A view of metabolomics from a chemometrics perspective

Rui CLIMACO PINTO

with

Johan TRYGG

- Umeå University – Sweden -
Overview

- Umeå chemometrics/bioinformatics group CLiC
- Metabolomics
- Integration of chemometrics in metabolomics
- Multivariate regression / Discriminant analysis
- OPLS and O2PLS framework
- Examples of chemometrics in metabolomics
Umeå, Sweden

University built in 1965
25 000 students / 4000 staff
Collaboration: Umeå Plant Science Center
Excellence centre in plant biology

- Hormone laboratory
- Transformation and tissue culture
- DNA sequencing & array facility
- Coffee room
- Separation
- Metabolomics
- Proteomics
- Chemometrics group
- New full security transgene facility
- New greenhouses for transgenic Poplar
- FT-IR
- NIR
- NMR
UMEÅ – CLiC

Objectives:
- Stimulate, organize and advance computer based modelling, tools and strategies to understand complex biological systems and e-bioscience
- Be the critical and missing link to the ongoing strong experimental research at Umeå University (UPSC, UCFB, FuncFiber, MIMS, UCMR and UCMM centers)
- Establish a unique bioinformatics/e-science profile in Umeå

Main research areas:
- Omics-technologies (mainly transcriptomics, proteomics, metabolomics)
- Network modeling, databases and visualization
- Structural biology and sequence analysis

Group leaders:
Antti, Henrik (Assoc. Prof.) - Predictive and Human Metabolomics
Hedenström, Mattias (Ass.Prof.) - Characterization of plant material and biofluids using NMR spectroscopy
Hvidsten, Torgeir (Ass.Prof) - A systems biology approach to model the transcriptional network in trees
Linusson Jonsson, Anna (Ass.Prof) - Probing molecular interactions of protein-ligand complexes guided by an integration of chemometrics and molecular modelling
Rydén, Patrik (Ass.Prof.) - Pathogenicity of Francisella tularensis
Sauer, Uwe (Assoc.Prof) - BioCrystallography and BioInformatics
Sjöström, Michael (Prof.) - Multivariate quantitative structure activity relationships (M-QSAR)
Stenberg, Per (Ass.Prof) - Mining functional DNA elements in eukaryotic genomes
Trygg, Johan (Assoc.Prof) - Chemometrics in metabolomics, ‘omics profiling and systems biology
Trygg group’s chemometrics in ’Bio-’

**Tree biology:** Functional genomics in transgenic Poplar trees
- Umeå Plant Science Center

**Disease diagnosis & biomarker identification**
- Rheumatoid Arthritis, Diabetes 1 & 2, Huntington, etc...

**Medicine (Post operative surgery):** Kidney transplant
- Monitor immune suppression vs toxicity with NMR spectroscopy

**Dietary:** Functional foods
- Health effect from food supplement with NMR & GC-MS spectroscopy

**Medical imaging by ultrasound**
- Study muscle tissue physiology and function in rehabilitation
Metabolomics
Metabolomics - definitions

Supporting thesis: Functional status of a complex biological system resides in the quantitative and qualitative pattern of metabolites in body fluids

- **Metabolome** - Complete set of metabolites to be found within a biological sample
- **Metabolite**
  - Small biological molecules, intermediates and products of metabolism
  - Primary: main functions (growth, development, reproduction)
  - Secondary: ecological function (ex. antibiotics and pigments)

- **Metabolomics** – systematic study of the unique chemical fingerprints that specific cellular processes leave behind (MS)
- **Metabonomics** - quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification (NMR)
Metabolomics

- Instrumental analysis
  - Mainly GC-MS, LC-MS, NMR
  - Also Raman, FTIR
  - Large amounts of data
- Use of chemometrics
- Disease diagnosis, functional genomics, toxicology, plant science, nutrition, pharmaceutical and environmental research, personalized medicine
- Today - trend in biological interpretation rather than only classify samples
Personalised medicine

- Personalized medicine – refine the empirical approach used in most clinical trials by incorporating powerful new diagnostics that can identify individual predictive characteristics and better control variability

Metabolomics in personalised medicine

- Drug metabolism pathways
- Definition of disease subsets
- Definition of groups of patients
- **Monitoring treatment response**
- Prevention
- Drug safety

J. Woodcock, Clinical Pharmacology & Therapeutics, 81 (2007) 164
### Metabolomics – steps

<table>
<thead>
<tr>
<th>Study design</th>
<th>Mass Spectrometry</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Problem formulation</td>
<td>(Extraction) Analysis</td>
</tr>
<tr>
<td></td>
<td>Study design</td>
<td>Analysis</td>
</tr>
<tr>
<td></td>
<td>Sample collection</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Mass Spectrometry</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extraction</td>
<td>Phasing + Baseline correction</td>
</tr>
<tr>
<td></td>
<td>Derivatization</td>
<td>Alignment</td>
</tr>
<tr>
<td></td>
<td>Analysis</td>
<td>Normalization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data processing</th>
<th>Mass Spectrometry</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td></td>
<td>Phasing + Baseline correction</td>
</tr>
<tr>
<td>Baseline correction</td>
<td></td>
<td>Alignment</td>
</tr>
<tr>
<td>Peak-picking or deconvolution</td>
<td></td>
<td>Normalization</td>
</tr>
<tr>
<td>Peak identification</td>
<td></td>
<td>Bucketing / peak-picking / deconvolution</td>
</tr>
<tr>
<td>Normalization</td>
<td></td>
<td>Scaling</td>
</tr>
<tr>
<td>Scaling</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical analysis</th>
<th>Mass Spectrometry</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataoverview (e.g. PCA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modelbuilding (OPLS, PLS, NN or other)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model optimization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model validation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictions and identification of biomarkers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biological interpretation</th>
<th>Mass Spectrometry</th>
<th>NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of perturbed metabolic pathways</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanistic explanation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow up studies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Biological samples

