Toxicological Evaluation of Complex Mixtures by Pattern Recognition: Correlating Chemical Fingerprints to Mutagenicity

Ingvar Eide,[1] Gunhild Neverdal,[1] Bodil Thorvaldsen,[1] Bjørn Grung,[2] and Olav M. Kvalheim[2]

[1]Statoil Research Centre, Trondheim, Norway; [2]Department of Chemistry, University of Bergen, Bergen, Norway

We describe the use of pattern recognition and multivariate regression in the assessment of complex mixtures by correlating chemical fingerprints to the mutagenicity of the mixtures. Mixtures were 20 organic extracts of exhaust particles, each containing 102–170 individual compounds such as polycyclic aromatic hydrocarbons (PAHs), nitro-PAHs, oxy-PAHs, and saturated hydrocarbons. Mixtures were characterized by full-scan GC–MS (gas chromatography–mass spectrometry). Data were resolved into peaks and spectra for individual compounds by an automated curve resolution procedure. Resolved chromatograms were integrated, resulting in a predictor matrix that was used as input to a principal component analysis to evaluate similarities between mixtures (i.e., classification). Furthermore, partial least-squares projections to latent structures were used to correlate the GC–MS data to mutagenicity, as measured in the Ames *Salmonella* assay (i.e., calibration). The best model (high r^2 and Q^2) identifies the variables that co-vary with the observed mutagenicity. These variables may subsequently be identified in more detail. Furthermore, the regression model can be used to predict mutagenicity from GC–MS chromatograms of other organic extracts. We emphasize that both chemical fingerprints as well as detailed data on composition can be used in pattern recognition. *Key words:* automated curve resolution, chemical fingerprint, complex mixtures, GC–MS, mutagenicity, PAH, pattern recognition, PLS. *Environ Health Perspect* 110(suppl 6):985–988 (2002).
http://ehpnet1.niehs.nih.gov/docs/2002/suppl-6/985-988eide/abstract.html

Toxicological evaluation of complex mixtures is really challenging. Simple mixtures may be handled easily with well-defined variables that may be combined differently to obtain the effect of each variable and possible interactions between them. The composition of the new mixtures may be determined by means of statistical experimental design, which is being used increasingly in mixture research. However, for practical reasons, this approach is only possible with a limited number of variables. With complex mixtures, the common strategy is to study the mixture as a whole or to fractionate the mixture (bioassay-directed fractionation). Mixture fractionation and recombination may be useful for identification of major bioassay-active fractions and possible interactions (*1*). Alternatives are spiking complex mixtures with individual compounds (*2,3*), lumping (*4*), and the "top 10" or "pseudo top 10" approaches (*5,6*). These approaches make it possible to treat the complex mixtures as simple mixtures. However, none of these approaches give information about all the individual compounds in the complex mixture. In fact, with complex mixtures, even a detailed and complete characterization may be impossible.

In a recent article we presented a new approach to correlate chemical fingerprints of very complex mixtures to mutagenicity (*7*). The mixtures were organic extracts of exhaust particles obtained from a variety of sources and were assumed to have different but partly overlapping compositions. Extracts of exhaust particles contain a variety of different polycyclic aromatic hydrocarbons (PAHs), nitro-PAHs, and oxy-PAHs (*8*). Many of these are mutagenic and carcinogenic. In addition, the extracts may contain saturated hydrocarbons.

In this article we show an example of a more complete strategy for the toxicological evaluation of very complex mixtures, demonstrated with one biological end point. Organic extracts of exhaust particles are characterized by full-scan GC–MS (gas chromatography–mass spectrometry). The GC–MS chromatograms are complex, as illustrated in the previous article (*7*), with significant overlap between peaks. Frequent scanning gives changes in spectra depending on whether the spectra are obtained from one, two, or more compounds (Figure 1). Based on this information, the complex GC–MS data are resolved into peaks and spectra for individual compounds using an automated curve resolution procedure (*7,9*). The resolved chromatograms are integrated, resulting in a predictor matrix. Principal component analysis (PCA) is used to evaluate similarities between mixtures (classification). The data matrix is also used as input to a multivariate regression model, which correlates the GC–MS data to the mutagenicity measured in the Ames *Salmonella* assay (*10*). Only the TA98 strain without the addition of liver enzymes was used in the present example. Partial least-squares (PLS) projections to latent structures (*11*) is used for the regression modeling, as it overcomes the problems of intercorrelated predictor variables and data matrices where the number of variables exceeds the number of samples (*12,13*). The regression model identifies those peaks that co-vary with the observed mutagenicity. These peaks may subsequently be identified chemically from their spectra. Furthermore, the regression model can be used to predict mutagenicity from GC–MS chromatograms of other organic extracts. This is an attractive possibility, as bioassays are generally more resource demanding and require larger samples than chemical characterization.

