

Improving the Y for QSAR

Using Chemometrics to Improve Biological Results

Presented at the UK QSAR meeting. 12th October 2004

by

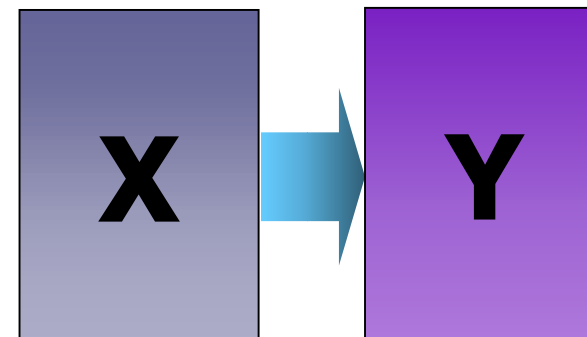
Mark Earll Umetrics UK Ltd.



QSAR



- Most focus of QSAR is on X
 - Chemical descriptors
 - Measured physical properties
 - Spectroscopic measurements
- What about the Y?
 - Biological measurements
 - Often 'one off' measurements
 - Derived from curve fitting
 - Optimisation of many Y's
 - Activity, 2ndary activity, ADME



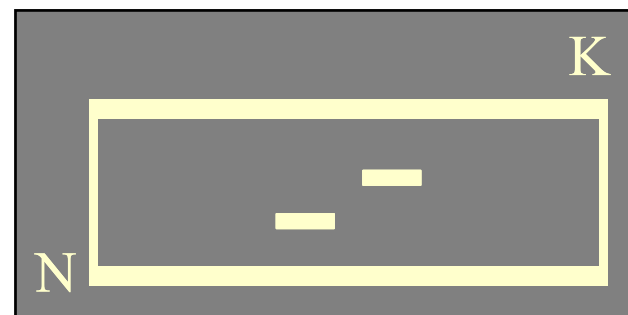


Chemometrics



- What is chemometrics?
 - Data Projection Methods
 - PCA, SIMCA, PCR, PLS, PLS-DA
 - Takes advantage of redundancy and correlation in data

- Developed for difficult to analyse tables of data
 - Many variables
 - Low No. of Observations
 - Noisy data
 - Missing data
 - Many responses



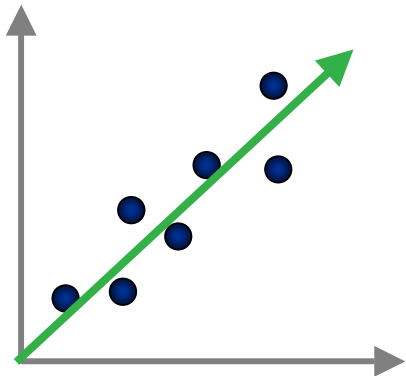


Principal Components Analysis

- Data visualisation and Simplification
 - Information comes from the **Correlation structure** of the data
 - Mathematical principle of Projection to lower dimensions

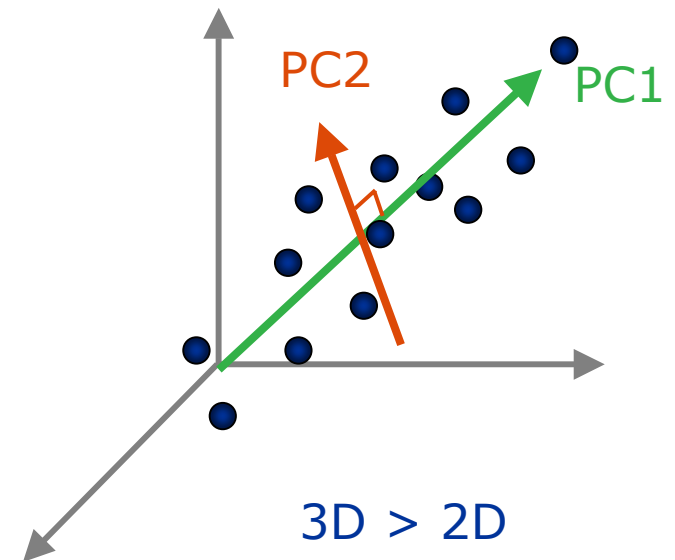
2 Variables

V1	V2
1	1.3
2	2.3
3	2.7
4	3.9
...	...



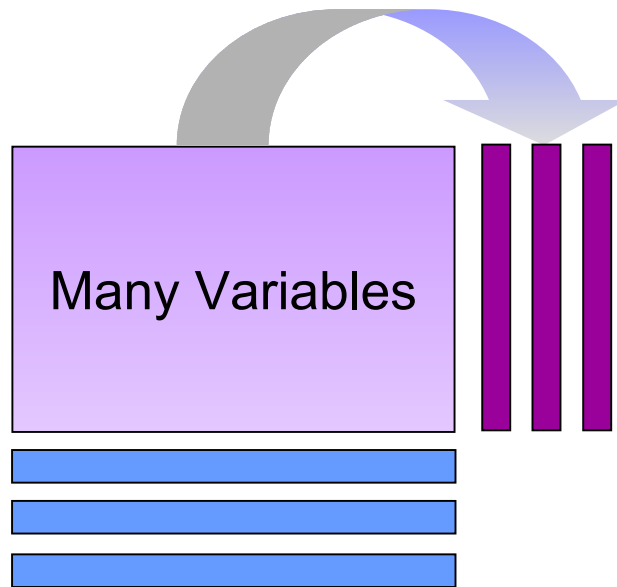
Many Variables

V1	V2	V3	Vn
1	1.3	0.4	
2	2.3	1.2	
3	2.7	2.1	
4	3.9	4.6	
...





PCA continued..



- SCORES:
- Fewer 'Latent' Variables
- Concise summary of the old
- Finds observation correlations
- Separates signal from noise

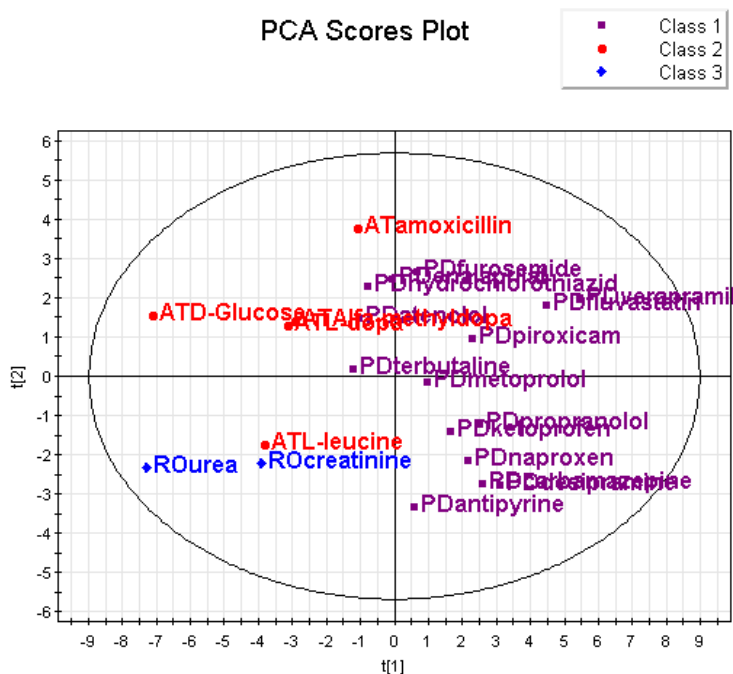
- LOADINGS:
- Summary of Variable correlations
- How 'Latent' variables relate to original ones

- Plots of scores and loadings give a pictorial representation of a dataset
- Correlations between **variables** and **observations** are easily seen...



