Statistics and NMR: a case study on human pancreatic cancer

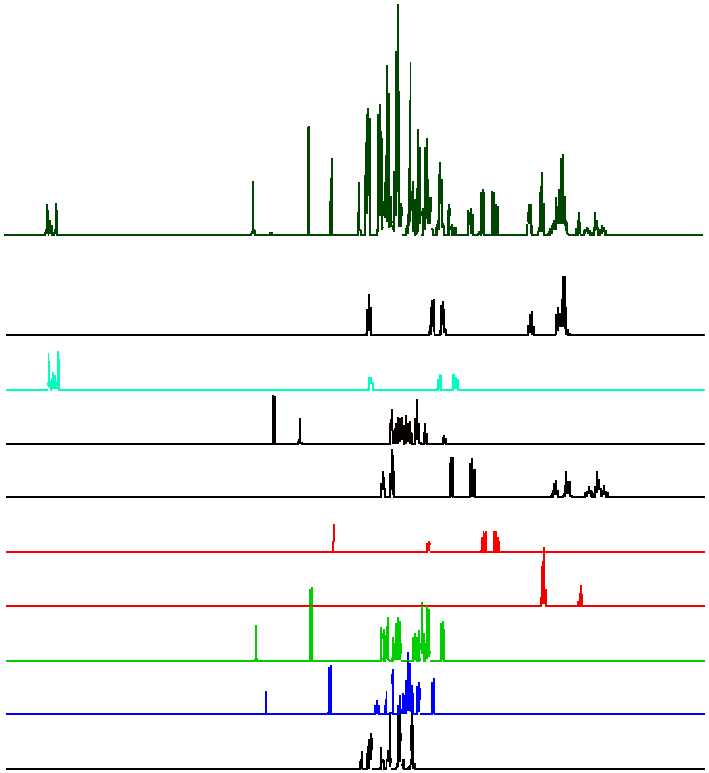
Claudia Napoli – Bruker Italia S.r.l.

13.06.12

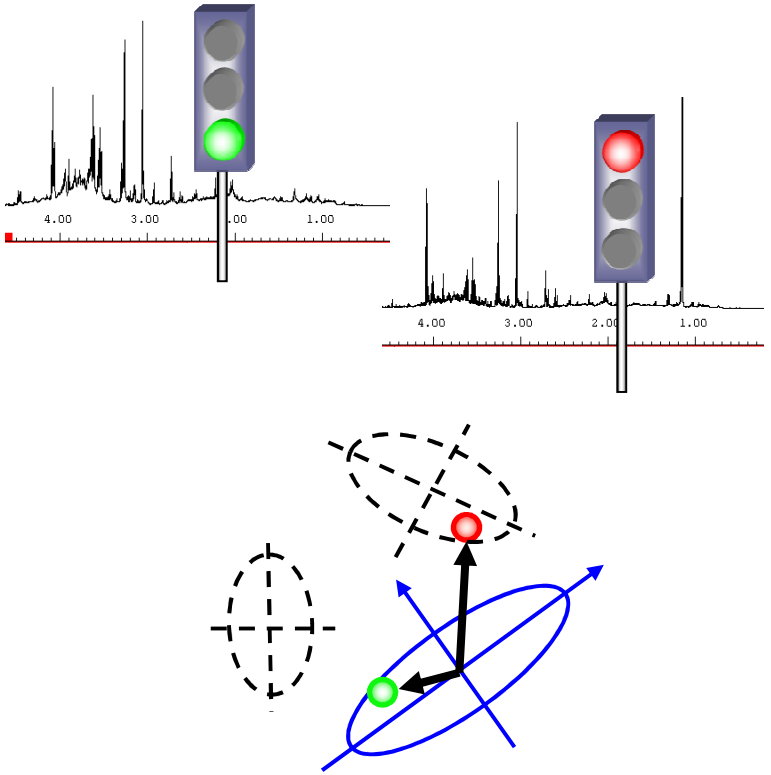
Summary

- Principal Component Analysis
- Metabolic trajectories
- PDAC example
- Discriminant Analysis
- Classification and cross validation of the model
- Targeted analysis
- Conclusions

Two approaches in data analysis ...