Materials and Methods

Organic extracts of exhaust particles. Twenty different organic extracts of exhaust particles were selected and assumed to have different but partially overlapping composition. Samples were obtained from the combustion of heating oils and gas in boilers. Dichloromethane (DCM; Merck, Darmstadt, Germany; >99.8%) was used as the solvent (*7*).

Ames* Salmonella *assay. Prior to mutagenicity testing, a volume of each of the DCM extracts was evaporated to dryness under dry nitrogen and completely dissolved in dimethylsulfoxide (Merck, Darmstadt, Germany; >99.8%). The standard plate incorporation assay described by Maron and Ames (*10*) was used for mutagenicity testing. A volume of 100 µL test solution was added to each plate. The *Salmonella typhimurium* strain TA98 was obtained from B. Ames (University of California, Berkeley, California, USA). The mutagenicity testing was performed without the addition of metabolizing system. Mutagenicity was expressed as revertants per milligram of particulate matter (PM). The values are based on the slopes of the regression lines of the dose–response curves from two independent assays, each at five doses with three parallels at each dose [details in Eide et al. (*7*)].

Gas chromatography–mass spectrometry. A volume of 0.5–1 mL of each DCM extract was spiked with 3.24 µg d_8-naphthalene

This article is part of the monograph *Application of Technology to Chemical Mixture Research.*

Address correspondence to I. Eide, Statoil Research Centre, N-7005 Trondheim, Norway. Telephone: 47 73584595. Fax: 47 73967286. E-mail: ieide@statoil.com

The authors are grateful to R. Arneberg, Pattern Recognition Systems, Bergen, Norway, for valuable support.

Received 18 December 2001; accepted 31 May 2002.

(Cambridge Isotope Laboratories, Woburn, MA; 99%). The volume of each extract was then reduced under a gentle stream of nitrogen. The samples were analyzed by GC–MS. A Fisons GC8000 (Fisons Instruments, Manchester, UK) equipped with a 100-m Petrocol DH fused silica capillary column (0.25 mm i.d., 0.5 µm film thickness; Supelco, Bellefonte, PA, USA) was used for sample introduction into the mass spectrometer. The GC program began at an initial temperature of 40°C, ramped to a final temperature of 320°C at 4°C/min, and held for 20 min. A Fisons MD800 quadrupole mass spectrometer (Fisons Instruments) operated in the EI mode (70 eV) was used to obtain mass spectra. The instrument was operated at 1.3 scans/sec from 40 to 450 m/z (i.e., full-scan mode) to obtain structural information from all important fragments [details in Eide et al. (7)].

