A stochastic model for the synthesis and degradation of natural organic matter part II: molecular property distributions

Stephen E. Cabaniss · Greg Madey · Laura Leff ·
Patricia A. Maurice · Robert Wetzel

Abstract A stochastic biogeochemical model has been developed to simulate the transformation and degradation of natural organic matter (NOM) using an agent-based algorithm which treats each molecule as a separate and potentially unique entity. Molecules react when a pseudo-random number is lower than the calculated reaction probability in a given time step; repeated time steps simulate the transformation of precursor molecules into a complex NOM assemblage. The data for each molecule—elemental and functional group composition—can be used to calculate many properties directly and exactly for each molecule in the assemblage, e.g., molecular weight (MW), fraction of aromatic C (Ar), and charge at pH 7 (Z). Empirical quantitative structure activity relationships (QSARs) are developed which permit the estimation of thermodynamic quantities K_{ow} (the octanol–water partition coefficient) and pK_a (acidity) for each molecule. Root mean square errors for these QSARs are 0.39 log units for log K_{ow} and 0.45 log units for pK_a. Distributions of both exactly calculated (MW, Ar, Z) and estimated thermodynamic (K_{ow}, pK_a) properties are examined and compared with published experimental data. Molecular weight distributions from size exclusion HPLC experiments on aquatic NOM are quantitatively similar to simulation results. pH titrations and polarity distributions from reversed-phase HPLC are qualitatively similar to simulation results. This agreement suggests that the agent-based model can be used to explore hypotheses regarding both compositional and thermodynamic properties of NOM.

Robert Wetzel—deceased.

S. E. Cabaniss (✉)
Department of Chemistry, University of New Mexico,
Albuquerque, NM 87131, USA
e-mail: cabaniss@unm.edu

G. Madey
Department of Computer Science, University of Notre
Dame, Notre Dame, IN, USA

L. Leff
Department of Biological Sciences, Kent State University,
Kent, OH, USA

P. A. Maurice
Department of Geology and Civil Engineering,
University of Notre Dame, Notre Dame, IN, USA

R. Wetzel
Department of Environment Science and Engineering,
University of North Carolina, Chapel Hill, NC, USA

Abbreviations

a_i	constant value in a linear equation
amu	atomic mass units
Ar	aromaticity, the fraction of non-carbonyl sp^2 C in a molecule or group of molecules

HPLC   High pressure liquid chromatography
K_{ow} octanol water partition coefficient
LFER   linear free energy relationship
MW   Molecular weight
N   number of predictor variables and slopes in a QSAR
NOM   Natural organic matter
pK_a $-\log K_a$, where K_a is the acidity constant
QSAR   Quantitative structure-activity relationship
r   correlation coefficient
RMSE   root mean square error
s   standard deviation
SE-HPLC   Size exclusion HPLC
Z   charge on a molecule (at pH 7)
ZD   charge density on a molecule, Z/MW (at pH 7)
#X   number of atoms of element X in a given molecule

Introduction

Natural organic matter (NOM) is a naturally occurring assemblage of molecules derived from plants, microorganisms and animals, often modified by environmental processes so that the original natural products synthesized by those organisms are unrecognizable. Unraveling the complex structures within the NOM assemblage is a challenging analytical problem (Schmitt-Koplin 1998; Cook et al. 2003; Hatcher et al. 2001). Yet, NOM has attracted attention in spite of its complexity, as it influences a variety of environmental processes, including pollutant solubilization and transport, microbial respiration and growth, light penetration in surface waters, and mineral dissolution (Aiken et al. 1985; Chiou et al. 1986; Findlay and Sinsabaugh 2003; Zepp et al. 1998; Chorover and Amistadi 2001; Namjesnik-Dejanovic and Maurice 2001).

Studies of NOM properties and its effects on environmental processes are necessarily conducted on complex assemblages, either whole or fractionated (e.g., the 'humic' and fulvic' fractions of dissolved organic matter). Although most reported properties are therefore bulk averages (Perdue and Ritchie 2004), the fundamental basis of these effects and properties is molecular, and it should be possible to extrapolate bulk properties and effects from the behavior of individual molecules. In many cases, the bulk properties are simply the summed or average contributions of individual molecules- for example, molecular weight, elemental composition and acidity. In other cases, properties like hemi-micelle formation in soil (Wershaw 1992) and non-linear optical phenomena (Wang et al. 1990; Power et al. 1986; Goldstone et al. 2004) may arise from intermolecular interactions which in theory can be predicted from the properties of individual molecules.

AlphaStep and NOMSim are agent-based models which simulate the transformation and degradation of NOM using a stochastic kinetic algorithm (Xiang et al. 2004, 2006; Cabaniss et al. 2005; Huang et al. 2005). As discussed in Part I of this work, this modeling approach treats NOM as a set of interacting molecules derived from given precursors- tannins, lignin, terpenoids, proteins, etc. (Figs. 1 and 2 of Cabaniss et al. 2005). Each molecule is represented as an individual software structure, or agent, characterized by its elemental and functional group composition (Table 1 in Cabaniss et al. 2005) and by a set of derived or calculated properties which include reaction probabilities (Tables 2 and 5 in Cabaniss et al. 2005). Time is treated discretely as a series of steps (6 min each is the default) and for each step each molecule is tested for the possibility of a reaction by comparing the reaction probabilities to a pseudo-random number. If a reaction occurs, the molecule is modified or eliminated as appropriate for that particular reaction (Table 4 of Cabaniss et al. 2005). Reaction probabilities depend not only upon the molecular composition, but upon environmental parameters like pH, light intensity, temperature and enzyme activity (Table 3 of Cabaniss et al. 2005). Thus, a simulation which begins with multiple copies of only 2 or 3 different structures potentially may generate thousands of different structures over the course of several simulated months of reactions.

The current version of the agent-based model simulates a simplified oxidative environment with user-specified levels of water, extracellular enzymes, microbes and sunlight. This implementation allows 12 transformations, most of which have multiple pathways, i.e., oxidation of an alkene might occur thermally, photolytically or enzymatically, with the total probability of reaction equal to the sum of the probabilities of the individual pathways. The reactions were selected because they are postulated to occur in

aerobic soils and waters, but the current set is not comprehensive, and does not include reactions specific to phosphorus or sulfur groups. The model counts the frequency of each transformation (Cabaniss et al. 2005), and typical simulations are dominated by a series of oxidative reactions, supplemented by condensations (in soil), amide hydrolyses, and microbial utilization. Over the course of a 7-month incubation, most of the precursor material is removed from the system either by microbial uptake or direct mineralization to CO_2. There is no conceptual reason the model cannot simulate reducing conditions by employing suitable reduction and addition reactions, and these will be included in a future implementation.

For each molecule (or agent), structural properties like molecular weight, percent aromatic carbon or equivalent weight are calculated exactly for each molecule (Table 2 in Cabaniss et al. 2005). Any property which can be derived directly from elemental composition and functional group information is available in principle; for example, elemental ratios or the ratio of phenolic to alcoholic carbon can be calculated for each molecule and treated either as an overall average for comparison with bulk data or treated as individual molecules as in a van Krevelen plot.

However, reaction energetics including kinetic and thermodynamic properties cannot be calculated directly and exactly from composition data alone. If the model is to be used to predict thermodynamic behavior of NOM, then it must have some mechanism for estimating thermodynamic properties from structural information.

This manuscript examines three properties directly derivable from structure—molecular weight, aromaticity, and charge at pH 7 and two estimated thermodynamic properties—the octanol–water partition coefficient (K_{ow}) and the acidity constant (K_a). Quantitative structure activity relationships (QSARs) for estimating the thermodynamic properties from molecular structure are calibrated and evaluated. Distributions of these properties for two NOM simulations are presented and compared with available experimental findings.

Molecular weight MW

Molecular weight has been related to several aspects of NOM reactivity, including metal complexation, adsorption onto minerals, hydrophobic compound solubilization, microbial 'lability' and penetration into nanopores (Cabaniss et al. 2000). In addition, some workers have found correlations between molecular size, aromaticity and ultraviolet light absorption (Chin et al. 1994). Although MW averages and distributions are easily calculated from molecular formulae, experimental measurements of average MW of NOM samples have proven problematic, and reported results depend on the analytical method used. The size-exclusion HPLC method (Chin and Gschwend 1991; Zhou et al. 2000) gives MW averages in reasonable agreement with ultracentrifugation and vapor–pressure osmometry (Perdue and Ritchie 2004) and has the added advantage of providing MW distribution data as well.

Aromaticity Ar

Aromaticity is defined as the fraction of C atoms in an NOM sample which have sp^2 hybridization, excluding carbonyl and carboxyl C. These unsaturated C structures are thought to absorb visible and solar UV radiation (Traina et al. 1990; Goldstone et al. 2004), enhance the association of NOM with other aromatic compounds through π–π interactions (Chin et al. 1997; Wijnja et al. 2004), mediate electron transfer reactions (Scott et al. 1998) and promote disinfection byproduct formation during drinking water chlorination (Reckhow et al. 1990; Liang and Singer 2003). Like MW, Ar is readily calculated from structural information; unlike MW, a single analytical method (^{13}C NMR) is generally accepted as the standard experimental technique for estimating Ar. However, quantitative measurements of Ar depend both on the specific pulse sequence used and the method of integration, so that more recent results are considered more reliable (Cook and Langford 1998; Perdue and Ritchie, 2004).

