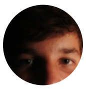
Identifying residues in natural organic matter through spectral prediction and pattern matching of 2D NMR datasets. Magn Reson Chem



MAGNETIC RESONANCE IN CHEMISTRY
Magn. Reson. Chem. 2004; **42**: 14–22
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/mrc.1308

MRC

Identifying residues in natural organic matter through spectral prediction and pattern matching of 2D NMR datasets

Andre J. Simpson,[1]* Brent Lefebvre,[2] Arvin Moser,[2] Antony Williams,[2] Nicolay Larin,[2] Mikhail Kvasha,[2] William L. Kingery[3] and Brian Kelleher[3]

[1] Department of Physical and Environmental Sciences, University of Toronto, Scarborough Campus, Toronto, Ontario M1C 1A4, Canada
[2] Advanced Chemistry Development Inc., 90 Adelaide Street West, Suite 600, Toronto, Ontario M5H 3V9, Canada
[3] Department of Plant and Soil Sciences, Mississippi State University, Starkville MS 39762, USA

Received 24 June 2003; Revised 27 August 2003; Accepted 30 August 2003

This paper describes procedures for the generation of 2D NMR databases containing spectra predicted from chemical structures. These databases allow flexible searching via chemical structure, substructure or similarity of structure as well as spectral features. In this paper we use the biopolymer lignin as an example. Lignin is an important and relatively recalcitrant structural biopolymer present in the majority of plant biomass. We demonstrate how an accurate 2D NMR database of ∼600 2D spectra of lignin fragments can be easily constructed, in ∼2 days, and then subsequently show how some of these fragments can be identified in soil extracts through the use of various search tools and pattern recognition techniques. We demonstrate that once identified in one sample, similar residues are easily determined in other soil extracts. In theory, such an approach can be used for the analysis of any organic mixtures. Copyright © 2004 John Wiley & Sons, Ltd.

KEYWORDS: NMR; 2D databases; pattern matching; complex mixtures; environmental chemistry

INTRODUCTION

Natural organic matter (NOM) in soils is the most abundant reserve of carbon on the Earth, and is central to biodiversity, water quality, sustainable food production and global warming. Knowledge of chemical structures within soils is vital to an understanding of the reactivity and role of natural organic matter in major environmental processes. However, resolution of these highly complex mixtures represents a major challenge to contemporary analytical science. While progress has been made employing a range of chemical and spectroscopic approaches, the techniques involved often provide only basic compositional data, or are highly selective and unrepresentative of the whole sample.[1] NMR spectroscopy allows the investigation of natural mixtures in their entirety. It has become common practice to extract soils, mostly with a base (sodium hydroxide), to yield an organic extract. Solution-state NMR has proven to be a very versatile and powerful technique for the study of these extracts.[2–16] However, the 1D NMR datasets of soil extracts contain considerable and often continuous overlap. Spectral matching of these 1D data sets is virtually impossible. Thus, when working with very complex natural mixtures, it is essential to introduce at least a second dimension in order to reduce the spectral overlap. However, even with two dimensions the datasets produced are extremely complicated, making interpretation by hand exceedingly time consuming, or in many cases impossible. Therefore, in order to advance the understanding of complex natural mixtures, there is a need for accurate and efficient techniques to extract as much information as possible from simple 2D NMR datasets, such as COSY, TOCSY and HMQC, which are easy to acquire on soil extracts. Matching of 2D NMR datasets offers numerous advantages over analogous 1D peak matching. 1D peak searches are difficult for complex mixtures that exhibit extensive spectral overall in a single dimension; furthermore, 2D peak searches offers additional information, such as heteronuclear connectivities and long-range homonuclear couplings, which allow assignments of unknown fragments with a greater degree of confidence. Although there are numerous 1D FT-NMR reference databases available,[17–19] the generation of 2D NMR databases is more difficult and time consuming. Numerous software packages have been developed that can perform pattern recognition in 2D,[20–22] but the lack of comprehensive, commercially available 2D NMR databases at present makes this difficult for heterogeneous mixtures, which contain a range of uncharacterized structures.

This paper describes procedures for the synthesis of an accurate and searchable 2D NMR database from drawn

*Correspondence to: Andre J. Simpson, Department of Physical and Environmental Sciences, University of Toronto, Scarborough Campus, Toronto, Ontario M1C 1A4, Canada.
E-mail: andre.simpson@utoronto.ca

Contract/grant sponsor: Canadian Foundation for Climate and Atmospheric Sciences (CFCAS); Contract/grant number: GR-327.

structures. In this paper we use lignin as an example. Lignin is an important structural biopolymer that occurs in the majority of plant biomass.[23] It is known to be relatively recalcitrant in the environment and hence is likely to accumulate in soils.[24–28] The structure of lignin is known to be highly heterogeneous and hundreds of fragments have been identified as constituents in its overall structure.[29] Lignin contains 1,4-, 1,3,4- and 1,3,4,5-substituted rings that are connected by various linking groups often containing oxygen substituents. Softwoods, such as pine, contain very few 1,3,4,5-substituted aromatic rings (syringyl), whereas in hardwoods, such as oak, the cross-linking of these units provides greater structural rigidity. We demonstrate how an accurate 2D NMR database of ∼600 2D spectra of lignin fragments can be easily and quickly constructed, and then show how some of these fragments can be identified in soil extracts through the use of various search and pattern recognition tools. Once identified in one sample, similar residues are easily identifiable in other soil extracts.

EXPERIMENTAL

Sample preparation

Pine forest fulvic acid (PFFA) was isolated from the A_h (surface) horizon of a pine forest site in the Harvard Forest at Petersham, MA, USA. The PFFA was obtained by exhaustively extracting soil residue with NaOH at pH 12.6. The residue was that which remained after previous sequential, exhaustive extractions using first 0.1 M sodium pyrophosphate (adjusted to pH 7.0 with HCl) and then 0.1 M sodium pyrophosphate at pH 10.6. The humic acids and FAs were separated by precipitation with HCl, and FAs were then isolated using XAD-8 and XAD-4 resins in tandem as described elsewhere.[30] The PFFA studied here represents ∼5% by weight of the total extractable organic material. Oak forest fulvic acid (OFFA) was isolated from the A_h horizon of a soil located in Uragh Wood at Lough Inchiquin, Kenmare, County Kerry, Ireland, using the same procedure. The OFFA studied here represents ∼5% by weight of the total extractable organic material. Total oak forest soil extract (TOFSE) was isolated from the same Irish soil. The soil was exhaustively extracted with 0.1 M NaOH, the extract was filtered through a 0.22 μm filter under nitrogen, passed repeatedly over Amberlite IR-1200H+ cation-exchange resin to help remove paramagnetic species and then freeze-dried. The TOFSE studied here represents the total alkaline extractable organic material.

