph.bham.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: epgce1.ph.bham.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: ramses.dsic.upv.es:2119/jobmanager-pbs-biomedg
- Destination: t2ce02.physics.ox.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: ce1.pp.rhul.ac.uk:2119/jobmanager-pbs-biomedgrid
- Destination: ce01.kallisto.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: grid10.lal.in2p3.fr:2119/jobmanager-pbs-biomed

- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-6hr
- Destination: ce01.grid.acad.bg:2119/jobmanager-lcgpbs-biomed
- Destination: scaicl0.scai.fraunhofer.de:2119/jobmanager-lcgpbs-biomed
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-12hr
- Destination: grid-ce.ii.edu.mk:2119/jobmanager-lcgpbs-biomed
- Destination: gw39.hep.ph.ic.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: grid0.fe.infn.it:2119/jobmanager-lcgpbs-grid
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-12hr
- Destination: grid10.lal.in2p3.fr:2119/jobmanager-pbs-biomed
- Destination: t2ce02.physics.ox.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: prod-ce-01.pd.infn.it:2119/jobmanager-lcglsf-grid
- Destination: ce01.grid.acad.bg:2119/jobmanager-lcgpbs-biomed
- Destination: testbed001.grid.ici.ro:2119/jobmanager-lcgpbs-biomed
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-3hr
- Destination: lcg06.sinp.msu.ru:2119/jobmanager-lcgpbs-biomed
- Destination: ce01.isabella.grnet.gr:2119/jobmanager-pbs-biomed
- Destination: ce2.egee.cesga.es:2119/jobmanager-lcgpbs-biomed
- Destination: obsauvergridce01.univ-bpclermont.fr:2119/jobmanager-lcgpbs-biomed
- Destination: dgc-grid-40.brunel.ac.uk:2119/jobmanager-lcgpbs-short
- Destination: dgc-grid-40.brunel.ac.uk:2119/jobmanager-lcgpbs-short
- Destination: testbed001.grid.ici.ro:2119/jobmanager-lcgpbs-biomed
- Destination: ce01.isabella.grnet.gr:2119/jobmanager-pbs-biomed
- Destination: ce01.ariagni.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: obsauvergridce01.univ-bpclermont.fr:2119/jobmanager-lcgpbs-biomed
- Destination: ce01.marie.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: ce01.pic.es:2119/jobmanager-lcgpbs-biomed
- Destination: t2ce02.physics.ox.ac.uk:2119/jobmanager-lcgpbs-biomed
- Destination: ce01.kallisto.hellasgrid.gr:2119/jobmanager-pbs-biomed
- Destination: mu6.matrix.sara.nl:2119/jobmanager-pbs-short
- Destination: mars-ce.mars.lesc.doc.ic.ac.uk:2119/jobmanager-sge-12hr
- Destination:
- Destination:
- Destination:
- Destination:
- Destination: grid001.ics.forth.gr:2119/jobmanager-lcgpbs-biomed

