Biogeochemistry (2005) 76: 319–347
DOI 10.1007/s10533-005-6895-z

A stochastic model for the synthesis and degradation of natural organic matter. Part I. Data structures and reaction kinetics

STEPHEN E. CABANISS[1,*], GREG MADEY[2], LAURA LEFF[3], PATRICIA A. MAURICE[4] and ROBERT WETZEL[5]

*[1]Department of Chemistry, University of New Mexico, Albuquerque, NM 87131, USA; [2]Department of Computer Science, University of Notre Dame; [3]Department of Biological Sciences, Kent State University; [4]Department of Geology and Civil Engineering, University of Notre Dame; [5]Department of Environmental Science and Engineering, University of North Carolina; *Author for correspondence (e-mail: cabaniss@unm.edu; phone: 505-277-4445)*

Received 2 September 2004; accepted in revised form 4 May 2005

Key words: Agent-based model, Carbon, Humic, Natural organic matter, Stochastic

Abstract. Here we present a stochastic biogeochemical model for the formation, transformation and mineralization of natural organic matter (NOM). The model is agent-based, with each software agent representing a single molecule of defined composition. Molecular properties and reactivities are estimated from composition and environmental parameters. Environmental parameters including temperature, pH, light intensity, dissolved O_2, moisture and enzyme activities are user controlled. Time is treated in discrete steps, and during each step potential reaction probabilities are evaluated for each molecule based on its structure and the environmental parameters. When reactions occur, the molecular composition is modified accordingly. The model uses small natural products and biopolymers for inputs, and the composition of the molecules produced is constrained only by the inputs and reaction stoichiometries, not by pre-defined structures. Example simulations using the program AlphaStep are presented, in which the breakdown of biopolymers and the condensation of small molecules both lead to molecular assemblages with elemental composition and average properties similar to those of aquatic NOM. This batch-reactor model can be expanded to include spatial information and environmental feedback.

Abbreviations: A – Arrhenius constant; CE – capillary electrophoresis; E_a – activation energy, J mol^{-1}; HPLC – high-pressure liquid chromatography; IHSS – International Humic Substances Society; k – rate constant; M_n – number-average molecular weight, amu; M_w – weight-average molecular weight, amu; NOM – natural organic matter; NMR – nuclear magnetic resonance; P – probability of a reaction occurring; PRN – pseudo-random number; R – universal gas constant in J mol^{-1} K^{-1}; RSD – relative standard deviation; SRFA – Suwannee River fulvic acid; T – temperature in Kelvins; t – time in hours (Δt = change in time); UV-Vis – ultraviolet-visible

Introduction

Natural organic matter (NOM) is ubiquitous in terrestrial, aquatic, and marine ecosystems, playing a crucial role in the biogeochemistry of aquatic and terrestrial systems (Schnitzer and Khan 1972; Aiken et al. 1985; Hessen and Tranvik 1998; Findley and Sinsabaugh 2003). Here NOM is defined as those organic molecules, which have been expelled or detached from the organism which generated them and their subsequent organic reaction products. Thus, neither living nor dead organisms nor portions thereof are considered NOM. After more than 50 years of research, we have learned a considerable amount about NOM functional behavior, including pH buffering (Perdue et al. 1984; Avena et al. 1999), metal complexation (Cabaniss and Shuman 1988; Xue and Sigg 1993), organic pollutant solubilization (Chiou et al. 1986; Chin et al. 1997), adsorption onto mineral surfaces (Gu et al. 1996; Zhou et al. 2001), alteration of mineral precipitation and dissolution (Namjesnik-Dejanovic et al. 2000; Chorover and Amistadi 2001), bioavailability to heterotrophs (Amon and Benner 1996; Wetzel et al. 1995), and promotion of photochemical reactions (Zafiriou et al. 1984; Hoigné 1990; Vaughan and Blough 1998). Structural understanding of NOM has also improved with the development of spectroscopic (Bortiatynski et al. 1996; Cook et al. 2003), chromatographic (Chin and Gschwend 1991; Namjesnik-Dejanovic and Cabaniss 2004) and mass spectrometric (Brown and Rice 2000; Klaus et al. 2000) techniques to provide information on functional groups, backbone structures, molecular size and charge.

However, expanding knowledge in these areas has highlighted the remaining deficiencies. Our understanding of NOM structure contains two persistent gaps, the twin problems of variability within a single NOM sample and variability among different samples collected at different times or locations. Variability within a single sample, also referred to as structural heterogeneity, is well attested to by the complexity of NOM behavior and spectra, as well as the failure of powerful separation methods like high pressure liquid chromatography (HPLC) and capillary electrophoresis (CE) to isolate and identify even 10% of the NOM in typical systems (Saleh 1989; Schmitt-Kopplin et al. 1998). We do not know how similar the molecules in NOM are to each other. Do most NOM molecules have similar carbon 'skeletons', differing principally in specific functional groups and average size? Conversely, do the carbon 'skeletons' differ greatly according to the precursor material, with similar collections of functional groups imparting similar reactivity?

Spatial and temporal heterogeneity is more obviously linked to the problem of NOM development from living matter-from what precursors and by what biochemical pathways is NOM formed? Does the degradation of biopolymers like lignin make a smaller or larger contribution to the NOM pool than the transformation of small natural products like tannins and terpenoids (Leenheer and Rostad 2004)? How quickly do these transformations occur, and how are they linked to microbial and chemical processes in the environment? Without understanding both the structural heterogeneity and the evolution of NOM, we cannot reasonably hope to predict the outcomes of environmental processes in which it plays a key role.

Given the importance of these questions, it is worthwhile asking why our understanding of NOM structure and development has lagged so far behind our ability to predict properties like pK_a values, metal complexation, and organic pollutant partitioning, all of which can be modeled reasonably well for specific NOM samples in the laboratory. For example, Cu(II) complexation by Suwannee River fulvic acid (SRFA, a reference NOM sample available in purified form from the International Humic Substances Society, IHSS) can be predicted within 0.10 pCu units, but the identity of the binding groups and their importance in environmental settings (at lower metal:ligand ratios) is uncertain (Brown et al. 1999).

These gaps arise partly from our desire for simple, predictive models. Functional models, whether they represent metal complexation, photo-generation of hydroxyl radicals, or NOM adsorption onto mineral surfaces, are typically 'parsimonious', i.e., they use a minimum number of reaction parameters (e.g., Zafiriou et al. 1984; Bartschat et al. 1992; Gu et al. 1995, 1996). While these models may be efficient predictors of laboratory data, the parameters are usually average estimators representing bulk properties. Some models, calibrated over larger data sets, will employ 'major' and 'minor' components, e.g., the 'weak' and 'strong' binding sites of metal complexation models or the 'reactive' and 'unreactive' carbon pools in soil NOM models (see Legovic 2001); however, these are still aggregate parameters and represent only a tiny component of the structural heterogeneity we know to be present.

Another reason for our relatively poor understanding of NOM structural heterogeneity has been our reliance on analytical methods that average over the entire NOM assemblage. Spectroscopic analyses by infrared and UV-Vis absorbance, fluorescence and one-dimensional ^{1}H and ^{13}C NMR necessarily provide weighted average data. Elemental composition and functional group determinations are usually interpreted in terms of 'average' molecules. These averages are routinely reported, although the inability to separate and purify the NOM samples by HPLC or CE indicates averaging over hundreds, if not thousands, of molecular structures.

The use of 'average' values to represent the complex NOM mixture is problematic, since often the extreme values in a distribution are more significant in terms of environmental reactivity than the central tendency. For example, 'average' values of Cu(II) binding constants by NOM isolates are routinely reported in the range of $K = 10^4 - 10^7$ from a variety of laboratory studies, but measurements under field conditions indicate some molecules have binding constants $> 10^{10}$ (Xue and Sigg 1993). This discrepancy arises because the 'average' values include binding by numerous weak ligand groups in the NOM mixture, while in the field the low Cu:NOM ratio means that only the very strongest sites in the mixture are filled. Similarly, NOM bioavailability to bacteria cannot be accurately represented as an average value, since some portion of the NOM is typically recalcitrant and not readily available, while other portions are quickly degraded (Wetzel et al. 1995; Amon and Benner 1996). Moreover, different bacteria are more or less well adapted to utilize certain components of the NOM mixture (Esham et al. 2000). This poor understanding of NOM heterogeneity and variability has significant practical consequences in other areas as well; global carbon cycling, organic pollutant and radionuclide solubilization and transport, evolution of soils, and bioavailability of nitrogen and phosphorus all involve NOM.

The development of NOM from its biological precursor compounds is both an interesting biogeochemical problem and an important aspect of predictive environmental modeling. Carbon cycling models based on average properties of various organic carbon pools can be parameterized to match overall fluxes (e.g., EPIC, Williams et al. 1985; Daisy Hansen et al. 1990; DyDOC Michalzik et al. 2003) but are too simplistic to represent the heterogeneous structure of NOM and its complex behavior in the environment. On the other hand, molecular models employing connectivity maps or electron density are too computationally intensive to be useful for large-scale environmental simulations (Schulten and Leinweber 2000).

Stochastic kinetics Simulations of chemical kinetics often proceed by representing each reaction as a deterministic differential equation (rate law) and numerically integrating the resulting set of equations. Although satisfactory for many purposes, this approach has the disadvantage that since the number of differential equations to be solved equals the number of possible reactions, even systems of modest complexity can require considerable computational time. An alternative and equally correct simulation of reaction kinetics treats the system as a set of individual molecules with reaction probabilities P (Erdi and Toth 1989). The best-known stochastic simulation algorithm is that of Gillespie (1976), which treats time as a continuous variable and can efficiently handle large numbers of reacting molecules.

An alternative stochastic algorithm used here treats time as a set of discrete ‘steps’ of duration Δt, computing the probability of reaction for each molecule during that time and then checking the probability versus a pseudo-random number (PRN) to determine whether a reaction occurs. For a first-order reaction, the probability of reaction is the product of the macroscopic rate constant k (units of inverse time) and Δt,

$$P = k\Delta t \tag{1}$$

As Δt approaches zero, this equation becomes exactly true; for finite values of Δt, a useful guideline is that P should remain ≤ 0.01, a condition maintained by using shorter time steps for more rapid reactions (Morton-Firth 1998). For thermal (non-photochemical) reactions, variation of k over typical environmental temperature ranges (0–40 °C) can be calculated using the Arrhenius equation,

$$k = Ae^{-E_a/RT} \tag{2}$$

where A is the Arrhenius frequency constant, E_a is the activation energy, R is the gas constant and T is the temperature in Kelvins.

