Journal of Computer-Aided Molecular Design, **16:** 711–726, 2002.
KLUWER/ESCOM
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Megavariate analysis of hierarchical QSAR data

Lennart Eriksson[1,*], Erik Johansson[1], Fredrik Lindgren[2], Michael Sjöström[3] & Svante Wold[3]
[1]*Umetrics AB, POB 7960, 907 19 Umeå, Sweden;* [2]*Umetrics AB, Malmö Office, Storgatan 22, SE-211 42 Malmö, Sweden;* [3]*Institute of Chemistry, Umeå University, Umeå, Sweden*

Received 8 May 2002; accepted in final form 19 November 2002

Key words: PCA, PLS, hierarchical modelling, multivariate analysis, megavariate analysis, QSAR, statistical molecular design (SMD)

Summary

Multivariate PCA- and PLS-models involving many variables are often difficult to interpret, because plots and lists of loadings, coefficients, VIPs, etc, rapidly become messy and hard to overview. There may then be a strong temptation to eliminate variables to obtain a smaller data set. Such a reduction of variables, however, often removes information and makes the modelling efforts less reliable. Model interpretation may be misleading and predictive power may deteriorate.

A better alternative is usually to partition the variables into blocks of logically related variables and apply hierarchical data analysis. Such blocked data may be analyzed by PCA and PLS. This modelling forms the base-level of the hierarchical modelling set-up. On the base-level in-depth information is extracted for the different blocks. The score vectors formed on the base-level, here called 'super variables', may be linked together in new matrices on the top-level. On the top-level superficial relationships between the X- and the Y-data are investigated.

In this paper the basic principles of hierarchical modelling by means of PCA and PLS are reviewed. One objective of the paper is to disseminate this concept to a broader QSAR audience. The hierarchical methods are used to analyze a set of 10 haloalkanes for which K = 30 chemical descriptors and M = 255 biological responses have been gathered. Due to the complexity of the biological data, they are sub-divided in four blocks. All the modelling steps on the base-level and the top-level are reported and the final QSAR model is interpreted thoroughly.

Introduction

To understand the world around us, as well as ourselves, we need to measure many things, many variables, many properties of the systems and processes we investigate. Hence, data collected in science, technology, and almost everywhere else are multivariate, with multiple variables measured on multiple samples or at multiple time points. Multivariate data, accurately measured on intelligently selected observations and variables, contain much more information than univariate data, and hence an adequate multivariate characterization is a necessary first step in their investigation.

Traditionally, the phrase *multivariate analysis* has meant multiple linear regression (MLR), linear discriminant analysis (LDA), canonical correlation (CC), factor analysis (FA) and principal component analysis (PCA) applied to independent variables, i.e., cases where the variance-covariance matrix, X′X, has full rank and is fairly well conditioned [1,2].

In QSAR and bioinformatics practice, however, we often assume that our systems are driven by inherent, latent, variables, which are few compared with the number of observed variables, K. Hence, the data and variance-covariance matrices, X and X′X, do not have full rank, but rather have the practical rank, A. Methods used in this situation are PCA [1,3] for projecting X down onto a few latent variables, SIMCA [4] and PLS-DA [5] for classification, and PCR [6] and PLS [7] for latent variable regression. The latent variable

[*] To whom correspondence should be addressed. E-mail: lennart.eriksson@umetrics.com

models are philosophically different in objectives and formulation from the traditional multivariate models with independent variables [8,9].

To distinguish between these two types of situation (with related data, models, and data analytical methods), we have started to refer to the latter as megavariate [10,11]. *Megavariate analysis* models data in terms of multiple latent variables, to give results that are *multivariate*. This is a new nomenclature that will be used in cases where we feel a need to distinguish between the situation where X is full rank and the more common megavariate situation where X has a much lower rank than both the number of variables (K) and the number of observations (N).

Hierarchical PLS and PCA are two emerging megavariate techniques, which simplify the interpretation when dealing with very many variables [12–16]. In such a situation, plots and lists of loadings, weights, and coefficients, tend to become messy and the results are often difficult to overview. Instead of reducing the number of variables, and thus reducing the validity of the model, a better alternative is often to divide the variables into conceptually meaningful blocks and apply hierarchical PCA or PLS. With hierarchical modelling, each block of variables is first summarized by a few score vectors (‘super variables’). Next, these scores are collected and put together in a new matrix, consisting only of the score vectors. This new matrix is then modelled with PCA or PLS (Figure 1).

In QSAR modelling such blocks of data may correspond to different regions of the modelled molecules and different kinds of variables (size descriptors, polarity descriptors, . . .), or variables obtained with different probes in 3D-QSAR. Analogously, the data may also be naturally blocked on the Y-side, e.g., by biological responses relating to acute toxicity, sub-acute toxicity, carcinogenicity, etc.

The objective of this contribution is to disseminate the concept of hierarchical PCA and PLS, and demonstrate their utility for analyzing complex biological and QSAR data. We will use a data set drawn from risk assessment of existing chemicals. For each compound in this data set a total of 255 endpoint measurements are available. The biological data (Y) are divided into four blocks (Figure 2), i.e., endpoint data relating to (i) in vivo acute toxicity, (ii) in vivo subacute toxicity, (iii) in vitro measurements, and (iv) environmental (atmosphere and sediment) persistence.

Example data set: SIRAC

In a joint ***S***wedish ***I***talian project for ***RA***nkning of ***C***hemicals (SIRAC) – lasting roughly a decade between the years 1987 and 1997 – a strategy for systematic analysis and priority ranking of chemicals occurring in the environment was applied to a class of 58 saturated halogenated aliphatics [17]. The training set for this class of compounds consisted of 10 representative compounds, selected by statistical molecular design [18–20]. This set was subjected to a broad chemical and biological characterization.

The selected training set consist of the following 10 chemicals (the numbering of the compounds is preserved from previous publications):

(2) CH_2Cl_2, dichloromethane
(3) $CHCl_3$, trichloromethane
(7) CCl_3F, fluorotrichloromethane
(11) CH_2Cl-CH_2Cl, 1,2-dichloroethane
(15) $CHCl_2$-$CHCl_2$, 1,1,2,2-tetrachloroethane
(30) CH_3-CH_2Br, bromoethane
(33) CH_3-$CHBr_2$, 1,1-dibromoethane
(39) CBr_3F, fluorotribromomethane
(48) CH_3-$CHCl$-CH_3, 2-chloropropane
(52) CH_3-CH_2-CH_2-CH_2Br, 1-bromobutane

The experimental data generated within the SIRAC-project have been reported in the literature together with their QSAR analysis [17–31]. (All the data are downloadable at www.umetrics.com). However, an evaluation of the *complete* body of data has never been published. Hence, this is an incentive behind this paper, i.e., to ‘finally’ communicate the modelling results of the entire SIRAC-project.

Chemical descriptors ($K = 30$) used to map the training set

The following 30 descriptors were used: (x1) molecular weight, Mw; (x2) boiling point, Bp; (x3) melting point, Mp; (x4) density, D; (x5) refractive index; (x6) van der Waals volume, vdW; (x7) octanol-water partition coefficient, log P; (x8), ionization potential, Ip; (x9) number of carbon atoms, nC; (x10) number of bromine atoms, nBr; (x11) number of chlorine atoms, nCl; (x12) number of fluorine atoms, nF; (x13) retention GC system 1, GC1; (x14) retention GC system 2; GC2; (x15) retention HPLC system 1, LC1; (x16) retention HPLC system 2, LC2; (x17) rate constant in Finkelstein reaction, Kf; (x18) relative response to an FID, RFID; (x19) heat of formation, Hf; (x20) electronic energy, EE; (x21) core-core repulsion, CR;

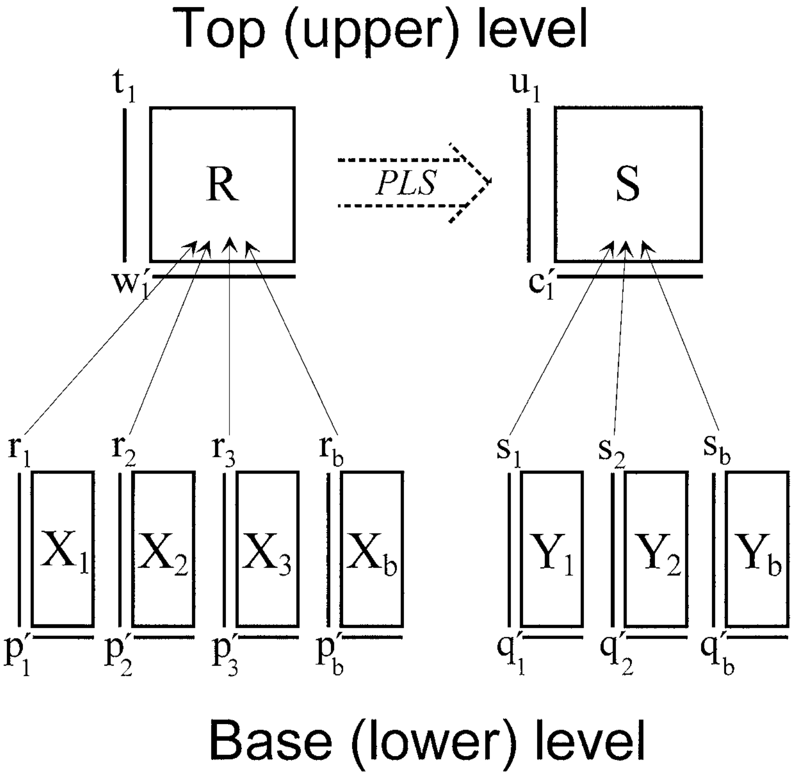

Figure 1. Schematic overview of hierarchical modelling. In the lower (base) level each block of data is modelled locally by either a PCA- or PLS-model. Each block is summarized by one or more score vectors ('super variables'). Note that different number of 'super variables' can be used for the different blocks. The computed 'super variables' are then 'moved' to the upper (top) level and united to form the new 'X' and 'Y' data matrices, depicted as 'R' and 'S', respectively. All conventional PLS statistics and diagnostics are retained.