- *And 50 more sites ….*
- *Production based on benchmarks*

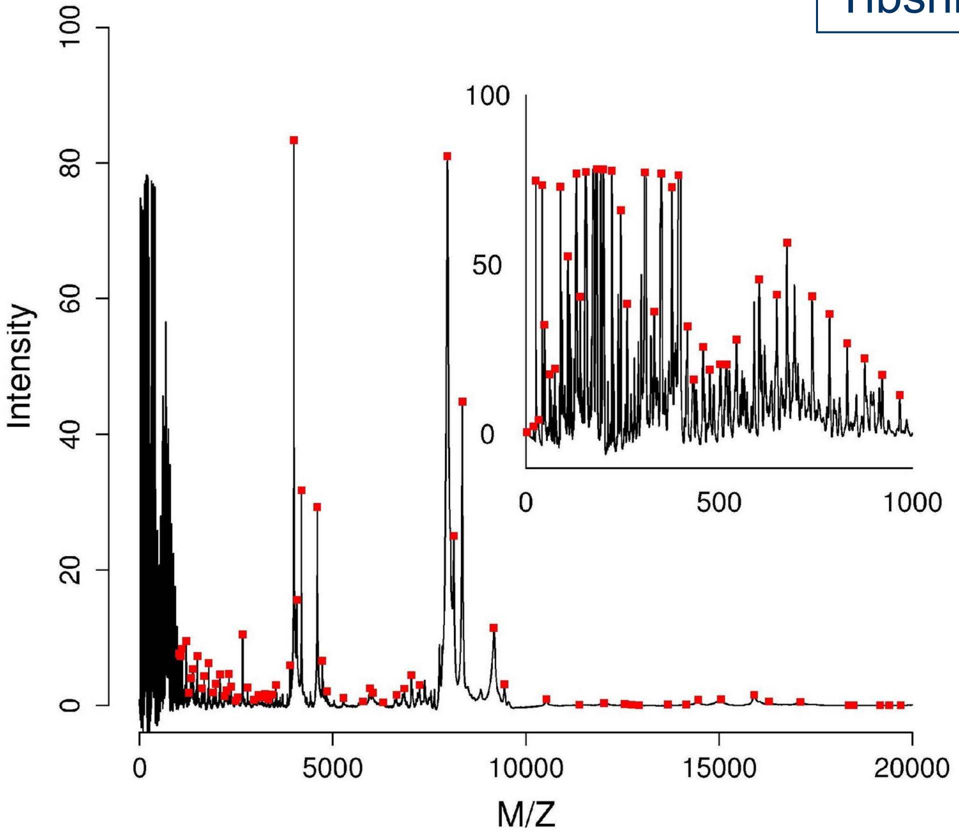
High-Throughput: Summary

- **Predictive profiling for high-throughput proteomics**
  - Selection Bias
    - Computational procedures for complete validation (BioDCV)
  - Biomarker Lists: reproducibility, stability, correlation
    - Modify machine learning algorithm to directly link selection to target functions (new kernel methods, or maybe simpler classifiers)
    - Consider the problem of batch preprocessing for true reproducibility
    - Use simulator to tune systems Use ensemble methods to achieve stability of selected lists
- **Applications**
  - Grid Computing as a viable resource for prediction with Mass spectrometry (SELDI-TOF, MALDI-TOF)

Details

Preprocessing – peak identification

Yasui 2003
Tibshirani 2004



Extracting Common Features

- Using peaks across multiple spectra can generate thousands of features.

- The number of examples required to learn a “reasonable” hypothesis increases exponentially with the number of features.

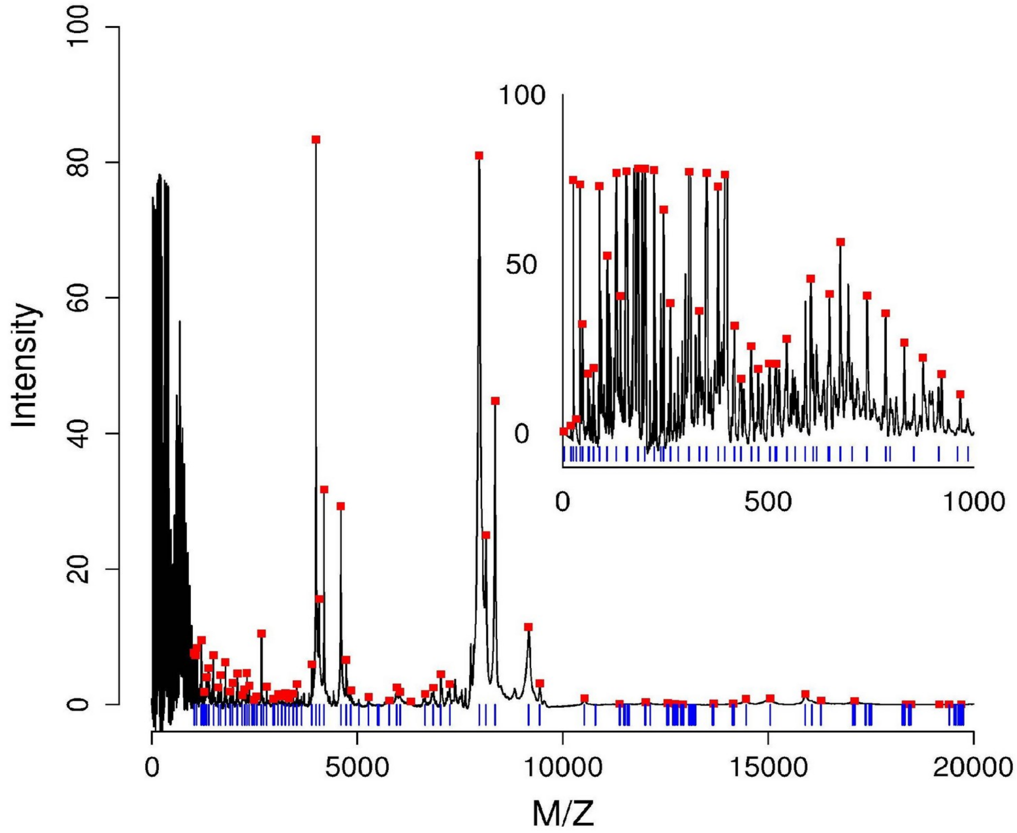
- Clustering reduces these features and has a rough correction for spectrometer resolution or drift of m/z between spectra.

