*Tetrahedron* Vol. 47, No. 22, pp. 3655-3664, 1991
Printed in Great Britain

0040-4020/91 $3.00+.00
Pergamon Press plc

# LOGIC FOR STRUCTURE DETERMINATION

Jean-Marc Nuzillard* and Georges Massiot

Laboratoire de Pharmacognosie, URA 492 au CNRS, Faculté de Pharmacie, Université de Reims, 51 rue Cognacq-Jay, 51096 Reims, France

(*Received in USA* 28 *November* 1990)

Abstract : A program for structure elucidation of organic molecules is described. The main source of input data relies on NMR carbon-proton correlation spectra. The methodology is illustrated for two indole alkaloids.

The automation of organic structure determination[1] is a current topic for which attention will certainly grow in forthcoming years. The well-known DENDRAL[2] project was initially intended to analyse mass spectral data replacing the blind screening of a huge spectral data bank by the intelligent use of a fragmentation rule set. The application of artificial intelligence concepts allowed a kind of "biomimetic" reasoning path. A program for structure elucidation is a chemical structure generator constrained by spectral data. The quality of the fit between the data and that predicted from the produced structures guides the production mechanism.
The choice of a particular spectroscopic method is governed by the ability to predict a relationship between spectra and structures. $^{13}C$ NMR spectroscopy quickly became the method of choice because the chemical shifts are reasonably calculated by series of additivity rules in various structural situations. The links established by this means are nevertheless far from being unequivocal, and moreover the additivity of substituent effects is only a poor approximation. The situation is even worse for $^{1}H$ NMR due to the great influence of the molecular environment on chemical shifts. Despite these drawbacks the programs coming from the DENDRAL project combining MS, NMR and substructural information supplied by the user have performed impressively.

Before the 80's, proximity relationships between nuclei were available from $^{1}H$ and $^{13}C$ NMR spectra with selective irradiation of $^{1}H$ multiplets, with nOe measurements giving indications about stereochemistry. This information directly provides substructural data without any reference to chemical shift knowledge, certainties then replacing assumptions. The tremendous technical step forward accomplished by the NMR techniques during the last decade brought chemists easy-to-record and easy-to-read proximity relationships through COrrelation SpectroscopY (COSY). Formally a set of selectively decoupled $^{1}H$ spectra is equivalent to a

$^1H$–$^1H$ COSY[3] spectrum but the latter is much more easily run and analyzed due to its graphical appearence. The automated use of two-dimensional (2D) spectra involves first the translation of the raw data into a set of peak coordinates; the corresponding "peak picking" algorithm is or will become a part of all modern spectrometers. The set of frequency coordinates produced could be used either for spectral assignment if the structure is known or else for structure determination. Automated assigments of $^1H$ resonances within peptides is a problem under current investigation[4]; combined with 2D nOe measurements and molecular mechanics this methodology is aimed at computing tertiary structures from samples in the liquid phase. The consideration of $^1H$–$^{13}C$ correlations gives more information about molecular structures than the corresponding $^1H$–$^1H$ ones, and substructural information about carbon atoms allows a straighforward tracing of the molecular skeleton. Their detection by the 1D or 2D INADEQUATE[5] spectra is however technically limited to highly concentrated (liquid !) samples due to the weak natural abundance and gyromagnetic ratio $\gamma$ of the $^{13}C$ nuclei; the required amount of substance is seldom available, especially if it is of natural origin. Automated structure generation from 2D INADEQUATE spectra has been studied[6], giving interesting solutions, even for highly symmetric molecules.

A technically practical and structurally informative approach relies on $^1H$ -$^{13}C$ COSY spectra for direct (HMQC[7] sequence) and long-range (HMBC[8] sequence) correlation detection. Obtained from proton spectra modulated by $^{13}C$ chemical shifts (the so-called inverse mode[9]) the theoretical sensitivity improvement of these spectra is $(\gamma_H/\gamma_C)^{3/2}$ (about 8), compared to INEPT-based methods. These techniques have been successfully applied to the analysis of complex biomolecules such as steroids, peptides and carbohydrates. A computer assisted spectral assignment program using HMBC and HMQC data will be reported[10]. A structure determination program named LSD (Logic for Structure Determination) relying on the same principles constitutes the subject of the present article.

