Statistically Integrated Metabonomic-Proteomic Studies on a Human Prostate Cancer Xenograft Model in Mice

Mattias Rantalainen, Olivier Cloarec, Olaf Beckonert, I. D. Wilson,
David Jackson, Robert Tonge, Rachel Rowlinson, Steve Rayner,
Janice Nickson, Robert W. Wilkinson, Jonathan D. Mills, Johan Trygg,
Jeremy K. Nicholson, and Elaine Holmes

Taru Tukiainen
Helsinki University of Technology

Outline

- Metabonomics
- Integrating omics data
- PLS, OPLS, O2PLS
- Prostate cancer
- Study design
- Results
- Discussion
- Comments

Metabonomics

- Definition:

  ‘the quantitative measurement of the time-related multiparametric response of living systems to pathophysiological stimuli or genetic modification’

  Nicholson & al., Nat Rev Drug Discovery 1, 153 (2002)

- Provides complementary information to that obtained from genomics, transcriptomics and proteomics
- Conducted on biological samples which represent the biochemistry of the whole system, e.g., urine and blood plasma and serum
- NMR (nuclear magnetic resonance) and MS key technologies

^{1}H NMR metabonomics

- ^{1}H NMR as a metabonomic tool
  - Specific yet non-selective
  - Little or no sample preparation
  - Rapid and non-destructive
  - Small sample sizes
  - Spectra highly reproducible

- Chemometrics methods (e.g. PCA and PLS) most common analysis methods

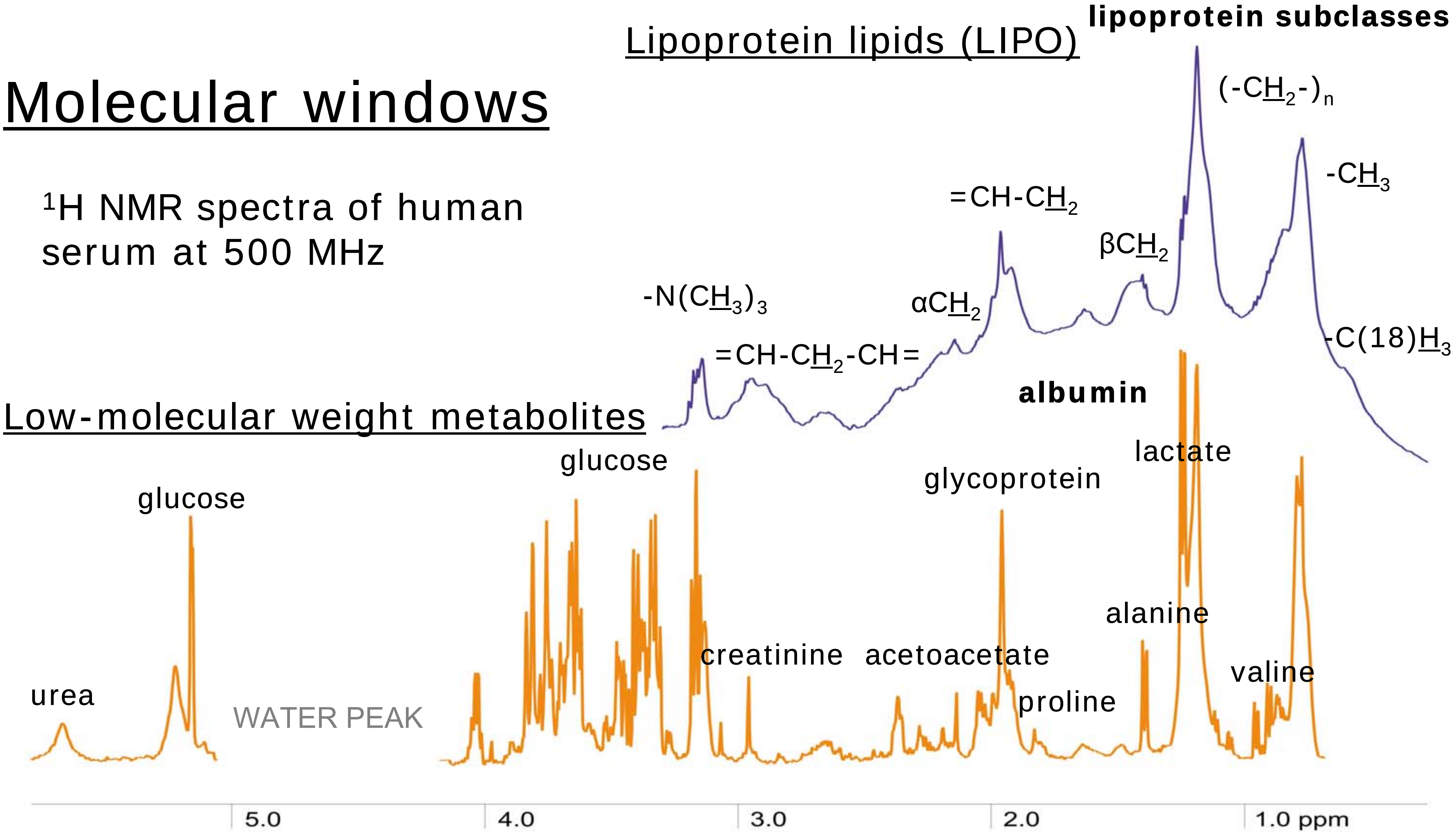
^{1}H NMR spectra

Molecular windows

^{1}H NMR spectra of human serum at 500 MHz

Lipoprotein lipids (LIPO)

Low-molecular weight metabolites



Integrating omics data

- Why?
  - Overview of all the biological processess
  - Improved undestanding of the biological system by defining how variables relate to each other

- Problems?
  - Mammalian biocomplexity
  - Requires a wide range of technical expertise

Partial least squares (PLS)

- Modelling technique that combines features from PCA and multiple regression
- Goal: to predict Y (matrix of observations) from X (matrix of predictors) and to describe their common structure
- Finds components from X that are also relevant for Y
- PLS decomposes both X and Y as a product of orthogonal scores and loadings

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F}$$

T and U are score matrices (latent variables), P and Q loading matrices, E and F matrices of residuals

- Orthogonal score vectors are created by maximising the covariance between different sets of variables (sets of columns from X and Y)
  - i.e., obtain pair of vectors t = Xw and u = Yc with the constraints that $w^T w = 1$, $t^T t = 1$ and $t^T u$ be maximal
- When the first score vectors (t and u) are found, they are subtracted from X and Y, respectively, and the procedure is re-iterated until X becomes a null matrix

Partial least squares (PLS) cont.

