Chemosphere 138 (2015) 176–182

Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Linking mutagenic activity to micropollutant concentrations in wastewater samples by partial least square regression and subsequent identification of variables

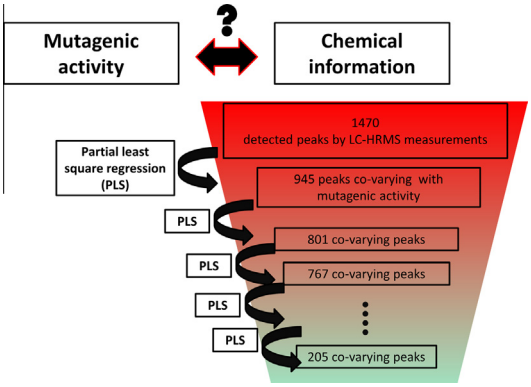
Christine Hug [a,b,*], Moritz Sievers [a], Richard Ottermanns [b], Henner Hollert [b], Werner Brack [a], Martin Krauss [a]

[a] UFZ – Helmholtz Centre for Environmental Research, Department of Effect-Directed Analysis, Permoserstr. 15, 04318 Leipzig, Germany
[b] RWTH Aachen University, Department of Ecosystem Analyses, Institute for Environmental Research, Worringerweg 1, 52074 Aachen, Germany

HIGHLIGHTS

- A sequence of wastewater treatment plant effluent samples was analyzed by LC–HRMS and Ames Fluctuation assay.
- Compounds co-varying with the mutagenicity were identified by "virtual" effect-directed analysis.
- Peak lists for identification were reduced by 86% using partial least squares projections.
- Compounds co-varying with mutagenicity were characterized and some identified.
- Identification and characterization of these compounds indicated an industrial source of mutagens.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form 16 May 2015
Accepted 22 May 2015

Keywords:
LC–HRMS
Mutagenicity
"virtual" effect-directed analysis
Feature selection

ABSTRACT

We deployed multivariate regression to identify compounds co-varying with the mutagenic activity of complex environmental samples. Wastewater treatment plant (WWTP) effluents with a large share of industrial input of different sampling dates were evaluated for mutagenic activity by the Ames Fluctuation Test and chemically characterized by a screening for suspected pro-mutagens and non-targeted software-based peak detection in full scan data. Areas of automatically detected peaks were used as predictor matrix for partial least squares projections to latent structures (PLS) in combination with measured mutagenic activity. Detected peaks were successively reduced by the exclusion of all peaks with lowest variable importance until the best model (high R^2 and Q^2) was reached. Peaks in the best model co-varying with the observed mutagenicity showed increased chlorine, bromine, sulfur, and nitrogen abundance compared to original peak set indicating a preferential selection of anthropogenic compounds. The PLS regression revealed four tentatively identified compounds, newly identified 4-(dimethylamino)-pyridine, and three known micropollutants present in domestic wastewater as co-varying with the mutagenic activity. Co-variance between compounds stemming from industrial wastewater and mutagenic activity supported the application of "virtual" EDA as a statistical tool to separate toxicologically relevant from less relevant compounds.

* Corresponding author at: UFZ – Helmholtz Centre for Environmental Research, Department of Effect-Directed Analysis, Permoserstr. 15, 04318 Leipzig, Germany.
E-mail address: christine.hug@ufz.de (C. Hug).

http://dx.doi.org/10.1016/j.chemosphere.2015.05.072
0045-6535/

1. Introduction

Over the last decades an increasing number of pollutants has been detected in surface water. The majority of these substances originates from human use and enters surface water via treated and untreated wastewater (Reemtsma et al., 2006; Nikolaou et al., 2007). Simultaneously, the biological characterization of wastewater treatment plant (WWTP) effluents revealed genotoxic, mutagenic, estrogenic, and other effects (Claxton et al., 1998; Miège et al., 2009). Mutagenic and genotoxic effects were reported to trigger long term effects on the survival of species (Anderson and Wild, 1994), but only limited information is available about compounds responsible for these effects.

To identify compounds with adverse effects in the environment, both chemical and biological characterization was combined to establish cause-effect relationships between compounds present and biological effects (Vermeirssen et al., 2010; Smital et al., 2011). To this end, two different approaches have been suggested. (1) Effect-directed analysis (EDA) is based on an experimental reduction of mixture complexity using chromatographic techniques to isolate active fractions for toxicant identification (Brack, 2003). Site-specific effect-relationships between chemical contamination and mutagenicity were investigated in several EDA studies (Reifferscheid et al., 2011; Gallampois et al., 2013). However, EDA is still costly and time-consuming and thus so far its application on a larger scale is limited. (2) An alternative approach using multivariate statistics to reduce the complexity of environmental contamination by correlation of effects with chemical analytical signals has been propagated as "virtual" EDA (Eide et al., 2002, 2004). In contrast to EDA, which focuses on a small number of samples, "virtual" EDA needs to be applied to larger sets of samples in space or time. Although this approach does not provide cause-effect relationships as such it has the potential to extract relevant information out of large datasets to derive hypotheses, which can be confirmed experimentally.

A large number of often co-varying variables (components) and a too small number of samples are major limitations for the application of multivariate statistics to environmental samples. PLS analysis may overcome this limitation since it tolerates data sets with co-varying predictors, large numbers of variables exceeding the number of observations (Kettaneh-Wold, 1992) and strongly co-varying, collinear data matrices (Wold et al., 1984).