Peak Alignment - Clustering

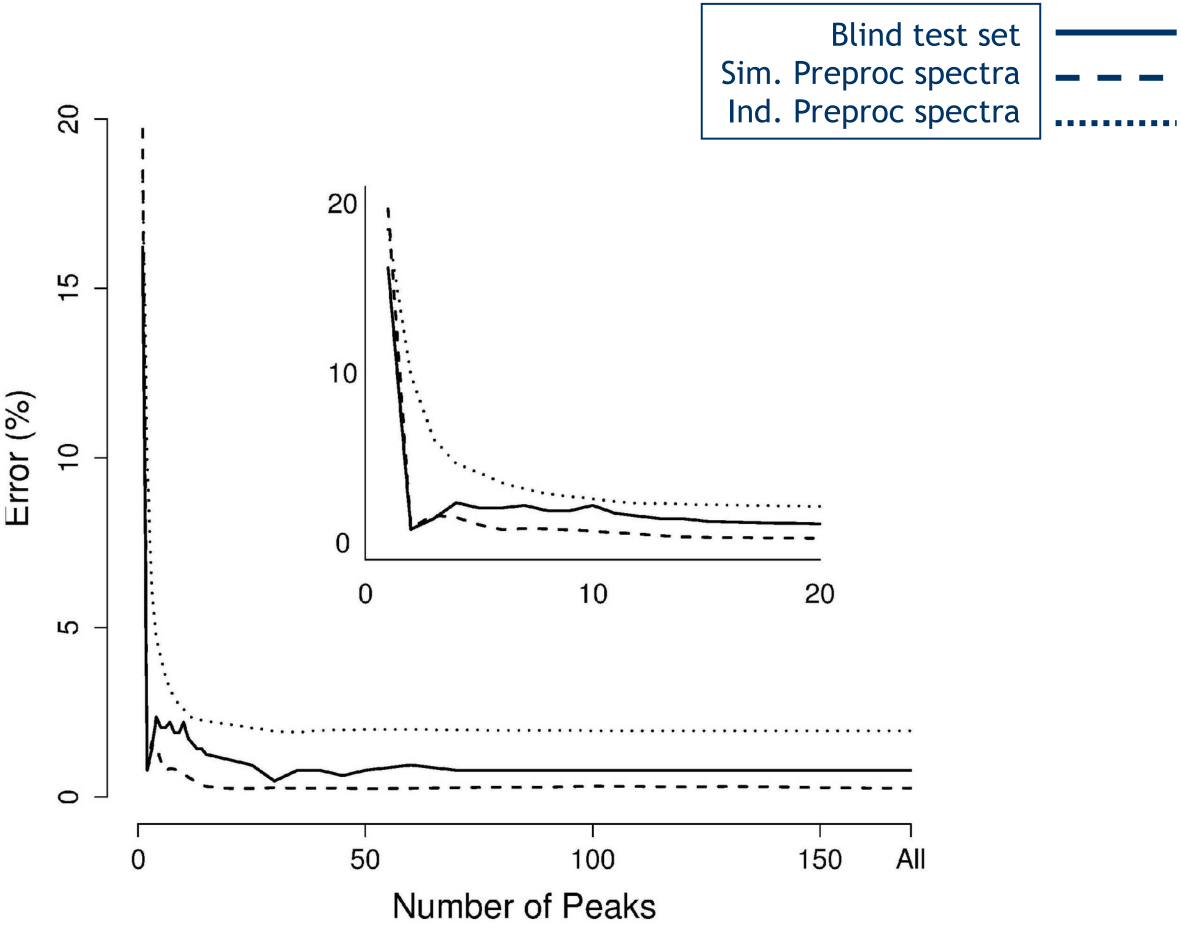
(Tibshirani 2004)