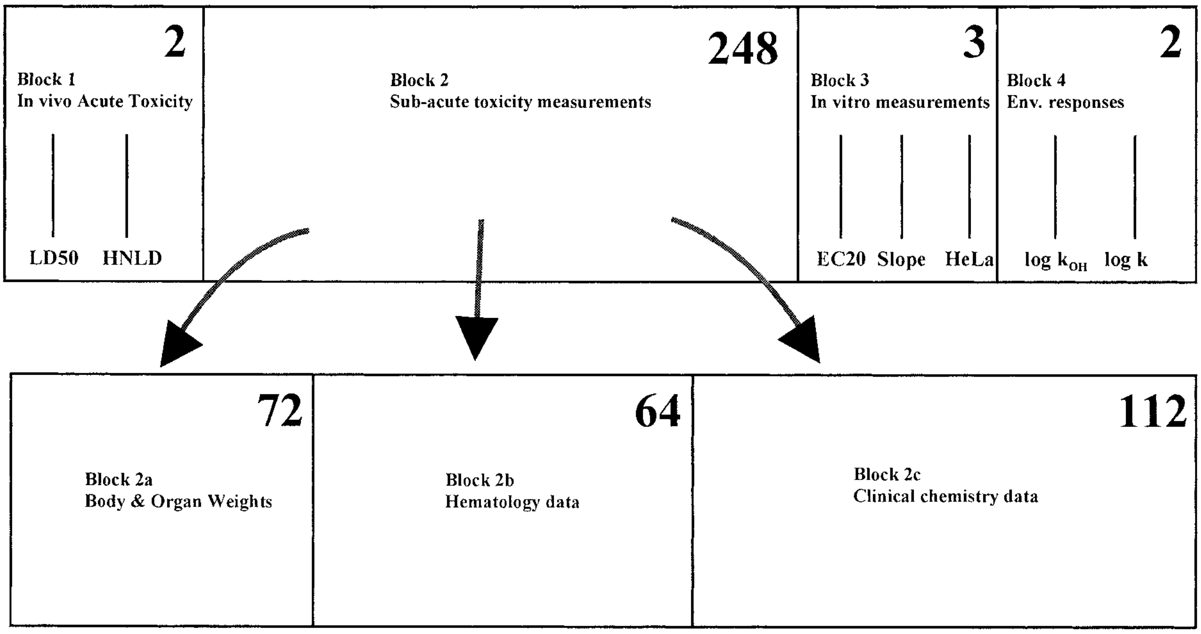

Figure 2. Blocking of Y-data in terms of four blocks. Blocks 2a–2c were individually analyzed on the base-level.

(x22) dipole moment, Dip; (x23) energy of HOMO, HOMO; (x24) energy of LUMO, LUMO; (x25) electronegativity, EN; (x26) hardness, HA; (x27) nucleophilic delocalizability, Dn; (x28) electrophilic delocalizability, De; (x29) polarizability, Pi; (x30) surface area, SA. Experimental and other training set selection details are found in references [21 and 22].

Biological responses ($M_{\text{Total}} = 255$) used to map the training set

The total number of Y-variables is 255. Due to the complexity of this material we decided to sub-divide the biological data into four blocks (Block 1 – Block 4), see also Figure 2. Moreover, because of the size of the second block (248 variables), it was further split into three sub-blocks (Block 2a, 2b, and 2c; Tables 1–3). In the hierarchical data analysis, the following block structure was used:

Block 1: *In vivo* Acute Toxicity, $M_{\text{Block1}} = 2$
- Log Acute toxicity to rat, 'LD50'; [published in references 22 and 27].
- Log Highest non-lethal dose to mouse, 'HNLD', [23].

Block 2: *In vivo* Sub-acute toxicity in albino rat (28 days), $M_{\text{Block2}} = 248$
- Block 2a: Body and Organ weights, 72 responses [29].
- Block 2b: Hematology Data, 64 responses [29].
- Block 2c: Clinical Chemistry Data, 112 responses [29].

Block 3: *In vitro* measurements, $M_{\text{Block3}} = 3$
- Cytotoxicity to Chinese hamster V79 cells, 'EC20'; [25].
- Genotoxicity to V79 cells using DNA precipitation assay, 'Slope'; [25].
- Cytotoxicity to human HeLa cells, 'HeLa', IC50; [30].

Block 4: Environmental responses, $M_{\text{Block4}} = 2$
- Atmospheric persistence, log rate constant with OH-radical, 'k(oh)'; [22].
- Soil/sediment persistence, log k for reductive dehalogenation in sediment/water mixture, 'log k'; [31].

In the measurement of Block 2-data, each training set compound was tested using five groups of rat, i.e., Controls, Low dose, Intermediate dose, High dose, and Satellites, with 5 female and 5 male animals in each group. Dosage levels were determined in advance in screening experiments. The exposure was 28 days and administration was done orally. The satellite group comprises animals exposed to the highest dose, but living 14 days longer, to investigate their ability to recover. This part of the biological data was pre-treated by (1) averaging inside each group, females and males separately, and by (2) standardization against controls [29].

Data analytical methods

In this paper we have used the software SIMCA-P, version 10 [32], and its implementation of standard and hierarchical PCA [1,4] and PLS [4,7].

Principal Components Analysis, PCA

PCA shows the correlation structure of the data matrix **X**, approximating it by a matrix product of lower dimension (**TP′**), called the principal components plus a matrix of residuals (**E**).

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{1}$$

Geometrically, this corresponds to fitting a line, plane or hyper plane to the data in the multidimensional space with the variables as axes. The scaling of the variables specifies the length of the axes of this space. **T** is a matrix of scores that summarizes the X-variables, and **P′** is a matrix of loadings showing the influence of the variables. **E** is a matrix of residuals; the deviations between the original values and the projections. The residual standard deviation (RSD) can be computed for observations and variables. The RSD of an observation (rows in **E**) is also called the observation distance to the PC model (DModX). The RSD of a variable relates to the variable relevance in the PC model. SIMCA iteratively computes one principal component at a time, comprising a score vector $\mathbf{t_a}$ and a loading vector $\mathbf{p'_a}$.

Prior to the analysis data were pre-processed by means of mean-centering and scaling to unit variance.

Partial least squares projections to latent structures, PLS

The regression extension of PCA is called PLS. PLS works with two matrices, **X** (e.g., chemical descriptors) and **Y** (e.g., biological responses), and has two objectives, namely to well approximate **X** and **Y**, and to model the relationship between them. The chemical variation in the predictor block (**X**) is summarized by the X-scores, **T**, and the corresponding variation in the

Table 1. Variable identities of Block 2a, body and organ weights (g).

Identity	Sat(F)	High(F)	Int.(F)	Low(F)	Sat(M)	High(M)	Int.(M)	Low(M)
total body weight	1	10	19	28	37	46	55	64
left adrenal gland	2	11	20	29	38	47	56	65
right adrenal gland	3	12	21	30	39	48	57	66
left ovary	4	13	22	31	–	–	–	–
right ovary	5	14	23	32	–	–	–	–
left kidney	6	15	24	33	40	49	58	67
right kidney	7	16	25	34	41	50	59	68
liver	8	17	26	35	42	51	60	69
spleen	9	18	27	36	43	52	61	70
left testicle	–	–	–	–	44	53	62	71
right testicle	–	–	–	–	45	54	63	72

Comment: The weights of the organs have been converted to relative weights by division with the total body weight (g). Four doses were used, satellite, high, intermediate and low. These doses were administered to female (F) and male (M) animals.