The major goal of PLS analysis is the discrimination of variables co-varying with the response from those not co-varying. PLS has been demonstrated to provide meaningful correlation of chemical fingerprints of exhaust particle extracts and mutagenicity, but also of biotic indicators of surface water quality and landscape conditions (Eide et al., 2002; Nash and Chaloud, 2011).

This study presents "virtual" EDA as a tool for the exploitation of combined chemical and toxicological monitoring data for the characterization of chemicals correlating with observed effects. The exercise was conducted using a sequence of six weekly taken effluent samples from a mixed industrial and municipal WWTP effluent exhibiting varying mutagenicity in the Ames Fluctuation Test after activation with S9. To this end, mutagenicity testing was combined with a LC–HRMS non-target screening (Hug et al., 2014) supplemented with a suspect screening for a set of (pro)-mutagenic aromatic amines since this compound group requires S9 activation and has been made responsible for mutagenicity in surface waters in the past (Kataoka et al., 2000; Fukazawa et al., 2001). Mutagenic activity and all detected peaks were subjected to a PLS regression to filter out those peaks co-varying with the mutagenicity by a stepwise dimensionality reduction of the model. Differences between peaks identified as co-varying with the mutagenicity and all detected peaks were determined and evaluated for their plausibility.

2. Materials and methods

2.1. Sampling and extraction

The effluent from the WWTP Bitterfeld-Wolfen, Saxony-Anhalt, Germany was grab-sampled at the outlet into the river Mulde once a week for six weeks using 5 L aluminum containers. Samples were named as sample 1 to sample 6 based on their order of sampling and stored for up to 5 days at −20 °C before extraction.

After filtration through glass fiber filters (GF/F Whatman), a volume of 20 L of WWTP effluent (pH 7–7.8) was extracted by solid-phase extraction (SPE) using 4 g of Chromabond HR-X sorbent (Macherey-Nagel) in Omnifit columns (Diba Industries Ltd) and a preparative HPLC pump (NoraPrep 200, Merck). After percolation through the SPE cartridge the aqueous phase was adjusted to pH 3 with formic acid and extracted a second time. The sorbent from both extractions was eluted with 300 mL of methanol, 200 mL of methanol containing 0.2% formic acid and 200 mL of methanol/acetone (80:20; v:v). The combined extracts were neutralized, evaporated to dryness and redissolved in methanol to a concentration factor of 1000. The ability of the SPE method to extract compounds with a wide range of physico-chemical properties has been shown before (Hug et al., 2014). Defined aliquots of samples were dried under nitrogen and re-dissolved in dimethyl sulfoxide (DMSO) for biological assessment. Prior to LC–MS analysis an aliquot of the methanol extract was filtered using a PTFE syringe filter (0.45 μm, Macherey-Nagel) and diluted with three parts of bidistilled water. A blank sample from 10 L of bidistilled water was prepared applying the described extraction procedure.

2.2. Mutagenicity testing and prediction

The Ames Fluctuation Test (AFT) was performed as described in Reifferscheid et al. (2012) with slight modifications using a TA98 tester strain with and without metabolic activation by S9 on 384-well microplates. Mutagenic samples induced a reversion from the auxotrophic to the prototrophic genotype indicated by a color change of the pH indicator bromocresol purple. Spontaneous reversion was evaluated in eight replicates with DMSO as negative control. The sensitivity of the TA98 strain was monitored with a dilution series of 2-nitrofluorene for tests without and 2-aminoanthracene for tests with metabolic activation. The mutagenic activity (given as revertants per L wastewater in L methanolic extract) was determined by fitting the number of positive wells to an exponential equation, see equation S1 and Supplementary Material (SM) S1.6. (Gallampois et al., 2013). The mutagenic activity was determined in triplicates with eight dilutions in three independent tests.

2.3. LC–MS/MS analyses and data processing

The liquid chromatography separation was performed with an Agilent 1200 UPLC system equipped with a Kinetex™ Core-Shell C18 column (100 mm × 3.0 mm; 2.6 μm; Phenomenex) with a linear gradient elution with water and methanol both containing 0.1% formic acid with a flow of 0.2 mL/min. The initial fraction of 10% methanol was increased after 3.2 min to 95% within a linear gradient in 17.8 min. Menthanol was maintained at 95% for 20 min followed by a re-equilibration for 9 min. The LC system was connected to an ion trap-Orbitrap hybrid instrument (LTQ Orbitrap XL, Thermo Scientific). Analytes were ionized by electrospray ionization in positive (ESI+) and negative (ESI−) ion mode.

Full scan chromatograms were acquired with a nominal resolving power of 60,000 (full width height maximum referenced to m/z 400, $FWHM_{400}$) and a mass deviation <5 ppm.

Data-dependent high-resolution product ion spectra (HRMS/MS) were recorded in a separate run. Parent peak lists were assembled from identified masses co-varying with mutagenicity (see Section 2.4) and masses of suspect promutagens. Collision induced fragmentation spectra were recorded at normalized collision energies of 35% and 50% and higher-energy collisional dissociation spectra at 50%, 70%, 90%, 120% at a nominal resolving power of 15,000 ($FWHM_{400}$).

For statistical analysis, ESI± data files were converted to the mzXML format and transformed to centroid data using the ProteoWizard v3.0 software (Chambers et al., 2012). Peak detection was performed in R, v2.15 (RDevelopmentCoreTeam, 2010), using the packages `XCMS` (Smith et al., 2006), and `RANN` (Kemp and Jefferis, 2012), see S1.4. Peaks in different samples were grouped based on their accurate mass and retention time applying an m/z width of 0.0015. Tables with peak groups defined by retention time and accurate mass were further processed in Microsoft Excel. Peaks with a sample to blank area ratio <10 were excluded from analysis.