Biochemical analysis of endogenous metabolites

Data

Challenge in modern biology: maximizing information
GC-TOF/MS-based metabolomics platform

Sample preparation

- Silylation: 30 ul MSTFA + 1h
  N-Methyl-N-trimethylsilyl trifluoroacetamide

- Methoxymation:
  vortex mix 10 min + 16h RT
  30 ul methoxyamine/pyridine

- Speed Vac Concentrator:
  200 ul supernatant

- Mix/ultrasonicate:
  vibration mill/centrifuge

- Add methanol:
  700 µl

- Add water:
  200 µl

- Human plasma:
  100 µl

Analysis + Data processing

- Agilent 6980 GC
- Agilent 7683 Autosampler

- Scans ~ 15000
- Masses ~ 750
- Samples ~ 50


NMR and GC / LC-MS methods - Umeå

- Trees
- Arthritis in human (>300) and rats (>200)
- Diabetes (2 mouse models, >200 samples)
- Huntington disease
- LC-MS methods for amino-acids in final phase of development. Adding compounds
- LC-MS for lipids and hormones in preparation
- Bacterian and human cell cultures for analysis in GC / LC
Integration of Chemometrics in metabolomics

- DOE, MVD
- PCA
- MCR
- OPLS (OPLS, O2PLS, OPLS-DA)
• **Chemometrics – reduced to a data modelling tool**
  – ANOVA- analysis of variance (hypothesis testing)
  – Overview of data (Principal component analysis)
  – Two class discrimination (PLS-DA, SIMCA)

• **Metabolomics – reduced to NMR/MS based technique**
  – ... with many interesting case studies, samples

• **Chemometrics + Metabonomics**
  – Samples + NMR/MS based characterisation + PCA/PLS-DA

—Is this enough?

Not many papers had been published...

...that aim for the whole chain of planning, sampling, experimental characterisation, modelling, visualisation and interpretation...

... especially, regarding validating the hypothesis made based on models.

Integration Chemometrics/Metabolomics
providing information for studying complex systems

1. **Define the aim**
   - What do we want?
   - What is known already / what more knowledge is needed?

2. **Selection of objects**
   - Design of Experiments (DOE)
     - Samples, time points, replicates...

3. **Sample preparation and characterisation**
   - Experimental protocol optimization
     - Extraction, derivatization, instruments parameters optimization...
     - Randomization of samples for GC/LC/NMR analysis by day, disease/control...
   - Data processing
     - Align peaks, correct baseline, curve resolution, normalisation, scaling

4. **Evaluation/Validation of collected data**
   - Exploratory analysis
   - Multivariate design
   - Interpretation & Visualization
   - Class-specific study
   - Dynamic study
1. Define the aim

– What do we want? Example for disease diagnostics:

<table>
<thead>
<tr>
<th></th>
<th>Metabolomics / metabonomics</th>
<th>Metabolic fingerprinting</th>
<th>Metabolite profiling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
<td>Comprehensive analysis with identification and quantification of as many metabolites as possible in a biological system, done in an unbiased way</td>
<td>Fast classification of samples based on metabolite data, without necessarily quantifying or identifying the individual metabolites.</td>
<td>Quantification of a number of pre-defined metabolites</td>
</tr>
<tr>
<td><strong>Potential use</strong></td>
<td>Diagnosis + biomarker discovery + biological understanding</td>
<td>Diagnosis method</td>
<td>Diagnosis + biomarker discovery + biological understanding</td>
</tr>
</tbody>
</table>

– What is known already / what more knowledge is needed?

- Literature review
- Known biomarkers
- Other extraction procedures, solvents, instruments

2. Selection of objects

- Design of Experiments (DOE)

Dynamic studies
- Allow slow/fast responders
- Different sampling times

Define samples, repetitions

Reduce residual variability

Study design

<table>
<thead>
<tr>
<th>Sampling period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>v1</td>
<td>v2</td>
<td>v3</td>
<td>v4</td>
<td>v5</td>
<td>v1</td>
<td>v2</td>
<td>v3</td>
</tr>
</tbody>
</table>

Time
DoE: Greenhouse design study

”biological variation”

- Experimental design
  - Initial conditions
  - Growth conditions
  - Position in greenhouse
  - Harvesting conditions
  - Grinding / Storage
  - Sample preparation

Greenhouse overview

Observed vs Predicted height (cm)

Variable influence

Results
3. Sample preparation and characterization

3.1. Experimental protocol optimization

- Solvents for extraction, derivatization, instruments parameters optimization...
- Randomization of samples for GC/LC/NMR analysis by day, disease/control...