The combination of HMQC and HMBC correlation sets yields a fictitious carbon-carbon correlation spectrum. An HMQC correlation, noted $^1J$ ($C_A$, H) and a $^nJ$ ($C_B$, H) HMBC correlation prove the existence of a n-1 bond path between the nuclei $C_A$ and $C_B$ (Figure 1), described as a correlation caused by a fictitious $^{n-1}J$ $^{13}C$–$^{13}C$ coupling constant. In a similar way a $^nJ$ ($H_A$, $H_B$) correlation from $^1H$–$^1H$ COSY, and $^1J$ ($C_A$, $H_A$) and $^1J$ ($C_B$, $H_B$) correlations from HMQC are equivalent to the fictitious $^{n-2}J$ ($C_A$, $C_B$) correlation (Figure 2). The evaluation of the coupling path length n is the major problem to solve in order to convert correlations into structures. The HMBC experiment can be setup in order to record exclusively $^2J$ (C, H) and $^3J$ (C, H) correlations, leading only to $^1J$ (C, C) and $^2J$ (C, C) fictitious correlations. A $^2J$ ($H_A$, $H_B$) gives no information and is easily detected by inspection of the HMQC spectrum, $H_A$ and $H_B$ being bound to the same carbon atom. An $^nJ$ ($H_A$, $H_B$) involving a large coupling constant (J > 5 Hz) is a proof for the existence of a bond between $C_A$ and $C_B$ ($^1J$ proximity). A well resolved small coupling constant is assignable either to a $^3J$ ($H_A$,

$H_B$) or to a $^4J$ ($H_A$, $H_B$) leading to $^1J$ and $^2J$ fictitious correlations as well. Very small $^nJ$ ($H_A$, $H_B$) leading to $^{n-2}J$ ($C_A$, $C_B$) with n > 2 are not taken into account. Thus HMBC, HMQC and $^1H$ - $^1H$ COSY spectra yield bonds and fictitious $^nJ$ ($C_A$, $C_B$), n remaining undetermined but equal to either 1 or to 2.

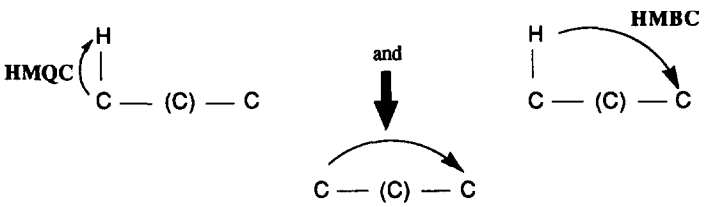



Figure 1. Fictitious carbon carbon correlations deduced from HMQC and HMBC spectra