For a second-order reaction between molecules R_1 and R_2, the probability of molecule R_1 reacting depends not only on a rate constant and Δt, but on the proximity of a molecule R_2. In a well-mixed reactor, this can be expressed as

$$P = k[R_2]\Delta t = k'\Delta t \tag{3}$$

(Shimizu and Bray 2001). For cases where R_2 is in great excess or otherwise held constant (e.g., buffered pH), the reaction is most easily treated as pseudo-first order, and the pseudo-first order constant k' is the product of the true first order k and $[R_2]$, as in Equation (3). However, in a spatially aware model or a system in which the bulk $[R_2]$ term is not fixed *a priori*, $[R_2]$ must be replaced by a probability term for finding a molecule R_2 within a suitable reacting 'distance'. When multiple reactions are possible for a given molecule, the probabilities are summed (summed p should still be <0.01) and compared to a single PRN; if a reaction has occurred, then the same PRN can be used to determine which one.

This discrete time step method is less efficient than Gillespie's approach for large numbers of identical molecules, but may be more efficient when the number of possible molecules and reactions exceeds the actual number. For example, in the simulation of cell signaling, a small number of protein molecules may each exist in 2^{20} possible states due to folding, phosphorylation, etc.; while the macroscopic or Gillespie approaches would require 2^{20} reaction calculations, the discrete time algorithm uses only as many as there are molecules in the simulation (Morton-Firth 1998). In addition, the discrete-time algorithm is easily adaptable to an agent-based implementation which incorporates spatial variability and transport.

Here we present a stochastic model that allows, for the first time, forward modeling of the evolution of NOM structure and properties. Forward modeling does not aim to fit the data by finding optimal values of adjustable parameters, but rather makes reasonable assumptions about both processes and parameters and then calculates the consequences of these assumptions. This model, encoded as the program AlphaStep, represents individual molecules as discrete software agents of specified elemental and functional group composition. The formation of NOM from biological precursor compounds such as lignin, tannins, terpenes and proteins is simulated using a discrete-time algorithm in which specific reaction probabilities are calculated from molecular composition and environmental parameters like temperature, light and enzyme activities. The reactivity and aggregate properties of the resulting NOM assemblage can then be calculated over time and compared with laboratory and/or field data. As examples, we compare biopolymers and small molecules as possible NOM precursors, simulating their transformations in surface water and soil environments, respectively.

AlphaStep implements only chemical and biological reactivity, simulating a well-mixed batch-reactor; it lacks both spatial variability and the ability to change environmental parameters either on a regular basis (e.g., diurnal variation) or as part of a feedback loop (e.g., changes in bacterial abundance or enzyme activity due to NOM availability or utilization). However, the reactions and algorithms described here can be readily incorporated in spatially and ecologically aware models which would be more environmentally realistic (e.g., Xiang et al. In press), and could be tailored for either terrestrial or aquatic environments.

Model description

An agent-based model simulates a complex system by representing separate interacting sub-units (birds, data packets, etc.) as software 'agents' with specific properties and rules of behavior (Parunak et al. 1998; Huang et al. 2005). Each agent interacts with its environment, including other agents, according to these rules, and in so doing may change properties during the simulation. Heterogeneity of behavior is thus inherent to an agent-based model, and overall properties are calculated by combining information from all the agents.

For purposes of this model, 'the NOM assemblage' (or simply 'NOM') is defined as the set of all molecular agents (or simply 'molecules') in the simulation. Precursor molecules (or simply 'precursors') are molecules of biological origin and defined structure. These are provided as starting materials for each simulation, and in general are expected to change structure during the course of the simulation. Precursors are assumed to have been generated by an organism, and only enter the NOM assemblage (i.e., enter the simulation) when they separate from the organism via excretion or decay. Some NOM molecules will have the structure of unmodified precursors, although under reactive conditions and after long reaction times these are expected to be only a small percentage of the total assemblage.

Each molecular agent has a defined structure (composition) and calculated reaction probabilities. These reaction probabilities are a function of both molecular structure and the biogeochemical environment, as represented by a set of environmental parameters. The probabilities and the transformation of the molecules through reaction are the 'rules of behavior' for these molecular agents. A simulation proceeds through a series of time 'steps' by testing each molecule for possible reactions, modifying the molecular structure accordingly if a reaction occurs, and recalculating reaction probabilities for the new structure. In addition to producing new molecular structures as a product of each reaction, the program reports reaction frequencies (counts) and calculates aggregate properties of the NOM assemblage for comparison with experiment.

Molecular structure

Each molecule in the simulation is composed of a number of C, H, N, O, P and S atoms and a number of functional groups (Table 1). Since this representation contains less information than the full atomic connectivity map commonly used in molecular modeling, thousands of molecules can be simulated at once. Because stereochemical information is not represented, the program cannot distinguish between enantiomers or between branching isomers like *n*-hexane and methyl pentane. However, functional group information allows the program to distinguish between functional isomers like butanol and di-ethyl ether.

The choice of functional groups included represents a compromise between the complexity of the NOM mixture and the need for a manageable list of parameters. The current list of functional groups was selected based on prevalence in natural products and biopolymers with as few sub-groups as were consistent with simple reaction rate predictions. For example, although carboxylic acids are divided into total acids and aromatic acids (with aliphatic acid readily calculated from the difference), no distinction is made between a purely aliphatic ketone, like cyclohexanone, and a ketone adjoining a benzene ring. Alcohols and ring structures are likewise divided into aliphatic and aromatic, but amines, aldehydes, ethers, esters, etc., are not. The list of functional groups can be readily expanded to accommodate these or other 'special' structures as needed. For example, it might be useful to include quinone moieties as special 'functional groups' for simulations of electron-transfer properties (Scott et al. 1998).

Table 1. Elemental and functional group composition of precursor compounds.

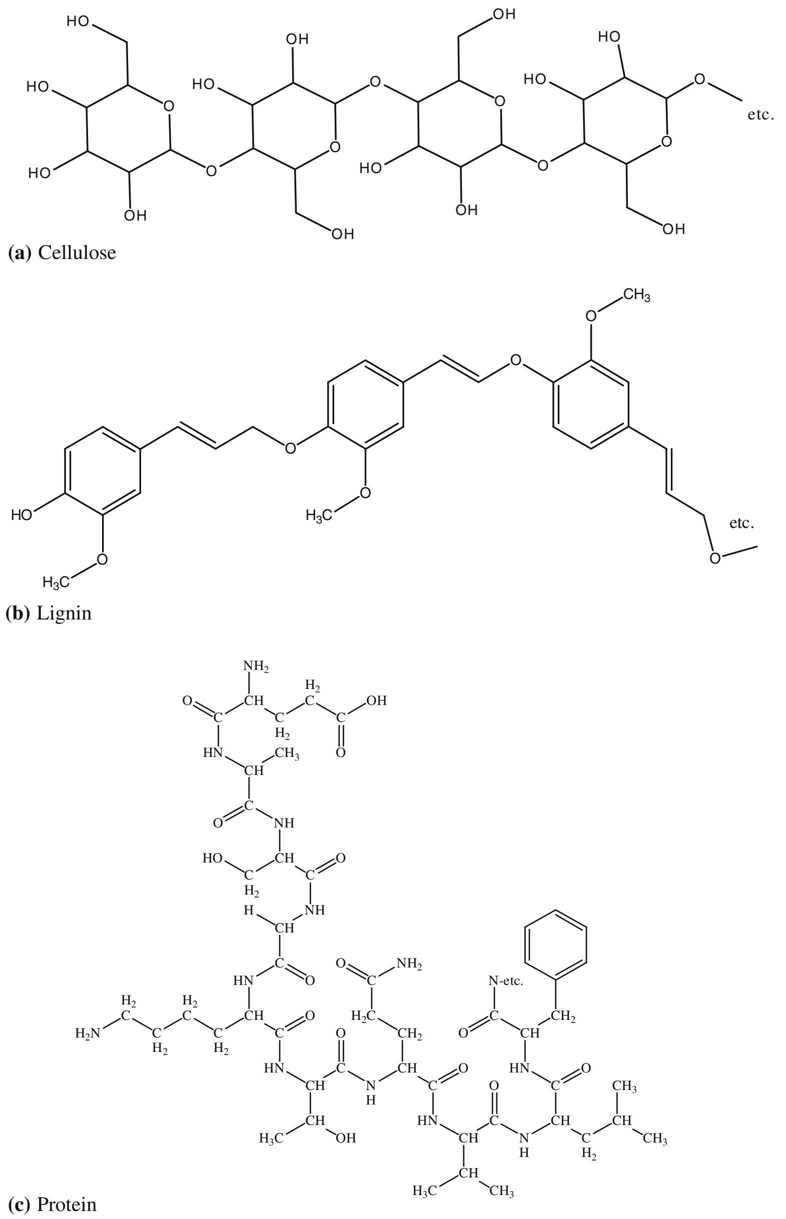
Elemental Composition			Starting materials (precursor compounds)			
	Protein	Cellulose	Lignin	Tannin	Terpenoid	Flavonoid
Carbon	240	360	400	14	20	15
Hydrogen	382	602	402	10	30	12
Nitrogen	60	0	0	0	0	0
Oxygen	76	301	81	9	2	6
Sulfur	0	0	0	0	0	0
Phosphorus	0	0	0	0	0	0
Functional groups						
C=C bonds	15	0	160	6	2	6
Rings	5	60	40	2	3	3
Phenyl rings	5	0	40	2	0	2
Alcohols	10	182	2	5	0	4
Phenols	0	0	1	5	0	3
Ethers	0	119	79	0	0	1
Esters	0	0	0	1	0	0
Ketones	0	0	0	0	0	1
Aldehydes	0	0	0	0	0	0
Acids	6	0	0	1	1	0
Aromatic Acids	0	0	0	1	0	0
Amines	6	0	0	0	0	0
Amides	54	0	0	0	0	0
Thioethers	0	0	0	0	0	0
Thiols	0	0	0	0	0	0
Phosphates	0	0	0	0	0	0

The six precursor molecules represented in Table 1, three macromolecules and three smaller natural products, were chosen to reflect various possible source and formation pathways of NOM (Robinson 1983; Leenheer et al. 2003; Leenheer and Rostad 2004). The cellulose molecule (Figure 1a) represents a chain of 60 D-glucose units, linked together through carbons 1 and 4. Actual cellulose molecules are larger and insoluble; the average molecule described here is a smaller soluble fragment, presumably lysed from the parent molecule enzymatically. The lignin molecule (Figure 1b) represents an oligomer of 40 conferyl alcohol units condensed together via ether linkages (implying dehydration). This is a vast oversimplification of the variety of monomeric units in lignin, but retains the important features of high aromaticity and numerous aryl–alkyl ether linkages. The 'average' protein molecule (Figure 1c) is a 50-residue peptide, 5 residues each of glutamic acid, lysine, glutamine, serine, threonine, glycine, alanine, valine, leucine, and phenylalanine. The terpenoid molecule is abietic acid (Figure 2a), a slightly oxidized diterpenoid composed of 4 isoprene residues. The flavonoid molecule is the flavanone fustin (Figure 2b), a yellow plant pigment. Meta-digallic acid (Figure 2c) is a hydrolyzable tannin which is more water-soluble than either abietic acid or fustin (Robinson 1983).