Peak lists were screened for 44 pro-mutagenic amines which have been identified in environmental samples before (Table S1). Reference standards, retention times and HRMS/MS spectra were available for 22 of these pro-mutagens, others were screened for in a suspect screening process as described by Hug et al. (2014). ESI+ chromatograms were searched for exact masses of $[M+H]^+$ ions of the 44 amines with a tolerance of ±5 ppm. Peaks were evaluated by a query in MassBank (Schulze et al., 2012) and an analysis of HRMS/MS spectra by MetFrag (Wolf et al., 2010).

2.4. Multivariate analysis

Multivariate data analysis was performed in R using the packages `PLS` (Mevik et al., 2013), `ade4` (Dray and Dufour, 2007), `vegan` (Oksanen et al., 2012), and `gclus` (Hurley, 2012). Logarithmic transformed predictor matrix was analyzed by principal component analyses (PCA) (Jackson, 2003) for possible outliers (details see S1.7).

After a screening for known pro-mutagens, all peaks were analyzed to reveal co-variances between mutagenicity and peaks by PLS regression (Wold et al., 1984). In a PLS regression it is assumed that mutagenic activity does not co-vary with all variables, but a few latent variables. These latent variables, in our case chromatographic peaks are identified.

PLS analysis only performs sufficiently, if peaks co-varying with the mutagenicity are detected in a minimum fraction of samples included in analysis (Kettaneh-Wold, 1992). Thus, only peaks detected in at least three samples with a minimum intensity of $1 \cdot 10^5$ were considered for predictor matrix X for statistical analysis. In the predictor matrix X each row represents one sample and each column one peak defined by retention time and accurate mass. Not detected peaks are represented by a zero. The response vector y was the mutagenicity of all samples.

Thus, after scaling to unit variance peaks or peak groups which co-vary with mutagenicity are discriminated by PLS regression from non-co-varying peaks by variables' influence on projection (VIP), which is calculated based on the weighted sums of the absolute regression coefficients. Variables' importance is proportional to its reduction in the sums of squares.

Starting with a PLS model including all variables the complexity of the model is sequentially reduced by eliminating variables with an influence on projection below 0.8 (Thomas et al., 1998). The goal was the establishment of models with a maximum predictive power (Q^2), a maximum correlation coefficient of the matrix X and vector y (R^2Y) and a minimum of variables (peaks) involved. The parameter Q^2 represents the coefficient of determination from the cross-validation (Wold, 1978), thus the predictive power of the model.

A coincidental selection of variables was excluded by the selection of 500 peak sets containing 205 randomly selected peaks, which were analyzed by PLS regression. Q^2 values from PLS models based on the randomly chosen subsets were compared to values of subsets systematically selected by VIP.

2.5. Analysis of peaks selected by PLS

Lists of all peaks and of peaks selected by PLS were analyzed using the R package `nontarget` to detect ^{13}C, ^{37}Cl, ^{81}Br, ^{15}N, ^{29}Si, and ^{34}S isotope peaks (Loos, 2012), see S1.5. The peaks detected as co-varying with mutagenicity by PLS were compared for their mass and retention time distribution, for the detection of isotope peaks and common substructures with the whole data set.

Compounds co-varying with the mutagenic activity were subjected to the identification process described by Hug et al. (2015). Molecular formulas within a ±5 ppm window were transformed from accurate m/z values to masses of the neutral molecules assuming an ionization as $[M+H]^+$ in ESI+ and as $[M-H]^-$ in ESI−. Molecular formulas were calculated by the QualBrowser including the elements C, H, N, S, O, P, F, Cl, Br, Si, and in ESI+ Na. The number of C, Cl, Br, N, S, and Si atoms was gained from the isotope patterns with a tolerance of ±1 for C and ±50% for N due to the low intensity of the N isotope peak. Molecular formulas were checked for their plausibility by the Seven Golden Rules (SGR) fast checker (Kind and Fiehn, 2007). HRMS/MS spectra were acquired for all peaks with confirmed molecular formula. The spectra were visually and automatically evaluated for fragments by MetFrag (Wolf et al., 2010) in combination with a ChemSpider query (mass tolerance 5 ppm) and a search in MassBank (Schulze et al., 2012). Retention time was predicted for all candidates applying a LSER retention time prediction model with an admissible tolerance of the 95%-prediction interval (Ulrich et al., 2011; Hug et al., 2014). Predicted retention time was not considered for candidate structures dissociated at the chromatographic conditions. Log K_{OW} and pK_a values were calculated in JChem for Excel version 6.2 (Chemaxon, Hungary).

The mutagenic activity of candidate structures was predicted using the VEGA software (http://www.vega-qsar.eu), details see S1.6.

3. Results and discussion

3.1. Mutagenic activity and chemical composition of WWTP effluent

All samples showed mutagenic activity in the AFT only after metabolic activation by S9 mix. The order of mutagenic activity after metabolic activation with S9 (revertants/mL wastewater in mL wastewater extract in brackets) was: Sample 1 (1.77) > Sample 4 (1.57) > Sample 3 (1.11) > Sample 2 (0.94) > Sample 5 (0.70) > Sample 6 (0.57) see Table S2.

3.2. Dataset for statistical analysis

After no known pro-mutagens could be identified, all detected peaks were analyzed by PLS regression. By automated peak picking in `XCMS` 949 peaks defined by accurate mass and retention time fulfilled the criteria stated above in ESI+, 565 in ESI−. Thus, the predictor matrix for the PLS analysis contained 1470 peaks. On average, 1347 peaks (1213–1430) were detected per sample.

