A Strategy for Ranking Environmentally Occurring Chemicals. Part VI. QSARs for the Mutagenic Effects of Halogenated Aliphatics

Lennart Eriksson,[a] Sven Hellberg,[a] Erik Johansson,[b] Jörgen Jonsson,[a] Michael Sjöström,[a] Svante Wold,[a] Rune Berglind[c] and Britt Karlsson[c]

[a] Research Group for Chemometrics, Department of Organic Chemistry, University of Umeå, S-901 87 Umeå, Sweden, [b] Umetri AB, P.O. Box 1456, S-901 24 Umeå, Sweden and [c] National Defence Research Establishment, NBC-Defence Research, S-901 82 Umeå, Sweden

Eriksson, L., Hellberg, S., Johansson, E., Jonsson, J., Sjöström, M., Wold, S., Berglind, R. and Karlsson, B., 1991. A Strategy for Ranking Environmentally Occurring Chemicals. Part VI. QSARs for the Mutagenic Effects of Halogenated Aliphatics. – Acta Chem. Scand. 45: 935–944.

A strategy for the systematic analysis and priority ranking of environmental chemicals has been applied to a class of 58 halogenated aliphatic hydrocarbons. A training set of ten compounds representing this class, was selected by statistical design. The training set compounds were then subjected to biological testing in the *Salmonella typhimurium* reverse mutation assay (Ames test). The measured biological data, recorded as dose–response curves, were analyzed to determine the mutagenic potency (slope of the initial portion) and the mutagen dose (MD_{50}) required to increase the number of revertants above the background by 50%. For each compound, four mutagenic potency estimates and four MD_{50} values were determined, all originating from the tester strains TA 100 and TA 1535 with and without metabolic activation. The obtained responses were analyzed with multivariate techniques to give QSAR models relating the mutagenic potency data to the physico-chemical properties of the compounds. Finally, the derived QSARs were used to predict the mutagenic potencies and the MD_{50}s for the non-tested compounds in the class.

With the aim of developing a rational ranking for the toxicity testing of environmental pollutants, we have outlined a strategy based on statistical experimental design and multivariate modelling of the relation between chemical descriptor data and biological responses.[1–3] The strategy consists of six consecutive steps, of which the first is the division of chemicals into classes of structurally similar compounds. Once the first step has been conducted, the remaining steps are carried out on a class-by-class basis. Briefly, steps 2–6 are: (2) characterization of the chemical and structural variation within a class, (3) selection of a series of compounds – the training set – on which to base a quantitative structure–activity relationship (QSAR), (4) biological testing of the training set, (5) calculation of QSAR model, and, finally, (6) experimental validation of a developed QSAR on a set of validation compounds.

The proposed strategy was first applied to a class of 58 halogenated aliphatic hydrocarbons (see Table 1), resulting in the selection of a training set with 10 compounds.[4] Several biological endpoints were measured on these training set compounds, and the data were modelled by appropriate QSARs. So far, for this set of compounds, QSARs have been developed for the acute toxicity (LD_{50}) to rat,[4,5] the highest non-lethal dose to mouse,[4,5] and the genotoxic effect on DNA in Chinese hamster V79 cells.[6] To date, a fourth biological endpoint has been measured, namely the mutagenicity of the training set compounds as evidenced by the *Salmonella typhimurium* reverse mutation assay (Ames test).[7] The present article reports the QSAR-analyses of these mutagenicity data. Moreover, the question of how to evaluate test data from the Ames test in a meaningful, quantitative fashion, is discussed.

In the Ames assay, histidine auxotrophs (requiring histidine for growth) are exposed to a test substance on a Petri dish. The measurement made is the number of colonies on each plate, which reflects the number of bacteria having reverted to histidine prototrophy (independence). To produce a reliable dose–response curve for a chemical, it has been recommended that the test be performed over a concentration range of three orders of magnitude.[7] For most mutagens, the dose–response curve increases linearly at low doses, and then, as the dose increases, the curve may flatten and eventually turn downwards owing to effects of cytotoxicity.[7,8] Mutagenicity test results are evaluated as the number of revertants per microgram of test compound, taken from the linear portion of the dose–response curve.[7]

While much effort is continually made to improve and refine the experimental protocol of the Ames test, comparatively little attention has been focused on standardizing methods for representation and quantitative analysis of measured data.[8] In an early paper of Weinstein and Lewinson,[9] a statistical procedure was outlined, assuming the revertant colony formation at any dose to follow a Poisson process. Stead *et al.*[10] also adopted this concept and

Table 1. The 58 compounds belonging to the AX-class.

No.[a]	Compound	No.[a]	Compound
1	CH_3Cl	**30**	CH_3CH_2Br
2	CH_2Cl_2	**31**	$CHBr_3$
3	$CHCl_3$	**32**	$CH_3{-}CH_2F$
4	$CHCl_2F$	**33**	$CH_3{-}CHBr_2$
5	$CHClF_2$	**34**	CBr_2ClF
6	CCl_4	**35**	CH_2Br_2
7	CCl_3F	**36**	CH_3I
8	CCl_2F_2	**37**	CH_2BrCl
9	CH_3Br	**38**	$CBrF_3$
10	$CH_3{-}CH_2Cl$	**39**	CBr_3F
11	$CH_2Cl{-}CH_2Cl$	**40**	$CH_2Br{-}CH_2F$
12	$CH_2Br{-}CH_2Cl$	**41**	$CH_3{-}CHF_2$
13	$CH_2Cl{-}CHCl_2$	**42**	$CH_3{-}CH_2I$
14	$CH_3{-}CCl_3$	**43**	$CH_2Br{-}CH_2{-}CH_2Br$
15	$CHCl_2{-}CHCl_2$	**44**	$CH_2Br{-}CH_2{-}CH_2F$
16	$CHCl_2{-}CCl_3$	**45**	$CH_3{-}CH_2{-}CH_2Br$
17	$CCl_3{-}CCl_3$	**46**	$CH_3{-}CHBr{-}CH_3$
18	$CCl_2F{-}CClF_2$	**47**	$CH_3{-}CH_2{-}CH_2Cl$
19	$CH_2Br{-}CH_2Br$	**48**	$CH_3{-}CHCl{-}CH_3$
20	$CH_3{-}CHCl_2$	**49**	$CH_3{-}CH_2{-}CH_2I$
21	$CClF_2{-}CClF_2$	**50**	$CH_3{-}CHI{-}CH_3$
22	$CH_3{-}CHCl{-}CH_2Cl$	**51**	$CH_2Br{-}CH_2{-}CH_2{-}CH_2Br$
23	$CH_2Cl{-}CHCl{-}CH_2Cl$	**52**	$CH_3{-}CH_2{-}CH_2{-}CH_2Br$
24	$CH_3{-}CH_2{-}CH_2F$	**53**	$(CH_3)_3{-}CBr$
25	$CH_2F{-}CH_2{-}CH_2F$	**54**	$CH_3{-}CH_2{-}CH_2{-}CH_2Cl$
26	$CH_3{-}CF_2{-}CH_3$	**55**	$(CH_3)_3{-}CCl$
27	$CH_2Cl{-}CHCl{-}CHCl_2$	**56**	$CH_3{-}CH_2{-}CH_2{-}CH_2I$
28	$CH_2F{-}CF_2{-}CH_2Cl$	**57**	$(CH_3)_3{-}CI$
29	$CH_3{-}CF_2{-}CH_2Cl$	**58**	$CH_3{-}CH_2{-}CHI{-}CH_3$

[a] The numbers of the compounds are the same as in the previous parts.[4–6] The training set compounds are Nos. **2**, **3**, **7**, **11**, **15**, **30**, **33**, **39**, **48** and **52**.

