Stochastic Generator of Chemical Structure. 1. Application to the Structure Elucidation of Large Molecules

Jean-Loup Faulon†

Sandia National Laboratories, P.O. Box 5800, Albuquerque, New Mexico 87185-0710

Received April 7, 1994®

This paper presents an original computer-assisted structure elucidation system based on a stochastic approach. Using a randomized technique, it is shown that the number of chemical structures that match a set of analytical data can be approximated in a reasonable computational time. Furthermore, it is demonstrated that a sample of three-dimensional models can be generated and be statistically representative of the entire population of potential models. The analytical data introduced in the system can be derived from experimental techniques as diverse as elemental analysis, functional group analysis, ^{1}H, ^{13}C, and ^{29}Si NMR, mass spectrometry, pyrolysis, gas chromatography, pycnometry, N_2 and CO_2 adsorption, mercury intrusion, and SAXS. The stochastic structure elucidation system is applied to macromolecular compounds that are studied in biochemistry (lignin), geochemistry and fuel science (coal), and material sciences (amorphous silica gel). For these compounds as well as for other amorphous chemical structures cited in the paper the proposed stochastic approach is the first technique that correlates a large diversity of analytical data and three-dimensional molecular models.

INTRODUCTION

Molecular modeling is an important research tool in many fields of chemistry. Several techniques have been developed, among them *ab initio* calculations,[1] semiempirical calculations,[2] and molecular simulations.[3] Molecular simulations such as *molecular mechanics*[3a] (MM), *molecular dynamics*[3b] (MD), *systematic conformational search*[3b] (CS), *simulated annealing*[3b] (SA), and *Monte Carlo*[3b] (MC), are the cornerstones of most commercially available molecular modeling products. The aim of molecular simulations is to study the three-dimensional organization of chemical compounds. While MM, MD, and CS are *deterministic* methods, exploring a part of or the entire conformational space of molecules, SA, MC, or the *genetic algorithm*[3c] (GA) are *stochastic* methods, which randomize the exploration of the conformational space.

It is imperative to recall that most of the molecular simulations cannot be used if the structural formula (i.e., the connectivity between atoms) of the studied chemical compound is unknown. In fact, before using any molecular simulations technique, one has to define a reasonable starting three-dimensional conformation and therefore know the structural formula of the studied molecule.[4] While molecular simulations are able to modify and optimize the three-dimensional organization of chemical structures, they are generally unable to change the connectivity between atoms.

There are many research areas in chemistry in which the studied chemical compounds have unknown structural formulas. For example, the present paper refers to compounds that are studied in biochemistry (lignin), geochemistry and petroleum chemistry (humic acids, kerogen, and asphaltene), geochemistry and fuel sciences (coal), and materials sciences (amorphous silica gels). Although the connectivity between the atoms is not known for these compounds, molecular simulations are currently used to study their three dimensional organization.[5-7] To overcome the problem of determining a structural formula and an initial three-dimensional organization (which is the input for all molecular simulations), three approaches can be taken—*conventional*, *deterministic*, and *stochastic*.

The Conventional Approach. This approach is commonly taken by chemists when trying to elucidate an unknown chemical structure. With the conventional approach, a structural formula is first inferred from a set of analytical data. The structural formula is retrieved manually through a repetitive trial-and-error process that consists of matching the structure in construction with the analytical data. In fact, the trial-and-error process is the every day task performed by the structural chemist, and almost all known chemical structures have been found this way. In the fields covered by the present paper, the trial-and-error process has been applied with lignin,[8] kerogen and asphaltene,[9] and coal.[10] The second step of the conventional approach consists of taking the structure produced in the first step and generating a three-dimensional model. This task can be performed using a molecular builder provided by any molecular modeling software. As examples of applications of the conventional approach coupled with molecular simulations, the reader can consult the following references for lignin[5] and coal.[6] Although there have been many applications and successes with this approach, there are at least two disadvantages when the conventional approach is applied for large molecules. (1) The process to build a structure is accomplished through manual fitting and is usually prohibitively time consuming for large molecules. (2) When many structures can be built from the same analytical data, the reason one structure is chosen over another is often not clearly defined. If the structures are chosen arbitrarily (as were lignin,[8] kerogen,[9] and coal[10]), it is impossible to draw definite conclusions regarding a compound modeled with the conventional approach.

The Deterministic Approach. This approach consists of retrieving *all* the structural formulas and eventually the corresponding three-dimensional structures from a set of structural analytical data. For the past 25 years, there have been many attempts to automate the deterministic approach. Several techniques and computer programs have been proposed under the generic name *computer-assisted structure elucidation* (CASE[11]). The programs that have been developed are based on artificial intelligence and graph theory and attempt

† E-mail address: loup@faulon.sandia.gov.

® Abstract published in *Advance ACS Abstracts*, August 15, 1994.

0095-2338/94/1634-1204$04.50/0

to mimic the work of a chemist elucidating a structure. The first CASE program was published by Lederberg *et al.*[12a] and was able to enumerate all the acyclic structures from a molecular formula. This program was part of the DENDRAL project[12b] and was the precursor of CONGEN[12c,d] and GENOA,[12e,f] the first expert systems ever published. With CONGEN and GENOA any type of organic structure can be treated. Both programs are able to enumerate the isomers of a molecular formula and are also able to generate structures with more restrictive constraints, such as molecular fragments (pieces of the molecular structure with known connectivity between their atoms). However, the technique employed is more a heuristic than a systematic algorithm, and a preknowledge of part of the results is necessary; GENOA used a precompiled catalogue of 3000 elementary cyclic structures. Furthermore, the proof of irredundancy and exhaustivity of the structure generation was never published, and differences were found between the results of these programs and other techniques.[13a] More systematic are the approaches of the CHEMICS,[13a-f] ASSEMBLE,[14a,b] and COMBINE[14c] structure generators. These programs are based on the concept of the connectivity stack,[13a] which allows an exhaustive and unique enumeration.[13b] The starting point of these programs is a set of segments representing the unknown compound, a segment is a small molecular fragment containing one, two, or three atoms. To enumerate the isomers, an exhaustive permutation of all segments is processed. With the concept of the connectivity stack, redundancies can be avoided without cross-checking the solutions. Using this method, exhaustivity and irredundancy of the solutions can be easily proven; in fact, all the permutations are considered, and all redundant structures are rejected.

There are two central issues that have to be raised with CASE systems. The first issue regards the ability of such systems to treat redundant information. In the real world of structure elucidation, chemical structural data is highly redundant; consequently, the molecular fragments input in the CASE systems generally overlap. The problem of overlapping fragments was first studied by Dubois *et al.*[15a,b] This work led to the development of the program DARC-EPIOS,[15c] which can retrieve structural formulas from overlapping ^{13}C NMR data. Similar techniques have also been applied with the COMBINE,[14c] and ACCESS[16] programs, while GENOA[12e,f] uses a more general technique, which consists of preprocessing the required molecular fragments by determining all possible nonoverlapping combinations. More recently, a new strategy has been proposed, named structure generation by reduction.[17] All the CASE programs cited above generate a chemical structure by assembling atoms or molecular fragments together and consequently creating bonds. Structure generation by reduction does not create bonds but removes bonds from a *hyperstructure*. Initially, the hyperstructure contains all the possible bonds between all the required atoms and molecular fragments. The fragments can overlap. As bonds are removed, the continued containment of each fragment is tested until a valid chemical structure is obtained. Based on the concept of structure generation by reduction, the COCOA[17a] and the GEN[17b] programs have been developed.

In common with all the solutions that have been proposed to resolve the problem of overlapping fragments is the increase of the computational complexity of the process of structure elucidation. Because the overlapping problems are treated by supplementary computational operations, they increase the global computational time of CASE systems. In fact, computational complexity is the second central issue of structure elucidation programs. All the CASE systems cited above have a time complexity that increases exponentially with the number of input atoms. For this reason the CASE systems are up till now inefficient when working with a large number of atoms. The largest structure retrieved by the previously cited CASE studies contains 28 non-hydrogen atoms (Diasin structure—$C_{21}H_{24}O_7$); this structure was elucidated by GENOA.[12f] Computational complexity in structure elucidation systems has recently been studied, and optimized techniques have been proposed.[18-21] Based on graph theory, the proposed techniques generate an exhaustive and irredundant list of structural formulas. However, checking if a structure is irredundant is computationally time demanding; in fact, in the worst case, this process has an exponential time complexity.[20] Hence, the aim of the optimized programs is primarily to limit the number of combinations that generate duplicate structures. With one of these optimized techniques, a structure containing 122 non-hydrogen atoms (fragment of lignin, $C_{116}H_{126}O_6$) has been elucidated.[21] Although, these techniques can treat larger structures, the number of atoms or molecular fragments that can be processed by any deterministic CASE system is still limited due to the exponential complexity of the problem of structure elucidation. In other words, even with today's fastest computers, the deterministic approach is not applicable for macromolecules.

The Stochastic Approach. Fundamental to this approach is the following question: Is it necessary to generate all of the structural formulas (corresponding to a set of analytical data) in order to study the three-dimensional physical characteristics of an unknown compound? Related to this is the question of whether or not the concept of a unique structural formula has a physical or chemical significance for amorphous macromolecules such as lignin, coal, kerogen, or silica gel. Performing a structure elucidation using a stochastic approach is similar to studying the conformational space of a chemical structure using stochastic methods such as SA, MC, or GA methods. 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 set of analytical data.

The purpose of the present paper is to prove that by using a stochastic approach, it is possible to generate a sample of three-dimensional molecular models that statistically represents the entire population of all the possible models that can be built from a set of analytical data. In the next sections, a new structure elucidation system (the SIGNATURE program[21]) is presented; the SIGNATURE program is believed to be the first stochastic CASE system.