- Complete hierarchical clustering on log(m/z) axis over all spectra
- Build a dendrogram
- Cut at treshold T
  → induces centroids position

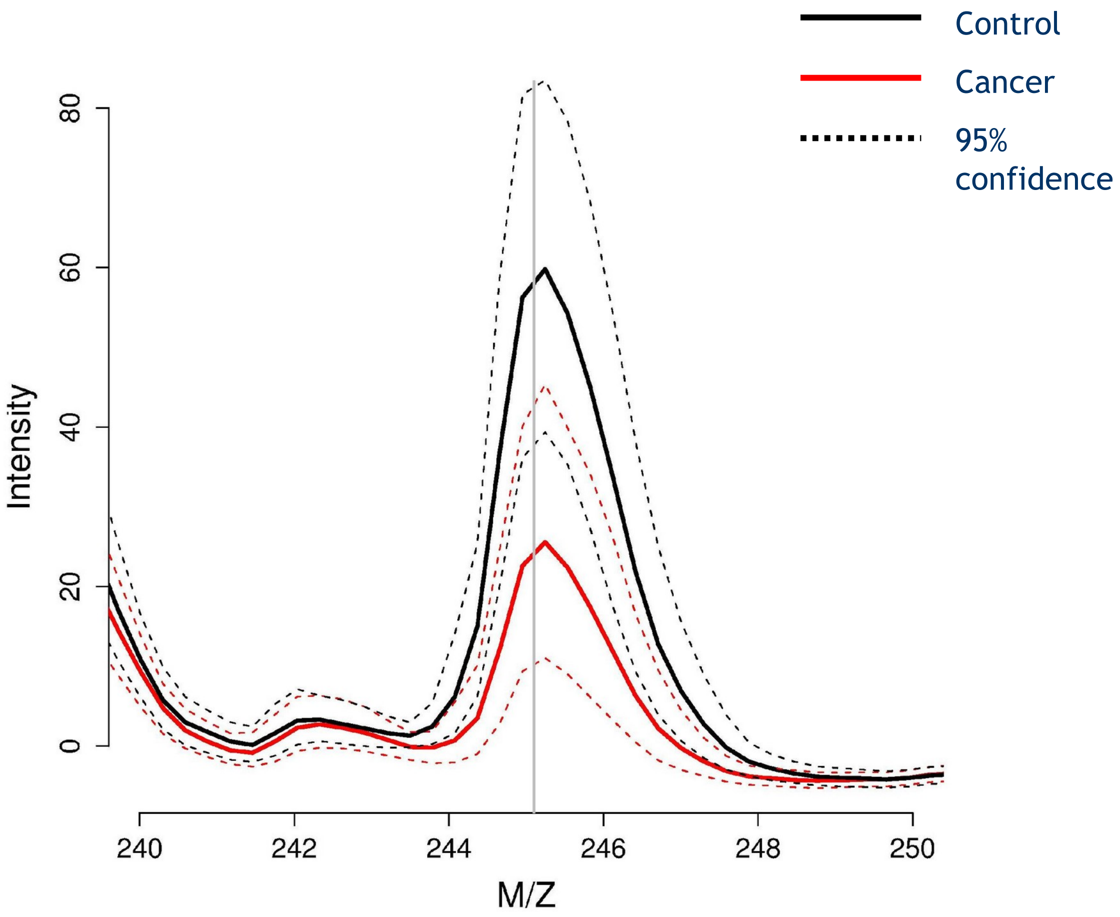
Spectra with extracted centroids



Error Curve vs. Features number



Top discriminant peaks


Control
Cancer
95% confidence
Intensity
M/Z