Solvent DoE

<table>
<thead>
<tr>
<th>ID no</th>
<th>run order</th>
<th>methanol</th>
<th>ethanol</th>
<th>acetonitrile</th>
<th>acetone</th>
<th>chloroform</th>
<th>plasma</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1‡</td>
<td>20, 14, 18</td>
<td>800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N2</td>
<td>6</td>
<td>0</td>
<td>800</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N3</td>
<td>31</td>
<td>0</td>
<td>800</td>
<td>800</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N4</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>800</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N5</td>
<td>24</td>
<td>600</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>N6</td>
<td>18</td>
<td>0</td>
<td>600</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>N7‡</td>
<td>3</td>
<td>0</td>
<td>600</td>
<td>0</td>
<td>0</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>N8‡</td>
<td>19, 12, 21</td>
<td>0</td>
<td>0</td>
<td>600</td>
<td>0</td>
<td>200</td>
<td>0</td>
</tr>
<tr>
<td>N9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>735</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>N10</td>
<td>23</td>
<td>0</td>
<td>535</td>
<td>0</td>
<td>0</td>
<td>265</td>
<td>0</td>
</tr>
<tr>
<td>N11</td>
<td>15</td>
<td>0</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>535</td>
<td>0</td>
</tr>
<tr>
<td>N12</td>
<td>30</td>
<td>0</td>
<td>535</td>
<td>0</td>
<td>0</td>
<td>265</td>
<td>0</td>
</tr>
<tr>
<td>N13</td>
<td>25</td>
<td>0</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>535</td>
<td>0</td>
</tr>
<tr>
<td>N14</td>
<td>16</td>
<td>0</td>
<td>535</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N15</td>
<td>27</td>
<td>0</td>
<td>265</td>
<td>535</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N16‡</td>
<td>29, 9, 17</td>
<td>665</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>135</td>
</tr>
<tr>
<td>N17</td>
<td>4</td>
<td>535</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>265</td>
<td>0</td>
</tr>
<tr>
<td>N18</td>
<td>5</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>535</td>
<td>0</td>
</tr>
<tr>
<td>N19</td>
<td>13</td>
<td>535</td>
<td>0</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N20</td>
<td>22</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N21</td>
<td>32</td>
<td>535</td>
<td>265</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N22</td>
<td>8</td>
<td>265</td>
<td>535</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N23</td>
<td>7</td>
<td>0</td>
<td>235</td>
<td>235</td>
<td>235</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>N24‡</td>
<td>26, 2, 33, 10</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
3. Sample preparation and characterization

3.1. Experimental protocol optimization

- Solvents for extraction, derivatization, instruments parameters optimization...
- Randomization of samples for GC/LC/NMR analysis by day, disease/control...

Derivatization DoE

<table>
<thead>
<tr>
<th>expt no.</th>
<th>methanol, μL</th>
<th>incubationa °C, min</th>
<th>extraction min</th>
<th>incubationb °C, min</th>
<th>oximation °C, h</th>
<th>silylation °C, h</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>700</td>
<td>0, 10</td>
<td>1</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N2</td>
<td>700</td>
<td>70, 30</td>
<td>1</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N3</td>
<td>700</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N4</td>
<td>700</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N5</td>
<td>700</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N6</td>
<td>700</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N7</td>
<td>700</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N8</td>
<td>700</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N9</td>
<td>700</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N10</td>
<td>900</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N11</td>
<td>900</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N12</td>
<td>900</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N13</td>
<td>900</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N14</td>
<td>900</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N15</td>
<td>900</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N16</td>
<td>900</td>
<td>70, 30</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N17</td>
<td>900</td>
<td>0, 10</td>
<td>3</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N18</td>
<td>800</td>
<td>0, 10</td>
<td>2</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N19</td>
<td>800</td>
<td>0, 10</td>
<td>2</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
<tr>
<td>N20</td>
<td>800</td>
<td>0, 10</td>
<td>2</td>
<td>0, 10</td>
<td>20, 16</td>
<td>20, 1</td>
</tr>
</tbody>
</table>

*a Temperature and duration before extraction. *b Temperature and duration after extraction.
3. Sample preparation and characterization

3.2. Data processing
- Align peaks by a reference spectrum
- Region selection
- Baseline correction
- Normalisation
- Scaling
- Multivariate curve resolution (ex: GC-MS)
Data pre-processing
Methods in GC-MS, LC-MS, NMR

• Baseline correction
• Alignment
• Time-window setting (GC-MS, LC-MS)
• MCR
Multivariate curve resolution

resolve hyphenated data into chromatographic and spectral profiles.

Solves: \( X = C \cdot S^T + E \)

- \( C \) = Chromatographic profiles
- \( S \) = Spectroscopic profiles

\( X = C \cdot S^T = C_1 \cdot S_1^T + C_2 \cdot S_2^T + \ldots + C_n \cdot S_n^T + E \)
4. Evaluation/validation of collected data

- Overview of data
- Exploratory analysis
- Multivariate design
- Class-specific study
- Dynamic study
- Visualization
- Interpretation
Principal Component Analysis
Overview, outliers, groups, tendencies
Overview of data

GC/MS metabolite profiles

Samples

Wildtype

Transgenic

PCA
Overview & data exploration
Example: PCA on GC/MS spectra on human plasma

Outlier detection

- Drift / robustness

Two phase problem
Chloroform / Acetone

Tendences observed
PCA for Multivariate design

Example for choice of calibration and validation sets

Groupings in data
Select subset from each meaningful cluster

Selection from a database
Diverse selection
Multivariate method – Get results

Many different methods to choose from

**Linear methods**

- Full rank methods
  - Multiple Linear Regression (MLR)
  - Stepwise MLR
  - Ridge Regression

- Latent variable regression methods
  - Principal Component Regression (PCR)
  - **Partial Least Squares (PLS)**
  - **Orthogonal Projections to Latent Structures (OPLS)**

