LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-738170

Synthesis route attribution of sulfur mustard by multivariate data analysis of chemical signatures

K. H. Holmgren, S. Hok, R. Magnusson, A. Larsson, C. Astot, C. Koester, D. Mew, A. K. Vu, A. Alcaraz, A. M. Williams, R. Norlin, D. Wiktelium

September 7, 2017

Talanta

Disclaimer

Synthesis route attribution of sulfur mustard by multivariate data analysis of chemical signatures

Karin Höjer Holmgren[a], Saphon Hok[b], Roger Magnusson[a], Andreas Larsson[a], Crister Åstot[a], Carolyn Koester[b], Daniel Mew[b], Alexander K. Vu[b], Armando Alcaraz[b], Audrey M. Williams[b], Rikard Norlin[a], Daniel Wiktelius[a,*]

[a] Swedish Defence Research Agency, FOI, CBRN Defence and Security, Cementvägen 20, 90182 Umeå, Sweden
[b] Forensic Science Center, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California, 94550, United States
* Corresponding author at: Swedish Defence Research Agency, FOI, CBRN Defence and Security, Cementvägen 20, 90182 Umeå, Sweden. E-mail address: daniel.wiktelius@foi.se (Daniel Wiktelius)

ABSTRACT
A multivariate model was developed to attribute samples to a synthetic method used in the production of sulfur mustard (HD). Eleven synthetic methods were used to produce 66 samples for model construction. Three chemists working in both participating laboratories took part in the production, with the aim to introduce variability while reducing the influence of laboratory or chemist specific impurities in multivariate analysis. A gas chromatographic/mass spectrometric data set of peak areas for 103 compounds was subjected to orthogonal partial least squares - discriminant analysis to extract chemical attribution signature profiles and to construct multivariate models for classification of samples. For one- and two-step routes, model quality allowed the classification of an external test set (16/16 samples) according to synthesis conditions in the reaction yielding sulfur mustard. Classification of samples according to first-step methodology was considerably more difficult, given the high purity and uniform quality of the intermediate thiodiglycol produced in the study. Model performance in classification of aged samples was also investigated.

Keywords:
Chemical forensics
Multivariate data analysis
Sulfur mustard
Synthesis method attribution

Abbreviations:
HD = Distilled sulfur mustard
CWA = Chemical warfare agent
CWC = Chemical weapons convention
CAS = Chemical attribution signature
TDG = Thiodiglycol
OPLS-DA = Orthogonal partial least squares – discriminant analysis
PCA = Principal component analysis
PC = Principal component

This work is supported by the Swedish Civil Contingencies Agency under Contract 2014-5170 and the Department of Homeland Security, Science and Technology Directorate - Homeland Security Advanced Research Projects Agency - Chemical and Biological Division under Contract # HSHQPM-16-X-00102.

1 Introduction

Sulfur mustard (2,2'-dichloroethyl sulfide or bis(2-chloroethylsulfide), HD) is a vesicant commonly employed as a chemical warfare agent (CWA). Large amounts of HD were produced and stockpiled in several countries before the ratification of the Chemical Weapons Convention (CWC) in 1997 [1]. The convention bans the production, storage and use of CWAs, and the destruction of declared stockpiles is still an ongoing endeavor. HD is a relevant threat agent because its synthesis is a relatively simple process, and there are several synthetic routes that can be used for its production in large scale [4-6]. While access to the required precursors and reagents is limited by export control and the obligations of the State Parties of the CWC, many of these materials have widespread industrial applications, making barriers of trade less effective in controlling their use.

The medical effects of HD exposure include blistering of skin and eye damage, and have a delayed onset. At higher doses, severe burns occur for which there is no effective treatment. Inhalation of HD causes lung damage that may induce chronic lung disease. After extensive exposure, HD inflicts lethal injuries [2].

HD was deployed on a large scale in the battlefields of the First World War, and has since then been used on several occasions [2]. Evidence of its use in the Syrian Civil War has recently been put forward by the Fact Finding Missions of the Organization for the Prohibition of Chemical Weapons [3]. The use of CWAs in modern conflicts calls for forensic analytical tools to enable sourcing of chemical threat agents. For national legal systems and courts of law, the absence of validated forensic methods is equally troublesome when trying to convict terrorists accused of perpetrating chemical attacks. Chemical attribution methods would potentially enable linkage between different samples and retrospective determination of the production method for chemical threat agents.

Chemical attribution signatures (CASs) can be extracted by analysis of a material. These are "the impurities, byproducts, and degradation products that can be used to readily differentiate samples of a material and associate them with a particular source of reagents, type of equipment, synthetic pathway or reaction condition", as defined by Mazzitelli et al.[7]. Stemming from forensic studies on narcotics [8], this concept has been applied in the area of CWAs and toxins in conjunction with statistical models for attribution based on various types of analytical data [7, 9-18].

Multivariate data analysis tools were used to build a calibration model for the retrospective determination of HD production methods in this study. A relevant range of authentic samples of CWA material to investigate is, by nature, unavailable. For this reason, the preparation of representative sample sets via multiple synthetic methods was carried out by several chemists in two different laboratories (see author affiliations). Working in collaboration, we also attempted to reduce the influence of external markers not relating to HD synthesis in our data sets. Experience from related work on Russian VX proved that even a limited set of samples could produce a valid predictive model using gas chromatography/mass spectrometry (GC/MS) data [19]. We wanted to apply the basic methodology of the previous study to attribute a CWA to a synthetic method, where the set of samples were more homogeneous, and the concentration of CAS useful for attribution could be expected to be smaller. To the best of our knowledge, this is the first study of production route attribution for HD.

2 Experimental

2.1 Safety and legal matters

HD is a highly toxic compound, and most reagents needed for its production are hazardous materials. Rigorous protective measures must be in place when preparing HD to avoid serious, possibly fatal, injuries. Appropriate decontamination of equipment is required to prevent exposure of personnel and the environment. The synthesis of CWA, in general, is controlled by national and international legislation, and unless correct authorization is in place, the work will be considered criminal.

