



Tutorial for the structure elucidation of small molecules by means of the LSD software

Jean-Marc Nuzillard, Bertrand Plainchont

► To cite this version:

Jean-Marc Nuzillard, Bertrand Plainchont. Tutorial for the structure elucidation of small molecules by means of the LSD software. *Magnetic Resonance in Chemistry*, 2018, 56 (6), pp.458 - 468. 10.1002/mrc.4612 . hal-01904950

HAL Id: hal-01904950

<https://hal.univ-reims.fr/hal-01904950>

Submitted on 9 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tutorial for the structure elucidation of small molecules by means of the LSD software

Jean-Marc Nuzillard*, Bertrand Plainchont

University of Reims, UMR CNRS 7312.

Keywords: NMR; ¹H; ¹³C; ¹⁵N; computer-assisted structure elucidation; artificial intelligence.

Abstract.

Automatic structure elucidation of small molecules by means of the “Logic for Structure Elucidation” software (LSD) is introduced in the context of the automatic exploitation of chemical shift correlation data and with minimal input from chemical shift values. The first step in solving a structural problem by means of LSD is the extraction of pertinent data from the 1D and 2D spectra. This operation requires the labeling of the resonances and of their correlations; its reliability highly depends on the quality of the spectra. The combination of COSY, HSQC and HMBC spectra results in proximity relationships between non-hydrogen atoms that are associated in order to build the possible solutions of a problem. A simple molecule, camphor, serves as an example for the writing of an LSD input file and to show how solution structures are obtained. An input file for LSD must contain a non-ambiguous description of each atom, or atom status, which includes the chemical element symbol, the hybridization state, the number of bound hydrogen atoms and the formal electric charge. In case of atom status ambiguity, the pyLSD program performs clarification by systematically generating the status of the atoms. PyLSD also proposes the use of the nmrshiftdb algorithm in order to rank the solutions of a problem according to the quality of the fit between the experimental carbon-13 chemical shifts and the ones predicted from the proposed structures. To conclude, some hints toward future uses and developments of computer-assisted structure elucidation by LSD are proposed.

Introduction.

The structure elucidation of small organic molecules is still a process in which computers play a marginal role in the present area of artificial intelligence (AI) revival, even though computer-aided structural elucidation (CASE) was one of the first playgrounds of AI.^[1,2] Nuclear magnetic resonance (NMR) and mass spectrometry (MS) are two complementary analytical methods that allow chemists to *deduce* molecular structures from spectra. The pioneering AI works aiming to create CASE software involved both methods and were based on the reproduction of the reasoning mechanism of experienced chemists. Even though MS spectra are considered as less informative than NMR spectra in what concerns the access to fine structural details, MS gives access to molecule gross formula and therefore defines the nature and number of the atoms that constitutes a molecule, which is of capital importance for a successful structure elucidation. MS also reveals the constitution of possible molecular fragments. The following lines emphasize on the role of NMR in CASE systems. A recent review article on the automated interpretation of NMR spectra for small organic molecules in solution provides an overview of currently available academic and commercial CASE software as well as bibliographic references of review articles.^[3]

Liquid state NMR spectra, as those commonly used in small molecule studies, are ruled by two fundamental parameters: chemical shifts (δ) and scalar coupling constants (J). Chemical shifts are defined for the NMR active nuclei of molecules (^1H , ^{13}C , and when applicable ^{15}N , ^{19}F , ^{31}P ...) and reflect the modification of their interaction with the intense static B_0 field of the spectrometer by the electronic cloud. The δ value of a nucleus in a molecule is independent of B_0 and is strongly related to the structural features of the nucleus environment, such as the electronegativity of neighboring atoms or the vicinity of multiple bonds or of aromatic systems. A particular chemical shift value is usually compatible with several molecular fragments. A set of such values, possibly obtained for different nucleus types, and a fragment association algorithm may lead to the assembly of one or more chemically plausible structures. The first CASE systems were based on this atom centered approach. The proposed structures may be validated or not by the ability to account for the observed scalar couplings. Such couplings occur through an indirect nucleus-electron-nucleus magnetic interaction and their intensity J is related to the distance n between nuclei measured as the number of separating bonds. The nJ values globally decrease with n but, apart in some well-defined cases, it is not possible to unambiguously deduce n from the magnitude of nJ . Any pair of NMR active nuclei within a molecule presents a scalar coupling but in practice it cannot be detected when its value falls below resonance line width. The couplings through 4 bonds in ^1H - ^1H and ^1H - ^{13}C nuclei pairs are often difficult to detect, unless double or aromatic bonds are present along the path or under particular geometrical arrangement of the bonds. Longer range couplings ($n > 4$) are even more rarely detected. The existence of significantly large scalar couplings constitutes therefore a strong constraint for the validation of structures automatically generated by assembly of chemical shift derived fragments. This approach was the one proposed in 1991 by Christie and Munk in their SESAMI software.^[4] The same year, we proposed a radically new approach to structure elucidation in which scalar couplings drive the formation of bonds in the "Logic for Structure Determination" software (LSD).^[5] The present tutorial article is devoted to the LSD software.