OVERVIEW OF THE SIGNATURE PROGRAM

At the heart of any structure elucidation system is the analytical data. There is a large variety of chemical and physical analytical techniques that give structural information. Because the computer program presented in this paper has mainly been applied to elucidate "random" chemical structures, the present discussion is purposely confined to experimental techniques used to characterize amorphous solids. Table 1 presents a list of such techniques; these techniques are routinely employed in biochemistry, geochemistry, petroleum, fuel, and material sciences. The list given in Table 1 is not exhaustive,

Table 1. Common Analytical Techniques Used To Characterize Amorphous Solids

technique[a]	data	data type
elemental analysis	atomic ratios	2D, quantitative
functional group analysis	functional groups ratios	2D, quantitative
CPMAS ^{13}C NMR	structural fragments centered on carbon atoms	2D, quantitative
block decay ^{13}C NMR	structural fragments centered on carbon atoms	2D, quantitative
CPMAS ^{29}Si NMR	structural fragments centered on silicon atoms (Q^n distribution)	2D, quantitative
CRAMPS ^{1}H NMR	structural fragments centered on hydrogen atoms	2D, quantitative
2D COSY ^{1}H–^{1}H NMR	structural fragments centered on hydrogen atoms	2D, qualitative (quantitative)
INADEQUATE 2D ^{13}C NMR	structural fragments centered on carbon atoms	2D, qualitative
HMQC 2D ^{1}H–^{13}C NMR	structural fragments centered on carbon and hydrogen atoms	2D, qualitative
MS/MS	molecular formula	2D, quantitative
GC/MS and Py/GC/MS	structural formula of molecular fragments	2D, qualitative
pycnometry (Hg or He)	density	3D, quantitative
N_2 and CO_2 adsorption	pore volume, surface area, fractal dimension	3D, quantitative
mercury intrusion	micropore volume distribution, fractal dimension	3D, quantitative
SAXS and SANS	surface area, fractal dimension	3D, quantitative

[a] All the NMR techniques listed are solid state NMR. CPMAS = cross-polarization with magic angle spinning, CRAMPS = combined rotation and multiple pulse spectroscopy, INADEQUATE = incredible natural abundance double quantum transfer experiments, HQMC = heteronuclear multiple quantum coherence, MS = mass spectrometry, GC = gas chromatography, Py = pyrolysis, SAXS = small angle X-ray scattering, SANS = small angle neutron scattering.

but it is the list of analytical techniques that have been used by the SIGNATURE program.[22] Each technique listed in Table 1 provides quantitative or qualitative data and gives 2D or 3D information on the studied compound. The classification quantitative/qualitative is conventional in chemistry. More important is the classification 2D/3D adopted in Table 1. 2D data are information relative to the topology of the unknown chemical structure. For example, elemental analysis gives the percentage of each element present in the studied structure. In biochemistry, geochemistry, fuel, or petroleum sciences, pyrolysis followed by gas chromatography and mass spectrometry (Py/GC/MS) is routinely employed to infer the structural formulas of the molecular fragments of an unknown macromolecular compound.[24] In short, 2D data characterize the connectivity between atoms, and 2D data are independent of the three-dimensional conformation of the unknown structure. There are analytical techniques that characterize the three-dimensional organization of a molecular structure. These techniques provide 3D data. One example of these techniques is the pycnometry that measures the density. 3D data depend on the three-dimensional organization of the studied compound, and as discussed in the introduction, in most cases, the three-dimensional organization of a molecular model cannot be studied if the structural formula of this model is unknown. Therefore, as shown in Figure 1, the stochastic structure generator presented in this paper processes 2D data before 3D data.

Four distinct tasks are performed by the SIGNATURE program (cf. Figure 1). The purpose of the first task, the *signature equation*,[26] is to calculate the list of molecular fragments and interfragment bonds that constitute the models. Roughly, the signature equation consists of matching 2D qualitative data with 2D quantitative data in order to compute an exhaustive and nonoverlapping list of molecular fragments and interfragment bonds. 2D quantitative data are not exact values; there is a standard deviation associated with each datum. It is the task of the expert using the SIGNATURE program to input these standard deviations. Furthermore, if the molecular formula of the studied compound is unknown, the user of the program inputs the average number of atoms. Most of the time, there are several lists of molecular fragments and interfragment bonds that correspond to the given sets of 2D data and standard deviations. The goal of the signature equation is to determine the "best" list, i.e., the list that minimizes the deviation between the model and the 2D quantitative data. Once a list of molecular fragments and interfragment bonds is determined, a structural formula can be obtained by connecting the fragments with the corresponding interfragment bonds. At that stage, the structure to be constructed is much like a jigsaw puzzle; one knows the pieces of the puzzle and the ways these pieces are connected together. Generally, several structural formulas can be constructed. The second task, the stochastic structure generation, evaluates the number of possible structural formulas and generates a sample of structural formulas that statistically represents the entire population of possibilities. The stochastic structure generator constructs the structural formulas in a three-dimensional space. During the generation process, the expert using the system inputs the sample size and can impose some structural constraints, such as avoiding the formation of double bonds or forcing the generator to build five or six membered rings. Once the sample of models is constructed, the third task, the 3D simulations, submits each model to molecular orbital calculations or molecular simulations. After the optimized 3D models are produced, 3D physical properties are calculated for the models and compared to the corresponding 3D analytical data (cf. Table 1). Finally, the sample is statistically analyzed by the fourth task. The purpose of the statistical interpretation is to determine the optimal sample size needed for statistical significance and to extrapolate the physical properties calculated for the sample by the 3D simulations to the entire population of possible models. The statistical analysis may indicate that an increase of the sample size is necessary or that the generation of an entirely new sample is necessary where the models are better matched with the overall set of analytical data. To generate a new sample of models, the expert using the program can decide to modify the structural constraints during the stochastic structure generation.

The purpose of the following sections is to present the theoretical grounds of the four main tasks presented above. Technical information concerning the SIGNATURE program is given in the appendix.

THE SIGNATURE EQUATION

The *signature* is a systematic codification system for 2D analytical data, which can be compared to the SMILES notation system.[25] This concept, which was first defined and applied in the limited context of kerogen macromolecules,[26]

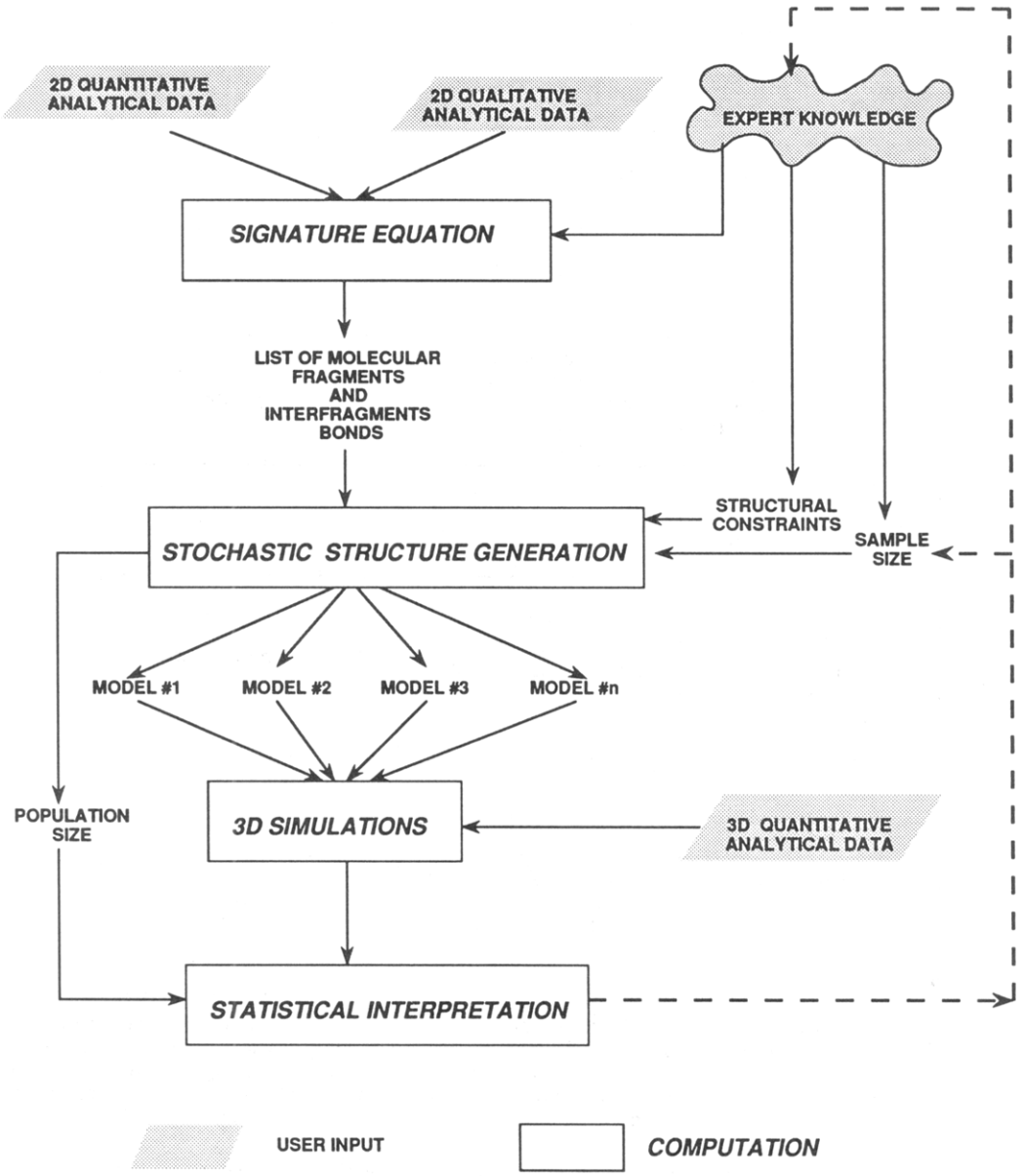


Figure 1. General scheme of the SIGNATURE program.

is expanded in the present paper. Prior to defining the signature, some terminology based on graph theory has to be defined. The reader not familiar with the notation used in chemical graph theory can consult, for example, the book of Trinajstic.[27]

In the following discussion, V represents a set of vertices, E a set of edges, and C the set of elements of the periodic table. A molecule is represented by the graph $G = (V,E)$, where the elements of V are the atoms and the edges of E are the bonds. Let $c()$ be the function that associates a vertex to an element. Every element has a valence, which represents the number of covalent bonds that can be formed with this element. An **atomic graph** is a graph representing a molecule which is not necessarily entirely built (i.e., part of the connectivity can be unknown). In formal language, an atomic graph $G = (V,E)$ is an unlabeled and nonoriented graph colored by the elements of C which verify the equation

$$\forall\, v \in V, \text{degree}(v) \leq \text{valence}\,(c(v)) \qquad (1)$$

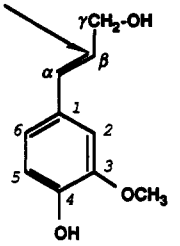
in other words the number of bonds of a given vertex is less or equal to the valence of the corresponding atom.

A vertex is **saturated** if its degree is equal to the valence of the associated element. An atomic graph is saturated if all vertices are saturated. Every covalent molecule is a saturated atomic graph.

Let v be a vertex of the atomic graph $G = (V,E)$, and let $T_\lambda(v)$ be the spanning subtree of height λ, rooted on v. The λ-***signature*** of v, $(\sigma_\lambda(v))$ is defined by the following relation:

$$\sigma_\lambda(v) = c[T_\lambda(v)] \qquad (2)$$

in other words, the λ-signature of v is the set of chemical elements of the subtree $T_\lambda(v)$. When applied to molecular structures, the subtree $T_\lambda(v)$ can be viewed as a molecular fragment centered on the atom v reduced to a limited environment of radial distance λ. These type of fragments are named FREL;[28a] they were first defined by Dubois *et al.*[28b] and are used in the DARC-EPIOS system.[15c] Using the FREL terminology, the λ-signature of an atom v, is the string of characters formed by the chemical elements of the FREL of radial distance λ, centered on v. In order to obtain a standard notation for the signature, the chemical elements of the FREL are read in deep-first order, and the branches



Coniferyl alcohol

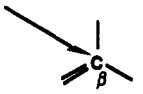
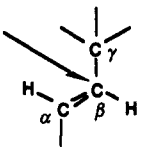
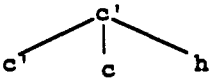
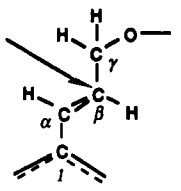
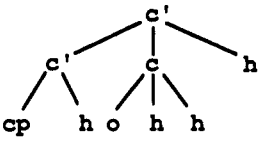
λ	FREL (fragment centered on atom β of radial distance λ)	Rooted spanning subtree ordered in lexicographic order using the atom type notation	λ-signature
0		c'	$\sigma_0(\beta)$ = c'
1			$\sigma_1(\beta)$ = c'(c' c h)
2			$\sigma_2(\beta)$ = c'(c'(cp h) c(o h h) h)

Figure 2. The signatures of an atom. Coniferyl alcohol is the main monomer precursor of gymnosperm lignin. The arrow indicates the chosen atom (β). In the signature notation, the chemical elements are differentiated by atom types. The atom types used in Figures 2, 3, and 4 are "cp" for aromatic carbon, "c' " for carbon double bonded, "c" for aliphatic carbon, "o" for oxygen single bonded, "h" for hydrogen, and "." for radical (i.e., bonding site). The total list of atom types used by the SIGNATURE program is the list of atom types defined in INSIGHT 2.3 (Biosym) molecular modeling products. Each signature is written in lexicographic order with "cp" > "c' " > "c" > "o" > "h" > "." .

of the FREL are ordered from the left to the right in lexicographic order. Examples of *signatures* of atoms are given in Figure 2.