Each molecule also has chemical and physical properties which are calculated from its composition. Exactly calculated properties, e.g., the molecular weight, percent 'aromatic' carbon (non-carbonyl sp^2 carbon) and elemental composition, are derived without any uncertainty from the structural information. Estimated properties depend on empirical or semi-empirical relationships, e.g., octanol–water partition coefficient and pK_a, and will be considered in a subsequent paper.

Many common experimental measurements determine aggregate properties of the NOM assemblage, rather than properties of individual molecules. AlphaStep obtains average values of exactly calculated properties directly from the individual molecular structures- the properties include average elemental composition, average fraction of 'aromatic' carbon, number-average and weight-average molecular weight, and others listed in Table 2.

The formulae in Table 2 are self-explanatory in most cases, but those dealing with acidity and 'aromatic' C involve hidden assumptions. Molecular charge at pH 7, and hence the average charge, is calculated by assuming that all carboxylic acids and no phenols are deprotonated, and that all amines are protonated. This agrees reasonably well with typical pK_a values of carboxylic acids and phenols, but ignores the possibility of more acidic aromatic amines. However, since carboxylic acids are the principal charged group in NOM, this assumption should have only a small effect on the overall calculated charge. The calculation of equivalent weight similarly ignores amines with low pK_a values, considering only carboxylic acids. Finally, the aromaticity calculation assumes that all 'aromatic' C (experimentally defined by ^{13}C NMR) occurs in C=C double bonds, ignoring possibilities like C≡C triple bonds and cyclopentene anions. However, these are expected to be very minor structures in NOM.

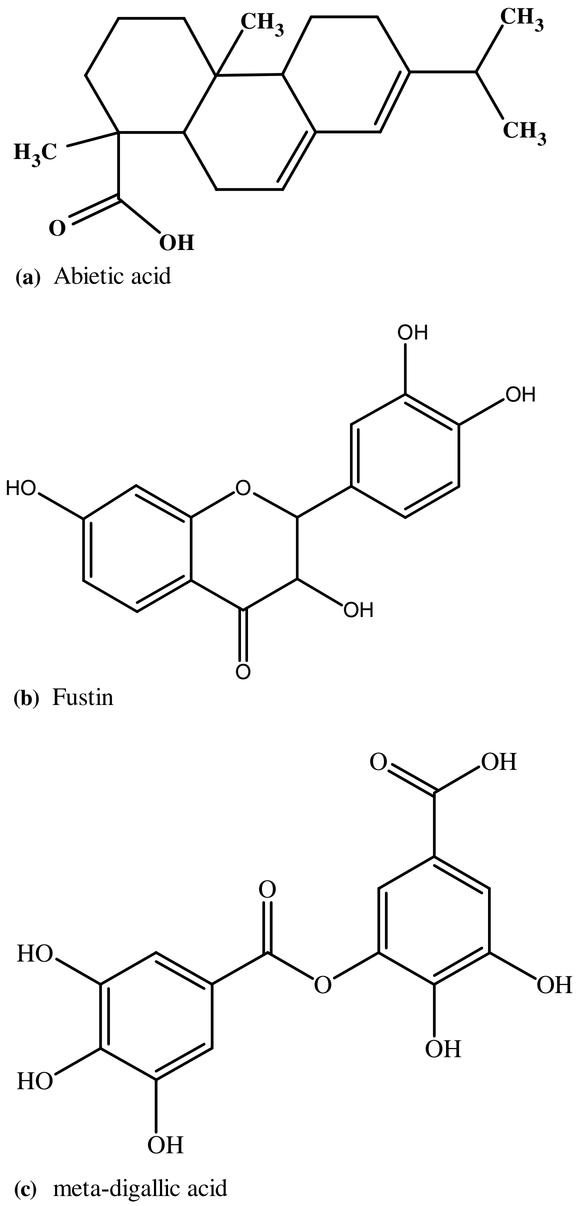


(a) Cellulose

(b) Lignin

(c) Protein

Figure 1. Macromolecular precursor structures (a) cellulose fragment composed of D-glucose sub-units (b) lignin fragment composed of condensed coniferyl alcohol sub-units (c) protein fragment composed of Glu, Lysine, Glm, Ser, Thr, Gly, Ala, Val, Leu, and Phe residues.



(a) Abietic acid

(b) Fustin

(c) meta-digallic acid

Figure 2. Small molecule precursor structures (a) abietic acid, a diterpenoid (b) fustin, a flavonoid pigment (c) meta-digallic acid, a hydrolyzable tannin.

Table 2. Formulae for calculating aggregate properties of molecular assemblages.

N = number of molecules

$\mathrm{AW_E}$ = atomic weight of element E
MW = molecular weight
Subscript i refers to ith molecule in the assemblage

$$M_\mathrm{E} = \text{mass of element E} = \sum_{i=1}^{N} \mathrm{AW_E}(\#\ \mathrm{E})_i$$

$$M_\mathrm{T} = \text{total mass of all molecules} = \sum M_\mathrm{E} = \sum_{i=1}^{N} \mathrm{MW}_i$$

$$\mathrm{Wt\%}_E = \text{weight percent of}\quad \mathrm{E} = \frac{M_\mathrm{E}}{M_\mathrm{T}}$$

$$M_n = \text{number average molecular weight} = \frac{M_\mathrm{T}}{N}$$

$$M_\mathrm{W} = \text{weight average molecular weight} = \frac{\sum_{i=1}^{N} \mathrm{MW}_i^2}{M_\mathrm{T}}$$

$$Z_\mathrm{n} = \text{average molecular charge at pH 7} = \frac{\sum(\#\ \text{amine})_i - \sum(\#\ \text{acid})_i}{N}$$

$$\mathrm{EW} = \text{average equivalent weight} = \frac{M_\mathrm{T}}{\sum(\#\ \text{acids})_i}$$

$$\mathrm{Ar} = \text{fraction aromatic C} = \frac{2\sum(\#\ \mathrm{C{=}C})_i}{\sum(\#\ \mathrm{C})_i}$$

Environmental parameters

Reaction rates, and therefore probabilities, can be influenced by a variety of biological, physical and chemical factors. The current version of AlphaStep contains four biological parameters and five physical/chemical parameters to represent the larger number of actual rate-influencing factors in the environment (Table 3).

Table 3. User-controllable simulation parameters.

Physical/Chemical	
Temperature	T, Temperature from 0 to 80 °C, affects thermal (dark) reaction rates
Water activity	A H_2O, Scaled 0–1, affects hydrolysis and hydration reactions
Light intensity	I, μmol cm^{-2} h^{-1}, affects oxidation and decarboxylation rates
pH	pH, affects hydrolysis, dehydration and decarboxylation rates
Dissolved O_2	[O_2] mM, affects oxidation rates
Biological	
Bacterial utilization	B, Scaled 0–1, affects rate of microbial utilization
Protease activity	E_P, Scaled 0–1, affects amide hydrolysis rate
Oxidase activity	E_O, Scaled 0–1, affects oxidation rates
Decarboxylase activity	E_D, Scaled 0–1, affects decarboxylation rate

The chemical and physical parameters are all measurable quantities. Temperature is permitted to range from 0 to 100 °C with a default of 24.8 °C (300 K), and influences all dark reactions through an Arrhenius relationship (Equation 2). Light intensity refers to the photon flux between 290 and 320 nm, and is given in μmol cm^{-2} h^{-1} with a default of 1 μmol cm^{-2} h^{-1}. Light intensity influences all photo-reactions in a linear fashion, and no wavelength dependence is currently implemented. The pH defaults to 7, but can range from 0 to 14; OH^- catalyzes hydrolysis reactions, while H^+ catalyzes hydrolysis and hydration/dehydration reactions and enhances decarboxylation rates. Dissolved oxygen, [O_2], can range from 0 to 5 mM, but defaults to 0.1 mM, below saturation for typical surface waters. Dissolved O_2 is the only oxidant explicitly represented, and thus controls the rate of all oxidation reactions; other oxidizing species–$^1O_2$, OH·, Fe^{3+}, etc., are assumed to be derived from and proportional to [O_2]. The activity of water is 1.0 in aqueous solution; in soils it can vary from 0 to 1, proportional to the humidity. Water is required for hydrolysis reactions and for hydration.

In contrast, the biological parameters are all unitless activities normalized to 1 for maximum reaction speed. Bacterial abundance (B), represents the population of viable heterotrophs, and controls the rate of 'utilization' and eventual mineralization of organic molecules. The current version of the model lumps all types of heterotrophs together and gives them an 'average' utilization behavior, with no feedback to account for growth or death in the bacterial community. The enzyme activities for protease (E_P), oxidase (E_O), and decarboxylase (E_D) each represent a class of enzymes, so their catalytic properties are less specific than real enzymes. For example, E_P increases the rate of all amide bond cleavage reactions, not simply those near a particular amino acid residue.

Reactions: transformations and probabilities

The very large number of possible reactions in an NOM mixture is a formidable obstacle to quantitative modeling. No database of reaction k values contains all the possible structures, let alone all possible reactions, of the molecules in NOM. Instead, AlphaStep estimates reaction probabilities based on individual molecular structure and environmental parameters. This procedure also eliminates the requirement that final product structures be known *a priori*, and allows forward modeling with a minimum of preconceived notions of NOM structure.

Alphastep currently simulates the twelve reactions shown in Tables 4 and 5. A reaction transformation is defined as a change in molecular structure which can result from one or more different reaction pathways or mechanisms. For example, the oxidation of an aldehyde to a carboxylic acid can occur by an enzymatic or a photochemical pathway; the two pathways have different probabilities and depend on different environmental parameters, but the

Table 4. Molecular transformations simulated.