Charge Z and charge density ZD

Molecular charge and charge density strongly affect both the solubility and ionic interactions (metal binding, sorption) of NOM. Otherwise hydrophobic molecules can be made water-soluble or show surfactant characteristics if they are charged. Metal ion

complexation by NOM is thought to have a strong electrostatic component generated by the negative molecular charges (Bartschat et al. 1992; Avena et al. 1999). The pH dependence of NOM sorption to Fe oxide surfaces also appears to have a substantial electrostatic component (Zhou et al. 2001). The negative charge on NOM arises largely from carboxylic acids (Leenheer et al. 1995a and b) most of which are deprotonated at pH 7 (Cabaniss 1991), and the positive charge (if any) arises from amine groups. Here the charge Z on a molecule at pH 7 is defined as the number of amine groups minus the number of carboxylic acid groups, and charge density ZD as Z/MW.

Octanol-water partition coefficient K_{ow}

Molecular polarity is linked to the ability of NOM to interact with and solubilize hydrophobic molecules (Chiou et al. 1986) and also to NOM sorption onto mineral surfaces (Maurice et al. 1998; Maurice and Namjesnik-Dejanovic 1999; Lumsdon et al. 2005). K_{ow} is often used as a surrogate parameter for molecular polarity, at least insofar as interaction with water is concerned (Schwarzenbach et al. 2003). K_{ow} is defined as the equilibrium partitioning constant between aqueous and octanol phases, so that for a compound A A_{oct}

$$A_{aq} \Leftrightarrow A_{oct} \qquad K_{ow} = \frac{[A]_{oct}}{[A]_{aq}}. \qquad (1)$$

Because of the utility of K_{ow} in pharmaceutical and environmental applications, large databases of >10,000 compounds have been compiled for this parameter (Hansch and Leo 1979; Meylan and Howard 1995; ACD Labs 2004). In addition, numerous methods of estimating log K_{ow} from molecular structure have been developed (Lyman et al. 1990; Schwarzenbach et al. 2003), many based upon the fragment approach of Hansch and Leo (1979). NOM polarity (also known as hydrophobicity or hydrophilicity) is sometimes linked to adsorption onto XAD resins (Aiken et al. 1992), but this is an operational classification without a clear mechanistic relationship to molecular structure. Dejanovic and Cabaniss (2004) recently published a method for estimating the log K_{ow} distribution within dissolved NOM samples, and reported that at pH 4 most of the

NOM assemblage had K_{ow} values between 10 and 100, indicating fairly polar structures.

Acidity constant K_a

The charge on NOM molecules is a function of pH, and in acidic and neutral solutions this charge is principally due to deprotonation of carboxylic acid groups (Leenheer et al. 1995a and b). The site acidity constant K_a of a carboxylic acid group R–COOH is defined by the reaction expression

$$R{-}COOH \Leftrightarrow R{-}COO^- + H^+$$
$$K_a = \frac{[R{-}COO^-]\{H^+\}}{[R{-}COOH]}. \qquad (2)$$
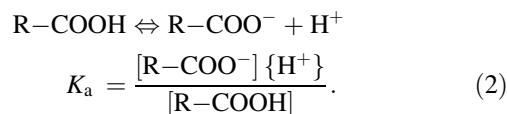
The negative log of K_a, or pK_a, is used as a convenient indicator of acid strength and typically ranges from 2 to 6 for most carboxylic acids. Like K_{ow}, pK_a is of such practical and theoretical importance that numerous predictive methods have been devised, beginning with the early work of Hammett (1937) and proceeding through the table-based calculations of Perrin et al. (1981) to more sophisticated expert systems (ACD Labs 2004) and quantum chemical simulations (da Silva et al. 1999; Toth et al. 2001).

Calibration of quantitative structure activity relationships

Quantitative structure-activity relationships provide a rapid method of estimating reaction constants (equilibrium or kinetic) for a compound of known structure. QSARs are often based upon linear free energy relationships (LFERs)—the idea that a reaction ΔH^o or ΔG^o can be expressed as the weighted sum of the energy terms of related reactions. Thus, the predicted quantity is usually log K (or ln K) rather than a K value, since at constant temperature the log K is proportional to ΔG^o (Carey and Sundberg 2000; Schwarzenbach et al. 2003). QSARs differ from LFERs in that the sum may include structural terms, not only reaction energies. For a typical QSAR, the estimated reaction K is calculated from a linear equation of the type

$$\text{Log } K = a_0 + \sum_{i=1}^{N} a_i x_i. \tag{3}$$

where x_i are the N predictor variables (numbers of a particular structural fragment, for example), a_i are the N coefficients (slopes) for these variables, and a_0 is the intercept. Devising a robust and useful QSAR requires a calibration data set with variability in K and molecular structure similar to the group of 'unknown' compounds (i.e., known structure but unknown K) to which the QSAR is to be applied. The calibration data set is then used to select an appropriate set of independent variables x_i and the corresponding 'best fit' a_i and a_0. The choice of possible x_i is not necessarily unique, so several different QSARs may be devised for a given calibration set.

Numerous QSARs have been devised for predicting log K_{ow} and pK_a, but they use molecular structure information not available in this agent-based data model—for example, proximity of two functional groups in a pK_a QSAR (Perrin et al. 1981), or the length of alkyl chains in a log K_{ow} model (Schwarzenbach et al. 2003). In this paper, QSARs for log K_{ow} and pK_a have been devised which

(1) use only the structural information available in the AlphaStep/NOMSim data model–elemental composition and functional group counts,
(2) can be used on molecules of relatively large size, and
(3) minimize the root mean square error (RMSE) for the calibration data set.

The general statistical procedure applied to a candidate set of predictor variables was to use linear regression to fit Eq. 3 to the calibration data set with all candidate variables, obtaining not only the coefficients a_i but also their standard deviations s_{a_i}. Then the least significant variable, defined as that having the highest ratio of s_{a_i} to absolute value for a_i, was discarded. This procedure was repeated until all remaining variables had s_{a_i} less than the absolute value of a_i.

Log K$_{ow}$ QSAR

A calibration data set was chosen which contains 71 compounds composed only of C, H, O and N atoms with log K_{ow} values ranging from –1.05 for N-methyl acetamide up to 4.21 for di-benzoether. Although this data set contains 6 hydrocarbons, most of the compounds had one or more functional groups—12 alkyl carboxyls, 12 aromatic carboxyls, 12 alcohols, 20 phenols, 6 aldehydes, 6 ketones, 6 ethers, 4 amines, 4 amides and 8 esters. Experimental log K_{ow} values were obtained from Hansch et al. (1995), as reported by EPI Suite (EPA 2000). The KowWin v. 1.67 program (Meylan and Howard 1995) predicts log K_{ow} with a RMSE (root mean square error) of 0.26 log units and a maximum absolute error of 0.76 log units for this set of compounds. This RMSE is comparable to those reported for larger data sets and for alternative algorithms, which suggests that the data set, though small, is representative.

The initial attempt to predict log K_{ow} used atom/fragment equations similar to those of Meylan and Howard (1995) but containing only composition data available in AlphaStep. Although the predictive power of this method was good for small molecules, it gave experimentally unattainable values for K_{ow} (|log K_{ow}| > 50) for macromolecules, and therefore did not meet the second criterion above.

An alternative approach uses carbon-normalized independent variables, for example the oxygen to carbon atomic ratio #O/#C (where # indicates the number of atoms of a given element in the molecules). In addition, (#C)$^{1/2}$ was introduced as an independent variable as an indicator of overall molecular size. This C-normalized approach may be more reasonable for larger, flexible molecules that can fold into conformations which minimize contact between water and non-polar (hydrophobic) structures—the classic example of which is protein folding.

Using this C-normalized approach and linear regression, various combinations of independent variables were tested until a satisfactory QSAR was obtained. Table 1 shows the values and standard deviations of a_i for the final predictive equation fitted to the calibration data set, which gives an $r^2 = 0.88$, standard error of 0.43 log units, RMSE of 0.39 log units, and a maximum absolute error of 1.14 log units. The plot of predicted versus experimental log K_{ow} for the calibration data (Fig. 1a) is approximately linear. The |log K_{ow}| for AlphaStep's three macromolecular precursors are all <25 log units, much smaller than those from the fragment-based

Table 1 Regression parameters for log K_{ow} prediction

Variable x_i	Coefficient a_i (Std. Dev. of a_i)
Intercept	–1.53 (0.89)
$(\#C)^{1/2}$	1.32 (0.24)
#H/#C	0.518 (0.177)
#O/#C	–4.88 (0.76)
#Alkyl COOH /#C	5.16 (1.33)
#Aryl COOH/#C	6.27 (1.51)
#Alcohol/#C	–1.98 (0.81)
#Phenol/#C	0.633 (0.77)
#Aldehydes/#C	1.092 (0.878)
#Ketones/#C	–2.58 (1.02)
#Amines/#C	–6.25 (1.37)
#Amides/#C	–5.95 (1.12)
#Esters/#C	3.37 (1.64)

Note: #X refers to the number of atoms (C, H, or O) or functional groups in a particular molecule
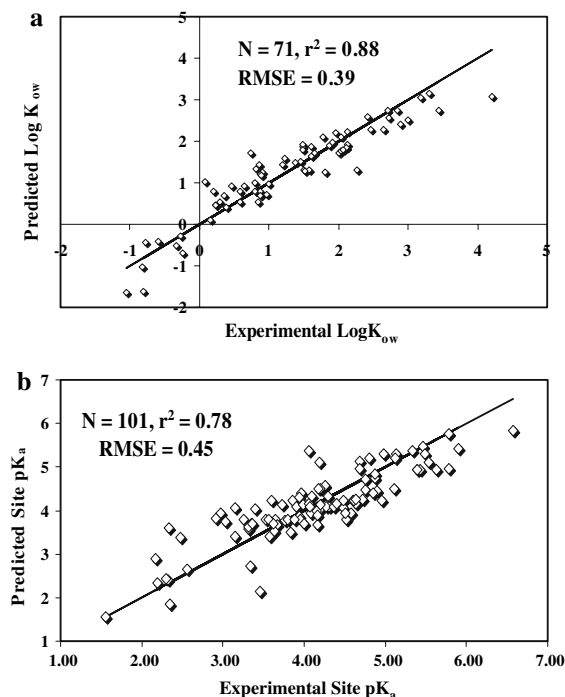
Fig. 1 Points are predicted using equation 3 versus experimental data for the calibration data set. Line is 1:1 slope, predicted = experimental. RMSE = Root mean square error. (a) log K_{ow} data, parameters from Table 1. (b) pK_a data, parameters from Table 2

QSAR. Overall predictive capability within the calibration data set (RMSE <0.4 log units) is reasonable, but not as good as the more sophisticated and extensively calibrated KowWin model (RMSE 0.26 log units).