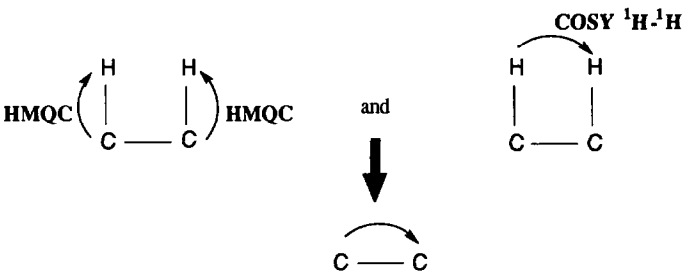


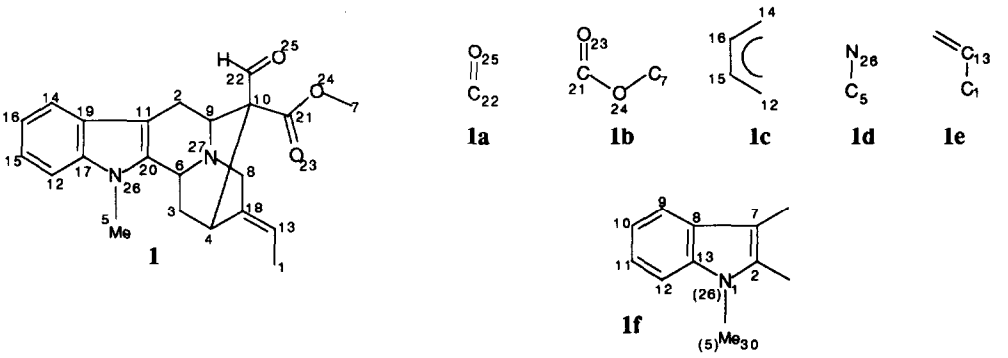

Figure 2. Carbon-carbon bond deduced from HMQC and COSY spectra.

The PROLOG[11] (PROgramming in LOGic) language has been retained for the translation of the structure elucidation algorithm. This language allows easy programming of complex logical operations. FORTRAN's equivalent would require (for us!) a painful writing task. PROLOG may be viewed as a data base management language equipped with an inference engine. The programmer's working area contains facts (unconditionally true assertions) and rules, describing how certainties can be deduced from basic facts. A problem is immediately solved if its solution is present as a fact, otherwise it is decomposed into sub-problems that have to be solved separately. All solutions of each query are systematically searched through the backtracking process. During the resolution free variables may be assigned a value, allowing a kind of "parameter-passing" between requests. These features are those of the so-called "first-order logic" programming. With PROLOG the programmer need not worry about data types or memory allocation. The LSD program is written in PROLOG (Xilog V.2, development version, supplied by ACT INFORMATIQUE 12, rue de la montagne S$^{te}$ Geneviève 75005 PARIS); it is run on a GOUPIL G5 computer (PC-compatible) working at 16 MHz with a 640 Kbytes memory.

The smaller the number of observed correlations, the larger the number of compatible structures! Supplementary information must therefore be introduced in the data base to input constraints to the structure generator. The molecular formula has to be known as well as a status description for each non-hydrogen atom. The status of an atom consists of its order number (arbitrary and given by the user), its hybridization state ($sp^3$ or $sp^2$, not sp presently), its valency and multiplicity (number of bound hydrogen atoms). For carbon atoms, this information is available from the $^{13}C$ and DEPT subspectra[12]; similar data have to be provided by the user for heteroatoms. The hybridization state of carbon atoms resonating around 100 ppm is not clear-cut; in this spectral area either regular $sp^2$ or acetalic $sp^3$ carbons signals are present. The various corresponding possibilities have to be proposed by the user. Special properties of atoms, deduced from elementary spectral analysis, may be indicated, mainly consisting of specifications about the status of the neighboring atoms. The program is able to handle groups of superimposed carbon signals if the status of each member of the group is known. Correlations between groups are treated as special properties of the atoms belonging to these groups. If fully assigned substructural units have been recognized, their bonds should be added to the data base as starting points for the resolution process. If, for any reason, an unassigned or partly assigned fragment is known to be present in the structure, it may be introduced as a collection of bonds between atoms having their own numbering system; the status of these atoms can only partly be described.