Example: NIPALS PLS algorithm

Initialise vector u with random numbers.

Repate the following steps until convergence

1) $\mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$	4) $\mathbf{c} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
2) $\|\mathbf{w}\| \to 1$	5) $\|\mathbf{c}\| \to 1$
3) $\mathbf{t} = \mathbf{X}\mathbf{w}$	6) $\mathbf{u} = \mathbf{Y}\mathbf{c}$

Loadings p and q are calculated as coefficients of regressing X on t and Y on u

$$\mathbf{p} = \mathbf{X}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t}) \quad \text{and} \quad \mathbf{q} = \mathbf{Y}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$$

Score vectors are used to deflate the matrices X and Y

$$\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^T \quad \text{and} \quad \mathbf{Y} = \mathbf{Y} - \mathbf{t}\mathbf{t}^T\mathbf{Y}/(\mathbf{t}^T\mathbf{t}) = \mathbf{Y} - \mathbf{t}\mathbf{c}^T$$

Reiterate until X becomes a null matrix.

Estimate of the PLS regression model

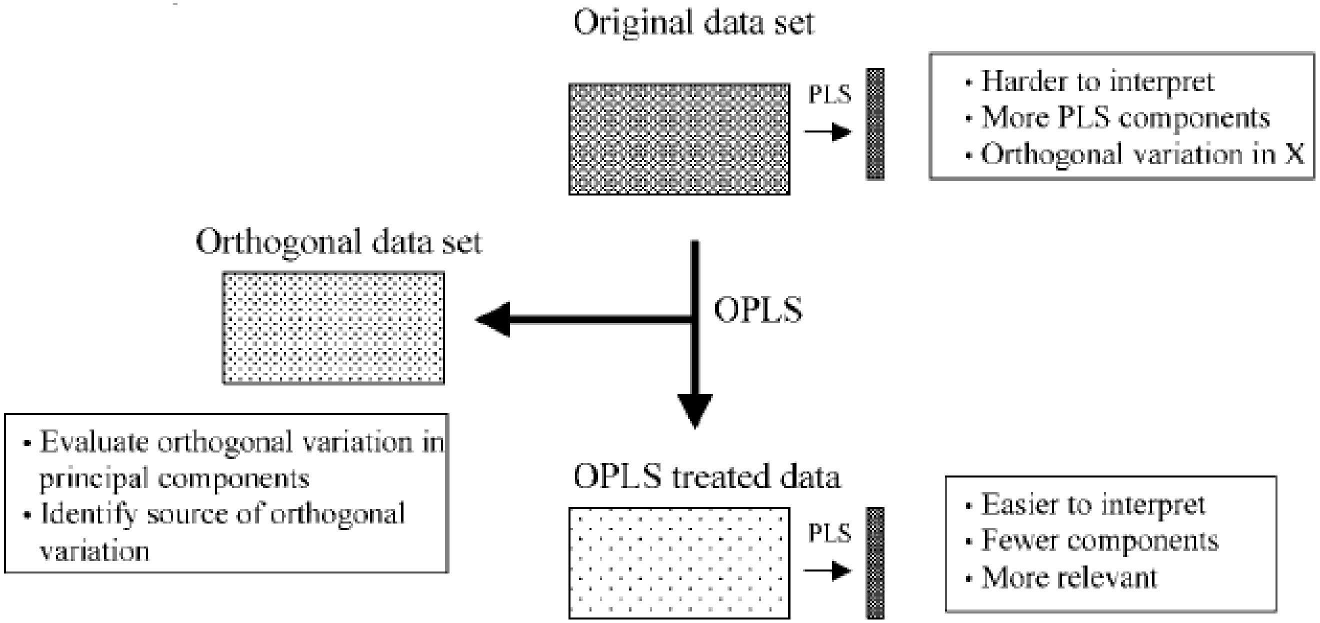
$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B} = \mathbf{T}\mathbf{T}^T\mathbf{Y} = \mathbf{T}\mathbf{C}^T$$

B represents the regression coefficients

Orthogonal projections to latent structures (OPLS)

- Similar method to PLS but with an integrated Signal Correction filter
- Removes systematic variation from an input data set X (predictors) not correlated, i.e., *orthogonal*, to the response matrix Y (observations)
- Modification of the NIPALS PLS algorithm
- Benefits:
  - Improves interpreation of PLS models
  - Reduces model complexity
  - Allows the non-correlated variation to be further analysed

Orthogonal projections to latent structures (OPLS) cont.



- Harder to interpret
- More PLS components
- Orthogonal variation in X

- Evaluate orthogonal variation in principal components
- Identify source of orthogonal variation

- Easier to interpret
- Fewer components
- More relevant

Figure 1. Overview of orthogonal projections to latent structures (O-PLS).

O2PLS

- Modification of OPLS
- Allows modelling and prediction in both directions between the data matrices X and Y
- Separates the X-Y related (predictive) variance and the structured noise (orthogonal) present in the data
- Modification of the NIPALS PLS algorithm

O2PLS cont.

