

# AlphaStep User Guide

Version 0.3, October, 2003

## I. Introduction

AlphaStep is a Windows-based program which simulates the environmental transformation of biological molecules into NOM (Natural Organic Matter, sometimes referred to as humic substances) and the eventual consumption/destruction of that NOM. AlphaStep simulates a variety of chemical and biological transformations in a well-mixed reactor, but does not simulate any type of transport; it can be thought of as one component of a more comprehensive model, under development, which will include both transport and biological community terms. AlphaStep is coded in Delphi 6 and runs under Windows XP. Its development has been supported by a grant from the National Science Foundation ().

AlphaStep allows the user to control the input biomolecules (number and structure) and a variety of chemical (pH, dissolved oxygen), physical (light intensity, temperature) and biological (enzyme activities, microbial consumption) parameters. The program simulates chemical and biological transformations using a stochastic algorithm, and then allows the user to examine changes in aggregate parameters (elemental composition, molecular weight, percent aromaticity) of the entire mixture of molecules with time. The user may also examine the properties of individual molecules at the end of the simulation.

AlphaStep is intended to be useful to scientists and engineers interested in NOM for theoretical or practical reasons. However, it is more useful for asking and testing qualitative questions about transformation mechanisms than for making quantitative predictions about specific environmental rates. Possible questions to explore include

*How does a change in pH affect the molecular weight of NOM over a period of weeks or months?*

*Is it possible for NOM to form by condensation of small biomolecules? By degradation of large biomolecules?*

*Can we expect different elemental composition of NOM from wet and dry soils? Different molecular weights?*

AlphaStep is intended partly as a stand-alone application to allow ecologists, geochemists and environmental scientists to explore possible chemical routes of NOM transformation. However, its principal function is to allow testing of the biogeochemical reactions in a simple form (i.e., without spatial data) and provide feedback which will improve the more general, spatially-aware model. *Please* send your comments, suggestions, or strange-looking results to Steve Cabaniss [cabaniss@unm.edu](mailto:cabaniss@unm.edu). As a user, your input is needed to 'evolve' and improve the program codes.

## II. General algorithm description

AlphaStep simulates biogeochemical transformations of organic molecules using a stochastic, agent-based approach. The user defines a suite of input biomolecules from a list of possibilities which includes lignin, cellulose, protein, and smaller natural products. Although the program can handle up to 10,000 molecules, a substantially smaller number (~1,000) is recommended. The user also must define the physical, chemical and biological conditions which will govern the rates of reaction- these include pH, isolation, temperature, enzyme activities and microbial consumption level. The stochastic reaction algorithm treats each molecule separately, calculating probabilities of various reactions from the molecular composition and prevailing conditions. As molecules react, their composition and properties change in a chemically sensible way. That is, each molecule may undergo only specified transformations which result in 'legal' organic molecular structures. However, the structures of the newly created molecules are not restricted in any other way- these resulting molecules have no pre-determined structure and may be wholly unanticipated by the user. Properties of the resulting molecules- both individual and aggregate- may be displayed and inspected.

Each molecule is represented as an individual entity, or agent, with particular composition and properties. The composition of each molecule includes elemental abundance (i.e., the molecular formula) and the abundance of each functional group. No connectivity data are included, so that the program cannot consider branching or proximity effects. Molecular properties are calculated from this composition data. Some of these calculations are straightforward (molecular weight, acid content, percent aromatic carbon) while others rely on empirical correlations ( $K_{ow}$ ) or assumptions (microbial 'availability').

