Supporting Information

Mamadou S. Diallo, Andre Simpson, Paul Gassman, Jean Loup Faulon,
James H. Johnson, Jr , William A. Goddard III, and Patrick G Hatcher

3-D Structural Modeling of Humic Acids Through Experimental Characterization, Computer Assisted Structure Elucidation and Atomistic Simulations. 1. Chelsea Soil Humic Acid

Environmental Science and Technology

I. Experimental Methods and Procedures

I. 1. Humic Acid Extraction

Chelsea soil humic acid (HA) extraction was based on a standard procedure developed by the International Humic Substances Society (1). Briefly, aliquots of 10 g of Houghton muck were mixed overnight with 200 mL of 0. 1 N NaOH in 250 mL poly(propylene) centrifuge tubes under continuous stirring. Following centrifugation for 30 minutes at 5000 rpm, the supernatant of each centrifuge tube was decanted and stored at 4 ºC. The combined supernatants were subsequently mixed with concentrated HCl (adjusted to solution pH of 1.0) for 12 hours. After centrifugation, the supernatants were discarded and each residual Chelsea soil HA precipitate was collected and stored at 4 ºC. This dissolution-precipitation cycle was repeated three times. After the fourth dissolution step, the Chelsea soil HA-NaOH extracts were filtered (twice) through a 0.22 μm polyethersulfone membrane filter under N_2 pressure. The filtrates were then acidified to pH 1.0 and centrifuged. The residual Chelsea soil HA extracts were freeze dried and stored in a glass bottle. All solutions used during the extraction procedures were prepared with N_2 saturated distilled and deonized water and only exposed to N_2 gas head space. The extracted Chelsea soil HA samples were characterized by elemental analysis (2), diffuse reflectance FT-IR spectroscopy, 1-D ^{1}H and ^{13}C solution NMR spectroscopy, 2-D solution NMR (TOCSY and HMQC) spectroscopy and electrospray ionization (ESI) quadrupole time-of-flight (Qq-TOF) mass spectrometry.

I. 2. Diffuse Reflectance FT-IR Spectroscopy

Aliquots of Chelsea soil HA were placed over P_2O_5 at 45 ºC overnight to remove excess moisture. A sample of 30 mg of the dried HA sample was mixed with 400 mg of spectroscopic grade KBr using a Wig-L-Bug amalgamator. The mixture and reference sample were, respectively, loaded into the sample and reference cups of a diffuse reflectance cell of a Bruker FT-IR (IFS66v/S) spectrometer. After evacuation to 13 millitorr, high resolution (4 cm^{-1}) reflectance spectra for Chelsea soil HA and the reference KBr sample were acquired over the mid infrared region (4000-400 cm^{-1}) using a KBr beamsplitter and a mercury cadmium telluride (MCT) detector with a velocity of 40 Hz. After background subtraction, the sample reflectance spectrum was converted to absorbance using the Kubelka-Munk transform (KKT). The frequencies of peak absorbances were determined by picking the peak centroid frequency, whereas those for the shoulders were picked slightly away from the overlapping peaks.

I. 3. 1-D and 2-D Solution NMR Spectroscopy

Prior to the NMR experiments, aliquots of Chesea soil HA were dissolved in water and mixed with 70 mL of 0.1 M NaOH. The resulting solutions were passed 3 times through an amberlite IR-1000H ion exchange resin to remove the paramagnetic metal ions. Aliquots of the subsequently freeze dried Chelsea soil HA were placed over P_2O_5 for 48 hrs at 45 ºC to remove any additional moisture. NMR data were acquired using a Bruker Avance 400 MHz NMR spectrometer fitted with a QNP ^{1}H, ^{13}C, ^{15}N and ^{31}P probes. 1-D ^{13}C NMR and ^{13}C Polarization ENhancement During Attached Nucleus Testing (PENDANT) experiments were carried out with 100 mg of sample dissolved in D_2O/NaOD (990 µL D20; 10 µL NaOD). 1-D ^{13}C NMR (60,000 scans) were acquired using inverse gating with 12 s recycle delay, and processed with 75 Hz line broadening. PENDANT (200,000 scans) were acquired using a 2 s recycle delay and a J1 (^{1}H-^{13}C) of 145 Hz. The spectrum was processed using 75 Hz line broadening. All other experiments were carried out at very low concentrations (1 mg of sample dissolved in 1 mL DMSO-d_6) in an attempt to (i) minimize the effects of residual paramagnetic ions and (ii) increase the relaxation and resolution of the signals present. 1-D ^{1}H NMR (15,000 scans) were carried out with 2 s recycle delay and processed with 1 Hz line broadening. Deuterium exchange was achieved by the addition of a D_2O (~0.05 mL) to the HA sample dissolved in DMSO-d_6. Total Correlation Spectroscopy (TOCSY) (400 scans) were acquired using a 80 ms mixing time, TD (F1) 1024 and