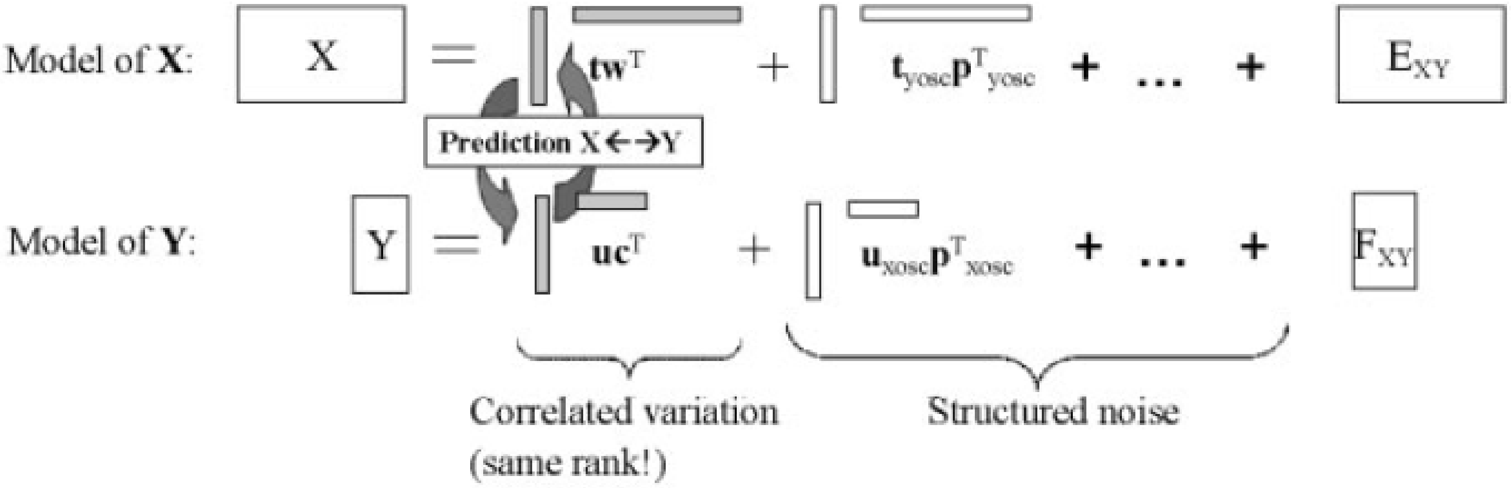


Figure 3. O2-PLS provides a model of both **X** and **Y**. Each model can have a different number of structured noise components, but the jointly correlated X–Y components (**T**,**U**) will always be of the same rank. It is also predictive in both ways.

Prostate cancer

- Prostate: a gland in the male reproductive system
- In UK around 30 000 men a year are diagnosed with prostate cancer, 10 000 die of it
- Affects most frequently men over age 50
- Diagnosis based on biomarkers
  - Prostate specific antigen (PSA), the ’gold’ standard
  - Carcinoembryonic antigen (CEA)
- Biomarkers unreliable, high false-negative and false-positive discovery rates
- **Need to identify and validate more biochemical and molecular biomarkers**

Study design

10 mice of which 5 animals recieved a prostate cancer tumor transplant

Blood plasma collected on day 30

Metabonomics

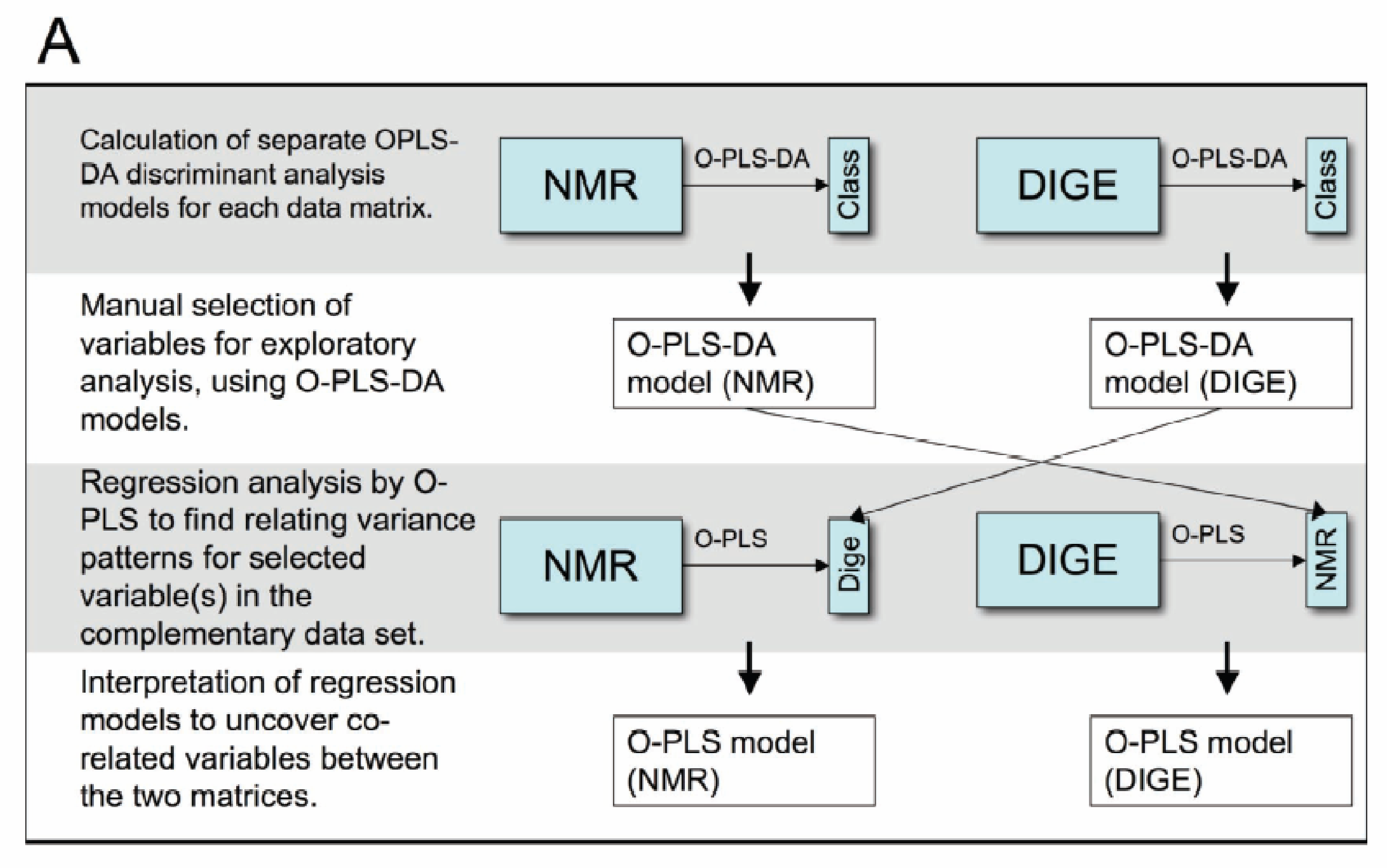
- ^{1}H NMR of blood plasma at 600 MHz
  - 1D (Lipoprotein lipids) spectrum
  - CPMG (Low-molecular weight metabolites) spectrum

Proteomics

- 2D-Gel analysis of blood plasma
- Identification of protein spots of interest by LC-MSMS and Mascot

Methods

O-PLS-DA



A

Calculation of separate OPLS-DA discriminant analysis models for each data matrix.