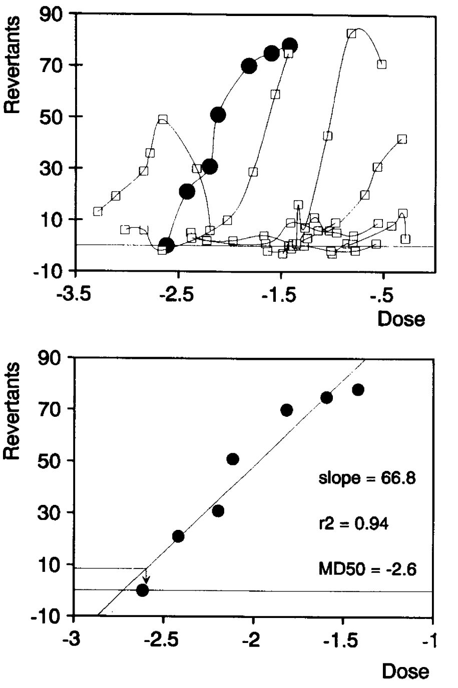

Fig. 1. Scatter plot for TA 1535 with metabolic activation, showing the distribution of dose–response curves of the nine compounds tested. The lower graph is an enlargement for 2-chloropropane (No. **48**), illustrating the evaluation procedure of a typical dose–response curve. The unit for the *x*-axis is log (μmol compound per gram agar).

fitted non-linear dose–response functions with up to four parameters. Using rather elaborate statistics, Margolin and coworkers[11] also fitted several parameters to adjust for non-linearity. A rather different and empirical approach was taken by Bernstein and coworkers.[8] Assuming a linear relationship between dose and response in the initial part of a dose–response curve, they based their statistical analysis solely on this region reasoning that it contains most of the interpretable information about mutagenicity. Moreover, they argued that the curvature of a dose–response curve beyond the linear region, depends on many underlying mechanisms that are not well understood and are likely to vary from chemical to chemical. Thus, Bernstein *et al.* suggested that the slope of the initial linear part should be used as a quantitative measure of the mutagenic potency of a compound. Rather recently this was also adhered to by McCann *et al.*[12]

According to the recommendations of Ames *et al.* and Bernstein *et al.*, we have focused our analysis of the mutagenicity data of the halogenated aliphatics to the initial region of the measured dose–response curves. However, knowing the slope of a dose–response curve is not the whole story. Some additional quantity is needed to locate it along the dose axis (cf. Fig. 1). To achieve this, linear regression was used to estimate the slope of the initial part of a dose–response curve, and the resulting model subsequently utilized to estimate a mutagen dose (MD_{50}) causing 50% increase in the number of revertants with reference to the background level. Thus, the slope is expected to give one estimate of the mutagenic potency and the MD_{50} another. However, it should be noted that the MD_{50} also provides, indirectly, information about the relative cytotoxicity among the compounds, as this is linked to the position of the dose–response curve along the dose axis.

Materials and methods

Biological data. The biological testing of the training set compounds (step 4 of the strategy) was carried out according to the revised procedure described by Ames *et al.*[7] However, owing to the volatility of the compounds, the testing had to be carried out in a special chamber allowing uniform conditions during the experimental period. Mutagenicity test results were obtained for nine out of the ten training set compounds. For one compound, fluorotrichloromethane (No. **7**) reliable test results could not be obtained owing to experimental difficulties (extreme volatility).[13] The compounds were tested with four standard tester strains, TA 98, TA 100, TA 1535 and TA 1537, at seven doses plus control, with five replicates at each dose. The actual range of doses used for each compound were determined in advance in screening tests checking for cytotoxicity. Also, the experiments were conducted with and without the standard S9-mix to investigate the degree of metabolic activation. We refrain from giving more experimental details, since they have been published separately.[13]

The evaluation of the mutagenicity data only concerns the strains TA 100 and TA 1535, since the compounds of interest only caused base-pair substitution (TA 100 and TA 1535) and no frameshift mutagenesis (TA 98 and TA 1537).[7] In theory this implicates four dose–response curves for each compound; two tester strains (TA 100 and TA 1535) times two treatments (with and without the S9-mix). Each dose–response curve was examined to find the initial linear portion and in most cases the identification of such a region was obvious. The results from the regression analy-

Table 2. Slope and MD_{50} estimates[a] for the training set compounds.