Let $G = (V,E)$ be an atomic graph; i.e., a molecular fragment or a molecule. The λ-signature of G, $(\sigma_\lambda(G))$ is the sum ofthe λ-*signatures* of all the vertices of G:

$$\sigma_\lambda(G) = \sum_{v \in V} \sigma_\lambda(v) \qquad (3)$$

Examples of signatures of molecules and molecular fragments are presented in Figure 3.

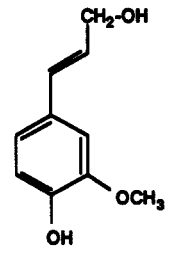
Let e be an edge of the atomic graph $G = (V,E)$ (i.e., a bond in the corresponding molecule). The λ-signature of e, $(\sigma_\lambda(e))$ is the difference between the λ-signature of the graph $G = (V,E)$ and the λ-signature of the graph $G\text{-}e = (V,E\text{-}e)$:

$$\sigma_\lambda(e) = \sigma_\lambda(G) - \sigma_\lambda(G - e) \qquad (4)$$

Examples of signatures of bonds are given in Figure 4. From the definition of the signatures of a bond, one may notice that signatures can also be calculated for chemical reactions. In fact, the molecule associated with the graph $G\text{-}e$ can be viewed as a radical structure, and the creation of the bond e can be interpreted as a radical reaction mechanism. Hence, the λ-signature of a chemical reaction is the difference between the λ-signatures of the products and the λ-signatures of the reactants.

As already mentioned, the 2D analytical data can be coded with signatures. The 2D quantitative data provide a counting of atoms or molecular fragments centered on atoms; therefore, to each datum can be associated a FREL and consequently a signature. For example, elemental analysis gives atomic ratios. These can be coded with *0*-signatures. To each identified peak in a NMR spectrum can be associated a FREL of radius generally equal to 1 or 2; therefore each NMR datum can be coded in the form of a *1*- or a *2*-signature. An example of this is shown in Table 2, which lists the 2D quantitative data and the associated signatures for a lignin compound. The 2D quantitative data for this compound were taken from Hatcher.[29]

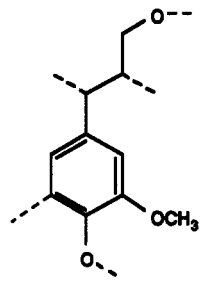
Signatures can also be calculated for 2D qualitative data. 2D qualitative data are composed of molecular fragments and interfragment bonds. Molecular fragments are part of the studied structure with known connectivity between their atoms; therefore, molecular fragments can be represented by atomic graphs, and signatures can be calculated for each of them using eq 3. Interfragment bonds are bonds between fragments, in other words, interfragment bonds are edges between atomic graphs; consequently, signatures of each



Coniferyl alcohol

σ_0 = 6cp + 2c' + 2c + 3o + 12h

σ_1 = cp(cp cp c') + 2cp(cp cp o) + 3cp (cp cp h)
+ c'(c' cp h) + c'(c' c h)
+ c(c' o h h) + c(o h h h)
+ o(cp c) + o(cp h) + o(c h)
+ 3h(cp) + 2h(c') + 5h(c) + 2h(o)



Guaiacyl (lignin molecular fragment)

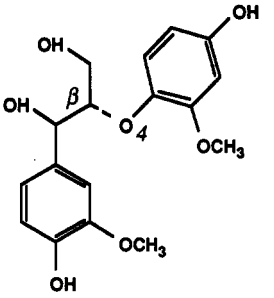
σ_0 = 6cp + 4c + 3o + 9h

σ_1 = cp(cp cp c) + 2cp(cp cp o)
+ 2cp(cp cp h) + cp(cp cp .)
+ c(cp c h .) + c(c c h .)
+ c(c o h h) + c(o h h h)
+ o(cp c) + o(cp .) + o(c .)
+ 2h(cp) + 7h(c)

Figure 3. The signatures of a molecule or a molecular fragment. Cf.: caption of Figure 2 for the signature notation. Guaiacyl is derived from coniferyl alcohol and is the main molecular fragment of gymnosperm lignin macromolecule. The dashed lines represent the bonding sites. The other molecular fragments of gymnosperm lignin are in order of importance *p*-hydroxyphenyl (same structure as guaiacyl without methoxy functional group) and syringyl (same structure as guaiacyl with two methoxy functional groups in positions 3 and 5).

interfragment bond are obtained using eq 4. Figures 3 and 4 list signatures of some molecular fragments and interfragment bonds for the lignin example. The overall set of available qualitative data for lignin are reviewed in Faulon and Hatcher.[23d]

The purpose of the signature equation is to compute the quantities of each molecular fragment and each interfragment bond in order to best match the 2D quantitative data. On one hand, quantitative data provide λ-signatures of the unknown structure. On the other hand, λ-signatures can be calculated for each molecular fragment and each interfragment bond that are present in the studied compound. Once the appropriate amount of molecular fragments and interfragment bonds have been determined, the following relation is true: λ-signatures of molecular fragments + λ-signatures of interfragment bonds = λ-signatures of the unknown structure, within the standard deviations associated with the 2D analytical data. Let $\sigma_\lambda(S)$ be the λ-signature of the unknown structure, and let $\sigma_\lambda^\epsilon(S)$ be the set of standard deviations corresponding to $\sigma_\lambda(S)$, the quantity x_i of each molecular fragment f_i ($1 \le i \le I$), and the quantity y_j of each



β—O-4 interfragment bond

σ_0 = c + o - c - o = 0

σ_1 = c(c c o h) + o(cp c) - c(c c h .) - o(cp .)

Figure 4. The signatures of a bond. The chosen bond is the dashed line. Cf.: caption of Figure 2 for the signature notation. The β–O-4 bond is the main interfragment bond that occurs in gymosperm lignin. The other interfragment bonds are β–5, β–1, β–β, and α–O-γ.

interfragment bond b_j ($1 \le i \le J$) can be calculated by solving the following system of equations:

$$\sigma_0(S) - \sigma_0^\epsilon(S) \le \sum_{i=1}^{I} x_i\sigma_0(f_i) + \sum_{j=1}^{J} y_j\sigma_0(b_j) \le \sigma_0(S) + \sigma_0^\epsilon(S)$$
$$\sigma_1(S) - \sigma_1^\epsilon(S) \le \sum_{i=1}^{I} x_i\sigma_1(f_i) + \sum_{j=1}^{J} y_j\sigma_1(b_j) \le \sigma_1(S) + \sigma_1^\epsilon(S)$$
$$\cdots\cdots\cdots\cdots$$
$$\sigma_\lambda(S) - \sigma_\lambda^\epsilon(S) \le \sum_{i=1}^{I} x_i\sigma_\lambda(f_i) + \sum_{j=1}^{J} y_j\sigma_\lambda(b_j) \le \sigma_\lambda(S) + \sigma_\lambda^\epsilon(S) \quad (5)$$

where x_i and y_j are the unknowns, and 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 positive integer numbers. The value of λ depends on the 2D analytical data, λ is the largest radial distance of all the FREL identified by the 2D analytical techniques. For the set of systems to which the SIGNATURE program has been applied, λ ranged between one[23a–d] and two.[23e] Hence, in eq 5, there are generally more unknowns than equations; one may expect several solutions from eq 5. Nevertheless, only the best solution is interesting; i.e., the solution that minimizes the difference between the sum of the signatures of the molecular fragments and interfragment bonds, and the signature of the studied structure. To determine the best solution, one needs to resolve the following system of equations expressed in a vectored form

$$\min\{\|\sum X - \sigma(S)\|\}, \quad \sum X \le \sigma(S) + \sigma^\epsilon(S), \quad \sum X \ge \sigma(S) - \sigma^\epsilon(S), \quad X \text{ integral} \quad (6)$$

where Σ is the matrix of signatures of molecular fragments and interfragment bonds

$$\sum = \begin{matrix} \sigma_0(f_1) & \sigma_1(f_1) & \cdots & \sigma_\lambda(f_1) \\ \vdots & \vdots & & \vdots \\ \sigma_0(f_I) & \sigma_1(f_I) & \cdots & \sigma_\lambda(f_I) \\ \sigma_0(b_1) & \sigma_1(b_1) & \cdots & \sigma_\lambda(b_1) \\ \vdots & \vdots & & \vdots \\ \sigma_0(b_J) & \sigma_1(b_J) & \cdots & \sigma_\lambda(b_J) \end{matrix}$$

where $\sigma(S)$, and $\sigma^\epsilon(S)$ are the signatures and associated standard deviations of the unknown structure $\sigma(S) = (\sigma_0(S),$

..., $\sigma_\lambda(S)$), and $\sigma^\epsilon(S) = (\sigma_0^\epsilon(S), ..., \sigma_\lambda^\epsilon(S))$, and where X is the vector of the unknowns $X = (x_1,...,x_I, y_1,...,y_J)$.

Equation 6 is named the signature equation. The reader may recognize that the signature equation is an *integer linear programming* (ILP) problem.[30] Several deterministic[30a] and stochastic[30b] techniques can be applied to resolve ILP problems. The most straightforward technique is the *systematic enumeration*. This technique is deterministic and is similar to the CS method employed to study the conformational space of molecules. The systematic enumeration is used by the SIGNATURE program, and presents the advantage of listing all the solutions of eq 5, while resolving eq 6. Hence the expert using the program can compare the various solutions to each other. The solution of the signature equation that minimizes $\Sigma X - \sigma(S)$ is given in Table 3 for the lignin example of Table 1.