NMR → O-PLS-DA → Class

DIGE → O-PLS-DA → Class

Manual selection of variables for exploratory analysis, using O-PLS-DA models.

O-PLS-DA model (NMR)

O-PLS-DA model (DIGE)

Regression analysis by O-PLS to find relating variance patterns for selected variable(s) in the complementary data set.

NMR → O-PLS → Dige

DIGE → O-PLS → NMR

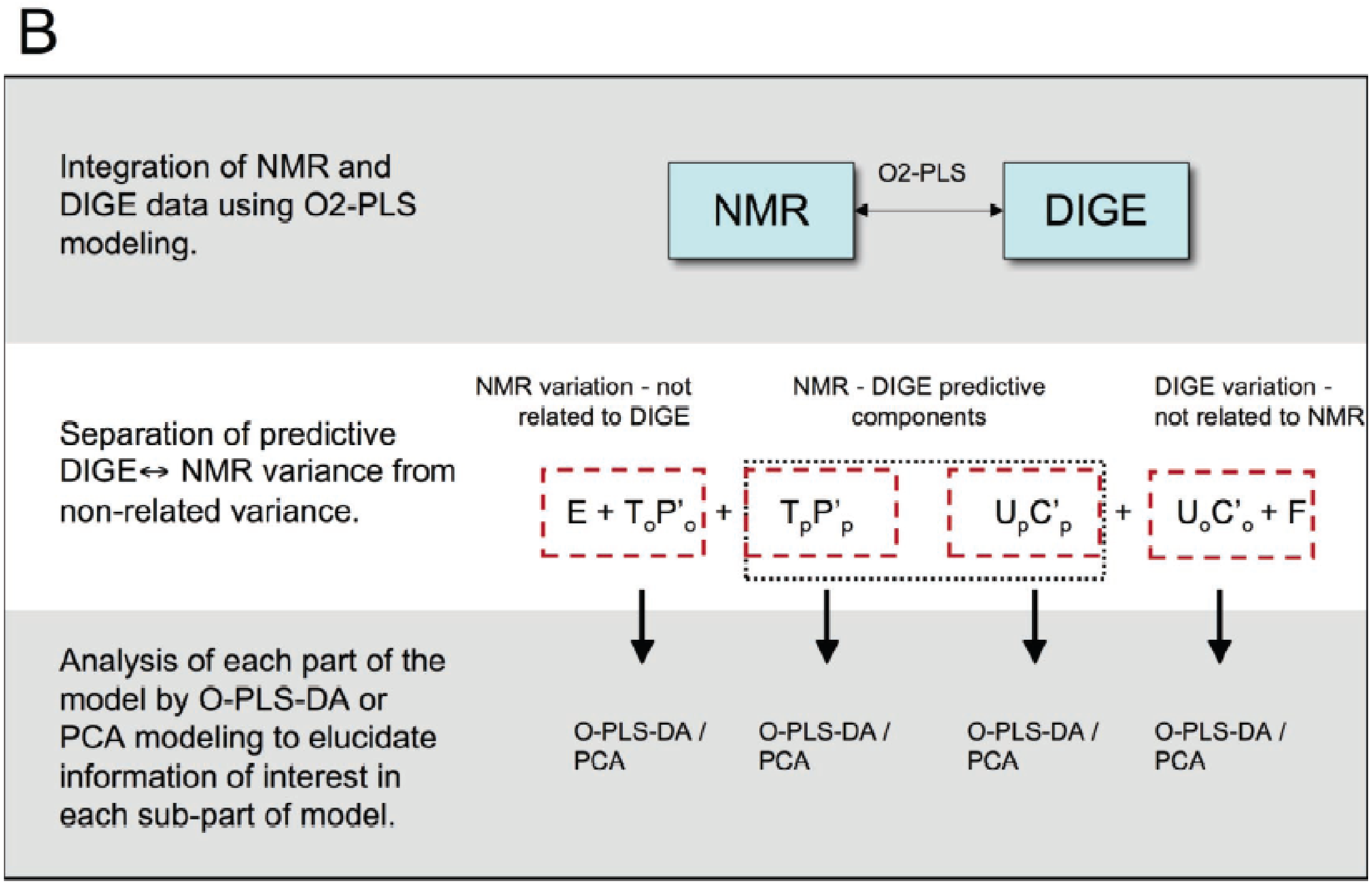
Interpretation of regression models to uncover co-related variables between the two matrices.

O-PLS model (NMR)

O-PLS model (DIGE)

All models validated by 5-fold cross validation

O2-PLS



B

Integration of NMR and DIGE data using O2-PLS modeling.

Separation of predictive DIGE↔ NMR variance from non-related variance.

NMR variation - not related to DIGE
NMR - DIGE predictive components
DIGE variation - not related to NMR