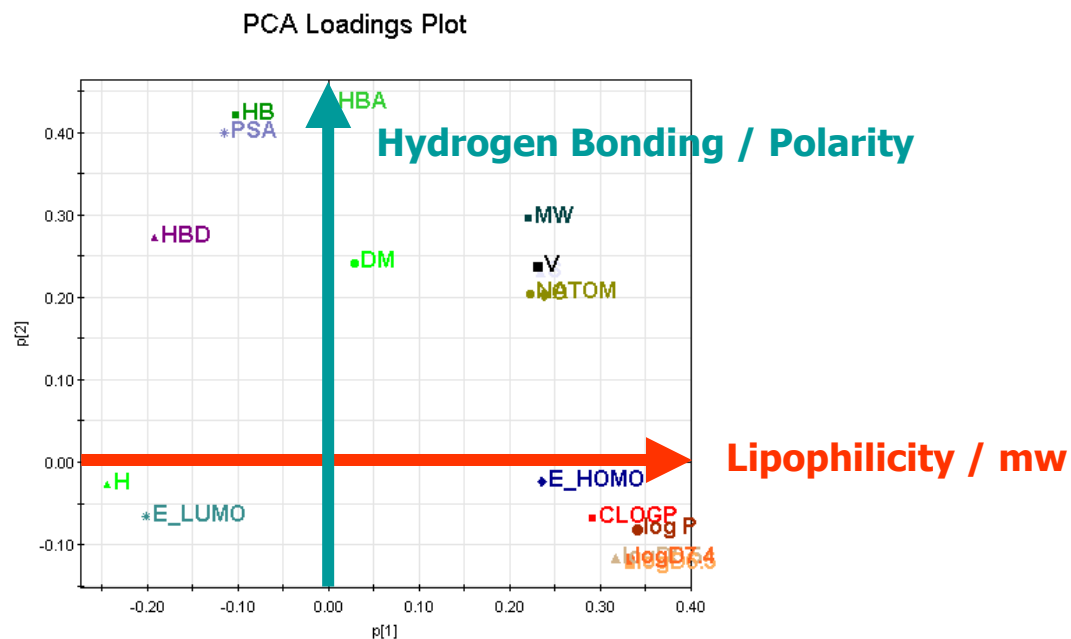
PCA of Human permeability data

- Human Intestinal Absorption (Winiwarter et al 1998)
 - Route of absorption vs Physchem parameters
 - Observation and variable correlations easily visualised



D-Crit [3] = 1.56302

SIMCAP+ 10.0 - 18/09/02 15:12:17



SIMCAP+ 10.0 - 18/09/02 15:17:11



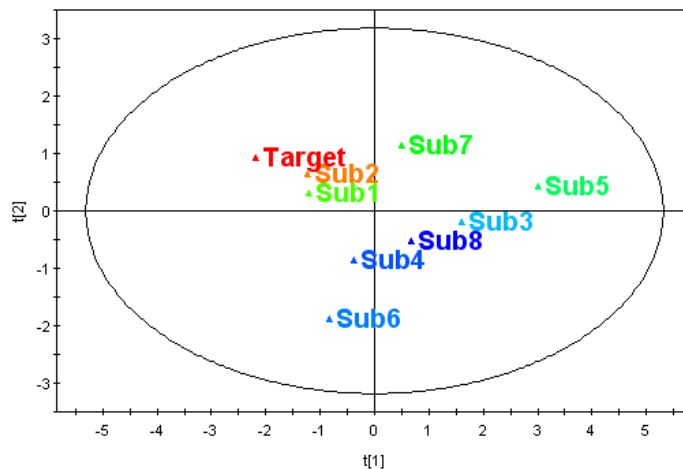
Lead finding – Simple PCA



- Multiple biological tests
 - Eight compounds
 - PLUS a theoretical target
 - Lead identification
 - Correlations in activity tests?
 - Redundancy vs. economy

Dataset: Aldrich							
	1	2	3	4	5	6	7
1	Primary ID	Name	Test1	Test2	Test3	Test4	Test5
2	1	Target	55	20	50	70	45
3	2	Sub1	49	23	55	44	67
4	3	Sub2	53	30	50	62	69
5	4	Sub3	72	42	71	21	90
6	5	Sub4	38	33	64	15	68
7	6	Sub5	93	55	82	35	94
8	7	Sub6	66	40	43	18	35
9	8	Sub7	72	20	73	39	85
10	9	Sub8	75	25	50	10	100

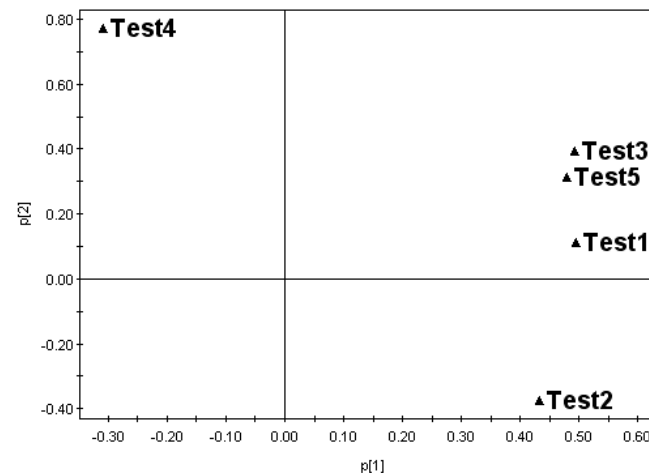
Scores Plot
Colored according to value of Test4



Ellipse: Hotelling T2 (0.95)

SIMCAP+ 10.5 - 06/09/2004 12:01:4

Loadings Plot



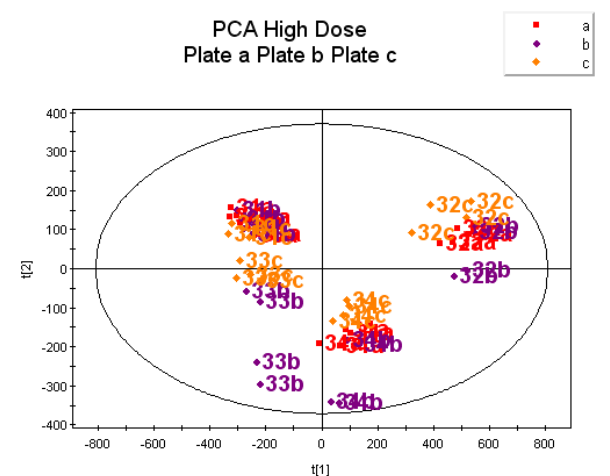
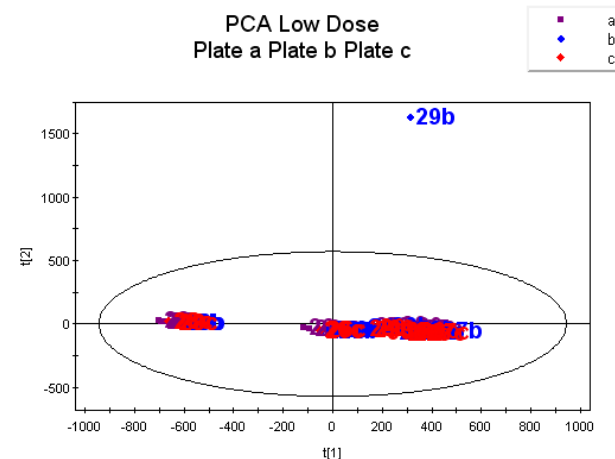
SIMCAP+ 10.5 - 06/09/2004 12:02:54



PCA for QC of Assays



- PCA very sensitive to outliers
 - Detection tool for QC of assays
 - Outliers
 - Trends (biological variation should be random!)
- 'Triomics' = Genomics, Proteomics, Metabonomics
 - Frequent problems with reproducibility
 - Plate to Plate / Gel to Gel variations
 - Operator to Operator variations
- HTS Plating problems
 - Dispensing
 - Shaking
 - Manufacturing defects



Ellipse: Hotelling T2 (0.95)