Data matrix and curve resolution. Signals from compounds eluting before the internal standard (d_8-naphthalene) were not used, as these were assumed to be nonmutagenic. The remaining matrices were split into smaller ones, each one containing only one cluster of coeluting peaks. Mass numbers containing only background were deleted using a shape criterion for the masses. Finally, mass numbers with intensity <1% of the maximum intensity of the peak cluster were deleted. The whole procedure was automated. The curve resolution was performed with the recently developed MS Resolver from Pattern Recognition Systems (Bergen, Norway) (*9*). It is based on a modified version of the Gentle iteration method (*14,15*). To match the resolved spectra to ascertain that the same compound is represented by the same variable number in all samples, similarity between spectra was evaluated for peaks that appeared within a time interval of 4 min (peak position of the internal standard varied less than 2 min between different samples). For each resolved spectrum, only the 10 most significant intensities are used for the similarity matching. This ensures that small noise-dominated mass numbers have no influence on the matching procedure. A similarity index of 0.8 was used for the similarity matching (*7,9*). After this initial calculation of all possible similarities, median spectra were constructed from the spectra found to represent the same component. Next, a second matching procedure was performed between all unmatched spectra and the median spectra. Finally, a similarity check was performed among the median spectra. Components left unmatched after this procedure appear in only one sample. These were discarded from further analysis. The integrated areas of the remaining resolved chromatograms were calculated, resulting in a predictor matrix of size 20 × 472. In the predictor matrix, each row represents one sample and each column one compound, the latter identified by its mean retention time. A schematic illustration of the curve resolution and the data matrix is shown in our previous article (*7*). To create the final predictor matrix **X**, each resolved peak area was multiplied by the added amount of internal standard and divided by the area of the internal standard and the amount of PM used in the sample. Each value in the **X** matrix represents micrograms compound per milligram particles. The response vector **y** contains 20 values representing mutagenicity (revertants per milligram particles). The internal standard is used only to adjust the concentration of the samples in relation to each other and in relation to the mutagenicity data [details in Eide et al. (*7*)]. One internal standard has been considered sufficient for this purpose. Deuterated naphthalene was chosen because it was relatively easy to resolve from the other peaks in the chromatogram.

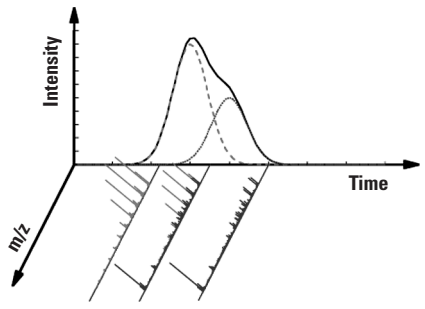

Figure 1. Illustration of curve resolution of GC–MS data based on changes in spectra.

Pattern recognition and regression. Multivariate data analysis and modeling were performed with Simca-P 8.0 for Windows (Umetrics, Umeå, Sweden). PCA (*16*) was performed on the **X** matrix for outlier detection by means of loading plot, and for the evaluation of similarities between mixtures by means of score plot (classification). Multivariate regression was performed with PLS (*11*). PLS finds the relationship between the response vector **y** (or matrix **Y**) and the matrix **X** (predictor variables) by simultaneous projections of both the *X* and *Y* spaces to a plane or hyperplane (*12,13*). The data were centered and scaled to unit variance before the PLS analysis. At first a model with all variables was calculated. The variables with the lowest variable importance were eliminated. This procedure was repeated until the best model was obtained, both with respect to correlation coefficients (shown as r^2 *Y*) and prediction properties (shown as Q^2). The latter are obtained after cross-validation (*17*) and are important to avoid overfit. Intercorrelations between *x*-variables were evaluated by r^2 *X*. To ascertain that the correlations were not simply due to chance, the model was validated by performing PLS with all 472 *x*-variables after randomizing the values in the **y** vector (10 permutations), as described in our previous article (*7*).

Table 1. Number of compounds and mutagenicity of each sample.

Sample no.	No. of compounds	Mutagenicity (revertants per mg PM)
1	127	182
2	157	266
3	129	402
4	141	501
5	102	349
6	134	504
7	150	504
8	143	184
9	127	331
10	151	286
11	137	287
12	143	371
13	139	94
14	143	83
15	128	123
16	158	151
17	155	98
18	166	114
19	170	143
20	160	91

Results

The total number of different compounds in the 20 samples is 472 when applying a similarity index of 0.8. The **X** matrix contains one row for each of the 20 samples, and one column for each compound that is identified by its retention time only. On average, 143 peaks (ranging from 102 to 170) were resolved from each sample, as shown in Table 1. This implies that 70% of the values in the **X** matrix are zero. A value of zero in the matrix means that the compound is not present in the corresponding sample. Table 1 shows the mutagenicity of each sample ranging from 83 to 504 revertants per milligram PM.

