Introduction to Chemometrics

Torbjörn Lundstedt
KI 080609
Chemometrics

• Use of mathematical and statistical methods for selecting optimal experiments
  Statistical experimental design
  Design of Experiments (DoE)

• Extracting maximum amount of information when analysing multivariate (chemical) data
  E.g. classification, (process) monitoring, multivariate calibration, Quantitative Structure-Activity Relationships (QSAR)
Why perform experiments?

- Increase understanding of observed phenomenon(s)
- Identify what is important for influencing an investigated system
- Find experiments (compounds) with desired properties
- Make predictions about the outcome of new experiments
DoE – terminology

• **Experimental domain**
The experimental area studied, area where model is valid

• **Factors**
Controlled variables which can be varied independently and have an impact on the result in the experiments (“X-block”)

• **Independent variables**
Same as factors

• **Quantitative variables**
Continuous variables – Independent variables which can be adjusted to any value over a specified the range
DoE – terminology

• Qualitative or Discrete Variables
  Independent variables which describe non-continuous variation, e.g. type of solvent, cell medium A or B

• Responses
  Variables which are observed and a result from changing independent variables (“Y-block”)

• Dependent Variables
  Same as responses
DoE – terminology

- **Residuals**
  The difference between the observed response and the response predicted from the model

- **Uncontrolled or background variables**
  Known variables which are not possible/desirable to alter

- **Unknown variables**
  Currently unidentified variables
Aim of Modelling

• Present the result in a clear and interpretable way – graphics very useful!
• Extract as much information as possible from the experiments
• Provide a “correct” conclusion – validate!
• Indicate which new experiments to perform and the probable outcome of these
Model types

• Fundamental models
  (hard models, “global models”)

\[ E = mc^2 \quad y = y_0 e^{-kt} \quad U = IR \]

• Empirical models
  (soft models, “local models”)
  Taylor expansions (polynomials of different complexity)

\[ y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_{12} + e \]
Models

Mathematical Equation Describing a System

- Chemistry
- Biology
- Physics
- Economics
- etc.
Soft Modelling
Smaller Parts of the Universe is Modelled

A smaller experimental domain is investigated
Linear Model

\[ y = b_0 + b_1 x_1 + b_2 x_2 + e \]

\[ e = \]

The residual, the part of the data the model does not explain

Important for validating the model
Second Order Interaction Model

\[ y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + e \]

- \( b_{12} \) Interaction term, resulting from the effect of two variables.
- Skews the surface.
Quadratic Model

\[ y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + e \]

- \( b_{11} \) Quadratic term, resulting from the effect of one variable. Curves the surface.
Establishing the Model

• Generally a model is based on a set of experiments, where some output (response or responses) has been measured.

• In the experiments different factors, variables, are investigated at different levels, i.e. settings (e.g. temp., conc., logP, nos. C, etc.)
Limitation of Models

• Usually only local validity (soft models) (interpolation/extrapolation)

• **All models are wrong**
  ... but some are still very useful!

• How should the experiments be performed in order to gain as much information as possible?
Change One Separate factor at a Time

Examine $x_1$
Time

Examine $x_2$
Temperature

Optimum?
Best yield?
COST
Change One Separate factor at a Time

If there is an interaction ➔ Not the optimum!
Statistical Experimental Design

Change all variables at the same time
– Find Directions –
Statistical Experimental Design

• Planning of experiments to perform in order to extract as much information as possible with as few experiments as possible

• Analysis of the result – modelling
Experimental Strategy

Most important: **Definition of Aim(s)**

- **Problem formulation:**
  - What is the aim?
  - What is desired?
    (yield, purity, activity, robustness)

- **Familiarization**
  - What is known?
  - What is unknown?
  - Test experiments
The time you spend in the beginning to define a project you will have back with interest in the end!
Screening designs: Full Factorial Designs

- Every level of a factor is investigated at both levels of all the other factors
- It is a balanced (orthogonal) design
- $k$ factors (experimental variables) gives with a 2 level full factorial design
  $2^k$ experiments
Full Factorial Designs