TD (F2) 256 with Time-Proportional Phase Incrementation (TPPI). Gaussian processing was carried out with a line broadening of –1 and a gaussian broadening of 0.003 in both dimensions. Additional processing was carried out using a sine-squared function with phase shift of 90° in both dimensions. Heteronuclear Multiple Quantum Coherence (HMQC) (1024 scans) were acquired using a BIRD pulse train, TPPI, TD (F1) 1024, TD (F2) 256 and J1 (^{1}H-^{13}C) of 145 Hz. F1 was processed with a sine-squared function with phase shift of 90° while F2 was processed with a gaussian broadening of 0.005 and line broadening of –1.

I. 4. Electrospray Ionization (ESI) Quadrupole Time-of-Flight (Qq TOF) Mass Spectrometry

The ESI Qq TOF mass spectra of Chelsea soil HA was acquired using the Micromass Q-Tof™ II mass spectrometer of the Campus Chemical Instrument Facility at Ohio State University. This mass spectrometer equipped with an orthogonal electrospray source (Z-spray). To limit the the number of multiply charged peaks, the mass spectrometer was operated in the positive ion mode. Polyalanine and alanine were used as calibration standards within the mass range 100 – 2000 m/z. Chelsea soil HA was first dissolved in a 10^{-2}M NaOH at pH = 12.0. An aliquot of the dissolved HA was then diluted with a 50:50 % water-methanol solution and infused into the electrospray source at a rate of 5 - 10 μL min^{-1}. Optimal ESI conditions were: capillary voltage 3000 V, source temperature 110^{o} C and a cone voltage of 60 V. Nitrogen was employed as ESI gas. Q1 was set to optimally pass ions from m/z 100 – 2000 and all ions transmitted into the pusher region of the TOF analyzer were scanned over m/z 200-2000 with a 1 s integration time. Data were acquired in a continuous mode (10-15 minutes) until acceptable averages were obtained**.**

II. Computer Assisted Structure Elucidation of Organic Geomacromolecules

Organic geomacromolecules such as humic acids (HAs), fulvic acids, lignin, peat, kerogen, shale, asphaltenes, etc are ubiquitous in nature. Because these complex and multifunctional compounds are "operationally defined", the development of reliable 3-D structural models for these compounds has been a major challenge in environmental chemistry, soil chemistry, organic geochemistry and petroleum chemistry. Three approaches may be used to generate 3-D structural models for complex organic geomacromolecules: *conventional*, *deterministic* and *stochastic*.

II, 1. The Conventional Approach.

The conventional approach is commonly used to elucidate the structure of an unknown compound. With the conventional approach, a structural model is inferred from a set of analytical data through a repetitive trial-and-error process that consists of matching the postulated structure with the analytical data. Virtually, the structures of all chemicals known to date have been elucidated using the conventional approach. There are, however, two major problems associated with this conventional approach. First, the process is carried out manually in most cases; thus, it is time consuming and not very reliable for multifunctional geo-macromolecules such as humic acids, fulvic acids, lignin, peat, kerogen, shale, asphaltenes, etc. Second and most importantly, the conventional approach does not provide any means of selecting the appropriate isomers when numerous structural models can be inferred from the same set of analytical data. Thus, reliable results may be difficult to achieve when structural models of organic geomacro-molecules generated with the conventional approach are used in subsequent calculations of their physicochemical properties by computational chemistry.

II. 2. The Deterministic Approach.

This approach is predicated upon retrieving all the structural models that can be built from a given set of quantitative and qualitative data. For the past 25 years, there have been many attempts to automate the deterministic approach. Several computer programs have been proposed under the generic name of computer-assisted structure elucidation (CASE). Most of these CASE programs are based on artificial intelligence and graph theory and attempt to mimic the work of a chemist elucidating a structure (3-10). The ability of a CASE program to treat redundant information is a critical issue in structure elucidation. Structural input data tend to be highly redundant. Thus, the molecular fragments used as input to CASE program generally overlap. The treatment of overlapping fragments usually results in an exponential increase of computational times as the number of input atoms increase. A number of investigators have attempted to optimize existing CASE programs to limit the generation of duplicate structural models (9-10). Although these optimized CASE program can handle relatively large structures, the size of a structural model that can be generated by any deterministic CASE program is still limited by the exponential increase of computational time associated with structure elucidation process. Consequently, the deterministic approach to structure elucidation is ill suited for complex and multi- functional organic geomacromolecules such as humic acids (HAs).