Figure 2 shows the score plot (t[1] vs. t[2]), and illustrates that 10 of the samples (located close to the center of the plot) are

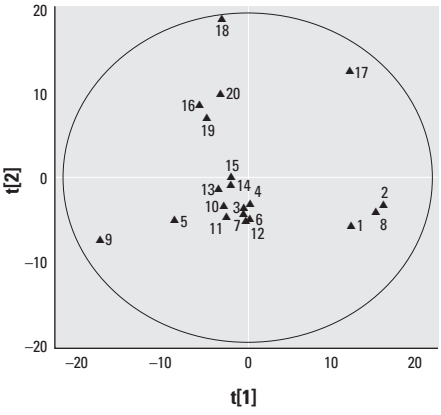

Figure 2. Score plot from PCA illustrating similarities and dissimilarities among samples.

similar in composition. Samples 1, 2, and 8 are similar to each other and different from the others. Furthermore, samples 16, 19, and 20 constitute a third class with mutual similarities. On the other hand, samples 5, 9, 17, and 18 are unique in composition. Ideally, for the purpose of multivariate regression, the samples should be evenly distributed in the score plot, approaching the situation obtained with statistically designed experiments.

Table 2 shows the overall results of the PLS analyses, starting with all 472 *x*-variables. By removing the variables with the lowest importance, the models are generally improved. The best model is obtained with 41 variables, resulting in very high correlation coefficient and prediction properties. Relatively good models can also be obtained with as few as nine variables. Simultaneously with improving the regression models, r^2X is increased to 0.62 for the best model, indicating some intercorrelations between the remaining predictor variables. These intercorrelations can be broken only by adding more samples with different compositions to the calibration model. The best regression model was obtained with four PLS components (latent variables). Adding more PLS components only improves the model insignificantly with regard to fit and may deteriorate the prediction properties of the model (overfit).

Figure 3 shows the observed versus predicted response values for the 20 samples obtained by the model with 41 variables. The regression model can be used to predict mutagenicity from GC–MS chromatograms of other samples, provided the samples are within the model domain. We emphasize, however, that the regression model should be improved by expanding the matrix with more samples that are different from the ones used in the present work.

Figure 3 reflects differences in mutagenicity between the samples. Although samples 7 and 14, for example, are very different in mutagenicity, they have very similar composition, as illustrated by the score plot in Figure 2. However, the difference may be seeen from other principal components (only the first two shown in Figure 2).

Table 2. Correlation (r^2) and prediction (Q^2) properties of 11 different models with different number of variables.

No. of variables	r^2X	r^2Y	Q^2
472	0.11	0.72	0.36
114	0.24	0.78	0.61
88	0.32	0.96	0.70
55	0.42	0.93	0.76
46	0.57	0.99	0.81
41	0.62	0.99	0.87
29	0.55	0.92	0.82
22	0.60	0.93	0.86
14	0.68	0.90	0.82
12	0.68	0.89	0.84
9	0.71	0.86	0.80

Discussion

This work is the second attempt to combine curve resolution of complex GC–MS data and multivariate calibration with a toxicological parameter as the response. Compared with the previous study (*7*), the present data matrix was based on a larger number of samples that were slightly less complex than previous samples. The samples were relatively well distributed with respect to composition and mutagenicity. Although the number of samples was relatively low, and a total of 472 different variables were identified, the regression model with the most important 41 variables was very good, as evaluated by correlation coefficient, prediction properties obtained after cross validation, and a low number of required PLS components (latent variables). The good model is a result of samples with different composition, different number of variables, and different mutagenicity. The model may therefore be used to predict mutagenicity of other extracts of soot particles from their GC–MS chromatograms. However, before prediction, a PCA should be performed, and the score plot should be used to verify that the new samples are within the calibration domain described by the 41 variables. Otherwise, another regression model, obtained with a higher number of variables from the calibration set, should be considered.

The multivariate data analysis gives an empirical model that identifies the peaks that co-vary with mutagenicity. An evaluation of the regression model with 41 variables shows that some of the *x*-variables co-vary (shown by r^2X), which is expected as the total number of variables exceeds the number of samples. This implies that nonmutagenic compounds may correlate with mutagenic ones, and as a consequence, also with mutagenicity. To break these intercorrelations between *x*-variables and improve the model, more samples with different compositions are required for the calibration model.