Table 2. Variable identities of Block 2b, hematology parameters.

Identity	Sat(F)	High(F)	Int.(F)	Low(F)	Sat(M)	High(M)	Int.(M)	Low(M)
white blood cells (wbc) [10^9/1]	73	81	89	97	105	113	121	129
red blood cells (rbc) [10^9/lf	74	82	90	98	106	114	122	130
hemoglobin (hgb) [g/l]	75	83	91	99	107	115	123	131
erythrocyte volume fraction (EVF) [%]	76	84	92	100	108	116	124	132
mean corpuscular volume (MCV) [fl]	77	85	93	101	109	117	125	133
mean corp. hemoglobin (MCH) [pg]	78	86	94	102	110	118	126	134
mean corp. hem. Conc. (MCHC) [g/l]	79	87	95	103	111	119	127	135
platelets (PL T) [10^9/1]	80	88	96	104	112	120	128	136

See comment to Table 1 for further explanation of the structure of the variables.

response block (**Y**) is described by the Y-scores, **U**. Basically, what PLS does is to maximize the covariance between **T** and **U**. For each model dimension, a weight vector **w**, is computed, which reflects the contribution of each X-variable to the modelling of Y, in that particular model dimension. The resulting X-weight matrix, **W**, is important since it reflects the structure in **X** that maximizes the squared covariance between **T** and **U**. The corresponding matrix of Y-weights is designated **C**. Additionally, a matrix of X-loadings, **P**, is calculated in order to deflate **X** appropriately.

The decomposition in PLS of **X** and **Y** can be described as:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}; \mathbf{Y} = \mathbf{TC}' + \mathbf{F} \tag{2}$$

The set of PLS regression coefficients can then be computed according to:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{C}' \tag{3}$$

Subsequently, an estimate of **Y**, $\hat{\mathbf{Y}}$, is obtained as:

$$\hat{\mathbf{Y}} = \mathbf{XW}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{C}' = \mathbf{XB} \tag{4}$$

Prior to the data analysis data were pre-processed by means of mean-centering and scaling to unit variance.

Hierarchical PCA and PLS models

The idea with hierarchical PCA and PLS is to work with blocking of data in order to improve clarity and interpretability. Both these methods operate on two or more levels (Figure 1). On the lower level the details of each block are accounted for. This analysis provide the score vectors ('super variables'), which are used to construct the data matrices of the upper level. On the upper level, a simple overview-type-of-relationship between rather few super variables is developed.

On each level, 'standard' PLS or PCA-scores and loading plots, as well as residuals and their summaries

Table 3. Variable identities of Block 2c, clinical chemistry data.

Identity	Sat(F)	High(F)	Int.(F)	Low(F)	Sat(M)	High(M)	Int.(M)	Low(M)
K	137	150	163	176	189	202	215	228
Na	138	151	164	177	190	203	216	229
Cl	139	152	165	178	191	204	217	230
Urea	140	153	166	179	192	205	218	231
TSP	141	154	167	180	193	206	219	232
Alb	142	155	168	181	194	207	220	233
A/G	143	156	169	182	195	208	221	234
Chol	144	157	170	183	196	209	222	235
AP	145	158	171	184	197	210	223	236
Alat	146	159	172	185	198	211	224	237
Ca	147	160	173	186	199	212	225	238
P	148	161	174	187	200	213	226	239
Mg	149	162	175	188	201	214	227	240
Glucose	241	242	243	244	245	246	247	248

Comment: K, Na, Cl, Ca, P, Mg, Urea, Chol (Cholesterol), represent the concentration in serum [mM]. TSP is total serum protein [g/l]. Alb is albumin [g/l]. A/G is the albumin/globulin ratio. AP is the activity of alkaline phosphatase [μkat/l]. Alat is the activity of alanine-aminotransferase [μkat/l]. See comment to Table 1 for further explanation of the structure of the variables.

such as DModX, are available for model interpretation. This allows an interpretation focused on pertinent blocks and their dominating variables. For further details reference is given to the literature [14,15,32].

Prior to the hierarchical analysis, data were preprocessed by means of mean-centering and scaling to unit variance.

Results

The SIRAC data-set will here be analyzed according to the following three-stage hierarchical approach:

(i) Base-level PCA of endpoint data
(ii) Top-level PCA of the base-level PCA results
(iii) Top-level PLS (QSAR) modelling relating a multivariate set of 30 physico-chemical descriptors (X) to the multivariate endpoint data (Y).

In the base-level PCA, stage (i), we will not extract more than two components for each block because of the low number of compounds in the training set (N = 10). Also note that only Blocks 2a, 2b, and 2c are large enough to warrant base-level analysis.

Base-level modelling of body and organ weights (Block 2a)

A two-component PCA model was fitted to the 72 variables of Block 2a. This model accounted for 51% of the variance ($R^2X = 0.51$; $A = 1 \Leftrightarrow 0.28$; $A = 2 \Leftrightarrow 0.23$). Figure 3a displays the score plot of this model. This plot shows that exposure to $CHCl_3$ (compound no. 3), CH_2Cl-CH_2Cl (no. 11) and $CHCl_2$-$CHCl_2$ (no. 15) strongly influence body and organ weights. The other three graphs (Figures 3b–3d) originate from the same loading plot but with different parts of the primary variable name displayed. In Figure 3b we can see some grouping according to sex. Compounds 11 and 15 inflict lowered organ weights in females. There is no grouping according to dose (Figure 3c), but some grouping according to organ (Figure 3d). Primarily adrenal glands, ovaries, kidneys, and spleen are positioned in the positive end of the first principal component.

Base-level modelling of hematology parameters (Block 2b)

A two-component PCA model accounting for 54% of the variance was calculated ($R^2X = 0.54$; $A = 1 \Leftrightarrow 0.33$; $A = 2 \Leftrightarrow 0.21$). Figure 4 provides some results for this model. The score plot (Figure 4a) shows that exposure to CH_2Cl-CH_2Cl (11), and CH_3-$CHCl$-CH_3 (48) influence hematology most strongly. In the top right plot (Figure 4b) we see no grouping according to sex. Similarly, there is no grouping according to dose (Figure 4c, bottom left). However, there is some grouping according to hematology parameter