**Non-Linear methods**

- Neural Networks (NN)
- Support Vector Machines (SVM)
- Regression trees
Validation = f(Prediction, Interpretation)

- **Prediction** is part of the statistical validation, many tools exist
  - External predictions (RMSEP value), cross-validation
  - Many are familiar with these

  **Examples:**
  1. Predict concentration of active substance in tablet production with NIR spectroscopy
  2. Predict viscosity in pulp using NIR spectroscopy
  3. Predict severity of coronary heart disease (CHD) on biofluids with NMR
  4. Predict biological activity from amino acid sequence (QSAR)

- **Interpretation** is part of the chemical / biological validation (what does it mean?)
  - No direct quantifiable measure as RMSEP exists
  - Model interpretation (e.g. regression coefficients)
    - Pure constituent spectrum
    - "Sequence motif"
    - "Functional profile"
  - Not as common, requires much more effort (communication between disciplines)

  **Both are related & complementary in validating models/results**
**Validation in disease diagnostics**

<table>
<thead>
<tr>
<th>Statistical results valid from statistical point of view</th>
<th>Biological results are relevant to study</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Prediction of <strong>validation dataset</strong> (not CV).</td>
<td>- <strong>Identification</strong> of differentially regulated metabolites and their associated metabolic pathways.</td>
</tr>
<tr>
<td>- <strong>3 classes</strong>: Controls, disease and related disease control group.</td>
<td>- Establish whether the results are in accordance with known facts or are spurious, e.g. products of uncontrolled factors.</td>
</tr>
<tr>
<td>- Realistic measure for the error in the classification of new samples from the <strong>same patient population</strong>.</td>
<td></td>
</tr>
<tr>
<td>Will <strong>NOT</strong> guard against sampling bias nor drift in analytical instruments.</td>
<td></td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td><strong>Recommended</strong></td>
</tr>
<tr>
<td>- <strong>Follow-up study</strong> in a separate population, analyzed separately in a <strong>different lab</strong>.</td>
<td>- <strong>Follow-up study</strong> in a separate population analyzed separately in a <strong>different lab</strong>.</td>
</tr>
<tr>
<td>- Realistic measure of the expected error in classification of <strong>new patients</strong>.</td>
<td>- <strong>Only reliable way</strong> to reveal whether the observed metabolic perturbations are in fact a <strong>product of the investigated disease</strong>, or a product of sample bias.</td>
</tr>
<tr>
<td>- <strong>Guard against sampling bias and drift</strong> in analytical instruments.</td>
<td></td>
</tr>
</tbody>
</table>

Multivariate calibration
Discriminant analysis / classification
Multivariate calibration, MC
Model the relation between two blocks of data

**Samples** - Powders, molecules, industrial process samples, plasma, tissue…

**Sample characterisation** - Spectrometers (NIR, UV, IR, NMR, MS), chromatography, chemical descriptors, gene-arrays, metabolites

- Focus modelling towards known information (concentration, groupings)
- Model the relation between blocks of data (same samples, different spectra)

**Linear prediction model:** \( y = Xb + f \)

**Focus:** How to solve for \( b \)?

**Objective:** Provide good fit to estimate \( y \), good predictions for future samples
Example: One component system

Spectral profile of Y-predictive component

X matrix
Example: Modeling 1-component model

- **PLS regression**
- **Ridge Regression**
- **Linear Neural Net**

**Observed vs Predicted**

**b coefficients**
But... Chemical / biological data are complex

- Lots of unknown systematic variation – mostly due to poor knowledge...
  - strong dietary, environmental, hormonal variations, etc...
  - Experimental variation, sampling, instrumental variation
  - Input material varies with supplier

- Measured signal is the sum of many contributing factors
  - Pharmaceutical tablet formulation (e.g. binders, fillers, active drug, lubricant)
  - Human urine sample (e.g. genetics, diet, gender, age, stress, disease)
  - Plant biotech / Pulp & paper (e.g. wood species, cellulose & lignin content, water, age)
  - In QSAR the molecular descriptor profile is a function of its chemical and biological property/activity/function
Example: simulation with two component system (overlap)

Spectral profile of Y-predictive component

Spectral profile of Y-orthogonal component

X matrix

PLS

$y_1 \perp y_2$

$y_1$

$y_2$

$y_1$

$y_2$
Example: Two Gaussian peaks
Model interpretation by coefficient profile

PLS regression

Ridge Regression

Linear Neural Net

Negative dips observed!
PLS - Regression coefficients $[b_1 \ b_2 \ ...]$, one for each Y-variable

what do they mean?

$y_1 = Xb_1 + f_1$

1. The regression coefficient vector $b$ does not represent the estimated pure constituent spectrum

2. Its profile must be **orthogonal to all other known and unknown** constituents in $X$
   (Otherwise it will not be good for prediction)

Model overview

PLS, MLR, PCR, RR etc...
PLS NIPALS (1980’s)
Wold, Martens and colleagues

\[ X = TP' + E \]
\[ y = Tc' + f \]
PLS model

Example: Single-Y, two component system

94% variation 3.7% variation

Regression coefficients, $b$

$w_1, w^*_1$

$w^*_2$

$p_1$

$p_2$
What to do, and interpret?

1. Use preprocessing filters
   - MSC, SNV, 1,2nd derivatives, wavelet, Fourier, etc
     • Can remove pertinent information, loadings...