The set of predictive terms a_i in Table 1 is chemically plausible once their interactions are considered. The coefficient for $(\#C)^{1/2}$ is positive, which indicates that larger molecules are typically less water soluble. The coefficient for the H/C ratio is also positive, consistent with the known non-polar behavior of aliphatic hydrocarbons, and the O/C ratio is very negative, consistent with the more polar behavior of O-substituted molecules. It may appear counter-intuitive that several of the O-containing functional groups, especially the carboxylic acids, have large positive coefficients; however, these terms are added to the elemental ratio terms, so that the net effect of these groups is to decrease log K_{ow} (increase polarity). For example, an aromatic acid group actually adds ($-4.88 \times 2 + 0.52 + 6.27$) = –2.97 per C, so it increases polarity more than a phenol or aldehyde group.

One notable *caveat* in thinking about these quantities is that NOM molecules are typically ionized at environmental pH values (typically pH 4–10), while these log K_{ow} values represent uncharged molecules. Since we know the average negative charge on Suwannee River fulvic acid persists even below pH 2 (Leenheer et al. 1995a and b), it might require pH 1 or below to obtain uncharged NOM molecules. It is therefore more appropriate to think of the log K_{ow} values as an index of molecular polarity or hydrophilicity, rather than as equilibrium constants used for partitioning calculations at environmental pH.

pK_a QSAR

NOM contains three common acid groups, carboxylic acids (pK_a range 2–6), phenols (pK_a range 9–14) and aliphatic amines (pK_a range 8–13). In general, natural waters have pH values between 4 and 9, so the latter two groups are generally protonated and the overall charge on NOM is typically provided almost entirely by carboxylate groups. In addition, the pK_a values of phenols and aliphatic amines have been studied less extensively and are less reliable than carboxylic acid pK_as, partly because of analytical difficulties in

determining pK_a values >12. Consequently, this work focuses on the prediction of pK_a for carboxylic acids as both the most useful and most likely to succeed.

Estimation of acidity presents a fundamentally different problem from the estimation of polarity (K_{ow}), since it is a property of a specific functional group rather than an entire molecule. Thus while a given molecule has only a single K_{ow} value, it may have multiple acid groups and consequently multiple K_a values. This difference requires an estimation algorithm capable of generating several related acidity constants for a single molecule without knowledge of their relative proximity, a severe disadvantage relative to other QSARs which use full molecular structure data.

This difference is manifested in the interpretation of experimental data on polyprotic acids. Acidity constants for polyprotic acids are typically reported as 'site constants', defined for a single specific acidic proton on a specific structure, or as 'thermodynamic constants', defined in terms of any acidic proton which is removed from any structure with the same charge to create a conjugate base of a more negative charge (Tanford 1961; Cantor and Schimmel 1980). For monoprotic acids the experimentally measured thermodynamic pK_a is equal to the site pK_a, while for polyprotic acids the thermodynamic K_a values are weighted sums of the site K_a values. If all the carboxyl groups in a molecule are equivalent (e.g., phthalic acid or succinic acid), then simple statistical weights can be used to convert site K_as to thermodynamic K_as, and vice versa. For example, in a symmetric diprotic acid the thermodynamic K_a for removing the first proton is just the sum of the two site K_as, so a site K_a is just one half of the thermodynamic K_a. For non-identical sites it is not possible to calculate site K_a from thermodynamic K_a without additional information. However, if the two sites are very different (site K_a values differ by a factor of 10 or more) it may be reasonable to assume that each thermodynamic K_a corresponds to a single site K_a (for example, in 2-ketoglutaric acid).

A calibration data set was selected with 41 monoprotic, 22 diprotic, 4 triprotic and 1 tetraprotic acids for a total of 101 pK_a values. In compiling the calibration data set for the pK_a QSAR, polyprotic acids were statistically corrected for identical or nearly identical sites (Tanford 1961); if the sites were structurally very different, site K_as were assumed to

equal thermodynamic K_as. The pK_a values at 0.10 ionic strength and 20–25°C were used when available, or obtained at other ionic strengths and adjusted to 0.1 using the extended Debye–Huckel approach (Smith et al. 1998). This set contains 32 aromatic carboxyl groups and 4 amino acids, as well as phenol, alcohol, ether and non-carboxyl carbonyl groups. The pK_a values of these acids were estimated by the free expert system SPARC (Hilal et al. 1995) and by the commercial ACD I-Lab software (Advanced Chemistry Development 2004) with RMSE of ∼0.25 for the monoprotic acids but higher RMSE for the polyprotic acids.

Successful pK_a QSARs must address the issue of proximity, in that groups near a carboxylic acid can affect its pK_a by electron withdrawal or donation, by changing solvent properties or by intramolecular hydrogen bonding. Since the AlphaStep data model lacks information on linkages, this problem is treated probabilistically. The higher the density of a particular functional group on molecule, the more likely it is to be located 'near' a particular carboxyl group. This probability is represented in the pK_a QSAR equations by 'normalizing' functional group numbers on a per carbon basis; for example, the effect of alcohols is represented by the variable (#OH/#C) rather than simply (#OH). Normalizing to overall MW gives similar but slightly inferior fits to the data.

A pK_a QSAR was obtained by testing various combinations of linear and parabolic normalized terms and discarding variables which had minimal effect on the resulting standard deviation of the model. The final QSAR equation has one constant term and nine variables, seven linear and two parabolic, as shown in Table 2. Note that #COOH is set to 0 for monoprotic acids, so that this term affects the prediction only for the polyprotic acids. The overall RMSE for the calibration data set is 0.45 log units, with an $r^2 = 0.78$. While this fit is not as good as the log K_{ow} prediction, either in terms of RMSE or r^2, this may be due to inherent difficulties in modeling polyprotic acids. Both the internet-based prediction programs SPARC and I-LAB have higher RMSE for diprotic acids (0.57 and 0.31 for the acids in this calibration set, respectively) than for monoprotic acids. Prediction error using this QSAR is comparable to the errors in quantum mechanical predictions in continuum solvent (Toth et al. 2001).

Table 2 Regression parameters for pK_a prediction

Variable x_i	Coefficient a_i (Std. Dev. of a_i)
Intercept a_0	2.81 (0.43)
#O/#C	1.65 (0.60)
#H/#C	1.21 (0.61)
#COOH/#C[a]	–1.31 (0.53)
Charge/#C	–5.68 (0.46)
#Ether/#C	–1.29 (1.04)
#Carbonyl/#C	–3.46 (0.86)
#Alcohol/#C	–1.49 (0.56)
(#O/#C)^2	–1.04 (0.24)
(#H/#C)^2	–0.28 (0.22)

[a] #COOH set to 0 for monoprotic acids based on linear

equation $\log K = a_0 + \sum\limits_{i=1}^{N} a_i x_i$

Note that while the pK_a values predicted by the QSAR are site constants, they can be converted back into thermodynamic pK_as by the inclusion of suitable statistical factors as needed (see above). Since the calibration data set contained mono- through tetra-protic acids, simulated molecules with up to 4 COOH groups are treated as mono-, di-, tri- or tetra-protic acids with their appropriate statistical factors—i.e., the carboxyl sites on each molecule are assumed to be similar. In addition, molecules with >4 COOH groups are treated as tetra-protic acids except that additional COOH groups above the fourth are assigned pK_{a2} (for the fifth and ninth groups), pK_{a3} (for the sixth and tenth groups), pK_{a1} (for the seventh and eleventh groups) and pK_{a4} (for the eighth and twelfth groups) and so forth. Thus, a molecule with 9 COOH groups would have the calculated pK_{a2} assigned to 3 of the COOH groups and pK_{a1}, pK_{a3}, and pK_{a4} assigned to two COOH groups each.

The QSAR parameters are not mechanistic, but are qualitatively consistent with known properties of carboxylic acid compounds. The functional group coefficients in Table 2 are generally negative, that is they lower the pK_a (strengthen the acid) as expected; this is because O-containing functional groups typically withdraw electron density from the acid site, stabilizing the carboxylate anion (conjugate base) form. Carbonyl groups (ketones and aldehydes) strengthen the acid more than ethers or alcohols, consistent with the differences between small acids like pyruvic (2-oxopropanoic, $pK_a = 2.48$) and lactic (2-hydroxy propanoic, $pK_a = 3.86$). The largest

coefficient in terms of absolute value is for charge, which is also expected since placing a negative charge on the molecule greatly increases the energy of further ionization (negative charge times the negative coefficient has a net positive effect on pK_a). The positive values associated with the O:C ratio are overbalanced by the combination of the (O:C)^2 term and the functional group terms-adding an O as a functional group has a net acid strengthening affect. In contrast, the H:C ratio coefficient is not generally balanced, so that adding H atoms (i.e., making the molecule more reduced) leads to weaker acids.

pK_a values for AlphaStep precursor molecules are not available in standard reference collections (NIST 1998; Advanced Chemistry Development 2004). Comparing precursor results to structurally similar molecules, the QSAR developed here predicts a pK_a of 4.22 for abietic acid and SPARC predicts pK_a 4.76 (adjusted for $I = 0.1$), compared to the experimental pK_a of 4.66 for the structurally similar cyclohexanoic acid. Other molecules not included in the calibration data set are predicted reasonably well—e.g., butanoic acid (pK_a predicted 4.68, experimental 4.63) and EDTA (pK_{a1} predicted 2.42, experimental 2.0).