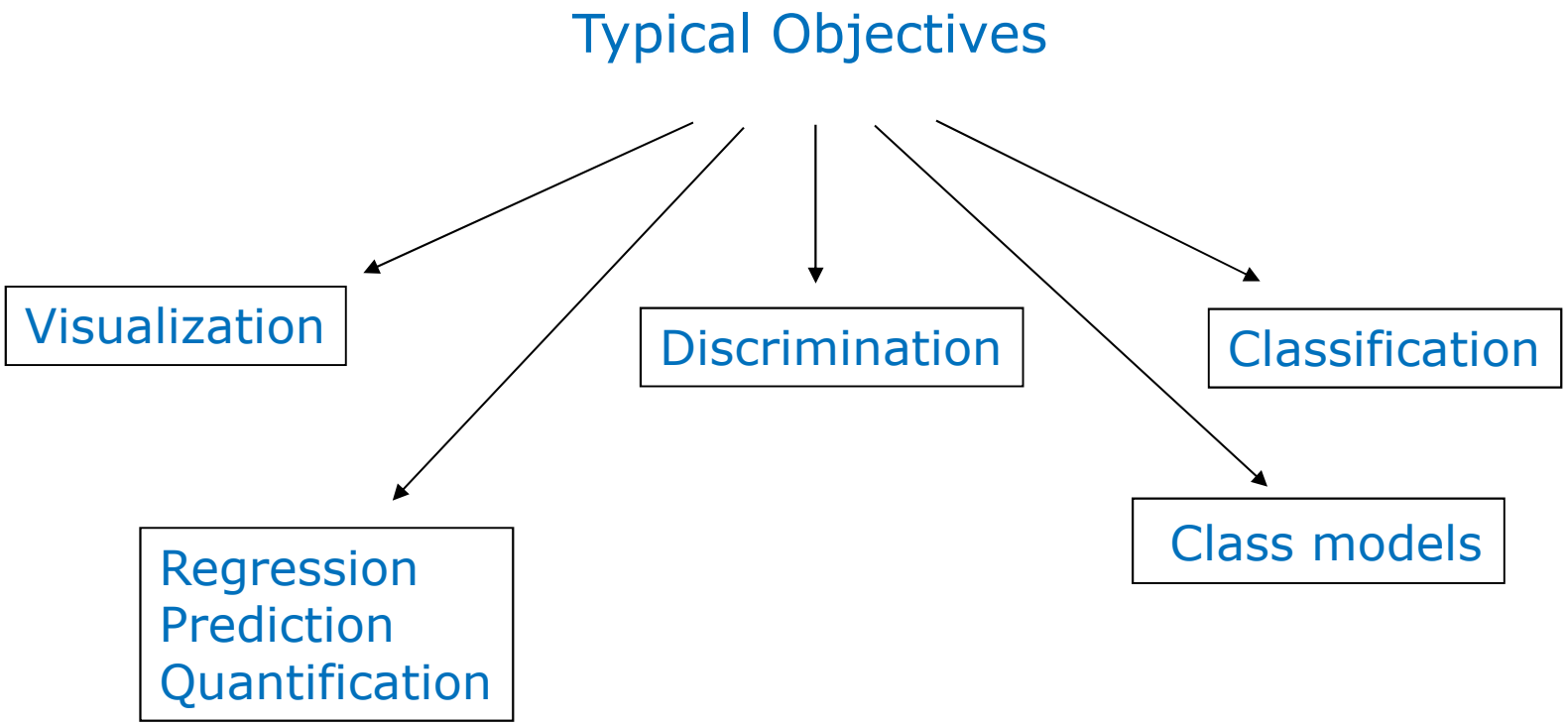
Metabolic Profiling



Fingerprinting: statistics

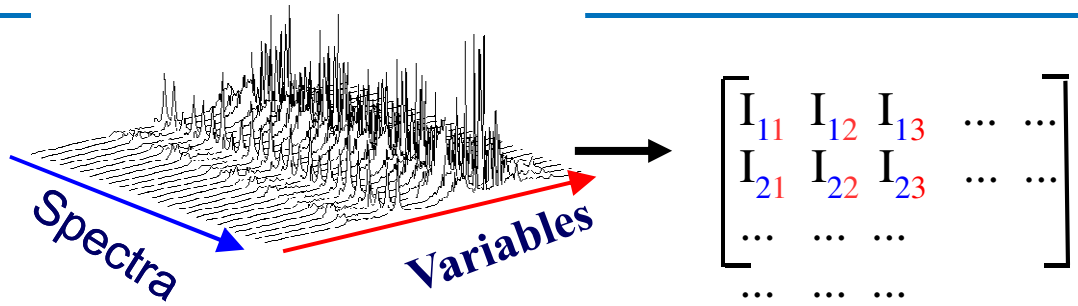
These two approaches can be combined in order to get complementary information!