Most common: investigate in two levels

The experiments in a design with two variables

The experiments in a design with three variables

More variables – hyper cube
### Full Factorial Designs

#### $2^k$ Experiments

For two variables

<table>
<thead>
<tr>
<th>Exp. No.</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

For three variables

<table>
<thead>
<tr>
<th>Exp. No.</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

For four variables

<table>
<thead>
<tr>
<th>Exp. No.</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>10</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>12</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>14</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>16</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Simple to generate** – similar pattern no matter the number of variables to investigate!
Analysis of result
Multiple Linear Regression (MLR)

- Regression method using a least squares fit "Classical regression"
- Requires independent variables in the X-block
- Separate model for each Y response
  Coefficients for each y analysed – variable influence can be identified
Always look at the raw data – e.g. a replicate plot

- Each point represent an experiment
- Exp. no 15 performed in replicate
- Variation in overall response, not in replicate
Cross-validation gives $Q^2$

Parts of the data is held out and a model is built on the remaining \( \Rightarrow \) repeated until all data has been kept out once

\[
Q^2 = 1 - \frac{\sum(Y_{\text{obs}} - Y_{\text{pred}})}{\sum(Y_{\text{obs}} - Y_{\text{average}})}
\]

\( \text{etc.} \)
Model diagnostics

<table>
<thead>
<tr>
<th>Nature of data</th>
<th>$R^2$</th>
<th>$Q^2$</th>
</tr>
</thead>
</table>
| Chemical       | Acceptable: $\geq 0.8$ | Acceptable: $\geq 0.5$  
|                |            | Excellent: $> 0.8$        |
| Biological     | Acceptable: $> 0.7$  | Acceptable: $> 0.4$     |

- The goal is **not** to optimise $Q^2$
- A stable and interpretable model which can be used for predictions is desired
- A lousy model can still provide useful information
Useful plots

• Replicate plot
• Design matrix
• Residuals
• Coefficients
• ANOVA tables/plots
• Contour plots
• More…
Candy production – ”sega råttor”

**X (independent) – variables**
- Amount sugar (g)
- Amount glucose (g)
- Amount $H_2O$
- Amount Gelatine
- Amount $H_2O$
- Mix $H_2O$/gelatine speed
- Mix $H_2O$/gelatine time
- Mix $H_2O$/gelatine heat
- Mix 2 speed
- Heat 114
- Cool temperature
- Colour
- Flavour

**Y (dependent) – variables**
- Colour
- Taste
- Sweetness
- ”Seghet”
- Form
- Size
Full Factorial Designs

Sega råttor... Problem!

With an increasing number of variables the required number of experiments rapidly becomes impractical to handle...

<table>
<thead>
<tr>
<th>Number of variables</th>
<th>Number of Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>8</td>
<td>256</td>
</tr>
<tr>
<td>10</td>
<td>1024</td>
</tr>
<tr>
<td>12</td>
<td>4096</td>
</tr>
<tr>
<td>14</td>
<td>16384</td>
</tr>
<tr>
<td>16</td>
<td>65536</td>
</tr>
</tbody>
</table>
Fractional Factorial Designs (FFD)

• One solution is to use a smaller part – a fraction – of the full factorial design
• Possible to greatly reduce the number of experiments
• Still investigate the defined experimental domain well
• $2^{k-p}$ experiments required, $k =$ number of variables, $p$ the size of the fraction
Factorial Designs

(\(2^3\)) Full

(\(2^{3-1}\)) Fractional

Maximum volume

Maximum volume with a minimum number of experiments
Statistical Experimental Designs

Example of different types of designs

- Full factorial designs
- Fractional factorial designs
- Plackett-Burman designs (special case of FD)
- D-optimal designs
- Taguchi designs
- Central Composite Designs (CCC and CCF)
- Mixture Designs
- Simplex Designs
Multivariate analysis

- PCA
- PLS
- MVD
Chemometrics

- Use of mathematical and statistical methods for selecting optimal experiments
  Statistical experimental design and optimisation
- Extracting maximum amount of information when analysing chemical data
  Multivariate data analysis

Multivariate design
Combining statistical experimental design and multivariate data analysis – a tool in drug discovery
The (scientific) world today