SIMCAP+ 10.5 - 07/09/2004 11:53:29

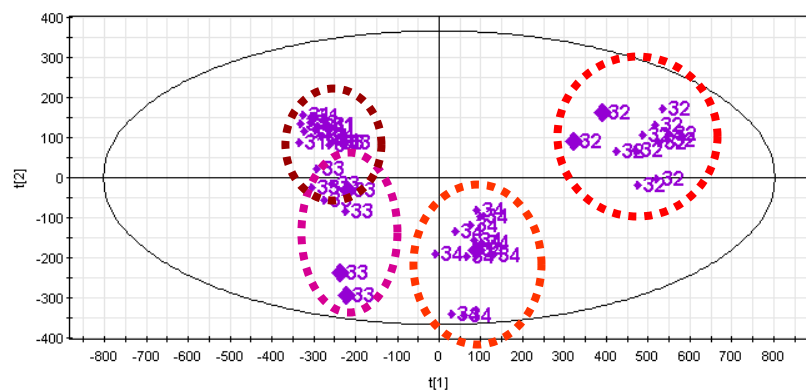


Genomics example: Grid repeats



- PCA used to compare chip repeats vs animal variation
 - Is the between animal variation greater than the replicate error?
 - How reproducible are repeats? (10 grid repeats 4 Doses 1600 variables)

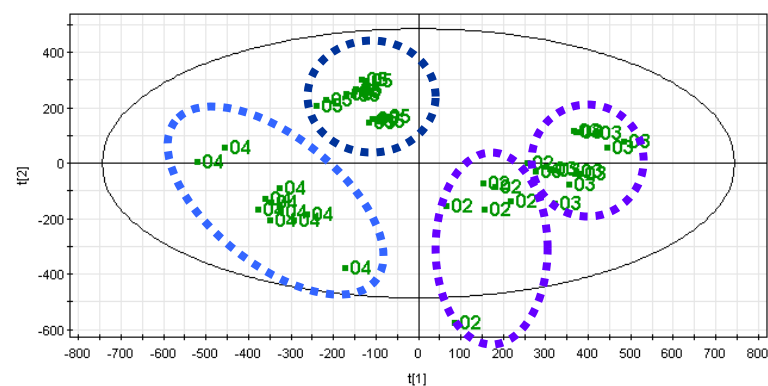
PCA High Dose Repeats Scores plot t1 / t2



D-Crit [2] = 1.20937

SIMCAP 9.0 - 29/08/02 13:18:56

PCA Control Group Repeats Scores plot t1 / t2



SIMCAP 9.0 - 29/08/02 17:10:07

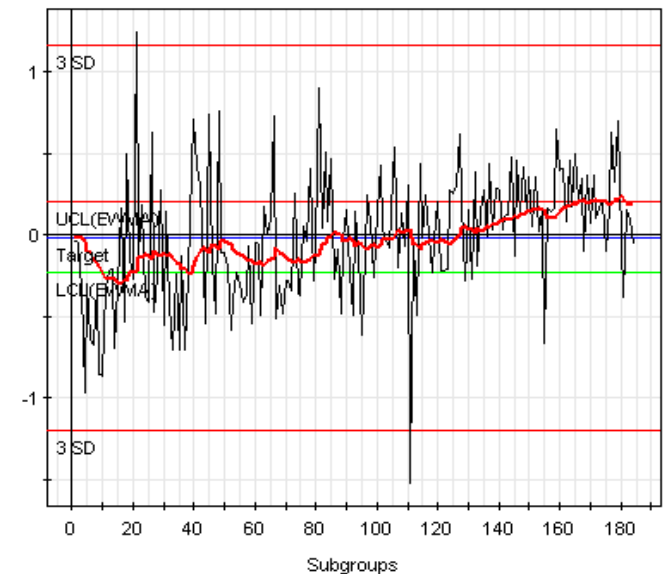
- Genetic variability between animals $>$ between repeats
 - Averaging by animal will reduce data but not information
 - Possibility of using lower number of repeats in future?



Long Term QC



- Repeats and standards essential
 - Repeats: must try to capture all possible sources of variation
 - Continual improvement: Detect source and fix!
- Example: Make PCA of ALL control rats
- Plot scores / DMODX and look for trends due to
 - time
 - batch of animals, handling
 - season
 - personnel
 - equipment
 - (weather!)
- Essential to keep good records for later analysis

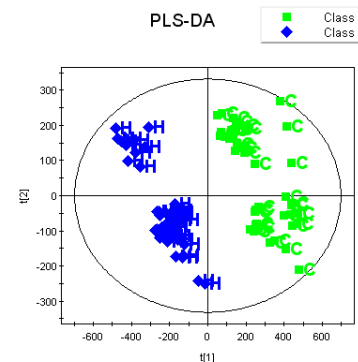




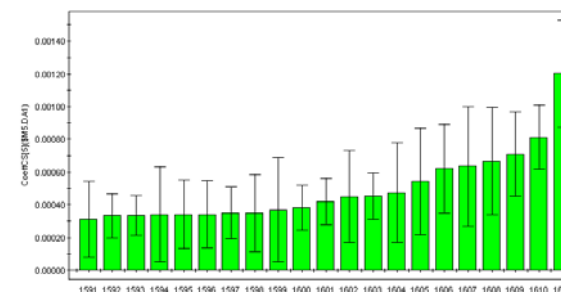
Validation of Triomics data



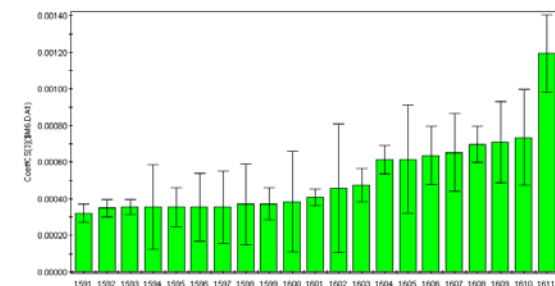
- Validation is essential !!!
- Separation of groups
 - PLS-Discriminant Analysis
 - Control/Treated Diseased/Healthy
- HOW PREDICTIVE?
- Coefficients show potential biomarkers, genes or proteins
- Must make model on new data (or left out)
 - Compare separation / Classification success
 - Compare coefficients (they should be same!)



PLS-DA Coeffs
Most Upregulated TS1



PLS-DA Coeffs
Most Upregulated TS2





DoE: Reporter Gene Assay



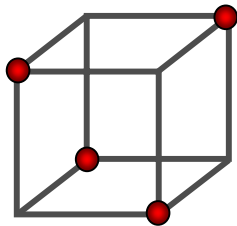
- DoE used to improve bio-assays
 - DOE is a set of optimised experiments where each factor is varied independently
 - Determines the influence of each factor and interactions
 - Can dramatically reduce the number of experiments required
- Inhibition of inflammatory response – reporter gene assay
 - 6 Factors: No of cells, Amount of PMA, Amount of Ionomycin, Stimulation time, Buffer volume, sample:substrate ratio
 - 1 Response: Signal to Background
 - **Screening phase** 2^{6-2} FFD = 16 experiments (c.f. 64 expts)
 - “Which are the important factors?”
 - **Optimisation phase** CCF = 17 experiments
 - “Which are the best settings of these 3?”
 - **Robustness testing phase** 2^{5-1} FFD = 16 experiments (c.f. 32)
 - “How sensitive is the method to small changes?”



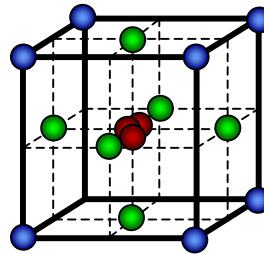
Reporter Gene cont..