Peaks were mainly detected at retention times between 2.2 and 5 min and 15 and 30 min (Fig. 1), which is corresponding to log K_{OW} values between −1.8 and −1.2 and 1.0 and 4.3,

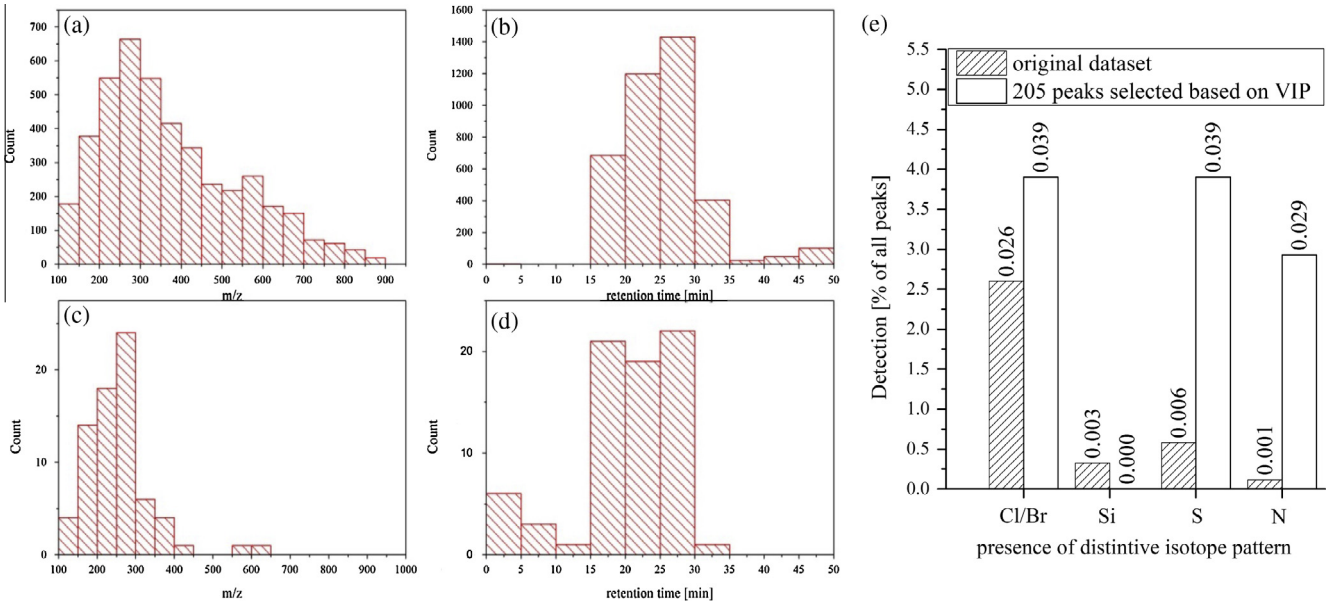


Fig. 1. Distribution of m/z values and retention time of peaks in original dataset (a and b) and of the 205 peaks selected based on VIP (c and d); (e) detection of isotope pattern by R nontarget in the original dataset and for peaks selected based on VIP. Frequency of detection is given as percent of all peaks in the evaluated dataset. Chlorine and bromine peaks are not distinguished by the R non-target analysis, because the mass difference of M + 2 isotope peaks of chlorine and bromine is below 3 ppm for molecules with m/z > 300 a.u. and intensity evaluations are only based on the M + 2 isotope peaks.

respectively (SM S2.3). Peaks eluting between 2.2 and 5 min indicate that compounds are highly hydrophilic, eluting close to the chromatographic dead time of the column. The evaluation by the R `nontarget` package assigned that 2.6% of the detected compounds contained Cl or Br, 0.3% Si, 0.6% S, and 0.1% N. The samples likely contain a higher number of compounds containing Cl, Br, Si, and N, because low intensity isotope peaks might not be reliably found by automated peak detection. The underestimation of the number of isotope peaks was assumed to be the same for all peaks and peaks selected by PLS regression, thus the results from isotope peaks are appropriate for a comparison of datasets.

3.3. Multivariate evaluation

Before PLS regression, the 6 × 1470 predictor matrix of areas of automatically detected peaks was analyzed for potential outliers (see S2.4). Based on the quantile-quantile plots of the first two PCs of the PCA, none of the samples was identified as outlier, see Fig. S3. Then the predictor matrix was analyzed for the co-variance of automatically detected peaks with mutagenic activity by a PLS regression. By the consecutive removal of all variables with an influence on projection <0.8 (Wold et al., 2001) the number of variables was reduced from 1470 to 51 with the intermediate levels shown in Table 1. Concurrently, the fraction of the mutagenic activity explained by the model (R^2Y) and the cross-validates prediction R^2 (Q^2) increased to a maximum for the R^2Y of 1.00 and 0.92, respectively. The further reduction of variables led to the decrease of R^2Y and Q^2. The PLS model with 62 compounds showed the highest Q^2 value of all models, whereas the R^2Y decreased from 1.00 to 0.70 by the reduction of variables from 205. This is also visible in Fig. S6 showing the predicted versus the measured mutagenic activity. Thus, we assume the model with 205 components to be the model with best predictive power and fit as it has the lowest residues between predicted and experimental mutagenicity and a higher Q^2 than for other models with fewer variables.