Unimolecular	
Splitting	
Ester hydrolysis	Adds H_2O, splits molecule in two pieces, one with new acid group and one with new alcohol
Amide hydrolysis	Adds H_2O, splits molecule in two pieces, one with new acid group and one with new amine
Modifying	
Alkene hydration	Adds H_2O and one alcohol group, removes one C=C
Alcohol dehydration	Removes H_2O and one alcohol group, adds one C=C
Weak C=C oxidation	Oxidation of C=C to 1,2 diol. Adds H_2O_2 and two alcohol groups, removes one C=C may convert phenyl ring into non-aromatic ring
Strong C=C oxidation	Oxidative cleavage of C=C. Adds O_2 and two C=O groups, removes one C=C may cleave a ring or split molecule
Alcohol oxidation	Removes 2H atoms and one alcohol group, adds an aldehyde or ketone
Aldehyde oxidation	Adds one O atom and a carboxylic acid, removes one aldehyde.
Decarboxylation	Removes CO_2, eliminating an acid group.
Removing	
Microbial utilization	Removes the entire molecule from the simulation
Bimolecular	
Ester condensation	Combines one molecule with an acid group and a second with an alcohol to form an ester. H_2O, one acid and one alcohol are removed, one ester group added
Aldol condensation	Combines one molecule with an aldehyde and a second with an aldehyde or ketone

transformation of the reacting molecule is the same for either. A transformation is defined by the change in elemental composition (adding or losing atoms) and functional groups given in Table 4. The probability of a reaction leading to the transformation of a given molecule is calculated from the molecular structure and the environmental parameters using the equations in Table 5.

Ester hydrolysis splits a reacting molecule into two new molecules, forming an alcohol and a carboxylic acid (assuming no ring esters). The relative size of the two successor molecules is determined randomly, with the larger successor molecule retaining 50–80% of the mass. The new acid group is assigned to the larger successor, the new alcohol group to the smaller. Ester hydrolysis is frequently slow ($t_{1/2} > 100$ years) by a neutral (non-catalyzed) pathway, somewhat faster by acid catalysis-k values 0.1–10 M^{-1} h^{-1}. On the other hand, the base-catalyzed hydrolysis can be quite fast, with typical second order rate constants for esters of acetic acid 200–10,000 h^{-1} water (Mabey and Mill 1978). Although the un-catalyzed pathway tends to have a slightly smaller activation energy, a typical E_a value (Kirby 1972) was used for the more significant base-catalyzed pathway in Table 5, Equation (1). These values give an overall rate constant of 5.6×10^{-4} h^{-1} at pH 7.0, but this increases rapidly with increasing pH so that at pH 9.0 the rate constant is 5.4×10^{-2} h^{-1}.

Table 5. Calculation of reaction probabilities.

(1) Ester hydrolysis
$k' = (\#\ \text{Esters})\ [H_2O]\ A\ e^{-\Sigma_a/RT}\ (1 + b\ [H^+] + c\ [OH^-])$
$A = 6 \times 10^5\ h^{-1}$ $E_a = 60$ kJ mol^{-1} $b = 10^4$ $c = 3 \times 10^8$

(2) Amide hydrolysis
$k' = (\#\ \text{Amide})\ [H_2O]\ A\ e^{-\Sigma_a/RT}\ ([H^+] + b\ [OH^-] + E_p)$
$A = 6 \times 10^6\ h^{-1}$ $E_a = 50$ kJ mol^{-1} $b = 10$

(3) Alkene hydration
$k' = (\#\ C=C)\ [H_2O] A\ e^{-\Sigma_a/RT}\ [H^+]$
$A = 2 \times 10^{13}\ h^{-1}$ $E_a = 80$ kJ mol^{-1}

(4) Alcohol dehydration
$k' = (\#\ OH) A\ e^{-\Sigma_a/RT}\ [H^+]$
$A = 10^{12}\ h^{-1}$ $E_a = 80$ kJ mol^{-1}

(5) Weak C=C oxidation
$k' = (\#\ C=C)\ (A_{enz}[O_2]E_O\ e^{-\Sigma_a/RT} + A_{photo}[O_2]I)$
$A_{enz} = 5 \times 10^9\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1}
$A_{photo} = 5 \times 10^7\ h^{-1}\ M^{-1}$

(6) Strong C=C oxidation
$k' = (\#\ C=C)\ (A_{enz}[O_2]E_O\ e^{-\Sigma_a/RT} + A_{photo}[O_2]I)$
$A_{enz} = 10^9\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1} $A_{photo} = 10^7\ h^{-1}\ M^{-1}$

(7) Alcohol oxidation
$k' = (\#\ OH)\ (A_{enz}[O_2]E_O\ e^{-\Sigma_a/RT} + A_{photo}[O_2]I)$
$A_{enz} = 5 \times 10^9\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1}
$A_{photo} = 10^7\ h^{-1}\ M^{-1}$
If # O > # C then $k' = 0$

(8) Aldehyde oxidation
$k' = (\#\ HC=O)\ (A_{enz}[O_2]\ E_O\ e^{-\Sigma_a/RT} + A_{photo}[O_2]I)$
$A_{enz} = 5 \times 10^9\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1} $A_{photo} = 10^7\ h^{-1}\ M^{-1}$

(9) Decarboxylation
$k' = (\#\ COOH)\alpha_H\ (A_{enz}\ E_D\ e^{-\Sigma_a/RT} + A_{photo}\ I)$
$A_{enz} = 5 \times 10^7\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1}
$A_{photo} = 10^2\ h^{-1}\ M^{-1}$ $\alpha_H = [H^+]/(10^{-4} + [H^+])$

(10) Microbial utilization
If MW > 1000, then use
$k' = B\ Q\ (0.1 - 10^{-4}\ MW)$
$Q = (\#\ O/\#\ C) + 0.002\ (\#\ P + \#\ N)$
If MW < 1000, then $k' = 0$

(11) Ester condensation
$k' = (\#\ COOH)\ [H^+]\ A\ e^{-\Sigma_a/RT}$
$A = 5 \times 10^{10}\ h^{-1}\ M^{-1}$ $E_a = 60$ kJ mol^{-1}
2nd molecule must have alcohol group

(12) Aldol condensation
$k' = (\#\ COOH)\ ([OH^-] + [H^+])\ A\ e^{-\Sigma_a/RT}$
$A = 1 \times 10^{10}\ h^{-1}\ M^{-1}$ $E_a = 50$ kJ mol^{-1}
2nd molecule must have aldehyde or ketone group

Amide hydrolysis splits a reacting molecule at an amide linkage, destroying the amide link and creating an amine and a carboxylic acid. Splitting into successor molecules is treated in the same way as for ester hydrolysis. Uncatalyzed amide hydrolysis is neglected in rate calculations because it is extremely slow. The acid-catalyzed rate constants are small 0.001–0.1 M^{-1} h^{-1} – and the base-catalyzed constants are only slightly larger 0.01–0.5 M^{-1} h^{-1} (Talbot 1972; Mabey and Mill 1978). For example, the half-life of acetamide in a sterile, pH 7 solution is estimated to exceed 1000 years. However, bacteria and fungi synthesize and exude a variety of proteases which catalyze the reaction effectively; the rate expression in Table 5 is typically dominated by the enzymatic term.

Alkene (C=C) hydration adds water to form an alcohol, and both requires water and an acid catalyst. E_a is taken from typical values for small alkenes (Boyd et al. 1960; Schubert and Lamm 1966; Chiang and Kresge 1985). The equation in Table 5 gives an acid-catalyzed second order rate constant of 0.19 M^{-1} h^{-1}, or a negligible rate of 2×10^{-8} h^{-1} at pH 7.

Alcohol dehydration is the reverse of alkene hydration, the molecules gaining a C=C double bond and losing an alcohol by eliminating H_2O. Like hydration, dehydration typically proceeds by an acid-catalyzed pathway. Equation 4 in Table 5 lumps together dehydration of primary, secondary and tertiary alcohols, even though the rates differ considerably (tertiary dehydrate fastest; primary, slowest). Steric factors also cause large ($> 10^4$) changes in rate (Boyd et al. 1960; Noyce and Lance 1962). The secondary alcohol cyclooctanol dehydrates with a somewhat large activation energy (~80–90 kJ/mole) in a reversible fashion, with acid-catalyzed second order rate constants of $\sim10^{-4}$ to 10^{-3} M^{-1} h^{-1}. The tertiary alcohols 2,3 dimethyl-2-butanol and amyl alcohol dehydrate much more rapidly, with an acid-catalyzed rate constants as large as 0.01–1 M^{-1} h^{-1}. The equation in Table 5 gives an acid-catalyzed rate constant of 0.009 M^{-1} h^{-1}, or a negligible rate of 9×10^{-10} h^{-1} at pH 7.

Weak oxidation of an alkene transforms the C=C bond into a 1,2 diol- the molecule loses a C=C, but gains two alcohols, 2 H atoms and 2 O atoms. The reaction may occur photochemically via reactive oxygen species (H_2O_2 or HO·) or by an enzymatic mechanism (Schwarzenbach et al. 2003). E_a for the enzymatic pathway has the typical value of 50 kJ/mole, which gives a pseudo-first-order rate constant of 0.00085 h^{-1} at 25 °C, 0.1 mM dissolved O_2 and $E_{ox} = 1$ for a molecule with 1 C=C bond. Assuming the same conditions and $I = 1.0\times10^{-6}$ mol h^{-1} cm^{-2}, the photochemical rate constant is 0.005 h^{-1}.

Strong oxidation of an alkene (oxidative cleavage) splits a C=C linkage into aldehydes or ketones, depending on the substituents of the original C=C; if the C=C is on a ring, the ring will be opened, otherwise, the molecule is split. In this version of Alphastep, the cleavage always creates aldehydes and the probability that a ring will be opened is the fraction of all C=C bonds that occur in phenyl rings for that molecule. Splitting into two successor molecules is handled in the same way as hydrolysis of esters.

The rate of thermal oxidation by O_2 in the absence of enzymatic catalysis is assumed to be negligible. This leaves two possible pathways – enzymatic catalysis by oxidases and photochemical oxidation via activated oxygen species. In each case, O_2 is assumed to be the ultimate source of the actual oxidant, which might be an activated oxygen species like H_2O_2 or hydroxyl radical, OH. The enzymatic pathway assumes an E_a of 50 kJ/mole, which gives a pseudo-first-order rate constant of 0.000172 h^{-1} at 25 °C, 0.1 mM dissolved O_2 and $E_{ox} = 1$ for a molecule with 1 C=C bond. If we assume that the reaction rate constant of a 'typical' molecule with 1 C=C bond with hydroxyl radical is ~0.1 h^{-1} in full sunlight (Schwarzenbach et al. 2003), and that 5% of these reactions lead to a 'strong oxidation' transformation, then the photochemical rate constant for this reaction would be 0.0010 h^{-1} in full sunlight.