2.2 Synthesis

HD was prepared using 11 synthesis routes (R1-11, Fig. 1). All syntheses were performed by three chemists producing two replicates per route, and at least two replicates of each route were produced in each laboratory. Thus, for each route, a total of six replicates were made and the total number of preparations were 66. Known and established synthetic methods were tested and slightly adjusted so that crude HD of the highest possible quality was obtained, without purification [4-6, 20]. A detailed protocol for each route was devised to ensure that a homogenous sample set was produced by both laboratories. Among specified parameters were starting materials, reagents, solvents, reaction time, temperature, scale, concentration, agitation method and isolation of the crude material (i.e. evaporation, phase separation or filtration). All chemicals were obtained from Sigma-Aldrich in "puriss p.a." or "reagent plus" quality, where possible, to reduce introduction of impurities specific to these materials. Reactions were performed at a scale yielding 0.3-1 g of product. Nine of the routes (R1-9) were two-step syntheses involving the key intermediate thiodiglycol (TDG, Fig. 1). For R1-3 and R7-9, the first step of the synthesis involved double addition of hydroxyethylene to sulfur, either by reaction of sodium sulfide with 2-chloroethanol, or ring opening of ethylene oxide with hydrogen sulfide. For R4-6, thioethanol was reacted with ethylene oxide to give TDG. Chlorination of TDG with hydrochloric acid (R1, 4, 7), phosphorous trichloride (R2, 5, 8), or thionyl chloride (R3, 6, 9) in the second step produced HD. Route 10 relied on addition of sulfur and chloride to ethylene by sulfur monochloride to give HD in a single step. Sulfur dichloride generated *in situ* was used to make the same transformation in R11. The complete details of these syntheses are not given here to prevent proliferation. For reaction details please contact the authors for guidance and references.

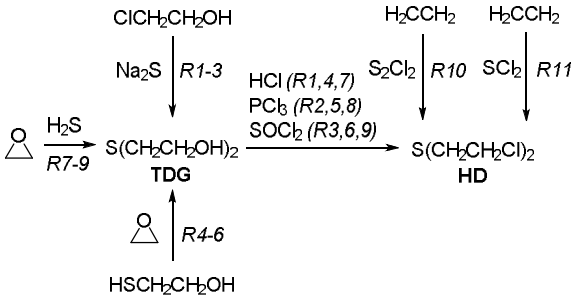

Fig. 1. Preparative synthesis routes (*R1-11*) of HD discussed in this work.

2.3 Aging of HD preparations

The main sampling time point was one week after synthesis (t = 7 d). 44 out of the 66 HD preparations were also sampled directly after synthesis (t = 0) and again after six months (t = 6 m) to evaluate the effect of aging on the CASs. The preparations were stored at approx. 21 °C, in screw cap vials, in the dark.

2.4 Sample preparation

An aliquot of 5 µL was withdrawn from the HD preparations and dissolved in dichloromethane (DCM, 6.35 mL) to make a 1 mg/mL stock solution. Aliquots thereof were used for analysis, either directly or after derivatization or dilution to 0.1 mg/mL. Sample volume was in all cases 200 µL. Samples were stored at −20 °C, in the dark, until analyzed. Dibenzothiophene (20 µL of a 10 µg/mL DCM solution) was used as an internal standard (IS) in all analyses for semiquantification of compounds. TMS-derivatized samples (0.1 mg/mL) were prepared by treatment with BSTFA (20 µL BSTFA:TMCS 99:1 (Thermo Scientific, Bellefonte, PA, USA), 20 µL of 1 mg/mL HD stock solution, 140 µL DCM, 20 µL IS solution, and heating for 60 minutes at 60 °C before analysis). Samples of TDG were prepared in acetronitrile at 10 mg/mL for direct analysis, and aliquots thereof were diluted to 1 mg/mL and TMS-derivatized as described above. Stock solutions (representing 22 of the 66 syntheses) were

also exchanged between laboratories; samples were shipped in insulated containers, with frozen gel packs to maintain a cool temperature, and time in transit was approximately one week.

2.5 Analysis – compound identification and data processing

2.5.1 GC/MS Methods

Samples were analyzed in electron ionization (EI) mode on an Agilent 7890A GC/5975C MSD equipped with a HP-5MS column (30 m long, 0.25 mm i.d., 0.25 µm film thickness, Agilent Technologies). The sample (1 µL) was introduced by splitless injection at 250 ºC, with helium as carrier gas at a constant flow of 1.0 mL/min. The GC oven was kept at 40 ºC for 1 min, followed by a 10 ºC/min increase to 300 ºC and held at 300 ºC for 5 min (32-min runtime). A solvent delay of 3.30 or 6.60 min was used for underivatized or derivatized samples, respectively. The MSD was set to scan 29-500 m/z at a speed of 3.08 scan/sec. The temperature of the transfer line, ion source and quadrupole were set at 280 ºC, 230 ºC and 150 ºC, respectively. Samples were analyzed also in chemical ionization (CI) mode on an Agilent 6890 GC/5973 MSD. Isobutane was used as CI gas, in positive mode, with a scan range of 60-550 m/z. The ion source was set at 300 ºC. All other settings were the same as in EI mode. Underivatized samples of TDG were analyzed in EI mode on an Agilent 5890 GC/5972 MSD with split injection (ratio 50:1), ion source temperature of 173 °C, scan range 29-450 m/z at a speed of 1.81 scan/sec. The GC oven was kept at 60 °C for 3 min, followed by a 16 °C/min increase to 325 °C giving a runtime of 19.5 min. A solvent delay of 3.0 min was used. All other conditions were the same as mentioned above. Purity analysis of HD (0.1 mg/mL) and TDG (10 mg/mL) was made using the appropriate GC/MS method, as described above. Purity was assessed by integration of the total ion chromatogram (Chemstation RTE integrator, 5 data points sampling, minimum area count 1) and expressed as a percentage of the target analyte area relative to the sum of the areas of all of the chromatographic peaks.