II. 3. The Stochastic Approach.

The stochastic approach to structure elucidation is conceptually similar to the search of the conformational space of a chemical compound by Monte Carlo simulations or simulated annealing to find its lowest energy conformations (11). However, in the case of structure elucidation, the search space is no longer composed of an infinite number of all possible conformations, but is composed of the finite number of all possible structural isomers that can be constructed from a given set of analytical data. Faulon (11) has shown that, by using a stochastic approach, it is possible to generate all the 3-D models or a sample statically representative of the entire population of structural models that can be built from a given set of analytical data. This new computer assisted structure elucidation software (SIGNATURE) is based on the *signature* descriptor, a molecular codification system similar to the SMILES molecular representation system (11). This concept, which was first presented and applied in the context of structural elucidation of organic geo-macromolecules by Faulon (12), is summarized next. To help the reader who is not familiar with the field of *Chemical Graph Theory* grasp the usefulness of the signature descriptor in structure elucidation, below we provide definitions for a number of key concepts.

Molecular graph. A molecule can be represented by a graph $G=(V_G,E_G,C,c_G())$, where the vertices of V_G represent its atoms and the edges of E_G represent its bonds. The function $c_G()$ associates every atom of G to an element of C, where C can be the set of elements of the periodic table or any set of atom types provided by an empirical force field (FF) such as the Dreiding FF (13). Because every element of C has a valence (i. e., the number of covalent bonds that can be formed with this element), a chemical graph representing a given molecule or molecular fragment is not necessarily saturated. Thus, a molecular graph $G=(V_G,E_G,C,c_G())$, is formally an undirected graph colored with the function $c_G()$ over the elements of C verifying the equation:

$$\forall x \in V_G,\ deg(x) \leq valence(c_G(x)) \qquad 1$$

where $deg(x)$ is the degree of vertex x and $valence(c_G(x))$ is the valence of its associated atom ($(c_G(x))$. A vertex is saturated if its degree is equal to the valence of the associated element. A molecular graph is saturated if all its vertices are saturated. Every covalent molecule can be represented by a saturated molecular graph.

Signature-tree. Let $G=(V_G,E_G,C,c_G())$ be a molecular graph, the h-signature-tree ($^h\sigma$–tree) of an atom x of V_G, is a tree describing the neighborhood of x in G up to distance h. More precisely, the $^h\sigma$–tree of x, $^h\sigma_I(x)=(V(^h\sigma_I(x)),E(^h\sigma_I(x)),C,c_\sigma())$ is a rooted tree on x, where the first layer is composed of the neighbors of x, the second layer is composed of the neighbors of the first layer, and this recursively up to layer h.

Signature of an atom. Let $G=(V_G,E_G,C,c_G())$ be a molecular graph and let x be a atom of V_G. The signature of height h of atom x is a canonical representation of the h–σ–tree, $^h\sigma_G(x)$, colored by the function $c_\sigma()$. Since there is a one-to-one mapping between signatures and signature-trees we use the same notation, $^h\sigma_G(x)$, to represent both objects.

Signature of a molecule. The signature of an atom can essentially be viewed as a string of characters over an alphabet of atom types. For a given height h, the list of all possible atomic signatures, although large, is finite. Consequently, any given molecule/molecular fragment can be represented by its coordinates in a vector space where the base vectors are the distinct atomic signatures. This enables us to define the signature of a molecule/molecular fragment as the linear combination of its atomic signatures:

$$^h\sigma(G) = \sum_{x\in V_G} {}^h\sigma_G(x) = \sum_{i=1}^{^hK_G} {}^h\alpha_i \; {}^h\sigma_G(^hX_i) \qquad 2$$

where $^h\sigma_G(\,^hX_i)$ is a base vector, $^h\alpha_i$ is the number of atoms having the signature of the base vector, and ^hK_G is the number of base vectors.