$E + T_oP'_o$ + $T_pP'_p$ $U_pC'_p$ + $U_oC'_o + F$

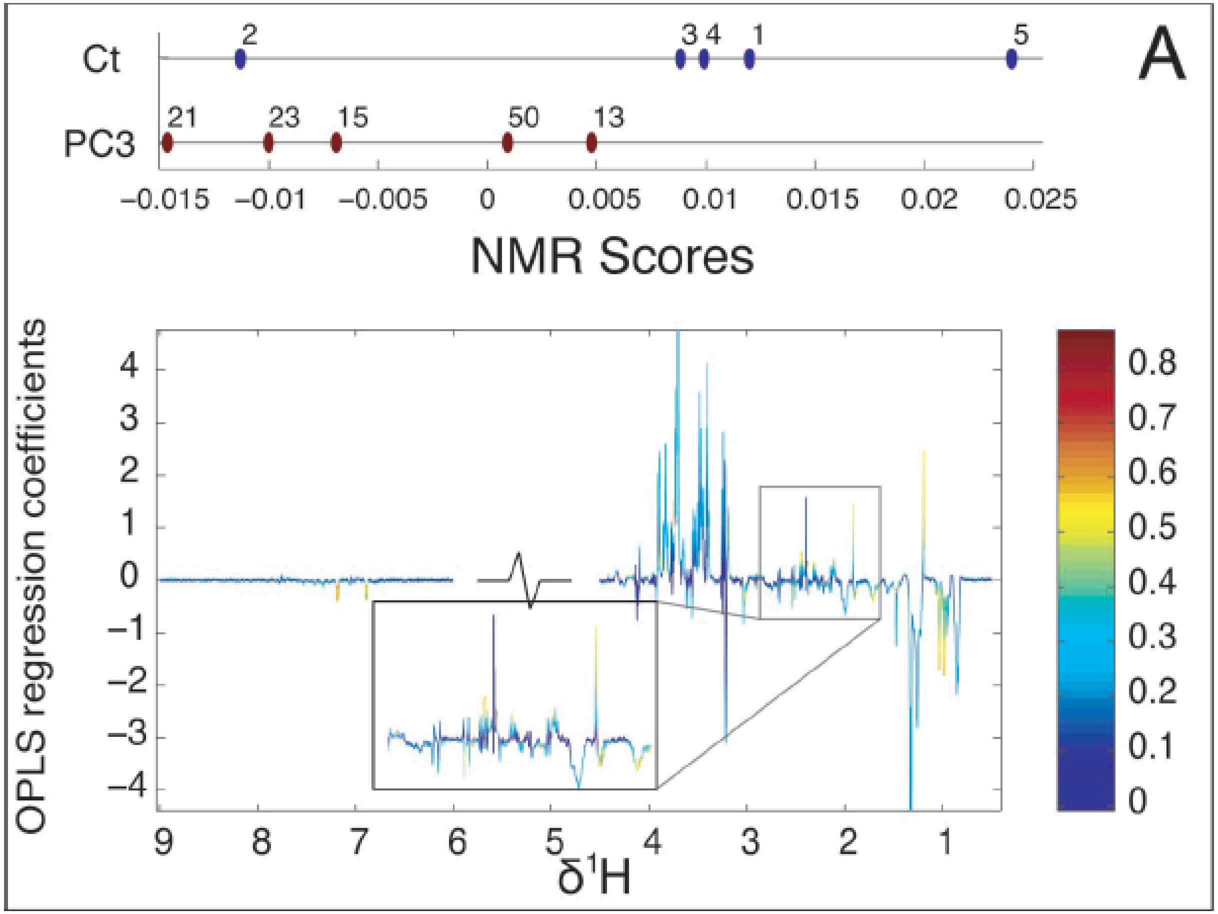
Analysis of each part of the model by O-PLS-DA or PCA modeling to elucidate information of interest in each sub-part of model.

O-PLS-DA / PCA
O-PLS-DA / PCA
O-PLS-DA / PCA
O-PLS-DA / PCA

All models validated by 5-fold cross validation

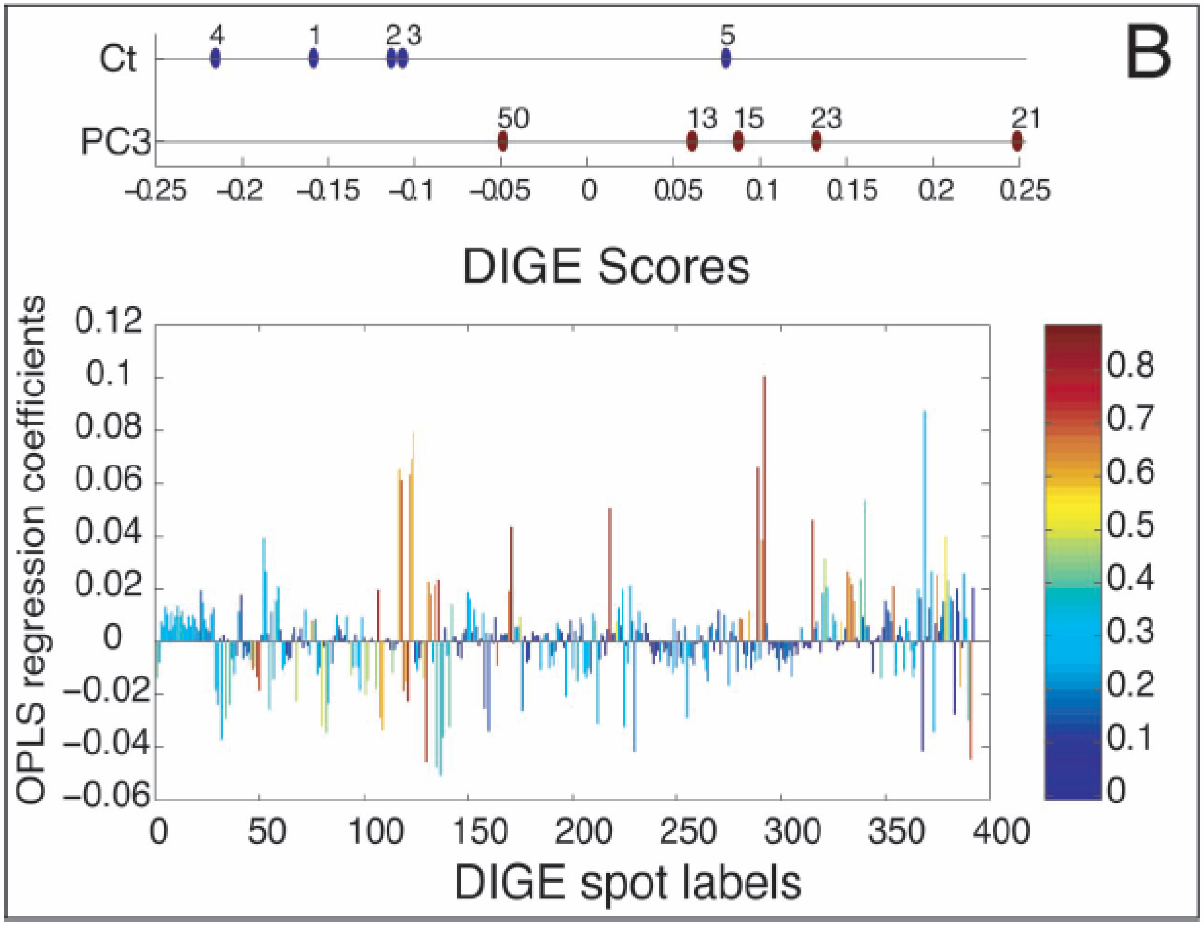
Results

OPLS of NMR data



Metabolites that changed the most between the groups:
valine
isoleucine
glutamine
leucine
lysine
tyrosine
phenylalanine,
glucose
3-D hydroxybutyrate
and acetate