2. Avoid this variation
   - Improve instrument, sample preparation, and so on ...
     • Requires much knowledge, often not realistic

3. Why not ...
   Separately model the Y-predictive and Y-orthogonal variation?
   • Understand what’s going on!!
   • Orthogonal signal correction method [Wold S et al. 1998]
   • OPLS method [Trygg J & Wold S. 2002]
The O-PLS framework
Orthogonal Signal Correction (OSC)

OSC, Wold et al. (1998), Sjöblom et al. (1998), DOSC, Westerhuis et al. (2001)

• Basic idea, perform an ”inverse PLS model”:
  Remove structured noise (i.e. systematic) from $X$ not correlated to $Y$

$$X = t_{osc} p_{osc}^T + X_E$$

Estimate calibration model (e.g. PLS) based on the filtered $X_E$
Y-Orthogonal variation, what is it?

”Impact of nothingness” – Gottfries et al.

For example...

- Experimental problems
- Side reactions causing biproducts
- Non-linearities (e.g. kinetics)
- Within class variation
- Sampling issues
- and so on...

Orthogonal PLS (OPLS)

Focus modelling towards known information

\[ X = t_p p'_p + T_o P_o + E \]
\[ y = u_p c_p' + f \]

- Only a single Y-related component
- Used for two-class discriminant analysis

Trygg J.; Wold, S.; Orthogonal projections to latent structures (O-PLS), Journal of Chemometrics, 2002, 16, 119-128
Multi-block modeling

• Compare & Integrate X and Y in terms of....
  – Analytical platforms, Experimental conditions, Process step, Time (drift), Replication, Pre-treatments, ...

• Understand...
  – Overlap? What is jointly related?
  – What is unique for X, for Y?

Transcriptomics

O2PLS modelling

Metabolomics
Two block modeling
The O2-PLS model

Y-orthogonal variation

\[ T_o P^T_o \]

\[ X \]

\[ T_p \leftrightarrow U_p \]

\[ Y \]

\[ P_p T \]

Symmetric/ ~ PCA comp

\[ C_p T \]

X/Y joint variation

\[ X \rightarrow \]

\[ U_o C^T_o \]

X-orthogonal variation


PLS modeling vs OPLS modeling

PLS, MLR, PCR, RR etc...

- Mixes Y-orthogonal and Y-predictive variation
- Uni-directional, Models Y FROM X

OPLS

- Separates Orthogonal and Predictive variation
  (e.g. ‘between block’ from ‘within block’)
- Bi-directional, Models X AND Y
- Only uses predictive variation for modeling Y
Benefits of OPLS modeling

✓ Model diagnostics:
  – $R^2(XY)$: How much variation in X is correlated to Y, and vice versa?
  – $R^2(X_{yo})$: How much is not correlated to Y? (to X?)

✓ Model interpretation
  – More focussed components (plots) & easier interpretation
    • Predictive components ($T_pP_p^T$)
    • $Y$-orthogonal components ($T_oP_o^T$)
  – Pure profile estimation

✓ Model (prediction):
  – Understand & correct for faults/mistakes found in $Y$-orthogonal components
  – e.g. experimental, sampling

• Multi-block modeling ($X \leftrightarrow Y$)
  – Integrate, compare and filter multiple data tables
OPLS model

Example: Single-Y, two component system

Scores plot

Y-orthogonal

Predictive

Loading plot

$p_1c_1$

$p_2c_2$

$p_1$

$p_2$

49% variation

Predictive profile

Y-orthogonal profile

Predictive profile

$R^2X[1] = 0.492085$

$R^2X[2] = 0.489151$
OPLS model
Example: Two component system, where unknown variation is correlated to known $y$

Spectral profile of Y-predictive component

Spectral profile of Y-orthogonal component

$y_1$ and $y_2$ have 0.7 correlation

X matrix

Single $y$

$y_1$ and $y_2$

PLS

OPLS
Example: Two component system, where unknown variation is strongly correlated to known $y$.

**Difficult to relate PLS loadings to the variation it represents**
OPLS

Example: Two component system, where unknown variation is correlated to known y

X matrix

Spectral profile of Y-predictive component

Spectral profile of Y-orthogonal component

y1 and y2 have 0.7 correlation

Use 2 separate OPLS models?

OR

Use 1 OPLS model with multi-y?

Multiple y
Single-Y vs multi-Y OPLS models

Two single-Y OPLS models

- y1: 84% variation
- y2: 15% variation

Multi-Y OPLS regression

\[ K = B_p (B_p^T B_p)^{-1} \]

OPLS as a filter

Example: Calibration transfer of near infrared spectra

- Instrument A, B used to measure NIR spectra of an active pharma compound
- 15 batches specially selected to cover a variation of the water content
- A reference spectrum measured every second
- Water content varied from 1.38 to 4.47 wt./wt.% (Karl–Fischer titration)
- \( Y = \text{class } (-1,1) \ [\text{Instrument A vs Instrument B}] \)

**Wood’s anomaly** (diffraction gratings)

**OPLS as a filter**

Example: Calibration transfer of near infrared spectra

**PLS-DA**

**OPLS-DA**

Water region

\[ R^2_{X[1]} = 0.861133 \quad R^2_{X[2]} = 0.0625766 \]

\[ R^2_{X[1]} = 0.121249 \quad R^2_{X[2]} = 0.802461 \]
Example PAT: Binary powder

- Diffuse reflectance NIR spectroscopy
- Mixture of two powders with markedly different particle size
- 11 batches of powders, 0% to 100% in steps of 10%.