2.5.2 Quality control (QC) of chromatography/mass spectrometry and retention index calibration

A QC-sample was analyzed at the beginning of each sample sequence and after approximately every ten samples. The QC-sample contained eight compounds with different properties (to monitor performance of the GC/MS system) and a C_8-C_{28} hydrocarbon series (to perform a retention index (RI) calibration). The peak area of the internal standard was used as an additional indicator of instrument performance and to verify sample concentration. Blank samples (DCM or DCM+BSTFA) were run after every fourth sample.

2.5.3 Compound identification

Identification of compounds was done by MS spectra matching (match factor ≥ 900) against mass spectra libraries (NIST 11 Version 2.0g and OPCW Central Analytical Database (OCADv.19_2017)). Mass spectra for compounds not found in libraries were manually interpreted with the help of information on molecular weight (MS-CI), isotope ratios, retention index, synthesis conditions and published data of related compounds [21-27]. Considering the number of compounds present at low concentration, a selection was made based on peak area and distribution in the sample set. Tentative identifications were possible in many cases, but a known or tentative structure was not required for inclusion of a compound in the data set. A table of 103 included compounds and a summary of their MS data, is available in the supporting information. HD itself was not considered for inclusion since its intensity in the chromatograms was well above the dynamic range of the detector in most samples. Some compounds could be dismissed after assertive identification, as they clearly could not be considered true CAS for route attribution and their source was known (e.g. tetrachloroethylene (solvent impurity), butylated hydroxytoluene (solvent stabilizer), 1-tetradecene and two other higher order alkenes (contaminants in the laboratory environment), and bis(trimethylsilyl)ether and tris(trimethylsilyl) phosphate (artifacts of injector contamination)).

2.5.4 Peak table construction

GC/MS-EI data was evaluated using AMDIS (Version 2.70 2011, NIST) and processed to generate a peak table of areas in a sample (observation) – compound (variable) matrix. A target library was created in which MS spectra and RI were used to assign compounds to peaks in the chromatograms. The compounds included in the peak table were identified by their mass spectrum (match factor set to 70) and RI (a RI-offset of ± 20 units was tolerated). A peak area of at least 30000 and a minimal signal to noise ratio of 20 were required for inclusion in the data set. The peak areas were normalized by division by the total area in each chromatogram. From the master peak table, subsets were selected for assembly of the multivariate models presented in the results section.

2.6 Statistics

The software SIMCA (ver 14. Sartorius Stedim Biotech) was used for multivariate statistic evaluation of the data. Principal component analysis (PCA) was first used to overview the class-specific information within the data-set, followed by classification of synthesis routes with orthogonal partial least squares discriminant analysis (OPLS-DA) [28]. Each individual value in the normalized peak table was transformed with log (x + 0.00001). Different types of scaling were evaluated during preliminary modeling; it was found that mean centering of variables, without further scaling, was optimal in this case.

3 Results and discussion

3.1 Synthesis

Synthesis commenced with preparation of TDG for two-step routes (R1-9, Fig. 1). All syntheses produced TDG as a clear viscous oil in purities ranging from 89 – 99.5% in good to quantitative yield. All routes produced HD as a clear to dark brown oil of low viscosity. Conversion in the second step of Routes 1-9 was assessed by GC/MS analysis of TMS-derivatized samples of the HD-preparations, where the absence of bis(TMS)TDG in all cases indicated complete consumption of TDG. The crude weight yield was above theoretical for all preparations which indicated mass balance for the transformation and reagents used. Crude product purity ranged from 36.5-99.8%; the median purity of all replicates was 90.5%. Small amounts of solid material were noted in preparations from Routes 3, 6, 9 and 10.

3.2 Analysis of TDG preparations

Direct analysis of the TDG preparations showed no compounds of interest for further discussion, except 2,2'-dithiodiethanol (data not shown). This compound was found in higher concentrations in samples produced by Routes 7-9. No other compound was present in levels above 1% of the total area of the TIC. Analysis of derivatized samples of TDG preparations gave a similar result, with the addition of some longer chain TDG analogues with further glycol/thioethanol increments, which were detected in low levels (< 2% total area TIC).

3.3 Analysis of underivatized samples of HD preparations

3.3.1 Composition of samples and compound levels

An overview of the composition of selected samples is given in Fig. 2. In general the intensity of peaks were low, thus a sample concentration of 1 mg/mL was required in analysis of the HD-preparations. R10 and 11 contained the largest variety of compounds and the highest levels of impurities. The highest concentration of CAS candidates was found in samples of R11. Routes 2, 3, and 5-9 were intermediate with respect to the number and concentration of impurities. Samples of R1 and R4 presented few compounds at levels of 3 µg/mL or below. The low intensity of candidate CASs was taken as an indication that the carefully designed synthesis protocols had delivered samples that would challenge our analytical methods, and make statistical analysis demanding.