Before closing this section, it is important to point out that the signature equation resolves the overlapping problems inherent to structure elucidation systems (cf. introduction section). In fact, each solution of the signature equation is an exhaustive list of nonoverlapping molecular fragments. The exhaustive and the nonoverlapping characters can easily be proved. If a fragment is missing or two fragments overlap, the list of molecular fragments does not match the 2D quantitative data, and there is no solution to eq 6. As previously mentioned, eq 6 is solved by using an enumeration technique; in other words, all the possible X vectors are tested. According to eq 6, if $|\Sigma X - \sigma(S)| \leq \sigma^\epsilon(S)$, the solution X is output; otherwise, the program outputs the deviation $\Sigma X - \sigma(S)$. Hence, the expert using the program can easily detect which FREL are in excess or are missing among the fragments. The FREL that are in excess lead to outputs where $\Sigma X - \sigma(S) > \sigma^\epsilon(S)$. The missing FREL lead to outputs where $\Sigma X - \sigma(S) < -\sigma^\epsilon(S)$. FREL that are in excess are the parts of the fragments that overlap. In order to obtain a solution to eq 6, the expert using the program has to remove these FREL from the fragments. FREL are missing when the list of fragments that is input into eq 6 is not exhaustive. This situation is not unusual, since analytical techniques are not necessarily able to detect all the molecular fragments present in a macromolecular compound. When such a problem occurs, the expert using the system adds the missing FREL to the list of molecular fragments. In other words, the SIGNATURE program can accept molecular fragments that are reduced to a single atom. In the worst case, where there is no 2D qualitative data, the 2D qualitative data input in the SIGNATURE program are composed of the FREL derived from the 2D quantitative data.

THE STOCHASTIC STRUCTURE GENERATION

The inputs of the stochastic structure generator are a list of molecular fragments and interfragment bonds, a set of structural constraints, and a sample size. Once the molecular fragments are connected together with appropriate interfragment bonds, and with respect to the structural constraints, the resulting structural formula represents the unknown compound. In fact, there are several ways of connecting the molecular fragments together and, therefore, several possible structural formulas. Consequently, the outputs of the stochastic structure generator are the number of structural formulas that can be constructed and a random sample of these formulas built in a three-dimensional space. The method employed to count and to generate at random the structural formulas is based on the *equivalent classes algorithm*.[21] This algorithm was originally developed to compute the structural isomers of a molecular formula. There are several techniques to transform structural formulas into three-dimensional molecular models.[31] The specific molecular model builder used by the SIGNATURE program is described and applied in Faulon *et al.*[26a,c]

The equivalent classes algorithm is a deterministic structure generator that calculates all the graphs that can be constructed by adding edges to an initial graph. In the present case, the initial graph is composed of the molecular fragments, and the edges added by the algorithm are the interfragment bonds. The resulting graphs are the structural formulas. They are connected, saturated (cf. definition in previous section), nonisomorphic (i.e., structurally different), and must respect the given set of structural constraints. The structural constraints are chosen by the expert using the SIGNATURE program. The possible constraints are as follows: (1) Avoid the construction of a double, triple, or conjugated bond. (2) Forbid a bond between two atoms that belong to the same fragment. (3) Force or forbid the construction of n-membered rings (n = 3–6).

The *equivalent classes algorithm* is based on the concept of graph automorphism. Two atoms of the same molecule are equivalent if an automorphism (i.e., a permutation) exists between these two atoms that does not change the connectivity of the molecule. All the atoms that are equivalent belong to the same class. All the atoms of a given molecule can be partitioned into equivalent classes. When extended between two graphs, the automorphism function is called an isomorphism. Hence, if a permutation can be found between the atoms of two molecules, and if the bonds of the molecules can be associated by the same permutation, the two molecules are isomorphic. Mathematical definitions of automorphism and isomorphism for atomic graphs are given in Faulon.[21]

The *elementary operation* of the *equivalent classes algorithm* is to calculate all the different ways of saturating a given vertex (atom) by adding to this vertex one or several edges (bonds). The *elementary operation* checks that each bond added belongs to the list of interfragment bonds and respects the structural constraints. To prevent the generation of unconnected final saturated graphs, the *elementary operation* verifies that after each bond added, no saturated subgraphs have been formed. The final saturated graphs may be isomorphic if the *elementary operation* is processed on a arbitrary series of unsaturated vertices.[32] However, if all vertices that belong to the same equivalent class are processed one after another, the *equivalent classes algorithm* guarantees under minor restrictions the nonisomorphism of the final saturated graphs.[33] Hence, to generate an exhaustive and irredundant set of structural formulas, the *equivalent classes algorithm* process is as follows.[34] First the equivalent classes of the initial graph are computed. Then, for a given equivalent class, the algorithm computes all the nonisomorphic graphs by applying the *elementary operation* to all the vertices of the class. The resulting graphs are not necessarily saturated. However, the classes are now different because some edges have been added. For each resulting graph the algorithm determines the new equivalent classes. The algorithm is recursively applied until there are no more unsaturated vertices. The *equivalent classes algorithm* is illustrated in Figure 5a.

When randomized, the deterministic *equivalent classes algorithm* becomes a stochastic structure generator. In the deterministic version of the algorithm, the *elementary operation* constructs all the graphs that can be generated by saturating a given vertex. Furthermore, the *elementary operation* is repeated for all the graphs generated. In the

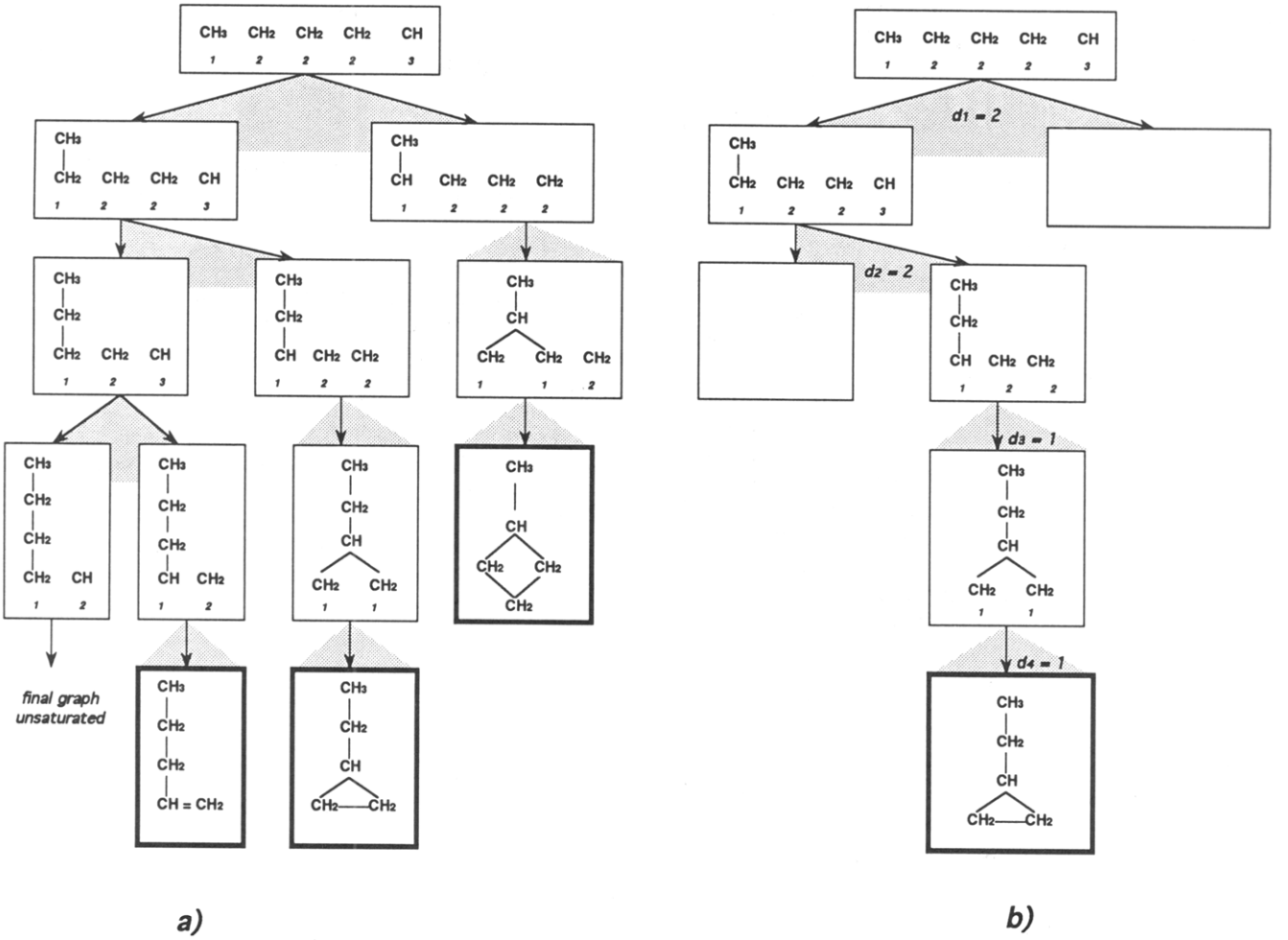


Figure 5. Graphical illustration of the equivalent classes algorithm. Each node (rectangle) is an atomic graph. The leaves of the trees (bolded rectangles) are saturated graphs. The numbers *1*, *2*, and *3* indicates the equivalent classes of the unsaturated vertices. Each shaded triangle represents the computation of all the graphs that saturate the vertices of a given equivalent class (arbitrarily, this class is always class *1*): (a) deterministic algorithm and (b) stochastic algorithm. The number of saturated graphs approximated by the stochastic algorithm is $d_1 \times d_2 \times d_3 \times d_4 = 4$.

stochastic version of the algorithm, the *elementary operation* consists of choosing at random one graph among the graphs saturating a given vertex. The same computation is repeated for the selected graph. The stochastic structure generator ends when all vertices have been saturated. The stochastic structure generator is illustrated in Figure 5b.

All the structural formulas that are constructed by the stochastic structure generator have to be structurally different. In other words, all the corresponding saturated graphs must be nonisomorphic. As already mentioned, the equivalent classes algorithm guarantees the nonisomorphism of the saturated graphs. Therefore, two saturated graphs constructed by the stochastic structure generator are nonisomorphic if their series of intermediate unsaturated graphs are different. This requirement can easily be verified if the series of random numbers used to select the intermediate graphs is different each time the stochastic structure generator is run.

One notices that counting the number N of possible structural formulas is equivalent to counting the number of leaves on the backtracking tree represented in Figure 5a. Although it is not possible to exactly calculate the number of leaves of the backtracking tree in a polynomial computational time, Knuth[35] demonstrated that it is possible to approximate this number using an algorithm that runs in a time bounded by λd, where λ is the height of the tree, and d is the degree of the tree (i.e., the maximum number of children of the nodes). The algorithm proposed by Knuth can be implemented using the stochastic structure generator described above. The Knuth algorithm computes all the children of a given node and repeats this computation for one child selected randomly. The algorithm stops when a leaf is reached. Let d_i be the number of children of a given node x_i. Knuth demonstrated that $\hat{N} = \prod_i d_i$ is an unbiased estimation of the number of leaves on any backtracking tree. Applied to our problem, d_i is the number of graphs that can be constructed by saturating a given vertex x_i. To calculate $\hat{N}$, one needs to run the stochastic structure generator and store the d_i values, until a saturated graph is generated. Since $\hat{N}$ can be calculated using the stochastic structure generator; $\hat{N}$ is an unbiased estimation of the number of possible structural formulas.