- \( X = \) NIR spectra (SNV) in the range 1080-2025 nm
- \( Y = \) % binary mix of powders

Figure: Schematic overview of the vertical cone mixer and the fibre-optic probe set-up.

R\(^2\)[1] = 0,983807  R\(^2\)[2] = 0,0137598  Ellipse: Hotelling T\(^2\) (0,95)
R\(^2\)[1] = 0,901427  R\(^2\)[2] = 0,0662814  Ellipse: Hotelling T\(^2\) (0,95)
Example PAT: Binary powder
Non-linearities transparent in OPLS loading profiles

PLS loading profiles (p)

OPLS loading profiles (p)
OPLS-derived methods

- Bifocal OPLS (BIF-OPLS)
- Kernel OPLS
- Multi-block modeling OPLS
Three blocks of data (X/Y/Z)
BIF-OPLS
(near publication)
Non-linear modeling techniques
 Kernel-OPLS

- There are situations where linear modeling techniques are insufficient
  - Biological and chemical systems, image analysis, etc.

- Many alternatives exist for prediction and classification
  - Artificial neural networks (ANNs)
  - Bayesian networks
  - Support Vector Machines (SVMs)
  - Kernel-based Partial Least Squares (KPLS)

- **K-OPLS**
  - Benefits are related to the interpretation of $Y$-predictive and $Y$-orthogonal scores
  - Not possible with KPLS or SVMs

Kernel-based methods

- Kernel-based methods utilize $\Phi(\mathbf{X})$ instead of $\mathbf{X}$ to predict $\mathbf{Y}$
- The function $\Phi(\cdot)$ extends $\mathbf{X}$ into a high-dimensional space (feature space)
- In this higher-dimensional space, a linear model is used for regression or classification
- The model is non-linear in the original space

$$K_{i,j} = \kappa(x_i, x_j)$$

Multi-block modeling OPLS
(in development)
Visualisation of OPLS model
STOCSY & S-plot: correlation and covariation combined into one plot

STOCSY (NMR)
Cloarec et al.

S-plot (NMR, MS, etc...)
Wiklund et al.

S-plot, Human Control vs Ra

Line plot

Scatter plot
Covariation and correlation

• **Covariation** is the measure of how much two variables vary together (strength)
  - Covariation is scale dependant (i.e. dependant upon the size of variability of the two variables)
  - Can hold positive, 0, and negative values

\[
\text{Cov} (t, y) = [(t)^T(y)] / (N-1)
\]

• **Correlation** = Fit is a dimensionless measure of covariation
  - Correlation is scale invariant (i.e. not dependant upon the size of variability of the two variables)
  - Can hold values between -1 to +1

\[
\text{Corr} (t, y) = [\text{Cov} (t, y) / (\|t\| \|y\|)] (N-1)
\]
Understand the most influential metabolites related to class separation

→ S-plot of the OPLS predictive component

Correlation
\[(p_{1p cpr})\]

S-plot

Confidence interval of variables based on Jack-knifing estimation

Covariation \[(p_{1p})\]
Understand the most influential metabolites (putative) NOT CORRELATED to class separation

⇒ S-plot of the OPLS orthogonal component
Examples
2-class separation OPLS
Disease diagnosis:
Rheumatoid Arthritis – brief background

• Worldwide prevalence of approximately 1%
• **Autoimmune disease**, the body attacks itself, aetiology largely unknown
• Treatment; irreversible disease, no known cure, medication to maintain mobility and ease pain
• Early diagnosis critical
  – More successful treatment with early medication
• Diagnosis for rheumatoid arthritis
  – Physical examination, antibodies (today not specific for RA), X-ray, MRI
• **New diagnostic tools are needed...**
Two class separation - Rheumatoid arthritis
Blood serum samples from 40 individuals (20 RA/20 Control)

**Within group dynamics**

Group separating direction
Specific metabolites for healthy and diseased
Rheumatoid arthritis: Control vs. RA
Understand biochemical differences

- Significant (subset) metabolites for separation of RA samples from healthy controls.
  - Variables represent endogenous metabolites

Up regulated in Ra

Down regulated in RA
RA: Comparison of the human case and animal models

- Great overlap of metabolites between humans and animals
  - Different metabolites show overlap in different animal models
  - Allows for identification of relevant animal models
  - Selection of model system for treatment studies

<table>
<thead>
<tr>
<th>BM</th>
<th>Human Rheumatoid Arthritis</th>
<th>Mouse Collagen Induced Arthritis</th>
<th>Rat Adjuvant Induced Arthritis</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC001</td>
<td>↑</td>
<td>na</td>
<td>Na</td>
</tr>
<tr>
<td>EC002</td>
<td>↑</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>EC003</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC004</td>
<td>↑</td>
<td>0/↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC005</td>
<td>↓</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>EC006</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC007</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC008</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>EC009</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC010</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC011</td>
<td>↓</td>
<td>0/↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC012</td>
<td>↓</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>EC013</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC014</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>EC015</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC016</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC017</td>
<td>0</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC018</td>
<td>↑</td>
<td>↑</td>
<td>↓/?</td>
</tr>
<tr>
<td>EC019</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC020</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC021</td>
<td>↑</td>
<td>↑/?</td>
<td>↑</td>
</tr>
<tr>
<td>EC022</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC023</td>
<td>0</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC024</td>
<td>↑</td>
<td>↓</td>
<td>0/↑</td>
</tr>
<tr>
<td>EC025</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC026</td>
<td>0/↑</td>
<td>↓</td>
<td>↑</td>
</tr>
</tbody>
</table>
RA: Comparison of therapies in animal model

- Metabolites levels are affected by administered therapeutics
  - New drug (X) restore levels in more metabolites compared to MTX*
  - Useful in development of novel drugs
  - Tool in clinical studies to verify therapeutic effect in clinical studies
  - Concomitant development of novel drug and diagnostic test, theranostics?