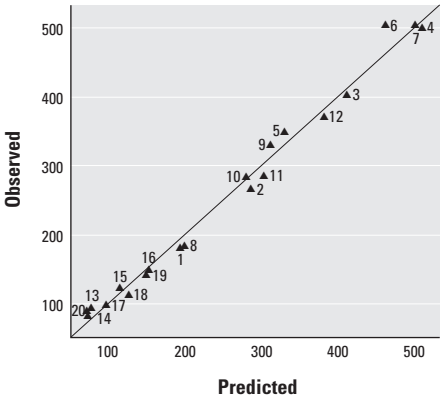

Figure 3. Observed versus predicted mutagenicity (revertants/mg PM) of 20 samples. Regression model with 41 variables.

Generally, samples to a regression model can be selected from score plots (*10,16*). Ideally, the samples should be evenly distributed in the score plot, approaching the situation obtained with statistically designed mixtures. Because naturally occurring mixtures are not statistically designed, a limited number of mixtures (samples) can be selected for mutagenicity testing and incorporated into the regression model from their GC–MS patterns and score plots.

Compared with bioassay-directed fractionation (*18*) on a column according to polarity, for example, the pattern recognition approach may be regarded as a "virtual" bioassay-directed fractionation. However, it cannot be done on one sample only. It requires a number of samples with different but overlapping composition. In this work, PLS defines a group of individual compounds that co-vary with mutagenicity, regardless of the physical–chemical properties of the compounds. As a consequence, the major contributors to mutagenicity are contained in one virtual fraction together with other compounds that co-vary with mutagenicity. The compounds in this virtual fraction may subsequently be identified chemically, as the number has been decreased significantly. After identification, they may be evaluated with respect to their mutagenicity. Those believed to be mutagenic may be used as variables in new experiments to identify their contribution to the overall mutagenicity, as well as possible interactions. This implies that the pattern recognition approach also may be used as the basis for the top-10 or pseudo top 10 approach (*5,6*). The pattern recognition approach itself does not give information about interactions unless the number of mixtures (samples) exceeds the total number of variables, permitting interaction terms to be included in the regression model. However, good linear regression models indicate additivity, and possible interactions may be insignificant.

This work emphasizes the strategy for toxicological evaluation of complex mixtures. Only the *Salmonella* TA98 strain without the addition of liver enzymes was used. Thus, only direct-acting mutagens such as nitro-PAHs and oxy-PAHs will be detected, whereas compounds that require metabolic activation by cytochrome P450 enzymes will not contribute (*8*). Consequently, the response matrix should be expanded, especially for risk assessment purposes. Other *Salmonella* strains should be included, e.g., TA100 and strains deficient in nitroreductases, and the mutagenicity testing should be carried out both with and without the addition of

metabolizing enzymes [this is discussed in more detail in our previous article (*7*)]. Furthermore, the application of toxicogenomics (*19*), for example, will generate multiple responses, i.e., a **Y** matrix with many response variables. PCA and PLS are most useful for analysis of **X** and **Y** matrixes with multiple predictor variables and responses, respectively.

We emphasize that the concept of pattern recognition also can be used when compounds are properly identified and quantified, e.g., by GC–MS SIM. However, it requires much more work. The advantage is higher sensitivity and accuracy in the quantification. Consequently, compounds present in very low concentrations may not be detected by full-scan GC–MS and it cannot be completely ruled out that some of these may be extremely mutagenic and contribute to the overall mutagenicity.

Pattern recognition for regression purposes can also be used on compounds other than PAHs and for end points other than mutagenicity. Figure 4 outlines the complete strategy for toxicological evaluation of complex mixtures based on pattern recognition. Detailed characterization, or fingerprinting, the latter after curve resolution and similarity matching, is used to create the **X** matrix (predictor variables). PCA is used for classification purposes. Toxicity or mutagenicity testing generates data to the **Y** matrix (responses). PLS is used for regression modeling to identify the peaks or compounds that co-vary with the responses. The peaks may subsequently be identified from their pure spectra and may be used in more detailed studies of impact and possible interactions. Furthermore, new samples should be evaluated by PCA to ascertain that they are within the regression domain before toxicity or mutagenicity is predicted.