To find how accurate $\hat{N}$ is, one can repeat the calculation of $\hat{N}$ by running the stochastic structure generator for different series of random numbers. More precisely, the stochastic structure generator can be run until the deviation between the successive $\hat{N}$ numbers is lower than a prechosen value. Expanding on this idea, Jerrum *et al.*[36] and Sinclair[37] proposed a more sophisticated technique. They demonstrated that from a sample of n leaves generated at random, the exact total number of leaves N can be approximated by $\hat{N}$, such that

$$\text{Prob}\left[\frac{\hat{N}}{1+\alpha} \le N \le (1+\alpha)\,\hat{N}\right] \ge 3/4, \qquad \alpha \in (0,1] \quad (7)$$

Furthermore, they proved that the value of n, required to

verify eq 7 is linearly proportional to $\lambda^3 d^3/\alpha^2$, where λ and $d = \max_i d_i$ have the same definition as above. The results of Jerrum *et al.* and Sinclair are theoretically important; nonetheless, when applied to our problem, the sample size required is generally too large. Hence, the preferred technique consists of running the stochastic structure generator until the deviation of $\hat{N}$ reaches a prechosen value.

THE 3D SIMULATIONS

Prior to simulating any three-dimensional physical characteristic, one needs to determine optimum three-dimensional conformations for each model constructed by the stochastic structure generator. The SIGNATURE program does not provide conformational searching capabilities; however, the program is interfaced with several molecular modeling products (cf. Appendix for details). It is the task of the expert using the program to decide which software and which type of molecular orbital calculations or molecular simulations are appropriate. If the simulations carried out reveal several optimum conformations for a given model, these optimum conformations have to be added to the sample of models. Once optimum conformations have been obtained, the 3D simulator calculates for each model of the sample the density, the pore volume distribution, the surface area, and the fractal dimension of the surface.

The analytical techniques that experimentally determine the three-dimensional physical characteristics are listed in Table 1. The methods employed by the SIGNATURE program to simulate the three-dimensional physical characteristics have already been presented and applied in the context of coal macromolecules.[23a-c] The same methods have also been employed to determine the density and the surface area of aryl-bridged polysilsesquioxane models.[23e]

THE STATISTICAL INTERPRETATION

The second task of the SIGNATURE program approximates the number of possible models that can be constructed from a set of 2D analytical data and then constructs a sample of these models. The purpose of the statistical interpretation is to determine if the sample constructed statistically represents the entire population of all possible models. The models constructed are randomly generated and structurally different; therefore, the task performed by the stochastic structure generator is a *simple random sampling without replacement*[38] (SRSWOR).

With the SRSWOR statistical theory,[38a] it is possible to define an optimal size for a sample and extrapolate the mean sample value of a certain characteristic to the whole population (i.e., to all the possible molecular models). According to this theory, a sample of average structures can be defined, representing the studied compound. It is also possible to evaluate certain characteristics of this sample, such as those calculated by the 3D simulator, and to extrapolate these characteristics to the whole population of models.

Consider a population of size N from which a simple random sample of size n is drawn, without replacement. As applied in our study, N is the number of possible molecular models, and n is the number of structures randomly built by the stochastic structure generator. Let x be the sample mean and let X and S^2 be the population mean and variance for a certain characteristic (the true density for example). It is known that x is an unbiased estimation and the following are true

$$X = x \tag{8}$$

$$\mathrm{Var}(x) \leq (N-n)\,S^2/Nn \tag{9}$$

By imposing the restraint

$$\mathrm{Var}(x) \leq V^* \tag{10}$$

for a prechosen constant V^*, n is determined to satisfy eq 9. The n required is

$$n \geq NS^2/(S^2 + NV^*) \tag{11}$$

Prior information on S^2 is needed to determine the sample size using eq 11. Unfortunately such information is lacking in our problem. For example, it is not possible to know the variance of the density, for all the molecular models that can be constructed. In such an instance the following stepwise procedure of Cochran[38b] is used to determined the optimal value for the sample size: 1. Take an initial sample size n_1 (≥ 2). In fact, n_1 is the sample size arbitrarily chosen in task 2, by the expert using the SIGNATURE program. Calculate s_1^2, the variance of this sample. s_1^2 is an unbiased estimation of S^2. 2. Calculate

$$n_2 = (s_1^2/V^*)(1 + 2/n_1) \tag{12}$$

$$n = \max\{n_1, n_2/(n_2/N + 1) + 1\} \tag{13}$$

3. Take $n - n_1$ additional observations if necessary. Note from eq 12 and 13 that $n = n_1$ if

$$V^* \geq s_1^2(1 + 1/N - n_1/N)(1 + 2/n_1)/(n_1 - 1) \tag{14}$$

The procedure of Cochran[38b] correlates the variance of the mean value for a certain characteristic and the sample size. Hence, for an arbitrary sample size, it is possible to determine the mean value for the population (eq 8), the variance of this mean value (eq 14), and the population variance using the following equation (obtained from eq 11):

$$S^2 \leq V^* Nn/(N-n) \tag{15}$$

The value of V^* obtained from eq 14 may be larger than expected because the sample constructed by the stochastic structure generator is too small. In this case, a new sample size is chosen using eq 13. If the new sample size is greater than the initial sample size, additional molecular models are constructed using the stochastic structure generator.

The population mean X may not match the corresponding 3D analytical data for a given characteristic. In such a case the sample constructed by the stochastic structure generator does not represent the studied compound, and a new sample has to be generated. To build a new sample, the expert using the program may decide to select a new list of molecular fragments and interfragment bonds (task 1). The expert can also choose to construct molecular models with specific structural constraints during the stochastic structure generation (task 2).

The stochastic technique used in the present paper is based on the SRSWOR sampling theory. However, other stochastic methods can be employed, for example, SA, MC, or GA. GA (*genetic algorithm*) is particularly interesting because it can eliminate the feedback loop of the SIGNATURE program that links the statistical interpretation to the expert using the system. More precisely, let us assume that a sample of models has been constructed by the SIGNATURE program. In a GA scheme a fitness is calculated for each model to measure how accurately the model matches the overall set of analytical

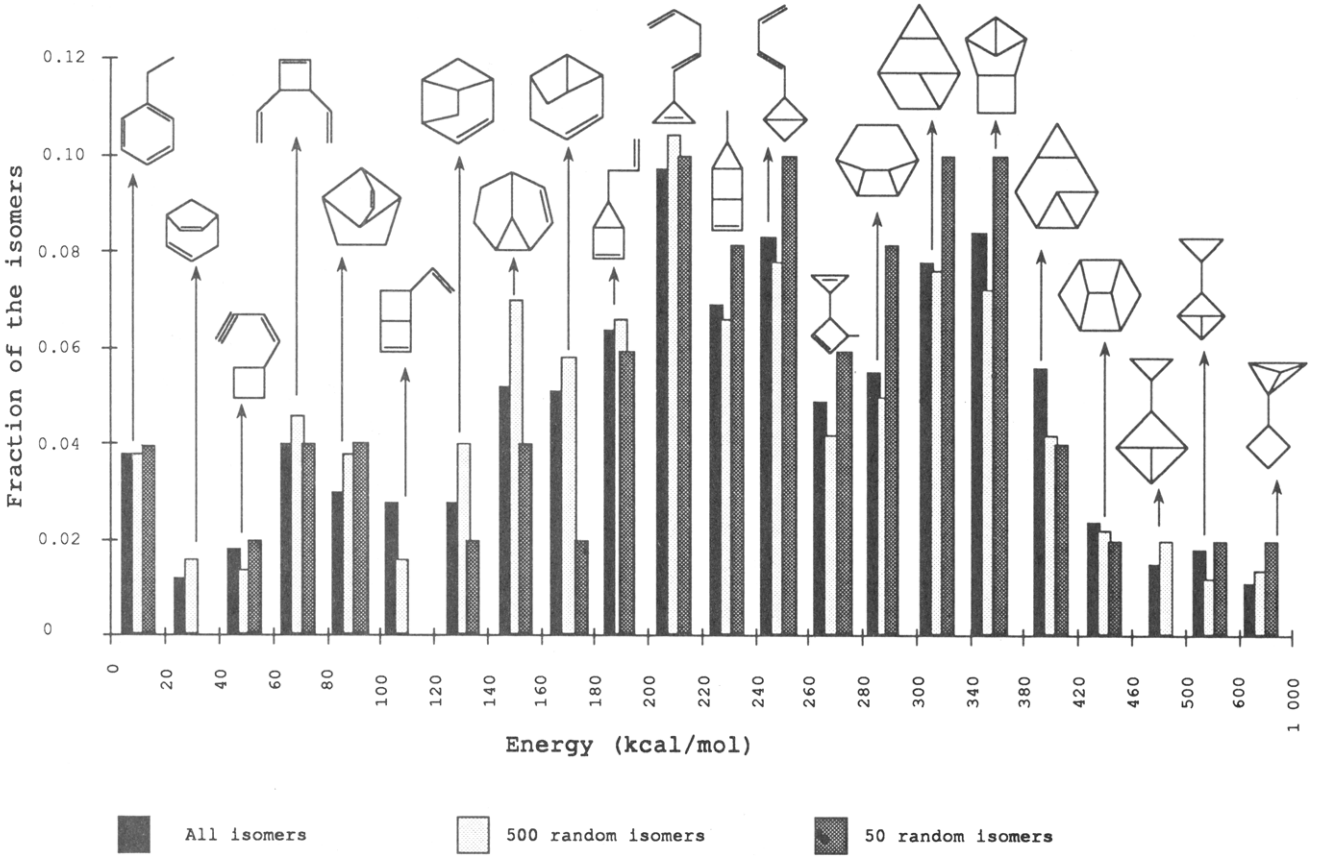


Figure 6. Potential energy distribution of the isomers of C_8H_{10}. The potential energies were calculated using the MM technique described in the text. For each energy range, one example of the structural isomers is drawn.

data. To guarantee the survival of the fittest, models with bad fitness are rejected. Then, by mating and mutating the resulting models, a new generation of models is generated. GA is recursively applied until the overall fitness of the models reaches a desirable value. The implementation of GA into the SIGNATURE program is currently studied.

RESULTS AND DISCUSSION

The purpose of this section is to present four examples of applications of the SIGNATURE program: these are the study of the potential energy distribution of the isomers of C_8H_{10} and the structural elucidation of gymnosperm lignin, vitrinite from coal, and arylene-bridged polysilsesquioxane. The reader interested by the chemical and structural details regarding the last three compounds can consult the following references for coal,[23a-c] lignin,[23d] and polysilsesquioxane.[23e]

The first example is a study case that was designed to quantify how well the stochastic structure generator performs. In this example, the potential energy distribution of the isomers of C_8H_{10} were analyzed using the deterministic and the stochastic versions of the structure generator presented above. One may notice that in the present example there is no need to resolve the signature equation. In fact, the structures to be generated are all composed of eight carbon atoms and 10 hydrogen atoms; these atoms constitute the list of molecular fragments. The deterministic version of the structure generator computed 4008 different isomers of C_8H_{10} (not including eventual stereoisomers). To generate these isomers the deterministic version of the structure generator ran for 353.0 s CPU time on a SGI Personal Iris Workstation. For comparison, the stochastic version of the structure generator estimated the number of isomers to be 3399. To calculate this number the Knuth algorithm was run for 16.5 s CPU time until the deviation between the estimates was lower than 1000.