Comp. No.	Dose-range[b]	TA 100 with S9	TA 100 without S9	TA 1535 with S9	TA 1535 without S9
2	−1.0 to −0.3	slope = 103.6 $MD_{50} = -0.70$ $r^2 = 0.95$ $n = 5$	slope = 120.6 $MD_{50} = -0.69$ $r^2 = 0.97$ $n = 5$	slope = 21.4 $MD_{50} = -0.49$ $r^2 = 0.97$ $n = 5$	slope = 0 MD_{50} = 0[c] $r^2 = 0.04$ $n = 7$
3	−1.5 to −0.6	slope = 0 MD_{50} = 0[c] $r^2 = 0$ $n = 7$	slope = 0 MD_{50} = 0[c] $r^2 = 0$ $n = 7$	slope = 0 MD_{50} = 0[c] $r^2 = 0$ $n = 7$	slope = 0 MD_{50} = 0[c] $r^2 = 0$ $n = 7$
11	−1.6 to −0.3	slope = 27.4 $MD_{50} = -0.47$ $r^2 = 0.67$ $n = 7$	slope = 27.4 $MD_{50} = -0.74$ $r^2 = 0.76$ $n = 7$	slope = 33.7 $MD_{50} = -1.19$ $r^2 = 0.91$ $n = 7$	slope = 28.3 $MD_{50} = -1.13$ $r^2 = 0.91$ $n = 7$
15	−2.4 to −1.0	slope = 10.6 $MD_{50} = -1.03$ $r^2 = 0.83$ $n = 5$	slope = 17.2 $MD_{50} = -1.47$ $r^2 = 0.60$ $n = 4$	slope = 1.9[d] MD_{50} = 0[c] $r^2 = 0.14$ $n = 7$	slope = 1.8[d] MD_{50} = 0[c] $r^2 = 0.48$ $n = 7$
30	−1.4 to −0.5	slope = 113.2 $MD_{50} = -1.26$ $r^2 = 0.90$ $n = 7$	slope = 101.8 $MD_{50} = -1.23$ $r^2 = 0.97$ $n = 7$	slope = 97.3 $MD_{50} = -1.39$ $r^2 = 0.85$ $n = 7$	slope = 95.6 $MD_{50} = -1.31$ $r^2 = 0.94$ $n = 7$
33	−2.4 to −0.6	slope = 44.0 $MD_{50} = -0.93$ $r^2 = 0.87$ $n = 5$	slope = 26.4 $MD_{50} = -0.66$ $r^2 = 0.55$ $n = 5$	slope = 2.6[d] MD_{50} = 0[c] $r^2 = 0.27$ $n = 7$	slope = 5.3[d] MD_{50} = 0[c] $r^2 = 0.61$ $n = 7$
39	−3.3 to −2.2	slope = 25.5 $MD_{50} = -2.94$ $r^2 = 0.76$ $n = 5$	slope = 29.0 $MD_{50} = -2.73$ $r^2 = 0.92$ $n = 4$	slope = 52.9 $MD_{50} = -3.43$ $r^2 = 0.93$ $n = 5$	slope = 38.6 $MD_{50} = -3.64$ $r^2 = 0.90$ $n = 5$
48	−2.6 to −1.4	slope = 56.1 $MD_{50} = -2.00$ $r^2 = 0.93$ $n = 7$	slope = 55.9 $MD_{50} = -2.29$ $r^2 = 0.93$ $n = 7$	slope = 66.8 $MD_{50} = -2.60$ $r^2 = 0.94$ $n = 7$	slope = 55.3 $MD_{50} = -2.71$ $r^2 = 0.96$ $n = 7$
52	−3.0 to −1.4	slope = 87.0 $MD_{50} = -1.50$ $r^2 = 0.71$ $n = 4$	slope = 43.5 $MD_{50} = -1.14$ $r^2 = 0.74$ $n = 4$	slope = 60.8 $MD_{50} = -2.31$ $r^2 = 0.83$ $n = 5$	slope = 32.1 $MD_{50} = -1.80$ $r^2 = 0.66$ $n = 5$

[a] In the QSAR calculations, the slope and MD_{50} estimates are given the following variable numbers: (29) slope TA 100 with metabolic activation (MA), (30) slope TA 100 without MA, (31) slope TA 1535 with MA, (32) slope TA 1535 without MA, (33) MD_{50} TA 100 with MA, (34) MD_{50} TA 100 without MA, (35) MD_{50} TA 1535 with MA, and (36) MD_{50} TA 1535 without MA. For comparative purposes also the following information is included. The spontaneous revertant rates were 41 ± 12 (TA 100) and 11 ± 5 (TA 1535) revertants per plate. On average the positive control 2-aminoanthracene caused 229 ± 77 (TA 100 with MA), 100 ± 27 (TA 100 without MA), 46 ± 24 (TA 1535 with MA) and 11 ± 7 (TA 1535 without MA) revertants per plate. A second positive control, sodium azide, only tested without MA, caused 141 ± 60 (TA 100) respective 101 ± 36 (TA 1535) revertants per plate. [b] The lowest and highest dose for each compound, given as log (μmol compound per gram agar). [c] Approximated value, see the text for explanation. [d] Regression with near zero slope, which was not used to calculate any MD_{50} values. Instead, the MD_{50} was approximated as zero, see footnote *c*.

ses of the linear parts are listed together in Table 2, with corresponding correlation coefficients and numbers of doses used in the calculations. The resulting regression line was then used to estimate an MD_{50} value (cf. Fig. 1). It is common practice to calculate the dose giving rise to the double background of revertants (100 % increase), but we used 50 % to stay within the domain of the experimental results. In some cases the calculation of an MD_{50} estimate was not straightforward because the dose–response curve was flat, with zero, or close to zero, slope. However, it was possible to establish reasonable approximations to replace missing observations. Although such approximations may, at first sight, appear rough, they are far more informative than missing values in the QSAR analysis. For 1,1,2,2-tetrachloroethane (No. **15**) and 1,1-dibromoethane (No. **33**) the dose–response curves were flat in TA 1535 and therefore no direct MD_{50} estimates were obtainable within the current testing ranges. A similar phenomenon was observed for dichloromethane (No. **2**) in TA 1535 without metabolic activation. Parallel testing in TA 100, however, indicated these compounds to be active. Hence it seemed reasonable to anticipate that MD_{50} values might also be determinable in TA 1535, but at slightly higher doses than the ones already used. We utilized the TA 100 MD_{50} values to get an idea of the appropriate magnitude, which turned out to be a dose in the order of −0.5 (in log [μmol compound per gram of agar]). To avoid over-estimation of the missing values, the approximate TA 1535 MD_{50} values were set to dose 0 (same units as above), which is just outside the overall dose-range being spanned by the training set compounds (cf. Fig. 1). Lastly, we turned to the compound trichloromethane (No. **3**) for which all four dose–response curves were flat. This compound was tested in the dose-interval −1.2 to −0.4, rather close to the dose −0.5 previously identified. Hence, based on the same arguments as above, it was decided to assign the value of 0 to the missing MD_{50} values of trichloromethane.

In summary, the biological responses comprised eight variables (endpoints), namely four estimated slopes (Nos. **29–32**) and four MD_{50} values (Nos. **33–36**), see Table 2. As described above, the variables **33–36** in turn and order contained eight, eight, six and five calculated MD_{50} values, respectively. Consequently, they also contained one, one, three and four approximated values, respectively. To distinguish the slope variables from the MD_{50} variables, the former henceforth are called 'mutagenic potency' variables and the latter 'cytotoxicity' variables.

Chemical descriptor data. In this work, 14 chemical variables were used to describe the chemical and structural variation among the compounds in the AX-class training set. Their details have already been presented.[5] These 14 variables are: molecular weight (M_w, 1), boiling point (b.p., 2), melting point (m.p., 3), density (D, 4), refractive index (n_D, 5), van der Waals volume (V_{vdW}, 6), hydrophobicity (log P, 7), ionization potential (E_i, 8), log (retention times) from two gas chromatographic (GC1 and GC2, 9 and 10) and two liquid chromatographic (LC1 and LC2, 11 and 12) systems, the log (rate constant) for the Finkelstein iodide substitution reaction (k_f, 13) and the relative response to a flame ionization detector (R_{FID}, 14). Variables 9–14 were measured in our laboratory. To account for possible non-linearities in the response data from the Ames test, the quadratic terms of all variables were included in the QSAR analysis. Thus, the battery of chemical descriptors consisted of 28 (14 + 14) variables.