- Generating numbers to understand and quantify phenomenon's around us
- Many responses are measured, sometimes at regular time intervals
- "Large" data tables are generated
- Tools for viewing all data simultaneously are needed

\[
\begin{align*}
\text{\(N\) observations} & \quad \text{\(X\)} \\
\text{\(K\) variables} & \\
\end{align*}
\]
Principal Component Analysis (PCA)

- A projection method – extract information (variance) from large data sets
- Creates “windows” in a multidimensional space (matrix with several variables correlated to each other)
- Graphical plots to interpret the result; identify classes, patterns, outliers, etc.
Data pre–treatment

- **Mean-centring**: Subtract average for each variable
- **Scaling to unit variance**: $1/SD_{\text{variable}}$

Raw data → Moving the coordinate system to the origin → The variables have an equal chance to influence the PCA model

Default setting is mean centering and unit variance scaling
PCA – graphical description

Investigating three variables, e.g. formula weight, melting point and log P
PCA

Scores (observations)

Loadings (variables)
## PCA - A graphical example

### Raw data

<table>
<thead>
<tr>
<th>No.</th>
<th>Gender</th>
<th>Shoe Size</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female1</td>
<td>37</td>
<td>168</td>
</tr>
<tr>
<td>2</td>
<td>Female2</td>
<td>36</td>
<td>166</td>
</tr>
<tr>
<td>3</td>
<td>Male1</td>
<td>42</td>
<td>185</td>
</tr>
<tr>
<td>4</td>
<td>Female3</td>
<td>38</td>
<td>171</td>
</tr>
<tr>
<td>5</td>
<td>Male2</td>
<td>41</td>
<td>174</td>
</tr>
<tr>
<td>6</td>
<td>Male3</td>
<td>43</td>
<td>180</td>
</tr>
</tbody>
</table>

**Mean**: 39.5, 174
Centring of data

Raw data

<table>
<thead>
<tr>
<th>No.</th>
<th>Gender</th>
<th>Shoe Size</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female1</td>
<td>37</td>
<td>168</td>
</tr>
<tr>
<td>2</td>
<td>Female2</td>
<td>36</td>
<td>166</td>
</tr>
<tr>
<td>3</td>
<td>Male1</td>
<td>42</td>
<td>185</td>
</tr>
<tr>
<td>4</td>
<td>Female3</td>
<td>38</td>
<td>171</td>
</tr>
<tr>
<td>5</td>
<td>Male2</td>
<td>41</td>
<td>174</td>
</tr>
<tr>
<td>6</td>
<td>Male3</td>
<td>43</td>
<td>180</td>
</tr>
</tbody>
</table>

Mean 39.5 174

Centred data

<table>
<thead>
<tr>
<th>Shoe Size</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.5</td>
<td>-6</td>
</tr>
<tr>
<td>-3.5</td>
<td>-8</td>
</tr>
<tr>
<td>2.5</td>
<td>11</td>
</tr>
<tr>
<td>-1.5</td>
<td>-3</td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>3.5</td>
<td>6</td>
</tr>
</tbody>
</table>

The mean is calculated for each variable. The centring subtracts the mean from values of each variable.
Plotting the data
Raw data vs Centring

Raw data

Centring

Ad scaling!
1. Fit a line to the data points through the origin.

2. Make a perpendicular projection to the principal component for all data points.

3. Measure the distance from the origin to the projections.

Score values \( (t_i) \)
4. Measure the angle, $\alpha$, between the principal component and each variable.

5. Calculate $\cos(\alpha)$

$\text{Loadings (} p_i \text{)}$
6. DModX – Distance to the Model in X (X = the data table)

Finding deviating observations
Comparison between the “graphical” PCA and the PCA obtained from SIMCA