LSD started as a list of PROLOG rules for the transformation of NMR data into structures.^[6] Most of the concepts on which the present version of LSD relies were elaborated at that time. Execution speed and software portability considerations led to a migration from declarative programming in PROLOG to procedural programming in C language. The practical resolution of structural problems lead to further software improvements, resulting in the present version that can be freely downloaded.^[7] The name LSD designates the structure generation algorithm, *lsd*, and the companion programs *outlsd*, *genpos*, *m_edit*, and *mol2ab*, as well as the documentation and example files. *Outlsd* translates solution files produced by *lsd* into SD files,^[8] SMILES chains,^[9] and 2D coordinate files in a format that can be converted by *genpos* to PostScript structure description files or modified by *m_edit*, a rudimentary structure drawing editor. PyLSD, a software layer above LSD

that is written in the Python computer language, was created in order to remedy to limitations of LSD. PyLSD is also freely available.^[10]

Experimental data for *lsd* input

In the best possible situation, the high-resolution mass spectrum of an unknown substance produces a reliable exact gross formula. It will be assumed here that this requirement is met. Each peak in the ^{13}C NMR spectrum of the unknown is associated to an index (starting at 1) which is also associated to a carbon atom number in the structure of the unknown. The number of ^{13}C NMR resonances is equal to the number of carbon atoms if the ^{13}C chemical shift values are all different for all carbon atoms. Two resonances with identical chemical shifts must therefore receive different index values in order to consistently relate atom indexes and resonances indexes. Resonance superposition may happen for symmetry reasons or may be purely accidental. The numbering of ^1H NMR resonances is best achieved using the ^1H - ^{13}C 2D HSQC spectrum that pairs, or equivalently said, correlates the chemical shifts of ^1H and ^{13}C nuclei that are directly bound. A ^1H resonance receives an index that is identical to the one of the carbon it is attached to, thus defining an index for what organic chemists call a position. Two anisochronous ^1H nuclei in a methylene group are thus identified by the same index. Other atoms than carbons and hydrogens must also receive arbitrary indexes. Hydrogen atoms bound to nitrogen atoms may be identically indexed if a ^1H - ^{15}N HSQC spectrum is available. Slowly exchanging hydrogens in alcohol groups, most often visible for compounds dissolved in $\text{DMSO}-d_6$, give rise to reasonably sharp resonances and may be also arbitrarily numbered according to the numbering of sp^3 , hydrogen bearing oxygen atoms; these ^1H NMR signals are most often very helpful in the structure elucidation process. The HSQC spectra should be preferentially recorded in the so-called multiplicity edited mode that produces 2D peaks of opposite signs for the resonances of the carbon atoms bound to an odd or even number of hydrogens. Such a spectrum, in conjunction with the 1D ^1H and ^{13}C spectra gives access to the exact number of hydrogen atoms bound to each carbon atom, which in itself constitutes a highly useful structural information about the unknown. The indexing of ^1H resonances is not as straightforward as the one of ^{13}C resonances because of possible partial or complete ^1H multiplet overlapping, while ^1H broadband decoupling leaves a single resonance line for each type of ^{13}C nucleus. The recent advent of pure shift ^1H NMR spectra will certainly open the way to a simplified protocol for the 1D and 2D NMR spectra annotation.^[11]

The ^1H - ^1H COSY spectra are most often recorded in single phase modulation mode using gradients for coherence pathway selection. The time-domain data are filtered using sine-arch multiplication in order to reduce the peak tails caused by the large dispersive contributions to peak

shape, thus resulting in a large but generally not critical sensitivity loss in the corresponding magnitude mode spectra. It might be useful to obtain a rough evaluation of the magnitude of the active coupling that is associated to COSY cross-peak. $|^nJ|$ with $n>3$ are rarely greater than 3 Hz. Such information can be found in phase-sensitive double quantum filtered COSY spectra, even though the phase correction of such spectra requires some care that is not necessary for the processing of magnitude mode spectra. Alternatively, degrading the resolution of the COSY spectrum at the processing stage may be a (dirty) way to eliminate the COSY cross peaks that originate from small, long-range couplings. The not so rare superposition of ^1H multiplets make the COSY spectrum sometimes not as useful as expected. Most of the structural information used by *Isd* comes from the HMBC spectra. The echo-antiecho version of the “classical” gradient-enhanced HMBC pulse sequence, with improved rejection of 1J coupling signals, gives satisfactory results in terms of sensitivity and artifact elimination. The 1D ^1H , 1D ^{13}C , 2D ^1H - ^1H COSY, 2D ^1H - ^{13}C HSQC and 2D ^1H - ^{13}C HMBC spectra thus constitute the minimal set of necessary experimental NMR data that may be completed by 2D ^1H - ^{15}N HSQC, 2D ^1H - ^{15}N HMBC, 2D ^1H - ^{13}C H2BC, 2D ^1H - ^{13}C 1,1-ADEQUATE or even 2D ^{13}C - ^{13}C INADEQUATE spectra (see Fig. 1).