Principal components analysis (PCA). PCA[14] is a projection method that combines variables to a few, independent (orthogonal) underlying dimensions, with the purpose of obtaining an overview of the dominant patterns or major trends in the data table. Here, PCA is used to analyze the (9 × 8) biological response matrix. In PCA, a data matrix (say **X**) is decomposed into means ($\bar{x}_k$), scores (t_{ia}), loadings (p_{ak}) and residuals (e_{ik}) according to eqn. (1).

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia}p_{ak} + e_{ik} \qquad (1)$$

Here, the elements x_{ik} are the biological response data with index i denoting compounds and k the endpoints. The score t_{ia} describes the location of compound i for the ath principal component (PC), and the loading p_{ak} indicates how much and in what way (positive or negative correlation) a variable k contributes to this PC. The first PC describes the structure associated with the major variance in the data, the second PC the second largest variance, and so on. To determine the number of significant PCs [A in eqn. (1)], the cross-validation technique[15] is used. This achieves optimal predictive ability without overfitting of the model.

Partial least-squares projections to latent structures (PLS). The PLS method[16] is used to relate the biological response matrix (**Y**) to the systematic variation in the chemical descriptor data (matrix **X**), i.e. to establish QSARs for the AX-class. PLS is similar to PCA but calculates separate PLS-components for each matrix **X** and **Y**. Thus, the chemical descriptor variables are projected down on a low-dimensional subspace simultaneously with the projection of the biological activity variables onto the same subspace. In this way a model is obtained providing a good approximation of the **X**- and **Y**-matrices *and* high correlation between the two. As in PCA, the number of significant components is determined by cross-validation.[15]

Results

PCA of the biological response matrix. Before performing QSAR analysis of a multitude of biological response data, it is informative to check the underlying dimensionality of such variables by means of PCA.[17] The PCA of the 9 (compounds) by 8 (endpoints) response data matrix resulted in a two-component model (according to cross-

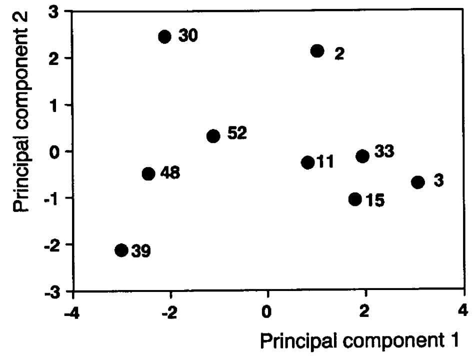

Fig. 2. Score plot from the PCA of the 9 × 8 biological response matrix with PC2 plotted versus PC1. For the numbering of the compounds, see Table 1.

validation) accounting for 79 % (49 + 30) of the total variance. The resulting score plot (Fig. 2) summarizes the biological data of the nine compounds. The comparatively weakly cytotoxic (high MD_{50} values) and non- or low-mutagenic (zero slope) compounds trichloromethane (No. **3**), 1,2-dichloroethane (No. **11**) and 1,1,2,2-tetrachloroethane (No. **15**) are all situated in the lower right-hand corner. Hence, this region is fairly 'safe' from an overall cytotoxic and mutagenic perspective. Upwards and especially to the left of the plot, one finds compounds that are highly cytotoxic or mutagenic. The most cytotoxic compound, fluorotribromomethane (No. **39**), is located in the lower left-hand corner, whereas the most mutagenic, bromoethane (No. **30**) is positioned in the upper left-hand corner. This leads to the conclusion that the PCs jointly describe the mutagenicity and relative cytotoxicity of the compounds. Thus, the mutagenic potency is changed as one moves diagonally from the lower right-hand corner (compounds **3**,**15** etc.) to the upper right corner (compound **30**), whereas an increase in the relative cytotoxicity is connected to the other diagonal going from the upper right (compounds **2** and **3**) to the lower left-hand (compound **39**) part.

Judging from the appearance of the score plot, the two phenomena – relative cytotoxicity and mutagenicity – are little correlated to each other. This is also corroborated by the corresponding loading plot (Fig. 3), displaying the behaviour of the eight response variables in the PC analysis. Three prominent groups of variables are clearly seen, two containing the mutagenic potency variables (Nos. **29–32**) and one the four cytotoxicity (MD_{50}) variables (Nos. **33–36**). The cytotoxicity variables dominates the first dimension, but the contribution from the TA 1535 slope variables (Nos. **31** and **32**) is not negligible. The mutagenic potency variables 29 and 30 (TA 100) are the most influential for the second PC. It is interesting to note the tight grouping of the four MD_{50} variables. This is a strong indication that they have a similar information content which justifies the approximations made to replace some missing MD_{50} values. Taken together the four endpoints stabilize each other and thereby facilitate the extraction of their intrinsic systematic variation. The overall separation of all the variables indicate that they carry different information about the biological effects of the compounds. Hence, based on Fig. 3, it was decided to split the original biological response matrix into three parts and perform separate QSAR-analyses on each part (see below).

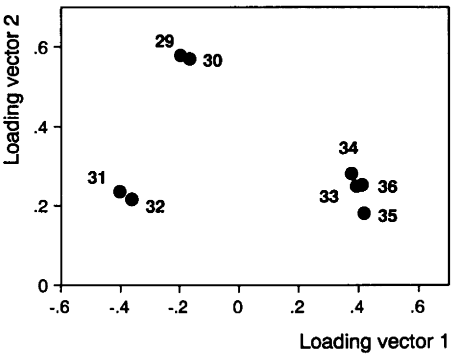

Fig. 3. Loading plot from the analysis of the biological response matrix. The variables are numbered as in Table 2.

QSAR analyses of the mutagenic potency variables (Nos. **29–32**). The PLS analysis of the TA 100 mutagenic potency variables (Nos. **29** and **30**) resulted in a two-dimensional model describing 79 % (61 + 18) of the variance in biological activity. As seen in Figs. 4(a)–(b), there is a fairly good agreement between observed and calculated mutagenic potencies for both endpoints; see also Table 3. Hence, the chemical descriptor variables are sufficient to model the variation in mutagenicity. It is evident from the corre-

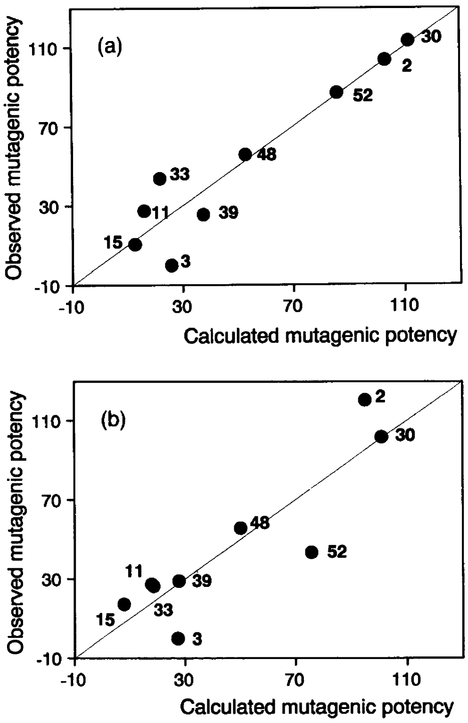

Fig. 4. Correlation plots with observed mutagenic potencies plotted against calculated values for TA 100 (a) with and (b) without metabolic activation. Notation as in Fig. 2.