Peaks included in the optimum PLS model (No. 7, Table 1) were characterized, screened for target compounds listed in Hug et al., 2014 and subjected to an identification process.

Table 1
Number of components used for PLS regression, number of peaks included in the PLS model, correlation coefficient of the matrix X and vector y (R^2Y), and cross-validates prediction R^2 (Q^2) of these PLS models and corresponding values for the validation of the model No. 7.

PLS No	PLS components	R^2Y	Q^2	Number of peaks
1	1	0.90	0.46	1470
2	1	0.90	0.73	945
3	1	0.90	0.65	801
4	1	0.90	0.71	767
5	1	0.91	0.82	643
6	2	0.99	0.85	337
7	**2**	**1.00**	**0.84**	**205**
8	1	0.70	0.92	62
9	1	0.88	−0.27	53
10	1	0.89	−1.83	51
Validation based on randomly chosen variables and PLS model No. 7	2		Min −1.22 Max −0.53	205

Values for the best PLS model are presented in bold.

For the validation of the established model, 500 groups of 205 peaks were randomly selected from the predictor matrix and evaluated by a PLS model. Thereby, a coincidental selection of 205 variables for the best PLS model was excluded. Table 1 and Fig. S7 show the obtained R^2Y and Q^2 values. The Q^2 values for randomly selected predictor matrices ranged from −1.22 to −0.53, compared to 0.84 for the model established by the VIP based selection of variables. Thus, peaks selection during the PLS was not coincidental, but peaks showed an above-average co-variance with the mutagenic activity.

3.4. Analysis of the peaks detected as co-varying with mutagenic activity

Compounds co-varying with the mutagenic activity were characterized and showed the following differences compared to all

peaks for m/z values, retention time, and the abundance of chlorine, bromine, nitrogen, and sulfur:

The distribution of masses in both, the restricted and the complete dataset, centered around 250, but the restricted data set of 205 peaks from PLS analysis includes fewer peaks with masses >300 m/z (Fig. 1a and c). Most of the peaks were detected in a retention time range of 15–30 min for both datasets, whereas a larger fraction of peaks with low retention times were detected for the dataset co-varying with mutagenic activity. For the original dataset more peaks were detected with a retention time >30 min (Fig. 1b and d). The retention time of compounds correlating with the mutagenicity between 15 min and 30 min corresponds to a log K_{OW} of 0.9–3.3 (see S2.3) and is in accordance with results from a previous study (Hug et al., 2015).

More peaks with S, N, and Br or Cl isotope pattern were detected in the dataset selected by PLS, see Fig. 1. The increased frequency of N, S, and halogens in the PLS dataset suggests a preferential selection of anthropogenic compounds and an exclusion of non-anthropogenic effluent organic matter compounds (e.g., fatty acids and other microbial metabolites), which mainly consist of C, H, and O. The increased presence of N in peaks correlating with mutagenicity is in agreement with predominance of N-containing compounds among known metabolically activated mutagens (Ahlberg et al., 2014; Kazius et al., 2005). However, none of the suspected pro-mutagenic aromatic amines was detected in the samples.

Seventeen peaks without a Lorentzian peaks shape and 35 with an intensity $<1 \cdot 10^5$ of the 205 peaks were excluded from further processing. For three of 153 remaining peaks the SGR fast checker did not confirm any possible molecular formula. Seventeen of the molecular formulas did neither have an entry in the ChemSpider, nor in the PubChem database. These peaks were put on hold since commercial standards for these compounds are not available and would prohibit for a straightforward identification. Candidates found for the molecular formula were ranked using the MetFrag score for the HRMS/MS spectra (Wolf et al., 2010) in combination with a ChemSpider search, a query in MassBank (Schulze et al., 2012) and the prediction of the chromatographic retention by a LSER model (Ulrich et al., 2011; Hug et al., 2014).

From the remaining 133 peaks and molecular formulas eight compounds could be identified or tentatively identified, which are shown in Table 2. For these compounds a molecular formula could be confirmed by HRMS/MS spectra and retention time prediction.

Three target compounds (diclofenac, benzothiazole-2-sulfonic acid, phenylbenzimidazole sulfonic acid) and two tentatively identified nontarget compounds from a previous study (6,7,8,9-tetrahydrobenzocyclohepten-5-one, naphthalenedisulfonic acid benzenesulfonamide) (Hug et al., 2014) were found among these 133 compounds. These tentatively identified compounds showed a good agreement of predicted and measured HRMS/MS spectra, retention time and the number of hydrogens exchangeable by deuterium, but their identity was not confirmed due to the absence of reference standards.

In addition, we newly identified 4-(dimethylamino)-pyridine (4-DMAP) co-varying with the mutagenic activity of the samples. 4-DMAP is produced by the chemical industry discharging via the WWTP, is used as nucleophilic catalyst in industrial processes (Grondal, 2003), but was reported as not mutagenic (Ogawa et al., 1988). An ethoxy(pyrrolidinyl)benzaldehyde and 2-ethoxy-N-(1-methyl-3-phenylpropyl)-acetamide were tentatively identified based on matching predicted and experimental retention times and HRMS/MS information. Literature spectra or commercially available standards for confirmation were lacking. All tentatively identified compounds were predicted to be not mutagenic by at least two out of three models, thus they are presumably not mutagenic. Based on these results, a synthesis of these compounds was not considered. The low number of positively identified compounds is caused by the lack of comprehensive LC–HRMS/MS databases, uncertainty in retention time prediction for ionic compounds at the chromatographic conditions like aromatic amines and the absence of reference standards as discussed in detail in Hug et al. (2014).