Prior to NMR spectroscopy, previously freeze-dried samples were further dried over P_2O_5 for 24 h at 40 °C to remove excess moisture. Samples (50 mg) were dissolved in DMSO-d_6 (1 ml) and transferred to 5 mm NMR tubes.

NMR spectroscopy

Data for the TOFSE and PFFA samples were acquired on a Bruker Avance 400 MHz spectrometer fitted with a QNP ^{1}H, ^{13}C, ^{15}N, ^{31}P probe. Data for the OFFA sample were acquired on a Bruker DRX 500 instrument fitted with a ^{1}H–^{13}C–^{15}N TXI probe. In each case, correlation spectroscopy (COSY) spectra [64 scans, TD (F_1) 1024, TD (F_2) 512] was acquired using a 45 °read pulse. Processing was carried out using an unshifted sine-squared function in both dimensions and a zero-filling factor of two. Spectra were projected using a magnitude calculation. Heteronuclear multiple quantum coherence (HMQC) spectra [128 scans, TD (F_1) 1024, TD (F_2) 512, ^1J(^{1}H, ^{13}C) 145 Hz] were acquired using a BIRD pulse train and TPPI.[31] F_1 was processed with a sine-squared function with a phase shift of 90 °and F_2 was processed with a Gaussian broadening of 0.01 and line broadening of −1. Standard ^{13}C NMR data were acquired using inverse gated decoupling, 60 000 scans and a 5 s recycle delay. The resulting spectrum was processed with an exponential multiplication with 60 Hz line broadening. DEPT90 data were acquired with a recycle delay of 2 s, 60 000 scans and assuming a ^{1}H–^{13}C one bond coupling of 145 Hz. Processing was performed using an exponential multiplication with 60 Hz line broadening. The projection from the HMQC experiment was calculated using an exponential multiplication with a 60 Hz line broadening in F_1.

Predictions and database generation

Structures were entered into ACD/ChemSketch, the structure-drawing interface for the NMR prediction packages utilized in this work.[19] The ^{1}H and ^{13}C NMR predictor packages that were used for this work were versions 6.12 and 6.13, respectively. 2D NMR predictions were carried out using ACD/Labs 2D NMR Predictor software which is integrated with the 2D NMR Processor software (Version 6.18). Simulations were carried out using a ^{1}H base frequency of 500 MHz, and a ^{13}C base frequency of 125 MHz. COSY spectra were predicted without diagonal cross peaks and a ^{1}H linewidth of 3 Hz. HMQC (^{1}H–^{13}C COSY) data were predicted using a ^{1}H linewidth of 3 Hz and a ^{13}C linewidth of 5 Hz. Only one-bond ^{1}H–^{13}C couplings were predicted in this study. More details are given in the Results and Discussion section.

Peak searches and pattern matching

Peak searches were carried out using the peak search tool in ACD/2D NMR Manager. Compounds that displayed similar ^{1}H–^{1}H and ^{1}H–^{13}C couplings to those in the NOM data were visually compared with the NOM NMR data. The NMR spectra for matched compounds were opened in ACD/2D NMR Processor from the database, and an overlay plot was created. Compounds showing poor correlation were manually removed from the database. More details are given in the Results and Discussion section.

RESULTS AND DISCUSSION

Database generation

An accurate and comprehensive 2D NMR database would be extremely useful in a number of disciplines but especially in environmental chemistry where the analysis of very complex mixtures is common. A comprehensive 2D NMR database in combination with pattern recognition tools would allow spectroscopists to extract structural information directly from 2D datasets of very complex and intact mixtures, in a semi-automated or even fully automated fashion. In turn, this would lead to a better understanding of the structural components present in the environment, and allow chemists to measure and predict their individual chemical properties and reactivities. This information is vital in order to understand and predict environmental processes at the larger scales.

The creation of a conventional 2D NMR database to represent the range of structures in soils and water is, however, extremely challenging. Extracts from soils and sediments will likely contain thousands or even hundreds of thousands of chemical species, many of which may have complex heterogeneous structures. The separation of these structures into 'pure' fractions has hindered scientists for decades. Even if the samples could be separated or representative analogues synthesized, acquiring the 2D data sets needed to construct a representative database may take decades. That said, it is feasible to create 2D NMR spectra through predictions. In this work, we constructed a relatively small (~600 spectra) database for fragments found in the biopolymer lignin, and then searched for these fragments in the COSY and HMQC data on selected soil extracts. The ACD/Labs suite of software was employed to generate the database and for the purpose of pattern recognition.

ACD/Labs prediction software employs a large internal 1D database containing over 3.2 million chemical shifts for over 300 000 structures when the contents of the ^{1}H and ^{13}C databases are combined. These data have been used to develop prediction algorithms and are the foundation of the prediction capabilities. If the exact structure being predicted is already present in the internal database, the measured chemical shifts from the database are used directly by the prediction algorithms. In all cases, the nearly 300 lignin structures considered in this study were already present in the internal 1D NMR database along with their experimentally measured chemical shifts. In cases where the structures are present in the database, the cross peak positions in the COSY and HMQC spectra are highly representative, and these predicted 2D spectra were directly updated to a user-defined database using the ACD/2D NMR Manager package, which allows spectral databasing. The finished database contained ~300 compounds. For each structure, the ^{1}H, ^{13}C, HMQC and COSY spectra were stored alongside the chemical structure together with various structural properties including the IUPAC name, accurate mass and chemical formula.