Table 3. Calculated biological activities for the training set compounds.

Comp. No.	Variable No.							
	29	30	31	32	33	34	35	36
2	103.2	95.2	15.2	3.8	−0.48	−0.50	−0.68	−0.42
3	25.8	27.3	8.7	0.2	−0.28	−0.22	−0.15	0.09
7[a]	52	48	47	37	−1.6	−1.8	−1.9	−1.9
11	16.0	17.9	21.8	13.6	−0.58	−0.61	−0.51	−0.32
15	12.6	8.0	−3.1	−4.8	−1.00	−1.45	0.11	−0.06
30	111.5	101.1	97.0	81.9	−1.12	−1.24	−1.49	−1.35
33	21.7	18.5	18.0	12.1	−0.82	−0.72	−0.46	−0.36
39	37.6	27.8	49.4	44.0	−2.94	−2.67	−3.41	−3.53
48	52.8	50.1	67.7	55.4	−2.12	−2.39	−2.73	−2.76
52	86.0	75.9	62.8	50.8	−1.49	−1.16	−2.09	−1.89

[a] Predicted values.

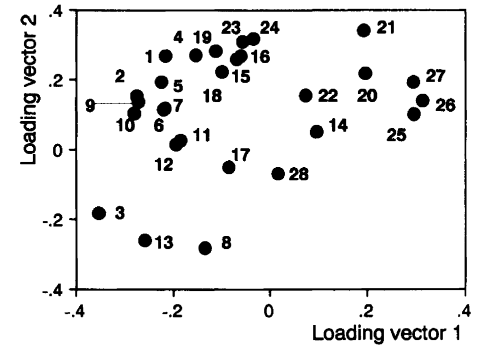

Fig. 5. Loading plot for the TA 100 mutagenicity QSAR. The chemical descriptors are: (1) M_w, (2) Bp, (3) Mp, (4) D, (5) n_D, (6) V_{vdV}, (7) log P, (8) E_i, (9) GC1, (10) GC2, (11) LC1, (12) LC2, (13) k_f, (14) R_{FID}, (15) M_w^2, (16) Bp^2, (17) Mp^2, (18) D^2, (19) n_D^2, (20) V_{vdV}^2, (21) [log P]^2, (22) E_i^2, (23) $GC1^2$, (24) $GC2^2$, (25) $LC1^2$, (26) $LC2^2$, (27) k_f^2, (28) R_{FID}^2.

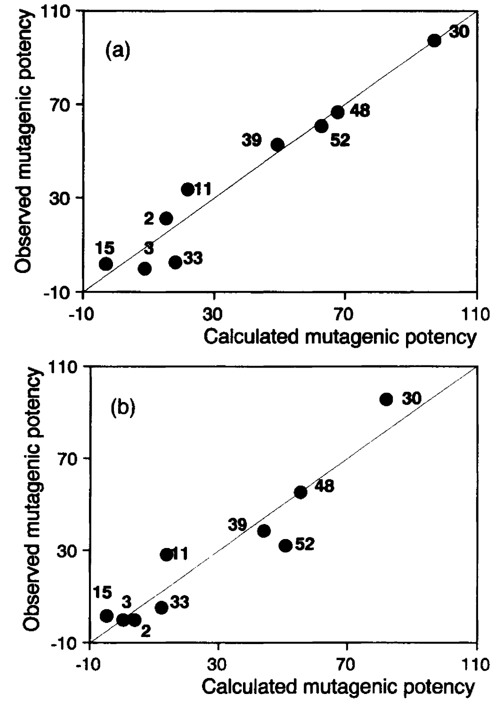

Fig. 6. Correlation plots with observed mutagenic potencies plotted against calculated values for TA 1535 (a) with and (b) without metabolic activation. Notation as in Fig. 2.

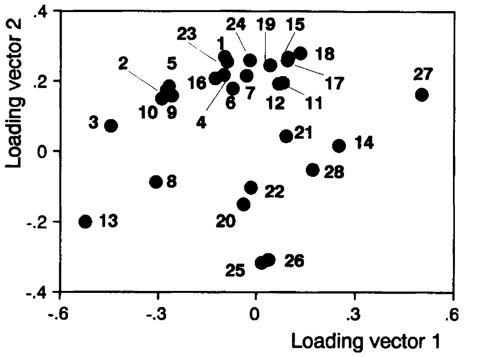

Fig. 7. Loading plot for the TA 1535 mutagenicity QSAR. Notation as in Fig. 5.

sponding loading plot (Fig. 5), that the first model dimension is dominated by hydrophobicity and bulk-describing variables, such as molecular weight (No. 1), boiling point (No. 2), melting point (No. 3), van der Waals volume (No. 6), log *P* (No. 7), and the two GC-variables (Nos. 9 and 10). Other important variables are the log k_f rate constant (No. 13), its quadratic term (No. 27), and the quadratic terms of the two LC-variables (Nos. 25 and 26). The ionization potential (No. 8), the boiling point, the density (No. 4) and the log k_f rate constant are among the most influential variables for the second dimension, together with the quadratic terms of log *P* (No. 21) and the two GC-variables (Nos. 23 and 24).

In the second PLS analysis, of the TA 1535 mutagenic potency variables (Nos. 31 and 32), a two-dimensional model was obtained which explained 89 % (68 + 21) of the variance in biological activity. This is slightly higher than in the previous model, suggesting the TA 1535 mutagenic potency variables to be better modelled by the chemical data than their TA 100 counterparts. Indeed this is the case, which may be seen in Figs. 6(a)–(b), where the observed mutagenicities are plotted versus the calculated ones (see also Table 3). The loading plot for this QSAR (Fig. 7) is quite different from that of the previous model. Many of the hydrophobicity and size-describing variables now appear in the second model dimension, and only partly contribute to the first. This applies to variables such as molecular weight (No. 1), boiling point (No. 2), density (No. 5), van de Waals volume (No. 6), log P (No. 7), and some of their quadratic terms. The first PLS component is strongly influenced by the melting point (No. 3), the ionization potential (No. 8), and the linear and quadratic terms of the log k_f rate constant (Nos. 13 and 27).