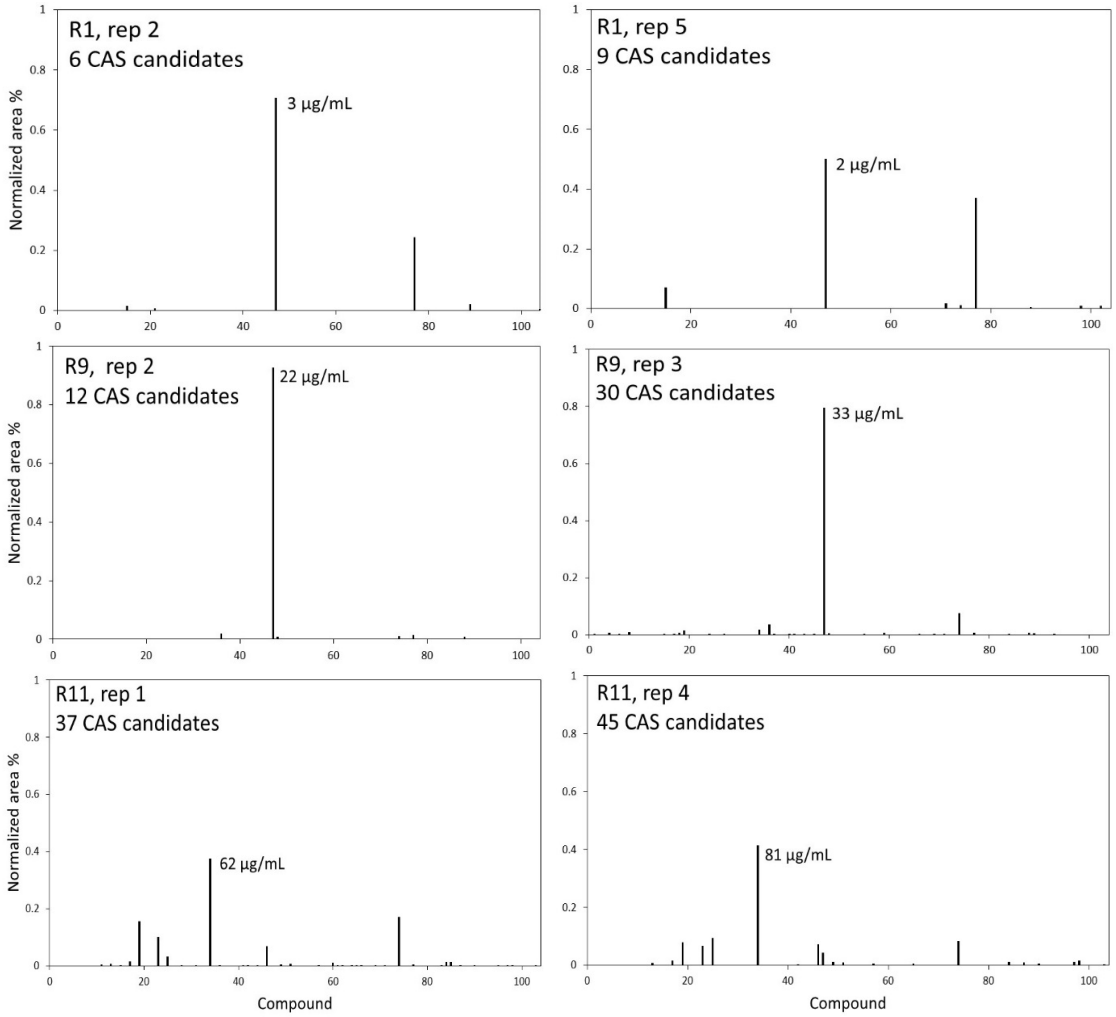

Fig. 2. Impurity profiles in two replicates of Routes 1, 9 and 11. Each bar represents the normalized area % of a compound, and its concentration in 1 mg/mL samples, approximated by semiquantification using IS. The number of CAS candidates in each sample is given on each chart.

3.3.2 Observations of compound distribution

Bis(2-chloroethyl)disulfide and sesquimustard (Fig. 3) were ubiquitous compounds, but the amounts varied between routes and replicates. 1,4-dithiane (Fig. 3) was also found at low levels (< 0.2 µg/mL) in almost all samples.

Manual inspection of chromatograms and the peak table revealed few markers pertinent to TDG synthesis methodology for R1-9 (Fig. 1). When grouping samples by first step, samples of R1-3 and R4-6 were characterized by the absence of such compounds, indicating that the intermediate TDG was of high, and similar, quality in these routes. Higher content of bis(2-chloroethyl)disulfide was recognized as a feature for R7-9 after chlorination, which paralleled findings already at the TDG stage (section 3.2).

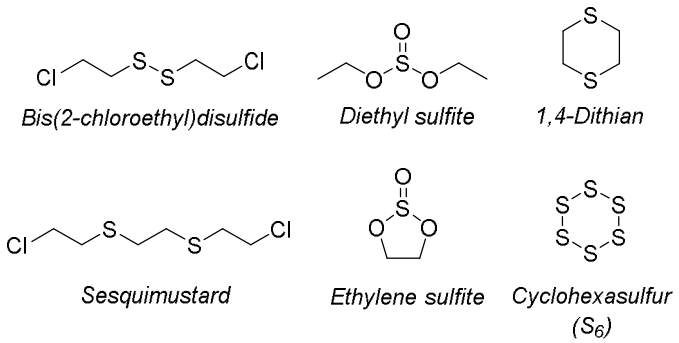

Fig. 3. Exemplar compounds found by GC/MS analysis of HD preparations.

Grouping samples of two-step routes (R1-9, Fig. 1) by the chlorination method revealed a compound distribution that was expected to be useful for classification of samples with the help of byproducts generated in the second step. Routes 1, 4 and 7 (chlorination by HCl) were characterized by scarcity and low intensity of CAS candidates, none of which were unique to these routes. While R1 and R4 were indistinguishable by manual inspection, R7 could be recognized by the presence of bis(2-chloroethyl)disulfide in higher levels, as discussed above. Preparations involving PCl_3 for chlorination (R2, 5, 8) contained traces of reaction intermediates in the form of phosphorous containing byproducts. Chlorination by $SOCl_2$ (R3, 6, 9) gave HD preparations with more impurities than those made with other chlorination methods. Compounds related to $SOCl_2$ were seen in samples of these routes, for example ethylene sulfite and diethyl sulfite (Fig. 3). Routes 10 and 11 have dissimilar impurity profiles to the other routes, which was expected considering the different reaction conditions for HD synthesis by these methods. The number of CAS candidates was large, typical compounds for both routes were elemental sulfur in the form of S_6 (Fig. 3), and other polysulfides. Samples of R10 and R11 also contained several HD analogues. Samples of Route 11 had several unique compounds, many of them were polychloroethyl derivatives.

In summary, manual inspection of the data-set indicated that route classification might be possible for groups of samples sharing the same reaction conditions in the HD-forming step. However, it was not apparent if there were compounds carrying information from the TDG preparations into the final samples of HD for two-step routes.