### III. User controllable values

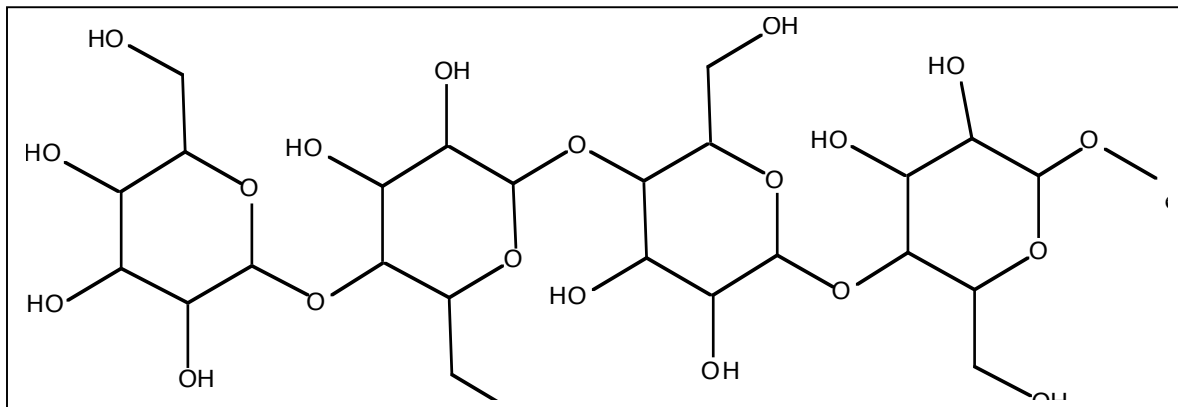
Parameters which control the simulation can be changed (within certain bounds) through the tabbed parameter fields in the upper left-hand corner of the main screen. Four tabbed pages are available, one each for the input molecules, physical/chemical parameters, biological parameters and the simulation control parameters. You can change the value of a given parameter by selecting the appropriate tabbed page, placing your cursor in the value display, and editing or replacing the given value. Your new value is registered when the cursor/mouse leaves the edit field; at that point, if the program detects an illegal (not-a-number) or out-of-bounds (for example, negative light intensity) value, it will refuse to accept it and simply re-display the old value.

#### A. Input molecules

The simulation provides six types of input molecules, which can be combined in any desired ratio. Although the program will accept up to 10,000 input molecules, I recommend beginning with 1000 or less. Note that each molecule is defined by its elemental composition and functional group content, not by connectivity. Thus, it cannot distinguish between 1-butanol and 2-butanol, or between catechol (1,2 dihydroxybenzene) and 1,4 dihydroxybenzene. It can distinguish between 1-butanol and diethyl ether because the former is an alcohol, the latter an ether.

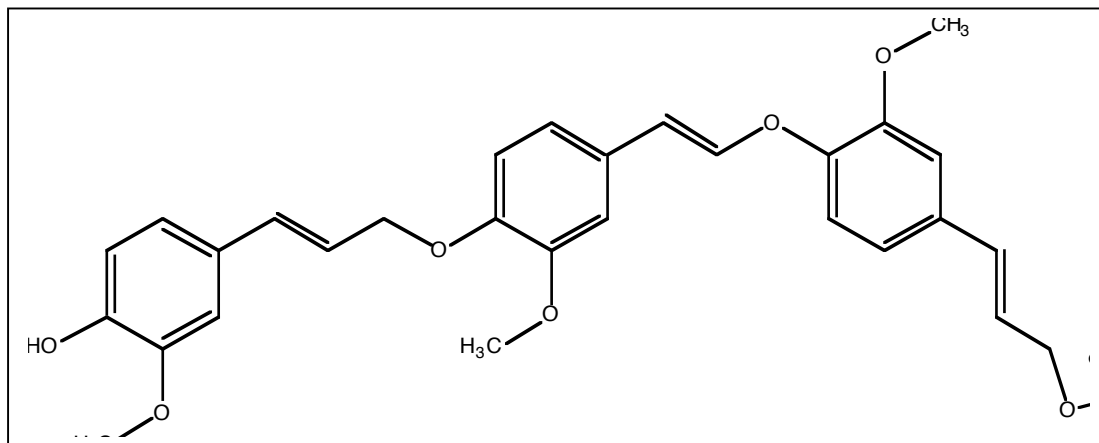
The available input molecules were selected to represent average input 'types', and have 'average' or representative properties, rather than necessarily corresponding to a specific natural product. Currently the program provides three biopolymers (cellulose, lignin and protein) and three smaller natural products (terpene, tannin, flavonoid)

**Cellulose** The cellulose molecule in this program is 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 overall elemental composition is  $C_{360}H_{602}O_{301}$ , and the molecule contains 60 rings, 182 alcohols and 119 ether linkages.





**Lignin** The average lignin molecule in this program is an oligomer of 40 coniferyl alcohol units condensed together via ether linkages (implying dehydration). Actual lignin molecules are larger and insoluble; the average molecule described here is a smaller soluble fragment, presumably lysed from the parent molecule enzymatically. The overall elemental composition is  $C_{400}H_{402}O_{81}$ , and the molecule contains 40 phenyl rings, 79 ether linkages, 1 alkyl alcohol and 1 phenol.