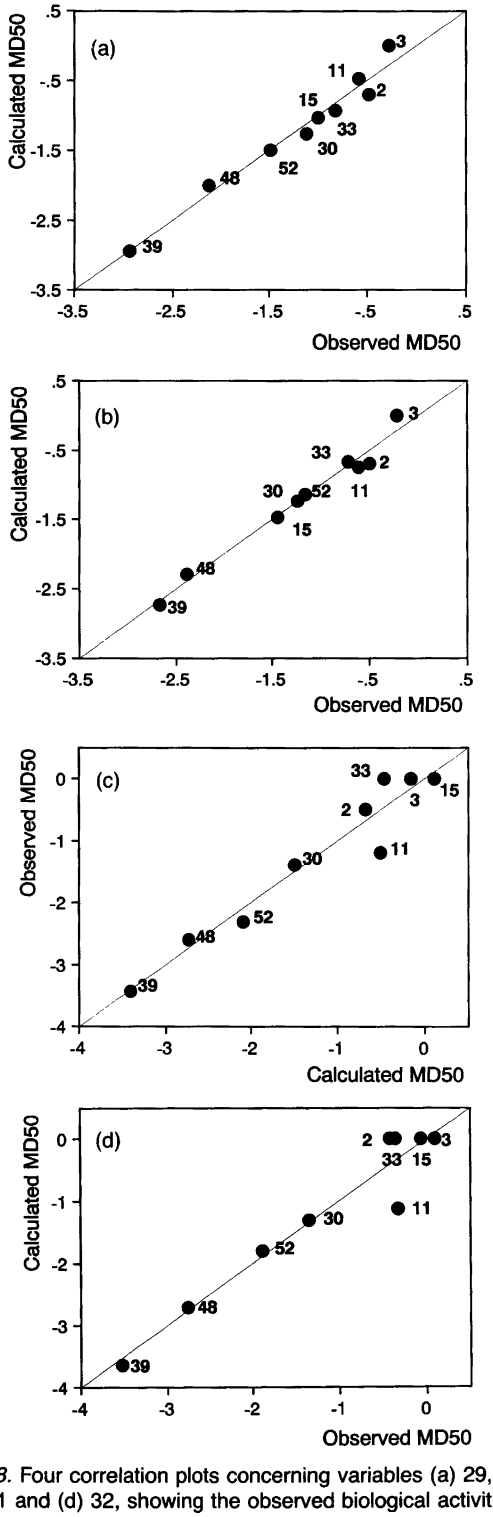

Fig. 8. Four correlation plots concerning variables (a) 29, (b) 30, (c) 31 and (d) 32, showing the observed biological activities versus the corresponding calculated values. Notation as in Fig. 2.

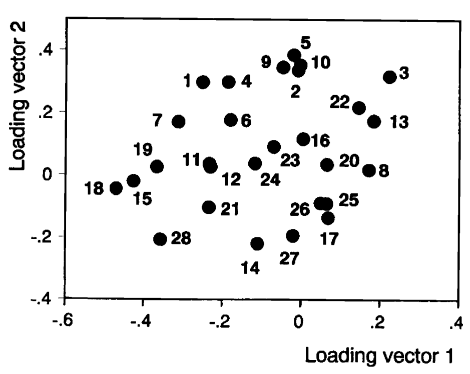

Fig. 9. Loading plot corresponding to Fig. 8. Notation as in Fig. 5.

QSAR analysis of the cytotoxicity variables (Nos. 33–36). The PLS analysis of the four MD_{50} variables gave a three-dimensional QSAR model, in contrast with the previous two-dimensional ones. This model explained 93 % of the variance in biological activity, and the contribution from each dimension was 64, 19 and 10 %, respectively. Figs. 8(a)–(d), shows the observed MD_{50} values plotted versus the corresponding values calculated by the QSAR model (see also Table 3). All four response variables are well explained by the QSAR model. The cytotoxicity QSAR is dominated by other variables than the two mutagenic potency QSARs (see Fig. 9). This is consistent with the finding that the different types of endpoint do not correlated well (Fig. 3). The first dimension is strongly dominated by log P (No. 7) and the quadratic terms $M_w{}^2$ (No. 15), D^2 (No. 18), $n_D{}^2$ (No. 19) and $R_{FID}{}^2$ (No. 28), indicating a non-linear relationship between the chemical and biological variables. In contrast, the second dimension is almost entirely influenced by the linear size and polarizability describing variables, i.e. molecular weight, boiling point, density, refractive index and the two GC-variables. Lastly, the third and minor dimension (not shown) which contains both linear and quadratic terms, is rather difficult to interpret.

Predictions for the remaining AX-compounds. The three derived QSARs for mutagenic potency and relative cytotoxicity were used to predict the biological activities of the 49 untested compounds belonging to the AX-class. This was accomplished by inserting their chemical descriptor data into the three models. All predicted values are listed in Table 4. Note that predictions have not been given for eleven compounds that deviate too strongly in chemical properties from the training set as indicated by their high residual standard deviation. The predictive abilities of the QSARs should be considered with some caution until they have been checked experimentally on a validation set of compounds. The predictions for the majority of the non-tested compounds are based on only 16 (8 + 8) chemical descriptor variables, because they have not yet been investigated in six chemical model systems used by us[5] (variables Nos. 9–14). The lack of descriptors is, however,

Table 4. Predicted biological activities for the AX-compounds not belonging to the training set.