<table>
<thead>
<tr>
<th>No.</th>
<th>Gender</th>
<th>“Graphical” PCA</th>
<th>p₁</th>
<th>SIMCA PCA</th>
<th>p₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female1</td>
<td>-6.5</td>
<td>Size = 0.39</td>
<td>-6.4945</td>
<td>Size = 0.35</td>
</tr>
<tr>
<td>2</td>
<td>Female2</td>
<td>-8.75</td>
<td>Height = 0.92</td>
<td>-8.7171</td>
<td>Height = 0.94</td>
</tr>
<tr>
<td>3</td>
<td>Male1</td>
<td>11.05</td>
<td></td>
<td>11.185</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Female3</td>
<td>-3.35</td>
<td></td>
<td>-3.3339</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Male2</td>
<td>0.6</td>
<td></td>
<td>0.51964</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Male3</td>
<td>6.85</td>
<td></td>
<td>6.841</td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2_X = 0.98033 \]

Calculation of \( R^2_X \)

\[ R^2_X = 1 - \sum \frac{(x_i - x_{ipred})^2}{\sum (x_i - \bar{X})^2} \]  
\[ \text{i.e.} \quad 1 - \frac{SS_{Res}}{SS_{Tot}} \]

or

\[ R^2_X = \frac{\sum (x_{ipred})^2}{\sum (x_i - \bar{X})^2} \]  
\[ \text{i.e.} \quad \frac{SS_{Pred}}{SS_{Tot}} \]
Score plot ($t_1$) - to evaluate the result
Questionnaire PCA example

- Questions in the form of ranking on a continuous scale or as yes/no

- 213 general questions about TV-programs, celebrities, food habits, ethical opinion, etc.

- “Limited time” for answering the questionnaire
Data – the chemistry department

- 14 persons ➔ observations
- 213 questions “describing” the chemists ➔ variables

<table>
<thead>
<tr>
<th>Observation 1</th>
<th>Variable 1</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIV</td>
<td>DIV</td>
<td>DIV</td>
</tr>
<tr>
<td>DIV</td>
<td>Se</td>
<td>Media</td>
</tr>
<tr>
<td>-1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>-3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>-3</td>
<td>-1</td>
<td>4</td>
</tr>
<tr>
<td>-3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>-5</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
Finding relations

- Relating the persons to the questions, i.e. the observations to the variables
Finding patterns

No children

Have children
PCA

- Same principals with a larger number of variables
- Principal components are always orthogonal (independent) from each other
- Each principal component summarises the data set by generating scores for the observations with corresponding loadings for the variables, i.e. scores and loadings should be compared to each other
- Can handle moderate amount of missing data (25%)
Determining the number of principal components

- $Q^2$ – cross-validated value indicating how well the model is able to predict the data (explained variance)

- Eigenvalue – the length of the principal component

- (Chemical) interpretation in the plots, e.g. $A=1$ ($A$ denotes the number of components)
Outliers – deviating observations

- DModX for 1
- Outlier in DModX
- t-value for 2
- DModX for 2
- Outlier in score space
- t-value for observation 1
PCA objectives

• Overview of data – always a good starting point (historical data, data from other sources)
• Identify patterns in the data set
• Identify important variables
• Identify outliers
• Understand how variables (loadings) and observations (scores) are related to each other
PCA objectives

• Classification & clustering – dividing the data set depending on the patterns (structure) of the data set
  – Treated vs. untreated
  – Before vs. after treatment (cross-over designs, difference in time)
  – Bioinformatics (proteins, enzymes)

• Classify new observations

• Summarise data with a fewer number of variables – generating "principal properties" design variables for multivariate design (starting materials or products, chemical libraries)
Soft Independent Modelling of Class Analogy

• Referred to as SIMCA classification
• Separate PCA models for each identified class
• Predictions of new objects in score space
• Predictions of new objects in DModX

• Use your and the knowledge of others…
PCA Modelling

X-matrix

Class I

Class II

New samples
Class III

PC1

$\mathbf{x}_1$

$\mathbf{x}_2$

$\mathbf{x}_3$
SIMCA Modelling, Class I

X-matrix

Class I

Class II

New samples Class III

Predict in to model

PCA Class I

PC1

x₁

x₂

x₃
PLS - Partial Least Squares
Projection to Latent Structures

• A projection method, "regression extension of PCA"

• Find the relation between the latent structure in X and latent structure in Y

• Maximize the covariance between the X block and the Y block

• PLS1 - one Y variable

• PLS2 - more than one Y variable
Analysis of the result

• Relating the variables (X-block) to the response or responses (Y-block)
• Need a regression method which can handle correlated X-variables
• Analyse many Y variables at the same time
**PLS**