Peak searches

COSY and HMQC spectra for the PFFA sample were selected for detailed analysis. Initially the processed spectra were imported into ACD/2D NMR Manager, manual peak peaking was performed and the peak information exported to a text file. Figure 1 shows the aromatic region of the PFFA COSY spectrum with the 20 major sets of cross peaks highlighted. Table 1 lists the chemical shifts of the cross peaks. Using the peak search tool in the ACD/Labs software, peaks were systematically matched using the following scheme:

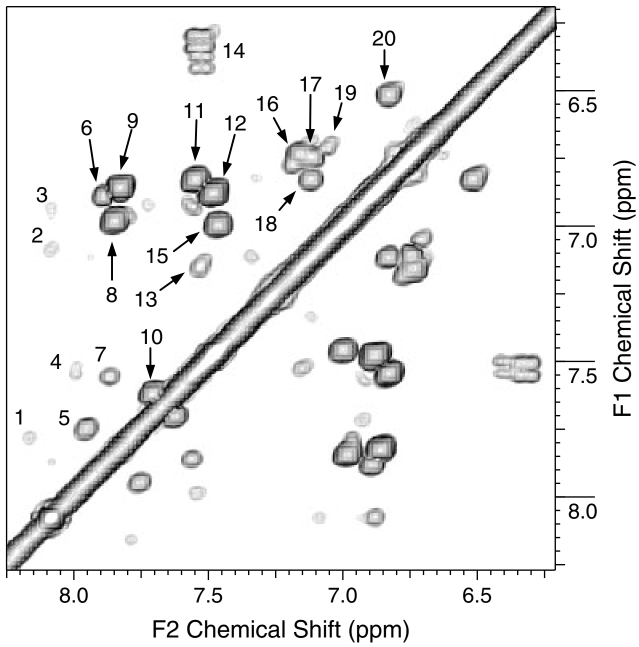
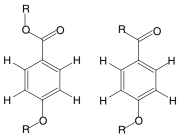


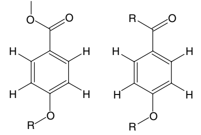
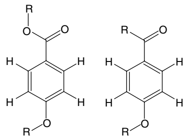
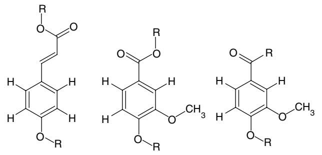
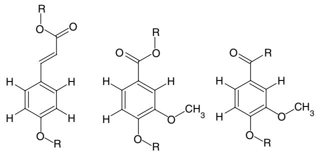
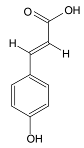
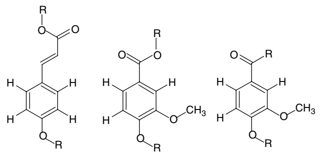
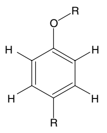
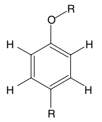
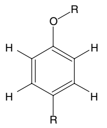
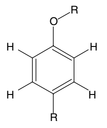
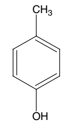
Figure 1. Expanded region of the COSY spectrum for the PFFA sample, with the major crosspeaks highlighted (corresponding chemical shifts are given in Table 1).

Table 1. Assignments of lignin residues in the PFFA sample: column 1 correlates with cross peaks observed in the COSY spectra (see Fig. 1) and columns 3–6 correlate with the number of compounds returned from the specific database searches

No.	F_2 (ppm)	F_1 (ppm)	TOCSY matches	TOCSY/HSQC matches	Visual pattern matching (TOCSY)	Visual pattern matching (HSQC)	Common structure unit or units[a]
1	8.16	7.79	0	0	0	0	Unlikely to be lignin derived
2	8.1	7.09	0	0	0	0	Unlikely to be lignin derived
3	8.08	6.94	0	0	0	0	Unlikely to be lignin derived
4	7.99	7.54	1	1	0	0	Unlikely to be lignin derived
5	7.95	7.75	0	0	0	0	Unlikely to be lignin derived
6	7.89	6.89	9	8	8	5	
7	7.86	7.56	1	1	0	0	Unlikely to be lignin derived

(continued overleaf)

Table 1. (*Continued*)

No.	F_2 (ppm)	F_1 (ppm)	TOCSY matches	TOCSY/HSQC matches	Visual pattern matching (TOCSY)	Visual pattern matching (HSQC)	Common structure unit or units[a]
8	7.85	6.98	11	11	11	8	
9	7.83	6.86	19	17	15	6	
10	7.72	7.62	0	0	0	0	Unlikely to be lignin derived
11	7.54	6.83	29	18	14	7	
12	7.48	6.88	29	20	11	7	
13	7.52	7.14	7	5	0	0	Undetermined structure
14	7.53	6.35	36	25	11	8	
15	7.47	7.0	29	25	14	8	
16	7.17	6.74	51	27	8	6	
17	7.11	6.75	49	26	8	5	
18	7.12	6.83	51	30	10	6	
19	7.05	6.70	63	35	14	8	
20	6.82	6.51	25	17	0	0	 Loose fit, assignment cannot be made with confidence

[a] R = H, methoxy or an aliphatic carbon in a lignin linking group.

1. Data for an individual COSY cross peak were input into the 'peak search' tool from the stored text file.
2. A search was performed (with the ^{1}H tolerance set at 0.2 ppm) and hits were returned and saved as a new 'COSY' database. This new database only contained compounds that exhibit a COSY cross peak similar to that in the NOM NMR data.
3. Data for corresponding HMQC cross peaks or regions were input into the 'peak search' tool.
4. A search was performed on the 'COSY' database, created at the end of step 2 (with the tolerances set to 0.2 and 3 ppm for proton and carbon, respectively), and hits were returned and saved as a new database. This new database contains only compounds that exhibit a COSY cross peak similar to that seen in the NOM NMR data and also display consistent HMQC couplings.
5. These steps were repeated for all cross peaks in Fig. 1.

At the end of the above procedure, 20 databases were produced, one for each cross peak in the COSY spectra (Fig. 1). Each database contains only those structures that display at least one cross peak that is similar in terms of both COSY and HMQC couplings to that in the PFFA sample. Columns 4 and 5 in Table 1 indicate the total number of compounds that match the COSY criteria (column 4) and combined COSY/HMQC criteria (column 5). Although the compounds in the databases display partial similarity to the cross peaks in the PFFA sample, this does not necessarily mean that the compounds are present in the sample. However, this peak searching procedure does considerably narrow the search to a small handful of compounds with which a more detailed visual inspection can be carried out.