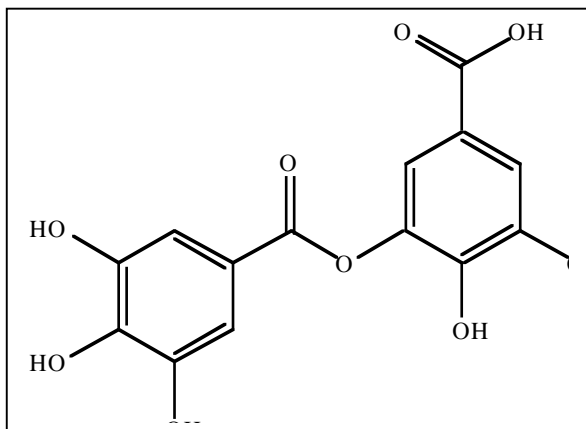


**Protein** The average protein molecule in this program is a peptide consisting of 50 residues, 5 each of the following:

Amino acid	R group
Glutamic acid	$CH_2-CH_2-COOH$
Lysine	$CH_2-CH_2-CH_2-CH_2-NH_2$
Glutamine	$CH_2-CH_2-CONH_2$
Serine	$CH_2-OH$
Threonine	$CHOH-CH_3$
Glycine	H
Alanine	$CH_3$
Valine	$CH-(CH_3)_2$
Leucine	$CH_2-CH-(CH_3)_2$
Phenylalanine	$CH_2-C_6H_5$

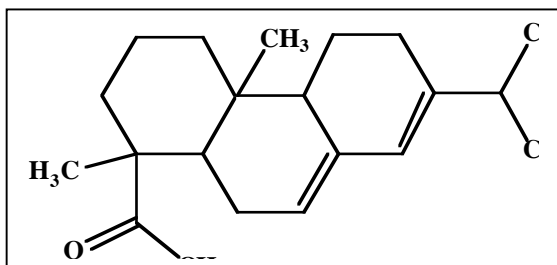
This protein molecules has molecular formula  $C_{240}H_{382}O_{76}N_{60}$  and contains 54 amides, 6 carboxylic acids, 6 amines, 10 alcohols, and 5 phenyl rings.

**Terpenoid** The terpenoid molecule in this program is abietic acid, a slightly oxidized diterpenoid (4 isoprene residues). It has molecular formula  $C_{20}H_{30}O_2$ , and the molecule contains 3 rings, 2 C=C bonds and 1 carboxylic acid group. Diterpenoids like this are produced by various pine species.



Abietic acid,  $C_{20}H_{30}O_2$

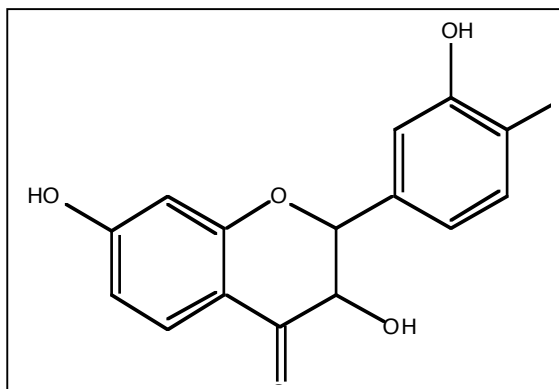
**Flavonoid** The flavonoid molecule in this program is fustin, a flavanone with molecular formula  $C_{15}H_{12}O_6$ . The molecule contains 2 phenyl rings, three phenols, an alky alcohol, an ether and a ketone. Fustin is a yellow plant pigment which is more water soluble than most terpenoids.



Fustin, a flavanone

**Tannin** The tannin molecule in this program is meta-digallic acid, a hydrolyzable tannin which is more water-soluble than either abietic acid or fustin. It has molecular formula  $C_{14}H_{10}O_9$ , and the molecule contains 2 phenyl rings, 5 phenols, 1 ester and 1 aromatic carboxylic acid group.

Meta-digallic acid



## *B. Physical and chemical parameters*

AlphaStep allows you to control several aspects of the simulation's 'environment'. Some of these parameters, like temperature, will affect the rate of nearly every reaction (although not to the same extent). Others, like the concentration of dissolved O<sub>2</sub> and the light intensity, will affect only certain reactions- e.g., oxidation reaction rates are enhanced by dissolved O<sub>2</sub>.

**Temperature:** The temperature can be set to any value between 0 and 100 degrees Celsius (although anything above 50 is quite unrealistic). Rates of 'thermal' (i.e., non-photochemical) reactions are quite temperature-sensitive, and the 'rule of thumb' for many reactions is to expect a doubling of rate for each 10 degree increase in temperature. This rule assumes an activation energy of ~50 kJ/ mole, and the thermal reactions in AlphaStep have activation energies ranging from 50 kJ/mole (enzymatic oxidation reactions) to 80 kJ/mole (hydration and dehydration). The latter will thus be somewhat more temperature sensitive, while photochemical reaction rates are independent of temperature.