Multivariate Statistics / Fingerprinting

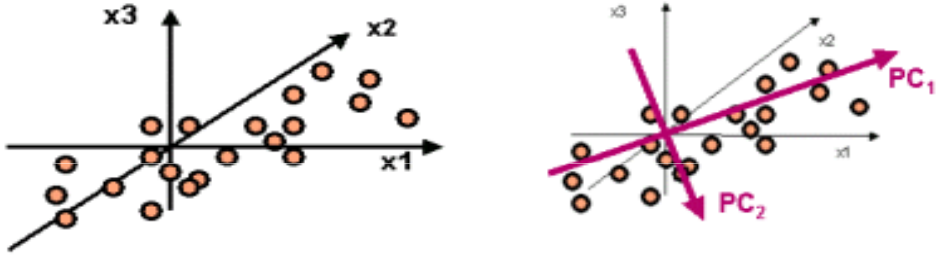


PCA workflow

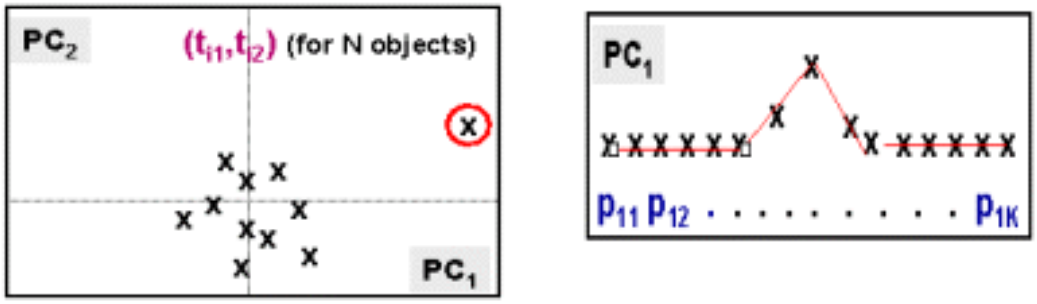
1. Bucket spectra



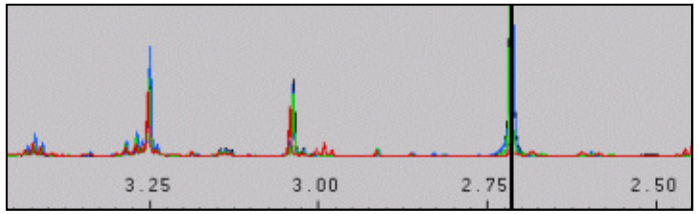
2. Do PCA



3. Analyze statistics plots

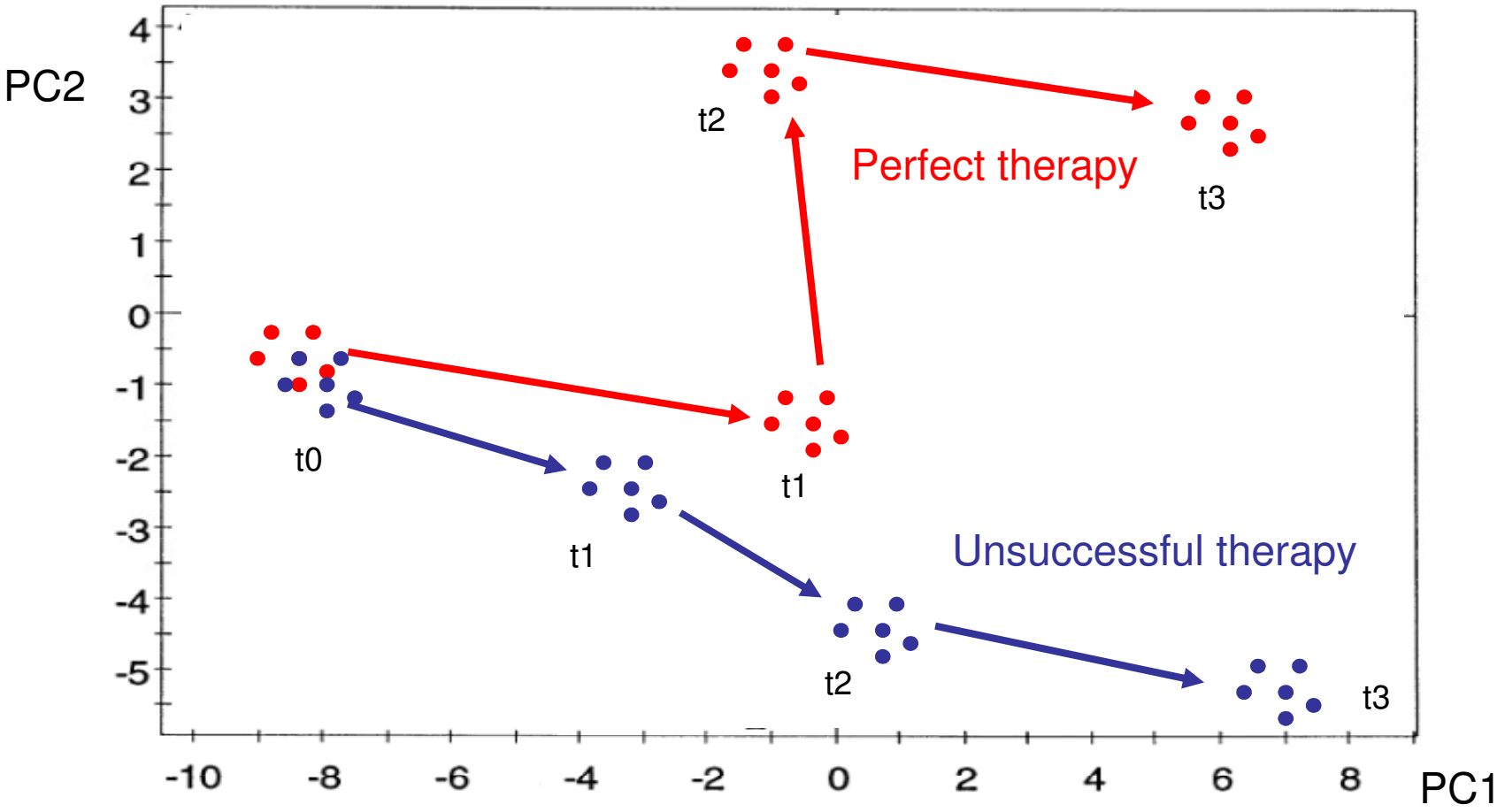


4. Which is the reason ?



Metabolic Trajectories →Dynamic Processes

How can we use metabolic trajectories for our purposes?



It requires a clear class separation in PCA

Our work on PDAC

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

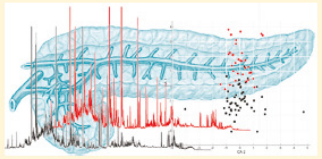
Urine Metabolic Signature of Pancreatic Ductal Adenocarcinoma by ^{1}H Nuclear Magnetic Resonance: Identification, Mapping, and Evolution

Claudia Napoli,[†,‡] Nicola Sperandio,[§] Rita T. Lawlor,[§] Aldo Scarpa,[§,∥] Henriette Molinari,[†] and Michael Assfalg*[,†]

[†]Department of Biotechnology and [∥]Department of Pathology and Diagnostics, University of Verona, Verona, Italy
[‡]Bruker Italia, Milan, Italy
[§]ARC-NET Center for Applied Research on Cancer, Verona University Hospital, Verona, Italy

Supporting Information

ABSTRACT: Pancreatic ductal adenocarcinoma (PDAC) has a dismal prognosis and is highly chemoresistant. Early detection is the only means to impact long-term survival, but screening methods are lacking. Given the complex and heterogeneous nature of pancreatic cancer, unbiased analytical methods such as metabolomics by nuclear magnetic resonance (NMR) spectroscopy show promise to identify disease-specific molecular fingerprints. NMR profiles constitute a fingerprint of the biofluid, reporting quantitatively on all detectable small biomolecules. NMR spectroscopy was applied to investigate the urine metabolome of PDAC patients ($n = 33$) and to detect altered metabolic profiles in comparison with healthy matched controls ($n = 54$). The spectral data were analyzed using multivariate statistical techniques. Statistically significant differences were found between urine metabolomic profiles of PDAC and control individuals ($p < 10^{-5}$). Group discrimination was possible due to average concentration differences of several metabolite signals, pointing to a multimolecular signature of the disease. The robustness of the determined statistical model is confirmed by its predictive performance (sensitivity = 75.8%, specificity = 90.7%). Additionally, the method allowed for a neat separation between spectral profiles of individuals with intermediate and advanced pathologic staging, as well as for the discrimination of samples based on tumor localization. NMR spectroscopy analysis of urinary metabolic profiles proved successful in identifying a complex molecular signature of PDAC. Furthermore, results of a descriptive-level analysis show the possibility to follow disease evolution and to carry out tumor site mapping. Given the high reproducibility and the noninvasive nature of the analytical procedure, the described method bears potential to impact large-scale screening programs.