Resolution issues

Resolution in 2D HSQC and HMBC spectra is a parameter of vital importance for structure elucidation, either manual or automatic. The actual tendency to delegate spectra recording to instrumental facilities that rely on sample changer operation and standard recording parameters specially optimized to save time and money may result in still usable spectra but also may not. Automatic shimming is a wonderful invention and no one could reasonably work without it, especially on high field NMR instruments. However, it might happen that, for any reason, the ^1H 1D spectrum does not present the expected resolution. In this case there is not point to go on with the recording of 2D spectra, considering that a poor magnet shimming always causes a sensitivity loss. A human being has to take a decision in order to obtain high quality spectra and not to leave the problems to the person in charge of deducing structures from low-quality spectra. It must be kept in mind that the acquisition time in the direct time domain of 2D experiments is available at nearly no cost. It is therefore unwise to record an FID in a 2D experiment with a short t_2^{max} acquisition time. At 600 MHz, a 16 ppm ^1H spectral width correspond to about 10 kHz, resulting in a 0.05 ms dwell time, so that 2K real data points are acquired in about 100 ms. There is generally plenty of signal to record 100 ms after the beginning of an FID, a signal that is simply thrown away by careless operation and whose recording would have improved the spectral resolution and sensitivity of HMBC and COSY spectra at the negligible cost of some hard disk space and processing time increase. However, the

HSQC spectra recorded with a cryoprobe constitute an exception because the necessary ^{13}C decoupling power and duration might not be compatible with a safe operation of the probe due to radio-frequency induced sample heating. Identically, a poor resolution in the ^{13}C chemical shift domain of HSQC and HMBC spectra brings an uncertainty on the indexing of some 2D peaks which is always detrimental to the quality of structure determination. The presence of close ^{13}C resonances calls for the application of a resolution enhancement strategy. The simplest one is simply the increase of t_1^{max} (also known as acquisition time in the indirect time dimension) by increasing the number of t_1 increments. Overall acquisition times may be then kept in practical limits by reducing accordingly the number of transients per FID, keeping in mind that the overall spectral signal to noise ratio is governed by the total number of recorded transients (and not by the number of transients per t_1 increment), especially in heteronuclear experiments for which signal decay in the indirect dimension is very slow. The presence of close resonances in the 1D ^{13}C NMR spectrum should therefore incite to set a sufficiently high number of t_1 increments in HSQC and HMBC spectra that would result in an appropriate F_1 resolution. Other alternatives for resolution enhancement in the indirect dimension include band-selective spectra,^[12] spectral aliasing^[13] and non-uniform sampling.^[14] To summarize this long paragraph, it is easier to obtain pertinent structures from good quality data.

Principle of structure assembly

Lsd makes a distinction between the hydrogen atoms of a molecule and all the other ones, also referred to as heavy atoms. *Lsd* connects the heavy atoms of a molecule, under the assumption that each hydrogen atom shares only one single bond with a heavy atom. A bond between two heavy atoms is first created as a single bond. The multiple character of bonds is determined only when the graph is completely created. This means that the hybridization state of each atom must be known as well as that the number of its neighbors (or the number of vertices of each node) and its formal electric charge. The number of neighbors of an atom is deduced from its status which includes the atomic symbol, the index (*vide supra*), the hybridization state, the number of attached hydrogen atoms (called multiplicity) and the electric charge. Chemical elements such as phosphorus or sulfur may be present in more than a single valence state. The valence must be appended to the atomic symbol for “less common values”, like in S4 for the sulfur atom of a sulfoxide or in S6 in a sulfone. The sulfur atom of a sulfoxide and of a sulfone must have their hybridization state artificially declared as sp^2 and sp because they bear one double bond and two double bonds, respectively. The indication of atomic formal electric charges was introduced in order to represent, among others, molecules

containing quaternary ammonium, amine oxide or nitro groups. A nitro group may be written either as $\text{--N}^+(\text{=O})\text{--O}^-$ or as --N5(=O)=O , with sp^2 and sp hybridized N and N5 atoms, respectively.

A COSY chemical shift correlation between two ^1H resonances through n bonds is related to a distance of $n-2$ bonds between the heavy atoms that are directly bond to the correlating hydrogens. A 2J COSY correlation is only visible for two resonances of anisochronous ^1H nuclei within a methylene group, a situation that is revealed in a straightforward manner by the ^1H - ^{13}C HSQC spectrum; such a COSY correlation does not bring any useful information on heavy atom connectivity. The same consideration holds for NH_2 groups, for which two different ^1H chemical shifts are paired in a ^1H - ^{15}N HSQC spectrum by a correlation with a single ^{15}N chemical shift. A 3J COSY correlation indicates the existence of a bond between two heavy atoms. Unless otherwise stated, *lsd* considers a COSY correlation as arising from a 3J coupling. In the same way, a nJ HMBC correlation reveals a path made of $n-1$ bonds between two heavy atoms. By default, *lsd* considers that the two heavy atoms are either bound together or both bound to the same unknown intermediate atom (see Fig. 1). The *lsd* algorithm sets bonds between the heavy atoms that are *de facto* bound according to the COSY spectrum and makes use of ^1H - ^{13}C and ^1H - ^{15}N HMBC data in order establish one-bond or two-bond connections between heavy atoms, as if 1J or 2J ^{13}C - ^{13}C and ^{13}C - ^{15}N chemical shift correlations were recorded. *lsd* works internally with such fictitious correlations between heavy atom resonances.

Once all correlation data are exploited by *lsd*, it rarely happens that all heavy atoms received the number of neighbors that their status prescribes. The so-called incomplete atoms are then systematically paired by *lsd* in order to obtain chemically correct structures.

The creation of bonds either from correlation data or from final atom pairing is placed under atom property control. Atom properties were initially designed in order to implement user-supplied atom neighborhood constraints based on chemical shift values. A shielded carbon atom may be constrained to have only carbon atoms as neighbors. A shielded carbon of a methyl group for which the ^1H resonances is a singlet may be forced to have a quaternary carbon as neighbor. Atom properties are used during structure generation each time a new bond is created.