Benzothiazole-2-sulfonic acid is used in rubber and dye production (Hunger, 2003; Kloepfer et al., 2005), but is also assumed to be an oxidation product of methylbenzothiazoles, thus stems from domestic and industrial sources (Reemtsma et al., 2002). Naphthalenedisulfonic acid benzenesulfonamides are used in dye production (Sakalis et al., 2004), as dispersants (Altenbach and Giger, 1995), and additives in cement and concrete (Dodson, 1990). Naphthalene sulfonic acids were detected in industrial wastewater (Storm et al., 1999) and their application is restricted to industry (Altenbach and Giger, 1995). The operator of the WWTP reported azo dye production for the time of sampling (personal communication), thus we would assume industrial application or production of azo dyes as most likely source of the naphthalenedisulfonic acid benzenesulfonamide. While the candidates were predicted as not mutagenic by prediction models, very similar structures were reported to be mutagenic (Chung and Cerniglia, 1992), thus degradation products, analogously produced or applied compounds might contribute to the mutagenic effect. 6,7,8,9-Tetrahydrobenzocyclohepten-5-one was reported as educt for the synthesis of pharmaceuticals (Boulos et al., 2012) or might stem from the synthesis of 4-methyltetralone in the industrial area connected to the WWTP. Thus, several compounds co-varying with the mutagenic activity originate completely or partially from industrial processes, some were assigned to specific branches.

4. Conclusions

While "virtual" EDA has been suggested earlier, a study demonstrating the potential power of this approach was still missing. In the present study the integration of LC–HRMS/MS analysis, mutagenicity testing and PLS multivariate statistics was able to reduce a large number of peaks detected in a WWTP effluent to only 14% of them co-varying with mutagenicity. The elemental composition of these 205 remaining peaks clearly discriminates the subset from non-co-varying peaks and supports their industrial origin. The selection of compounds eluting in the range between 15 and 30 min is in accordance with a conventional EDA study on samples from the WWTP effluent (Hug et al., 2015).

Compounds selected by PLS are assumed to either be mutagenic or to occur in concentrations co-varying with those of the mutagens causing the observed effect. The latter suggests the same sources or processes emitting both sets of chemicals. While the applied PLS regression should not be misunderstood as a tool to predict mutagenic activity from a limited set of chemicals, the statistical analysis supported the identification of suspicious industrial processes and is a basis for subsequent studies analyzing these processes for the formation of potential mutagens as candidate chemicals for toxicant identification. The number of available samples (6) was rather small in this study. The selectivity of this "virtual" fractionation method is assumed to rise with an increasing numbers of samples.

Major pre-condition of a promising application of "virtual" EDA is the existence of sequence of samples, which are of similar composition but with varying concentrations of individual chemicals and varying toxic effects. The approach may be used for temporal sequences of effluent water samples as in this study but are also promising for river basin, national or European scale monitoring as long as sampling and analysis are harmonized. While

Table 2
Compounds (tentatively) identified as identity of one of the 205 peaks co-varying with the mutagenic activity.

Status	Compound	Exact mass (amu)	t_r (min)	Ion detected	Molecular formula	Source	Mutagenic activity	
							Predicted	Exp.
Tent. Identified[a]	6,7,8,9-Tetrahydrobenzocyclohepten-5-one	161.0960	24.7	$[M+H]^+$	$C_{11}H_{12}O$	Industry	Not mutagenic	n.d.
Tent. Identified[a]	Naphthalenedisulfonic acid benzenesulfonamide	457.9675	16.5	$[M-H]^-$	$C_{16}H_{13}NO_9S_3$	Industry	Not mutagenic	n.d.
Identified	Benzothiazole-2-sulfonic acid	213.9638	17.0	$[M-H]^-$	$C_7H_5NO_3S_2$	Industry/ Domestic	Not mutagenic	Negative
Identified	Phenylbenzimidazole sulfonic acid	273.0337	15.7	$[M-H]^-$	$C_{13}H_{10}N_2O_3S$	Domestic	Not mutagenic	Negative
Identified	Diclofenac	294.0093	26.6	$[M-H]^-$	$C_{14}H_{11}Cl_2NO_2$	Domestic	Not mutagenic	Negative
Identified (nontarget)	4-(Dimethylamino) pyridine	122.0844	3.8	$[M+H]^+$	$C_7H_{10}N_2$	Industry	Not mutagenic	Negative
Tent. Identified	2-Ethoxy-4-(1-pyrrolidinyl)benzaldehyde/ 3-Ethoxy-4-(1-pyrrolidinyl)benzaldehyde	219.1259	25.7	$[M+H]^+$	$C_{13}H_{17}O_2N$	Industry[b]	Mutagenic (low reliability)	n.d.[c]
Tent. Identified	2-Ethoxy-N-(1-methyl-3-phenylpropyl)-acetamide	235.1572	20.7	$[M+H]^+$	$C_{14}H_{21}NO_2$	Not known	Not mutagenic	n.d.[c]

[a] Tentatively identified by Hug et al. (2014). 2-Methyl-1-tetralone was excluded as candidate for this compound by the acquisition of the reference standard, which resulted in a HRMS/MS spectrum deviating from the one in the sample. HRMS/MS spectra of 4-(dimethylamino)-pyridine in sample and reference standard are given in Figs. S8–S11.
[b] 4-(1-pyrrolidinyl)benzaldehyde is produced in the industrial area connected to the WWTP. Experimental (exp.) mutagenic activity was either negative in the AFT.
[c] Not determined (n.d.) because no reference standard was available.

demonstrating the potency of virtual EDA to exclude chemical signals without relevance for observed effects, this approach does not confirm cause-effect relationships between specific contaminants and observed (and co-varying) effects. Virtual EDA may be seen as a promising tool to support long term or large scale monitoring programs by deriving hypotheses on causing agents for further confirmation by more in-depth studies at specific sites or time points. Thus, virtual EDA may serve as a link in a chain of tools for the characterization, identification and assessment of complex environmental contamination.