KEYWORDS: pancreatic ductal adenocarcinoma, metabolomics, NMR spectroscopy, urine

- Pancreatic ductal adenocarcinoma (PDAC) is a very aggressive and resistant cancer, the fifth cause of tumore death in Europe.
- At present, the diagnosis is possible, in the 80% of the cases, at a very advanced disease.
- No specific tumor markers exists for PDAC

Samples

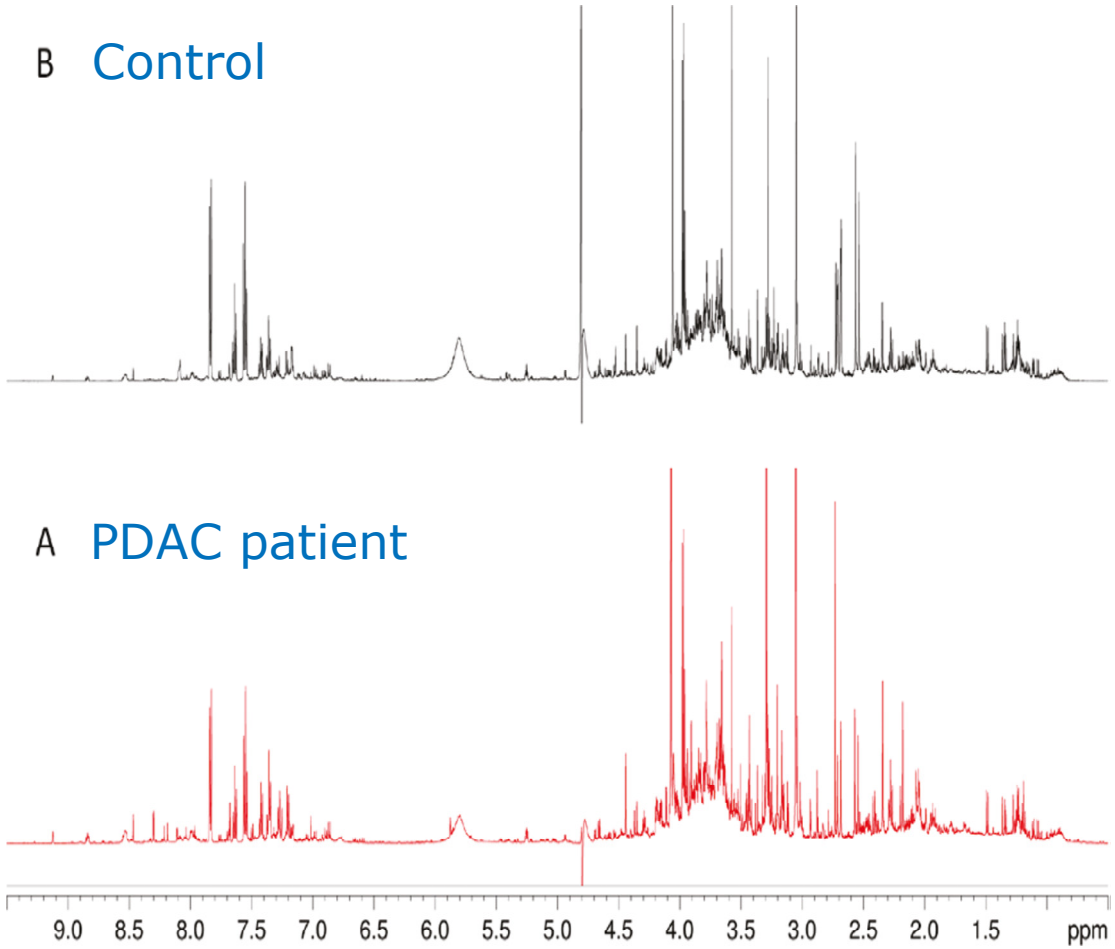
- Urine
- Plasma
- Bile
- Serum
- cerebrospinal fluid (CSF)
- seminal fluid
- amniotic fluid

- Urine represents an ideal biofluid for NMR based metabolomics since it contains the highest number of water-soluble metabolites
- Non invasive sampling and minimal sample preparation
- Limited experience exist with urinary metabolic markers for cancer diagnosis