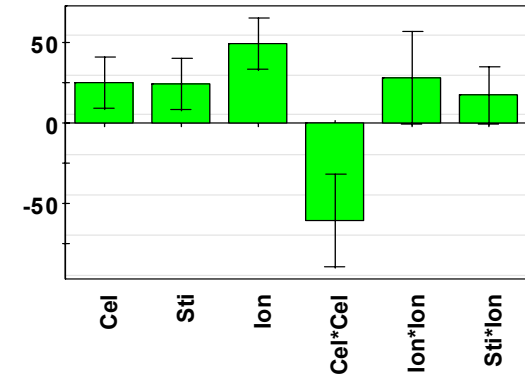
- Screening
 - 16 expts (2^{6-2} FFD)
 - Most important factors
 - Cell, Ionomycin, StTime



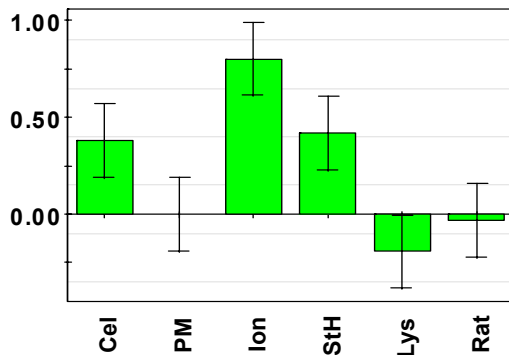
- Optimisation
 - 17 expts (CCF)
 - Best combination



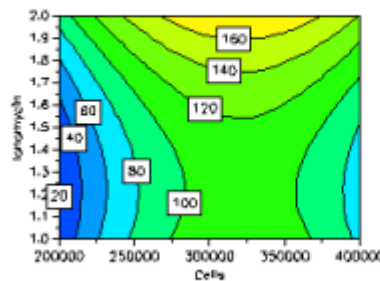
Investigation: Reporter Gene Assay RSM with CCF (MLR)
Scaled & Centered Coefficients for S/B



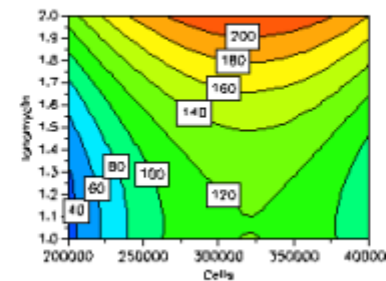
Investigation: Reporter Gene Assay Screening (MLR)
Scaled & Centered Coefficients for S/B~



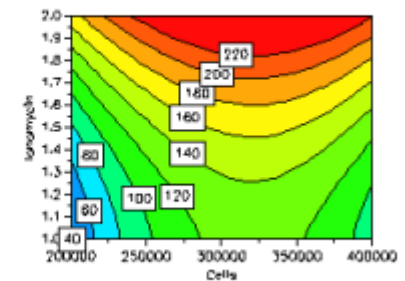
Investigation: Reporter Gene Assay RSM with CCF (MLR)
4D Contour of S/B



StimH = 4



StimH = 5



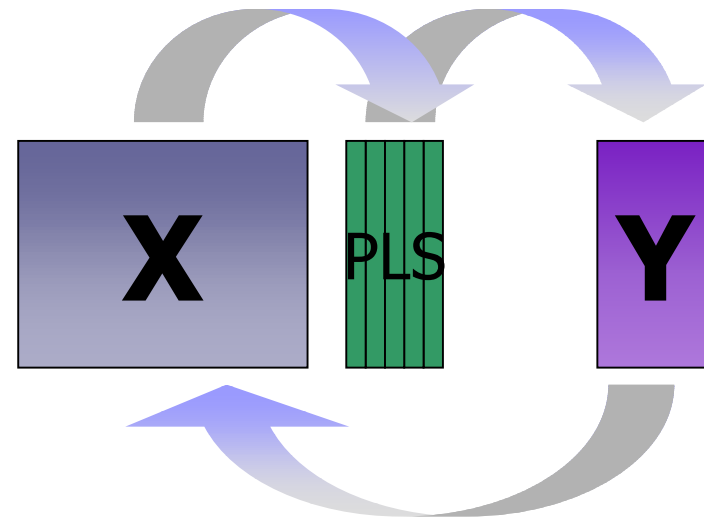
StimH = 6



Partial Least Squares – Multi Y



- PLS relies on a projection of X as does PCA
 - BUT guided by its relationship to Y
- A versatile regression method
 - Multi Y variables
 - Describes structure of X and Y
 - Robust to noise
- Interpretation
 - Latent variables = Main trends



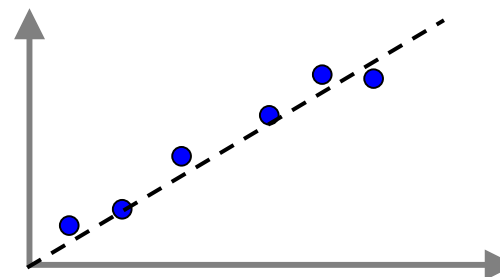


The forms of QSAR Y data



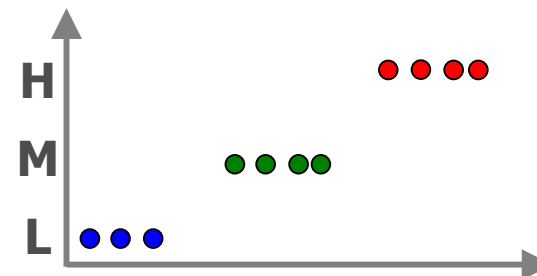
1. Continuous data

- One-off measurements
- Averages (May lose information)
- Curve Fits (IC50)



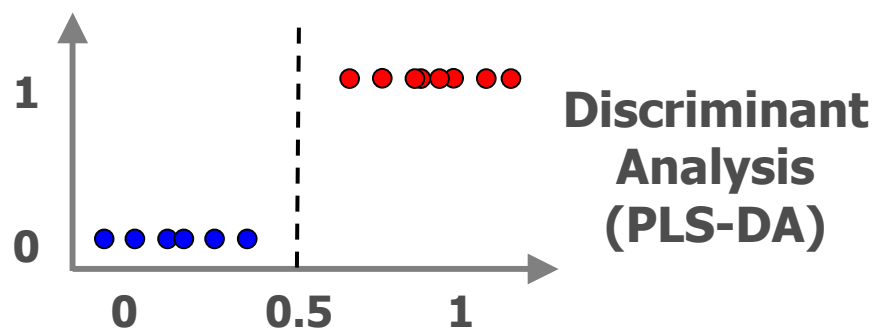
2. Multi-level Qualitative

- Low / Med / High



3. Qualitative

- Control / Treated
- Healthy / Diseased



4. Descriptive

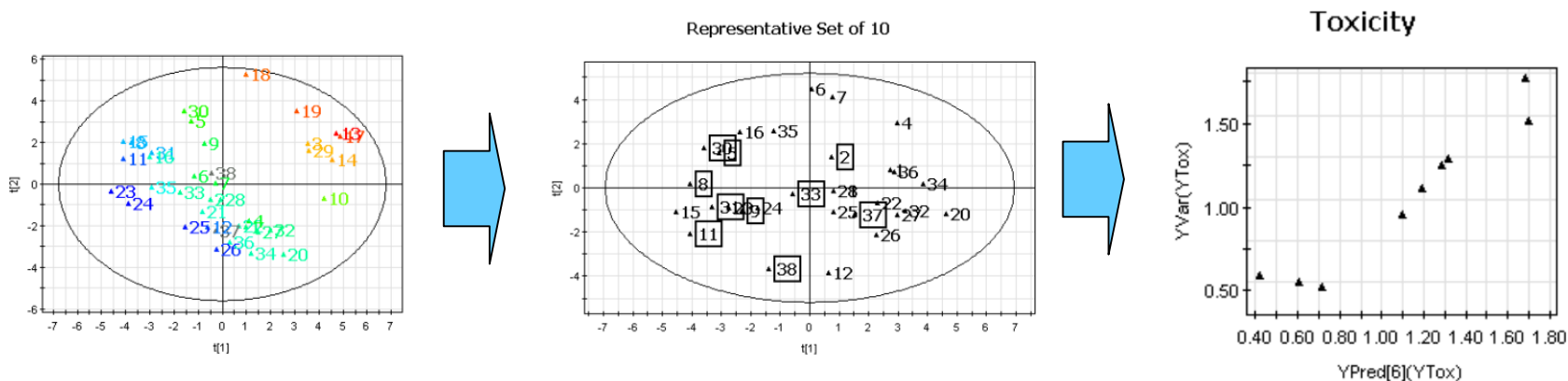
- Pathology data
- c.f. Sensory data

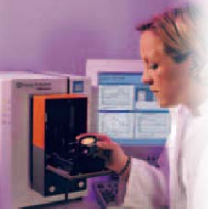


Statistical Molecular Design



- A somewhat heretical concept!
 - Improve quality of biological results by measuring fewer samples!
- Statistical Molecular design
 - PCA to characterise chemical space
 - Use Design of Experiments to select ***small representative subsets***
 - Smaller sample sets enable repeat biological measurements
 - Sufficient chemical variation captured and cause-effect models