Comp. No.[a]	Variable No.[b]							
	29	30	31	32	33	34	35	36
1	173	154	62	49	−3.1	−3.8	−3.8	−4.1
4	92	85	55	43	−1.4	−1.9	−1.4	−1.4
5	171	151	24	14	−0.3	−1.1	0.7	0.6
6	24	23	19	12	−0.6	−0.4	−0.4	−0.2
8	55	51	43	33	−2.9	−4.3	−2.3	−2.9
9	143	129	41	29	−1.0	−1.0	−1.4	−1.2
10	89	82	83	69	−2.6	−3.3	−3.2	−3.4
12	18	17	20	13	−0.5	−0.7	0.0	0.1
13	0	0	0	0	−0.4	−0.6	0.4	0.4
14	0	0	0	0	−0.6	−0.8	−0.2	−0.1
16	–	–	–	–	–	–	–	–
17	–	–	–	–	–	–	–	–
18	82	70	44	36	−1.6	−1.5	−1.7	−1.7
19	0	0	39	36	−1.5	−2.0	−0.6	−0.8
20	54	52	42	30	−1.1	−1.1	−1.5	−1.3
21	–	–	–	–	–	–	–	–
22	51	47	41	31	−0.8	−0.8	−0.9	−0.8
23	0	0	0	0	−1.1	−1.7	0.2	0.0
24	–	–	–	–	–	–	–	–
25	–	–	–	–	–	–	–	–
26	–	–	–	–	–	–	–	–
27	–	–	–	–	–	–	–	–
28	68	64	87	72	−1.7	−1.7	−2.5	−2.4
29	30	29	57	46	−1.8	−2.1	−2.2	−2.2
31	–	–	–	–	–	–	–	–
32	–	–	–	–	–	–	–	–
34	30	25	70	61	−1.5	−1.1	−1.8	−1.7
35	20	16	34	28	−1.8	−1.7	−1.8	−1.8
36	107	93	74	63	−0.7	−0.2	−0.8	−0.5
37	73	67	15	5	−0.3	−0.1	−0.5	−0.1
38	–	–	–	–	–	–	–	–
40	91	84	47	35	−0.9	−0.7	−1.4	−1.1
41	–	–	–	–	–	–	–	–
42	109	95	102	89	0.0	0.6	−0.1	0.3
43	0	0	0	0	−1.5	−2.1	0.0	−0.4
44	55	51	39	29	−0.6	−0.5	−0.7	−0.5
45	68	63	68	55	−0.8	−0.7	−1.1	−0.8
46	65	60	60	48	−0.8	−0.6	−1.1	−0.8
47	59	56	78	64	−1.8	−2.0	−2.4	−2.4
49	120	103	101	89	0.1	0.8	0.1	0.6
50	127	110	103	90	0.4	1.2	0.4	0.9
51	67	50	0	0	−1.4	−2.4	1.0	0.5
53	41	37	33	25	−0.4	−0.5	0.1	0.2
54	68	62	77	64	−1.9	−2.0	−2.4	−2.4
55	15	16	36	27	−1.6	−1.9	−1.6	−1.7
56	207	176	98	87	0.0	0.8	0.1	0.5
57	173	146	79	69	0.7	1.5	1.1	1.6
58	198	169	97	85	−0.2	0.6	−0.3	0.2

[a]Numbers as in Table 1. [b]Numbers and units as in Table 2. Predicted values are not given for eleven compounds, see the text for explanation.

not a computational problem since PLS can handle missing data, but may make the predictions less reliable. The 28 chemical descriptor variables were used for the training set compounds to improve the stability of the three QSARs, which in turn also stabilizes the predictions for the non-tested compounds.

To facilitate the interpretation of all the predicted values, they were inserted into the existing PC-model for the response data (see Fig. 10). It is seen that there are compounds predicted to be highly cytotoxic (lower left-hand corner) and highly mutagenic (upper right-hand corner), but none to be a combination of both (which would be equal to a position in the upper left-hand corner). In general, chemicals predicted to be highly mutagenic are those which are mono-iodinated, like for instance iodomethane (No. **36**), iodoethane (No. **42**), 1-iodopropane (No. **49**), 2-iodopropane (No. **50**) and the iodobutanes (Nos. **56–58**). Moreover, several mono-chlorinated com-

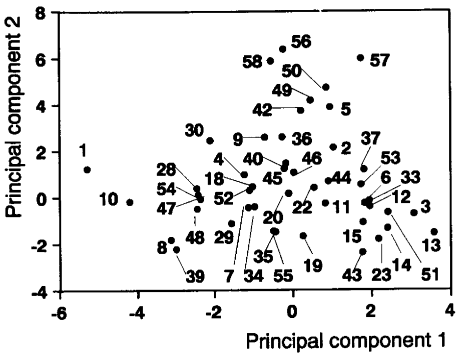

Fig. 10. Scatter plot visualizing the projection of the predictions in Table 4 onto the two-dimensional PC model of the biological response matrix. For the numbering of the compounds, see Table 1.

pounds, such as chloromethane (No. **1**), chloroethane (No. **10**), 1-chloropropane (No. **47**) and 1-chlorobutane (No. **54**), are classified as cytotoxic, together with some fluorinated compounds, such as fluorotrichloromethane (No. 7, training set member), dichlorodifluoromethane (No. **8**) and 1-chloro-2,2-difluoropropane (No. **29**). The monobrominated compounds, like bromomethane (No. **9**), 1-bromopropane (No. **45**) and 2-bromopropane (No. **46**), are predicted to be of intermediate cytotoxicity and mutagenicity. There are also many chemicals that are located in the 'safe' lower right-hand corner and hence classified as comparatively harmless. Some examples are the polychlorinated and polybrominated compounds 1,1,2-trichloroethane (No. **13**), 1,1,1-trichloroethane (No. **14**), 1-bromo-2-chloroethane (No. **12**), 1,2,3-trichloropropane (No. **23**), 1,3-dibromopropane (No. **43**) and 1,4-dibromobutane (No. **51**). Apparently, the models predict multiple halogenation with chlorine and/or bromine to give relatively non-cytotoxic and non-mutagenic compounds in the Ames test.

Discussion

QSARs are useful tools for predicting toxic effects of chemicals and for the identification of potentially hazardous ones. However, there are some aspects that deserve special attention to avoid ending up with QSARs of low predictive capability. First, according to their theoretical foundation,[18] QSAR models are only locally valid. Thus a given QSAR works only for chemically and biologically similar compounds. A suitable division of chemicals into classes of structurally similar compounds may solve this problem. Second, the selection of the training set of compounds by means of statistical design,[19] ensures that the structural domain of a class is spanned in a balanced manner. A major limitation of many QSAR models is the lack of consistently measured biological data for a well-defined series (training set) of compounds. Finally, the compiled data, both chemical and biological, should be analyzed by means of a multivariate data analytical method, such as PLS, which provides information about the structure of the data and the range of the class. All of the above given criteria have been taken into account or directly incorporated in the recently proposed strategy for QSAR development concerning environmental chemicals.[1–3]

Application of the strategy to the class of saturated halogenated aliphatics (the AX-class) lead to the identification of ten training-set compounds distributed in a balanced manner.[4] Subsequently, the compounds in the training set were subjected to biological testing. In the present work, experimental data from the Ames test are presented and evaluated. A number of approaches have been proposed for examining such dose–response data. Some methods are based on regression analysis of an initial (usually linear) part, but others on more elaborate statistical treatment of a whole dose–response curve. As a quantitative measure of the mutagenic potency of a compound, we used the slope of the initial linear part. Further, to describe the position of a dose–response curve on the dose axis, we calculated the mutagen dose (MD_{50}) inducing 50% increase in the number of revertants.

The eight obtained biological responses (four mutagenic and four cytotoxic) were initially analyzed by means of principal component analysis in order to explore the latent, underlying dimensionality of the response matrix. It was found that the response variables formed three prominent groups (cf. Fig. 3), two with the mutagenic potency variables and one with the cytotoxicity variables. Hence, the variables were split up and treated in three groups in the QSAR modelling. In the next few paragraphs, it was demonstrated that the three groups of endpoints could be modelled by means of the construction of three multiresponse QSARs. The results were rather satisfying agreements between observed biological activities and values calculated by the QSARs (Figs. 4, 6 and 8). Furthermore, it was evident that a non-linear relationship existed between chemical descriptor data and biological responses. The derived QSARs were two or three-dimensional and significant contributions occurred both from the linear and non-linear (quuadratic) variables. Primarily, however, the linear size and hydrophobicity-describing variables were the most important.