Experimental design

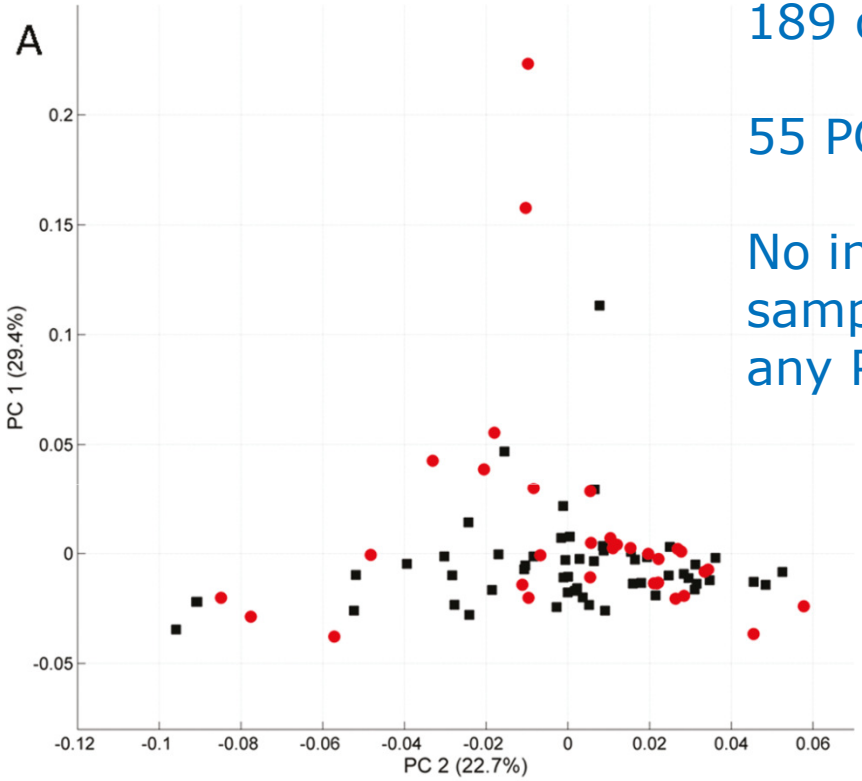
- 33 PDAC patients – 54 healthy controls
- Only male individuals selected
- Limited age range: 62+/- 6 years
- Clinical data collection included age, weight, weight loss, diagnosis, medical treatments and concomitant diseases
- Investigation of class discrimination and classification based on disease, staging and tumor site.

Statistical approach



- Principal Component Analysis (PCA)
- Multivariate Analysis of variance (MANOVA)
- Canonical Analysis (CA)
- K-Nearest Neighbor Classification (K-NN)
- Analysis of contribution of the original variables to group discrimination

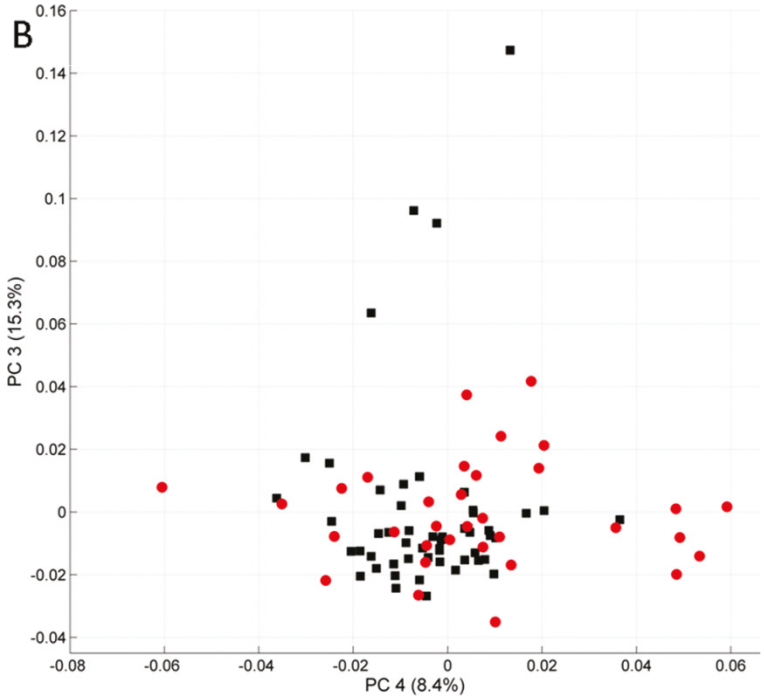
PCA



189 original variables (buckets)

55 PCs explaining 99.9% of total variance.

No indication of group clustering of the samples according to the health status, in any PC dimension.