**pH:** The pH can be set to any value between 0 and 14. Note that even in relatively dry soil environments, the pH of the soil moisture is a relevant parameter. All of the hydrolysis/condensation and hydration/dehydration reactions are acid-catalyzed, and the hydrolysis reactions are base catalyzed as well.

**Water activity:** The activity of water ranges from 0 (completely dry) to 1 (aqueous solution). Water is required for hydrolysis (amide and ester) and hydration reactions, which cannot proceed if this parameter is set to 0. The rates of these reactions will be slow if the water activity is low (for example, in a desert soil we might expect a water activity of 0.0001).

**Light intensity:** The light intensity can vary from 0 (darkness) to 0.01 Einsteins. The default value of 0.0001 corresponds to full sunlight at mid-latitudes. Increasing the light intensity will increase the rate of all oxidation reactions and decarboxylation. If no enzymatic reaction is possible, this increase will be simple proportionality.

**Dissolved O<sub>2</sub>:** The dissolved oxygen concentration can be set to any value between 0.0 (completely anoxic) and 4.0 mM. Note that the default value of 0.1 mM corresponds to 3.2 mg O<sub>2</sub> per liter which is below the value expected for equilibrium with atmospheric O<sub>2</sub>, which is closer to 0.3 mM for 20 degrees Celsius. Dissolved O<sub>2</sub> enhances the rates of all oxidation reactions (oxidation of alkenes, alcohols and aldehydes), which cannot proceed in an anoxic environment.

## *C. Biological parameters*

The biological parameters, unlike their physical and chemical counterparts, are unitless 'activities' ranging in value from 1 to 0.

## Bacterial Density

**Protease:** This parameter represents the activity of all enzymes which cleave the amide linkage (peptide bond) in proteins and smaller peptides. When set to a maximum value of 1 in a room temperature aqueous solution, enzymatic cleavage of an amide linkage has a rate constant of about  $0.01 \text{ hr}^{-1}$ , suggesting that a protein molecule with 50 amide bonds should react within an hour or two, but complete cleavage into individual amino acids would require a week or more. Note that water is required for proper enzyme function- no reaction is possible if the water activity is zero.

**Oxidase:** This parameter represents the activity of all enzymes which promote oxidation. This 'lumps together' quite a few different enzymes, since the oxidative cleavage of an alkene to form aldehydes requires very different enzymes from the oxidation of an aldehyde to a carboxylic acid. Oxidase activity works together with dissolved oxygen to oxidize C=C double bonds (mild oxidation to a diol or complete bond cleavage to a pair of aldehydes), alcohols (to aldehydes or ketones) and aldehydes (to carboxylic acids).

## Decarboxylase

*D. Simulation control*

## IV. User actions

AlphaStep currently supports six user actions, each of which corresponds to clicking a button on the main program screen.

### A. Create New

Pressing this button clears out the existing set of molecules and replaces it with a set defined on the “Input Molecules” tab. No data from previous calculations will be saved!

### B. Simulate

Pressing this button runs a simulation with the existing set of molecules and the parameters given in the various parameter tab fields (Physical/Chemical, Biological and Batch). Note that it does *not* replace the current set of molecules with a new ‘input’ set, so if you have just run a simulation then pressing this button effectively *continues* the same simulation. Also, note that the simulation cannot be interrupted, and no intermediate data are displayed, so you will experience an apparent program ‘pause’ when you hit the button- please be patient. The data grid will be re-written when the simulation is complete.

### C. Inspect

Pressing this button opens the Molecular Inspector window. While this window is open, the user may examine the composition and calculated properties of any molecule in the data set. Molecules and their values may not be changed or deleted, however. The main program window is inactivated until the user closes the Molecular Inspector window.

### D. Default

Pressing this button returns all user controlled parameters to their default values.  
simulation

Simulation Duration:	1000 hrs or ~42 days
Time step (DeltaT):	0.1 hour or 6 minutes

Physical and chemical parameters:

Temperature:	298 Kelvins or 25 °C
pH:	7, neutral
Water activity:	1 (assume aqueous solution)
Light intensity:	0.0001 Einsteins
Dissolved O <sub>2</sub> :	0.1 millimolar

Biological parameters:

(Note that unlike the chemical and physical parameters, these values have no units but are ‘normalized’ to a value of 1 for a highly active system.)