3.4 Analysis of TMS-derivatized samples of HD preparations

While the underivatized sample set of R1-9 contained no obvious CAS candidates attributed to the protocol for TDG synthesis, we hoped that some of the missing information would be found in the derivatized samples. Polar byproducts may well arise in synthesis of TDG, and derivatization of TDG samples did result in some presumptive CAS precursors (section 3.2). Disappointingly, these were not found in TMS-derivatized samples of the final HD preparations, nor were their anticipated chlorinated end-products. Polar CAS candidates such as bis(trimethylsilyl)phosphite and bis(trimethylsilyl)sulfite were indicative of the chlorinating reagent used for R2, 5, and 8 and R3, 6, and 9 respectively; however, this information was already conveyed in the underivatized sample set (section 3.3.2). HD related bis(chloroalkyl)sulfides amenable to GC analysis without derivatization were the expected impurities in R10 and 11, because no polar intermediates or reagents are involved in these syntheses. Indeed, analysis of derivatized samples from these routes gave no additional information. Taken together, we judged the information earned from derivatized samples to be of little value in multivariate analysis, and we chose not to include any derivatized compounds in the peak table.

3.5 Statistical model construction and validation with external test sets

3.5.1 Principal component analysis

Normalized peak areas of 103 compounds in 66 underivatized samples (t = 7 d) from the eleven synthetic routes was subjected to a principal component analysis (PCA, Fig. 4). The PCA model required 8 significant principal components (PCs) to explain 77% of the data variance ($R^2X = 0.77$, $Q^2X = 0.49$). Adding further principal components to cover more of the variation does not improve the predictive ability of the model. This indicates that much of the variance not included in the model is in fact chemical noice of substances not carrying any information on route attribution.. The score plot of the PCA confirmed manual observations of compound distribution among the samples (section 3.3.2), grouping samples with a similar impurity profile stemming from the HD forming step. The complete data set was considered suitable for subsequent OPLS-DA since there were no evident outliers.

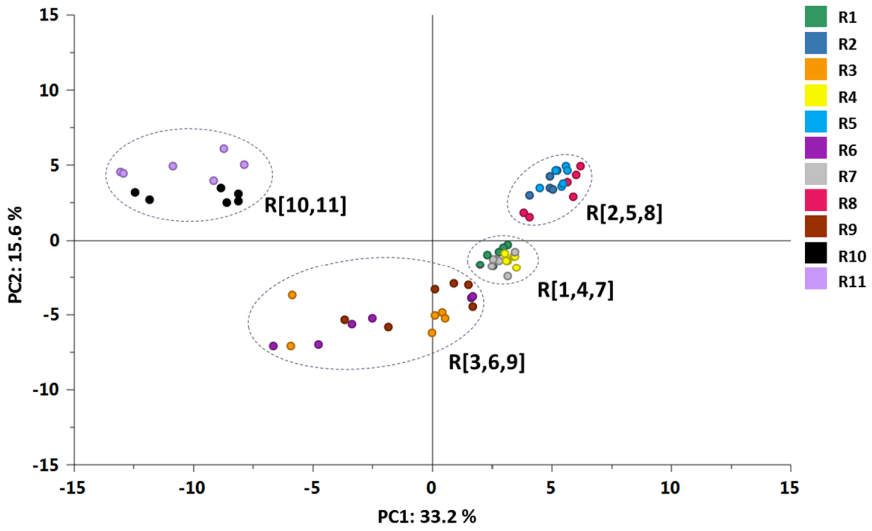

Fig. 4. Score plot representation of the result (PC1 vs PC2) from a PCA of data from all samples. Dashed ovals have been added to visualize our interpretation of the result. Details on Routes 1-11 are given in Fig. 1.

3.5.2 Construction of classification models by OPLS-DA

We had envisioned a single multivariate model that could classify samples from all eleven routes; however, when including all classes, the result was not of sufficient quality (103 compounds, 66 samples, 4 predictive and 3 orthogonal (4+3) PCs, $R^2X = 0.71$, $R^2Y = 0.40$, $Q^2X = 0.28$). The low quality of the model might be explained by inadequate resolution of the CAS profile component conveying information on TDG synthesis method, in particular for R1-6. This result suggested further replicates are needed to investigate if statistical resolution of CAS profiles from all routes is possible.

Models were then constructed where classes included samples grouped according to synthesis method in each step. Normalized peak data from the five groups R[1,4,7], R[2,5,8], R[3,6,9], R10, and R11 were first studied to extract attribution information emanating from the HD forming step. The model quality was sufficient (5 classes, 103 compounds, 66 samples, 4+3 PCs, $R^2X = 0.71$, $R^2Y = 0.95$, $Q^2Y = 0.89$, Fig. 5) to resolve the groups. Coefficient plots were used to extract compounds that were positively correlated to the groups of samples. Overall, the recognition pattern was complex and the presence and relative concentrations of many compounds contributed to the CAS profiles. Many of the observations made by manual inspection of the data set (section 3.3.2) were mirrored in the coefficient plots. Samples of HD preparations with HCl as chlorinating reagent (R[1,4,7]) had few significant CASs; among them were bis(2-chloroethyl)disulfide and sesquimustard. Phosphorous containing adducts were attributed to the synthesis conditions of R[2,5,8] by the model. Ethylene sulfite and diethyl sulfite were recognized as CAS for R[3,6,9]. Dissimilarities in the CAS profiles of

R10 and 11 that were difficult to assess in the normalized peak table became observable with the help of coefficient plots. Relative differences in levels of CASs present in samples of both routes seemed to constitute the basis for recognition. HD analogues that were positively correlated to R10 were also found in samples of R11, and had sporadic occurrence in samples from other routes as well. Other HD-related compounds were significant CASs for R11. The influence on model quality of the seven omitted compounds specified in section 2.5.3 was examined by including them in model construction; they were found to have no relevance. With these exceptions, no further exclusions were made. Considering that many compounds were not identified, we decided that further selection could not be done without bias.