PDAC patients

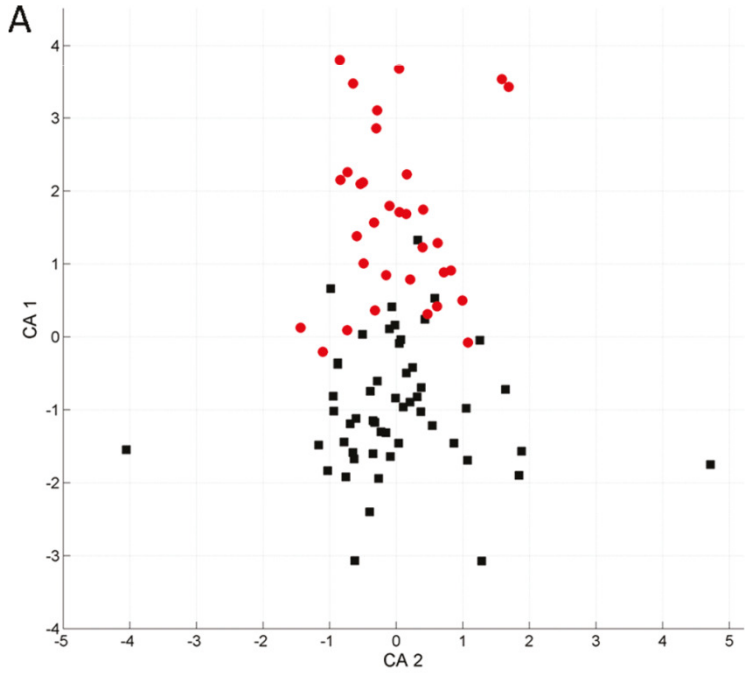
Controls

MANOVA + CA

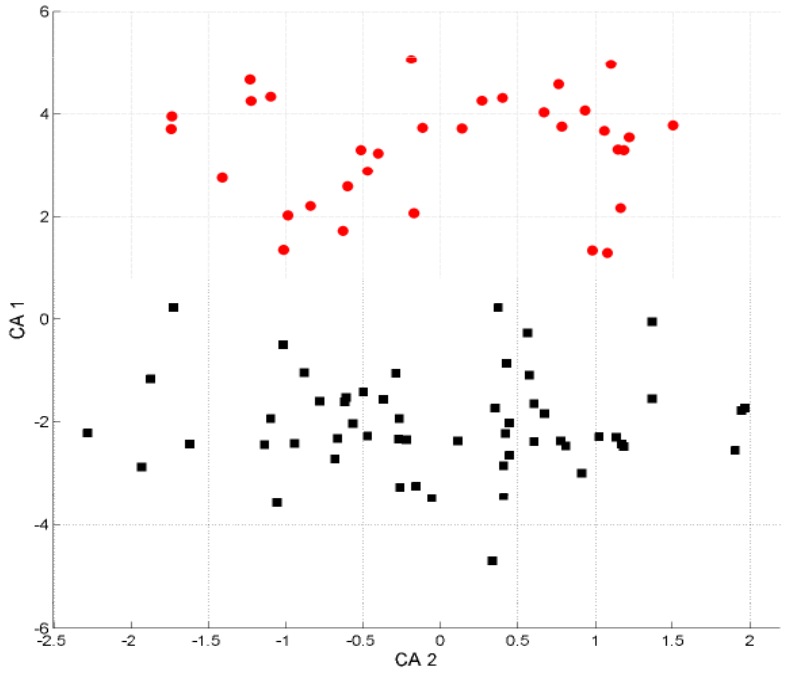
MANOVA → Group means dimensionality 1
→ 2 distinct classes

● PDAC patients
■ Controls

28 PCs (from PCA) taken into account (99% of total variance)



55 PCs (from PCA) taken into account (99.9% of total variance)



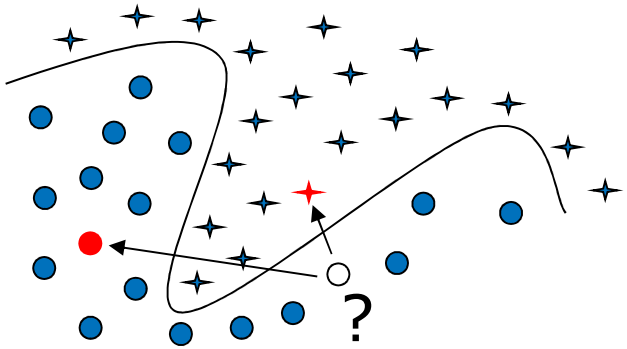
Classification method

- NCM (Nearest Class Mean)

The one used in our routines. It looks for the correct classification of the sample by comparing the descriptor value of the unknown data with the means of values of different defined classes.

- K-NN (K-Nearest Neighbours)

It is used in presence of particular distributions of data that make previous method inapplicable.