- Can handle many noisy collinear variables (compare with MLR)
- Tolerate moderate amounts of missing data (X and Y)
- Multiple responses modelled at the same time
- The result can be graphically visualized i.e. score plots and loading plots
PLS - Geometric Interpretation

- Each observation is represented by one point in the X-space and one in the Y-space
- As in PCA, the initial step is to calculate and subtract the averages; this corresponds to moving the coordinate systems
• The mean-centering procedure implies that the origin of each coordinate system is re-positioned.
**PLS - Geometric Interpretation**

- The first PLS-component is a line in the X-space and a line in the Y-space, calculated to
  - **a)** approximate the point-swarms well in X and Y
  - **b)** provide a good correlation between the projections \((t_1\) and \(u_1\))

- Directions are \(w_1\) and \(c_1\) and co-ordinates along these vectors are \(t_1\) and \(u_1\), respectively.
PLS - Geometric Interpretation

- The projection coordinates, $t_1$ and $u_1$, in the two spaces, $X$ and $Y$, are connected and correlated through the inner relation
  
  $u_{i1} = t_{i1} + h_i$ (where $h_i$ is a residual)

- The slope of the dotted line is 1.0
PLS predictions

- A new observation is similar to the training set if it is inside the tolerance cylinder in X-space.
- Then its projection on the X-model (t) can be entered into the T-U-relation giving a u-value for each model dimension.
- These values define a point on the Y-space model, which, in turn, corresponds to a predicted value for each y-variable.
PLS - Model diagnostics

SIMCA supports two internal model validation strategies

1. Cross validation
   To estimate the optimal model complexity

2. Response permutation test (Validate-option)
   To check the degree of overfit
Evaluation of $R^2$ and $Q^2$

- PRESS is the sum of squared differences between predicted and observed $y$-elements:
  \[ \text{PRESS} = \sum (y_{im} - \overline{y}_{im})^2 \]

- PRESS can be transferred into a dimensionless quantity, $Q^2$, which resembles $R^2$:
  \[ Q^2 = 1 - \frac{\text{PRESS}}{\text{SSY}_{\text{total}}} \]

- $R^2$ is always larger than $Q^2$
- High $R^2$ and high $Q^2$ is desired
- The difference between $R^2$ and $Q^2$ should not be too large:
  \[ R^2 = 1 - \frac{\text{SSY}_{\text{resid}}}{\text{SSY}_{\text{total}}} \]
PLS-DA
PLS Discriminant Analysis

- Adds information in a Y block indicating group belonging
- X variables important for separation can be identified
- Works for more than two groups

For example two groups
Group A; $y_1=1$, $y_2=0$
Group B; $y_1=0$, $y_2=1$
PLS – step by step

1. Problem definition / Objective
2. Collect data
3. Import data
4. Pre-treatments
5. Calculate model
6. Evaluate/Validate model
7. Analyse model
8. Suggest new experiments
Multivariate design
Combinatorial Chemistry

• Fast compound generation
  On solid phase, in solution, phage display, in parallel, in mixtures

• Analytical chemistry
  Purification, analysis methods

• ID/characterization
  Coding, LC-MS, NMR

• Fast biological testing
  High Trough-put Screening

• Increase structure diversity
  Synthesize and test a large number of compounds
Drug Discovery Aim: Reduce Development Time

6200 Synthesized

21 Subacute Toxicology

6.5 Tested in humans

2.5 Phase III clinical trials

One Approved drug

Time span 12.8 Years

COST ~ $350,000,000 Nature vol. 384, 1996, supp.
Information Drives the Drug Discovery Process

Aim: Gain information
  – the more
  – high quality data \(\rightarrow\)
  design
  – the quicker
  – the better

Guided learning

New or better drugs

FILED PATENT
\(\rightarrow\)
THE CLOCK IS RUNNING
The Combinatorial Explosion

- Rapidly generate a great number of possible compounds

- Commercially available reactants (ACD – Available Chemicals Directory, others)