Although the predictive power of this QSAR is not as good as the log K_{ow} QSAR developed above, it is comparable to ab initio predictions and only ~0.15–0.20 units worse than expert system algorithms using considerably more detailed information. The predictive capability of this QSAR is thus adequate for use with the agent-based algorithm employed here.

Results and discussion

Two simulations of NOM assemblages discussed here were described in Part I of this series (Cabaniss et al. 2005). Briefly, system 1 is a soil simulation which begins with small molecule precursors (abietic acid, meta-digallic acid and fustin) and incubates them at pH 5.0 in the dark. System 2 simulates the degradation of protein and lignin in sunlit water at pH 7.0, with somewhat lower enzyme activities. Bulk properties (average elemental composition, aromaticity, carboxyl acidity, and MW) of both simulations after 5000 h simulation time were similar to those reported in the literature (Cabaniss et al. 2005; Perdue and Ritchie 2004).
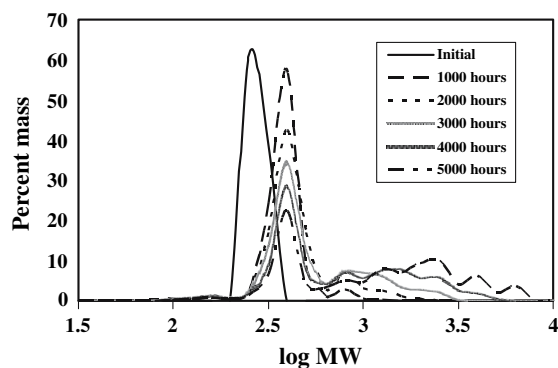
Although average properties are more easily and commonly measured, property distributions contain more information and are potentially more useful. Consider the problem of calculating the number of molecules small enough to enter a nano-pore; a M_n or M_w value is much less helpful than knowing the fraction of molecules below the desired size cut-off. In the case of thermodynamic properties—acidity for example—the presence of large numbers of weak sites with only a few strong acid sites may give a low 'average' K_a (high average pK_a) which is completely irrelevant when calculating charge at low pH.

One advantage of the agent-based modeling approach is that it provides complete distribution information for all properties which can be estimated. These distributions can be examined directly to provide insight into NOM composition and behavior, or used to simulate experimental data for predictive or testing purposes. In addition, two-dimensional plots or correlations of properties can reveal relationships between variables which are obscured in comparisons of average values.
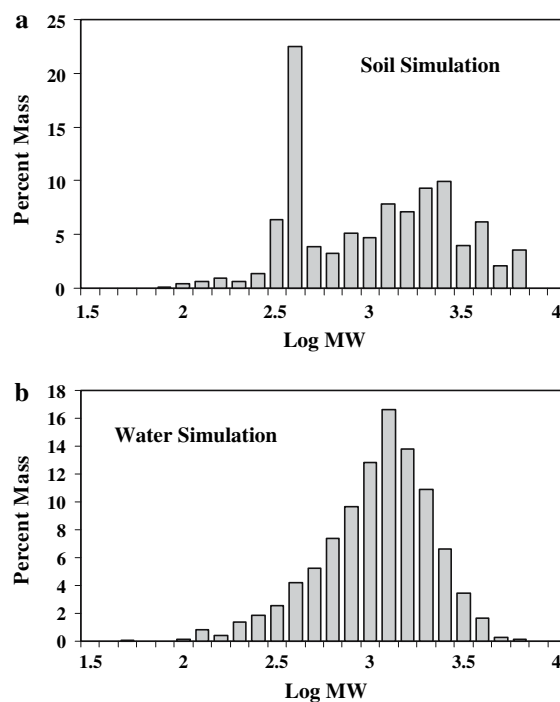
Molecular property distributions

Over the course of a 5000-h simulation, molecular weight distributions change from 'spikes' reflecting the small number of different precursor molecules (3 for soil system 1, 2 for water system 2) to broader curves more characteristic of NOM. Figure 2 shows how in the soil simulation the smaller precursors are combined over time to form larger molecules. The

initial change in MW (first 1,000 h) of the large 'spike' is due principally to oxidation reactions, which add weight without actual condensation or polymerization; subsequent increases in the numbers of high-MW compounds (MW > 600 amu, in this case) are principally due to condensations, while the decrease in the low MW 'spike' is also partly due to preferential utilization of the smallest molecules by the microbial community. Changes in the MW distribution in the water simulation are larger but less complex, the amide hydrolysis and oxidation reactions rapidly breaking the biopolymers into smaller units which then begin to recombine at longer times (data not shown).

After 5000 h of reaction time, the MW distributions for both simulations 1 and 2 have a broad peak near MW 1000. However, the distribution in system 1 is distorted by a sharper peak near MW 400 (log MW = 2.6, Fig. 3a) reflecting a large number of small precursor molecules which have not been significantly modified. In system 2 the precursor molecules have reacted more extensively, and the
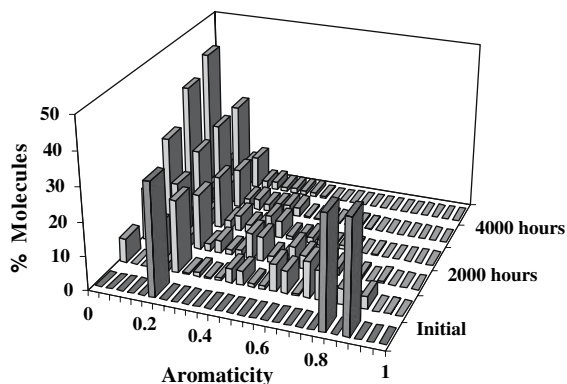

Fig. 2 Development of the molecular weight distribution over time in the soil simulation (system 1, dark incubation of abietic acid, meta-digallic acid and fustin at pH 5)

Fig. 3 Mass distribution as a function of log molecular weight (MW) after 5,000 h for (**a**) system 1, dark, acid pH incubation of abietic acid, meta-digallic acid and fustin and (**b**) system 2, sunlit incubation of protein and lignin at pH 7

mass distribution is much smoother and nearly Gaussian (Fig 3b). In each case ∼3% of the molecules have MW > 2000 amu (atomic mass units), but the system 1 assemblage has a higher proportion of molecules with MW < 200 amu (17% vs. 15% on a molecule basis, 4.3% vs. 2.8% on a per weight basis) in spite of its larger M_w.

The aromaticity distributions also reflect a larger percentage of unreacted molecules in simulation 1 after 5,000 h reaction time, shown by the pronounced peak at ∼10% aromaticity (Fig. 4a). In contrast, simulation 2 produced a much broader distribution between 7.5% and 17.5% at this time, but over 40% of the molecules had no aromatic carbons at all (Fig. 4b). This difference in the two distributions, obvious on visual comparison, is obscured in the nearly identical bulk aromaticity results of ∼10% in each case. Although less noticeable in the figures, the size of the high aromaticity 'tails' of these distributions differs significantly. The fraction of molecules with aromaticities >30% is 7.9% for simulation 1, but only 1.7% for simulation 2.
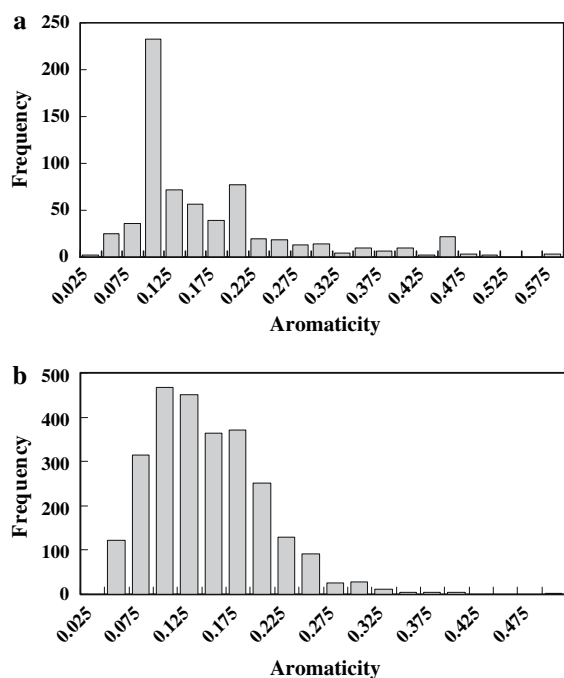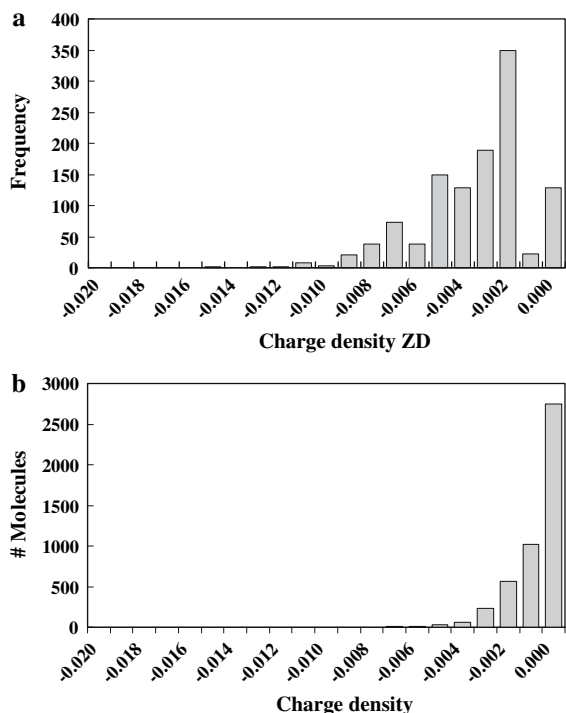