Pattern matching

An example of 'pattern matching' is shown in Fig. 2. The spectrum is imported from the database of matches for peak 14 (Fig. 1; from this point on this will be referred to as 'database 14') into ACD/2D NMR Processor and an overlay plot was created (see Fig. 2). The pattern matching is carried out by a visual inspection of the overlay plot, the fragment from the database is considered a hit if all the cross peaks match those in the sample (within ~0.1 ppm ^{1}H, ~3 ppm ^{13}C) and no extraneous peaks are observed from the database fragment. It is clear from Fig. 2 that the red 1,4-substituted cinnamic acid fragments of the molecule fit well with the data, whereas the remaining black fragments show a poor correlation. A number of peaks do not correlate well with the PFFA data and, as a result, the molecule as a whole must be considered a mismatch and is removed from database 14. However, the close match of the red fragments observed in Fig. 2 does still provide us with solid information. Region A (peak 14, Fig. 1), which results from protons on the double bond, and region B, resulting from protons on the aromatic ring of a 1,4-substituted structure, strongly correlate with 1,4-cinnamic acid-type molecules, and suggest that these moieties may be present in the sample. Indeed, this is supported when the remaining structures in the database (database for peak 14 from Fig. 1) are overlayed. All 24 structures in database 14 (note that all these structures show a peak consistent with the H−H and H−C couplings for peak 14 in the PFFA COSY and HMQC data; see Table 1) contain cinnamic moieties and in each case it is the protons on the double bond (see Fig. 2) that gives rise to a peak similar to region A (peak 14, Fig. 1). Of these 24 cinnamic structures, only the 1,4-substituted rings contain aromatic ring protons with resonances that correlate well with peaks in the PFFA sample; these fall in region B of Fig. 2. As a result, after pattern matching the entire database 14 with the COSY data, only 11 cinnamic moieties remain that exhibit good correlation with the PFFA and that do not contain extraneous peaks. To confirm further that these 11 cinnamic moieties fit with H−C couplings observed in the PFFA, pattern matching was also carried out with the HMQC data. Matching with the HMQC data shows that eight of the molecular fragments show good matches with peaks in the PFFA data whereas three units show extraneous peaks. Of the eight fragments remaining in database 14, all are of the form of 1,4-oxygen-substituted cinnamic acids/esters (see Table 1). By combining a series of database searches with visual pattern matching, it is possible to assign peak 14 to the protons on the double bond of a 1,4-oxygen-substituted cinnamic acid/ester. The assignments not only are based on both the COSY and HMQC data for all protonated positions in the molecular fragment but are also confirmed by numerous molecular species from the database. The presence of 1,4-cinnamic acid structures in soil extracts is expected considering that the production of 4-hydroxycinnamic acid is the first oxidative step in the synthesis of lignin, flavonoids, coumarins and other phenylpropanoids,[32] and 1,4-cinnamic units are major constituents of the lignin biopolymer itself.

By repeating the COSY and HMQC pattern matching process for all peaks in Fig. 1, it is possible to offer solid interpretations for many resonances (see Table 1), and demonstrate that most of the intense resonances in the COSY spectra likely originate from lignin-type residues. In the case of the PFFA, the assignments offered here have been further checked against HMBC and HSQC−TOCSY data (not shown) to ensure the assignments are feasible and to help validate the methods employed here. However, it is important to point out that collecting HMBC and HSQC−TOCSY spectra is very time consuming for soil extracts owing to the insensitivity of the experiments and relaxation during the longer delays employed. Although HMBC or HSQC−TOCSY data can be collected, in many cases such datasets may take a week or more to acquire such that the majority of couplings in the sample are detected. In comparison, COSY and HMQC data can both be collected in ~6 h and are more representative of the sample as a whole (see Fig. 3). This paper therefore focuses on the extraction of information from these simple datasets that can be applied more 'routinely' to soil extracts. In many cases the 'luxury' of long-range couplings is not available, hence there is a need for methods that extract as much information as possible from the simpler and more easily acquired datasets.

Testing the assignments

2D NMR spectral prediction itself provides an excellent tool with which to test the assignments offered in Table 1. If these assignments are reliable, it should be possible to replace the