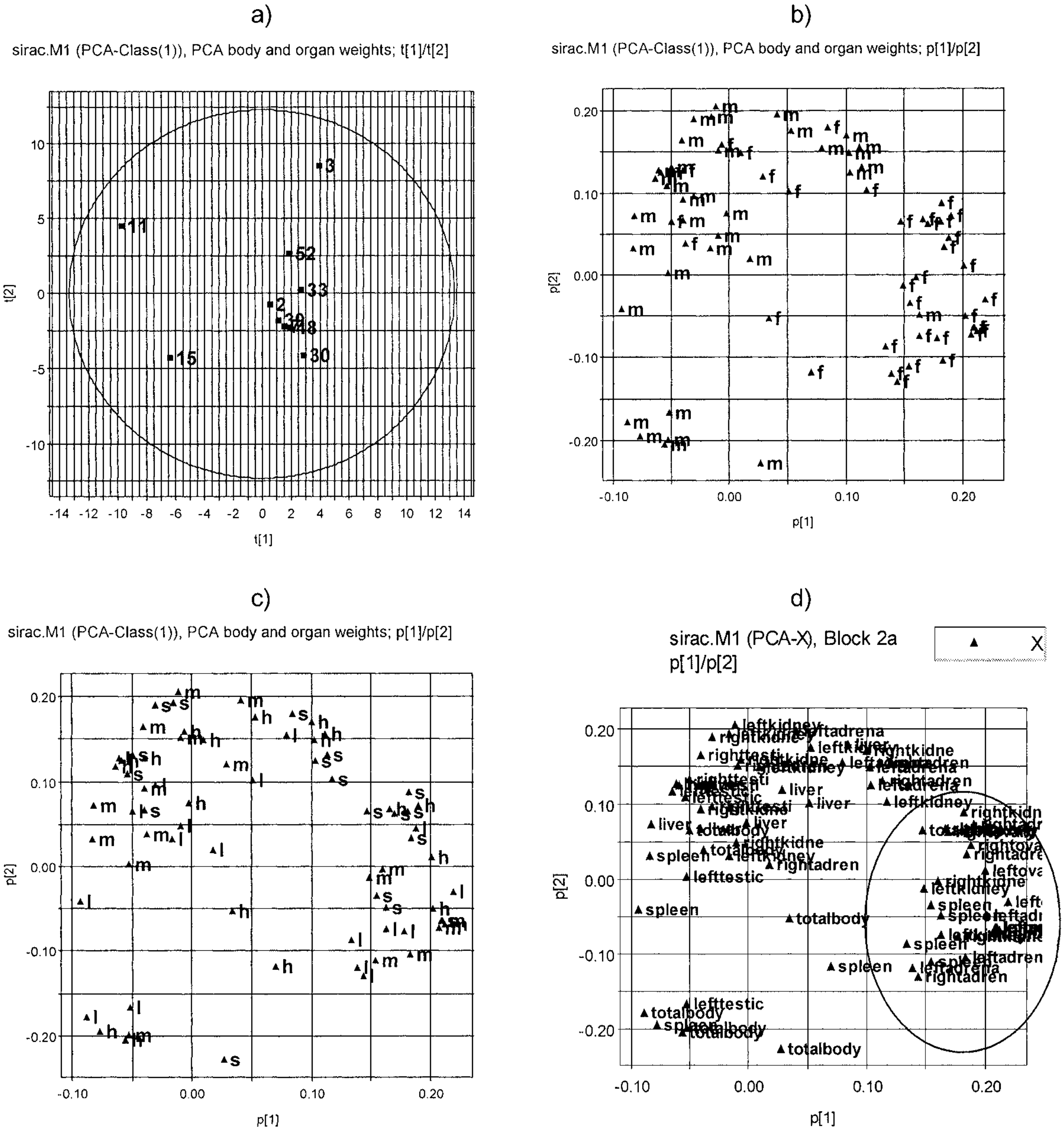

Figure 3. (a, top left). Score plot t_1/t_2 of the base-level PCA of Block 2a. The compounds are numbered in accordance with the text in Section 2. (b, top right). Loading plot p_1/p_2 corresponding to the previous figure. Sex is used as plot mark: f = female rat, m = male rat. (c, bottom left). Loading plot p_1/p_2 with dose level as plot mark: l = low, m = medium, h = high, s = satellite. (d, bottom right). Loading plot p_1/p_2 with inner organ as plot mark (see description in Table 1).

measured (Figure 4d, bottom right). Variables of the type rbc, hgb, and evf are predominantly located in the lower left-hand corner of the loading plot. Variables related to mcv, mch, mchc, and wbc are positioned in the lower right-hand part.

Base-level modelling of clinical chemistry data (Block 2c)

A two-component PCA model describing 40% of the variance was computed. ($R^2X = 0.40$; $A = 1 \Leftrightarrow 0.21$; $A = 2 \Leftrightarrow 0.19$). This model has the lowest explained variance of the three base-level models, a result, in part, explicable by the fact that this is the largest block.

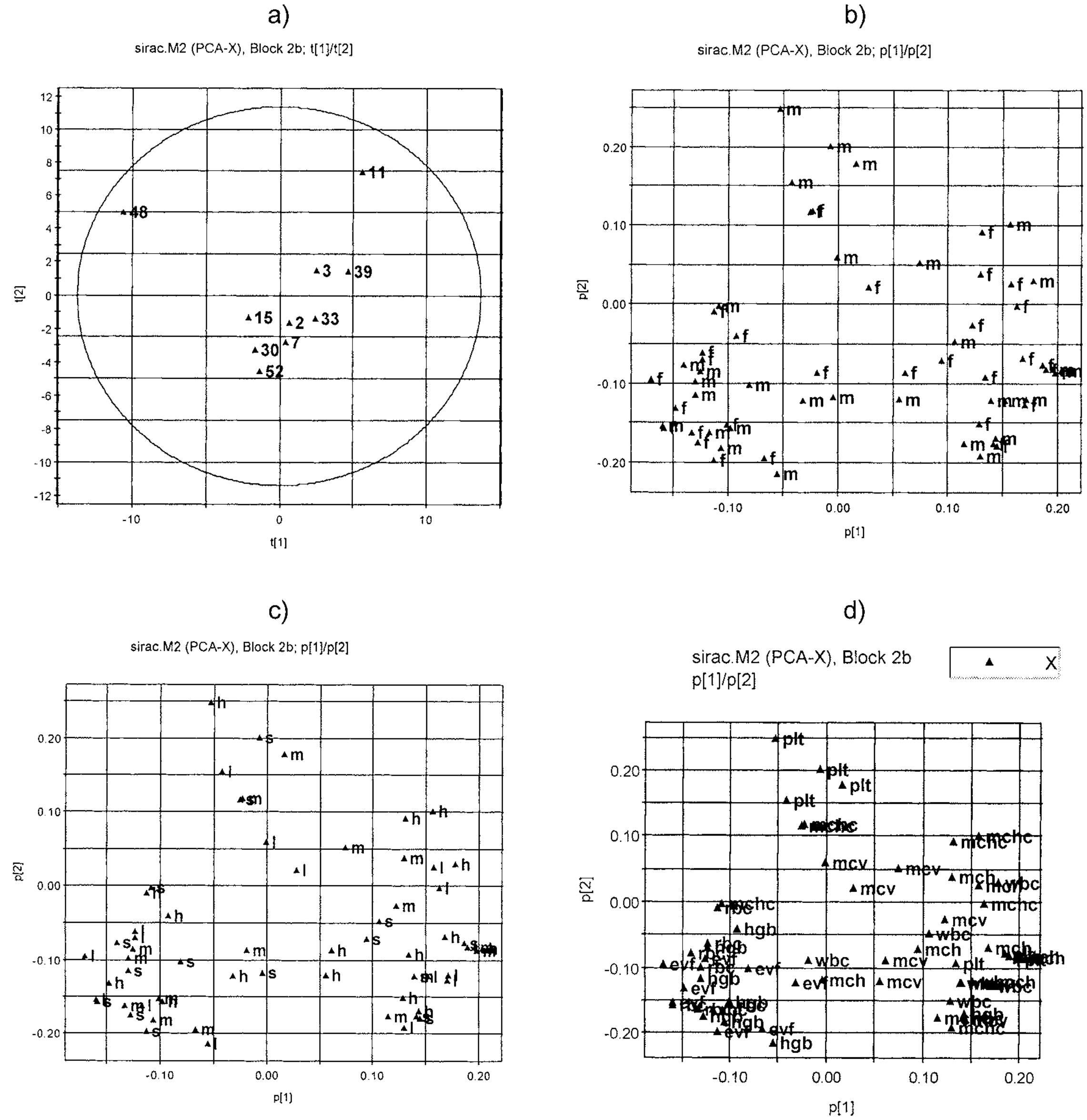

Figure 4. (a, top left). Score plot t_1/t_2 of the base-level PCA of Block 2b. Notation as in Figure 3a. (b, top right). Loading plot p_1/p_2 corresponding to the previous figure. Sex is used as plot mark: f = female rat, m = male rat. (c, bottom left). Loading plot p_1/p_2 with dose level as plot mark: l = low, m = medium, h = high, s = satellite. (d, bottom right). Loading plot p_1/p_2 with hematology parameter as plot mark. The hematology parameters are explained in Table 2.

The score plot (Figure 5a, top left) suggests that exposure to $CHCl_3$ (3) causes the largest changes in clinical chemistry profiles. In Figure 5b (top right) we see no grouping according to sex. Similarly, there is no grouping according to dose (Figure 5c, bottom left). Interestingly, however, there is some grouping according to clinical chemistry profiles (Figure 5d, bottom right). The liver enzymes (AP and Alat – boxed) are situated in the upper right-hand area of the loading plot. This is strong evidence that trichloromethane attacks the liver.

Top-level PCA of all toxicity data

The six computed 'super variables' of Blocks 2a–2c were merged with the variables of Blocks 1, 3, and 4, thus producing the data structure displayed in