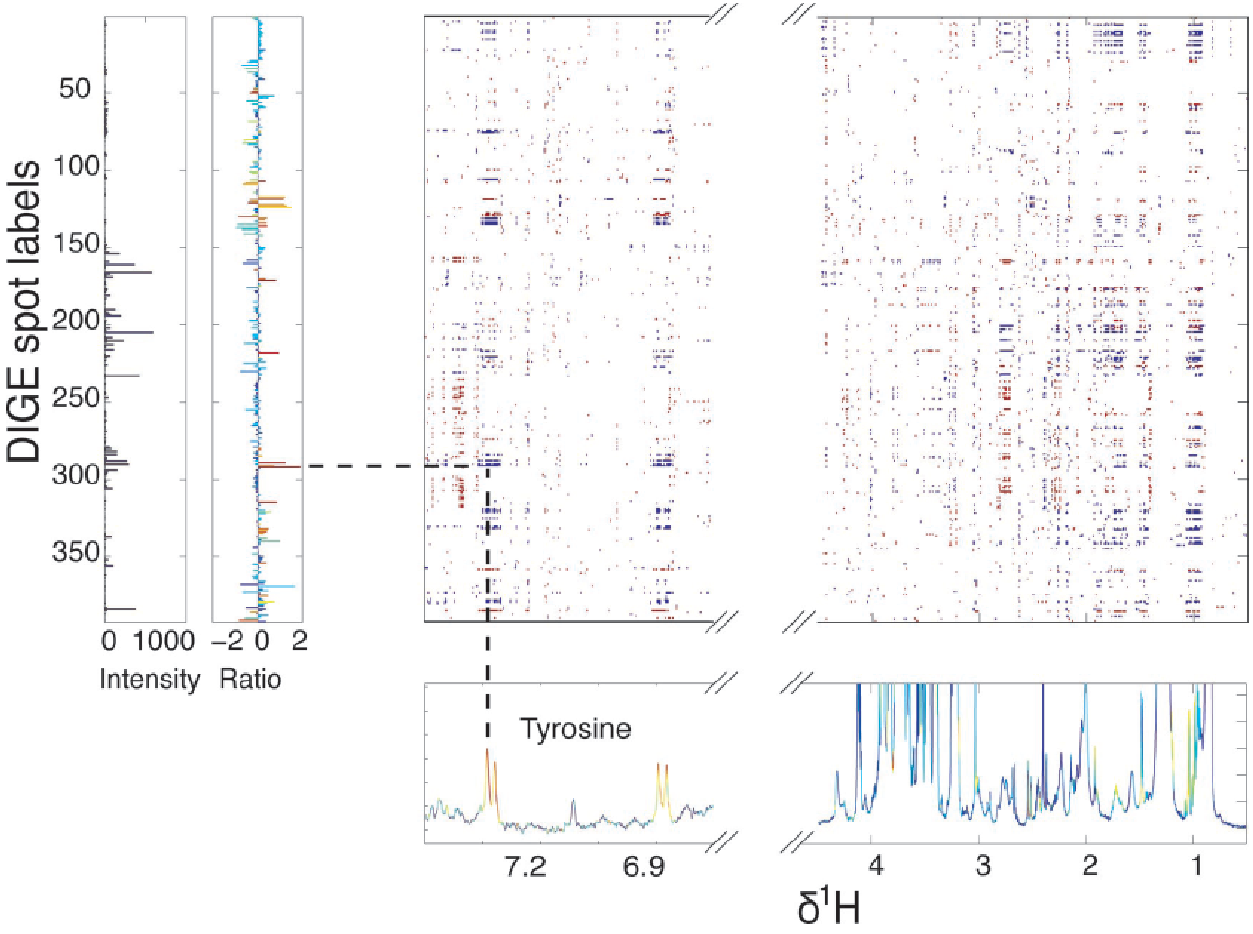
OPLS of 2D Gel data



Several proteins differentially expressed between the groups, including gelsolin and serototransferrin precursor, however, many of the proteins were not identified