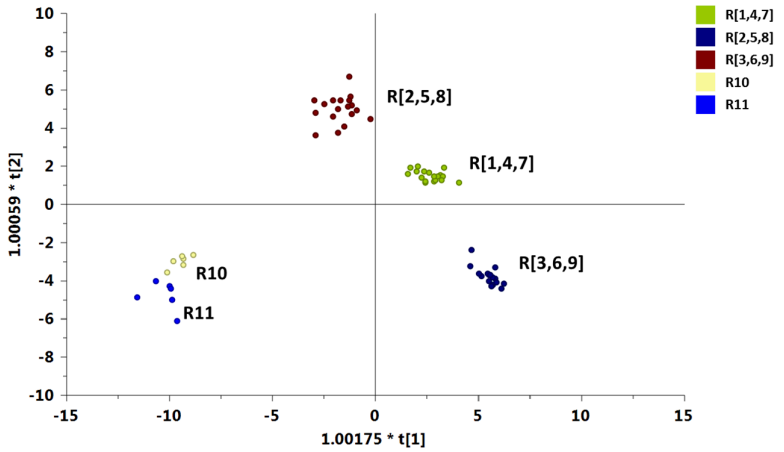

Fig. 5. Score plot representation of the result of OPLS-DA (PC1 vs PC2) of CAS profiles from samples grouped according to synthesis conditions in the HD forming step. Details on Routes 1-11 are given in Fig. 1.

OPLS-DA of groups of samples with the same synthesis method for preparation of TDG (R[1,2,3], R[4,5,6], and R[7,8,9] resulted in a model of intermediate quality (3 classes, 69 compounds, 53 samples, 2+3 PCs, $R^2X = 0.56$, $R^2Y = 0.75$, $Q^2 = 0.53$). This result was expected considering the lack of observations of compounds related to TDG formation discussed in sections 3.3.2 and 3.4. The result from cross-validation showed that the predicted values for samples belonging to each class were poorly centered around 1 (0.5-1.4, 0.4-1.3, and 0.4-1.2 for R[1,2,3], R[4,5,6], and R[7,8,9,] respectively). Furthermore, these ranges overlapped predicted value ranges for samples not belonging to the class, suggesting that the model could not identify CAS profiles. The CAS profile of R[4,5,6] seemed most difficult to recognize in the data set, whereas that of R[7,8,9] was closest to allow classification. These results confirm that the stringent protocols for TDG preparation yielded product in high and uniform purity in the synthesis replicates. Chemical analysis techniques of higher sensitivity may allow identification of relevant CASs to attribute samples to TDG synthesis conditions. Since the lower quality of the model did not allow the removal of a test set for validation, we did not pursue further statistical analysis of the data.

3.5.3 Validation of multivariate model for classification of samples according to synthesis method in the HD-forming step

An external test set consisting of 16 samples was arbitrarily selected from the five groups R[1,4,7], R[2,5,8], [R3,6,9], R10, and R11, so that at least four samples remained in each. Samples generated in both laboratories were present in the external test set. OPLS-DA model quality prevailed with 24% of the sample input removed (5 classes, 93 compounds, 50 samples, 4+3 PCs, $R^2X = 0.72$, $R^2Y = 0.94$, $Q^2X = 0.82$), and was used for classification of the samples in the test set. The model correctly assigned synthesis method in the HD-forming step for all 16 samples of the test set, with predicted values close to one (Table 1). There were no misclassifications; predicted values were close to zero for samples not belonging to a particular class.

Table 1. Prediction results of an external sample test-set by an OPLS-DA attribution model for synthetic method in the HD-forming step (Section 3.5.3). Details on routes are found in Fig. 1, sample grouping is described in Section 3.5.2.

Test sample (Route)	Predicted values				
	R[1,4,7]	R[2,5,8]	R[3,6,9]	R10	R11
1 (R1)	**0.80**	0.08	0.13	–0.04	0.03
2 (R2)	0.01	**1.02**	–0.02	0.02	0.02
3 (R2)	0.09	**0.98**	–0.05	–0.03	0.01
4 (R3)	0.00	–0.06	**1.23**	–0.27	0.09
5 (R4)	**1.06**	–0.01	–0.02	–0.02	–0.01
6 (R5)	–0.05	**1.06**	0.01	–0.01	–0.01
7 (R6)	–0.04	–0.02	**1.04**	0.06	–0.04
8 (R6)	0.00	0.01	**1.05**	–0.04	–0.02
9 (R7)	**1.10**	–0.11	0.01	–0.02	0.02
10 (R7)	**0.93**	0.01	0.00	0.01	0.05
11 (R8)	–0.13	**1.16**	–0.01	0.03	–0.05
12 (R9)	–0.11	–0.01	**1.02**	0.12	–0.02
13 (R10)	–0.10	0.06	0.08	**1.00**	–0.04
14 (R10)	–0.03	0.02	0.04	**0.81**	0.16
15 (R11)	–0.06	–0.09	0.06	0.09	**1.00**
16 (R11)	0.12	–0.05	–0.03	0.06	**0.90**

3.5.4 Model performance on data from samples of aged HD preparations

The performance of the original model built on data from all 66 HD preparations sampled one week after synthesis (t = 7 d) was evaluated by classification of samples collected at other time points. These were samples drawn directly after synthesis (t = 0), and upon aging of the preparations for 6 months (t = 6 m). The t = 0 samples had highly similar impurity profiles as those used to build the model, and indeed the model could correctly classify all t = 0 samples. We could conclude that the CAS profiles do not change markedly in a timeframe of seven days. Classification of the t = 6 m samples was more difficult with a misclassification rate of 33%. Aged samples of R[1,4,7], R10 and R11 were predicted correctly while many of R[2,5,8] and R[3,6,9] were misclassified. This result indicates a change in CAS profiles over time for some routes. In particular, samples of R[2,5,8] seemed to lose some of the CAS information, becoming more similar to R[1,4,7] over time. Crystallization of solid material may explain loss of some CAS. The CAS profiles of R[3,6,9] became more similar to those of R10 and 11, suggesting an enrichment in the number and intensity of CASs. These findings could be confirmed by comparing the normalized peak tables of the two sampling time points and by PCA of the combined data from t = 7 d and t = 6 m samples (11 compounds, 108 samples, 7 PCs, $R^2X = 0.67$, $Q^2 = 0.43$, Fig. 6). Importantly, chemical markers of aging were not the same for the various routes. The results indicate that the time elapsed from preparation until sampling is important to consider in synthesis route attribution of HD. They also support the notion that inclusion of many compounds in multivariate analysis relying on CAS profiles is important, because rich profiles could be less susceptible to alterations in sample composition. Further studies are needed to collect data from aged

samples to fully assess the stability of CAS profiles over time, and to build models that encompass such variations. The chemistry of HD upon aging has been studied previously [21, 25, 26].