Signature of a bond. Let $G=(V_G,E_G,C,c_G())$ be a molecular graph and let b be a bond/edge of E_G. Let $G\text{-}b=(V_G,E_G\text{-}\{b\},C,c_G())$ be the molecular graph in which the bond b has been removed. The h-signature of b is defined as follows:

$$^h\sigma(b)={}^h\sigma(G)-{}^h\sigma(G-b) \qquad 3$$

Signature of a reaction. Let $B=(V_B,E_B,C,c_G())$ and $E=(V_E,E_E,C,c_G())$ be two molecular graphs representing the reactants and products of the reaction $R: B\rightarrow E$. Note that signatures can be computed on graphs that are not necessarily connected, hence B and E can both be composed of several molecules. The h-signature of reaction R is given by the equation:

$$^h\sigma(R)={}^h\sigma(E)-{}^h\sigma(B) \qquad 4$$

The interested reader can consult reference 11 for examples of the formal representations of the pertinent signatures for lignin based molecular fragments such as coniferyl alcohol and guaiacyl.

The signature equation. For complex organic geo-macromolecules such as HAs, the signature descriptor provides a simple and robust means of coding (i) elemental analysis data as 0 level atomic signatures, (ii) quantitative $^{1}H/^{13}C$ NMR as 1 or 2 level atomic signatures and (iii) qualitative data (e.g., molecular fragments and interfragment bonds from FT-IR spectroscopy, qualitative 1-D/2-D NMR spectroscopy, ESI mass spectrometry, etc) as 1, 2 or higher level molecular signatures. Once these qualitative and quantitative data for the humic acid (HA) of interest have been coded into the pertinent signatures, the following conservation law provides the conceptual framework for the use of SIGNATURE in structure elucidation:

sum of h-signatures of molecular fragments + sum of h-signatures of interfragment bonds = sum of h-signatures of the HA of interest.

Let ${}^{h}\sigma(S)$ and ${}^{h}\sigma_{\varepsilon}(S)$ be the set of the experimentally derived input h-signatures and associated standard errors of the HA of interest. The quantity x_i of each molecular fragment f_i ($1 \bullet i \bullet I$), and the quantity y_j of each interfragment bond b_j ($1 \bullet i \bullet J$) can be calculated by solving the following system of equations

$$\left.\begin{aligned} {}^{0}\sigma(S)-{}^{0}\sigma_{\varepsilon}(S) \le \sum_{i=1}^{I} x_i\,{}^{0}\sigma(f_i) + \sum_{j=1}^{J} y_j\,{}^{0}\sigma(b_j) \le {}^{0}\sigma(S)+{}^{0}\sigma_{\varepsilon}(S) \\ {}^{1}\sigma(S)-{}^{1}\sigma_{\varepsilon}(S) \le \sum_{i=1}^{I} x_i\,{}^{1}\sigma(f_i) + \sum_{j=1}^{J} y_j\,{}^{1}\sigma(b_j) \le {}^{1}\sigma(S)+{}^{1}\sigma_{\varepsilon}(S) \\ \cdots\cdots\cdots\cdots \\ {}^{h}\sigma(S)-{}^{h}\sigma_{0}^{\varepsilon}(S) \le \sum_{i=1}^{I} x_i\,{}^{h}\sigma(f_i) + \sum_{j=1}^{J} y_j\,{}^{h}\sigma(b_j) \le {}^{h}\sigma(S)+{}^{h}\sigma_{\varepsilon}(S) \end{aligned}\right| \qquad 5$$

where I and J are the numbers of molecular fragments and interfragment bonds. Since the purpose of the SIGNATURE program is to construct molecular models, x_i and y_j are always positive integer numbers. Because of limited experimental data, the linear system given in Equation 5 is generally undetermined and has more than one solution. However, for the purpose of HA structure elucidation, we seek the best solution (Equation 6), i. e., that which minimizes the difference between the sum of the *signatures* of the molecular fragments and interfragment bonds, and the *signature* of the HA of interest:

$$\min \{ |\Sigma X - \sigma(S)| \},\ \Sigma X \bullet \sigma(S) + \sigma_{\varepsilon}(S),\ \Sigma X \bullet \sigma(S) - \sigma_{\varepsilon}(S),\ X \text{ integral} \qquad 6$$

$$\sum = \begin{vmatrix} {}^{0}\sigma(f_1) & {}^{1}\sigma(f_1) & \cdots & {}^{h}\sigma(f_1) \\ \vdots & \vdots & & \vdots \\ {}^{0}\sigma(f_I) & {}^{1}\sigma(f_I) & \cdots & {}^{h}\sigma(f_I) \\ {}^{0}\sigma(b_1) & {}^{1}\sigma(b_1) & \cdots & {}^{h}\sigma(b_1) \\ \vdots & \vdots & & \vdots \\ {}^{0}\sigma(b_J) & {}^{1}\sigma(b_J) & \cdots & {}^{h}\sigma(b_J) \end{vmatrix} \qquad 7$$