Oxidation of an alcohol converts it into an aldehyde or ketone, removing 2 H atoms in the process. The program assumes a 60% probability of a primary alcohol, which oxidizes to an aldehyde, and a 40% probability of a secondary alcohol, which oxidizes to a ketone. Tertiary alcohols do not oxidize. This reaction has no un-catalyzed, thermal pathway, but can proceed either by photochemical or oxidative enzymatic mechanisms. As for alkene oxidation, the enzymatic E_a is 50 kJ/mole and O_2 is assumed to be the ultimate oxidant. If we assume $[O_2] = 0.1$ mM, $T = 25$ °C and oxidase activity of 0.1, a molecule with one alcohol group has an enzymatic $k' = 9.8 \times 10^{-5}$ h^{-1}, or a half-life of nearly 300 days. Note that highly oxidized molecules containing more oxygen than carbon will not further oxidize by this reaction.

Oxidation of an aldehyde inserts an O atom into the C–H bond, creating a carboxylic acid. As for the oxidation of an alcohol, this reaction requires O_2 as the ultimate oxidant and occurs either by a photochemical or enzymatic pathway (enzymatic an activation energy of 50 kJ per mole). For a molecule with a single aldehyde group in solution with 0.1 mM $[O_2]$ and $I = 1.0 \times 10^{-6}$ mol h^{-1} cm^{-2}, this gives a photochemical $k' = 0.001$ h^{-1}, or a half-life of nearly 1 month.

Decarboxylation removes a carboxylic acid group from a molecule, leaving only the H atom, for a net loss of CO_2. In this program decarboxylation is limited to aliphatic acids, and is favored by protonation (the $[H^+]/(10^{-4} + [H^+])$ term in the probability equation). Metal complexation to form a charge-transfer complex favors decarboxylation, but has not been implemented in the current version of the program. Decarboxylation proceeds only via photochemical and enzymatic pathways. The decarboxylase enzyme pathway is assigned an E_a of 50 kJ per mole. If we assume a decarboxylase activity of 1, then at pH 7 the enzymatic rate constant is 2×10^{-5} h^{-1}; however, at pH 4.0 this increases to 9.8×10^{-3} h^{-1}.

The reaction designated 'microbial utilization' is not a chemical reaction at all, but represents the uptake of a small molecule by an organism, here assumed to be bacteria. In this case, the entire molecule is removed from the simulation; no specific assumption is made about whether the molecule is immediately mineralized to CO_2, H_2O, etc. The program assumes that molecules with

MW > 1000 amu are too large to be transported across the membrane, and that smaller molecules are taken up more quickly than larger molecules (the 0.1–10^{-4} MW term in the k' equation). Furthermore, it assumes that for molecules of the same size, more oxygenated molecules and molecules which contain the nutrients N and P will be taken up more quickly (the Q term in the rate equation). For glucose ($C_6H_{12}O_6$, MW 180) in a system with high bacterial abundance ($B = 1$), this gives k' for microbial utilization of 0.082 h^{-1}, or a half-life of 8.5 h. In contrast, a hydrocarbon like toluene would not be utilized ($k' = 0$), although oxidation reactions could lead to products which were utilizable; thus a k' of zero should not be thought of as meaning that the C in that molecule is biologically unavailable in the long term.

Ester condensation combines an acid-containing molecule with an alcohol-containing molecule to form a single product molecule with an ester linkage (plus water). The product molecule has all the elemental and functional group composition of the two reactants summed plus one ester group, except for the deletion of one acid, one alcohol, 2 H atoms and one O atom. Typically acid catalysis is required for the reaction to proceed. Kirby (1972) reviews acid-catalyzed rate constants for a related reaction, oxygen exchange in substituted benzoic acids, with typical values of $\sim$1–2 $M^{-1}\ s^{-1}$. In a 10 μM alcohol solution, this suggests rates of $< 0.01\ h^{-1}$ at pH 0 and much lower at neutral pH values. Here, we assume $A = 5 \times 10^{10}\ h^{-1}$ and $E_a = 60$ kJ mol^{-1}, the latter a typical value of the ΔH of activation from Kirby (1972). At pH 7.0, this corresponds to a value of $1.5 \times 10^{-7}\ h^{-1}$, or a half-life of 527 years. At pH 4.0. however, we obtain $P_1 = 1.5 \times 10^{-4}\ h^{-1}$ and a half-life of 0.53 years. Ester condensation in dilute solution at neutral pH is clearly very slow- however, if molecules are brought into close proximity (high local concentration) by adsorption or (hemi-) micelle formation in an acid environment, the rate could become appreciable.

Aldol condensation combines one molecule with an aldehyde and a second with an aldehyde or ketone, forming a C–C linkage between the two and converting the aldehyde group into an alcohol. The atomic composition of the new molecule is simply the sum of the compositions of the reacting molecules, since the gain of an H by creating an alcohol is balanced by losing an H from the second molecule (the ‘enol’ H). The reaction is typically base catalyzed, although an acid-catalyzed pathway has been included here. For the reaction probability of the first molecule, the Arrhenius constant is $1 \times 10^{10}\ h^{-1}\ M^{-1}$ and $E_a = 50$ kJ per mole, while the second molecule selected must contain an aldehyde or ketone group.

Algorithm

Each Alphastep simulation consists of an initialization phase followed by a series of discrete time steps, terminating at a specified simulation time. Figure 3 shows the basic components of the algorithm.

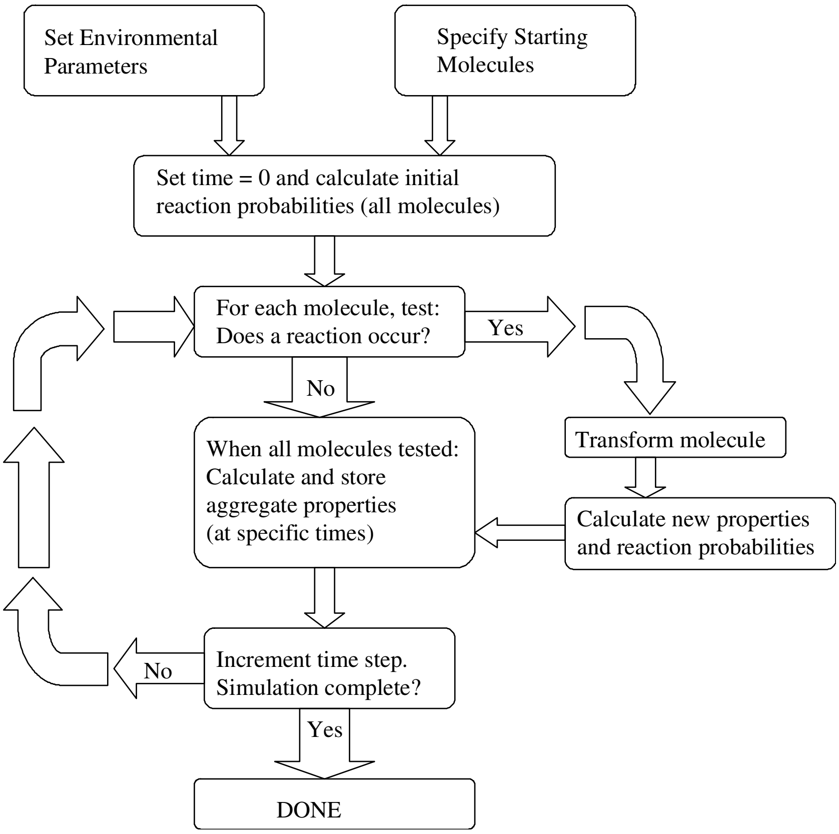

Set Environmental Parameters
Specify Starting Molecules
Set time = 0 and calculate initial reaction probabilities (all molecules)
For each molecule, test: Does a reaction occur?
Yes
No
Transform molecule
Calculate new properties and reaction probabilities
When all molecules tested: Calculate and store aggregate properties (at specific times)
Increment time step. Simulation complete?
No
Yes
DONE

Figure 3. Algorithm outline.

Initialization is performed interactively by setting simulation parameters for starting materials, environmental conditions and simulation controls. These may be done in any order and repeated or changed as necessary.

The user can specify any number of molecules of each starting material (precursor compound) from Table 1 up to 10,000 total molecules. However, due to memory limitations in the current version of AlphaStep the number of starting macromolecules- protein, cellulose, lignin- should be much smaller (a few hundred). On the other hand, the simulation can begin with several thousand of the smaller precursors- tannin, terpene, flavonoid.

The user characterizes the environment by specifying appropriate physical, chemical and biological parameters (Table 3). The default parameter values are each individually reasonable, but collectively highly reactive; high extracellular enzyme levels are probably not found in the same environments as high light intensity, for example.

Simulation control parameters allow the user to specify the duration of the simulation and the time step, both in hours. The default time step of 0.1 h should probably not be altered without a compelling reason, since decreasing it will slow down the simulation and increasing it may invalidate the 'small time' assumption for the fastest reactions. Simulation times up to 10^5 h (> 11 years) are possible. The frequency with which aggregate properties are calculated can be set in terms of the number of time steps, and with default values corresponds to every 50 h. Finally, the PRN generator can be set to begin each simulation with a specified 'seed'. This allows reproducible strings of PRNs to be used for testing, or for the string to be changed to look at the resulting variability between simulations. The random number generator has a 2^{32} cycle length.

Once the starting materials and parameters are specified, the main simulation loop depicted in Figure 3 is begun. The simulation time is set to zero and the aggregate properties (Table 2) of the NOM assemblage and the reaction probabilities of each molecule are calculated. Then for each time step, every molecule is tested for the possibility of a reaction by comparing a single PRN with the cumulative reaction probability, P (Xiang et al. In press). If the PRN is greater than P, no reaction occurs. If the PRN is smaller than P, then the number is compared with the individual reaction probabilities to determine which reaction occurs. If the reaction is second order, a second random number is used to determine which molecule, if any, the original molecule reacts with. The reacting molecule is transformed according to the reaction (Table 5) and then the individual properties and reactions probabilities of that molecule are re-calculated. When all molecules in the set have been tested, the time step is incremented, and aggregate properties are calculated if appropriate. This process is repeated until the simulation time reaches the desired duration.