In the next step, the QSARs were utilized to predict the mutagenicity and cytotoxicity of 38 non-tested compounds. All predicted values are listed in Table 4. These values were then projected onto the already existing two-dimensional PC-model for the response matrix. With this plot (see Fig. 10) it is fairly easy to get an overview of all the predictions. Simply by looking at the location of a compound in the score plot, one can tell whether it is likely to be mutagenic, cytotoxic or both. The score plot summarizes the properties of a substantial number of dose–response curves, both those actually observed for the training set compounds and those predicted for the non-tested compounds. The features it summarizes are the shape (slope, mutagenicity) and the location (cytotoxicity).

The derived QSAR models have not yet been validated experimentally. However, it is possible to compare our measurements and predictions with data found in the literature. Of the training-set compounds, trichloromethane (No. **3**) has been found to be non-mutagenic in a number of cases,[20-22] which agrees with our results. In contrast, results regarding dichloromethane are more varied, and it is claimed to be both mutagenic[23] and non-mutagenic.[24] We found dichloromethane (No. **2**) to be mutagenic in TA 100, but practically non-mutagenic in TA 1535. Furthermore, in an extensive compilation work by Ashby and Tennant[25] listing the mutagenicity in the Ames assay of some 220 chemicals, we were able to find data for four of the compounds considered. The training-set compound 1,1,2,2-tetrachloroethane (No. **15**) is classified as non-mutagenic which compares well with our experimental findings; inactivity in TA 1535 and very weak mutagenicity in TA 100. The compound 1,2-dibromoethane (No. **19**), is claimed to be mutagenic, which the mutagenic potency QSARs also predict. 1,1,2-trichloroethane (No. **13**), categorized as non-mutagenic, is also correctly predicted. The fourth compound, 1-chlorobutane (No. **54**), is not mutagenic but is predicted to be active. However, we note that Ashby and Tennant also failed to predict this compound. Our experimental set-up identified monohalogenated compounds to be mutagenic, like, for instance, 1-bromobutane (No. **52**). The mutagenic potency QSARs were trained in that way and consequently predict monohalogenation to induce mutagenicity. In yet another work, by Travis *et al.*,[26] we were also able to find observed data for four compounds. These literature data are weighted results from a battery of mutagenicity tests, including the Ames assay. As the nature of these data and our data differ slightly, some discrepancy may be expected. According to Travis and coworkers, two compounds, namely tetrachloromethane (No. **6**) and 1,2-dibromoethane (No. **19**), have low mutagenic activity, which implies that our predictions are in accordance with the measurements. Trichloromethane (No. **3**) was measured by us to be non-mutagenic, but is listed as to have low mutagenicity. Also 1,1,2-trichloroethane (No. **13**) is categorized as being slightly mutagenic, but we predict it to be non-mutagenic.

The comparison with literature data indicate that our QSARs have satisfactory predictive capabilities. However, the final judgement awaits experimental validation. The testing of a validation set of six compounds is being planned and will be started as soon as possible. Also note that the results given only provide information about one type of mutagenicity and cytotoxicity test, the Ames test. However, when combined with other test data related to other aspects, the predictions may be helpful in setting priorities for further biological testing.

Acknowledgements. Financial support from the Centre for Environmental Research (CMF), Umeå, the National Chemicals Inspectorate and the Swedish Natural Science Research Council (NFR), is gratefully acknowledged.

References

1. Jonsson, J., Eriksson, L., Sjöström, M., Wold, S. and Tosato, M. L. *Chemom. Intell. Lab. Syst. 5* (1989) 169.
2. Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. *Chemom. Intell. Lab. Syst. 7* (1989) 131.
3. Tosato, M. L., Vigano, L., Skagerberg, B. and Clementi, S. *Environ. Sci. Technol.* (1990). *In press.*
4. Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjöström, M., Wold, S. and Berglind, R. *Environ. Toxicol. Chem. 9* (1990) 1339.
5. Lindgren, F., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M. and Wold, S. *Quant. Struct.-Act. Relat. 10* (1991) 36.
6. Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Sjöström, M., Wold, S., Sandström, B. and Svensson, I. *Environ. Toxicol. Chem. 10* (1991) 585.
7. Maron, D. M. and Ames, B. N. *Mutat. Res. 113* (1983) 173.
8. Bernstein, L., Kaldor, J., McCann, J. and Pike, M. C. *Mutat. Res. 97* (1982) 267.
9. Weinstein, D. and Lewinson, T. M. *Mutat. Res. 51* (1978) 433.
10. Stead, A. G., Hasselblad, V., Creason, J. P. and Claxton, L. *Mutat. Res. 85* (1981) 13.
11. Margolin, B. H., Kaplan, N. and Zeiger, E. *Proc. Natl. Acad. Sci. 78* (1981) 3779.
12. McCann, J., Swirsky-Gold, L., Horn, L., McGill, R., Graedel, T. E. and Kaldor, J. *Mutat. Res. 205* (1988) 183.
13. Karlsson, B. *National Defence Research Establishment Report 803:3,* NBC-Defence Research, Umeå, Sweden 1990.
14. Jolliffe, I. T. *Principal Component Analysis,* Springer-Verlag, New York 1986.
15. Wold, S. *Technometrics 20* (1978) 379.
16. Dunn, W. J., III, Wold, S., Edlund, U., Hellberg, S. and Gasteiger, J. *Quant. Struct.-Act. Relat. 3* (1984) 131.
17. Wold, S. and Dunn, W. J., III. *J. Chem. Inf. Comput. Sci 23* (1983) 6.
18. Wold, S. and Sjöström, M. In: Chapman, N. B. and Shorter, J., Eds., *Correlation Analysis in Chemistry,* Plenum, 1978, pp. 2-54.
19. 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.
20. Trueman, R. W. *Prog. Mutat. Res. 1* (1981) 343.
21. Venitt, S. and Crofton-Sleigh, C. *Prog. Mutat. Res. 1* (1981) 351.
22. van Abbé, N. J., Green, T. J., Jones, E., Richold, M. and Roe, F. J. C. *Chem. Toxicol. 20* (1982) 557.
23. Ma, T. H., Harris, M. M., Anderson, V. A., Ahmed, I., Mohammad, K., Bare, J. L. and Lin, G. *Mutat. Res. 138* (1984) 157.
24. Buijs, W., van der Gen, A., Mohn, G. R. and Breimer, D. D. *Mutat. Res. 141* (1984) 11.
25. Ashby, J. and Tennant, R. W. *Mutat. Res. 204* (1988) 17.
26. Travis, C. C., Saulsbury, A. W. and Richter Pack, S. A. *Mutagenesis 5* (1990) 213.

Received December 21, 1990.