We are presently improving the methodology to use regression models to predict mutagenicity of new samples and to convert the pure spectra from each resolved peak to a format useful for library search for identification.

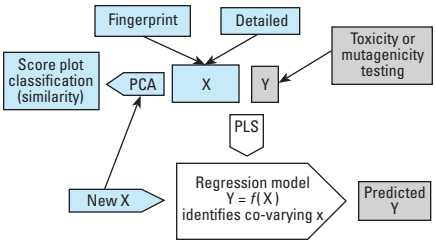

Figure 4. Schematic illustration of strategy for toxicologic evaluation of complex mixtures by pattern recognition.

References and Notes

1. Østby L, Engen S, Melbye A, Eide I. Mutagenicity testing of organic extracts of diesel exhaust particles after fractionation and recombination. Arch Toxicol 71:314–319 (1997).
2. Bostrøm, E, Engen S, Eide I. Mutagenicity testing of organic extracts of diesel exhaust particles after spiking with PAHs. Arch Toxicol 72:645–649 (1998).
3. Eide I, Johnsen HG. Mixture design and multivariate analysis in mixture research. Environ Health Perspect 106(suppl 6):1373–1376 (1998).
4. Verhaar HJM, Morroni JR, Reardon KF, Hays SM, Gaver DP, Carpenter RL, Yang SH. A proposed approach to study the toxicology of complex mixtures of petroleum products: the integrated use of QSAR, lumping analysis and PBPK/PD modeling. Environ Health Perspect 105(suppl 1):179–195 (1997).
5. Feron VJ, Groten JP, Jonker D, Cassee FR, van Bladeren PJ. Toxicology of chemical mixtures: challenges for today and the future. Toxicology 105:415–427 (1995).
6. Feron VJ, Cassee FR, Groten JP. Toxicology of chemical mixtures: international perspective. Environ Health Perspect 106(suppl 6):1281–1289 (1998).
7. Eide I, Neverdal G, Thorvaldsen B, Shen H, Grung B, Kvalheim O. Resolution of GC-MS data of complex PAC mixtures and regression modeling of mutagenicity by PLS. Environ Sci Technol 35:2314–2318 (2001).
8. Westerholm RN, Almén J, Hang L, Rannug JU, Egebäck KE, Grägg K. Chemical and biological characterization of particulate-, semivolatile-, and gas-phase-associated compounds in diluted heavy-duty diesel exhausts: a comparison of three different semivolatile-phase samplers. Environ Sci Technol 25:332–338 (1991).
9. Shen H, Grung B, Kvalheim OM, Eide I. Automated curve resolution applied to data from multi-detection instruments. Anal Chim Acta 446:313–328 (2001).
10. Maron DM, Ames BN. Revised methods for the *Salmonella* mutagenicity test. Mutat Res 113:173–215 (1983).
11. fWold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 5:735–743 (1984).
12. Kvalheim OM. Model building in chemistry, a unified approach. Anal Chim Acta 223:53–73 (1989).
13. Kettaneh-Wold N. Analysis of mixture data with partial least squares. Chemometrics Intelligent Lab Syst 14:57–69 (1992).
14. Grande B-V, Manne R. Use of convexity for finding pure variables in two-way data from mixtures. Chemometrics Intelligent Lab Syst 50:19–33 (2000).
15. Manne R, Grande B-V. Resolution of two-way data from hyphenated chromatography by means of elementary matrix transformations. Chemometrics Intelligent Lab Syst 50:35–46 (2000).
16. Jackson JE. A User's Guide to Principal Components. New York:John Wiley, 1991.
17. Wold S. Cross validatory estimation of the number of components in factor and principal components model. Technometrics 20:397–405 (1978).
18. Schuetzle D, Lewtas J. Bioassay-directed chemical analysis in environmental research. Anal Chem 58:1060A–1075A (1986).
19. Pennie WD, Tugwood GJA, Kimber OI. The principles and practice of toxicogenomics: applications and opportunities. Toxicol Sci 54:277–283 (2000).