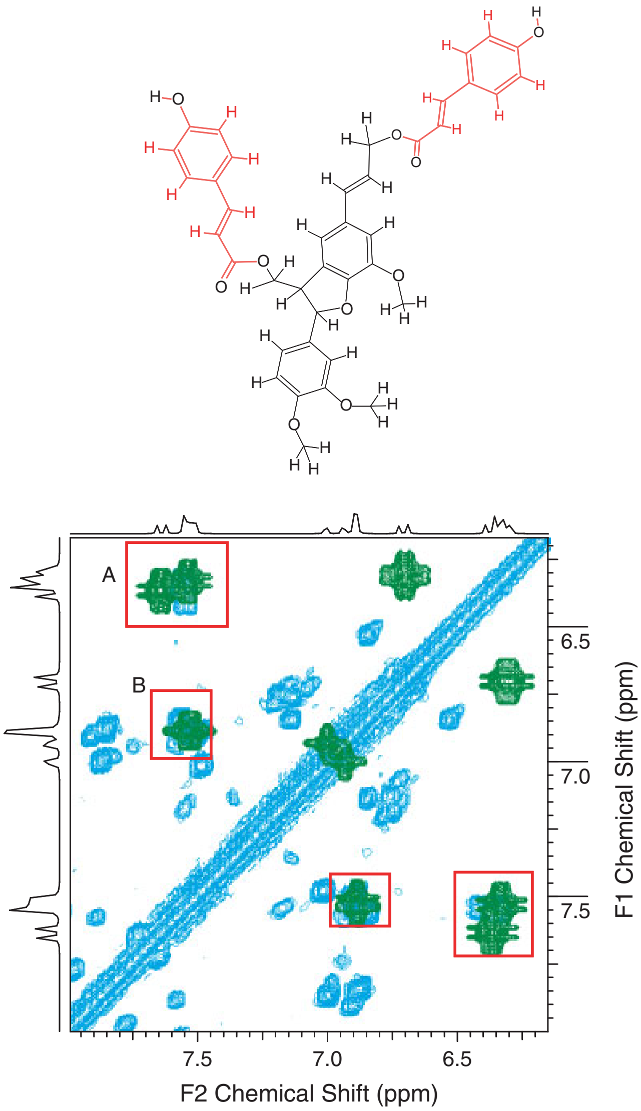


Figure 2. An example of 'pattern matching'. Blue COSY cross peaks are from the PFFA sample and green cross peaks are from the chemical structure that has been imported from the database. In this case the structure shown was imported from the database that contains matches with cross peak 14 in the PFFA COSY data (see Fig. 1). It can be seen that cross peak 14 (Fig. 1) clearly matches well with a cross peak in the molecule (region A in this figure), as do cross peaks in region B (this figure), which correspond to an area containing cross peaks 11, 12 and 15 (see Fig. 1). However, other parts of the molecule (region of structure in black) show poor correlations. While the entire molecule does not fit with the PFFA data, it is clear that the fragments in red, cinnamic acid moieties, show similar couplings and may be present in the PFFA sample.

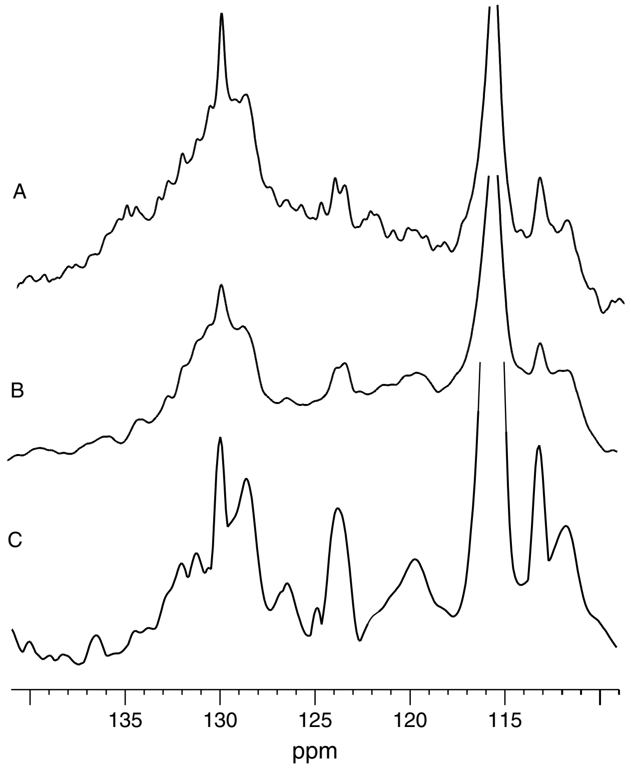


Figure 3. Comparison of the aromatic region of the PFFA sample. (A) Standard ^{13}C spectrum (all carbons); (B) DEPT-90 spectrum (protonated carbons only) (C) projection from the HMQC experiment (protonated carbons only). The conventional ^{13}C spectrum is presented only as a reference. Both the DEPT-90 and HMQC projections show only protonated carbons. An exact match between the DEPT and HMQC spectra is not expected owing to the different nature of the experiments and differences in acquisition and processing parameters (number of scans, time domain points, spectral resolution, differential effect of the same window function on data sets with different number of time domains points, etc.). However, the general similarity of the spectra is encouraging; the HMQC projection clearly shows that most (if not all) protonated aromatic carbon is represented in the HMQC experiment.

R group with various substituents and observe a good match with cross peaks in the PFFA data. In an attempt to test this theory, a synthetic mixture of ~50 different structures was created using structures of the type in Table 1. The R substituent was randomly chosen, and predicted COSY and HMQC spectra were created for the entire mixture. Figure 4 shows the overlay plots for the mixture of compounds and the PFFA data. In general, the correlation between the peaks in the PFFA data and those in the synthetic mixture is very good. A perfect correlation cannot be expected as the nature of the substituent R will slightly influence the exact chemical shift in the fragments. Additionally, small errors may be introduced for molecules that are not in the internal database for which predicted chemical shifts were used to create the 2D spectra. A good correlation for both the COSY and HMQC data clearly shows that all assignments shown in Fig. 1 are consistent with PFFA data. However, from just the COSY and HMQC data alone it is not possible to determine the exact nature of the R groups in the PFFA. This will be discussed in further publications, along with evidence from a range of additional NMR experiments.

Lignin fragments in soil organic matter

It is clear from assignments that many of the most intense cross peaks in the COSY spectra are consistent with residues commonly found in lignin. However, it is important to