To compute the energy distribution, the potential energy of each isomer was calculated using the MM simulations provided by Polygraf 3.21 (Molecular Simulations Inc.). All energies were minimized using the DREIDING force field[39] and using a conjugate gradient algorithm for 500 steps or until the root mean square between two successive conformations was lower than 0.1 (kcal/mol)/Å. The potential energy distribution of the 4008 isomers is shown in Figure 6. This distribution was obtained in 114 285 s of CPU time on a SGI Personal Iris Workstation. Most of the time was spent minimizing the potential energy (as mentioned above, it took only 353 s to generate all the isomers without minimization). Using the same hardware, it took 12 223 s to generate a potential energy distribution from a random sample of 500 isomers and 1344 s for a random sample of 50 isomers. Let f be the fraction of isomers of the total population having their potential energy in any given range of Figure 6. Let f_{500} be the same fraction obtained from the sample of 500 isomers, and let f_{50} be the fraction obtained from the sample of 50 isomers. It can be shown from Figure 6 that on average $|f_{500} - f| = 0.17f$, and $|f_{50} - f| = 0.35f$. Therefore, the sample of 50 isomers gives an acceptable approximation of the potential energy distribution. The same calculations were performed for all the hydrocarbons C_nH_{n+i}, with n varying between 2 and 8 and i varying between 0 and 2. For each of the previous hydrocarbons, it was concluded that a good approximation of the potential energy distribution can be obtained by generating a sample that represents only a small fraction of the total population.

Table 2. Signatures of the 2D Quantitative Analytical Data for a Gymnosperm Lignin Compound[a]

Analytical technique	Parameter[b]	Amount[b]	FREL[c]	0-signature[c]	1-signature[c]
EA	H	110.0	$H-$	110.0 h	
				+	
EA	O	36.0	$O-$	36.0 o	
				+	
NMR	f_a	58.0	$-C=$	58.0 cp	
				+	
NMR	f_{al}	42.0	$-C-$	42.0 c	
NMR	f_a^H	29.0	$H-C=C$		29.0 cp (cp cp h)
				+	
NMR	f_a^P	17.4	$O-C=C$		17.4 cp (cp cp o)
				+	
NMR	% OCH_3	8.2	$O-C-H$		8.2 c (o h h h)

[a] The analyzed sample is a gymnospermous degraded wood and was composed of lignin (75%) and of cellulosic materials (25%). The listed values taken from ref 29 correspond only to the lignin fraction: EA = elemental analysis and NMR = CPMAS ^{13}C solid state NMR. [b] f_a = aromatic carbon, f_{al} = aliphatic carbon, f_a^H = protonated aromatic carbon, and f_a^P = phenolic or phenolic ether. [c] Cf.: text and Figure 2 for notation.

The second example is the structural elucidation of lignin. Lignin is a biopolymer, which is an important constituent of terrestrial plants whose chemistry and reactivity is of great importance in many areas of energy research and industry. Lignin is thought to be biosynthesized through a random polymerization. 2D quantitative data relative to lignin have already been presented in Table 2. Qualitative 2D data were taken from the literature and were derived using 2D NMR techniques. Almost no 3D data are published for lignin; hence, our goal with this example is to find the lowest energy structures that match the 2D data. From the list of molecular fragments and interfragment bonds compiled in Table 3, the SIGNATURE program was directed to construct a sample of 10 random structures, each containing 27 monomers. The stochastic structure generator estimates that there were 1 481 760 possible different structural models (not including stereoisomers). From this population size and the sample generated, a statistical interpretation was carried out for the potential energy. The minimized potential energy was calculated for each model comprising the sample using the *quenched annealed dynamics* (QAD) molecular simulations provided by Polygraf 3.21. QAD were run using the DREIDING force field.[39] The initial temperature was set to 1200 K and then lowered from 1200 to 300 K, at 5 K/0.1 ps. The time step chosen for the integration of the Newton's equation of motion was 1 fs, and the QAD simulations were carried out for 60 ps. The minimized potential energy distribution of the 10 random lignin models is given in Figure 7. The population mean and deviation for the potential energy were calculated using eq 7, 13, and 14. The average potential energy was found to be 2041 ± 445 kcal/mol. The population deviation was found to be 1406 kcal/mol. Because of this high population deviation, one may wonder if lignin is really a random polymer. Furthermore, even with a small sample of 10 models, Figure 7 shows that there are lignin structures that have a much lower potential energy than the other.

Table 3. Best Solution of the *Signature Equation* for a Gymnosperm Lignin Model Containing 27 Monomers

molecular fragments[a]	amt	interfragment bonds[b]	amt
guaiacyl	20	β–O-4	18
p-hydroxyphenyl	7	β–5	4
syringyl	0	β–1	1
–OH	21	β–β	1
–H	67	α–O-γ	6
		5–5	4
		OH–α	21
		H–β	2
		H–5	15
		H–O-4	9
		H–O-γ	21

Deviation (Normalized for 100 Carbon Atoms)

$\sigma(S)^c$ =	ΣX^d =	$\lvert\Sigma X - \sigma(S)\rvert$ =
42.0 c	38.0 c	4.0 c
+	+	+
58.0 cp	62.0 cp	4.0 cp
+	+	+
110.0 h	111.1 h	1.1 h
+	+	+
36.0 o	35.8 o	2.0 o
+	+	+
29.0 cp (cp cp h)	29.2 cp (cp cp h)	0.2 cp (cp cp h)
+	+	+
17.4 cp (cp cp o)	18.1 cp (cp cp o)	0.7 cp (cp cp o)
+	+	+
8.2 c (o h h h)	7.7 c (o h h h)	0.5 c (o h h h)

[a] Cf.: Figure 3 and corresponding caption. [b] Cf.: Figure 4 and corresponding caption. [c] $\sigma(S)$ is the *signature* of the unknown compound (cf. Table 2). [d] ΣX is the sum of the *signatures* of the molecular fragments and interfragment bonds that are the solutions of the *signature equation*.

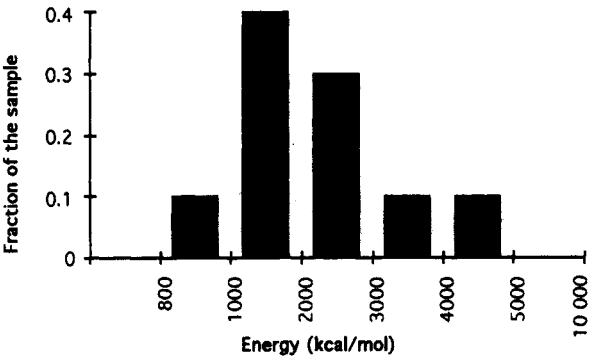


Figure 7. Potential energy distribution for a sample of 10 random lignin models. The potential energies were calculated using the QAD technique described in the text.

Because the major interfragment bond in lignin is the β–O-4 linkage type, the conformation of guaiacyl oligomers linked with this interfragment bond was systematically studied. There are two diastereoisomers with β–O-4 guaiacyl dimers, *threo* and *erythro*. This is due to the fact that the α and β carbon atoms are achiral (cf. Figure 4). For the same reason, there are 16 stereoisomers with β–O-4 guaiacyl trimers. All the stereoisomers of β–O-4 guaiacyl trimers were constructed, and their conformations were studied using QAD, as described previously. It was found that the stereoisomers *all-threo* and *all-erythro* had lower minimized energies than the others. There are 262 144 stereoisomers with β–O-4 guaiacyl decamers; it was not possible to analyze all of them. However, using the SIGNATURE program, a sample of 10 random stereoisomers was constructed. Each model was submitted to QAD. The average minimized potential energy of the random stereoisomers was found to be equal to 285 kcal/mol, while the minimized potential energy of the specific stereoisomers *all-threo* and *all-erythro* was lower than all the others. This

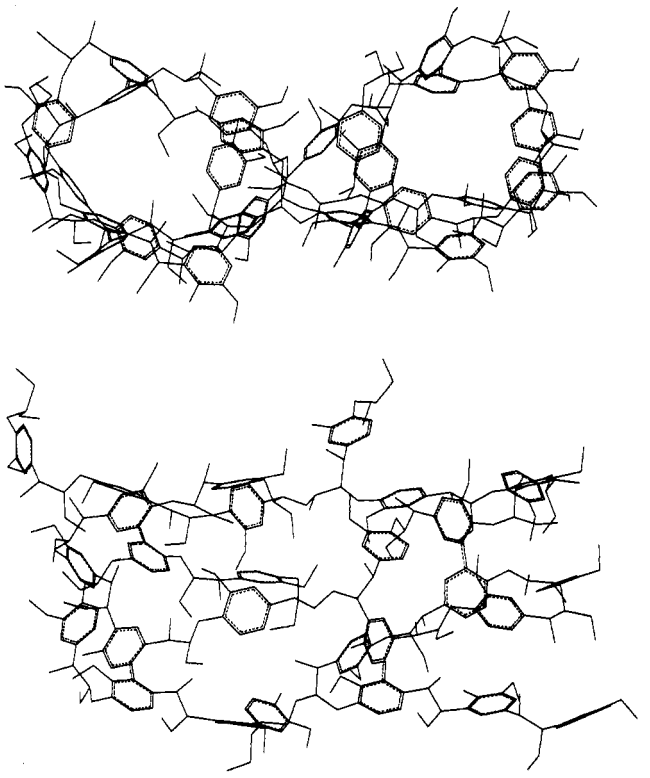


Figure 8. Proposed three-dimensional structural model for gynmospermous lignin. Molecular fragments and interfragment bonds forming the model are those listed in Table 3. The top figure is a three-dimensional representation of the structure in plain view. The bottom figure is the same structure viewed from the side. Both views have been obtained using the QAD simulations described in the text. Hydrogen atoms have been omitted for clarity of presentation.

energy was found to be equal to 250 kcal/mol. Therefore, it was concluded that the stereospecific *all-threo* and *all-erythro* oligomers are energetically favored. Furthermore, the most astonishing result was obtained by depicting the structures on a graphic screen; the lowest energy conformations of the *all-threo* and *all-erythro* β–O-4 guaiacyl oligomers displayed a helical order.

The SIGNATURE program was then directed to construct one structural model starting this time from a template containing two disconnected *all-erythro* helical oligomers. The initial helical structures were modified in order to match the quantities of molecular fragments and interfragment bonds given in Table 3. Like the 10 random structures, the resulting structure belonged to the 1 481 760 possible structural models. This structure was then submitted to QAD. The minimum of the potential energy (725.3 kcal/mol) was found to be much lower than the potential energies calculated for the 10 random structures. Furthermore, the helical order was preserved after molecular simulations (cf. Figure 8). It was concluded that according to the molecular simulations performed, a lignin structure constructed from a helical template appeared to be energetically favored over random structures.