Simulation results

System 1: hydrolysis and utilization of protein

The hydrolysis (proteolysis) and utilization of a protein is a relatively simple system with which to demonstrate the performance and reproducibility of the AlphaStep model. This simulation begins with 1000 protein molecules (described in Table 1) incubating in a dark aqueous solution at 25 °C and pH 7.0 in the presence of 0.10 mM dissolved $O_{2,}$ enzyme activities of 0.10 and bacterial abundance of 0.10. Each simulation of 1000 h of reaction time requires < 30 s of computer time on a 2 GHz Pentium 4 laptop; simulations were repeated with identical parameters but different random number seeds in order to establish standard deviations for the calculated results.

The simulation gives sensible results, as shown in Figures 4 and 5. The large protein molecules are quickly hydrolysed into smaller fragments which are then utilized by bacteria. The number average molecular weight drops quickly as the number of molecules increases, declining from > 5000 amu to < 1000 amu in

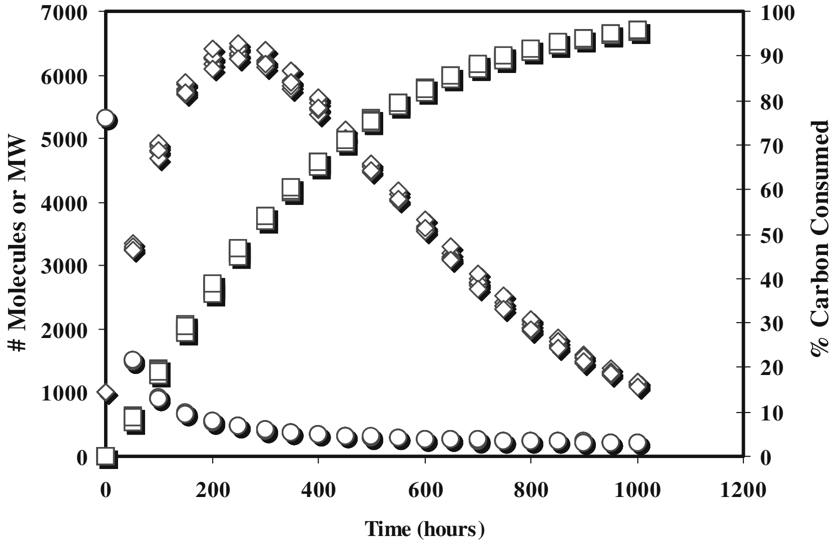


Figure 4. System 1: Protein hydrolysis and utilization. number of molecules (◇), number average molecular weight (○) and fraction of carbon utilized (□) as a function of time. Results are from three simulations with random number seeds 1, 2, and 3.

the first 100 h. M_n is slightly over 200 amu at 1000 h, indicating that nearly all amide linkages have been hydrolyzed and the remaining molecules are mostly amino acids and di-peptides (confirmed by inspecting individual molecule compositions). Utilization of the protein is ~50% (on a carbon basis) by 300 h, and >95% by 1000 h.

Comparing Figures 4 and 5 shows an advantage of monitoring both bulk properties and actual reaction rates, which are usually inferred from changes in concentration. The rate of carbon utilization in Figure 4 is not first or second order, but instead shows an increasing rate of utilization until ~200 h, at which time the rate peaks and begins to decline in an exponential fashion. The reaction count data in Figure 5 shows clearly that this 'lag phase' behavior is due to sequential reactions. The hydrolysis reaction is approximately first order in peptide linkages, as shown by the exponential decline of the reaction count data; the bacterial utilization cannot achieve maximum rate until a large number of small molecules have been created.

Note that this sort of general proteolysis would be extremely complex to model with deterministic differential equations: after one proteolysis reaction, each original molecule would produce two of nearly 100 possible peptides, and the second proteolysis would multiply the number of possible molecules by almost as much, so that after 10 proteolytic reactions the number of possible peptides would be staggeringly large.

Simulation results were generally consistent between runs with different PRN sequences. In a set of seven independent runs, most of the aggregate quantities had relative standard deviations (RSDs) of 2% or less, while the reaction count data had higher uncertainties (up to 15% in cases with at least 10 reactions per 50 h period). RSD was typically smallest for reactions with high counts and properties which averaged over the largest numbers of molecules, and were largest for very infrequent reactions or for assemblages with fewer molecules. Note the much smaller relative variation in the M_n and carbon utilization data in Figure 4 relative to the reaction count data in Figure 5. Run-to-run reproducibility will be better with larger numbers of molecules, but slower to calculate; simulating very small numbers of molecules is fast but can lead to highly irreproducible run-to-run variations (large RSD values).

System 2: condensation of small natural products in soil

The formation of humic-like materials in soil is more environmentally relevant than System 1, since it has been suggested that some humic substances, a major component of NOM, form by the condensation of small molecules in soils (Leenheer et al. 2003; Leenheer and Rostad 2004). Simulations began with 2000 molecules each of abietic acid, meta-digallic acid and fustin

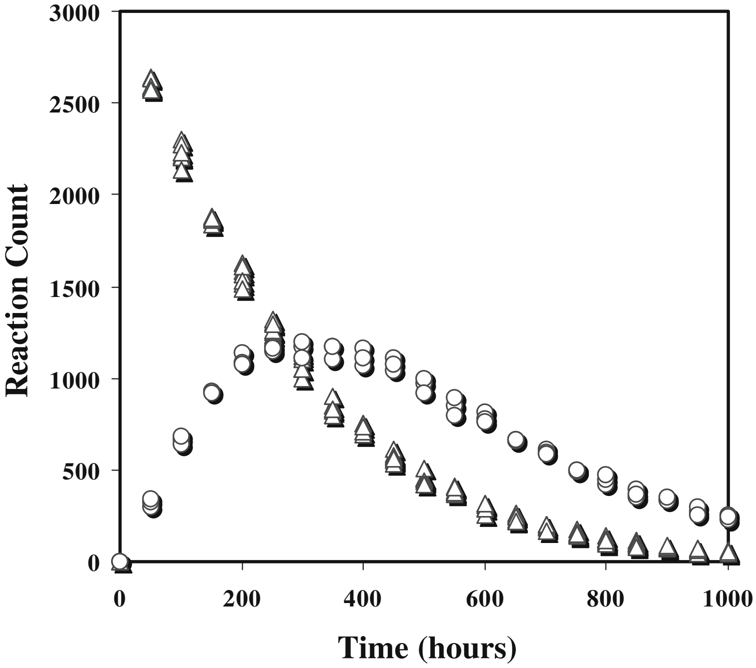


Figure 5. System 1: Protein hydrolysis and utilization. Numbers of hydrolysis reactions (△) and molecules utilized (○) per 50 h interval. Results are from three simulations with random number seeds 1, 2, and 3.

(Table 1, Figure 2) incubating in a dark, humid, well-oxygenated ($[O_2] = 0.30$ mM) soil at 25 °C and pH 5.0 with protease and oxidase activity of 0.10 and decarboxylase activity and bacterial abundance of 0.010. Simulation of 5000 h of reaction time requires < 90 s of computer time on a 2 GHz Pentium 4 laptop.

The simulated soil incubation produced an NOM assemblage at 5000 h with average composition quite different from the precursor. The average molecular weights doubled (M_n) or quadrupled (M_w) from ~300 amu into a range typical for NOM after 5000 h (Figure 6). On the other hand, the aromaticity (fraction of aromatic carbon) substantially decreased from 57% in the starting materials to ~11%, and the carboxylic acid content nearly doubled from 2.2 to 4.0 milliequivalents per gram (Figure 7). The carbon content dropped from 64% to 54% after 5000 h, while the oxygen content increased from 30 to 41% and the hydrogen content decreased slightly (5.7–5.4%).

Most of these changes could be attributed to enzymatic oxidation reactions (increased O content and acidity, decreased aromaticity) and microbial utilization (~2/3 by weight of the original carbon was utilized). The increase in molecular weights was due to a combination of condensation reactions and preferential microbial utilization of smaller molecules.

The NOM assemblage at 5000 h has aggregate properties similar to those of freshwater NOM (Table 6). All the aggregate parameters in Table 6 except for %N are within 2 standard deviations of the mean for experimental data (the precursor set contained no nitrogen), and most are within 1 standard deviation. The choice of 5000 h simulation time affects the similarity, of course; after 6000 h, the final assemblage would have M_n, M_w and acidity closer to the mean, but aromaticity would be further away. Since the current version of the

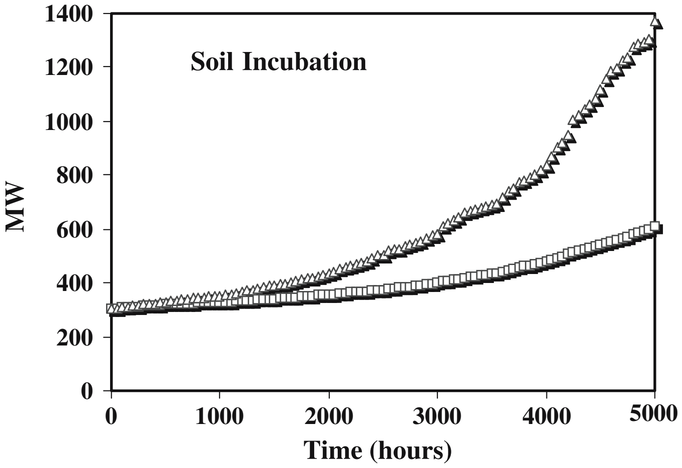


Figure 6. System 2: Soil incubation of small molecule precursors. Number average (triangles) and weight average (squares) molecular weight as a function of time.

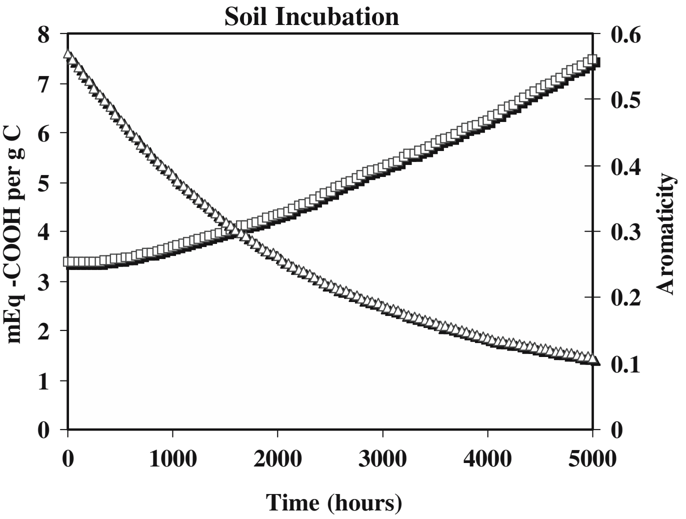


Figure 7. System 2: Soil incubation of small molecule precursors. Equivalent weight (□) and fraction aromatic carbon (△) as a function of time.