Methods of validation

- Test Set Validation
- Cross Validation
- Monte Carlo approaches

Confusion matrix:
Two class problem from bodyfluid nmr

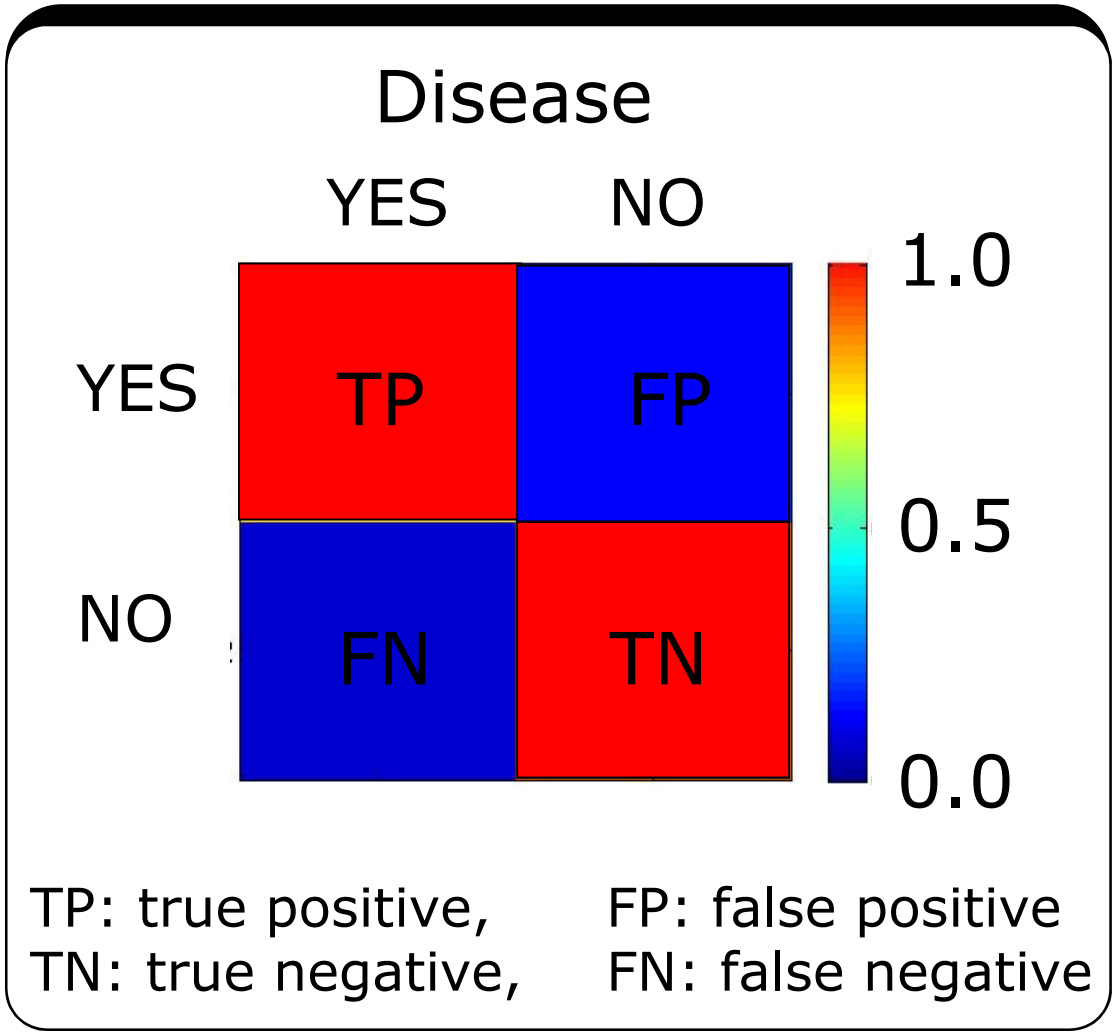
100 Randomsplits

N: total of subjects
Nd: total of diseased subjects
Nc: total of healthy subjects

Accuracy
Rate of correct tests
[(TP+TN)/N]

Sensitivity
Rate of diseased individuals with a positive test
(TP/Nd)

Specificity
Rate of nondiseased with a negative test
(TN/Nc)