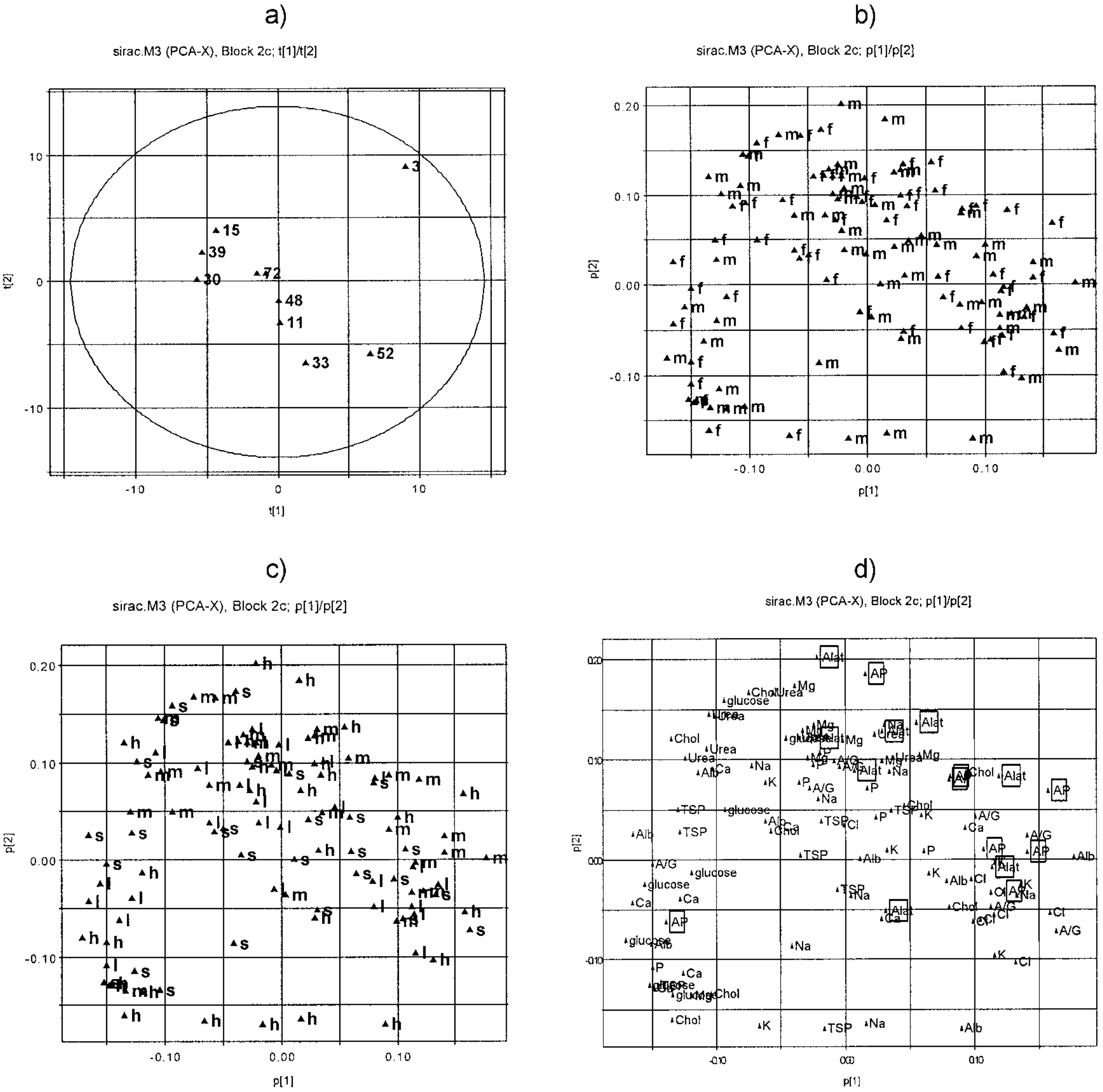

Figure 5. (a, top left). Score plot t_1/t_2 of the base-level PCA of Block 2c. Notation as in Figure 3a. (b, top right). Loading plot p_1/p_2 corresponding to the previous figure. Sex is used as plot mark: f = female rat, m = male rat. (c, bottom left). Loading plot p_1/p_2 with dose level as plot mark: l = low, m = medium, h = high, s = satellite. (d, bottom right). Loading plot p_1/p_2 with clinical chemistry variable as plot mark. The clinical chemistry variables are described in Table 3. The two liver enzymes AP and Alat are boxed, see text for deeper discussion.

Figure 6. As seen, there are 13 biological variables available on the top-level.

In order to overview the biological data, the two first principal components were calculated (R^2X = 0.57; A = 1 ⇔ 0.33; A = 2 ⇔ 0.24). From the score plot seen in Figure 7a it is obvious that compound 39 (CBr_3F) is the most extreme chemical. It is very toxic and reactive in the environment.

The interpretation is uncomplicated since the loading plot is ‘clean’. The first component reflects environmental persistence and toxicity and compounds 2, 39 and 48 contribute to the model in this direction. The second component models acute and sub-acute toxicity with compounds 3, 7 and 11 being the most prominent in this direction. Compounds 15, 30, 33, and 52 are less characteristic. We can also see that the two acute toxicity scales (LD50 and HNLD) are correlated.

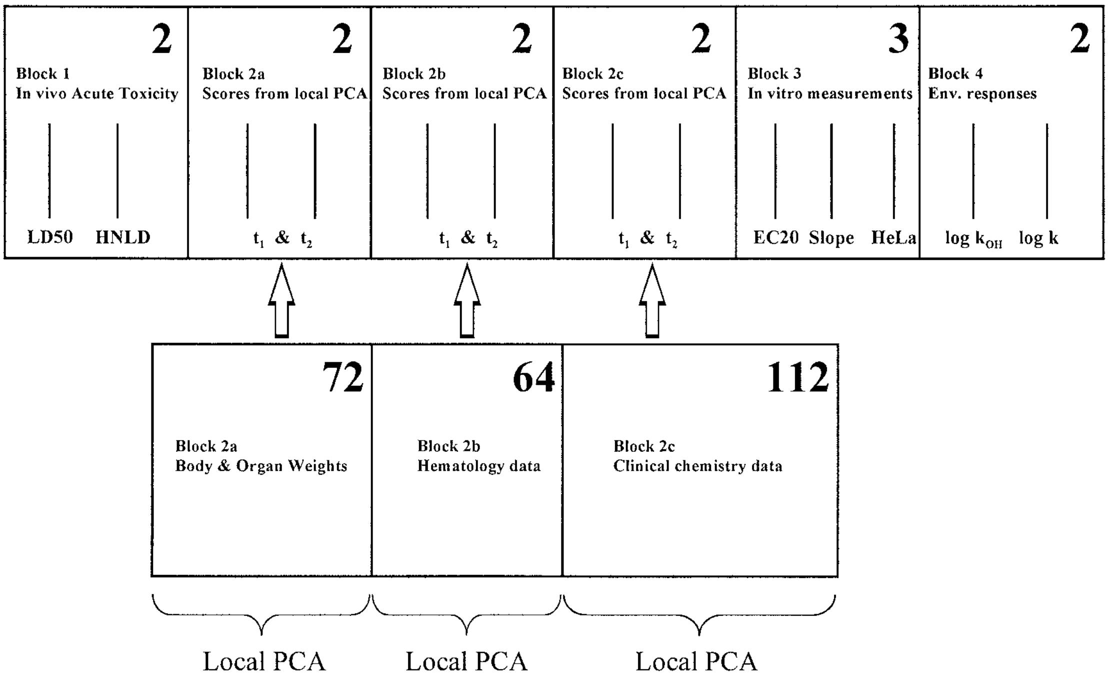

Figure 6. Preparation of Y-data for top-level modelling. Each of the three blocks 2a, 2b, and 2c, were summarized by means of two principal component score vectors. In total, the Y-matrix of the top level comprises 13 biological responses.

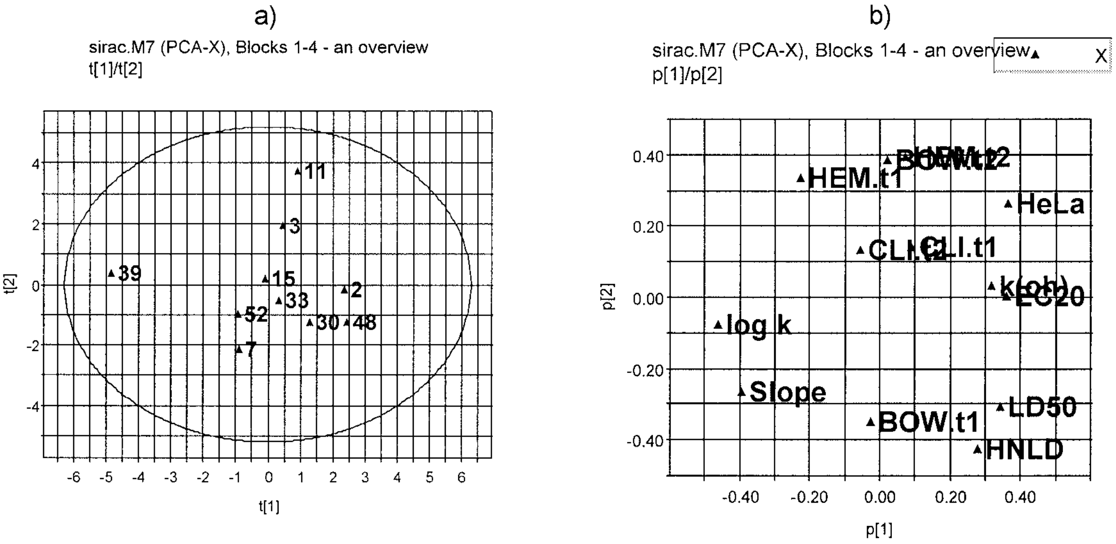

Figure 7. (a, left) Score plot t_1/t_2 of the top-level PCA of the Y-data. Notation as in Figure 3a. (b, right) Loading plot p_1/p_2 corresponding to Figure 7a. Except for the six ‘super variables’ the biological responses are abbreviated as presented in Section 2.2. The six ‘super variables’ are encoded as follows: BOW.t1 and BOW.t2, first and second score vectors of Block 2a; HEM.t1 and HEM.t2, first and second score vectors of Block 2b; CLI.t1 and CLI.t2, first and second score vectors of Block 2c.