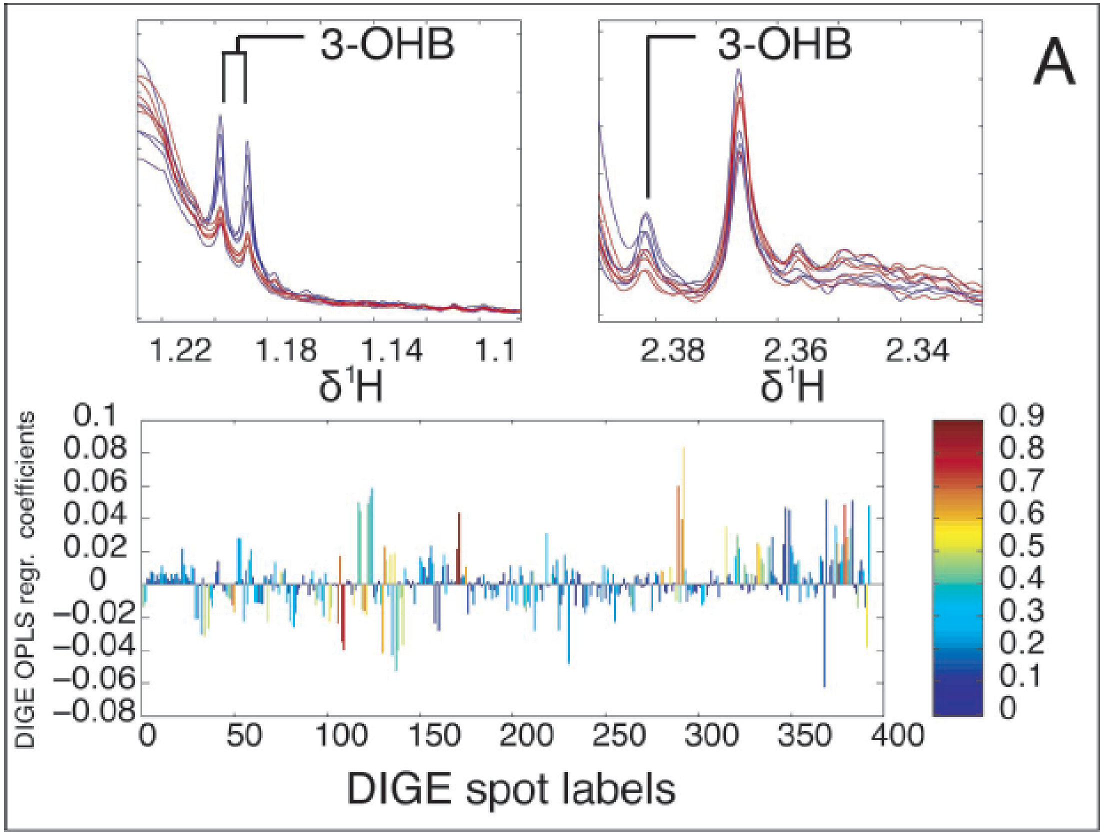
Correlation patterns between ^{1}H NMR and 2D Gel data

Correlation map:



Correlation patterns between ^{1}H NMR and 2D Gel data

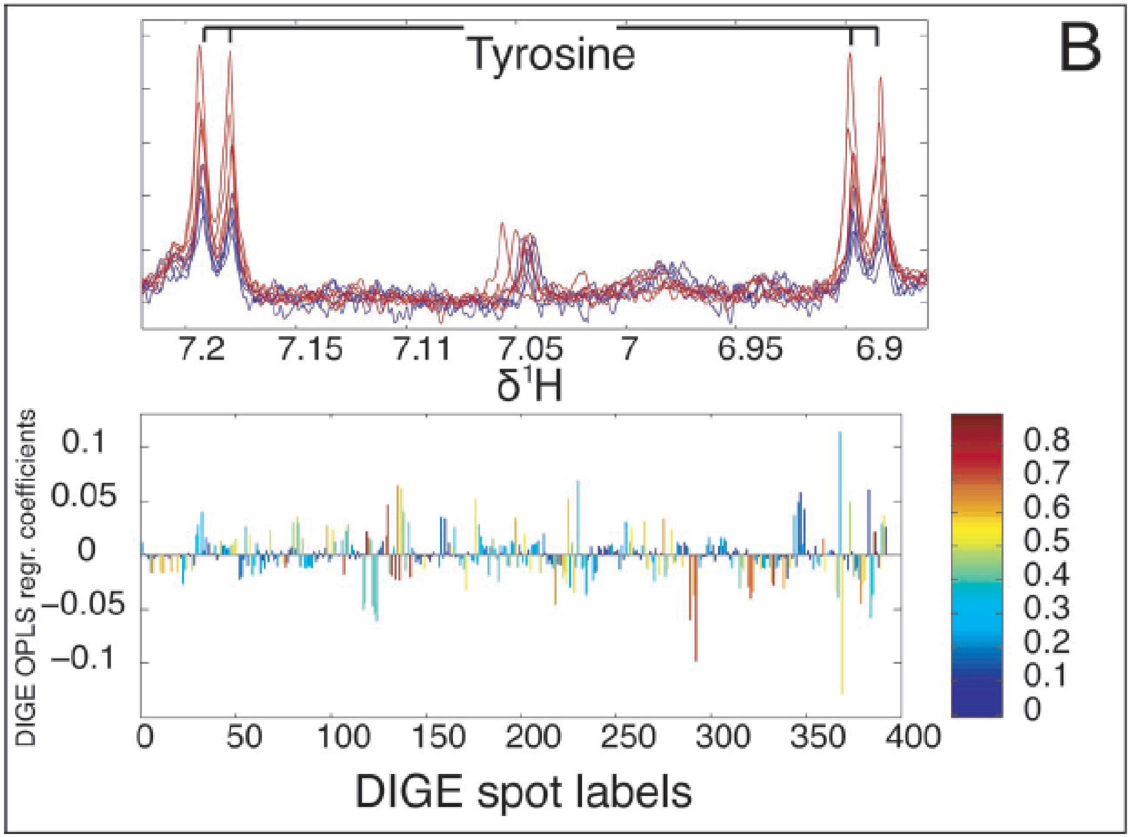
OPLS model between 2D Gel data and 3-D-hydroxybutyrate peaks:



Links, e.g., to serotransferrin precursor

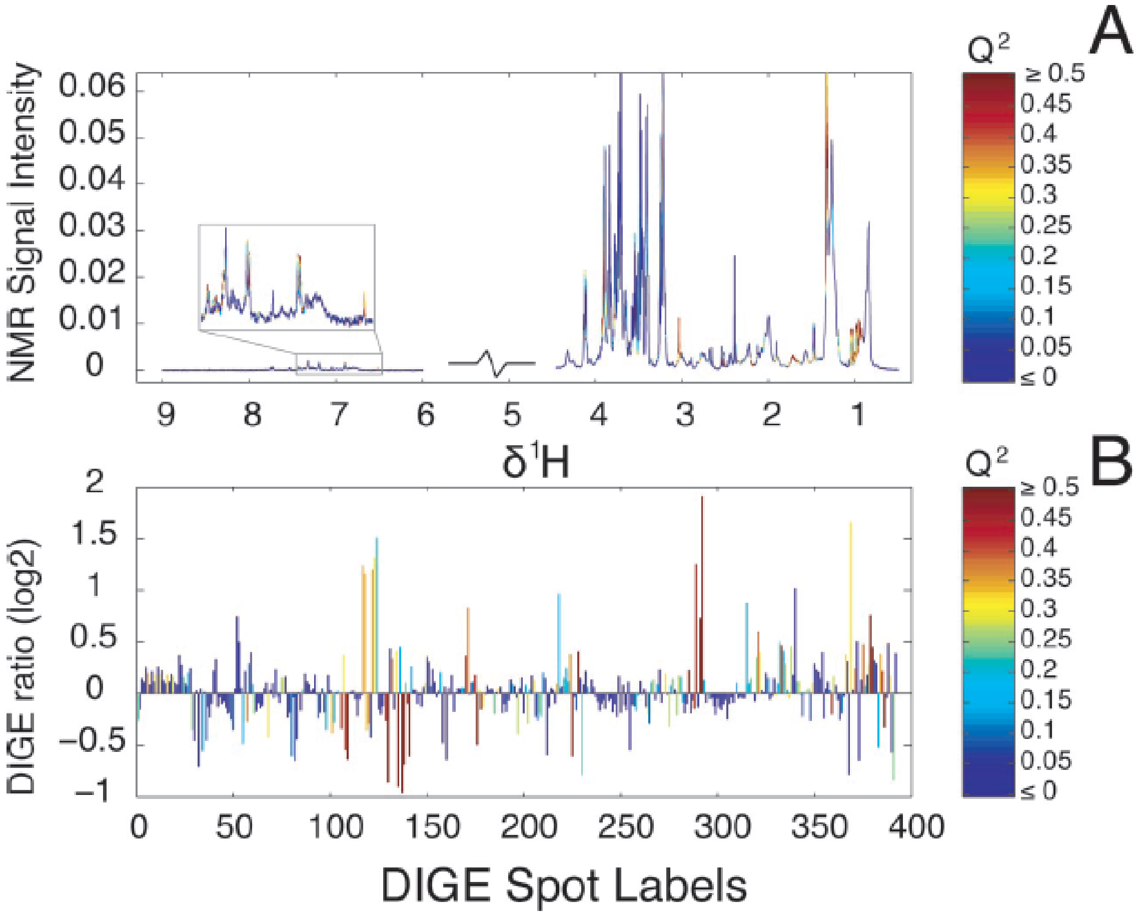
Correlation patterns between ^{1}H NMR and 2D Gel data

OPLS model between 2D Gel data and tyrosine peaks:

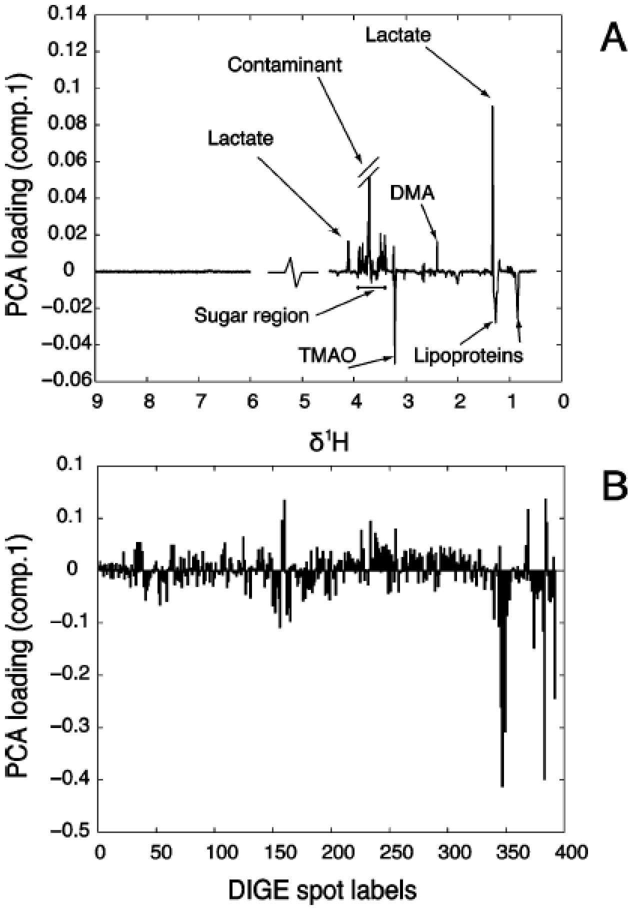


Links, e.g., to fibrinogen and gelsolin

Integration of ^{1}H NMR data and 2D Gel data using O2PLS



Analysis of the orthogonal and residual data by PCA



Discussion

- Methodological advances
  - First study to show that it is possible to statistically integrate proteomic and metabonomic data using OPLS
  - Method suitable for integration of all types of (omic) data
  - Cross-validation applied to the models allows the estimate the predictive ability of the models and thus ensures that the models are not over-fitted

- Biological advances
  - Clear differences between plasma metabolites and proteins between tumor transplanted animals and controls
  - Increased amounts of 3-D-hydroxybutyrate related to increased energy metabolism in the tumor?

Comments

- Methodological advances likely greater than the biological advances
- Very limited data set

- Had the animals fasted before blood plasma collection?
- Why was the 1D NMR data not used in combination with CPMG NMR data?
- Does this approach solve the problem of mammalian biocomplexity?

Summary

- Combining data from different omics platforms essential for better understanding of biological processess
- OPLS and O2PLS provide good means for integrating metabonomic and proteomic data, but the methods can be also applied for other types of (omics) data
- Variance described by the orthogonal components, i.e., systematic variation not related to the class, may be important and further exploited

Exercises

1. What are the benefits of OPLS and O2PLS compared to PLS? Are there any downsides in using these analysis methods?
2. Name at least one reason why MS would be a better tool for metabonomics than NMR.
3. What kind of (biological) difficulties there are in combining data from different omics platforms?

References

- Rantalainen et. al.: Statistically Integrated Metabonomic-Proteomic Studies on a Human Prostate Cancer Xenograft Model in Mice
- Nicholson et. al.: The Challenges of Modeling Mammalian Biocomplexity
- Trygg & Wold: Orthogonal projections to latent structures (O-PLS)
- Trygg: O2-PLs for qualitative and quantitative analysis in multivariate calibration
- Rosipal & Krämer: Overview and Recent Advances in Partial Least Squares
- Westerhuis et. al.: Assessment of PLSDA cross validation