Table 6. Comparison of average properties of aquatic NOM: Field data versus simulation

Property	Literature[a]		Simulation results[b]	
	Range	Mean ± SD	Soil	Surface water
% Carbon	42–57%	49.5 ± 3.3%	54%	44%
% Oxygen	34–53%	43.0 ± 4.1%	41%	49%
% Hydrogen	3.6–7.9%	5.0 ± 1.0%	5.3%	5.1%
% Nitrogen	0.4–5.4%	1.7 ± 1.0%	–	2.3%
M_n (amu)	400–2700	1107 ± 471	612	717
M_w (amu)	784–3320	1684 ± 530	1374	1173
% Aromatic C	10–43%	27 ± 11%	11%	10%
mEq COOH per g	2.7–10.0	5.2 ± 2.0	4.0	1.8

[a] Aquatic NOM data (including humic and fulvic acids) reviewed in Perdue and Ritchie (2004). M_n and M_w, by size exclusion, acid content by direct titration, % aromatic C by ^{13}C NMR.
[b] Simulation results after 5000 h. Soil results are System 1, Surface water results are System 2.

program omits environmentally important processes like the continual addition of starting material and NOM transport, the results cannot be thought of as rigorous. However, since freshwater NOM is often assumed to have a terrestrial origin, the correspondence between simulated and measured values suggests that AlphaStep is a reasonable chemical simulation.

System 3: photodegradation

Another proposed mechanism for the production of humic substances is the light-mediated transformation of biopolymers in surface waters. In this case, the simulation was begun with 400 molecules each of the protein and lignin fragments described in Table 1 and Figure 1. The model surface water has a neutral pH (7.0), 0.30 mM dissolved O_2 and is kept at 25 °C. Light intensity was a constant 2.0×10^{-8} mol cm^{-2} h^{-1}. Protease, oxidase and decarboxylase activities are 0.010, as is the bacterial abundance. Simulation of 5000 h of reaction time requires ~90 s of computer time on a 2 GHz Pentium 4 laptop.

The simulated surface water incubation also produced an NOM assemblage at 5000 h with aggregate properties which resembled aquatic NOM but were quite different from the lignin and protein precursors. After 5000 h, average molecular weights dropped from nearly 6000 amu for the precursor compounds to M_n = 717 amu and M_w = 1173 amu, within the range typical for aquatic NOM and fulvic acids (Figure 8, Table 6). Aromaticity also decreased rapidly from 55% in the starting materials to ~10%, which would correspond to significant bleaching of the UV-VIS spectrum. The carboxylic acid content more than tripled from 0.5 to 1.8 milli-equivalents per gram, although this is still lower than observed for aquatic NOM samples (Table 6). The elemental composition after 5000 h of incubation is similar to an oxygenated aquatic NOM. Carbon content dropped from 65 to 44%, oxygen content increased from 21 to 49% and the hydrogen content decreased significantly from 6.7 to 5.1% (Figure 9). The starting material was 7.1% nitrogen, reflecting the large quantity of protein in the precursor mixture; it dropped to 2.3% after 5000 h, slightly higher than the mean for aquatic NOM.

These changes result from a combination of proteolysis, oxidation and microbial utilization. Proteolysis and oxidative cleavage are each responsible for about half of the new molecules split off from the biopolymers. Microbial utilization of small molecules is a relatively minor reaction for the first 1000 h, averaging < 0.5 utilization per hour, but increases to ~2 utilizations per hour as the molecules become smaller and more oxidized. Figure 10 shows the progressive oxidation reactions, principally of the lignin precursor-oxidation of

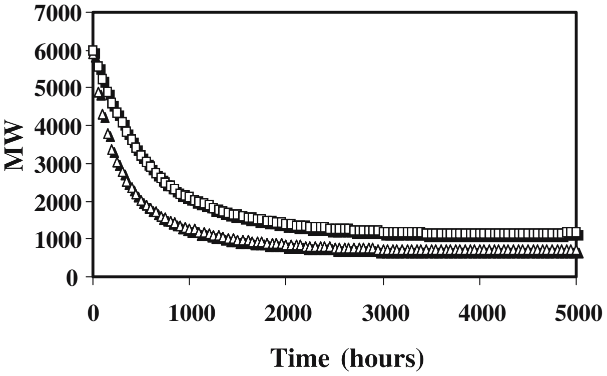


Figure 8. System 3: Surface water incubation of biopolymers: Number average (triangles) and weight average (squares) molecular weight as a function of time.

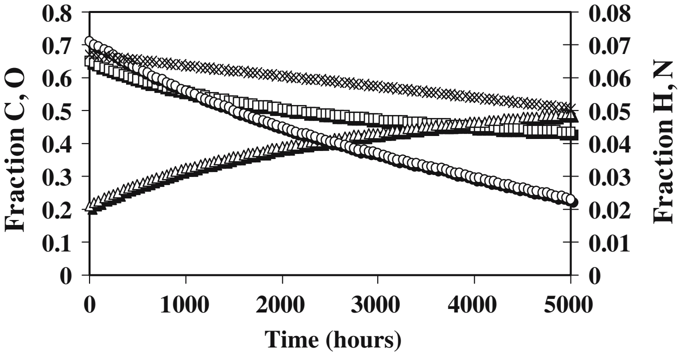


Figure 9. System 3: Surface water incubation of biopolymers: Fractional composition (by weight) of C (□), O (△), H (×) and. N (○).

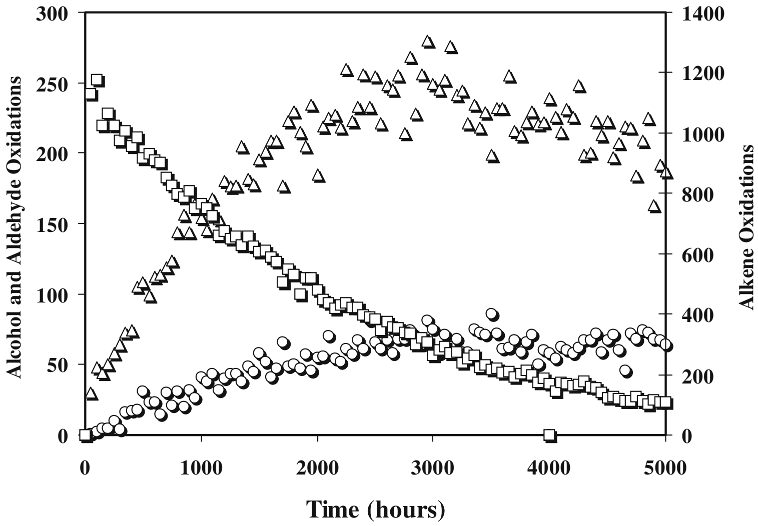


Figure 10. System 3: Surface water incubation of biopolymers: Reaction counts per 50 h interval for C=C oxidation to diol (□), alcohol oxidation to aldehyde or ketone (△), aldehyde oxidation to –COOH (○).

alkenes to diols is rapid initially, and as more alcohol groups are produced the oxidations to aldehydes/ketones and then oxidation of aldehydes to carboxylic acids become more frequent.

Overall, the NOM assemblage after 5000 h incubation has composition and properties similar to those of aquatic NOM (Table 6). The only property outside the range of reported values is the acidity, which is much lower than normal for aquatic NOM samples; simulated production of acid groups is apparently slow relative to the changes in MW, aromaticity and elemental composition.

Conclusions

The stochastic synthesis algorithm provides an acceptably fast and convenient model for chemical aspects of the production of NOM from naturally-occurring precursor molecules. Using only mechanistically sensible reactions and parameters, simple mixtures of natural products evolve into a complex assemblage of molecules with aggregate properties similar to NOM. The model can be used to provide chemical insight into possible transformations, and to test, in a simple way, hypotheses about origins of NOM.

The two environmental simulations presented here exemplify apparently contrasting theories of NOM formation – condensation and degradation. However, both the condensation of small tannin, terpenoid and flavonoid molecules in a model soil environment and the degradation of protein and lignin in a sunlit surface water produced molecular assemblages with bulk chemical properties similar to aquatic NOM.

The current version of the model oversimplifies or omits a number of environmentally important factors, including periodic (or continuous) precursor inputs, adsorption on or aggregation with solid surfaces (and the changes in reactivity this entails), environmental transport, abiotic and biotic reactions catalyzed by metal ions, specific enzyme production and decay rates, diurnal cycling and seasonal cycling of light and temperature, and feedback loops for environmental parameters, especially the microbial community. In addition, some of the current processes are overly simplistic- for example, the photolysis reactions have no wavelength dependence and do not consider the role of NOM in generating activated oxygen species. Nonetheless, the algorithm is useful for simple batch simulations, and with further enhancements should prove the basis for more realistic environmental simulations. The AlphaStep program and user manual can be downloaded from http://www.nd.edu/~nom/Software/software.html or obtained from the corresponding author.

Part II of this series will present the property distribution capabilities of AlphaStep, including the empirical algorithms which allow the program to estimate pK_a, polarity and Cu(II) binding constants for the NOM mixture.

Acknowledgements

Jerry Leenheer made key suggestions and comments on this work, especially regarding the inclusion of tannin, flavonoid and terpenoid precursor molecules. Funding was provided by the NSF Information Technology Research initiative and the Division of Environmental Biology via grant NSF DEB 0113570.

References

Aiken G.R., McKnight D.M., Wershaw R.L. and MacCarthy P. (eds) 1985. Humic Substances in Soil, Sediment, and Water – Geochemistry, Isolation, and Characterization. John Wiley & Sons, NY, pp. 692

Amon R.M.W. and Benner R. 1996. Bacterial utilization of different size classes of dissolved organic matter. Limnol. Oceanogr. 41: 41–51.

Avena M.J., Koopal L.K. and van Riemsdijk W.H. 1999. Proton binding to humic acids: electrostatic and intrinsic interactions. J. Colloid Interfac. Sci. 217: 37–48.

Bartschat T., Cabaniss S.E. and Morel F.M.M. 1992. An oligoelectrolyte model for cation binding by humic substances. Environ. Sci. Technol. 26: 284–294.

Bortiatynski J.M., Hatcher P.G. and Knicker H. 1996. NMR techniques (C, N, and H) in studies of humic substances. ACS Symposium Series 651: 57–77.