Microbial “density”	1
Protease activity	1
Oxidase activity	1

Decarboxylase activity

1

*E. Exit*

Pressing this button exits the program. IMPORTANT NOTE: No data is saved, so any unsaved results are permanently lost.

## V. Simulation Results

### A. Individual molecule

To see the composition or calculated properties of individual molecules, activate the Molecule Inspector window by clicking the 'Inspect' button. The Molecular Inspector window has four action buttons and four sets of informational fields, two for structural parameters and two for calculated properties. Note that individual molecule data is available only before or after, *not during*, a simulation.

The action buttons control which molecule is displayed, are quite simple. Each molecule has a number within the data set which is displayed near the top of the inspector window. The back and forward buttons simply look for the molecule with the next lower or higher number, respectively. The Go To button allows you to specify the number of the molecule you want to see (if no molecule exists with that number, the display does not change). Finally, the Exit button closes the Molecular Inspector window and returns you to the main program window.

The first set of informational fields is "Elemental Composition", and shows the number of atoms of each element (C,H,O,N,S,P) in the molecule. This is the fundamental structural composition of the molecule.

The second set of informational fields is the "Functional Groups", and shows the count for each type of functional group (alcohol, acid, C=C double bond, etc.). Functional groups are strongly related to molecular behavior and reactivity. Recall that dimethyl ether (volatile, toxic solvent) and ethanol (inebriating liquid) have the same elemental composition, C<sub>2</sub>H<sub>6</sub>O, arrange to form different functional groups.

The third set of fields is "Molecular Properties", and shows some of the calculated properties of the molecule, including molecular weight, charge at pH 7, percent "aromaticity", and various thermodynamic values (not yet implemented). Each of these calculated properties could be shown as a distribution for the entire ensemble or averaged (in some way) to produce a single aggregate value. For example, the weight-average and number-average molecular weights are both calculated from the molecular weights of the individual molecules.

The fourth set of fields is "Reaction Probabilities", which shows the overall probability of each type of chemical transformation in the program. The probabilities are those calculated under the most recent set of physical, chemical and biological parameters, and are not simply a property of the molecule (unlike the previous set, which contains values which are independent of these parameters).

### B. Aggregate properties

Aggregate properties are calculated for the entire set of molecules in memory. This can have the effect of 'averaging' over many molecules (and thus losing information), but these average properties are often those which can be measured experimentally. Aggregate properties are calculated at specific sampling points throughout a simulation, and the time-dependent values are recorded in the data grid at the bottom of the main program window. The data grid is

scrollable in both directions to allow visual inspection of the data, but cut-and-paste to the windows clipboard has not been implemented (yet).

The aggregate properties currently include

- N Molecule # of molecules in the simulation. This will change as molecules condense, split, or are consumed.
- MW<sub>n</sub> The number average molecular weight, calculated by dividing the total weight of all molecules by the number of molecules. This is the average molecular weight which can be experimentally determined from colligative properties (osmotic pressure, freezing point depression, etc.) or by SE-HPLC.
- MW<sub>w</sub> The weight average molecular weight, calculated by dividing the sum of the squares of all molecular weights by the total molecular weight. Note that this will only equal the number average molecular weight if all molecules have the same molecular weight- otherwise, MW<sub>n</sub> < MW<sub>w</sub>.
- Z average The average charge on each molecule at pH 7, calculated by summing the number of amine groups (positive charges) and subtracting the sum of the carboxylic acid groups (negative charges and dividing by the number of molecules.
- Equiv. Wt. The number of carboxylic acid groups in all molecules divided by the total mass of all the molecules. This is essentially the 'mass per acid group', and is typically in the range 200-500 amu for NOM.
- % Aromatic This *should* correspond to <sup>13</sup>C nmr measurements of NOM 'aromaticity'. The number of C=C double bonds is multiplied by 2, then divided by the total number of C atoms in the data set.
- %Element %C, %H, %O, etc. are simply the weight percentages of each element in the data set.

Aggregate properties can be plotted versus time on the graph at the upper right of the main program window. Use the selection box immediately under the graph to choose which variable you would like to plot. The graph has some limited zoom capability, but no printing capability.

The aggregate properties may be saved to a text file using the **Save** button on the right-hand side of the data grid. After pressing the button, you will be prompted to supply a file name, and if a name is given the most recent results (currently in the grid) are written to that file. ***This will overwrite any existing file of that name***, so be careful! The data are written as simple ASCII text, with the variable names in the first line and the time-dependent property data written in decimal format, one row per sampling time, below. This format should be easily readable by a spreadsheet program, but is not properly spaced for easy reading by the user.

## **VI. Tutorial**