Predicting chloroform production from organic precursors

Tom Bond [a, *], Nigel Graham [b]

[a] Department of Civil and Environmental Engineering, University of Surrey, Guildford, GU2 7XH, UK
[b] Department of Civil and Environmental Engineering, Imperial College London, London, SW7 2AZ, UK

ABSTRACT

Quantitative methods which link molecular descriptors for recognized precursors to formation of drinking water disinfection byproducts are scarce. This study aimed to develop a simple mathematical tool for predicting chloroform (trichloromethane) yields resulting from aqueous chlorination of model organic precursors. Experimental chloroform yields from 211 precursors were collated from 22 literature studies from 1977 onwards. Nineteen descriptors, some established and others developed during this study, were used as inputs in a multiple linear regression model. The final model, calibrated using five-way leave-many-out cross-validation, contains three descriptors. Two novel empirical descriptors, which quantify the impact of adjacent substituents on aromatic and enolizable chlorine substitution sites, were the most significant. The model has $r^2 = 0.91$ and a standard error of 8.93% mol/mol. Experimental validation, using 10 previously untested precursors, showed a mean discrepancy of 5.3% mol/mol between experimental and predicted chloroform yields. The model gives insight to the influence that specific functional groups, including hydroxyl, chlorine and carboxyl, have on chloroform formation and the relative contributions made by separate substitution sites in the same molecule. It is anticipated that the detailed approach can be updated and extended as new experimental data emerges, to encompass additional precursors and groups of disinfection byproducts.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

The discovery, in the 1970s, that chloroform and other trihalomethanes are generated from chlorination of natural organic matter during water treatment (Rook, 1974) surprised the scientific community. This breakthrough led to much research into disinfection byproducts and regulations for the four chlorinated and/or brominated trihalomethanes being introduced in the USA by the decade's end. The limit for total trihalomethanes is currently 80 μg L^{-1} in the USA, with five haloacetic acids, another group of halogenated disinfection byproducts, regulated at a total of 60 μg L^{-1}. Total trihalomethanes are also regulated in the EU at 100 μg L^{-1}. Initial concern about the trihalomethanes was based on results of a rodent bioassay which classified chloroform as a suspected human carcinogen (NCI, 1976). Independently, epidemiological studies have shown that long-term consumption of chlorinated drinking water is associated with an enhanced risk of developing bladder cancer, although the underlying reasons remain obscure (Hrudey, 2009). However, analysis of more recent toxicological evidence indicates that neither the trihalomethanes nor the haloacetic acids are plausible bladder carcinogens at typical drinking water concentrations (Hrudey, 2009). Thus, other disinfection byproducts such as the halobenzoquinones (Zhao et al., 2012), various other halogenated aromatic byproducts (Zhang et al., 2008), nitrosamines, haloacetonitriles and haloacetamides (Shah and Mitch, 2012) remain the focus of much current research attention, as these may be more toxicologically-significant.

The trihalomethanes and haloacetic acids are essentially viewed by regulators of drinking water quality as indicators of the total occurrence of chlorination disinfection byproducts (USEPA, 2015). They also remain the most-studied disinfection byproducts, particularly the trihalomethanes. Model compounds have been heavily used to elucidate mechanistic formation pathways and precursor characteristics from the early days of disinfection byproduct research (Bond et al., 2012; Rook, 1977). It was quickly appreciated that meta-substituted aromatic compounds are reactive chloroform precursors. For example, Rook (1977) reported that resorcinol (1,3-dihydroxybenzene) was converted into chloroform at an 85% mol/mol yield during chlorination, whereas yields from its regioisomers, catechol (1,2-dihydroxybenzene) and hydroquinone (1,4-dihydroxybenzene), were far lower at 0.5% and 1.5% mol/mol, respectively (see Fig. S2 for a simplified mechanism for

chloroform production from resorcinol). This variation can be (qualitatively) explained by the two activating hydroxyl groups promoting electrophilic substitution reactions at the ortho- and para-positions of the aromatic ring. However, the presence of additional substituents complicates this pattern, as depending on their identity and position, chloroform yields can either be enhanced or suppressed. Thus, de Laat et al. (1982), reported chloroform yields for pyrogallol (1,2,3-trihydroxybenzene), 4-hydroxycatechol (1,2,4-trihydroxybenzene) and phloroglucinol (1,3,5- trihydroxybenzene) as 0.1, 15.5 and 93% mol/mol, respectively. Certain aliphatic compounds, notably β-dicarbonyl compounds (Boyce and Hornig, 1980; Dickenson et al., 2008), including 3-oxopentanedioic acid (Table 1), also act as reactive trihalomethane precursors. Formation of trihalomethanes from carbonyl compounds can be likened to the haloform reaction, used for the synthetic preparation of trihalomethanes from methyl ketones (Larson and Weber, 1994). Its rate is controlled by the initial enolization of the organic precursor and the mechanism proceeds via electrophilic addition of chlorine at the carbon alpha to the carbonyl group (Fig. S1).