Boyd R.H., Taft R.W., Wolf A.P. and Christman D.R. 1960. Studies on the mechanism of olefin–alcohol interconversion The effect of acidity on the O^{18} exchange and dehydration rates of t-alcohols. J. Am. Chem. Soc. 82: 4729–4736.

Brown G.K., Cabaniss S.E., MacCarthy P. and Leenheer J.A. 1999. Cu(II) Binding by a pH-fractionated Fulvic Acid. Anal. Chim. Acta 402: 183–193.

Brown T.L. and Rice J.A. 2000. Effect of experimental parameters on the ESI FT–ICR mass spectrum of fulvic acid. Anal. Chem. 72: 384–390.

Cabaniss S.E. and Shuman M.S. 1988. Copper binding by dissolved organic matter I: Suwannee River fulvic acid equilibria. Geochim. Cosmochim. Acta 52: 185–193.

Chiang Y. and Kresge A.J. 1985. Mechanism of hydration of simple olefins in aqueous solution. J. Am. Chem. Soc. 107: 6363–6367.

Chin Y.P. and Gschwend P.M. 1991. The abundance, distribution, and configuration of porewater organic colloids in recent sediment. Geochim. Cosmochim. Acta 55: 1309–1317.

Chin Y.P., Aiken G.R. and Danielsen K.M. 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 C.T., Malcolm R.T., Brinton T.I. and Kile D.E. 1986. Water solubility enhancement of some organic pollutants and pesticides by dissolved humic and fulvic acids. Environ. Sci. Technol. 20: 502–508.

Chorover J. and Amistadi M.K. 2001. Reaction of forest floor organic matter at goethite, birnessite and smectite surfaces. Geochim. Cosmochim Acta 65: 95–109.

Cook R.L., McIntyre D.D., Langford C.H. and Vogel H.J. 2003. A comprehensive liquid-state heteronuclear and multidimensional NMR study of Laurentian fulvic acid. Environ. Sci. Technol. 37: 3935–3944.

Erdi P. and Toth J. 1989. Mathematical Models of Chemical Reactions. Princeton University Press, Princeton, NJ, pp. 259

Esham E.C., Ye W.Y. and Moran M.A. 2000. Identification and characterization of humic substances – degrading bacterial isolates from an estuarine environment. FEMS Microbiol. Ecol. 34: 103–111.

Findley S.E.G. and Sinsabaugh R.L. 2003. Aquatic Ecosystems Interactivity of Dissolved Organic Matter. Academic Press, San Diego, USA, pp. 512

Gillespie D.T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22: 403–4.

Gu B., Schmitt J., Chen Z., Liang L. and McCarthy J.F. 1995. Adsorption and desorption of different organic matter fractions on iron oxide: mechanisms and models. Environ. Sci. echnol. 28: 38–46.

Gu B., Mehlhorn T.L., Liang L. and McCarthy J.F. 1996. Competitive adsorption, displacement, and transport of organic matter on iron oxide: I. Competitive adsorption. Geochim. Cosmochim. Acta 60: 1943–1950.

Hansen S., Jensen H.E. and Nielsen N.E. 1990. Daisy – A Soil Plant Atmosphere System Model. Po-research from the National Agency of Environmental Protection No.A10, Denmark.

Hessen D.O. and Tranvik L.J. 1998. Aquatic Humic Substances: Ecology and Biogeochemistry. Springer-Verlag, Berlin, pp. 346

Hoigné J. 1990. Formulation and calibration of environmental reaction kinetics; oxidations by aqueous photooxidants as an example. In: Stumm W. (ed.), Aquatic Chemical Kinetics. John Wiley & Sons, Inc., pp.43–70.

Huang Y., Xiang X., Madey G. and Cabaniss S. 2005. Agent-based scientific simulation using Java/Swarm, J2EE, RDBMS and autonomic management technologies. Comput. Sci. Eng. 7: 22–29.

Kirby A.J. 1972. Hydrolysis and formation of esters of organic acids. In: Bamford C.H. and Tipper C.F.H. (eds), Comprehensive Chemical Kinetics (10) Ester Formation and Hydrolysis and Related Reactions. Elsevier Publ, NY pp. 57–208.

Klaus U., Pfeifer T. and Spiteller M. 2000. APCI-MS/MS: A powerful tool for the analysis of bound residues resulting from the interaction of pesticides with DOM and humic substances. Environ. Sci. Technol. 34: 3514–20.

Leenheer J.A., Nanny M.A. and McIntyre C. 2003. Terpenoids as major precursors of dissolved organic matter in landfill leachates, surface water, and groundwater. Environ. Sci. Technol. 37: 2323–2331.

Leenheer J.A. and Rostad C. 2004. Tannins and terpenoids as major precursors of Suwannee River fulvic acid. USGS Scientific Investigation Report 2004–5276. US Dept. of the Interior, Washington DC, pp. 16

Legovic T. 2001. 2001) WWW Server for Ecological Modeling. http://eco.wiz.uni-kassel.de/mod-info/index.html.

Mabey W. and Mill T. 1978. Critical review of hydrolysis of organic compounds in water under environmental conditions. J. Phys. Ref. Data 7: 383–415.

Michalzik B., Tipping E., Mulder J., Gallardo Lancho J.F., Matzner E., Bryant C.L., Clarke N., Lofts S. and Vincente Esteban M.A. 2003. Modelling the production and transport of dissolved organic carbon in forest soils. Biogeochemistry 66: 241–264.

Moran M.A. and Zepp R.G. 1997. Role of photoreactions in the formation of biologically labile compounds from dissolved organic matter. Limnol. Oceanogr. 42: 1307–1316.

Morton-Firth C.J. 1998. Stochastic Simulation of Cell Signaling. Pathways Doctoral thesis, Cambridge University, Cambridge, UKCB2 3EJ

Namjesnik-Dejanovic K., Maurice P.A., Aiken G.R., Cabaniss S.E., Chin Y.P. and Pullin M.J. 2000. 2000) Adsorption and fractionation of a muck fulvic acid on kaolinite and goethite at pH 3.7, 6 and 8. Soil Sci. 165: 545–559.

Namjesnik-Dejanovic K. and Cabaniss S.E. 2004. Reverse-phase HPLC method for measuring polarity distributions of natural organic matter. Environ. Sci. Technol. 38: 1108–1114.

Noyce D.S. and Lance C.A. 1962. The kinetics and mechanism of β-phenyl β-hydroxy propionic acid. J. Am. Chem. Soc. 84: 1635–1638.

Parunak H.V.D., Savit R. and Riolo L. 1998. Agent-based modeling vs. equation-based modeling: A case study and user's guide. Proceedings of Multi-agent systems and Agent-based Simulation. Gilbert N. (ed.), Springer-Verlag, Paris, pp. 10–25

Perdue E.M., Reuter J.H. and Parrish R.S. 1984. A statistical model of proton binding by humus Geochim. Cosmochim. Acta 48: 1257–1263.

Perdue E.M. and Ritchie J.D. 2004. Dissolved Organic Matter in Freshwaters in Treatise on Geochemistry vol. 5. Drever J.I. (ed.). , pp. 273–318

Robinson T. 1983. The Organic Constituents of Higher Plants. Cordus Press, North Amherst, MA, pp. 353

Saleh F.Y., Ong W.C.A. and Chang D.Y. 1989. Structural features of aquatic fulvic acids- analytical and preparative reversed phase HPLC separation with photodiode array detection. Anal. Chem. 61: 2792–2800.

Schmitt-Kopplin P., Garrison A.W., Perdue E.M., Freitag D. and Kettrup A. 1998. Capillary electrophoresis in the analysis of humic substances – facts and artifacts. J. Chromat. A 807: 101–109.

Schnitzer M. and Khan S.U. 1972. Humic Substances in the Environment. Marcel Dekker, NY, pp. 327

Schubert M.W. and Lamm B. 1966. The acid-catalyzed hydration of styrene. J. Am. Chem. Soc. 88: 120–124.

Schulten H.R. and Leinweber P. 2000. New insights into organic – mineral particles: composition, properties and models of molecular structure. Biol. Fert. Soils 30: 399–432.

Schwarzenbach R.P., Gschwend P.M. and Imboden D.M. 2003. Environmental Organic Chemistry. Wiley-Interscience, Hoboken, NJ, pp. 1313

Scott D.T., McKnight D.M., Blunt-Harris E.L., Kolesar S.E. and Lovley D.R. 1998. Quinone moieties act as electron acceptors in the reduction of humic substances by humics-reducing microorganisms. Environ. Sci. Technol. 32: 2984–2989.

Shimizu T.S. and Bray D. 2001. Computational cell biology – the stochastic approach. In: Kitano H. (ed.), Foundations of Systems Biology. MIT Press, Cambridge, MA, pp. 213–232.

Talbot R.J.E. 1972. The hydrolysis of carboxylic acid derivatives. In: Bamford C.H. and Tipper C.F.H. (eds), Comprehensive Chemical Kinetics (10) Ester Formation and Hydrolysis and related Reactions. Elsevier Publ, NY, pp. 209–294.

Vaughan D.J. and Blough N.V. 1998. Photochemical formation of hydroxyl radical by constituents of natural waters. Environ. Sci. Technol. 32: 2947–2953.

Wetzel R.G., Hatcher P.G. and Bianchi T.S. 1995. Natural photolysis by ultraviolet irradiance of recalcitrant dissolved organic matter to simple substrates for rapid bacterial metabolism. Limnol. Oceanogr. 40: 1369–1380.

Williams J.R. 1985. The EPIC model – an overview. In: DeCoursey D.G. (ed.), Natural Resources Modeling Symp. Pingree Park, CO.October 16–21, 1983, USDA, ARS, ARS –30.

Xiang X., Huang Y., Madey G., Cabaniss S.E., Arthurs L. and Maurice P.A. 1990. Modelling the evolution of natural organic matter with agent-based stochastic approach. Nat. Res. Model, in press.

Xue H.B. and Sigg L.L. 1993. Free cupric ion concentration and Cu(II) speciation in a eutrophic Lake. Limnol. Oceanog. 38: 1200–1213.

Zafiriou O.C., Joussot-Dubien J., Zepp R.G. and Zika R.G. 1984. Photochemistry of natural waters. Environ. Sci. Technol. 18: 358A–371A.

Zhou Q., Maurice P.A. and Cabaniss S.E. 2001. Size fractionation upon adsorption of fulvic acid to goethite: equilibrium and kinetic studies. Geochim. Cosmochim. Acta 65: 803–812.