Acknowledgments

The authors thank Angela Sperreuter and Margit Petre for their excellent technical support, Marion Heinrich for her outstanding support during the sampling campaign, MelisMuz for her support in selection of promutagens, Frédéric Sans-Piché, Janet Riedl, EginaMalaj and Carolina Vogs for their support in R. A free academic license for JChem, InstantJChem and the Calculator Plugins was kindly provided by ChemAxon (Budapest, Hungary). This study was performed within the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) and the EU FP7 project SOLUTIONS (grant agreement no. 603437).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chemosphere.2015.05.072.

References

Ahlberg, E., Carlsson, L., Boyer, S., 2014. Computational derivation of structural alerts from large toxicology data sets. J. Chem. Inf. Model. 54, 2945–2952.
Altenbach, B., Giger, W., 1995. Determination of Benzenesulfonates and Naphthalenesulfonates in waste-water by solid-phase extraction with graphitized carbon-black and ion-pair liquid-chromatography with UV detection. Anal. Chem. 67, 2325–2333.
Anderson, S.L., Wild, G.C., 1994. Linking genotoxic responses and reproductive success in ecotoxicology. Environ. Health Perspect. 102, 9–12.
Boulos, L.S., Abdel-Malek, H.A., El-Sayed, N.F., 2012. Synthesis of novel Benzosuberone derivatives using organophosphorus reagents and their antitumor activities. Z. Naturforsch B 67, 243–252.
Brack, W., 2003. Effect-directed analysis: a promising tool for the identification of organic toxicants in complex mixtures? Anal. Bioanal. Chem. 377, 397–407.
Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P., 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. 30, 918–920.
Chung, K.T., Cerniglia, C.E., 1992. Mutagenicity of azo dyes – structure–activity-relationships. Mutat. Res. 277, 201–220.
Claxton, L.D., Houk, V.S., Hughes, T.J., 1998. Genotoxicity of industrial wastes and effluents. Mutat. Res. 410, 237–243.
Dodson, V.H., 1990. Concrete Admixtures. Van Nostrand Reinhold, New York.
Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22, 1–20.
Eide, I., Neverdal, G., Thorvaldsen, B., Arneberg, R., Grung, B., Kvalheim, O.M., 2004. Toxicological evaluation of complex mixtures: fingerprinting and multivariate analysis. Environ. Toxicol. Pharmacol. 18, 127–133.
Eide, I., Neverdal, G., Thorvaldsen, B., Grung, B., Kvalheim, O.M., 2002. Toxicological evaluation of complex mixtures by pattern recognition: correlating chemical fingerprints to mutagenicity. Environ. Health Perspect. 110, 985–988.
Fukazawa, H., Matsushita, H., Terao, Y., 2001. Identification of co-mutagenic chlorinated harmans in final effluent from a sewage treatment plant. Mutat. Res.-Gen. Toxicol. Environ. 491, 65–70.
Gallampois, C.M.J., Schymanski, E.L., Bataineh, M., Buchinger, S., Krauss, M., Reifferscheid, G., Brack, W., 2013. Integrated biological-chemical approach for the isolation and selection of polyaromatic mutagens in surface waters. Anal. Bioanal. Chem. 405, 9101–9112.
Grondal, C., 2003. 4-Dimethylamino-pyridine (DMAP). Synlett, 1568–1569.
Hurley, C.; 2012. gclus: Clustering Graphics. R Package Version 1.3.1.
Hug, C., Krauss, M., Nüsser, L., Hollert, H., Brack, W., 2015. Metabolic transformation as a diagnostic tool for the selection of candidate promutagens in effect-directed analysis. Environ. Pollut. 196, 114–124.
Hug, C., Ulrich, N., Schulze, T., Brack, W., Krauss, M., 2014. Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. Environ. Pollut. 184, 25–32.
Hunger, K., 2003. Industrial Dyes: Chemistry, Properties, Applications. Wiley-VCH, Weinheim.
Jackson, J.E., 2003. A User's Guide to Principal Components. Wiley-Interscience, Hoboken, N.J..
Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P.M., Stevens, H.H., Wagner, H., 2012. vegan: Community Ecology Package.
Kataoka, H., Hayatsu, T., Hietsch, G., Steinkellner, H., Nishioka, S., Narimatsu, S., Knasmuller, S., Hayatsu, H., 2000. Identification of mutagenic heterocyclic amines (IQ, Trp-P-1 and A alpha C) in the water of the Danube River. Mutat. Res.-Gen. Toxicol. Environ. 466, 27–35.
Kazius, J., McGuire, R., Bursi, R., 2005. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48, 312–320.
Kemp, S.E., Jefferis, G., 2012. RANN: Fast Nearest Neighbour Search.
Kettaneh-Wold, N., 1992. Analysis of mixture data with partial least-squares. Chemometr. Intell. Lab. 14, 57–69.
Kind, T., Fiehn, O., 2007. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics 8.

Kloepfer, A., Jekel, M., Reemtsma, T., 2005. Occurrence, sources, and fate of benzothiazoles in municipal wastewater treatment plants. Environ. Sci. Technol. 39, 3792–3798.