Despite the extensive amount of research effort on this subject over the past ~40 years, quantitative methods to predict disinfection byproduct formation based on molecular descriptors are scarce. Hence, the aim of this study was to develop a simple mathematical method for predicting chloroform yields from model compounds, with the expectation that this will prove a useful screening tool for compounds which have not been tested experimentally. To achieve this, descriptors are required which quantitatively link reactive precursor functionalities to chloroform yields (and ultimately the key pathways leading from one to the other). A secondary aim was that such a mathematical framework would illuminate our knowledge of the characteristics of reactive precursors. Although quantitative structure activity relationship (QSAR) and quantitative structure–property relationship (QSPR) models are widely used in other fields, they have found limited use in disinfection byproduct research, despite having the potential to streamline research efforts (Chen et al., 2015). While not focussed on disinfection byproducts, the paper by Luilo and Cabaniss (2010) is noteworthy as it details a QSPR, validated using literature on 201 organic compounds, for predicting chlorine demand based on eight molecular descriptors. The same authors subsequently developed a model for predicting chloroform formation from organic precursors (Luilo and Cabaniss, 2011b), although this used a smaller subset of 117 model compounds.

2. Methods

2.1. Literature on chloroform formation

Chloroform formation data from 22 studies (Bond et al., 2009, 2014, 2016; Boyce and Hornig, 1980; Boyce and Hornig, 1983; Bull et al., 2006; Chaidou et al., 1999; Chang et al., 2011; de Laat et al., 1982; de Leer and Erkerlens, 1988; Dickenson et al., 2008; Gallard and von Gunten, 2002; Hong et al., 2009; Hureiki et al., 1994; Larson and Rockwell, 1979; Navalon et al., 2008; Norwood et al., 1980; Rook, 1977; Rule et al., 2005; Tawk et al., 2015; Tomita et al., 1981; Westerhoff et al., 2004) spanning the years 1977–2016 were collated and converted into units of % mol/mol where necessary. All studies measured chloroform formation from organic precursors under formation potential conditions, i.e. using excess chlorine. However, since there are no standard conditions for these tests, experimental conditions vary in the literature (Table S1). In this study only data collected at pH 7–8 using an excess of chlorine and contact times over 0.5 h were included. Thus, the modelling results only apply to these conditions, which are

representative of full-scale drinking water chlorination. Median conditions from the studies included were pH = 7, contact time = 24 h, temperature = 20 °C and chlorine dose = 20 mol/mol. One important difference with the chlorination of natural waters is that model compound studies are typically undertaken in the absence of bromide, and thus chloroform is the only one of the four chloro- and/or bromo-trihalomethanes monitored. In contrast, ambient bromide in natural waters leads to formation of varying amounts of brominated trihalomethanes.

For compounds tested in multiple studies, mean values were calculated and are given in Table S2. For example, chloroform yields from the well-studied precursor resorcinol have been reported 12 times, giving a mean value of 81.1% mol/mol and a standard deviation of 17.3 (Table S2). In total there were 69 compounds included with multiple chloroform yields. The mean of the standard deviations for these repeated compounds is 5.1% mol/mol. The final list of 211 precursors used for modelling, together with their chloroform yields, structures and alternative names, is given in Table S2.

2.2. Descriptor selection

Three descriptors used by Luilo and Cabaniss (2011b) to model chloroform formation and another three used by the same authors to model total organic halogen formation (Luilo and Cabaniss, 2011a) were included. Respectively these are the carbonyl index (CI), the difference between the sum of strong electron-donating groups and the sum of carbonyls per carbon in each molecule (EDCORH), the number of 1,3-activated aromatic carbons (OTactC) (chloroform descriptors) and the number of phenolic groups per carbon (ArOH:C), the square root of the number of heteroatoms (sqHeA) and the log of the hydrogen to carbon ratio (log H:C) (total organic halogen descriptors). Hammett and Taft constants account for substituent effects in aromatic and aliphatic compounds, respectively, and have been used widely in the development of QSARs and linear free energy relationships. In this study they were used in a manner following that described by Lee and von Gunten (2012) and Gallard and von Gunten (2002), who showed that the sums of Hammett or Taft for organic compounds can be quantitatively linked to rate constants for reactions with aqueous chlorine. Taft/Hammett constants were taken from published sources (Hansch et al., 1995; Perrin et al., 1981). Four descriptors were calculated by summing Taft constants for substituents around enolizable functionalities (Enol Taft and Enolizable Taft), alkenes (Alkene Taft) and amino acids (AmAc Taft). Hammett constants were summed to account for ortho-, meta- and para-interactions in aromatic compounds, as well as total interactions (HammettOrtho, Hammett Meta, Hammett Para and Hammett Sum). Finally, empirical constants were used to develop novel descriptors for five important precursor categories: alkenes (Alkene Score), enolizable aliphatics (Enolizable Score), aromatic ketones (Aromatic Ketone Score), beta-dicarbonyl (BDicarb Score) and aromatic compounds (Aromatic Score). These were calculated by giving each substituent around a potential chlorine substitution site a score, which were then multiplied together to give a combined score for the substitution site. Therefore, these descriptors quantify the influence of specific functional groups around a chlorine substitution site on chloroform formation.

For more complex molecules, scores for individual substitution sites were summed to obtain a total score for the whole molecule. Empirical substituent constants, used to derive a score for a substitution site, were selected to minimise regression residuals, in a similar fashion to some descriptors developed by Luilo and Cabaniss (2010). The 19 descriptors introduced above were used as inputs in a multiple linear regression model in SPSS, with no y-