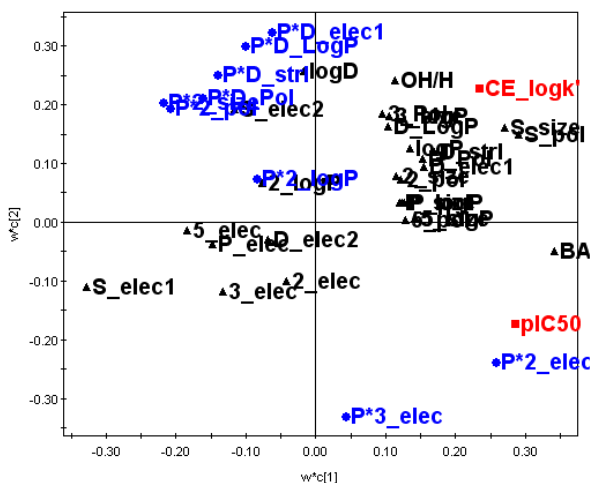




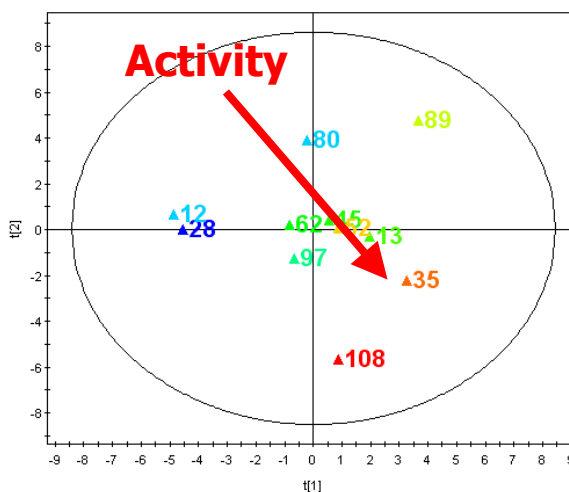
PLS of Thrombin Inhibitors



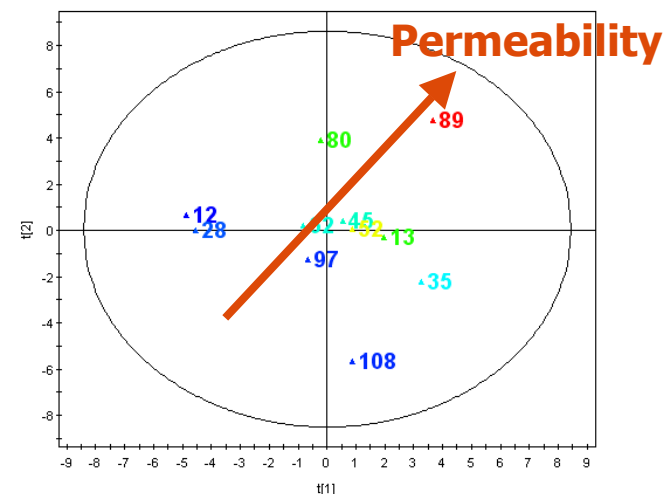
- 18 Representative compounds selected by SMD
- Multi Y optimisation
 - Primary activity IC50 Thrombin, (Secondary IC50 Trypsin)
 - Measure of membrane partitioning (by CE)



Thrombin1.M1 (PLS)
Colored according to value in variable Thrombin1.pIC50



Thrombin1.M1 (PLS)
Colored according to value in variable Thrombin1.CE_logk'



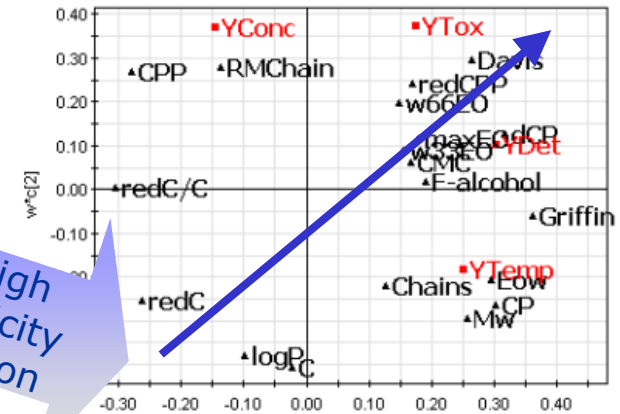
- PLS Model interpretation (Coeff and Loadings plots)
 - identifies desirable properties
 - future direction of synthesis defined



QSAR of Surfactants

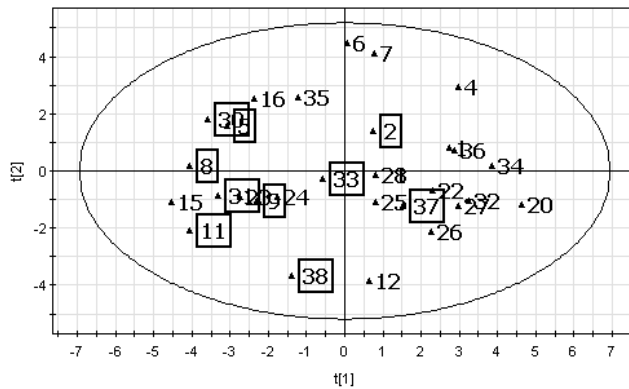
- 38 Surfactants made
 - Characterised by 18 properties
 - 10 selected for testing using DoE
- Measure 4 Y variables
 - Yconc, Ytox, Ydet, Ytemp
- Predict 28 untested

PLS Surfactant
 $w^*c[1]/w^*c[2]$

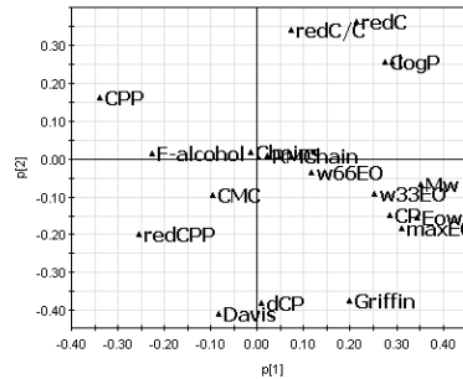


High toxicity region

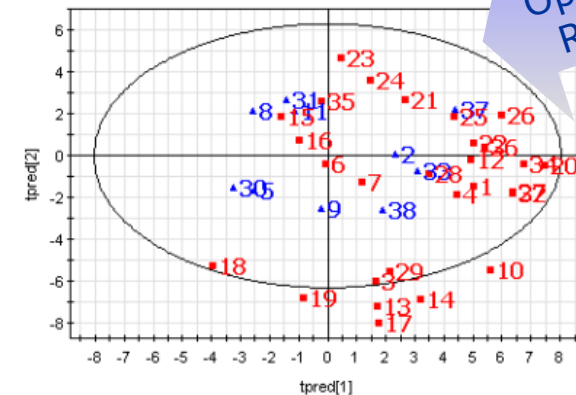
PCA of 30 Surfactants
Representative Set of 10



Surfactant p[1]/p[2]