Loos, M., 2012. nontarget: Detecting, combining and filtering isotope, adduct and homologue series relations in high-resolution mass spectrometry (HRMS) data.

Mevik, B.-H., Wehrens, R., Hovde Liland, K., 2013. pls: Partial Least Squares and Principal Component Regression.

Miège, C., Karolak, S., Gabet, V., Jugan, M.L., Oziol, L., Chevreuil, M., Levi, Y., Coquery, M., 2009. Evaluation of estrogenic disrupting potency in aquatic environments and urban wastewaters by combining chemical and biological analysis. Trac-Trends Anal. Chem. 28, 186–195.

Nash, M.S., Chaloud, D.J., 2011. Partial Least Square Analyses of landscape and surface water biota associations in the Savannah river basin. ISRN Ecol., 2011

Nikolaou, A., Meric, S., Fatta, D., 2007. Occurrence patterns of pharmaceuticals in water and wastewater environments. Anal. Bioanal. Chem. 387, 1225–1234.

Ogawa, H.I., Liu, S.Y., Sakata, K., Niyitani, Y., Tsuruta, S., Kato, Y., 1988. Inverse correlation between combined mutagenicity in Salmonella-Typhimurium and strength of coordinate bond in mixtures of Cobalt(Ii) Chloride and 4-substituted Pyridines. Mutat. Res. 204, 117–121.

RDevelopmentCoreTeam, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0.

Reemtsma, T., Weiss, S., Mueller, J., Petrovic, M., González, S., Barcelo, D., Ventura, F., Knepper, T.P., 2006. Polar pollutants entry into the water cycle by municipal wastewater: a European perspective. Environ. Sci. Technol. 40, 5451–5458.

Reemtsma, T., Zywicki, B., Stueber, M., Kloepfer, A., Jekel, M., 2002. Removal of sulfur-organic polar micropollutants in a membrane bioreactor treating industrial wastewater. Environ. Sci. Technol. 36, 1102–1106.

Reifferscheid, G., Buchinger, S., Cao, Z., Claus, E., 2011. Identification of mutagens in freshwater sediments by the Ames-fluctuation assay using nitroreductase and acetyltransferase overproducing test strains. Environ. Mol. Mutagen. 52, 397–408.

Reifferscheid, G., Maes, H.M., Allner, B., Badurova, J., Belkin, S., Bluhm, K., Brauer, F., Bressling, J., Domeneghetti, S., Elad, T., Flückiger-Isler, S., Grummt, H.J., Guertler, R., Hecht, A., Heringa, M.B., Hollert, H., Huber, S., Kramer, M., Magdeburg, A., Ratte, H.T., Sauerborn-Klobucar, R., Sokolowski, A., Soldan, P., Smital, T., Stalter, D., Venier, P., Ziemann, C., Zipperle, J., Buchinger, S., 2012. International round-robin study on the Ames fluctuation test. Environ. Mol. Mutagen. 53, 185–197.

Sakalis, A., Ansorgová, D., Holčapek, M., Jandera, P., Voulgaropoulos, A., 2004. Analysis of sulphonated azodyes and their degradation products in aqueous solutions treated with a new electrochemical method. Int. J. Environ. Anal. Chem. 84, 875–888.

Schulze, T., Schymanski, E., Stravs, M., Neumann, S., Krauss, M., Singer, H., Hug, C., Gallampois, C., Hollender, J., Slobodnik, J., Brack, W., 2012. NORMAN MassBank: towards a community-driven, open-access accurate mass spectral database for the identification of emerging pollutants. Norman Bulletin 3 NORMAN Network, 9–11.

Smital, T., Terzic, S., Zaja, R., Senta, I., Pivcevic, B., Popovic, M., Mikac, I., Tollefsen, K.E., Thomas, K.V., Ahel, M., 2011. Assessment of toxicological profiles of the municipal wastewater effluents using chemical analyses and bioassays. Ecotoxicol. Environ. Saf. 74, 844–851.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. 78, 779–787.

Storm, T., Reemtsma, T., Jekel, M., 1999. Use of volatile amines as ion-pairing agents for the high-performance liquid chromatographic-tandem mass spectrometric determination of aromatic sulfonates in industrial wastewater. J. Chromatogr. A 854, 175–185.

Thomas, D.R., Hughes, E., Zumbo, B.D., 1998. On variable importance in linear regression. Soc. Indic. Res. 45, 253–275.

Ulrich, N., Schüürmann, G., Brack, W., 2011. Linear solvation energy relationships as classifiers in non-target analysis—a capillary liquid chromatography approach. J. Chromatogr. A 1218, 8192–8196.

Vermeirssen, E.L.M., Hollender, J., Bramaz, N., van der Voet, J., Escher, B.I., 2010. Linking toxicity in algal and bacterial assays with chemical analysis in passive samplers deployed in 21 treated sewage effluents. Environ. Toxicol. Chem. 29, 2575–2582.

Wold, S., 1978. Cross-validatory estimation of number of components in factor and principal components models. Technometrics 20, 397–405.

Wold, S., Albano, C., Dunn, W.J., Esbensen, K., Hellberg, S., Johansson, E., Lindberg, W., Sjostrom, M., 1984. Modeling data tables by principal components and PLS – class patterns and quantitative predictive relations. Analusis 12, 477–485.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 58, 109–130.

Wolf, S., Schmidt, S., Müller-Hannemann, M., Neumann, S., 2010. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11.