<table>
<thead>
<tr>
<th></th>
<th>Vehicle</th>
<th>MTX</th>
<th>X 1mg</th>
<th>X 3mg</th>
<th>X 10mg</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC004</td>
<td>0/↑</td>
<td>↓</td>
<td>↓</td>
<td>0/↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC006</td>
<td>0/↑/?</td>
<td>0/?</td>
<td>0</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC007</td>
<td>↓</td>
<td>0/↑</td>
<td>0/↑</td>
<td>0/↓</td>
<td>↑</td>
</tr>
<tr>
<td>EC009</td>
<td>0</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC010</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC011</td>
<td>0</td>
<td>0/↓</td>
<td>↓</td>
<td>0/↓</td>
<td>↑</td>
</tr>
<tr>
<td>EC012</td>
<td>0/↓</td>
<td>↑</td>
<td>0/↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC013</td>
<td>↑</td>
<td>0/↑</td>
<td>0/↓</td>
<td>↑</td>
<td>0/↑</td>
</tr>
<tr>
<td>EC014</td>
<td>↑</td>
<td>0/?</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC015</td>
<td>0/↑</td>
<td>↑</td>
<td>0/↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC016</td>
<td>0</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>EC017</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC018</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>0/↑</td>
<td>0/↑</td>
</tr>
<tr>
<td>EC019</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC022</td>
<td>↑</td>
<td>↑</td>
<td>0/↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>EC023</td>
<td>↓</td>
<td>0/↓</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>EC024</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC025</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>EC026</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

*MTX, methotrexate
Multi-class separation OPLS
OPLS in multi class metabolomics
Example: Plant metabolomics on Poplar

\textit{PttPME1} expression was up and down regulated in transgenic aspen trees
PME enzyme activity in wood forming tissues was correspondingly altered

Lines in this study
\textbf{WT} poplar
\textbf{2B} – up regulated \textit{PttPME1} gene
\textbf{5} - down regulated \textit{PttPME1} gene

\textbf{Metabolomics} study of xylem and phloem, here only the xylem results are presented.
OPLS-DA model of Line 5 vs Wildtype

Score plot

**OPLS model**
1 predictive component
3 orthogonal components
R²X(p)=12%
R²X(o)=20%
Q²Y=80%
R²Y=96%

Within class variation

Between class variation
Understand the most influential metabolites (putative) related to class separation (transgene vs wildtype) → **S-plot** of the OPLS predictive component

S-plot

Confidence interval of variables based on Jack-knifing estimation

Correlation \((p_{1p \text{corr}})\)

Line 5 vs WT

Covariation \((p_{1p})\)
Understand the most influential metabolites (putative)
NOT CORRELATED to class separation

Orthogonal S-plot

Line 5 vs WT
Orthogonal S-plot
Multiblock modeling - O2PLS
Combined profiling projects at UPSC

Transcriptomics
- sample preparation
- RNA extraction
- labelling, hybridization
- scanning, (Scanarray)
- data extraction (GenePix)
- normalization, statistical treatment (UPSC-BASE)
- visualization and interpretation of data (MAPMAN, GeneSpring)

Proteomics
- Preparation and fractionation of samples
- LC/MS analysis and data processing
- Selection of differential expressed peptides derived from multivariate analysis
- LC/MS/MS analysis and identification of proteins by database search

Metabolomics
- Calculation of relative levels of metabolites that differ between samples
- visualization and interpretation of data

LC/MS analysis and identification of proteins by database search
Combined profiling of transgenic Poplar

DOE (Genotype/Internode)
Combined profiling of transgenic Poplar

Transcript data

Protein data

Metabolite data

Joint variation Transcript

Joint variation Protein

Deflation of joint variation

Same for Protein and Metabolite data
Combined profiling of transgenic Poplar

A

Joint internode effect

Internode gradient

Joint score vector 1

Joint score vector 2

Joint genotype effect

Separation of G3

Separation of G5

Joint score vector 1

Joint score vector 3

Transcript data set

Protein data set

Metabolite data set

Unique

Residual

29.6%

18.9%

34.8%

30.0%

32.3%

35.3%

51.5%

35.2%

32.4%
A combined profiling study of *Populus tremula × P. tremuloides*, investigating short-day induced effects at transcript and metabolite levels

-24 hybrid aspen (*Populus tremula × P. tremuloides*) trees
- growth chamber under long day conditions (12 h of PAR light (400 µEin m-2 s-1) and a 6-h daylength extension with low light (30 µEin m-2 s-1).
LD0 - Long day samples
SD6 – Short day
Dynamic modeling
Dynamic modeling

- Biological systems are dynamic processes that react to changes in their environment at both the cellular and organism levels.
- Modeling the time-related behavior of biological systems is essential for understanding the biology and underlying dynamics.