Predicted Scores for
Untested Surfactants



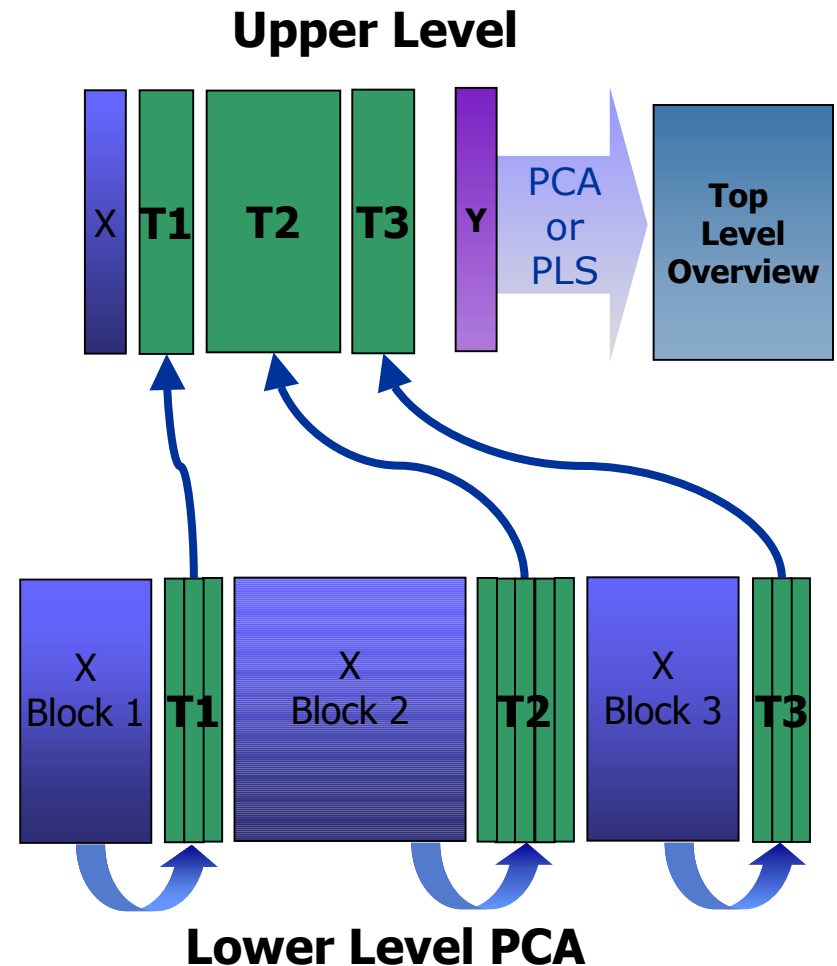
Optimum Region



Hierarchical Modelling

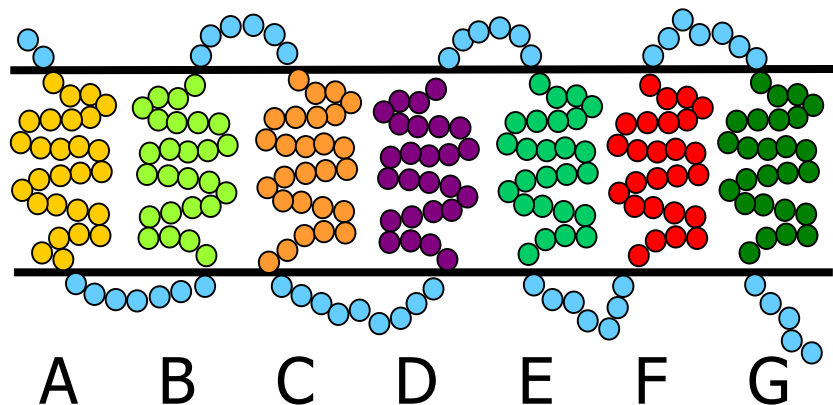


- Multi-step multivariate modelling
- Aim is simplification and ease of interpretation of large datasets
- Scores (or residuals) from one model used as basis for next
- May be used with PCA, PLS or PLS-DA
- Applications:
 - Structural Proteomics
 - Megavariate Y (trionomics)
 - Combining spectral techniques
 - LC-MS



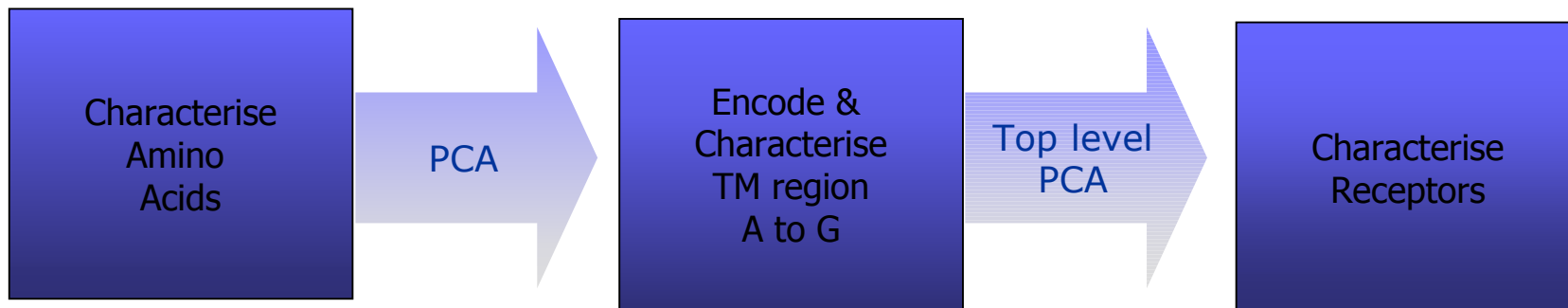


Receptor Structural Analysis



(Loops ignored TM regions only)

- 7TM GPCR receptors
- 3 level hierarchical analysis
- Interpretation made easier

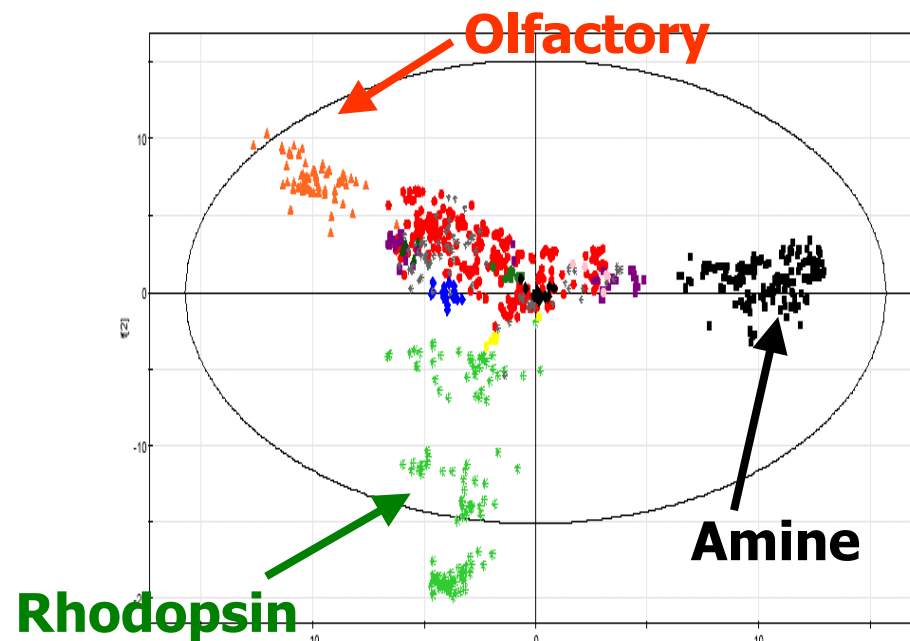




Multivariate Analysis of GPCR's



- Amino Acid Zz scales encoding 7 Transmembrane regions
- Hi-PCA Separates 12 Major classes, Subclasses, Species
- PLS-DA shows which AA's are conserved
- Rhodopsin often used as a 7TM model, is structurally quite different from other GPCR's