Bond creation by means of HMBC data consists in the repetition of a two-step process: correlation data selection and data exploitation.^[15] The order in which the correlations are selected has a strong impact on execution times. *lsd* selects first the correlations of the atoms for which the number of bonds that are missing in order to reach completion is the smallest. Completion of an atom is reached when all neighbors are known. Correlations of complete atoms are therefore selected in priority. *lsd* also tends to favor the selection of correlating atoms for which bonds have been recently established, so that bond formation concentrates around the same places, as much as possible. Once a proximity relationship between two heavy atoms X and Y is selected, *lsd* attempts first to bind X and Y, and, if possible, a new selection-exploitation step takes place. When all

consequences have been explored, *lsd* searches all atoms Z of the molecule for which the bonds X-Z and Z-Y can be established, and, if possible, a new selection-exploitation step takes place again. A third possibility for the exploitation of a correlation consists in leaving it unexploited. The user may specify a maximum number of correlations that can be ignored. This possibility allows *lsd* to find solutions even if HMBC correlations through more than three bonds are present. The systematic exploration of all possible interpretations of correlations places *lsd* in the category of the deterministic algorithms,^[16] by opposition to stochastic algorithms.^[17]

The double and triple bonds are placed in a structure only when all atoms are complete. *Lsd* stops the placement of multiple bonds once a first placement is found. This avoids to obtain the two possible double bond placements in an isolated aromatic ring as separate solutions. The default *lsd* behavior may therefore lead to the missing of realistic double bond placements in the not so frequent anti-aromatic rings. A structure that does not follow the Bredt's rule,^[18] meaning that a double bond is placed at the bridgehead of a small size bicyclic ring system, can be eliminated (or not) from the list of the solutions of a problem.

The solutions for which multiple bond placement and verification of Bredt's rule compliance was successful may be kept or discarded according to the result of a substructure search. Substructures are proposed from prior knowledge on sample origin as biogenesis for natural products or synthesis scheme for synthetic substances or simply from the recognition of a particular chemical shift value and/or coupling constant pattern that is associated to a particular structural feature.^[13,19]

A lack of resolution in the ^{13}C domain of an HMBC spectrum may lead to an uncertainty for the labelling of correlation peaks. The treatment of uncertain correlations by *lsd* is possible and may lead to identical solutions when two interpretations of the same uncertain data produce the same structure. The output of identical structures with identical assignments can be prevented by the storage of all solutions at the time they are produced and by the comparison of each new one with the already found ones. The production of identical structures with different resonance assignments can be avoided through the determination of InChi character strings;^[20] this option must not be used if solution ranking on the basis of chemical shifts values is performed because a pertinent assignment may be eliminated only because a less pertinent assignment was proposed earlier for the same structure.

Lsd and *pylsd* produce planar structures because they do not use any data that would bring information about the relative configurations of the various chirality elements a molecule may contain. The stereochemical assignment in small organic molecules may rely on the study of nuclear Overhauser effect,^[21] residual dipolar couplings,^[22] or ab initio chemical shift calculations.^[23]

Coding an *lsd* file

The LSD software distribution file includes ready-to-use compiled executable files of *lsd* for the two most common computer systems. The Linux version contains the source code and is easily compiled. Manuals for software installation and use, written in English and French, are available directly from the LSD web site and from the unpacked distribution files. The LSD software has no graphical interface for data input. The user has to describe a problem by writing a text file whose name has an *lsd* extension, at least for Windows systems. All details are given in the manual and the present article does not cover them all. A very simple compound, camphor, is used instead as example to illustrate the main aspects of *lsd* input file writing. The files in the Data directory (or folder) also provide some additional information. The ^1H , ^{13}C , COSY, HSQC and HMBC spectra of camphor were recorded; Fig. 2 presents the ^{13}C and HSQC spectra with the associated resonance numbering and Fig. 3 shows the annotated COSY and HMBC spectra. The content of the corresponding *lsd* input file is reported in Fig. 4. The gross formula of camphor is $\text{C}_{10}\text{H}_{16}\text{O}$. The ten ^{13}C NMR resonances are labelled from 1 to 10 in the order of decreasing chemical shifts and the ^1H resonances are labelled according to the HSQC spectrum. The carbon atoms are numbered identically to their ^{13}C resonance and the oxygen atom is numbered 11 in order to avoid “holes” in the numbering. One could argue here that the gross formula could have been deduced from NMR alone, but what is possible here is more an exception rather than a rule.