Top-level PLS (QSAR) modelling of all chemical (X) and toxicity (Y) data

In the final QSAR modelling the data were configured as outlined in Figure 8. In this case, PLS gave the unrealistic model complexity of nine components, all significant according to cross-validation. Since only 10 compounds are modelled, a maximum of 3–4 components should be used. With A = 4, explained variances are: X = 96%/Y = 60%. The results of this model are presented graphically in Figure 9.

The two score plots of the t/t-type (Figures 9a and 9c) demonstrate that the data set is indeed homogeneous and devoid of strong, influential outliers. This means that one of the most important conditions for QSAR modelling – homogeneity [33] – is fulfilled.

Additional t_1/u_1 - t_4/u_4 score plots are not given. However, they show good correlation between X and Y. The DModX and DModY statistics (no plots given) do not uncover any compounds with worryingly large residuals. Hence, we conclude that the data set is homogenous also in the relation between X and Y, and without sub-groupings and influential outliers.

Further interpretation of this model is conducted.

Discussion

General methodological aspects

Simplified interpretability

In many application areas of science and technology a large number of variables are routinely collected. These areas include e.g. multivariate calibration (spectroscopy, chromatography, electrophoresis), QSAR (2D, 3D), bioinformatics, and process modelling and monitoring. Projection methods such as PCA and PLS are very useful for the analysis of such data sets. In this article it is demonstrated how conventional PCA and PLS models can be configured by means of blocking of data to give hierarchical models.

The interpretation of a hierarchical model is simpler than that of a conventional unblocked model. By structuring the data in two or more levels, the model parameters are 'spread' across the different layers of a model – where the block level is the lowest level – giving fewer parameters to interpret at each level. The base level/top level modelling set-up thus creates a flexible framework whereby it is possible to overview the major trends and to zoom-in more detailed phenomena. On the highest level it will be possible to visualize relationships between blocks, or block-aggregates in models employing more than two model layers. Once the dominant blocks have been identified, these can be interpreted by inspecting plots of block scores and block loadings. On the lowest level more detailed information about relationships inside a particular block can be retrieved.

Besides simplified interpretability, blocked models offer a number of other interesting advantages of both practical and technical character. A few of these aspects are discussed in the ensuing three paragraphs.

An alternative to variable selection

First of all, we may understand the hierarchical modelling approach as an alternative to variable selection. This is of great importance in QSAR analysis, as there is often a strong temptation to remove both uninformative and redundant variables from large data sets in order to possibly improve the predictive power and interpretability of a QSAR-model. However, one should be very careful when reducing the number of variables, because this can be done in so many ways, so almost any result is possible. In variable selection, it is important to test the predictive power of the model on real new data, and not just cross-validation with the training set. In multivariate data, most of the X-variables contain at least some information about Y. Hence, one can hope for only a mild variable reduction, as usually not more than 20 to 30% of the variables have less information than the noise level [11].

Moreover, because of the correlations among the more important X-variables, one can continue to reduce the X-variables further than these 20 to 30% with no apparent decrease in fit. This makes the remaining X-variables take over importance from the ones that are deleted, and a serious bias is introduced. Thus the interpretation of the model shifts and some variables take the role as being the ones related to Y while other variables correlated to these have been deleted, and hence are forgotten in the interpretation. This also makes the prediction power of the model deteriorate because the correlations are not perfectly stable, and for new samples / molecules important variables are now missing in the model.

Also, one should remember that even seemingly unimportant variables still have a role in diagnosing outliers. Take, for instance, a variable that is almost constant in the training set, and hence will appear as unimportant in the QSAR-model. If a new compound has a value of this variable that substantially differs from its values in the training set, this is an indication that this compound is different, and hence predictions of its Y-values are doubtful. If one mechanically deletes all variables that do not contribute to the modelling of Y in the training set, one automatically decreases the possibility to find outliers among the new observations (compounds).

Hence, in summary, any means to avoid variable selection in QSAR is warranted, such as, blocking of data coupled with hierarchical analysis.

An alternative to block-scaling

As mentioned, prior to the analysis data were pre-processed by means of mean-centering and scaling to unit variance. This combination of pre-processing methods is sometimes referred to as 'auto-scaling'. One limitation of this scaling approach is that it does not consider whether variables are grouped in blocks of naturally related descriptors, or the number of variables in each such block. If, for example, UV-scaling is used, a large block of variables will dominate over a smaller block of variables, for purely numerical reasons. This is often not wanted.

We will now consider this problem. Consider a health related questionnaire completed by a number of people. In this questionnaire, there are ten questions regarding body fitness and exercising habits, 30 questions about food consumption patterns, five questions related to tobacco and alcohol usage, and three measured variables indicating body weight, blood pressure, and cholesterol level in blood.

The three measured variables probably convey more information about the health status of an individual than all the other variables put together. However, there is an apparent risk that the quantitative information of the measured variables will be masked by variation in the other 45 variables. With UV-scaling the former block would get a total variance of three, while the latter block is assigned a total variance of 45. Thus, even though scaling to unit variance has been performed, the questionnaire block will dominate over the measured one.

One way of addressing this problem is to employ *block-scaling* [11]. Also referred to as *battery-scaling*, this corresponds to down-weighting blocks of variables in relation to a selected basis scaling procedure. The basis scaling method is generally UV-scaling, especially when variables are markedly different in nature and numerical range. However, in multivariate calibration, procedures like no scaling or Pareto scaling may well be used as the basis for block-scaling [11].

Block-scaling can be done in many ways. In our experience, it is convenient to distinguish between soft and hard block-scaling. In soft block-scaling, each block of variables is scaled such that the sum of the variable's variances (after completed scaling) equals the square root of the number of variables in that particular block. Here, the additional scaling weight used is $1/(k_{\text{block}})^{1/4}$ – where k_{block} represents the number of variables in a block – which is multiplied by the basis scaling weight. Hard block-scaling involves even further down-weighting. With this approach the variables in a block are scaled such that the sum of their variances is unity. Here, the additional scaling weight used is $1/(k_{\text{block}})^{1/2}$.

With hierarchical models this scaling obscurity largely disappears. Due to the separate analysis of each block at the base level, each block will automatically acquire approximately the same influence on the upper level, as long as the number of score vectors calculated for each block is not markedly different. However, as outlined by Wold and co-authors [14], it might be desirable to weight blocks (in terms of the 'super variables') on the top level according to their importance. Such a focusing on important blocks can be accomplished by simple up- or downweighting. A procedure for this is suggested in reference 14.

Improved stability and predictive ability

As discussed by Wold and coworkers [14], hierarchical models often provide better predictions than the unblocked ones, as long as the data are divided into conceptually meaningful blocks. One reason for this is that blocked models are less sensitive to deficiencies in the scaling of the original variables and also easier cope with outliers.

In the SIRAC example, blocking of data was only imposed on the Y-variables and so we used PCA for their analysis. When hierarchical relationships exist also in the X-matrix, however, it is conceivable to replace PCA by PLS in the base-level modelling. This will lead to a projection of each block of X-data directed towards the information of relevance for modelling and predicting Y. This will, in turn, on the top-level, result in a stabilized PLS-model with improved modelling and predictive power.

It is noted that in the analysis of the example data, we could alternatively have used the X-matrix of chemical data and PLS in the base-level modelling step. The analysis would then have been formulated as an 'inverse' calibration with the roles of X and the respective Y-blocks swapped. For simplicity we here choose to work with PCA.

Discussion of SIRAC example

Non-trivial complexity of data

The SIRAC data set is unique in the sense that it contains very many responses, which far outnumber the chemical descriptors. The ratio M/K (number of Y-variables/number of X-variables) is 8.5. This means that the latent variable complexity is probably higher in Y than in X. Usually, in regression analysis the practical rank is higher in X than in Y. As mentioned in the foregoing section, by interchanging the X- and Y-blocks it would still be possible to run PLS on this data set. However, such an 'inverse' QSAR, formulated so as to model and predict chemical data (X) from biological data (Y), is rather exotic and perhaps also controversial.

Furthermore, regardless of any choice of data arrangement, the analysis of the full set of X- and Y-variables will create loading plots that are cluttered and arduous to interpret. Hence, it will be difficult to overview and understand the relationships between the X- and Y-blocks, and among the Y-blocks themselves. In our opinion, hierarchical modelling offers a good solution to this obstacle, by creating parsimonious models that are logically linked in two or more levels, and which are easy to interpret. The loading plots displayed in Figures 7 and 9 are clean and easily digested, compared with the almost overwhelming number of loading points plotted in Figures 3–5.