Fig. 4 Frequency distribution of molecules with respect to aromaticity after 5,000 h. (**a**) Simulated soil NOM (system 1); 486 molecules have no aromaticity (not shown). (**b**) Simulated surface water NOM (system 2); 2160 molecules have no aromaticity (not shown)
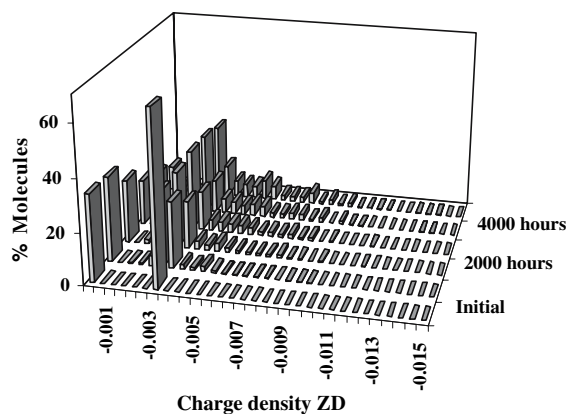

Fig. 5 Changes in aromaticity over time for the soil simulation (system 1, dark incubation of abietic acid, meta-digallic acid and fustin at pH 5). The unlabeled series are for 1,000, 3,000 and 5,000 h of simulation

The development of the soil aromaticity distribution over time reflects the relatively rapid oxidation of the highly aromatic precursors fustin and digallic acid (Fig. 5). Degradation products of 40–70% aromaticity appear quickly (1,000–2000 h) and are subsequently further oxidized to partially or fully aliphatic products. The high aromaticity precursor in the water simulation, a lignin fragment, also shows relatively rapid oxidation to less aromatic products while the less aromatic protein precursor is more slowly oxidized (data not shown).

Charge density distributions of the two simulations are qualitatively different, as might be expected from the presence of amine groups in the water simulation. After 5,000 h of simulation time, simulation 1 produced molecules which are principally anionic at neutral pH (Fig. 6a); only 11% of the molecules are neutral, and none are cationic. The slow, continued development of the high charge density 'tail' to this distribution begins early in the simulation and continues beyond the results shown here (Fig. 7). In contrast, degradation of lignin and protein in simulation 2 produced molecules which are mostly (57%) neutral at neutral pH (Fig. 6b), and a small but appreciable fraction of cationic molecules (2.4%). The neutral molecules are not necessarily uncharged; many are zwitterionic amino acids produced by protein hydrolysis. Consequently only 1.1% of the simulation 2 molecules have a charge density more negative than –0.005, corresponding to one carboxylate per 200 amu, while over 29% of the simulation 1 molecules have a higher negative charge density.
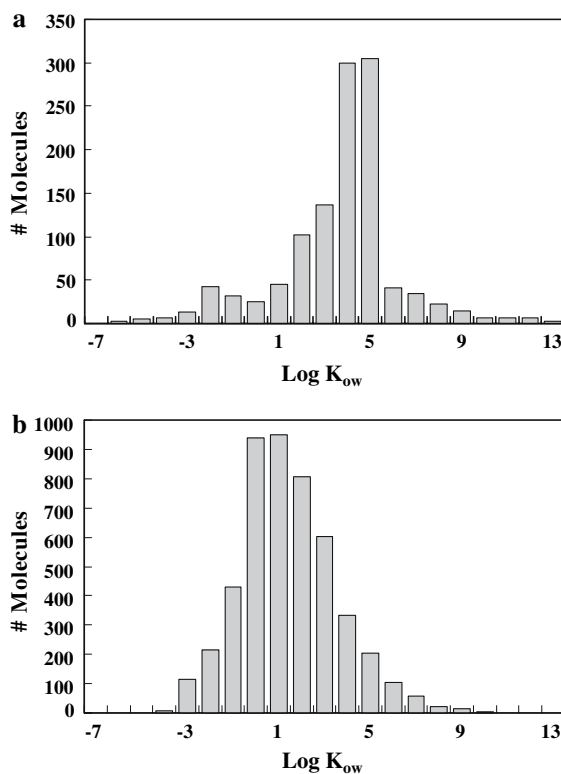
Fig. 6 Frequency distribution of molecules with respect to charge density ZD after 5,000 h. (**a**) Soil simulation (system 1) (**b**) Water simulations (system 2)

Fig. 8 Frequency distribution of molecules as a function of estimated log K_{ow} of the neutral form after 5,000 h. (**a**) Soil simulation (system 1) (**b**) Water simulation (system 2)

Fig. 7 Development of the molecular frequency distribution with respect to charge density ZD over 5,000 h for a soil simulation (system 1, dark incubation of abietic acid, meta-digallic acid and fustin at pH 5). The unlabeled series are for 1,000, 3,000 and 5,000 h of simulation

This may be a key difference in electrostatically driven processes like adsorption onto Fe oxides or complexation of hard metal cations like Al(III).

The polarity (log K_{ow}) distributions of both simulations are roughly monomodal after 5,000 h

simulation time (Fig. 8), but as for MW and Ar the simulation 1 distribution is distorted by the presence of unreacted or slightly reacted abietic acid precursor molecules, while the simulation 2 distribution is much broader and more nearly Gaussian. In addition, the simulation 2 molecules are generally more hydrophilic (peak centered near log K_{ow} 1–2) than the simulation 1 molecules (peak centered near log K_{ow} 4–5). Put another way, in simulation 2 over 70% of the molecules are more polar (hydrophilic) than 1-hexanol; in simulation 1 this percentage drops below 25%. The relatively hydrophilic nature of the NOM in simulation 2 may seem puzzling, given that the molecules in this simulation have on average a lower charge density at pH 7. This seeming discrepancy stems from three sources: First, the log K_{ow} estimates are for protonated molecules, so the carboxyl groups are not ionized. Second, many of the simulation 2 molecules contain amine groups which decrease the charge density but increase hydrophilicity. Thirdly, the average oxygen
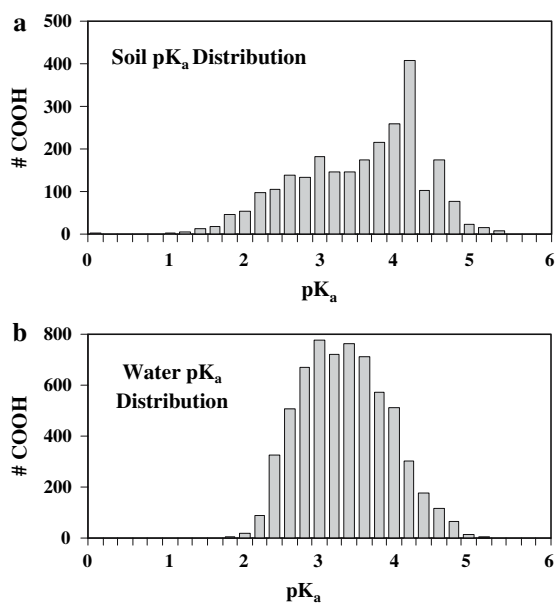
Fig. 9 Frequency distribution of carboxylic acid groups as a function of estimated pK_a after 5,000 h. (a) Soil simulation (system 1) (b) Water simulation (system 2)

content is somewhat higher in simulation 2, and these polar functional groups (not necessarily carboxyls) also increase hydrophilicity.

The pK_a distributions of both simulations after 5,000 h (Fig. 9) can be approximated as somewhat 'peaked' Gaussian distributions with a central tendency near pK_a 3.4. The soil distribution has a mean pK_a of 3.44 and a standard deviation (SD) of 0.84 log units; this distribution is somewhat distorted by a 'spike' near pK_a 4.2 which corresponds to unreacted abietic acid. The distribution is also skewed toward stronger acids while 2.9% of the acid groups have pK_a below 1.75 (2 SD below the mean), just above 0.5% have a pK_a above 5.13 (2 SD above the mean). Simulation 2 produced a narrower distribution with mean pK_a 3.24 and a standard deviation of 0.59 log units. However, this distribution is skewed toward weaker acids—only 0.6% of the acid sites have pK_a below 2.06 (mean –2 SD) while 2.9% have pK_a above 4.42 (mean + 2 SD). Both distributions are somewhat more 'peaked' than a true Gaussian, as indicated both by the facts that (a) in each case the 2 SD tails are less than the 5% of the total expected for a Gaussian distribution and (b) the kurtosis values are –0.29 and –0.45 for the soil and water simulations, respectively.
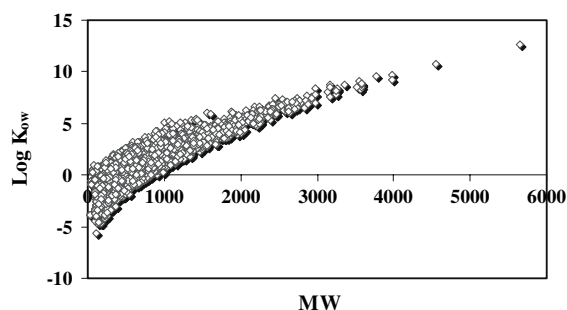
Relationships among molecular properties

Information about individual molecules in the NOM assemblage enables us to examine the relationships among molecular properties in detail. A confounding factor is that some properties are used in the QSARs for predicting others; for example, ZD is used in predicting pK_a, so a relationship between the two variables is expected.