The third example is the structural elucidation of vitrinite from bituminous coal. Coal is chemically and physically a highly heterogeneous material consisting mainly of organic matter (macerals) and some inorganic materials (minerals). The amount, the distribution, and the chemical structure of various macerals in coals depends on the chemical nature of the original coal-forming material and the conditions of coalification. The nature of each phase progressively changes during coalification (i.e., with rank). Vitrinite is derived from plant debris and is one of the major macerals in most American coals. As revealed by Py/GC/MS, at the bituminous rank, vitrinite is mainly composed of alkylbenzenes, alkylphenols, alkylnaphthalenes, and alkyldibenzofurans. The SIGNATURE program was asked to construct structural models for vitrinite at the bituminous rank. The models, arbitrarily composed of 333 carbon atoms, were constructed from a list of molecular fragments and interfragment bonds derived from Py/GC/MS results. The 2D quantitative data input into the program were given by elemental analysis and diverse solid state ^{13}C NMR techniques. At first, a sample of five structural models was constructed by the stochastic structure generator. The sample was generated without any additional structural constraints, and the number of possible models was estimated to be exponentially proportional to the number of fragments (e.g., 10^{60}). Each model comprising the sample was submitted to MD molecular simulations using the same procedure as for the QAD simulations mentioned above, with the exception that the temperature was maintained fixed at 300 K. For the five models constructed, helium density and micropore volume were calculated by the 3D simulator. The statistical analysis carried out on the five models clearly demonstrates that the structures constructed did not match 3D experimental data. The helium density was too high compared to the experimental values reported for vitrinite. No micropores were found in the models, which contradicted the fact that coal is a microporous material. Finally, the potential energy was at least 25 times larger than values reported by other molecular simulations studies of bituminous coal.[6] There are indirect supports from liquefaction experiment[10c] that coal is composed of clusters larger than the pyrolysis fragment introduced into the SIGNATURE program. These clusters are commonly named hydroaromatic clusters and were not present in the five models constructed. Hence, in a second attempt, the stochastic structure generator was constrained to build models by forming the maximum number of five or six membered rings. The purpose of this constraint was to generate larger clusters from the initial molecular fragments. The stochastic structure generator estimated that the number of possible models was equal this time to 319 318. A sample of 15 structural models was randomly constructed. Each model was submitted to MD molecular simulations. The helium density, the micropore volume, the micropore distribution, the surface area, and the fractal dimension of the surface were calculated for each model. The statistical interpretation demonstrated that the sample of 15 structures agreed with 3D analytical data. Further, the energetics and three-dimensional physical characteristics found for the different models were relatively close to each other, and the deviations were small. Therefore, it was concluded that for the characteristics investigated, a sample of 15 structures was sufficient to statistically represent the whole population of vitrinite models from bituminous coal.

The last study case is the structural elucidation of aryl-bridged polysilsesquioxane. Arylene-bridged polysilsesquioxanes have recently been synthesized by the hydrolysis and condensation of bis(triethoxysilyl)arylbenzene or terphenyl monomers.[40] Like sol-gel processed silicas, the arylene-bridged polymers are microporous, glass-like materials composed of aggregated particles. Surface areas of the dried gels (*xerogels*), determined by nitrogen and argon gas sorption porosimetry, ranged from 256–1000 m^2/g, and the majority of the porosity lies in the micropore region (mean pore diameters <20 Å). The SIGNATURE program was directed to construct structural models using elemental analysis and ^{29}Si NMR as 2D quantitative data, and the bis(triethoxysilyl)-

aryl benzene or terphenyl monomers as 2D qualitative data. When the structure generator was asked to construct models without additional structural constraints, the number of possible models was found to be exponentially proportional to the number of monomers (e.g., greater than 10^{16}). Nonetheless, a sample of 10 random models was constructed, and all the models displayed a lack of porosity. When the structure generator was constrained to form polycondensed cyclic systems the resulting sample of 10 random models matched the overall set of analytical data. Furthermore, it was demonstrated[23e] using graph theory that if polysilsesquioxane is indeed composed of polycondensed cyclic systems, the average number of silicon atoms ranges from 4 to 8. This study lead to the conclusion that polysilsesquioxane may not consist of random three-dimensional networks but may rather consist of aggregations of two-dimensional shell particles formed by polycondensed cyclic systems.

The above examples demonstrate that the SIGNATURE program is useful in the real world of structure elucidation. With the example of lignin, the program found that there are some molecular structures that have much lower potential energies than the others. Furthermore, these structures were found to be not as random and disordered than was generally believed. One can be confident that experiments (based for example, on Raman spectroscopy or ^{13}C NMR labeling) can be designed to verify or deny the helical order found by the program. With lignin, the SIGNATURE program plays fully its role of structure elucidation system by suggesting new ideas for studying this structure. With the examples of coal and polysilsesquioxane the program validated experimental evidence, which suggests that these structures are composed of polycondensed cyclic systems. In fact, when the program was asked to construct coal and polysilsesquioxane models without additional constraints the program did not find models that match the corresponding 3D analytical data. There are two situations where a sample of random models constructed from a set of 2D analytical data does not match the corresponding 3D analytical data. (1) 2D and 3D data are not compatible. (2) The 2D data introduced into the SIGNATURE program are insufficient to represent the structure of the unknown compound.

The first situation occurs when there is an experimental problem; for example, when the experimental samples used to derive the 2D data are different than the samples used to determine the 3D data. This situation is resolved by redesigning the experiments. It is easy to detect when the second situation occurs; if the 2D data do not provide enough information the number of possible models is extremely large (such as in the cases of coal and polysilsesquioxane). The second situation is solved by adding new 2D information taken, for example, from the literature. This information can be either input as new 2D quantitative or qualitative data or can be input as a structural constraint. As already mentioned, in the cases of coal and polysilsesquioxane there is experimental evidence, which suggests that these structures are composed of polycondensed cyclic systems. In both cases, when the SIGNATURE program was constrained to construct cyclic systems, the resulting samples of models match the overall set of analytical data.

CONCLUSION

In biochemistry, geochemistry, petroleum, fuel, and materials sciences there are needs for computer programs that elucidate large chemical structures. The stochastic structure generator presented in this paper is the first system able to correlate a large diversity of analytical data and three-dimensional macromolecular models. The main advantage that a stochastic approach has compared to a deterministic approach is a significant gain of computational time. In fact, the computational complexity of structure elucidation problems increases exponentially with the number of atoms (or molecular fragments), and structure elucidation probably belongs to a class of intractable computational problems.[41] Hence, as in the case of the traveling salesman computational problem, it is very improbable that an efficient deterministic algorithm will ever be found to resolve the problem of structure elucidation. Nonetheless, the present paper demonstrates that an efficient stochastic algorithm can approximate the solutions of structure elucidation. More precisely, a stochastic algorithm can be designed to run in a time polynomially proportional to the height and the degree of the backtracking tree of the solutions of structure elucidation. Consequently, the stochastic approach that has been presented makes it possible to study large molecular systems. In short, the SIGNATURE program appears to be an essential tool for anyone who wants to use molecular modeling techniques for a structurally unresolved macromolecular compound.

ACKNOWLEDGMENT

I am pleased to acknowledge the funding provided by the Institut Français du Pétrole, the U.S. Department of Energy, Sandia National Laboratories under contract DE-AC04-76DP00789, and Associated Western Universities. I am grateful to Dr. G. A. Carlson, Dr. P. G. Hatcher, Dr. J. M. Drappier, and M. Vandenckoucke for encouragement and constructive discussions.

APPENDIX

Prior to running the SIGNATURE program, the user has to enter the 2D structural data (cf. Figure 1). The 2D quantitative data are input using the signature notation (cf. section entitled The Signature Equation). These data are stored in a file format specific to the SIGNATURE program. The 2D qualitative data are composed of molecular fragments and interfragment bonds. Fragments and bonds are constructed using a molecular modeling software and stored in a library. The molecular modeling products that are currently interfaced with the SIGNATURE program are PCMODEL (Serena Software), POLYGRAF 3.21 (Molecular Simulations Inc.), and INSIGHT 2.3 (Biosym). Files can be read and written in the specific formats of these programs. The above commercial molecular modeling products are also used to optimize the conformations of the chemical structures built by the SIGNATURE program.

The SIGNATURE program is run in four steps. In the first step, from a file containing the 2D quantitative data and a library of molecular fragments and interfragment bonds, the signature equation is solved. This equation can lead to several solutions; each solution is stored in a separate file containing the model to be constructed (i.e., a list of fragments and bonds selected from the library) and the deviation between the model and the 2D quantitative data. In the second step, a sample of 3D models is constructed from the file containing the smallest deviation between model and 2D quantitative data. The output of the second step is a set of files each containing one 3D model and the number of all possible models (population size). In the third step, the conformations of the 3D models are optimized using the molecular modeling products mentioned above. Then, the 3D simulator of the SIGNATURE program is run to determine for each 3D

optimized model its density, pore volume distribution, surface area, and fractal dimension. The fourth step is the statistical interpretation. This step is not automated. However, the outputs of the 3D simulator can be read by the Microsoft Excel program, and this program can be used to automate the statistical equations (eq 8–15).

The SIGNATURE program is written in standard C computer programming language. The program has to be run with the UNIX operating system. The program has been implemented on Power Series, Personal Iris, and Indigo2 SGI workstations, SUN/SPARC workstations, and a CRAY YMP supercomputer. The size of the program including libraries of fragments and bonds and tutorial files is about 10 MB. A free copy of the program can be obtained by writing to the author of the present paper.

REFERENCES AND NOTES

(1) (a) Feller, D.; Davidson, E. R. Basic Sets for Ab Initio Molecular Orbital Calculations and Intermolecular Interactions. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 1–44.

(2) (a) Stewart, J. J. P. Semiempirical Molecular Orbital Methods. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 1, pp 45–82.

(3) (a) Burkert, U.; Allinger, N. L. *Molecular Mechanics*; ACS Monograph 177; American Chemical Society: Washington, DC, 1982. (b) Leach, A. R. A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Vol. 2, pp 1–56. See also references therein. (c) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. Conformational Searching Methods for Small Molecules. II. Genetic Algorithm Approach. *J. Comput. Chem.* **1993**, *14*, 1407–1414. See also references therein.

(4) This observation is also valid for *ab initio* and semiempirical calculations.

(5) (a) Gravitis, J.; Erins, P. J. Toplogical and Conformational Structure and Macroscopic Behavior of Lignin. *Appl. Polym. Sci.: Appl. Polym. Symp.* **1983**, 37, 421. (b) Elder, T. Application of Computational Methods to the Chemistry of Lignin. In *Lignin Properties and Materials*; Glasser, W. G., Sarkanen, S., Eds.; ACS Symposium Series 397, Washington, DC, 1989; pp 262. (c) Elder, T. Molecular Simulations of Lignin Oligomers. In *Viscoelasticity of Biomaterials*; Glasser, W. G., Hatakoyama, H., Eds.; ACS Symposium Series 489: Washington, DC, 1992; pp 370.