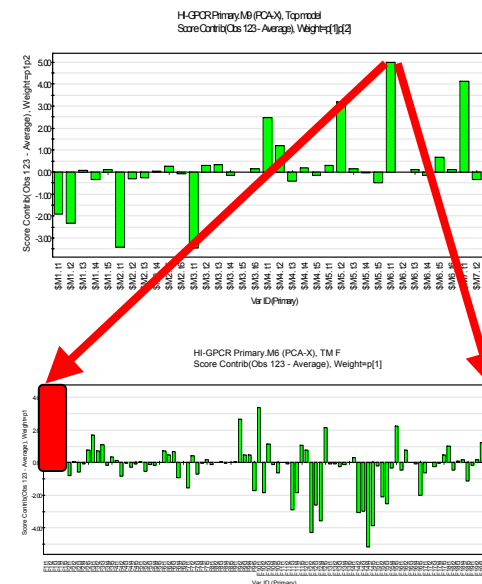




Receptor Model (cont)



- Interpretation
 - Drill down to base level Receptor -> Region -> Sequence
 - Example
 - Olfactory receptor: Region F score 1 is high
 - Amino acid pos 1 is high in properties z1 - z3
 - z1 = Lipophilic z2 = Large z3 = Acidic
- Predictivity
 - Training set 200 / prediction set 100 receptors
 - Percent correct
 - Amine 100%
 - Peptide 96%
 - Rhodopsin 99%
 - Olfactory 100%
 - Orphan 93%

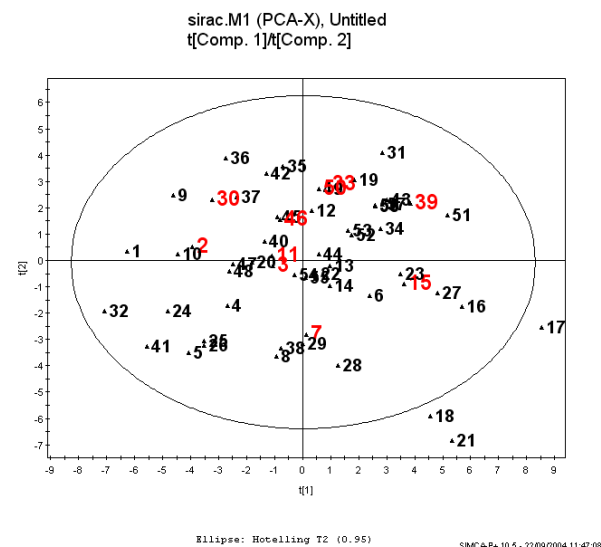




Hierarchical Modelling of Toxicity data



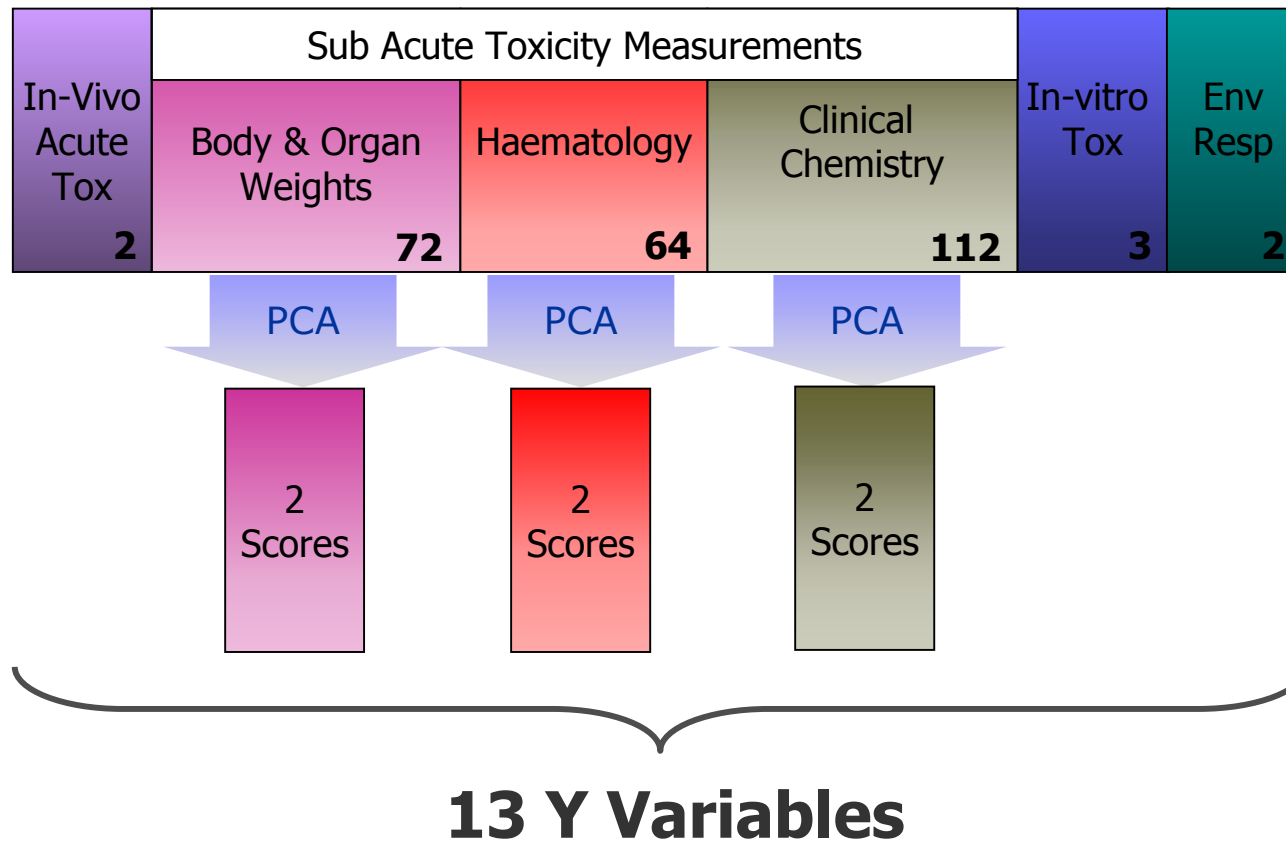
- SIRAC: Swedish-Italian project for Ranking of Chemicals
 - 1987-1995 Risk assessment of high production chemicals
 - 58 Halogenated aliphatic hydrocarbons
- Statistical Molecular Design
 - 30 Chemical descriptors
 - 10 Representative (non gas) molecules selected
 - Extensive biological testing on these 10
- **255 Y variables !!!!**
 - For QSAR modelling
 - Blocks of data will dominate
 - Hard to interpret





Hierarchical modelling

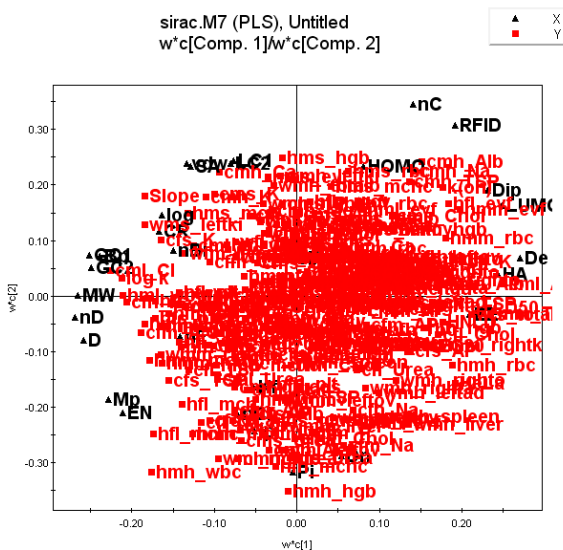
- 255 Y Variables -> 13



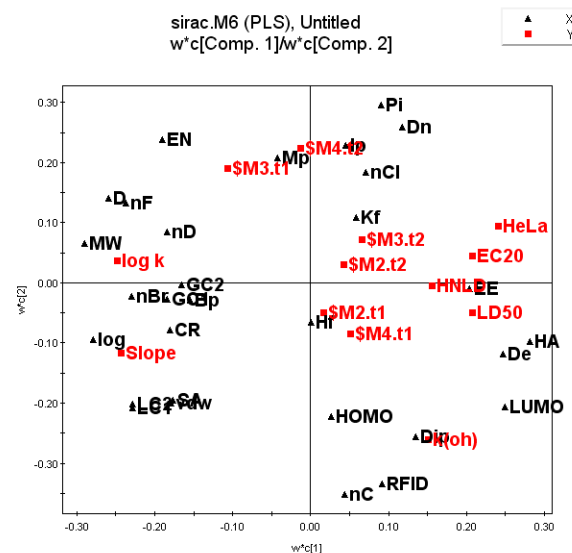


Hierarchical modelling 3

- PLS: 30X Variables 13Y Variables (6 are PCA summaries)
- Hierarchical model aids interpretation
 - Avoids domination of large blocks
 - PLS components can be interpreted