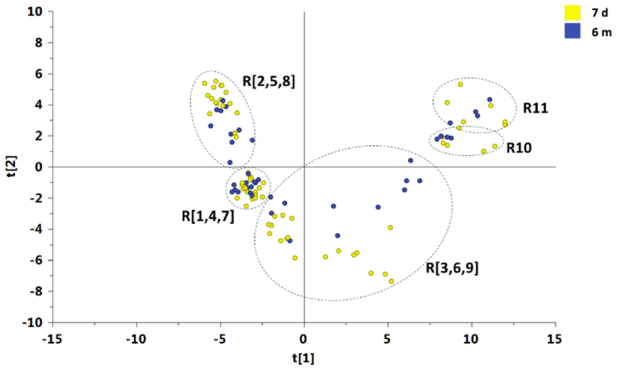

Fig. 6. Score plot representation of the result (PC1 vs PC2) from a PCA of data from samples of HD preparations aged one week (yellow) and 6 months (blue). Dashed ovals have been added to visualize our interpretation of the result.

4 Conclusion

A multivariate model for attribution of samples of HD preparations to the method used in the HD forming step was developed. The model encompassed the variation in CAS profiles obtained by GC/MS analysis of samples of HD preparations made in two laboratories by three chemists. While the classification of samples from synthesis routes involving five different conditions in the HD forming step could be done without ambiguities, there was not enough resolution in the data to assign samples to the methodology used for preparing TDG. This result probably reflects our ambition to create a high purity sample set by stringent synthesis protocols. Chemical analysis techniques of higher sensitivity should be investigated to obtain CAS profiles indicative of TDG synthesis method. The results of multivariate analysis in this study were confirmed by manual inspection of the data set, and correlated well with our understanding of sample composition and the reactions involved in the syntheses. We chose an approach where essentially all observed impurities were considered potential CAS, a necessity with respect to the number of unidentified compounds. While some compounds stood out as CASs of higher significance, the overall pattern in CAS profiles was unquestionably most important for synthesis route recognition. This was emphasized by challenging the model with samples of aged preparations, where alterations in CAS profiles over time made classification of samples to some routes more difficult. Further work is required to incorporate changes in CAS profiles from time of synthesis until sampling. Standard GC/MS techniques were used for analysis and commercially available software for data processing. Thus, our procedures can easily be applied and further developed by other designated laboratories in the field of CWA analysis.

Acknowledgment

The authors would like to thank the Swedish Civil Contingencies Agency [grant number 2014-5170], Sweden, the Department of Homeland Security, Science and Technology Directorate, Chemical Biological Division [grant number HSHQPM-16-X-00102], United States, and for financial support of the project. Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.

Disclaimer

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-JRNL-738170

Appendix A. Supporting information

List of compounds, with summary of MS data, example chromatograms.

Complete GC/MS data, tentative structure assignments and peak table is available on request from the authors.

References

[1] Organization for the Prohibition of Chemical Weapons, Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction, https://www.opcw.org/chemical-weapons-convention/ (accessed 17.05.30).
[2] K. Ghabili, P. Agutter, M. Ghanei, K. Ansarin, Y. Panahi, M. Shoja, Sulfur mustard toxicity: History, chemistry, pharmacokinetics, and pharmacodynamics, Crit. Rev. Toxicol. 41 (5) (2011) 384-403.
[3] K.-m. Ban, Letter dated 24 August 2016 from the Secretary-General addressed to the President of the Security Council, 2016. http://www.un.org/en/sc/documents/sgreports/2016.shtml (accessed 17.05.30).
[4] S.M. Somani, Toxicokinetics and toxicodynamics of mustard, in: S.M. Somani (Ed.), Chemical warfare agents, Academic Press, San Diego, 1992, pp. 13-43.
[5] Institute of Medicine (U.S.). Committee to Survey the Health Effects of Mustard Gas and Lewisite, Veterans at Risk : The Health Effects of Mustard gas and Lewisite, National Academy Press, Washington, 1993.
[6] R. Malhotra, K. Ganesan, K. Sugendran, R. Swamy, Chemistry and toxicology of sulphur mustard - A review, Def. Sci. J. 49 (2) (1999) 97-116.
[7] C. Mazzitelli, M. Re, M. Reaves, C. Acevedo, S. Straight, J. Chipuk, A systematic method for the targeted discovery of chemical attribution signatures: application to isopropyl bicyclophosphate production, Anal. Chem. 84 (15) (2012) 6661-6671.
[8] M. Morelato, A. Beavis, M. Tahtouh, O. Ribaux, P. Kirkbride, C. Roux, The use of forensic case data in intelligence-led policing: The example of drug profiling, Forensic Sci. Int. 226 (1-3) (2013) 1-9.
[9] J. Hoggard, J. Wahl, R. Synovec, G. Mong, C. Fraga, Impurity profiling of a chemical weapon precursor for possible forensic signatures by comprehensive two-dimensional gas chromatography/mass spectrometry and chemometrics, Anal. Chem. 82(2) (2010) 689-698.