The simplest relationship, that of linear correlation, was tested for each pair of the properties Ar, ZD, MW, and log K_{ow}. pK_a was not used in this test, since some molecules lack carboxylic acid groups, while others have more than one. The squared correlation coefficients are given in Table 3. Eight of the 12 r^2 values are below 0.1, and only 2 exceed 0.2. The charge density ZD does not correlate with any of the other three properties in either the soil or water simulation. On the other hand, the correlation of log K_{ow} with MW in both simulations ($r^2 = 0.31$ for soil, 0.77 for water) was expected, since $(\#C)^{1/2}$ is used in the QSAR for log K_{ow} prediction. The observed relationship is curved, rather than linear, consistent with the $(\#C)^{1/2}$ term (Fig. 10).

The smaller r^2 value of 0.18 in the water simulation between Ar and log K_{ow} is interesting, since the QSAR does not distinguish between aromatic and aliphatic C in these molecules and aromatic hydrocarbons are not less polar than aliphatic hydrocarbons. This correlation apparently derives from the presence of non-oxidized lignin precursor molecules. Lignin is relatively non-polar, and as long as it is not substantially oxidized it retains

Table 3 Linear correlation (r^2 values) of molecular properties

Property	MW	Ar	ZD	K_{ow}
System 1 (simulated soil NOM)				
1. Molecular weight MW	–			
2. Aromaticity Ar	0.03	–		
3. Charge Density ZD	<0.01	<0.01	–	
4. Partition coefficient K_{ow}	0.31	0.08	0.02	–
System 2 (simulated water NOM)				
1. Molecular weight MW	–			
2. Aromaticity Ar	0.15	–		
3. Charge Density ZD	0.05	0.04	–	
4. Partition coefficient K_{ow}	0.77	0.18	0.04	–

Fig. 10 Log K_{ow} plotted versus MW for simulated NOM molecules from surface water

Table 4 Linear correlation (r^2 values) of pK_{an}

Property	pK_{a1}	pK_{a2}	pK_{a3}	pK_{a4}
System 1 (simulated soil NOM)				
1. Molecular weight MW	0.26	0.22	0.07	0.01
2. Charge Density ZD	<0.01	0.09	0.19	0.17
3. Aromaticity Ar	0.04	0.28	0.23	0.23
System 2 (simulated water NOM)				
1. Molecular weight MW	0.28	0.42	0.36	0.27
2. Charge Density ZD	0.30	0.10	0.05	<0.01
3. Aromaticity Ar	0.13	0.16	0.18	0.10

both high aromatic content and high log K_{ow} values. This is also consistent with the small r^2 (0.15) between MW and log K_{ow} which is observed in the water simulation (with lignin) but which drops to only $r^2 = 0.03$ in the soil simulation.

Correlating pK_{an} values with whole molecule properties is a somewhat different undertaking, since a property might have one effect on certain types of acid groups but no effect, or the reverse effect, on others. For example, comparing the pK_{a1} values of small monoprotic acids with larger polyprotic acids in the calibration set might lead to the conclusion that larger molecules have stronger acid groups (lower pK_{a1} values); however, the weakest acid groups (highest pK_a values) in the calibration set are also found on larger polyprotic acids. In general, large negative charge density should be expected to weaken the pK_{a3} and pK_{a4} values, while the presence of multiple carboxyls or amine groups should strengthen pK_{a1} values. Aromatic acids are sometimes regarded as stronger than aliphatic acids—benzoic acid has a lower pK_a than acetic or cyclohexanoic acid, for example. However, inductive effects of electron withdrawing substituents can make aliphatic acids as strong or stronger than similarly substituted aromatic acids. Overall, the aromatic acids in the calibration data set are not stronger than the aliphatic acids—in fact the aliphatic acids have a slightly (0.2 log units) lower average pK_a. Consequently, no strong correlation between Ar and pK_{an} values was predicted.

In the soil simulation, relationships between the various pK_{an} values and other variables were as expected for MW and ZD but a little surprising for Ar (Table 4). pK_{a1} showed a minor correlation ($r^2 = 0.26$) with MW but not with Ar ($r^2 = 0.04$) or

ZD ($r^2 < 0.02$); this is consistent with the idea that the first pK_a of polyacids is typically lower as the number of carboxyl groups increases. In contrast, pK_{a4} correlated somewhat with ZD ($r^2 = 0.17$) and Ar ($r^2 = 0.23$) but not with MW ($r^2 < 0.01$); this is consistent with expected electrostatic effects of ZD, but unanticipated for Ar. pK_{a2} and pK_{a3} correlations values followed the trend of pK_{a1} and pK_{a4} for MW, but showed correlations with Ar comparable to pK_{a4}. pK_{a3} also showed slight correlation with ZD ($r^2 = 0.19$). These results are consistent with a pronounced charge effect on sequential deprotonations (pK_{a3} and pK_{a4}) in which increased negative charge on the molecule increases pK_a and a statistical enhancement of pK_{a1} on large, polyprotic acids. The correlation with aromaticity is negative—more aromatic molecules tend to have lower pK_a values except for pK_{a1}. This slight correlation seems to be a consequence of the large number of molecules with no aromatic C content but relatively high pK_a values (Fig. 11)- these may represent small alkyl polyacids
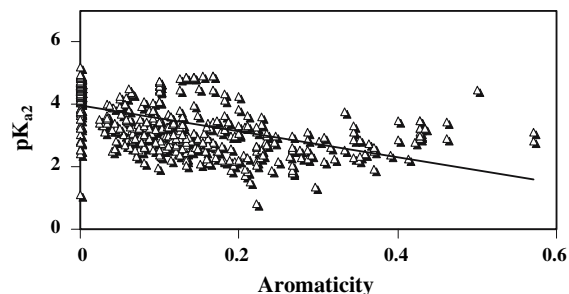
Fig. 11 Variation of pK_{a2} with aromaticity for 462 molecules (all the di-, tri-, and tetra-protic acids) in the soil simulation. The line is a linear regression 'fit' with $r^2 = 0.28$ and slope –4.15

with significant interactions among the carboxyl groups.

pK_{an} correlations in the water simulation are markedly different than those in the soil simulation (Table 4). Increasing MW tends to strengthen all the acid groups, not just pK_{a1}; r^2 varies from 0.27 to 0.42. Even more surprising, increasing negative charge (at pH 7) correlates with increased acid strength for pK_{a1} ($r^2 = 0.3$). This correlation may be a consequence of the chosen precursors—lower pK_{a1} values are found on tri- and tetra-protic degraded lignin fragments, which contain no amine groups; since the amine groups are responsible for positive charge, molecules with more net negative charge have lower pK_{a1}. Similarly, aromaticity is higher in the lignin fragments than in the protein fragments, so a small but consistent correlation between aromaticity and increased acid strength is noted.

Comparisons with experiment

Although this agent-based simulation is intended to provide mechanistic insight rather than to fit experimental data, it is nonetheless useful to compare the model predictions with our empirical knowledge of NOM. Reasonable qualitative agreement between simulation and experiment provides some confidence in the current model, although of course it does not constitute 'verification'. On the other hand, qualitative inconsistencies or gross quantitative discrepancies between simulation and experiment point out potential deficiencies in the conceptual model and/or the parameterization. Part I of this series showed that both simulations 1 and 2 produced molecular assemblages with bulk properties similar to NOM (Cabaniss et al. 2005). Comparisons of property distributions are more complex because the experimental NOM data often reflect averaged values which could arise from each of several property distributions. In other cases, no reliable method of comparison is available. While average (aggregate) measurements of charge, charge density and percent aromatic carbon are available (Cabaniss et al. 2005), current analytical methods cannot provide reliable distributions of these properties for complex NOM samples. Therefore, comparisons here will be restricted to MW, log K_{ow}, and pK_a.

The molecular weight distributions of simulated NOM are obtained from the molecular frequencies by multiplying the number of molecules of a given MW by that MW. This transforms the molecular frequency distribution into a mass distribution which is comparable to size-exclusion HPLC distributions, which are frequently observed to be log-normal with respect to MW (Cabaniss et al. 2000). While the mass distribution in the soil simulation has a sharp peak due to unreacted abietic acid after 5,000 h, the water simulation produced a monomodal, roughly log-normal mass distribution over that same time interval. Figure 12 compares the mass distribution from simulation 2 with the distribution from a SE-HPLC chromatogram of Ogeechee River FA (Cabaniss et al. 2000). The right-hand Y axis has binned mass data (0.2 log unit increments), while the left-hand Y axis shows absorbance, commonly assumed to be proportional to mass. The two distributions are quite similar, including the slight tailing on the low-MW side. The Ogeechee FA distribution peaks at slightly higher masses than the simulated NOM, consistent with its slightly higher-than average M_w and M_n (Cabaniss et al. 2000). Since the shape of the Ogeechee FA chromatogram is typical for aquatic NOM and FA samples, agreement between the model distribution and experimental (SE-HPLC) distributions is quite good.