As an example, the data base which has been processed by the program to deduce the structure of voachalotinal **1** is described hereafter. Compound **1** is an alkaloid isolated from a neo-caledonian Apocynaceae plant *Alstonia undulata*[13]. Its molecular composition is $C_{22}H_{24}N_2O_3$ according to mass and $^{13}C$ NMR spectroscopy. Carbon atom signals are labelled in order of increasing resonance frequency from 1 to 22. Since signals 19 and 20 are so close that their correlations are not resolved, they are considered as a group named a. The UV spectrum of **1** suggests that an intact indole chromophore is present. The two nitrogen atoms are assumed to be $sp^3$. A sharp downfield proton singlet at 8.95 ppm and a $^{13}C$ signal at 194 ppm are due to an aldehyde function. This observation allows the construction of the assigned unit **1a**, where carbon 22 is bound to oxygen 25 (arbitrary oxygen numbering). The presence of a methyl ester group is deduced from $^1H$ data (s, 3H, 3.7 ppm) and $^{13}C$ data (170 ppm and 52 ppm), leading to fragment **1b**. The hypothesis of the presence of an intact indole nucleus is supported by the aromatic proton pattern. The corresponding COSY peaks reveal the order in which these protons are arranged. Fragment **1c** is then deduced from the HMQC spectrum. The indolic nitrogen atom does not bear any protons (generally more deshielded than the aromatic protons), so the methyl carbon absorbing at 29.4 ppm is likely bound to it, leading to fragment **1d**. A common feature of the $^1H$ spectra of many indole alkaloids which exists in compound **1** is a quartet proton signal near 5.35 ppm correlating in the COSY spectrum with a methyl doublet signal near 1.6 ppm. This lead to the proposal of fragment **1e** from HMQC data. The partly assigned substructure **1f** describes the N-methylated indole nucleus. Order numbers are chosen according to biogenetic considerations[14] (except for the methyl group). Special properties involve carbon atoms number 12 and 14 which are respectively bound to atoms 16 and 15 (Fig. 3c), and each to a quaternary $sp^2$ atom. During resolution, when these quaternary $sp^2$ carbons are

identified they must be bound together and if this is not the case, they will automatically be bound together. These properties may be viewed as a definition of a partly defined substructure ; their rank as special properties will force their verification each time that atom 12 or 14 is concerned by the bond formation process. All this information is entered in the data base as PROLOG facts. Their encoding is straightforward except for special properties where knowledge of PROLOG syntax is necessary. Data are entered via the built-in editor of our PROLOG version or by any other text editor.




Finding out the solution structures is achieved through three successive steps (Figure 3) :

1) The correlations are assigned to $^{1}J$ or $^{2}J$ for bond generation. Another interpretation level is necessary when a correlation between a signal and a group of signals is processed. Correlations of complete atoms (i.e. having their expected number of neighbours) are considered first. If none exist the atoms which are the closest from completion will, in fact, be completed.

2) When all the correlations are used the remaining incomplete atoms are paired.

3) The occurrence in the solution of a partly assigned (or unassigned) given substructure is searched. The position of double bonds between $sp^2$ atoms is determined (only one disposition is found even if there are several). The solution is then directed to the output device (video display or disk file).

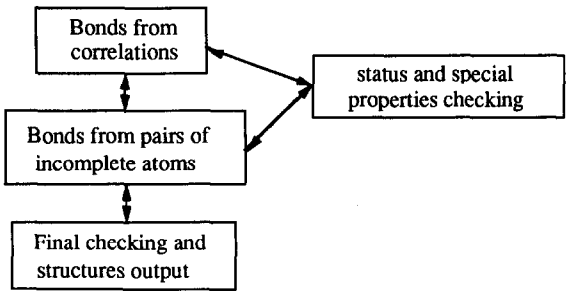



Figure 3. Main functions of the LSD program

In the first two steps, the bonding process continuously refers to the status and special properties data. The building and the maintenance of subsets of non-bound equivalent (i.e. of identical status) heteroatoms limits the search for different solutions representing the same structure. The solutions are PROLOG facts introduced into the data base by the program during resolution. They are : pairs of bound atoms, pairs of bound $sp^2$ atoms, hypotheses about correlations of groups and correspondances between atom order numbers and substructure numbers. Compound **1** affords two solutions in 5 min. 16 sec. of computing time; they are identical but differ in the assignments of the correlation between group a (signal 19 or 20) and signal number 2. In order to suppress the output of completely identical structures each new one is stored in a library, and if a structure is found to be identical to one previously found, it is rejected. Valid structures are finally converted into a format suitable for graphic output.