Base-level modelling provides detailed insight of each block

The Y-variables of the SIRAC data set were partitioned into four blocks, see the outline. The blocks enumerated 1, 3, and 4, were small and contained 2, 3, and 2 response variables, respectively. Consequently, they were not summarized by any 'super variable' on the base-level.

Conversely, the second block was very large, and it was in fact possible to fractionate this block even further as three sub-sets, here denoted 2a, 2b, and 2c. The disjoint PCA-models applied to these three blocks were able to highlight different haloalkanes as triggering different biological response profiles. It was clearly demonstrated that exposure to trichloromethane, 1,2-dichloroethane, and 1,1,2,2-tetrachloroethane had strong and characteristic impact on the body and organ weights (Figure 3a). It was concluded that the former two compounds chiefly affected female animals.

In a similar way, the analysis of Block 2b, pointed to 1,2-dichloroethane and 2-chloropropane as inducers of changes in hematology parameters (Figure 4a). There was some grouping in the variable space according to type of hematology parameter, but there was no clear grouping among the variables neither due to dose level nor sex.

The last base-level PCA-model (of Block 2c) identified trichloromethane as a strong liver attacking agent (Figure 5a). This was manifested by it eliciting elevated activity levels of two liver enzyme-related variables (coded as AP and Alat in Figure 5d).

Interpretation of Y-blocks 1, 3, and 4

Although the Y-variables incorporated in Blocks 1, 3, and 4, have not been explicitly modelled on the base-level in this article, it is still worthwhile to construe their meaning. As reported in papers [22, 23, 25, 27, 28, 30 and 31], one generally very potent training set member is the compound fluorotribromomethane (CBr_3F, no. 39). This compound is the most acutely toxic, genotoxic, and cytotoxic substance included in this set of compounds. Additionally, it is also very reactive in the two environmental response model systems here employed.

These extreme properties of compound no 39 is jointly expressed by the seven biological variables of Blocks 1, 3, and 4. If these variables were to be analyzed in conjunction with the other 248 Y-variables (of Blocks 2a–2c), these hallmarks of fluorotribromomethane would simply drown in the information of the sub-acute toxicity data (collected in Blocks 2a–2c), where this compound is not behaving remarkably in any way.

Thus, what we are trying to express, is a firm belief that one benefit of the hierarchical approach is that compound 39 is given a fair chance to be expressed in the modelling of the Y-data. This is because only six 'super variables' are used to encode the sub-acute toxicity data on the highest level, and this number well matches the number of Y-variables (i.e., seven) belonging to Blocks 1, 3, and 4.

Summarizing the Y-data: Top-level PCA-modelling

That compound 39 is indeed very potent is seen in Figure 7a, which represents a score plot of the first two PCA score vectors of the top-level model. The corresponding loading plot, seen in Figure 7b, suggests that

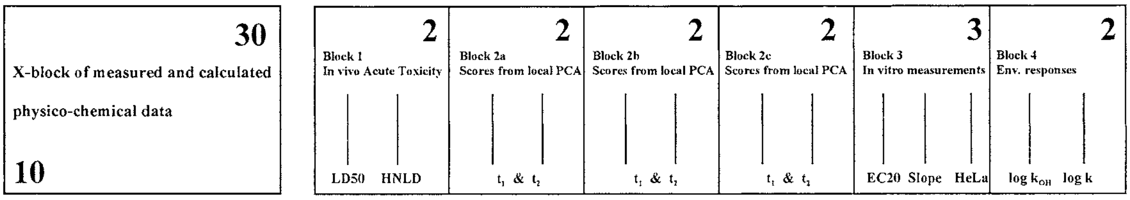

Figure 8. Arrangement of top-level QSAR data.

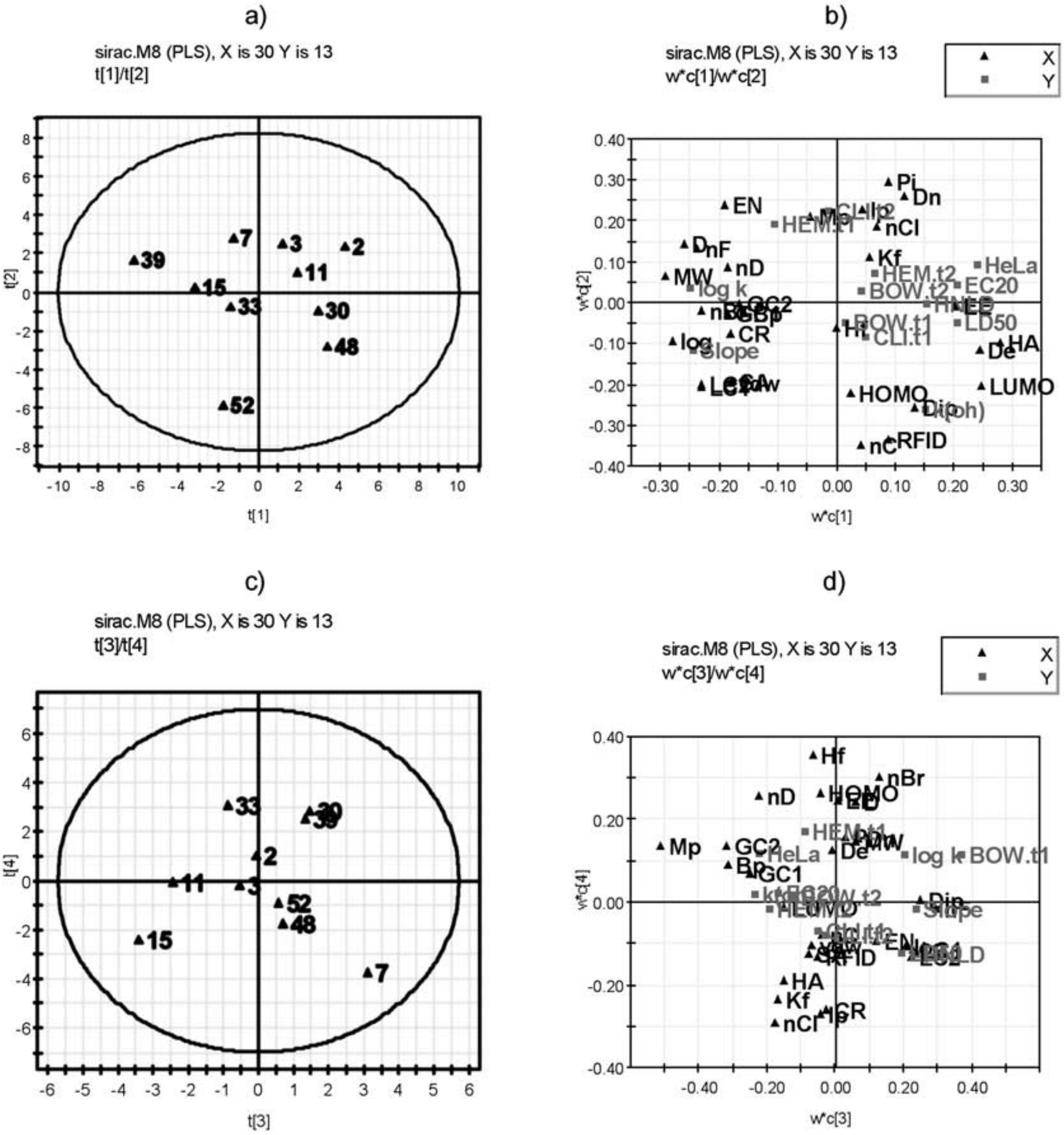

Figure 9. (a, top left). Score plot t_1/t_2 of the top-level PLS. Notation as in Figure 3a. (b, top right). Loading plot wc_1/wc_2 corresponding to Figure 9a. For a description of the variables, see the text, and the legend to Figure 7b. (c, bottom left). Score plot t_3/t_4 of the top-level PLS. Notation as in Figure 3a. (d, bottom right). Loading plot wc_3/wc_4 corresponding to the Figure 9c. Notation as in Figure 9b.

the first component can be seen as a general short-term toxicity scale as all short-term toxicity responses load strongly in this dimension. This component also models the two environmental responses well. The second component models the sub-acute toxicity data and to some extent the two acute toxicity responses (HNLD and LD50).

Compounds 2, 39, and 48 belong to those that influence the first component, whereas compounds 3, 7, and 11 have a strong impact in the second component. Compounds 15, 30, 33, and 52 have less characteristic biological features.

The final goal: Top-level PLS (QSAR) modelling

In the final stage, the battery of 30 X-variables was related to the ensemble of 13 Y-variables. Some model details of this model were reported. The idea is now to conduct a more careful interpretation of the QSAR-model. This interpretation is based on the two loading plots rendered in Figures 9b and 9d.