Simulated distributions of log K_{ow} are similar in shape, although not in magnitude, to the log K_{ow} distributions obtained from reverse phase HPLC (Dejanovic and Cabaniss 2004). Since the experimental data were obtained on partially deprotonated molecules at pH 4 and the simulated data represent
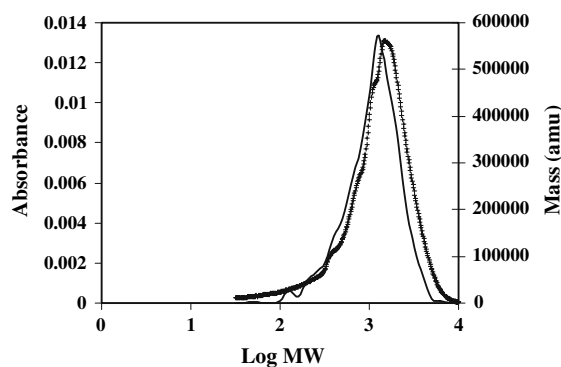
Fig. 12 Molecular weight distribution of simulated NOM (-, surface water simulation 2, right Y axis in atomic mass units) and the size-exclusion HPLC chromatogram of Ogeechee River FA (+, left Y axis, from Cabaniss et al. 2000)

a hypothetical uncharged state, an exact match cannot be expected. Simulation 1 gives a monomodal distribution centered near log $K_{ow} \sim 5.5$. The observed difference in the log K_{ow} between this and experimental data (centered between log K_{ow} 1 and 2) is consistent with the differences expected for an average negative charge of –2. On the other hand, simulation 2 gives a more nearly Gaussian distribution centered near log $K_{ow} \sim 1$. The distribution is much wider than is experimentally observed, however, which may be due to the charge effects discussed above (which would increase polarity of some simulated molecules) and the inability of the experimental technique to observe the most polar molecules (which would simply elute with the solvent peak). Overall, the agreement between simulation and experiment is encouraging but not quantitative.

Simulated pK_a distributions (Fig. 9) are qualitatively similar to interpretations of experimental results. pK_a distributions are not directly experimentally accessible, although they have been estimated from titration data using numerical methods (Perdue and Lytle 1983; Nederlof et al. 1993). In particular, several authors have found that a Gaussian distribution of pK_a provides a parsimonious representation of pH titration data; in general, only two Gaussian distributions—one for carboxyl acidity and one for phenolic acidity—can provide a reasonable 'fit' to the data (Perdue et al. 1984). Titrating Satilla River FA at ionic strength 0.1, Perdue and Lytle (1983) found the carboxyl acidity could be described by a Gaussian distribution with a mean of 3.62 and a standard deviation of 2.31. It should be noted that, like many pH titrations of NOM, this data set extends down only to pH ~ 3, so that the strongest carboxyls were never protonated during the experiment; this limits the reliability of the postulated pK_a distribution at low pK_a. Similarly, the higher-pK_a portion of the distribution overlaps with reported pK_a values for some strongly acidic phenols and amines, so that the 'carboxyl' acidity may include contributions from those functional groups. However, the central tendency of the Satilla River FA distribution is quite similar to that of the simulated NOM in Fig. 9 ($\mu = 3.44$ for the soil simulation; $\mu = 3.24$ for the water simulation), and the monomodal distributions in Fig. 9 are qualitatively similar to the Gaussian model.

However, calculated pK_a distributions are typically not unique; i.e., more than one distribution may fit the data to the same degree. Consequently, the most robust test of the simulated NOM is a direct comparison with experimental data. Titration data can be 'normalized' to account for different carboxyl concentrations by plotting the fraction of groups ionized, α, as a function of pH. In theory, α varies from 0 to 1 during a base titration as carboxyl groups are successively deprotonated. The titration plot for a single monoprotic weak acid is a classic sigmoidal curve in which α increases from 0.09 to 0.91 over the two pH-unit range bracketing the point pH = pK_a.

pH titrations of NOM from different sources tend to be similar in shape, showing some slight negative charge ($\alpha = 0.1$ to 0.2) even at fairly acidic pH, a featureless increase in α between pH 3 and 6, and a smaller increase between pH 6 and pH 8–9 leading to an 'endpoint' between pH 8 and 9. A second area of increasing negative charge at pH > 9 is not believed to correspond to the deprotonation of carboxylic acid groups (Perdue et al. 1984; Leenheer et al. 1995a), and can only be considered when successful QSARs have been developed for phenols and aliphatic amines (see comments above under pK_a QSAR.

Figure 13a compares pH titration data for terrestrial NOM with calculations based on the simulated soil NOM pK_a distribution shown in Fig. 9a. Experimental data from Cabaniss (1991) is for an isolated FA from Lake Drummond, NC, while the data from Dempsey and O'Melia (1983) is from a sub-fraction of FA from the same source. Both titrations are at ionic strength 0.1, although at different organic matter concentrations. The simulated titration curve was calculated for a series of pH values by averaging the individual α values for all carboxyl groups present. The simulated curve has an overall negative charge even at pH 2, as observed by Leenheer et al. (1995a) and the increase in α between pH 3 and pH 6 is less than expected for a simple monoprotic acid although greater than observed experimentally. Overall, the simulated titration has behavior similar to experiment near the midpoint (pH 3–4), but systematically overestimates α at higher pH (4.5–6.5). This overestimate might be due to the titration of non-carboxylic acids (phenols, amines) at the higher pHs, but is more likely due to a failure of the pK_a prediction algorithm to adequately represent the effects of increasing negative charge on larger
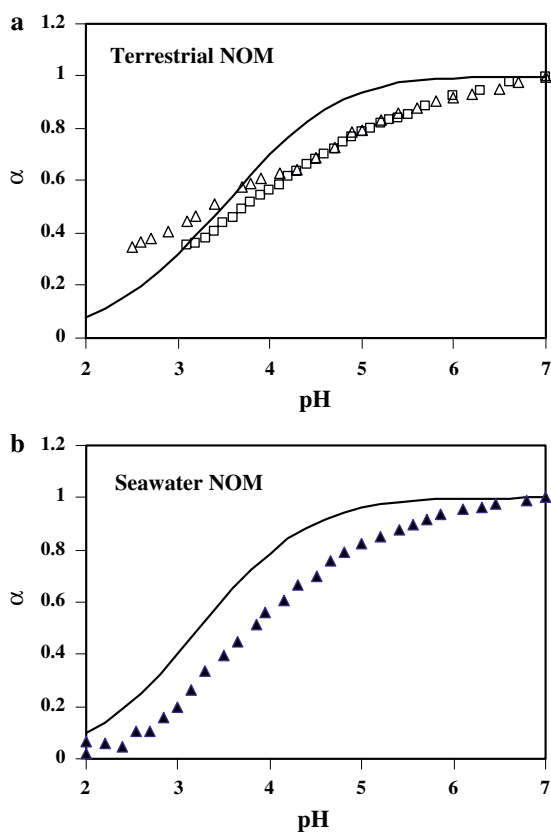
Fig. 13 Comparison of experimental and calculated titration curves, α versus pH. (**a**) Experimental data from Lake Drummond FA (□, Cabaniss, 1991) and a Lake Drummond FA sub-fraction #3 (△, Dempsey and O'Melia 1983), both at ionic strength 0.10. Model curves calculated for soil simulation (—). (**b**) Experimental data from Block Island Sound extract (▲, Huizenga and Kester 1979), model curves calculated for water simulation (—)

molecules (since only relatively small polyacids were included in the calibration data set).

Figure 13b compares pH titration data for seawater NOM with calculations based on the simulated water NOM pK_a distribution shown in Fig. 9b. Experimental data is for Block Island Sound NOM isolated on charcoal by Huizenga and Kester (1979) and titrated in simulated seawater. The simulated titration curve was calculated as above, and also has an overall negative charge at pH 2. Agreement between experiment and calculation is somewhat worse in this case than for the soil NOM, although this is partly due to the fact that Huizenga and Kester (1979) based their calculation on the assumption that the NOM was uncharged at pH 2. If typical values of α for pH 2 are assumed, the agreement is good for

lower pH but the model overestimates α at pH 4.5–6.5 as before.

Conclusions and future work

This agent-based model of organic matter transformations provides a good qualitative description of MW, polarity and pH data for NOM, even though the polarity and pK_a estimates are calibrated with model compounds rather than on NOM. However, it does not provide the quantitative 'goodness of fit' that is expected from a model calibrated on NOM data. The agreement of the shapes of log MW, log K_{ow} and pK_a distributions between the simulation and experiment is encouraging, but more work on both model and experiment may be needed for meaningful quantitative comparisons of the log K_{ow} distributions. The model has modest input requirements- a few physicochemical parameters (pH, temperature, etc.), a qualitative idea of microbial activity (e.g. enzyme activities) and a set of plausible precursor compounds. It should be useful for providing some insight into plausible mechanisms of transformations and the degree of heterogeneity expected within the NOM mixture.

Ongoing improvements in the model are dictated largely by the research interests of the authors, and include a priori prediction of metal complexation, incorporation of more sophisticated electrostatic terms, diurnal and seasonal cycling, a priori prediction of NOM adsorption onto mineral surfaces (and subsequent integration with transport models in porous media), and the effect of microbial community composition on NOM utilization. Other topics like phosphorus and sulfur transformation, hydrophobic compound solubilization and light attenuation could be incorporated by individuals or groups willing to design and calibrate suitable algorithms.