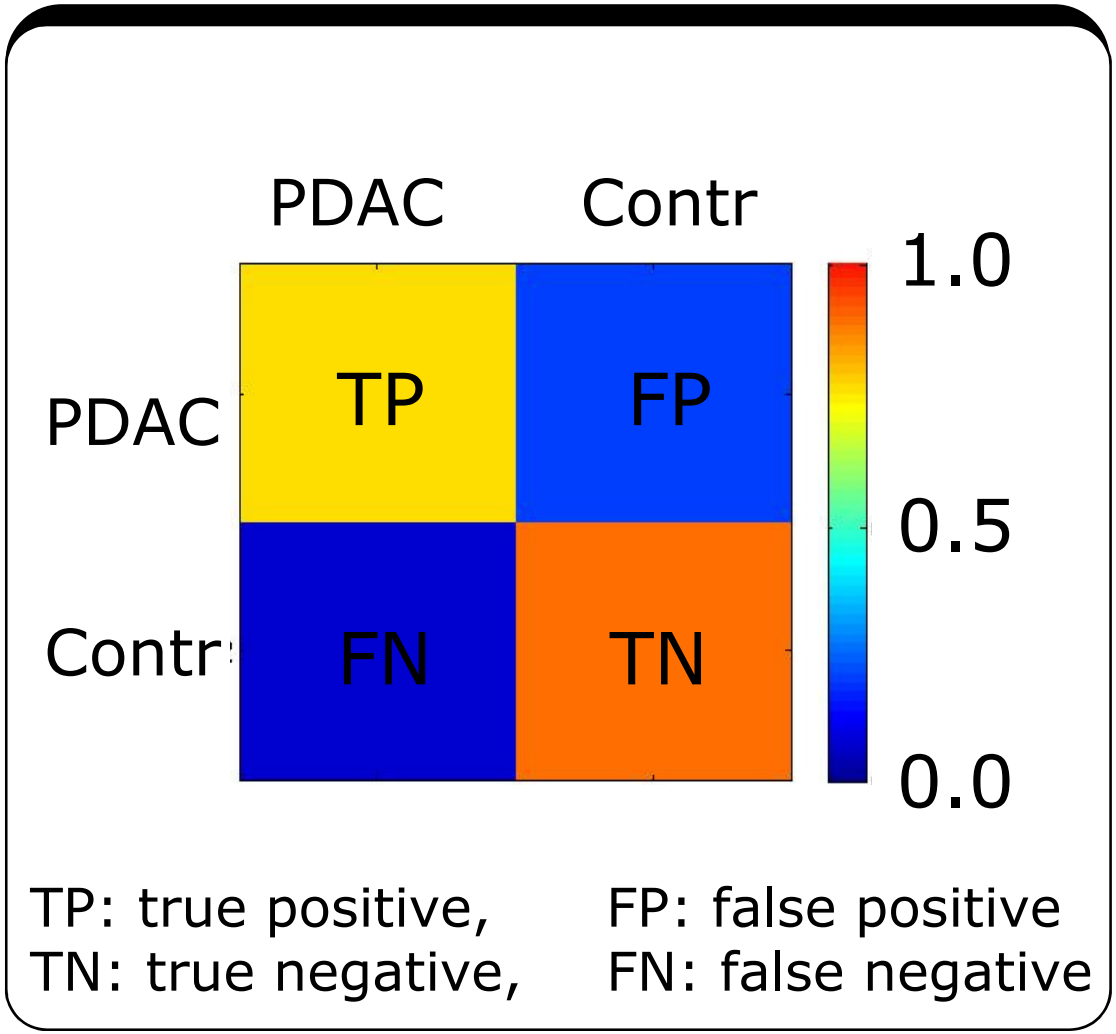
PDAC confusion matrix

Monte Carlo embedded CV

100 Randomsplits

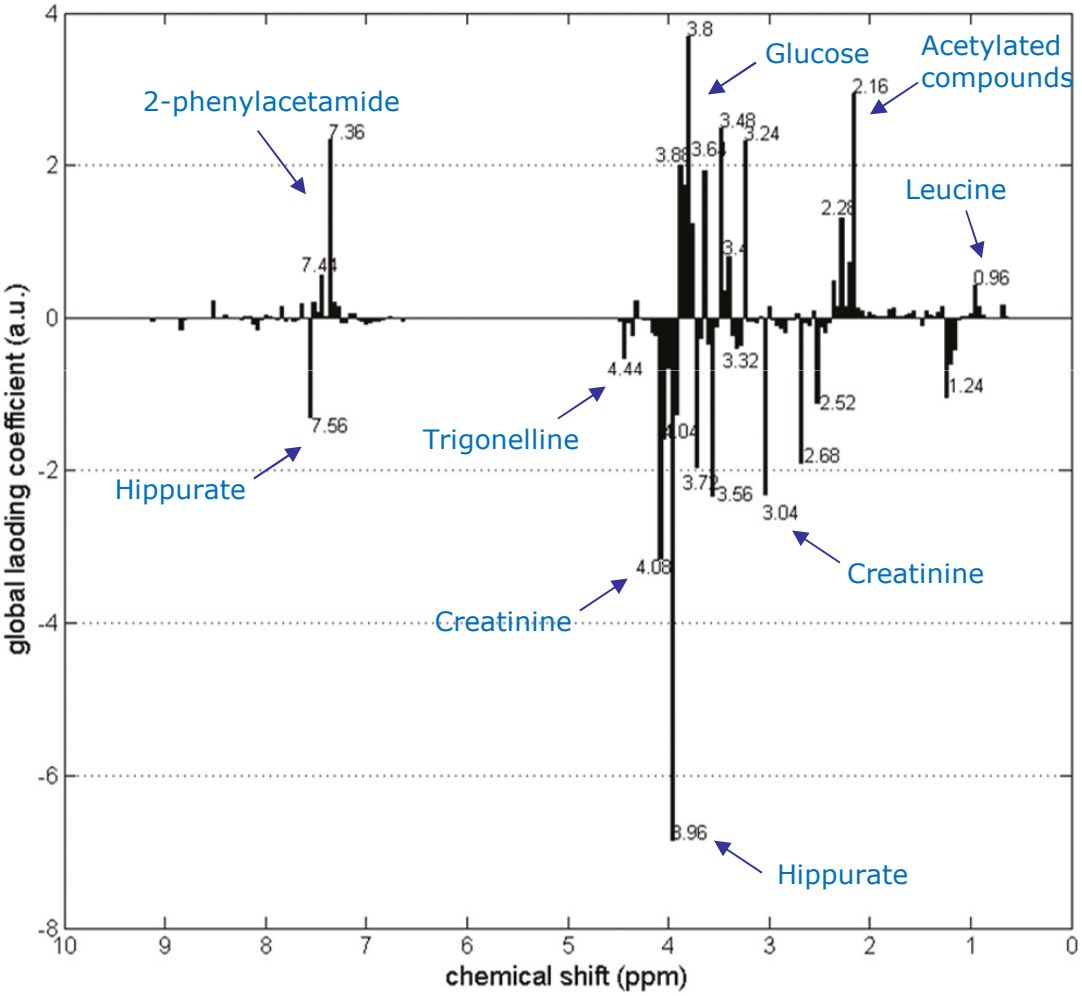
Accuracy: **85.1%**
Sensitivity: **75.8%**
Specificity: **90.7%**

Best classification rate achieved with **28 PCs**!



Contribution of the original spectral signals

CA → PCA → Original variables



Positive value indicate increased intensity in PDAC patients in respect to controls

Negative value indicate decreased intensity in PDAC patients in respect to controls

PDAC: tumor site and staging differentiation

An exploratory analysis was performed only on PDAC samples (33)

Application of PCA/CA on these samples allows to discriminate intermediate from advanced staging and the different tumor site (uncinate, head and body).

Both of these analyses were kept at the descriptive level, without attempting to evaluate the predictive ability of the models due to limited group size availability.

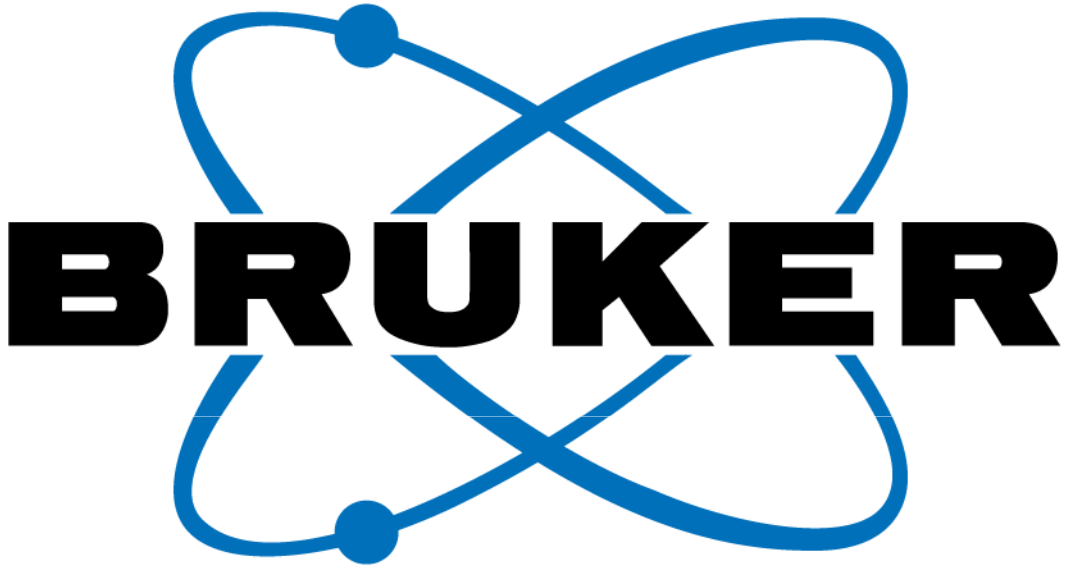
This successful investigation however points to the possible application of the NMR metabolomics method to differentiate different tumor localization and disease staging.

Thanks to…

- Dr. Michael Assfalg
  Dr. Henriette Molinari
  Department of Biotechnology, University of Verona

- Dr. Hartmut Schaefer
- Dr. Birk Schuetz
  Bruker BioSpin GmbH, Germany

Thanks for your attention!!



www.bruker.com