- Proprietary data bases

100 Diamines + 200 Ketones + 300 Carboxylic acids → 6,000,000 products

HTS → ID, verification → New lead compounds & new drugs
The Chemical "Space"

- $\sim 10^{200}$ organic molecules with a molecular weight of less than 850 g/mol
- $\sim 10^{40}$ organic compounds with “drug like properties”
- $10^{17}$ seconds have passed since the Big Bang
  Roughly 10 to 20 billion years ago
  – if you believe in that model…
Practical Limitations

• ~ One million compounds in mixtures
• ~1000 compounds in parallel synthesis
• Costs and practical obstacles to consider
  – Equipment
  – Synthesis
  – Work up
  – Biological testing
    (£ 0.1 - 2.0 per sample, 1996)
  – Disposal considerations
  – Personnel
    (salaries, training, etc.)
Important concepts

• **Maximum diversity**
  Chose as different compounds as possible

  *Diversity*: spread of compounds with a defined set of descriptors

  *Descriptor*: a variable characterizing a property of the compounds

• **Similarity**
  Chose structures similar to a known active compound, lead optimization
Best Selection?

Selection in a chemical space defined by two descriptors

All structures

Not only the number of compounds synthesised (experiments) that are of importance!
Design – screening

Increase training set – decrease risk of leverage due to deviating results
Follow up design
Multivariate design

- Multivariate characterization +
- Multivariate data analysis (PCA and PLS) +
- Factorial designs ➔

- MULTIVARIATE DESIGN (MVD)
- When used in drug development and with design in identified sub-cluster – Statistical Molecular Design (SMD)
Multivariate Design

• Tool for selecting representative structures
• Full factorial designs
• Fractional factorial designs
• D-optimal designs
• Works for more than two principal properties

• Factorial designs
  - synthesis optimisation
  - formulation optimisation
  - process optimisation
  etc.
Summary

• Historical data
  - Always a good starting point for analysis (PCA)
  - Determine data structure, preferred format/output
  - Get acquainted with the process and the data
  - Find “hidden” information
  - Good starting point for discussions

• Determine
  - The aim – is there a defined stop criteria (yield, purity, accepted batches, ID important/”sensitive” variables etc.)
  - Important to define prior to investigation → know when to stop
  - The experimental domain (variables, settings, responses)
Summary

• Design of Experiments (DoE)
  - Simplify analysis
  - Ensure a systematic variation in the investigated experimental domain
  - Small design within the defined limits
  - (Design in historical data)

• Analysis
  - PCA (SIMCA etc.)
  - PLS, PLS-DA

• Next step
  - Aim(s) reached
  - Important variables
  - New optimal experiments
Acure Pharma Business model

Consulting services
- Support chemistry
- Support IPR questions and problems
- Second opinion/news value
- Chemometrics
  PAT (Process Analytical Technology, FDA guidelines)
- Network

Drug development
- Proprietary compound library
- Finding financing for drug development
  - Company collaborations
  - Venture capital
  - Governmental funding (7FP, VINNOVA)
- Defined AcurePharma projects
- Exploratory research
Acure Pharma History

... 
AcureOmics AB September 2007 
VINNOVA grant BBB, May 2007 
Start-up company of the year 2007 
Action Pharma A/S – collaboration 
Uminova – in-licensing of Cancer project 
Research agreement with professor Sharma 
AnaMar Medical AB, Out licensing 
Carlsson Research – Research agreement, MVA 
Educational activities (SE, UK) 
Second opinion, Novelty search 
Consulting in research and development 
Acure Pharma Consulting AB ➔ Acure Pharma AB 
Future outlook

... 

... 

... 

Clinical trials Phase I

Identification of new early projects

Start up Ltd in UK? – EU project, 2008 – 2011

Consulting and educational activities

Establish additional research collaborations
## Project pipeline 2008

### Pre-clinical development

<table>
<thead>
<tr>
<th></th>
<th>Hits</th>
<th>Leads</th>
<th>Pre-CDs</th>
<th>CD</th>
<th>Phase I</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNF-alpha</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood Brain Barrier</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### RESEARCH COLLABORATIONS

- Adenosine 2a agonists
- MCR
- Serotonin modulators
- SCI
- Metabolic syndrom