Example: Functional foods study

- Functional foods: Foodstuffs with a documented health-promoting effect – besides energy addition
- Centre for Human Studies of Foodstuffs, Sweden
  - Inclusion/exclusion criteria
  - 9 individuals given prepared foodstuff
  - Multiple visits – document effect over time
Functional foods study:
Individual metabolism vs metabolic response to food intake

Individuals' metabolism baseline greater than the effect of foodstuffs
But... we are interested in the effect of foodstuffs
Dynamic (time-series) modeling

- In ‘omics (e.g. metabolic profiling) studies
  - the sampling rate and number of time points are often restricted (experimental, cost and biological constraints (< 4-15 time points)).
  - Chemometrics:
    - MSPC batch modeling (Antti et al)
    - ANOVA based modeling, e.g. ASCA (Smilde et al), ANOVA-PCA (Harrington et al)
    - Dynamic Bayesian networks (Kim et al)
    - Auto-regressive moving average (ARMA, Box et al)
    - SMART analysis (Keun et al)
    - Independent component analysis (Morgenthal et al)
    - PARAFAC (Forshed et al)

Existing strategies for modeling dynamic data rests on two major assumptions:
(1) The multivariate profile or fingerprint is comparable over all individuals.
(2) The global temporal behavior is aligned between all individuals.
Dynamic modeling

- Two alternative approaches using the OPLS model
  - Use OPLS property of single predictive components (+ Orthogonal components)
  1. Piece-wise dynamic modeling (Rantalainen et al)
  2. Dynamic modeling of individual effect profiles (Trygg et al)

![PCA model, individual 4](chart1)

![O-PLS model loading, individual 4](chart2)
Myo-inositol can have an effect on aminotransferase, supported by increase of ornithine, citrate and acetate. Myo-inositol has shown to have protective effect on cardiac dysfunction in diabetic rats.
Example: Dynamic modeling
Kidney transplant study
NMR profiles of human urine samples after surgery

1.) Principal component analysis (PCA) t1/t2 score

Post-operative time trajectory
Example: Dynamic modeling
Kidney transplant study
NMR profiles of human urine samples after surgery

2.) Collect ALL patient “recovery” profiles (Predictive OPLS component)

3.) PCA/SIMCA analysis of ALL patient “recovery” profiles (Predictive OPLS component)

4.) Interpretation
Concluding remarks

- Metabolomics has a promising future in different areas in the post-genomic era
- Chemometrics shall be used in all steps of the metabolomics pipelines
- New methods in chemometrics are needed to understand huge loads of information
- Multi-block modeling strategies needed
- OPLS approach is appropriate to model data from metabolomics
- O-PLS is a multivariate prediction method, similar to PLS,
  - separates two different types of variations in the modelled data
    - TpPp = X-Y related variation
    - ToPo = Y-Orthogonal variation in X (unique variation in X)
- Regression coefficient profile b should not be used for interpretation
- OPLS allows model diagnostics, prediction and interpretation
- Different strategies using OPLS/O2PLS are useful for different purposes
# Acknowledgements

<table>
<thead>
<tr>
<th>Chemistry dep, Umeå University</th>
<th>Umeå Plant Science Center, Umeå Univ, Sweden</th>
<th>Acure Pharma</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Bylesjö</td>
<td>T. Moritz</td>
<td>P. Lek</td>
</tr>
<tr>
<td>H. Stenlund</td>
<td>D. Eriksson</td>
<td>E. Seifert</td>
</tr>
<tr>
<td>R. Madsen</td>
<td>A. Johansson</td>
<td></td>
</tr>
<tr>
<td>M. Hedenström</td>
<td>A. Sjödin</td>
<td>AcureOmics</td>
</tr>
<tr>
<td>S. Wiklund</td>
<td>R. Nilsson</td>
<td>J. Gabrielsson</td>
</tr>
<tr>
<td>P. Jonsson</td>
<td>A Grönlund</td>
<td></td>
</tr>
<tr>
<td>H. Antti</td>
<td>S Jansson</td>
<td>Anamar Medical</td>
</tr>
<tr>
<td>Uppsala University</td>
<td>B Sundberg</td>
<td>G. Ekström</td>
</tr>
<tr>
<td>T. Lundstedt</td>
<td>G. Wingsle</td>
<td>SweTreeTechnologies</td>
</tr>
<tr>
<td>J. Olsson</td>
<td>J. Karlsson</td>
<td>M Hertzberg</td>
</tr>
<tr>
<td>Umeå University Hospital</td>
<td>V. Srivastava</td>
<td>K Johansson</td>
</tr>
<tr>
<td>SB. Rantapää,</td>
<td>R. Bahlerao</td>
<td>A Karlsson</td>
</tr>
<tr>
<td>GM Alenius</td>
<td>G. Sandberg</td>
<td>Umetrics AB,</td>
</tr>
<tr>
<td>Lund University &amp; MAS</td>
<td></td>
<td>E. Johansson</td>
</tr>
<tr>
<td>Å. Lernmark</td>
<td></td>
<td>L. Eriksson</td>
</tr>
<tr>
<td>L. Åkesson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imperial College</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J. Nicholson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. Holmes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Rantalainen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O. Cloarec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian (Univ. Siena and more)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Calderisi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Vivi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Tassini</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M Valensin,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Carmellini,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Cocchi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riken University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M. Kusano</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Swedish Foundation for Strategic Research (SSF)
Swedish Research Council
FORMAS FuncFibre
Knut & Alice Wallenberg Foundation
GlaxoSmithKline
AstraZeneca
MKS Umetrics