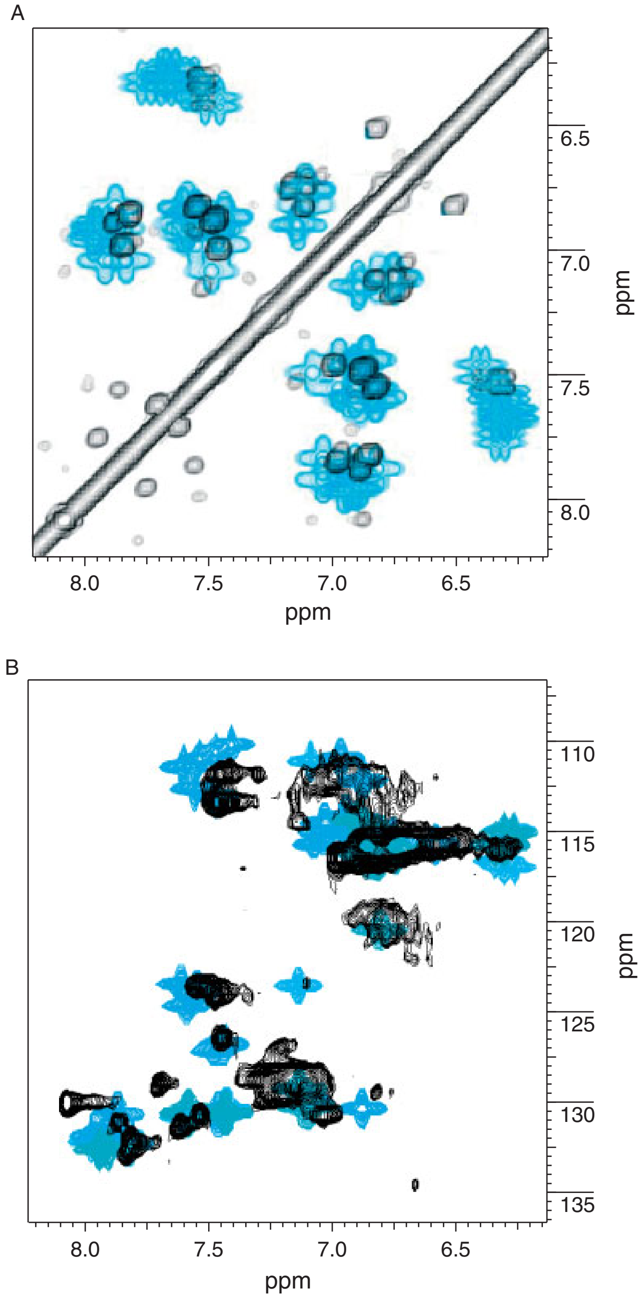


Figure 4. Overlay plot comparing the cross peaks from the PFFA sample with those for a mixture of ~50 compounds, of the type shown in Table 1. Black peaks represent those from the PFFA sample and blue peaks are from the 'synthetic' mixture. (A) COSY data; (B) HMQC data.

remember that the cross peaks in the COSY spectra may not represent all compounds present in soil extracts. For example, syringyl (1,3,4,5-substituted structures) in lignin will not give rise to a COSY cross peak as the aromatic protons at positions 2 and 5 are not strongly coupled. Thus, if the approach of searching the COSY spectra and then the HMQC spectra is used, syringyl units, which do not display cross peaks in the COSY but do display cross peaks in the HMQC spectra, will be overlooked. However, if such units are expected, then specific searches can be carried out directly on the HMQC data. Searches using the constructed database indicate that the protonated aromatic

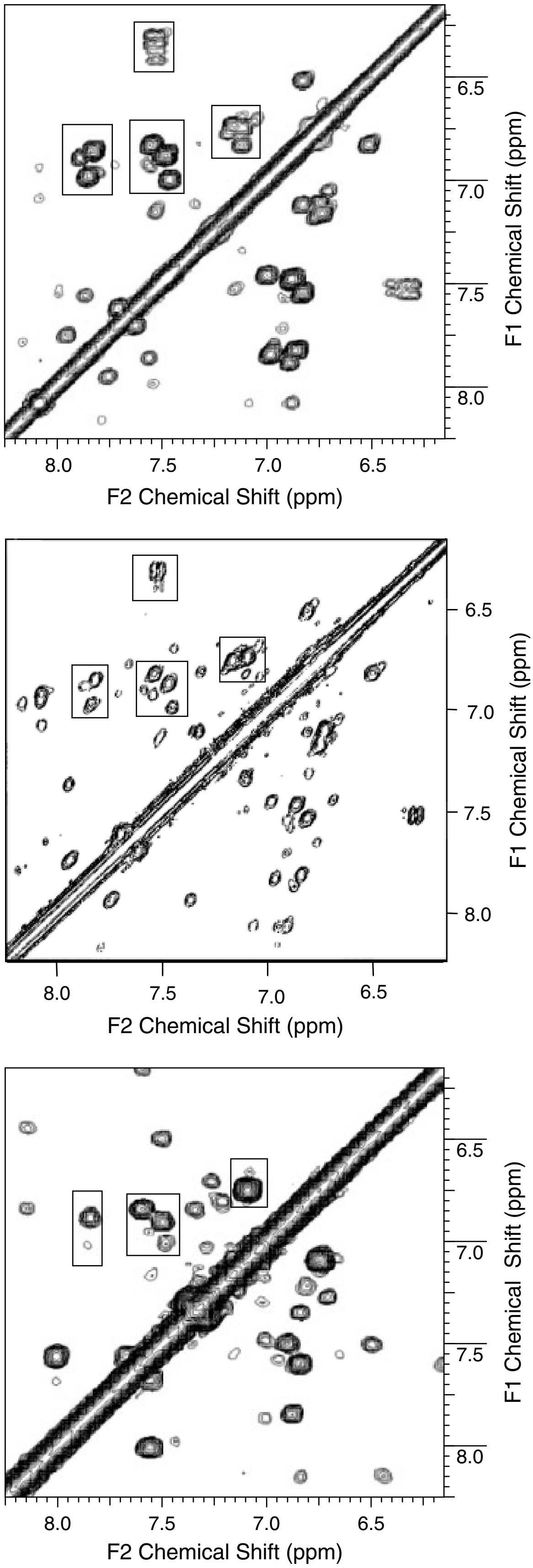


Figure 5. Comparison of the aromatic regions from the COSY spectra of the (top) PFFA, (middle) OFFA and (bottom) TOFSE samples. Resonances likely to be lignin derived are highlighted in each case.