SIMCA+ 10.5 - 22/09/2004 16:43:13



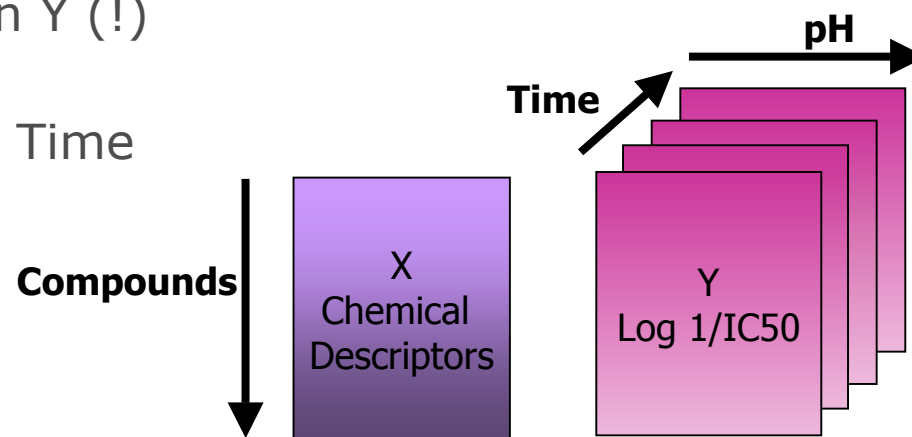
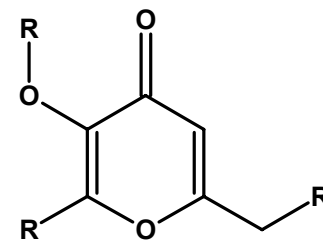
SIMCA+ 10.5 - 22/09/2004 16:43:36



3D Y : Time Resolved QSAR



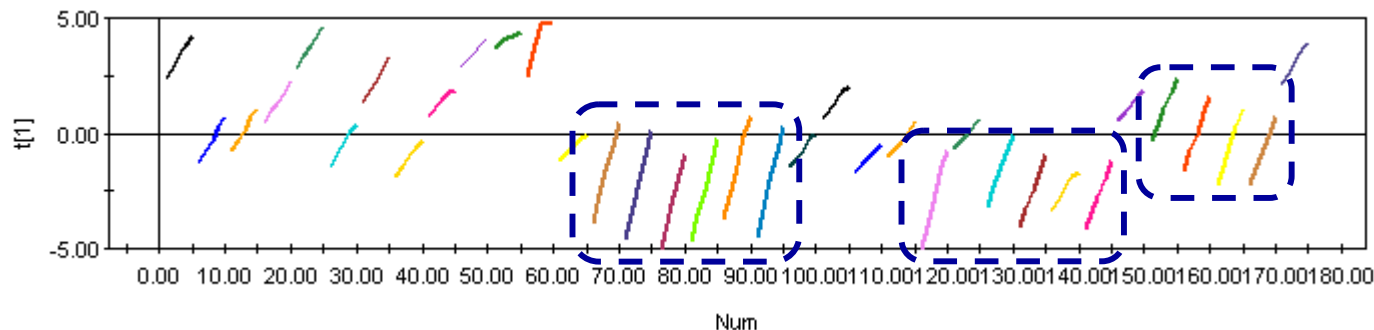
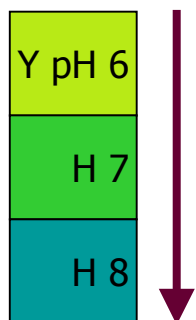
- Antibacterial 4-pyranones
- Activity against e.coli as a function of
 - Compound
 - Time
 - pH
- Unusual case of 3-Dimensions in Y (!)
- Which compounds have highest Time or pH dependent potency?
- Unfold data
 - Time wise
 - pH wise



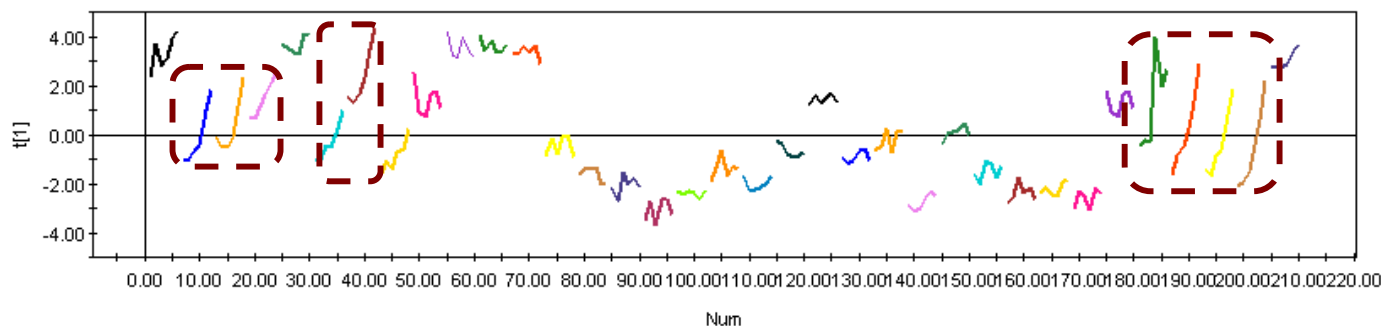
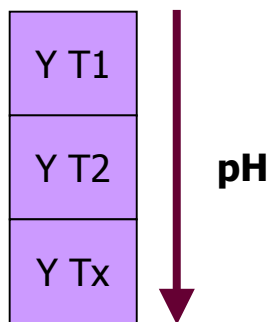


Batch modelling

- PLS batch model with time
 - Compounds with highest time-sensitive toxicity identified



- PLS batch model with pH
 - Compounds with the highest pH-sensitive toxicity identified





Conclusions

- Multiple and relevant Y's make successful QSARs !
 - The quality of the model depends on Y (biology) as much as on X (structure)
- PCA is a useful tool for QC of Biological (Y) data
 - Detection of trends, outliers, systematic variations
 - Summarises correlations in activity data, visually
- PLS is a versatile regression method
 - Multiple Y variables allows optimisation
 - Interpretable in terms of underlying chemistry & latent variables
- Hierarchical modelling simplifies massive datasets
 - 'Divide and Conquer' break into blocks
 - Easier interpretation
- DoE is an optimal way of designing experiments
 - Improving robustness of biological assays



Thanks & Acknowledgements

- **Umetrics**
 - Lennart Eriksson, Erik Johansson, Oliver Whelehan, Svante Wold, Nouna Kettaneh-Wold
- **Umeå University**
 - Johan Trygg, Henrik Antti
- **Imperial College**
 - Jeremy Nicholson, John Lindon, Elaine Holmes
- **AZ**
 - Anna Linusson
- **QSAR Soc / Misc**
 - Sally Rose, Patrick Barton
 - Torbjorn Lundstedt