An *lsd* input file is a succession of commands and comments. Comments are portions of text located between a semicolon and the end of the line it belongs to; they are generally present at the beginning of a file in order to document the origin of the compound under investigation. A file is conventionally divided into sections containing commands for execution control, heavy atom status description, bonds and correlation data, atom properties, and substructure definition. Sections may be separated by empty lines or comments for a good readability. A command always starts at the beginning of a line, has a 4-character mnemonic code and is followed by arguments separated by spaces. Considering camphor (see Fig. 4), the command “MULT 1 C 2 0” means that atom 1 is a carbon with a sp² hybridization state and no hydrogen atom attached (null multiplicity). The electric charge of the atom is an optional parameter, irrelevant in the present example, which may be specified as the last argument of the MULT command. The MULT name was historically given to the atom multiplicity property in the initial PROLOG code. The “SHIX 1 219.97” line is not used by *lsd* and should be considered here as a comment; it is used by *pylsd* for solution ranking but not as a constraint for structure generation. The SHIX code stands for “chemical SHift of X nuclei”, an X nucleus being any NMR sensitive nucleus but ^1H . The chemical shift of C-1 indicates it belongs to keto group whose oxygen is necessarily O-11. The bond between atoms 1 and 11 is given in the starting

point of the structure by the "BOND 1 11" command, even though this is not a necessity. The "HSQC 4 4" command states that C-4 (first argument) is bound to H-4 that stands for two hydrogens on C-4 at position 4 that are both numbered H-4. The HSQC section of the file looks non-informative due to the way hydrogen atoms are numbered but is kept as such to maintain compatibility with files in which number synchronization was not imposed. For bookkeeping only, the chemical shifts of ^1H nuclei may be stored in SHIH (chemical SHifts for ^1H nuclei) commands such as "SHIH 4 2.328"; due to the existence of two H-4 atoms, the chemical shift of the second H-4 is given by "SHIH 4 1.819". Neither *lsd* nor *pylsd* exploits these values presently. The COSY correlation of a H-6 with a H-7 resonance leads to write "COSY 6 7" in the file, which sets a bond between C-6 and C-7. The "HMBC 1 4" command indicates that the chemical shifts of C-1 and H-4 correlate in the HMBC spectrum, thus revealing either the existence of a C-1/C-4 bond or of a C-1/Z and Z/C-4 bond pair in which atom Z is unknown. The chemical shift similarity of C-4 and C-5 induces an ambiguity in the interpretation of the HMBC data relative to C-4 and C-5. Therefore, the spectrum shows that either C-4 or C-5 (or both) correlates with H-9, an observation that is coded as "HMBC (4 5) 9". The *lsd* algorithm systematically considers this command like "HMBC 4 9" and then like "HMBC 5 9" at the time it is exploited. The "HMBC (4 5) 4" command is automatically reduced to "HMBC 5 4" because C-4 cannot correlate with H-4 and is then considered as useless and eliminated from the correlation set because C-4 and C-5 are known to be bound from COSY data. In a similar way, the correlations that become explained during resolution are temporarily invalidated: the setting of an X/Y bond explains an X/Y correlation, as well as the formation of an X/Z bond if a Y/Z bond already exists. Interestingly, the resolution in HSQC spectra is sufficient to clearly differentiate the correlations of C-4 and C-5, while the same distinction is harder in the HMBC spectra. This observation may be explained by a higher resolution in the indirect dimension of the HSQC spectrum due to a smaller spectral width (typically 160 ppm instead of 240 ppm for the same number of t_1 increments) and by the intrusion of ^1H - ^1H scalar couplings in the indirect dimension of HMBC spectra. A resolution increase in the indirect dimension of the HMBC spectrum reduces the number of such ambiguous correlations with a possible reduction of the number of solutions. Finally, the last three lines of the *lsd* input file for camphor define properties of atoms C-8, C-9 and C-10. Each of these carbons are shielded and the hydrogen atom they bear resonate as singlets, so that it may be safely written that they are bound to quaternary carbons. The writing of properties proceeds through the definition of atom lists. "LIST L1 8 9 10" defines L1 as the list containing C-8, C-9 and C-10. "QUAT L2" defines L2 as the list of the quaternary carbon atoms of the molecule. "PROP L1 1 L2" means that each element in L1 has exactly one neighbor in list L2. In the present case, a property is set for each element of a list, but a property may be defined for a single atom. The LSD manual details all the possible ways to define lists and atom properties.

The writing of an input file for *lsd* is considered as fussy by many people who would have expected to enter data through a graphical user interface. The difficult point in the creation of such an interface is the extraction of correlation data from 2D NMR spectra. Human expertise is rather often required to achieve this task without introducing errors or to correct the outcome of an automated process. Moreover, motivated people are not deterred from using *lsd* simply because a text has to be written.

Practical structure generation

The structure of camphor can be deduced from the *lsd* input file named *camphor.lsd*, assuming that a terminal window (also called command window) is opened, that the working directory is the one that contains the *lsd* executable file, that executable files are allowed to be run from it, and that it contains the *camphor.lsd* file. Typing "*lsd camphor.lsd*" (without double quotes) in the command window and hitting the Enter key should produce a result similar to Fig. 5, the exact appearance of which depends on the operating system and on configuration flavors. A single solution is found and stored in the *camphor.sol* solution output file.

The camphor problem has indeed two solutions that are "assignment isomers", meaning that the two structures are the same but their assignment of the ^{13}C NMR resonances are different. By default, *lsd* does not display the two assignment isomers but only the first one that is found, the second one being considered as the repetition of the first one. The execution control command "DUPL 1" modifies this default behavior. Writing this command, preferentially before the MULT block, saving the file and rerunning *lsd* provides two solutions. The two solutions only differ by a permutation of C-2 and C-3. No argument in chemical shift correlation data can induce a preference for one or the other possibility; only chemical shift values make a difference (*vide infra*).