where $\sigma(S) = ({}^{0}\sigma(S), ..., {}^{h}\sigma(S))$, and $\sigma_\varepsilon(S) = ({}^{0}\sigma_\varepsilon(S), ..., {}^{h}\sigma_\varepsilon(S))$ are the vectors of input atomic/molecular signatures and associated standard errors , Σ is the matrix of signatures for the selected input molecular fragments (MFs) and interfragment bonds (IBs) and $X = (x_1,...,x_I, y_1,...,y_J)$ is the solution vector of Equation 6.

The reader may recognize that Equation 6, commonly referred to as the *signature equation*, describes formally an *integer linear programming* problem. The CASE program SIGNATURE uses two basic techniques to solve this problem: systematic enumeration and simulated annealing (11). The solution of Equation 6 involves a self-consistent iterative process. First, the user of SIGNATURE attempts to select the "best" possible list of input molecular fragments and associated interfragment bonds based on the results of the DRIFT, NMR and ESI MS data. Using this list an initial guess, the user attempts to obtain the best solution to Equation 6 [subject to the quantitative structural constraints] by varying the quantity x_i of each input molecular fragment (MF) f_i ($1 \bullet i \bullet I$), and the quantity y_j of each input interfragment bond (IB) b_j ($1 \bullet i \bullet J$). Thus, the user can only obtain the "best" solution to Equation 6 by identifying the list of input MF and IB that is "consistent" with the quantitative structural input data. Once the optimal molecular building blocks (i.e., types and amounts of MFs and IBs) have been determined, SIGNATURE generates all the 3-D models that are consistent with the input data by randomly connecting the pertinent MFs and IBs for the HA of interest. The users of SIGNATURE can also impose structural constraints such as generating 3-D structural with number average molecular weights within a specified range. Thus, SIGNATURE has the inherent capability to generate a sample of representative 3-D structural models for complex organic geo-macromolecules such as lignin (14) and asphaltenes (15) if the pertinent analytical data is available.

II. Literature Cited

1. International Humic Substances Society. http://www.ihss.gatech.edu/.
2. Huang, W.; Weber, W. J., *Env. Sci. Technol.*, **1997**, 31, 2562.
3. Carhart, R. E.; Smith, D. H., Brown, H. and Djerassi, C., *J. Am. Chem. Soc*., **1975**, 97, 5755
4. Smith, D. H., Gray, N. A. B. Nourse, J. G. and Crandel, C. W. *Anal. Chim. Acta*, **1981**, 133, 471.
5. Abe, H; Okuyama, T.; Fujiwara, F. and Sasaki, A. *J. Chem. Inf. Comput. Sci*., **1984**, 24, 22.
6. Kudo, Y. and Sasaki, S. *J. Chem. Inf. Comput. Sci*., **1985**, 25, 252.
7. Funatsu, K.; Miyabaski, N. and Sasaki, S., *J. Chem. Inf. Comput. Sci*., **1988**, 28, 9.
8. Oshima, T.; Ishida, Y.; Sato, K, and Sasaki, S. *Anal. Chim. Acta*, **1980**, 122, 95.
9. Bangov, I. P. *J. Chem. Inf. Comput. Sci*., **1994**, 34, 277.
10. Contreras, M. L.; Rozas, R. and Valdivias, R. *J. Chem. Inf. Comput. Sci*., **1994**, 34, 610
11. Faulon, J. L. *J. Chem. Inf. Comput. Sci*., **1994**, 34, 197.
12. Faulon, J. L., Vandenbroucke, M., Drappier, J. M.; Behar, F. and Romero, M. *Adv. Org. Geochem.*, **1981**, **16**, 981.
13. Mayo, S. L.; Olafson, B, D and Goddard, W. A. III. *J. Phys. Chem*, **1989**, 90, 8897.
14. Faulon, J.L.; Hatcher, P. G. *Energy & Fuels,* **1994,** *8*, 402.
15. Diallo, M.S; Cagin, T.; Faulon, J. L; Goddard, W. A III. *Asphaltenes and Asphalts. II. Developments in Petroleum Science Series,* 40 B, Eds, Yen T. F and Chilingarian, G. V., Elsevier Science, Amsterdam, Netherland, **2000**, 103.