[10] C. Fraga, B. Clowers, R. Moore, E. Zink, Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics, Anal. Chem. 82 (10) (2010) 4165-4173.
[11] H. Colburn, D. Wunschel, H. Kreuzer, J. Moran, K. Antolick, A. Melville, Analysis of carbohydrate and fatty acid marker abundance in ricin toxin preparations for forensic information, Anal. Chem. 82 (14) (2010) 6040-6047.
[12] C. Fraga, J. Wahl, S. Nunez, Profiling of volatile impurities in tetramethylenedisulfotetramine (TETS) for synthetic-route determination, Forensic Sci. Int. 210 (1-3) (2011) 164-169.
[13] C. Fraga, G. Acosta, M. Crenshaw, K. Wallace, G. Mong, H. Colburn, Impurity profiling to match a nerve agent to its precursor source for chemical forensics applications, Anal. Chem. 83 (24) (2011) 9564-9572.
[14] C. Fraga, L. Sego, J. Hoggard, G. Acosta, E. Viglino, J. Wahl, R. Synovec, Preliminary effects of real-world factors on the recovery and exploitation of forensic impurity profiles of a nerve-agent simulant from office media, J. Chromatogr. A 1270 (2012) 269-282.
[15] D. Behringer, D. Smith, V. Katona, A. Lewis, L. Hernon-Kenny, M. Crenshaw, Demonstration of spread-on peel-off consumer products for sampling surfaces contaminated with pesticides and chemical warfare agent signatures, Forensic Sci. Int. 241 (2014) 7-14.
[16] N. Mirjankar, C. Fraga, A. Carman, J. Moran, Source attribution of cyanides using anionic impurity profiling, stable isotope ratios, trace elemental analysis and chemometrics, Anal. Chem. 88 (3) (2016) 1827-1834.
[17] C. Fraga, K. Bronk, B. Dockendorff, A. Heredia-Langner, Organic chemical attribution signatures for the sourcing of a mustard agent and its starting materials, Anal. Chem. 88 (10) (2016) 5406-5413.
[18] E. Strozier, D. Mooney, D. Friedenberg, T. Klupinski, C. Triplett, Use of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection and random forest pattern recognition techniques for classifying chemical threat agents and detecting chemical attribution signatures, Anal. Chem. 88(14) (2016) 7068-7075.
[19] K. Höjer Holmgren, C. A. Valdez, R. Magnusson, A. K. Vu, S. Lindberg, A. M. Williams, A. Alcaraz, C. Åstot, S. Hok, R. Norlin, Part 1: Sourcing russian VX by multivariate statistics of chemical attribution signatures, Talanta (*In this issue*).
[20] E.M. Faber, G.E. Miller, beta-Thiodiglycol, Org. Synth. 12 (1932) 68-69.
[21] N. Munro, S. Talmage, G. Griffin, L. Waters, A. Watson, J. King, V. Hauschild, The sources, fate, and toxicity of chemical warfare agent degradation products, Environ. Health Persp. 107 (12) (1999) 933-974.
[22] L. Sachin, R. Karthikraj, K. Kumar, T. Sony, N. Raju, S. Prabhakar, Mass spectral studies on vinylic degradation products of sulfur mustards under gas chromatography/mass spectrometry conditions, Eur. J. Mass Spectrom. 21 (6) (2015) 791-800.
[23] F. Hsu, M. Winemiller, S. Bae, Mass spectral studies of 1-(2-chloroethoxy)-2-[(2-chloroethyl)thio]ethane and related compounds using gas chromatography-mass spectrometry (GC-MS) and gas chromatography-triple quad mass spectrometry (GC-QQQ), Phosphorus Sulfur 191 (6) (2016) 958-964.
[24] L. Zhou, H. Gu, G. Qu, Y. Zhong, Gas chromatography-mass spectrometric analysis of detoxicated products of mustard by chloramines, Chinese J. Anal. Chem. 25 (9) (1997) 1068-1071.
[25] M. Mazurek, Z. Witkiewicz, S. Popiel, M. Sliwakowski, Capillary gas chromatography-atomic emission spectroscopy-mass spectrometry analysis of sulphur mustard and transformation products in a block recovered from the Baltic Sea, J. Chromatogr. A 919 (1) (2001) 133-145.
[26] S. Hanaoka, K. Nomura, T. Wada, Determination of mustard and lewisite related compounds in abandoned chemical weapons (yellow shells) from sources in China and Japan, J. Chromatogr. A 1101 (1-2) (2006) 268-277.
[27] P.A. D'Agostino, L.R. Provost, Capillary column isobutane chemical ionization mass spectrometry of mustard and related compounds, Biomed. Environ. Mass. 15 (1988) 553-564.
[28] M. Bylesjö, M. Rantalainen, O. Cloarec, J. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemometr. 20 (8-10) (2006) 341-351.

FIGURE AND TABLE CAPTIONS

Fig. 1. Preparative synthesis routes (*R1-11*) of HD discussed in this work.

Fig. 2. Impurity profiles in two replicates of Routes 1, 9 and 11. Each bar represents the normalized area % of a compound, and its concentration in 1 mg/mL samples, approximated by semiquantification using IS. The number of CAS candidates in each sample is given on each chart.

Fig. 3. Exemplar compounds found by GC/MS analysis of HD preparations.

Fig. 4. Score plot representation of the result (PC1 vs PC2) from a PCA of data from all samples. Dashed ovals have been added to visualize our interpretation of the result. Details on Routes 1-11 are given in Fig. 1.

Fig. 5. Score plot representation of the result of OPLS-DA (PC1 vs PC2) of CAS profiles from samples grouped according to synthesis conditions in the HD forming step. Details on Routes 1-11 are given in Fig. 1.

Table 1. Prediction results of an external sample test-set by an OPLS-DA attribution model for synthetic method in the HD-forming step (Section 3.5.3). Details on routes are found in Fig. 1, sample grouping is described in Section 3.5.2.

Fig. 6. Score plot representation of the result (PC1 vs PC2) from a PCA of data from samples of HD preparations aged one week (yellow) and 6 months (blue). Dashed ovals have been added to visualize our interpretation of the result.