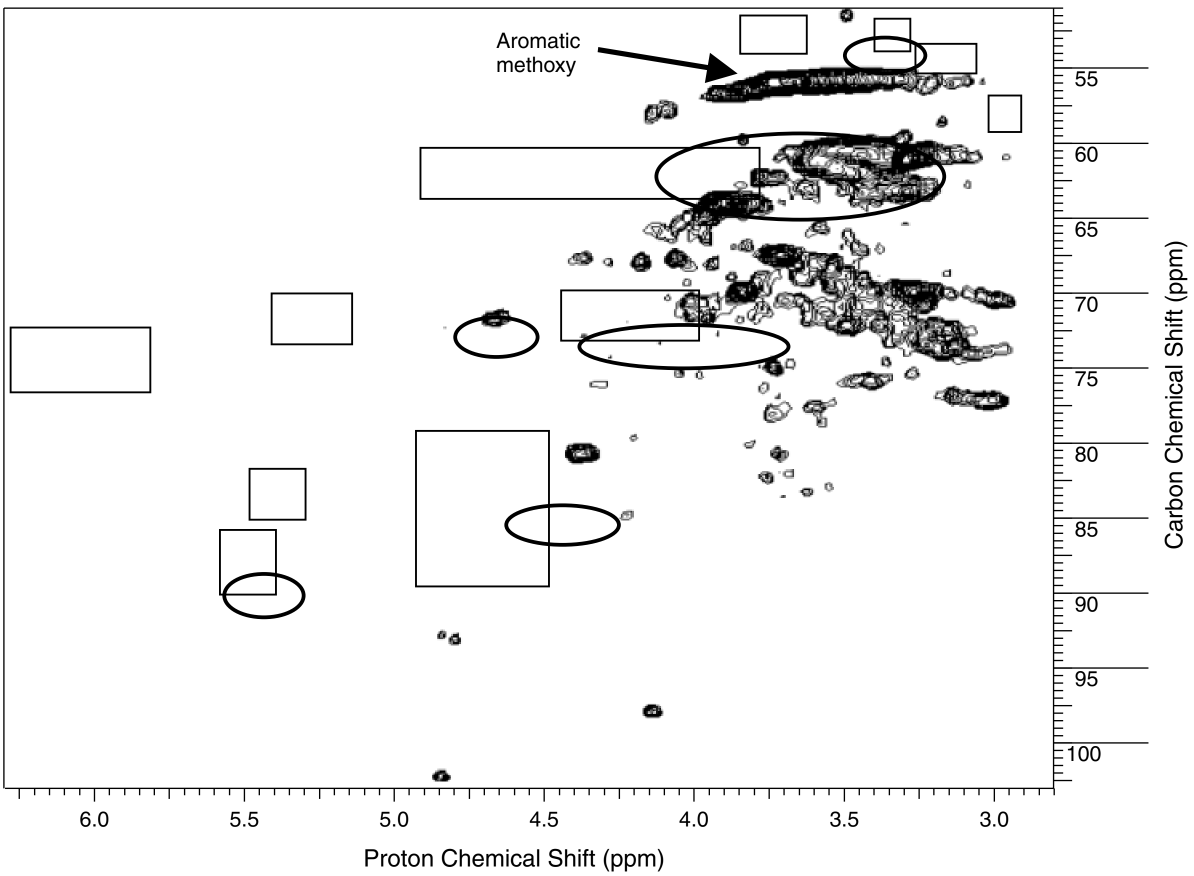


Figure 6. An expanded region of the PFFA HMQC spectrum. The ovals highlight the positions of lignin side-chains in underivatized lignin in DMSO-d_6[35] and the rectangles highlight the side-chains in acetylated lignin.[36]

carbons in syringyl units commonly found in lignin display chemical shifts in the region of 105–108 ppm (carbon) and 7.1–7.4 ppm (proton) in the HMQC spectra. This region in the PFFA shows no signal as expected, considering that the native vegetation on the site is pine, a softwood which by nature has a very low syringyl content. However, extracts from soils under hardwood forests do display signals in this region (data not shown), indicating the presence of syringyl units as expected.

Once lignin type residues have been identified in the soil extracts, similar resonances can be highlighted in different samples. Figure 5 compares the aromatic region of the COSY spectra for the PFFA, OFFA and TOFSE samples. Both the PFFA and OFFA have been extracted and fractionated by identical methods. Even though these samples were extracted from soils on either side of the Atlantic, the signatures in the aromatic region of the COSY spectra are very similar. Residues from lignin are clearly present in the oak forest sample, and additionally signals consistent with syringyl units are also present in the HMQC of this sample (data not shown). The TOFSE sample represents a total extract from the same soil. Signals consistent with the lignin residues are still apparent. However, the intensity of the peaks is very different for those observed in the fractionated FA, indicating that the various ratios of lignin type structures is different in this bulk extract. The exact nature of these differences cannot be explained from the data shown here and will be considered in further publications.

The presence of aromatic lignin-derived species in the soil extracts is clear; however, the presence of lignin side chains is not so clear. Searches to match common lignin side-chains in the database to resonances in the PFFA sample were unsuccessful. Closer inspection shows that signals are missing in the soil spectra where many lignin side-chains are expected (see Fig. 6). The reason for this is likely to be threefold. First, intact lignin is relatively hydrophobic and only partially soluble in the aqueous solvents commonly used for standard soil extracts. High molecular weight unaltered lignin fragments will likely be present in the so called 'humin' fraction, that is, the fraction not solubilized by alkaline solvents. Second, the fragments that are soluble in the alkaline extracting solvent are likely to be the smaller and more polar fragments that result from lignin degradation. Many of the lignin-derived structures in the PFFA (see Table 1) may well be present in the form of acids, explaining their solubility. In fact, the cleavage of the α–β bond in lignin side-chains to form carboxylic acids is one of the most important steps in lignin degradation.[33,34] This will lead to the release of a high proportion of alkaline soluble lignin aromatics without side-chains. Third, a large variety of side-chains are known to be present in the lignin structure. As a result, the NMR signals from each type of side-chain will be very weak in comparison with signals from the more 'repetitive' aromatic and methoxy species that are consistent throughout the lignin biopolymer. As a result, even if small fragments of lignin containing intact side-chains are present in the extract, it is feasible that the signals from the heterogeneous side-chains will be below the detection limit of the experiment. Only the more intense signals from the methoxy and the aromatic groups may be detectable. In reality, the explanation for the absence of lignin side-chains in the NMR data is likely to be a combination of all of the above.

Further considerations

In principle, the approach of combining database searches and visual pattern matching can be applied to any system. If the 1D NMR chemical shifts are available first hand in the literature or in the ACD/Labs internal databases, very accurate COSY and HMQC spectra can be simulated. It is also feasible to predict HMBC, TOCSY and INADEQUATE experiments. However, predicting 2D spectra containing long-range couplings are more difficult. The intensities of these long-range couplings are often weak in measured spectra and can be strongly influenced by molecular conformation, concentration and relaxation. Therefore, predicted spectra may not exactly match measured spectra in all cases, as the measured spectra may vary under different experimental conditions and for different components in the mixtures.