The solution file starts with a copy of the input file and is followed by a data block for each solution that contains the status of the atoms and the list of their bonds. The graphical display of the solutions requires an atom coordinate generator and a graphical structure renderer. The *outlsd* program is able to convert the solution file into a list of SMILES string, into an SD file with 2D atom coordinates, or into a coordinate file whose format was defined in the early days of the LSD software. According to the LSD manual, the production of the coordinate file of camphor is obtained by typing "*outlsd 6 < camphor.lsd > camphor.coo*". Rendering is carried out by typing "*genpos < camphor.coo > camphor.ps*", which produces a PostScript® document with one page per solution. The first solution is presented in Fig. 6. Because an SD file with embedded 2D coordinates can be produced by typing "*outlsd 7 < camphor.sol > camphor.sdf*", the user may rely on other chemoinformatic tools in order to obtain possibly better visual solution renderings. Structure diagram generation by *outlsd* is far

from perfection and structures are sometimes (often) unreadable. A tool called *m_edit* operates on coordinate and SD files created by *outlsd* in order to improve interactively the structure diagrams. The *m_edit* computer code is written in the Tcl/Tk programming language, for which the interpreter must be installed by the user. Fig. 7 shows the manually improved drawings of the two solutions obtained for camphor. Typing “solve camphor” in the terminal window chains the structure elucidation and solution display steps by successively calling *lsd*, *outlsd* and *genpos*.

Substructure search

If one has information about the camphor molecule and knows it might be a non-rearranged terpene, or equivalently that its structure includes two distinct isoprenic units, the commands in Fig. 8 select the solutions that fit with the provided sub-structural data. A sub-structure is defined by sub-atoms (SSTR commands) and sub-bonds between sub-atoms (LINK commands). The SSTR commands are similar to MULT commands but with noticeable differences. A sub-atom index starts with the letter S. The “A” element symbol stands for “any element symbol”. The hybridization state and multiplicity of sub-atoms may be chosen within sets of alternative values. A sub-structure can be made of non-connected parts like in the present example. However, an atom of the molecule will only be matched with a single sub-atom. This means that the basic sub-structure finding mechanism in *lsd* cannot deal with partly overlapping sub-structure elements. The writing of sub-structure files allows *lsd* to perform sub-structure search combinations such as “F1 or F2 and not F3”, if F1, F2 and F3 are related in the *lsd* input file to their respective sub-structure file names. The pinene example from the Data folder of LSD illustrates this feature. LSD presently contains a library of terpenic compound skeletons that was collected from the SISTEMAT knowledge base and a tool named *mol2ab* for the batch conversion of a set of MOL files into a collection of LSD substructure files.^[24]

Common pitfalls

A careful atom numbering and atom status determination, in accordance with the molecular formula, are two pre-requisites for a successful problem resolution. Experimental data over-interpretation constitutes the most common pitfall in the writing of a new input file for *lsd*. Even though it seems tempting to include the smallest correlation peaks of the HMBC spectrum in an input file, only those with an apparent high intensity should be introduced in a first run, the others being typed but inactivated by prefixing them with the comment delimiter (a semicolon). It happens that correlations from 4J couplings (a very long-range coupling, or VLRC) cannot be distinguished by their intensity from “regular” ones. The presence of VLRCs in an *lsd* input file and their attempted interpretation as correlations from 2J and 3J couplings result most of times in an empty list of

solutions. VLRCs had to be removed manually by the user with old LSD versions, so that all remaining HMBC commands are related to regular correlations. *Lsd* automates the elimination of VLRCs if an ELIM execution control command is present in the input file. An “ELIM x y” command allows *lsd* to eliminate up to x HMBC correlations that may arise from the existence of VLRCs through up to y bonds. The ELIM command operates during structure assembly: at the time an HMBC correlation is exploited for the creation of bonds, it is considered first as arising from a 2J coupling, then as a 3J coupling and finally as if it never existed if less than x correlations have already been eliminated. A resulting structure whose construction requires the elimination of a correlation through y+1 bonds or more is not considered valid. If a new *lsd* input file fails to produce a structure, and if the existence of a VLRC is suspected, then an ELIM command may be inserted for the elimination of a single correlation (x = 1). The value of x may be increased if the problem persists, at the price of an increase of the resolution time. By default, the ELIM command makes of any HMBC correlation a candidate for elimination. Each correlation may be associated to a range of coupling path lengths in order to override this default behavior. A “HMBC X Y 2 3” command indicates that the length of the coupling path between C-X and H-Y is comprised between 2 and 3, thus making it impossible to eliminate. A 1,1-ADEQUATE or a 2J H2BC correlation between C-X and H-Y is coded as “HMBC X Y 2” and forces the creation of a bond (see Fig. 1) while an INADEQUATE correlation between C-X and C-Y is directly coded as “BOND X Y”.

Incorrect constraints on atom status, atom properties and substructure presence result most of times in an absence of solution and the production of a message that advises the user to check for the pertinence of input data.