Five other compounds have been submitted to analysis by the LSD program : two sesquiterpenes and three indole alkaloids; resolution times were between 12 sec and 40 min. At the present time only one structure of a new compound has been determined by LSD. Compound **2**, extracted from the stem bark of *Peschiera buchtieni* among other known indole alkaloids[15], gives no clear mass spectral data and exhibits unusual proton chemical shifts. From $^1H$ and $^{13}C$ spectra it is deduced that compound **2** contains an N-methyl indole nucleus and an ethylidene side-chain, like compound **1**. The number of nitrogen atoms is supposed to equal 2, as in the great majority of monoterpenic indole alkaloids. From the $^{13}C$ data exactly five carbon atoms should bear heteroatoms, the odd number of carbon-bearing hydrogen atoms indicating that these heteroatoms globally bear an odd number of hydrogen atoms, the molecule containing two nitrogen atoms. The non-indolic nitrogen atom is either that of a secondary or tertiary amine or of a quaternary ammonium ion. For each of these possibilities there are two compatible numbers of oxygen atoms, as depicted in Figure 4.

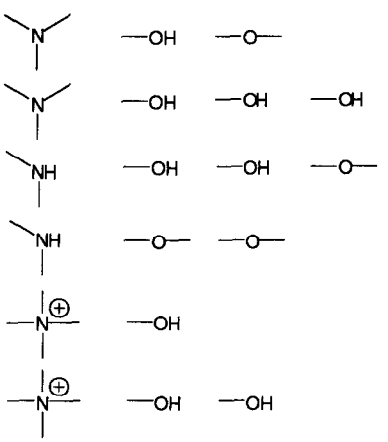



Figure 4. Status of heteroatoms in compound **2**

There should be six heteroatom-carbon bonds (the indole nucleus being not considered), one of these carbon atoms bearing two heteroatoms. The presence of an acetal group is excluded on the basis of chemical shifts. The six possible data sets are processed within 6 minutes yielding 20 solutions, listed in Figure 5.