Another important consideration is the effect of the solvent system on the predicted spectra. In this study, the soil extracts were run in DMSO-d_6. This, in the majority of cases, matches the solvent used to measure the 1D chemical shifts, which in turn were a basis for the 2D predictions.[17,19,29] Note that full reference to the origin of the chemical shift information is given for each compound contained in the ACD/Labs internal database. However, for other components in soils, such as sugars, peptides and tannins, compounds are often run in water and, for less polar compounds such as aliphatic chains, chloroform is often used (see References within ACD/Labs' HNMR DB and CNMR DB Version 7 for complete details). Thus, for accurate comparisons, it is important to solubilize the analyte (in this case the soil extract) into the same solvent as was used in the original study or to create user databases using data generated within the investigator's laboratory which can then be used by both the 1D and 2D prediction algorithms. For studies of sugars and peptides in soil extracts the use of D_2O as a solvent would make sense whereas chloroform would be a solvent of choice for the less polar aliphatic chains. Using such an approach will help reduce differences from solvent effects and thus minimize matching errors. If such an approach is employed, the combination of spectral prediction, database searching and pattern matching provides a powerful yet simplistic approach for the extraction of structural information from complex 2D datasets.

Acknowledgement

We thank the Canadian Foundation for Climate and Atmospheric Sciences (CFCAS), grant number GR-327, for providing funding for this project.

REFERENCES

1. Hayes MHB, McCarthy P, Malcolm RL, Swift RS. *Humic Substances II: in Search of Structure*. Wiley: New York, 1989.
2. Schmitt-Kopplin P, Hertkorn N, Schulten H-R, Kettrup A. *Environ. Sci. Technol.* 1998; **32**: 2531.
3. Wang L, Mao X, Yang Y. *Bopuxue Zazhi* 1998; **15**: 411.
4. Haiber S, Burba P, Herzog H, Lambert J. *Fresenius' J. Anal. Chem.* 1999; **364**: 215.
5. Morris KF, Cutak BJ, Dixon AM, Larive CK. *J. Anal. Chem.* 1999; **71**: 5315.
6. Fan TWM, Higashi RM, Lane AN. *Environ. Sci. Technol.* 2000; **34**: 1636.
7. Kingery WL, Simpson AJ, Hayes MHB, Locke MA, Hicks RP. *Soil Sci.* 2000; **165**: 483.
8. Haiber S, Herzog H, Burba P, Gosciniak B, Lambert J. *Environ. Sci. Technol.* 2001; **35**: 4289.
9. Haiber S, Herzog H, Burba P, Gosciniak B, Lambert J. *Fresenius' J. Anal. Chem.* 2001; **369**: 457.
10. Simpson AJ, Burdon J, Graham CL, Hayes MHB, Spencer N, Kingery WL. *Eur. J. Soil Sci.* 2001; **52**: 495.
11. Simpson AJ, Kingery WL, Shaw DR, Spraul M, Humpfer E, Dvortsak P. *Environ. Sci. Technol.* 2001; **35**: 3321.
12. Simpson AJ, Kingery WL, Spraul M, Humpfer E, Dvortsak P, Kerssebaum R. *Environ. Sci. Technol.* 2001; **35**: 4421.
13. Hertkorn N, Permin A, Perminova I, Kovalevskii D, Yudov M, Petrosyan V, Kettrup A. *J. Environ. Qual.* 2002; **31**: 375.
14. Simpson AJ. *Magn. Reson. Chem.* 2002; **40**: S72.
15. Simpson AJ, Salloum MJ, Kingery WL, Hatcher PG. *J. Environ. Qual.* 2002; **31**: 388.
16. Simpson AJ, Kingery WL, Hatcher PG. *Environ. Sci. Technol.* 2003; **37**: 337.
17. Pouchert CJ, Behnke J. *The Aldrich Library of ^{13}C and ^{1}H FT-NMR Spectral*. Aldrich Chemical: Milwaukee, WI, 1992; also available in electronic format from ACD/Labs, http://www.acdlabs.com/products/spec_lab/exp_spectra/spec_libraries/aldrich.html.
18. Hayamizu K, Yanagisawa M, Yamamoto O, Wasada N, Someno K, Tanabe K, Tamura T, Tanabe K, Hiraishi J. Integrated Spectral Data Base System for Organic Compounds, 2001; http://www.aist.go.jp/RIODB/SDBS/.
19. ACD/Labs ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P NMR prediction and databases, 2003; http://www.acdlabs.com/products/spec_lab/predict_nmr/.
20. Accelrys NMR Refine DGII, 2003; http://www.accelrys.com.
21. ACD/2D NMR Manager and Predictor, 2003; http://www.acdlabs.com/products/spec_lab/exp_spectra/2d_nmr/; http://www.acdlabs.com/products/spec_lab/predict_nmr/2d_nmr/.
22. Bruker Amix and Amix Tools, 2003; http://www.bruker-biospin.de.
23. Wardrop AB. *Lignins* 1971; 19.
24. Fustec E, Chauvet E, Gas G. *Appl. Environ. Microbiol.* 1989; **55**: 922.
25. Amalfitano C, Pignalosa V, Auriemma L, Ramunni A. *Soil Sci.* 1992; **43**: 495.
26. Del Valle HF, Rosell RA. *Arid Soil Res. Rehabil.* 1999; **13**: 239.
27. Olk DC, Dancel MC, Moscoso E, Jimenez RR, Dayrit FM. *Soil Sci.* 2002; **167**: 590.
28. Kiem R, Kogel-Knabner I. *Soil Biol. Biochem.* 2003; **35**: 101.
29. Ralph S, Ralph J, Landucci L. NMR Database of Lignin and Cell Wall Model Compounds, 1996; http://www.dfrc.wisc.edu/software.html.
30. Simpson AJ, Watt BE, Graham CL, Hayes MHB. *Spec. Publ. R. Soc. Chem.* 1997; **172**: 73.
31. Bax A, Subramanian S. *J. Magn. Reson.* 1986; **67**: 565.
32. Schalk M, Batard Y, Seyer A, Nedelkina S, Durst F, Werck-Reichhart D. *Biochemistry* 1997; **36**: 15 253.
33. Chang H-M, Chen C-L, Kirk TK. *Lignin Biodegradation: Microbiol. Chem. Potential Appl. [Proc. Int. Semin.]* 1980; **1**: 215.
34. Chen CL, Chang HM, Kirk TK. *Holzforschung* 1982; **36**: 3.
35. Heikkinen S, Toikka MM, Karhunen PT, Kilpelaeinen IA. *J. Am. Chem. Soc.* 2003; **125**: 4362.
36. Kilpelainen I, Ammalahti E, Brunow G, Robert D. *Tetrahedron Lett.* 1994; **35**: 9267.