Handling fuzzy atom status and ranking structures with pyLSD

The LSD software presents two important limitations: the obligation to assign a non-ambiguous status to each heavy atom and the lack of a ranking method when many solutions are compatible with the available data. The pyLSD software was written to remove these limitations (see Fig. 9). The *pylsd* algorithm processes data sets in which atom status parts include alternatives (or variants). Command such as “MULT 20 N (1 2 3) (0 1 2)” is not accepted by *lsd* but indicates to *pylsd* that no hybridization state and no multiplicity is known about trivalent neutral nitrogen atom N-20. The exploration of all atom status combinations is carried out under the constraint of the molecular formula, indicated by a “FORM” command. Other constraints may be imposed such as a value or a set of values of molecular electric charge and a maximum number of positively and negatively charged atoms. The molecular formula may itself contain ambiguous parts, thus meaning that the number of atoms of an element type may be given within a range. The possibilities for these

numbers are explored under the constraint of the monoisotopic molecular mass, either provided as an exact value or within an interval. The status of the atoms that are not explicitly related to a MULT command is taken from a default status set or within a restricted status set defined by the user. The treatment of problems with ambiguous molecular formula still requires validation and possible revision of the corresponding computer code.

Pylsd first establishes a list of molecular formulas and their corresponding sets of commands. Each element of this list is then expanded for atom status disambiguation, resulting in a collection of problems *lsd* solves sequentially.^[25] The solution files of these problems are grouped together in order to build a single solution file. All the solutions of a problem may be considered as equivalently valid with respect to the atom status, atom proximity relationships, atom property and sub-structural constraints. Any chemist would not consider them equally possible because a problem must have only one solution; a good way to rank them is to take into account the chemical shift values. Chemical shift predictors being software of economically strategic importance, it is difficult to find free ones that are embeddable and that present a wide applicability field and a good reliability. To the best of our knowledge, *nmrshiftdb* is the only one that fulfills these requirements.^[26] *Nmrshiftdb* can be used for free through its website but Java files are publicly available for the setting up of a stand-alone application. Each structure in the solution file is submitted to *nmrshiftdb* for ¹³C NMR chemical shift prediction if the *pylsd* input file contains at least a "SHIX" command relative to a carbon atom. The absolute values of the differences between the predicted ¹³C NMR chemical shifts and the experimental ones are summed to yield a score value that is expected to be minimal for the best of all proposed solutions. The higher is the score of a solution and the less likely it is. However, a chemical shift prediction purely based on atom connectivity, in the absence of configurational and conformational information, has necessarily a limited accuracy. The correct solution of a problem might therefore not be ranked at the first place and may even be ranked far from it if the structures present rare chemical functions for which the examples of the database used for prediction are not representative. This situation was encountered during the testing of *pylsd*, using the complex structure of the insect antifeedant azadirachtin. Conversely, the structure of hexacyclinol was ranked at the first place among ten structures.^[25] Regarding camphor, solution 1 in Fig. 7 is ranked first because the quaternary carbon that is adjacent to the carbonyl group is predicted to be more deshielded (C-2: observed 57.93 ppm, predicted 57.30 ppm) than the one carrying the geminal methyl groups (C-3: observed 47.02 ppm, predicted 46.60 ppm).

Molecular formula and atom status fuzziness is resolved in *pylsd* by a computer code written in the Python language that also takes care of problem resolution by *lsd*, solution collection, solution ranking, structure diagram generation, and solution display. Solution ranking by means of the *nmrshiftdb* algorithm is achieved by a computer code written in Java language and whose execution

requires a Java run-time environment (JRE). The consequence for the user is the necessity to install a JRE and a Python interpreter or to check for their availability. A file named `defaults.py` needs to be edited by the user so that *pylsd* finds the path to all the required software pieces.

Conclusion

This brief overview of the LSD software cannot replace the experience gained by the users who take the time to practice the writing of *lsd* input files for real-life problem resolution. It must be kept in mind that the software does not bring more knowledge than given as input, it simply rearranges correlation data into bonds under the supervision of neighborhood constraints given by atom properties and substructures. The benefit of the software is that all possibilities opened by data interpretation are considered and that none is left unexplored during the course of the reasoning. Readers are encouraged to build their opinion by themselves about the usefulness of such an automated approach of structure elucidation and at least to use it for the search for alternative structures when one is proposed for an unknown molecule. The publication of new structures in chemistry journals might be facilitated if the authors could provide structure proofs to reviewers that rely on computer assistance. Future developments might concern the merging of the *lsd* and *pylsd* algorithms, the writing of an interface software for data input and for the post-processing of solution sets. The present availability of efficient algorithms for substructure search, structure file formatting, structure diagram generation and rendering from free cheminformatic toolkits will also benefit to future implementations of LSD.

Acknowledgments. Financial support by CNRS, Conseil Regional Champagne Ardenne, Conseil General de la Marne, Ministry of Higher Education and Research (MESR) and EU-programme FEDER to the PIAneT CPER project is gratefully acknowledged. JMN and BP respectively thank CNRS and MESR for financial support.

References.

1. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Applications of artificial intelligence for organic chemistry: the DENDRAL project*. McGraw-Hill, New York, **1980**.
2. M. Carabedian, I. Dagane, J.-E. Dubois, *Anal. Chem.* **1988**, *60*, 2186–2192. DOI: 10.1021/ac00171a005
3. J.-M. Nuzillard, *eMagRes* **2014**, *3*. DOI: 10.1002/9780470034590.emrstm1384