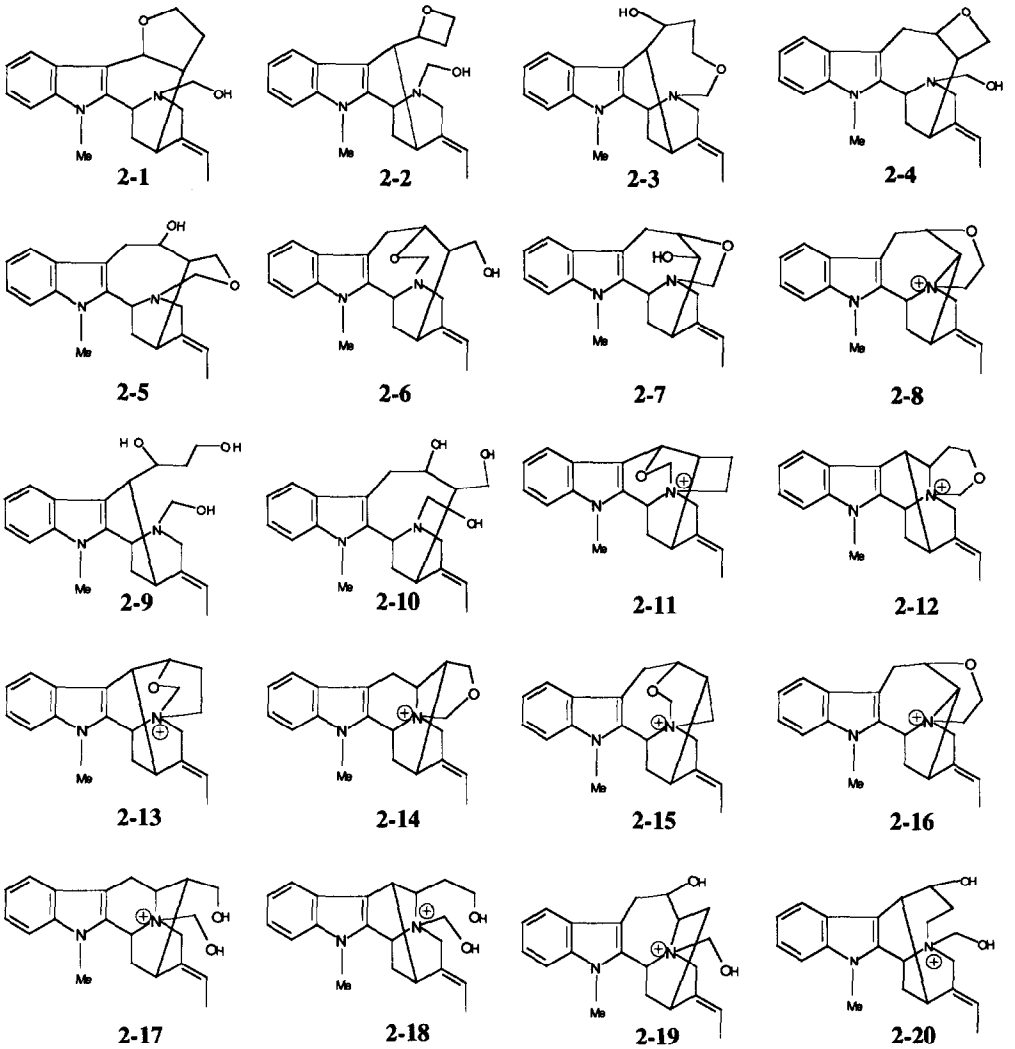



Figure 5. Solutions generated for compound **2**

At first the quaternary ammonium hypothesis was not considered and solution **2-6**, close to the known structure of affinisine **3** (see structure at bottom of page), was retained. The transformation of affinisine to **2-6** might involve (Figure 6) oxidation of a C-N bond, hydrolysis of the iminium salt, reduction of the carbonyl group to the alcohol, condensation of the amino group with formaldehyde, and addition of the alcohol to the iminium group.

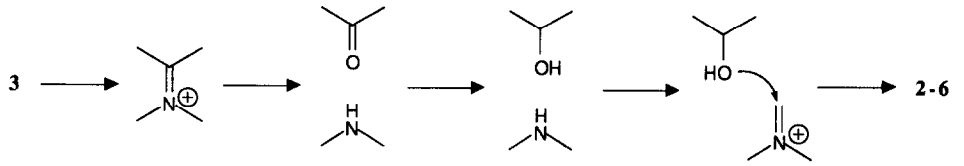
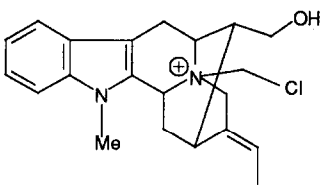
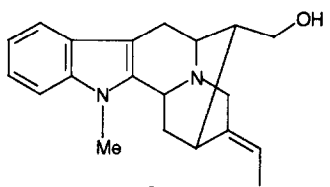



Figure 6. A possible biogenesis for compound **2-6**

Structure **2-6** is however not compatible with the ROESY[16] spectrum of compond **2** : two hydrogen atoms, separated by more than 5Å (measured from molecular models) exhibit a clear correlation peak. No other solution was satisfactory from the biogenetic point of view. Considering then the possibility of a quaternary nitrogen atom, solution **2-14** appeared as an intact affinisine molecule condensed with formaldehyde. Clearly such a molecule is unstable and cannot be isolated! According to the already observed condensation of a tertiary base with dichloromethane[17] during its extraction and/or purification process, structure **2** has finally been retained. This structure is confirmed by the synthesis of **2** by refluxing affinisine **3** in dichloromethane : the NMR spectrum of the crude reaction mixture shows the superimposition of the starting material signals and the characteristic peaks of **2**.