References

Advanced Chemistry Development (2004) www.acdlabs.com/products/phys_chem_lab/logp/

Aiken GR, McKnight DM, Wershaw RL, MacCarthy P (eds) (1985) Humic substances in soil, sediment, and water – geochemistry, isolation, and characterization. John Wiley & Sons, NY

Aiken GR, McKnight DM, Thorn KA, Thurman EM (1992) Isolation of hydrophilic organic acids from water using non-ionic macroporous resins. Org Geochem 18:567–573

Avena MJ, Koopal LK, van Riemsdijk WH (1999) Proton binding to humic acids: Electrostatic and intrinsic interactions. J Colloid Interf Sci 217:37–48

Bartschat BM, Cabaniss SE, Morel FMM (1992) An oligo-electrolyte model for cation binding by humic substances. Environ Sci Technol 26:284–294

Cabaniss SE (1991) Carboxylic acid content of a fulvic acid determined by potentiometry and aqueous FTIR. Anal Chim Acta 255:23–30

Cabaniss SE, Madey G, Leff L, Maurice PA, Wetzel RG (2005) A stochastic model for the synthesis and degradation of natural organic matter Part I. Data structures and reaction kinetics. Biogeochemistry 76:319–347

Cabaniss SE, Zhou Q, Maurice PA, Chin Y-P, Aiken GR (2000) A log-normal distribution model for the molecular weight of aquatic fulvic acids. Environ Sci Technol 34:1103–1109

Cantor CR, Schimmel PR (1980) Biophysical chemistry part III: behavior of biological macromolecules. W.H. Freeman, NY

Carey FA, Sundberg RJ (2000) Advanced organic chemistry part A: structure and mechanisms 4th ed. Kluwer Academic/Plenum Publishers, New York

Chin Y-P, Gschwend PM (1991) The abundance, distribution and configuration of porewater organic colloids in recent sediments. Geochim Cosmochim Acta 55:1309–1317

Chin Y-P, Aiken GR, O'Loughlin E (1994) Molecular weight, polydispersity, and spectro-scopic properties of aquatic humic substances. Environ Sci Technol 28:1853–1858

Chin Y-P, Aiken GR, Danielsen KM (1997) Binding of pyrene to aquatic and commercial humic substances: the role of molecular weight and humic structure. Environ Sci Technol 31:1630–1635

Chiou CT, Malcolm RT, Brinton TI, Kile DE (1986) Water solubility enhancement of some organic pollutants and pesticides by dissolved humic and fulvic acids. Environ Sci Technol 20:502–508

Chorover J, Amistadi MK (2001) Reactions of forest floor organic matter at goethite, birnessite and smectite surfaces. Geochim Cosmochim Acta 65:95–109

Cook RL, Langford CH (1998) Structural characterization of a fulvic acid and a humic acid using solid state ramp-CP-MAS C-13 nuclear magnetic resonance. Environ Sci Technol 32:719–725

Cook RL, McIntyre DD, Langford CH, Vogel HJ (2003) A comprehensive liquid-state heteronuclear and multidimensional NMR study of Laurentian fulvic acid. Environ Sci Technol 37:3935–3944

da Silva CO, da Silva EC, Nascimento MAC (1999) Ab initio calculations of absolute pK(a) values in aqueous solution I. Carboxylic acids. J Phys Chem A 103:11194–11199

Dejanovic KN, Cabaniss SE (2004) A reverse phase HPLC method for determining polarity distributions in natural organic matter. Envr Sci Technol 38:1108–1114

Dempsey BA, O'Melia CR (1983) Proton and calcium complexation of four fulvic acid fractions. In: Christman RF, Gjessing ET (eds) Aquatic and Terrestrial Humic Materials. Ann Arbor Science, Ann Arbor, MI

Environmental Protection Agency (2000) EPI Suite version 3.11

Findlay SEG, Sinsabaugh RL (eds) (2003) Aquatic ecosystems interactivity of dissolved organic matter. Academic Press, San Diego

Goldstone JV, Del Vecchio R, Blough NV, Voelker BM (2004) A multicomponent model of chromophoric dissolved organic matter bleaching. Photochem Photobiol 80:52–56

Hammett LP (1937) The effect of structure upon the reactions of organic compounds. J Am Chem Soc 59:96–103

Hansch C, Leo AJ (1979) Substituent constants for correlation analysis in chemistry and biology. John Wiley and Sons, New York

Hansch C, Leo A, Hoekman D (1995) Exploring QSAR hydrophobic, electronic, and steric constants. ACS professional reference book. American Chemical Society, Washington, DC

Hatcher PG, Dria KJ, Kim S, Frazier SW (2001) Modern analytical studies of humic substances. Soil Sci 166:770–794

Hilal SH, Karickhoff SW, Carreira LA (1995) A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pK$_a$s. Quantitative Structure-Activity Relationships 14:348–355

Huang Y, Xiang X, Madey G, Cabaniss SE (2005) Agent-based scientific simulation. Computing Sci. Engr 7:22–29

Huizenga DL, Kester DR (1979) Protonation equilibria of marine dissolved organic matter. Limnol Oceanog 24:145–150

Lyman WJ, Reehl WF, Rosenblatt DH (1990) Handbook of chemical property estimation methods. American Chemical Society, Washington, DC, pp 1-1 to 1-54

Leenheer JA, Wershaw RL, Reddy MM (1995a) Strong-acid, carboxyl-group structures in fulvic acid from the Suwannee River, Georgia. 1. Minor structures. Environ Sci Technol 29:393–398

Leenheer JA, Wershaw RL, Reddy MM (1995b) Strong-acid, carboxyl-group structures in fulvic acid from the Suwannee River, Georgia. 2. Major structures. Environ Sci Technol 29:399–405

Liang L, Singer PC (2003) Factors influencing the formation and relative distribution of haloacetic acids and trihalomethanes in drinking water. Environ Sci Technol 37:2920–2928

Lumsdon DG, Stutter MJ, Cooper RJ, Manson JR (2005) Model assessment of biogeochemical controls on dissolved organic carbon partitioning in an acid organic soil. Environ Sci Technol 39:8057–8063

Maurice PA, Namjesnik-Dejanovic K, Lower SK, Pullin MJ, Chin Y-P, Aiken GR (1998) Sorption and fractionation of natural organic matter on kaolinite and goethite. In: Arehart M, Hulston N (eds) Water-Rock Interaction IX. Balkema, Rotterdam, pp 109–113

Maurice PA, Namjesnik-Dejanovic K (1999) Aggregate structures of sorbed humic substances observed in aqueous solution. Environ Sci Technol 33:1538–1540

Meylan WM, Howard PH (1995) Atom/fragment contribution method for estimating octanol-water partition coefficients. J Pharm Sci 84:83–92

Namjesnik-Dejanovic K, Maurice PA (2001) Conformations and aggregate structures of sorbed natural organic matter on muscovite and hematite. Geochim Cosmochim Acta 65:1047–1057

Nederlof MM, De Wit JCM, Van Riemsdijk WH, Koopal LK (1993) Determination of proton affinity distributions for humic substances. Environ Sci Technol 27:846–856

Perdue EM, Lytle CR (1983) Distribution model for binding of protons and metal ions by humic substances. Environ Sci Technol 17:654–660

Perdue EM, Reuter JH, Parrish RS (1984) A statistical model of proton binding by humus. Geochim Cosmochim Acta 48:1257–1263

Perdue, EM Ritchie, JD (2004) Dissolved organic matter in freshwaters. In: Drever JI (ed) Treatise on Geochemistry, vol 5. pp 273–318

Perrin DD, Dempsey B, Serjeant PP (1981) pK_a prediction for organic acids and bases. Chapman and Hall, London

Power JF, Sharma DK, Langford CH, Bonneau R, Joussotdu-bioe J (1986) Photophysics of a well characterized humic substance. Photochem Photobiol 44:11–14

Reckhow DA, Singer PC, Malcolm RL (1990) Chlorination of humic materials: Byproduct formation and chemical interpretation. Environ Sci Technol 24:1655–1664

Schmitt-Koplin P, Garrison AW, Perdue EM, Freitag D, Kettrup A (1998) Capillary electrophoresis in the analysis of humic substances – Facts and artifacts. J Chromat A 807:101–109

Schwarzenbach RP, Gschwend PM, Imboden DM (2003) Environmental organic chemistry. Wiley-Interscience, Hoboken, NJ

Scott DT, McKnight DM, Blunt-Harris EL, Kolesar SE, Lovley DR (1998) Quinone moieties act as electron acceptors in the reduction of humic substances by humics-reducing microorganisms. Environ Sci Technol 32:2984–2989

Smith RM, Martell AE, Motekaitis RM (1998) NIST critically selected stability constants of metal complexes database. US Dept. Commerce, Gaithersburg, MD

Tanford C (1961) Physical chemistry of macromolecules. John Wiley and Sons, NY

Toth AM, Liptak MD, Phillips DL, Shields GC (2001) Accurate relative pK(a) calculations for carboxylic acids using complete basis set and Gaussian models combined with continuum solvation methods. J Chem Phys 114:4595–606

Traina SJ, Novak J, Smeck NE (1990) An ultraviolet absorbance method of estimating the percent aromatic content of humic substances. J Environ Qual 19:151–153

Wang ZD, Pant BC, Langford CH (1990) Spectroscopic and structural characterization of a Laurentian fulvic acid – Notes on the origin of color. Anal Chim Acta 232:43–49

Wershaw RL (1992) Membrane-micelle model for humus in soils and sediments and its relation to humification. Open file report, US Geological Survey, Washington DC

Wijnja H, Pignatello JJ, Malekani KJ (2004) Formation of pi-pi complexes between phenanthrene and model pi-acceptor humic subunits. Environ Qual 33:265–275

Xiang X, Huang Y, Madey G, Cabaniss S, Arthurs L, Maurice P (2006) Modeling the evolution of natural organic matter in the environment with an agent-based stochastic approach. Nat Resour Modeling J 19:67–90

Xiang X, Huang Y, Madey G, Cabaniss S (2004) A web portal for supporting environmental science research. In: Scharl A (ed) Environmental online communication. Springer, London

Zepp RG, Callaghan TV, Erickson DJ (1998) Effects of enhanced solar ultraviolet radiation on biogeochemical cycles. J Photochem Photobiol B: Biol 46:69–82

Zhou Q, Cabaniss SE, Maurice PA (2000) Considerations in the use of high-pressure size exclusion chromatography (HPSEC) for determining molecular weights of aquatic humic substances. Water Res 34:3505–3514

Zhou Q, Maurice PA, Cabaniss SE (2001) Size fractionation upon adsorption of fulvic acid on goethite: Equilibrium and kinetic studies. Geochim Cosmochim Acta 65:803–812