(6) (a) Carlson, G. A.; Granoff, B. Modeling of Coal Structure using Computer-Aided Molecular Design. In *Coal Science II*; Schobert, H. H., Bartle, K. D., Eds.; ACS Symposium Series 461, American Chemical Society: Washington, DC, 1991; pp 159. (b) Carlson, G. A. Computer Simulation of the Molecular Structure of Bituminous Coal. *Energy Fuels* **1992**, *6*, 771. (c) Nakamura, K.; Murata, S.; Nomura, M. CAMD Study of Coal Molecules. 1. Estimation of Physical Density of Coal Molecules. *Energy & Fuels* **1993**, *7*, 347–350. (d) Murata, S.; Nomura, M.; Nakamura, K.; Kumagai, H.; Sanada, Y. CAMD Study of Coal Model Molecules. 2. Density Simulation for Four Japanese Coals. *Energy & Fuels* **1993**, *7*, 469–472.

(7) Rustad, J. R.; Yuen, D. A.; Spera, F. J. Molecular Dynamics of Amorphous Silica at Very High Pressures (135 GPa): Thermodynamics and Extraction of Structures Through Analysis of Voronoi Polyhedra. *Phys. Rev. B* **1991**, *44*, 2108–2121. See also references therein.

(8) (a) Pearl, I. A. *The Chemistry of Lignin*; Dekker M: New York, 1967. (b) Adler, E. Lignin-Past, Present and Future. *Wood Sci. Technol.* **1977**, *11*, 169. (c) Glasser, W. G.; Glasser, H. R. Evaluation of Lignin's Structure by Experimental and Computer Simulations Techniques. *Pap. Puu* **1981**, *63*, 71.

(9) Behar, F.; Vandenbroucke, M. Chemical Modelling of kerogens. *Org. Geochem.* **1987**, *11*, 15–24.

(10) (a) Given, P. H. The Distribution of Hydrogen in Coals and its Relation to Coal Structure. *Fuel* **1960**, *39*, 147. (b) Solomon, P. R. Coal Structure and Thermal Decomposition. In *New Approaches in Coal Chemistry*; ACS Symposium Series No. 169; American Chemical Society: Washington, DC, 1981; 61. (c) Shinn, J. H. From Col to Single-Stage and Two-Stage Products: A reactive Model of Coal Structure. *Fuel* **1981**, *63*, 1187. (d) Hatcher, P. G.; Faulon, J. L.; Wenzel, K. A.; Cody, G. D. A Structural Model for Lignin-Derived Vitrinite from High-Volatile Bituminous Coal (Coalified Wood). *Energy Fuels* **1992**, *6*, 813.

(11) CASE is also the name of the structure elucidation system developped by Munk and co-workers, see refs 14 and 17. In the present paper the therm CASE is only employed to designate computer-assisted structure elucidation.

(12) (a) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. The number of Possible Organic Compounds. Acyclic Structures Containing C, H, O and N. *J. Am. Chem. Soc.* **1969**, *91*, 11, 2973–2976. (b) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; John Willey & Sons: New York, 1986. (c) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1975**, *97*, 20, 5755–5762. (d) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Application of Artificial Intelligence for Chemical Inference. 17. An Approach to Computer-Assisted Elucidation of Molecular Structure *J. Am. Chem. Soc.* **1975**, *97*, 5755–5762. (e) Carhart, R. E.; Smith, D. H.; Gray, N.; Nourse, J. G.; Djerassi, C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708–1718. (f) Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Crandell, C. W. The Dendral Project: Recent Advances in Computer-Assisted Structure Elucidation. *Anal. Chim. Acta* **1981**, *133*, 471–497.

(13) (a) Kudo, Y.; Sasaki, S. The connectivity Stack, a New Format for Representation of Organic Chemical Structures. *J. Chem. Doc.* **1974**, *14*, 200–202. (b) Kudo, Y.; Sasaki, S. Principle for Exhaustive Enumeration Of Unique Structure Consistent with Structural Information. *J. Chem. Inf. Comput. Sci.* **1975**, *16*, 43–49. (c) Oshima, T.; Ishida, Y.; Saito, K.; Sasaki, S. Chemics-UBE, A Modified System of Chemics. *Anal. Chim. Acta* **1980**, *122*, 95–102. (d) Abe, H.; Okuyama, T.; Fujiwara, I.; Sasaki, S. A Computer Program for Generation of Constitutionally Isomeric Structural Formulas. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 220–229. (e) Sasaki, S.; Kudo, Y. Structure Elucidation System Using Structural Information from Multisources: CHEMICS. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 252–257. (f) Funatsu, K.; Miyabayaski, N.; Sasaki, S. Further Cevelopment of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem Inf. Comput. Sci.* **1988**, *28*, 18–28.

(14) (a) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121–132. (b) Shelley, C. A.; Munk, M. E. CASE, a Computer Model of the Structure Elucidation Process. *Anal. Chim. Acta* **1981**, *133*, 507–516. (c) Lipkus, A. H.; Munk, M. E. Automated Classification of Candidate Structures for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9–18.

(15) (a) Dubois, J. E.; Carabedian, M.; Ancian, B. Automatic structural elucidation by carbon-13 NMR: DARC-EPIOS method. Search for a discriminant chemical structure-displacement relationship. *C. R. Acad. Sci. Ser. C* **1980**, *290*, 369–372. (b) Dubois, J. E.; Carabedian, M.; Ancian, B. Automatic structural elucidation by carbon-13 NMR: DARC-EPIOS method. Description of progressive elucidation by ordered intersection of sub-structures. *C. R. Acad. Sci. Ser. C* **1980**, *290*, 383–386. (c) Carabedian, M.; Dagane I.; Dubois, J. E. Elucidation by Progressive Intersection of Ordered Substructures from Carbon-13 Nuclear Magnetic Resonance. *Anal. Chem.* **1988**, *60*, 2186–2192.

(16) Bremser, W.; Fachinger, W. *Magn. Reson. Chem.* **1985**, *23*, 1056.

(17) (a) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87–93. (b) Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Fragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531–540.

(18) Bangov, I. P.; Computer-Assisted Structure Generation from a Gross Formula. 7. Graph Isomorphism: A Consequence of the Vertex Equivalence. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277–289.

(19) Contreras, M. L.; Rozas, R.; Valdivias, R. Exhaustive Generation of Organic Isomers. 3. Acyclic, Cyclic, and Mixed Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 610–616.

(20) Although there are algorithms for particular graphs such as trees and planar graphs that resolve the problem of graph-isomorphism in a polynomial time, for chemical structures the best known algorithms are asymptotically exponential, cf.: Razinger, M.; Balasubramanian, K.; Munk, M. E. Graph Automorphism Perception Algorithms in Computer-Enhanced Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 197–201.

(21) Faulon, J. L. On Using Molecular Graph-Equivalent Classes for the Structure Eluciation of Large Molecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 338–348.

(22) SIGNATURE stand for StochastIc GeNerator of chemicAl strucTURE.

(23) (a) Faulon J. L.; Hatcher P. G.; Carlson, G. A.; Wenzel, K. A. A computer-aided Molecular Model for High Volatile Bituminous Coal. *Fuel Processing Technology* **1992**, *34*, 277–293. (b) Faulon, J. L.; Carlson, G. A.; Hatcher, P. G. Statistical Model for Bituminous Coal: A Three-Dimensional Evaluation of Structural and Physical Properties Based on Computer-Generated Structures. *Energy Fuels* **1993**, *7*, 1062–1072. (c) Faulon, J. L.; Mathews, J. P.; Carlson, G. A.; Hatcher, P. G. Correlation between microporosity and fractal dimension of bituminous coal based on computer-generated models. *Energy Fuels* **1994**, *8*, 408–414. (d) Faulon, J. L.; Hatcher, P. G. Is there any order in the structure of Lignin? *Energy Fuels* **1994**, *8*, 402–407. (e) Faulon, J. L.; Loy, D. A.; Carlson, G. A.; Shea, K. J. Computer-aided Structure Elucidation for Arylsilsesquioxane Gels. *Comput. Mater. Sci.*, in press.

(24) Greenwood, P. F.; Zhang, E.; Vastola, F. J.; Hatcher, P. G. Laser Micropyrolisis Gas Chromatography/Mass Spectrometry of Coal. *Anal. Chem.* **1993**, *65*, 1937–1946.

(25) Weininger, D.; Weininger A. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

(26) (a) Faulon, J. L.; Vandenbroucke, M.; Drappier, J. M.; Behar, F.; Roméro, M. Modelization of Sedimentary Macomolecules: the Software X-MOL, *Rev. I.F.P.* **1990**, *45*, 161. (b) Faulon, J. L.; Vandenbroucke, M.; Drappier, J. M.; Behar, F.; Romero, M. 3D Chemical Model for Geological Macromolecules *Adv. Org. Geochem.* **1991**, *16*, 981. (c) Faulon, J. L. Prediction, Elucidation and Molecular Modeling: Algorithms and Applications in Geochemistry. Ph.D. Thesis, Ecole des Mines, Paris, 1991.
(27) Trinajstic, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(28) (a) FREL stand for **F**ragment **R**educed to an **E**nvironment that is **L**imited. (b) Dubois, J. E.; Panaye, A.; Attias, R. DARC System: Notion of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74–82.
(29) Hatcher, P. G. Chemical Structural Models for Coalified Wood (vitrinite) in Low Rank Coal. *Org. Geochem.* **1991**, *16*, 959.
(30) (a) Schrijver, A. *Theory of Linear and Integer Programming*. John Wiley & Sons: New York, 1986. (b) Stougie, L. *Design and Analysis of Algorithm for Stochastic Integer Programming*; CWI Tracts: Amsterdam, 1987.
(31) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
(32) Cf.: vertex algorithm in Table 1 in ref 21.
(33) The proof of the nonisomorphism is given by the theorems 1 and 2 in ref 21. To guarantee the nonisomorphism, the edges that are added by the equivalent classes algorithm must not be equivalent to the edges of the initial graph. In other words, the interfragment bonds must not already be present in the initial molecular fragments. In all the study cases made using the SIGNATURE program, this restriction was verified. In fact, the above restriction is easy to verify in the real world of structure elucidation. The molecular fragments are obtained by chemical or thermal degradation. During the degradation, specific bonds are broken (i.e., the interfragment bonds), and consequently these bonds are not present in the molecular fragments.
(34) For detailed algorithm cf. *classes algorithm* in Table 2 in ref 21.
(35) Knuth, D. E. Estimating the Efficiency of Backtrack Programs. *Mathematics of Computation* **1975**, *29*, 121–136.
(36) Jerrum, M. R.; Valiant, L. G.; Varizani, V. V. Random Generation of Combinatorial Structures from a Uniform Distribution. *Theoret. Comput. Sci.* **1986**, *43*, 169–188.
(37) Sinclair, A. *Algorithms for Random Generation and Counting: A Markov Chain Approach*; Brikhäuser: Boston, 1993.
(38) (a) Hajek, J. *Sampling from a Finite Population*. Marcel Dekker: New York, NY, 1981. (b) Cochran, W. G. *Sampling Techniques*, 3rd ed.; John Willey & Sons: New York, NY, 1977.
(39) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897.
(40) Shea, K. J.; Loy, D. A.; Webster, O. Arylsilsesquioxane Gels and Related Materials. New Hybrids of Organic and Inorganic Networks. *J. Am. Chem. Soc.* **1992**, *114*, 6700.
(41) Garey, M. R.; Johnson, D. S. *Computers and Intractability*; Freeman: San Fransisco, 1979.