**2**




**3**

Perspectives and conclusion

Several improvements are conceivable for data input and processing : an interactive scanning process for spectra would be helpful to produce more reliable information and a less rigid format for atoms status would be appreciated for molecules containing many heteroatoms. Processing is based on a recursive generation-and-test algorithm which is not the most efficient one when dealing with such constrained generation problems. Substructural information should act during the resolution process instead of being considered only in the final checking stage. All these developments are under study and results will be published in due course.

During the first forty years of the development of NMR, most of the efforts of chemists have been concentrated on the generation of chemical shifts values and on their understanding. The increase of the available magnetic fields now shows that chemical shifts (especially $^{1}H$ chemical shifts) depend on many factors and this constitutes a limitation to their use in computer assisted structure elucidation programs. Owing to recent developments in electronics, computer science and NMR understanding, new experiments have been developed which provide connectivity information. The LSD program presented here shows that this information is an adequate input into the logic of automated structure solving, and it is conceivable that in the years to come, full automation and linking of NMR experiments and of such programs will provide "off the screen" structures from pure samples of reasonable molecular weight.

## EXPERIMENTAL

All spectra have been recorded on a BRUKER AC300 NMR spectrometer modified to support reverse experiments. The HMBC and HMQC ( BIRDD9 and INVDR2LP microprograms) data were sets of 256 FIDs of 2K points. Direct and long-range coupling constants have been chosen equal to 135 Hz and 7 Hz respectively. An unshifted sine-bell window function has been applied in both dimensions prior to double real Fourier transform. Samples were not spun and four dummy scans were performed prior to data accumulation.

Acknowledgments

We thank C.N.R.S. for financial support and Mrs L. Le Men for her kind attention. Pr. I. UGI (Technische Universität München, Garching, Germany) and his coworkers are specially thanked for giving their molecular graphic output program.

## REFERENCES

1. Panaya, A.; Doucet, J.P.; Cayzergues, P.; Carrier, G.; Matthieu, G. *L'actualité chimique* **1988,** 103-11 and references cited therein.

2. Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. Application of Artificial Intelligence for Organic Chemistry. The DENDRAL project **1980,** Mc Graw-Hill, New York.

3. Aue, W. P., Bartholdi, E. and Ernst, R. R. *J. Chem. Phys.* **1976,** *64*, 2229-46.

4. Kraulis, P.J. *J. Magn. Reson.* **1989,** *84*, 627-33.

5. Bax, A.; Freeman, R.; Frenkiel, T. A. *J. Am. Chem. Soc.* **1981,** *103*, 2102-4.

6. Christie, B.D.; Munk, M.E. *Anal. Chim. Acta* **1987,** *200*, 347-62.

7. Bax, A.; Subramanian, S. J. *J. Magn. Reson.* **1986,** *67*, 565-9.

8. Bax, A.; Summers, M. F. *J. Am. Chem. Soc.* **1986,** *108*, 2093-4.

9. Müller, L. *J. Am. Chem. Soc.* **1979,** *101*, 4481-4.

10. Nuzillard, J. M.; Massiot, G. *Anal. Chim. Acta* **1990,** in press.

11. Clocksin, W. F.; Mellish, C. S. Programming in PROLOG **1981,** Springer Verlag, Berlin.

12. Doddrell, D. M.; T Pegg, D. T.; Bendall, M. R. *J. Magn. Reson.* **1982,** *48*, 323-7.

13. Pinchon, T. M.; Nuzillard, J. M.; Richard, B.; Massiot, G.; Le Men-Olivier, L.; Sévenet, T. *Phytochemistry* **1990,** *29*, 3341-4.

14. Le Men, J.; Taylor, W. I. *Experientia* **1965,** *21*, 508-11.

15. Extraction and chemical investigation of this plant material was performed in our laboratory by Zèches, M., Azoug, M. and Richard, B.

16. Bothner-By, A. A.; Stephens, R. L.; Lee, J. M.; Warren, C. D.; Jeanloz, R. W. *J. Am. Chem. Soc.* **1984,** *106*, 811-13.

17. Besselièvre, R.; Langlois, N.; Potier, P. *Bull. Soc. Chim. Fr.* **1972,** 1477-8.