These PLS-loadings indicate the following:

- Component 1: This component separates the molecules due to their molecular weight and hydrophobicity. The heaviest and most hydrophobic compounds are located in the left-hand part of Figure 9a. Many biological responses are found at the positive end of component 1. Hence, we conclude that toxicity correlates with hydrophobicity and size.
- Component 2: Chemical descriptors like polarizability, electronegativity and number of carbon atoms dominate in this component. This dimension reflects reactivity with the hydroxyl radical, i.e., persistence in atmosphere, and compounds with high electronegativity and polarizability are modelled to be highly reactive with the hydroxyl radical. This component also models the first summary score vector ('super variable') of hematology data (HEM.t1) and the second summary of clinical chemistry data (CLI.t2).
- Component 3: The third component is strongly influenced by the first summary of body and organ weights (BOW.t1). Chemical descriptors like dipole moment, melting point, and number of chlorine atoms best explain the behavior of BOW.t1 in this dimension.
- Component 4: The last component only accounts for 5% of the Y-variance, and it is most affected by X-descriptors like ionization potential and heat of formation.

A final remark about cross-validation

In any empirical modelling, it is essential to determine the correct complexity of a model. With many and correlated X- and/or Y-variables there is a substantial risk for 'over-fitting', i.e., getting a well fitting model with little or no predictive power. Hence a strict test of the significance of each consecutive PCA/PLS-component is necessary, and then stopping when components start to be small and non-significant.

Cross-validation (CV) is a practical and reliable way to test this significance [34–36]. CV is performed by dividing the data in a number of groups, and then developing a number of parallel models from reduced data with one of the groups deleted.

After developing a model, the deleted data are used as a test set, and the differences between actual and predicted data are calculated for the test set. The sum of squares of these differences are computed and collected from all the parallel models to form PRESS (predictive residual sum of squares), which is a measure of the predictive ability of the model. PRESS may also be transformed into the dimensionless quantity Q^2X or Q^2Y depending on which set of data is being forecast.

We have here developed every model using cross-validation, and these have all cross-validated significantly according to default cross-validation settings in the software used [32]. However, three phenomena complicate the use of cross-validation in this work.

First of all, there is a statistical molecular design underpinning the distribution of the points of the training set. Any exclusion of a molecule (compound) in cross-validation will, therefore, tend to erase the good spanning these molecules exhibit in the latent variable space of the experimental design. Hence, the preferred projection direction may change quite dramatically between different cycles in cross-validation, which will produce unstable and too pessimistic estimates of Q^2.

Secondly, in the top-level PCA and PLS models, some of the responses are orthogonal, because they were derived with PCA. Such orthogonal responses will (i) tend to give the top-level model more components, and (ii) they will also cross-validate worse than 'normal' responses.

Thirdly, in the base-level modelling of blocks 2a, 2b, and 2c, some molecules were found to have very characteristic, almost extreme, properties. It is conceivable that in cross-validation, the presence of some of these molecules may degrade the predictive power. However, it is emphasized that the objective of the base-level modelling was to accomplish a summary of the original variables, and not necessarily to optimize these three models' predictive power.

As a consequence of these three aspects, the Q^2s are less reliable and therefore also less useful than under more 'normal' modelling circumstances. The computed Q^2 (X and/or Y) values generally vary between 0.2 and 0.6.

Conclusions

This example shows how hierarchical multivariate analysis simplifies the interpretation of complex problems. The base level/top level modelling setup creates a flexible framework whereby it is possible to overview the major trends and to zoom-in onto more detailed phenomena. The detailed analysis of Blocks 2a, 2b, and 2c revealed that different compounds in the training set caused on-set of different biological response profiles. For instance, it was very clear in the analysis of Block 2c that trichloromethane (chloroform) triggered activity in the liver. The top level QSAR model highlighted the very special properties of compound 39 (fluorotribromomethane). Finally, we note that in the SIRAC-example hierarchical modelling was carried out only with regards to Y-data; generally, it is possible to conduct hierarchical modelling where both X and Y are engaged at both levels of modelling.

References

1. Jackson, J.E, *A User's Guide to Principal Components*, John Wiley, New York, 1991, (ISBN 0-471-62267-2).
2. Martens, H., and Naes, T., *Multivariate Calibration*, John Wiley & Sons, NY, 1989, (ISBN 0-471-90979-3).
3. Wold, S., Esbensen, K., and Geladi, P., Chemom. Intel. Lab. Syst., 2 (1987) 37.
4. Wold, S., Albano, C., Dunn, W.J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjöström, M., In: Kowalski, B.R. (Ed.) *Chemometrics: Mathematics and Statistics in Chemistry*, D. Reidel Publishing Company, Dordrecht, Holland, 1984.
5. Sjöström, M., Wold, S., and Söderström, B., *PLS Discriminant Plots*, Proceedings of PARC in Practice, Amsterdam, June 19-21, 1985.
6. Kalivas, J.H., J. Chemom., 13 (1999) 111.
7. Wold, S., Johansson, E., and Cocchi, M., In Kubinyi, H., (Ed.), *3D-QSAR in Drug Design, Theory, Methods, and Applications*, ESCOM Science Publishers, Leiden, 1993, pp. 523-550.
8. Burnham, A.J., Viveros, R., and MacGregor, J.F., J. Chemom., 10 (1996) 31.
9. Burnham, A.J., MacGregor, J.F., and Viveros, R., Chemom. Intel. Lab. Syst., 48 (1999) 167.
10. Eriksson, L., and Earll, M., The 14th European Symposium on Quantitative Structure-Activity Relationships, Bournemouth, UK, September 8-13, 2002.
11. Eriksson, L., Johansson, E., Kettaneh-Wold, N., and Wold, S., Multi- and Megavariate Data Analysis – Principles and Applications, Umetrics AB, 2001, ISBN 91-973730-1-X.
12. Berglund, A., De Rosa, M.C., and Wold, S., J. Comp.-Aid. Mol. Des., 11 (1997) 601.
13. Westerhuis, J., Kourti, T., and MacGregor, J.F., J. Chemom., 12 (1998) 301.
14. Wold, S., Kettaneh. N., and Tjessem, K., J. Chemom., 10 (1996) 463.
15. Rännar, S., MacGregor, J.F., and Wold, S., Chemom. Intel. Lab. Syst., 41 (1998) 73.
16. K. Janné, J. Pettersen, N.-O. Lindberg and T. Lundstedt, J. Chemom., 15 (2001) 203.
17. Tosato, M.L., Marchini, S., Passerini, L., Pino, A., Eriksson, L., Lindgren, F., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B. and Wold, S., Environ. Toxicol. Chem., 9 (1990) 265.
18. Gabrielsson, J., Lindberg, N.O., and Lundstedt, T., J. Chemom., 16 (2002) 141.
19. Linusson, A., Gottfries, J., Lindgren, F., and Wold, S., J. Med. Chem., 43 (2000) 1320.
20. Giraud, E., Luttman. C., Lavelle, F., Riou, J.F., Mailliet, P., and Laoui, A., J. Med. Chem., 43 (2000) 1807.
21. Eriksson, L., A *Strategy for Ranking Environmentally Occurring Chemicals*, Ph.D. Thesis, Umeå University, Umeå, Sweden, 1991.
22. Eriksson, L., Rännar, S., Sjöström, M. and Hermens, J.L.M., Environmetrics 5 (1994) 197.
23. Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjöström, M., Wold, S. and Berglind, R., Environ. Toxicol. Chem., 9 (1990) 1339.
24. Lindgren, F., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M. and Wold, S., Quant. Struct.-Act. Relat., 10 (1991) 36.
25. Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Sjöström, M., Wold, S., Sandström, B. and Svensson, I., Environ. Toxicol. Chem., 10 (1991) 585.
26. Eriksson, L., Hellberg, S., Johansson, J., Jonsson, J., Sjöström, M., Wold, S. and Berglind, R., Acta Chem. Scand. 45 (1991) 935.
27. Eriksson, L., Jonsson J. and Berglind, R., Environ. Tox. Chem. 12 (1993) 1185.
28. Eriksson, L., Sjöström, M. and Wold, S., Chemom. Intel. Lab. Syst., 14 (1992) 245.
29. Eriksson, L., Berglind, R., Larsson, R. and Sjöström, M., J. Env. Sci. Health, A28 (1993) 1123.
30. Eriksson, L., Sandström, B.E., Tysklind, M. and Wold, S., Quant. Struct.-Act. Relat., 12 (1993) 124.
31. Eriksson, L., Verboom, H. and Peijnenburg, W., J. Chemom., 10 (1996) 483.
32. SIMCA-P, version 10, Umetrics AB, www.umetrics.com.
33. Wold, S., and Dunn, III, W.J., J.Chem. Inf. Comp. Sci., 23 (1983) 6.
34. Wold, S., Technometrics, 20, (1978) 397.
35. Wakeling, I.N., and Morris, J.J., J. Chemom., 7 (1993) 291.
36. Clark, M.C., and Cramer, R.D.,Quant. Struct.-Act. Relat., 12 (1993) 137.