4. B. Christie, M. Munk, *J. Am. Chem. Soc.* **1991**, *113*, 3750–3757. DOI: 10.1021/ja00010a018
5. J.-M. Nuzillard, G. Massiot, *Tetrahedron* **1991**, *47*, 3655–3664. DOI: 10.1016/S0040-4020(01)80878-4
6. W. F. Clocksin, C. S. Mellish, *Programming in PROLOG*, Springer Verlag, Berlin, **2003**.
7. <http://eos.univ-reims.fr/LSD>
8. <http://download.accelrys.com/freeware/ctfile-formats/>
9. D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. DOI: 10.1021/ci00057a005
10. <http://eos.univ-reims.fr/LSD/JmnSoft/PyLSD>
11. R. W. Adams, *eMagRes* **2014**, *3*, 1–15. DOI: 10.1002/9780470034590.emrstm1362
12. C. Gaillet, C. Lequart, P. Debeire, J.-M. Nuzillard, *J. Magn. Reson.* **1999**, *139*, 454–459. DOI: 10.1006/jmre.1999.1808
13. G. B. Njock, D. E. Pegnyem, T.-A. Bartholomeusz, P. Christen, B. Vitorge, J.-M. Nuzillard, R. Shivapurkar, M. Foroozandeh, D. Jeannerat, *Chimia* **2010**, *64*, 235–240. DOI:10.2533/chimia.2010.1
14. M. Mobli, J. C. Hoch, *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *83*, 21–41. DOI: 10.1016/j.pnmrs.2014.09.002.
15. J.-M. Nuzillard, *Chin. J. Chem.* **2003**, *21*, 1263–1267. DOI: 10.1002/cjoc.20030211006
16. M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin, *J. Chem. Inf. Model.* **2006**, *46*, 1643–1656, DOI: 10.1021/ci050469j
17. J.-L. Faulon, *J. Chem. Inf. Comput. Sci.* **1997**, *36*, 731–740. DOI: 10.1021/ci950179a
18. J.-M. Nuzillard, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 723–724. DOI: 10.1021/ci00020a004
19. K. Jayaseelan, P. Moreno, A. Truszkowski, P. Ertl, C. Steinbeck, *BMC Bioinf.* **2012**, *13*, 106. DOI: 10.1186/1471-2105-13-106
20. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev, *J. Cheminf.* **2013**, *5*, 7. DOI: 10.1186/1758-2946-5-7
21. D. Yegor, D. Smurnyy, M. E. Elyashberg, K. A. Blinov, B. A. Lefebvre, G. E. Martin, A. J. Williams, *Tetrahedron* **2005**, *61*, 9980–9998. DOI: 10.1016/j.tet.2005.08.022
22. P. Trigo-Mouriño, A. Navarro-Vázquez, J. Ying, R. R. Gil, A. Bax, *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 7576–7580. DOI:10.1002/anie.201101739
23. N. Grimblat, A. M. Sarotti, *Chem. Eur. J.* **2016**, *22*, 12246–12261. DOI : 10.1002/chem.201601150
24. B. Plainchont, J.-M. Nuzillard, G. V. Rodrigues, M. J. Ferreira, M. T. Scotti, V. de P. Emerenciano, *Nat. Prod. Commun.* **2010**, *5*, 763–770.
25. B. Plainchont, V. de P. Emerenciano, J.-M. Nuzillard, *Magn. Reson. Chem.* **2013**, *51*, 447–453. DOI: 10.1002/mrc.3965
26. C. Steinbeck, S. Kuhn, *Phytochemistry* **2004**, *65*, 2711–2717. DOI: 10.1016/j.phytochem.2004.08.027

Figure captions.

Fig. 1. (top left) A COSY correlation between the ^1H resonances of H_i and H_j through a 3J coupling constant establishes a bond between the heavy atoms X_i and X_j , the latter being respectively identified by the HSQC correlations of their resonances with those of H_i and H_j . (top right) An HMBC correlation of the resonances of nuclei H_i and X_j and the HSQC between those of H_i and X_i establishes a proximity relationship between nuclei X_i and X_j : X_i and X_j are either bound together (dotted line) or bound to another atom in the molecule, indicated here as a question mark. (middle left) A H2BC correlation between the resonances of X_i and H_j indicates a $\text{X}_i - \text{X}_j$ bond if H_i and H_j share a 3J coupling; the H2BC spectrum may be seen as an HMBC spectrum in which only 2J X-H couplings intervene. (middle right) The $\text{H}_i - \text{X}_j$ 1,1-ADEQUATE correlation arises from the coupling between the adjacent X nuclei. (bottom) The INADEQUATE spectrum directly indicates the existence of a bond between the X nuclei, whatever their protonation state.

Fig. 2. Annotation procedure. (top) The ^{13}C NMR resonances are arbitrarily numbered in the order of decreasing chemical shifts. The resonance index is also an atom and a nucleus index. (bottom) The ^1H resonance numbering is determined by the ^{13}C resonance numbering and the HSQC correlation spectrum.

Fig. 3. The COSY and HMBC correlations are referenced according to the pairs of resonance indexes they concern. 2J COSY correlations do not have to be marked.

Fig. 4. The *lsd* commands for the resolution of the camphor problem. They are presented here in a tabular arrangement but are written in a single column in the *camphor.lsd* file.

Fig. 5. Terminal window, after the resolution of the camphor problem. Skeletal atoms is another name for heavy atoms. Only one solution is obtained because the generation of assignment isomers was disabled.

Fig. 6. Graphical rendering of the first solution of the camphor problem.

Fig. 7. Manually redrawn structures for the two solutions of the camphor problem. They only differ in a permutation of positions 2 and 3.

Fig. 8. (top left) Substructure element, made of two separated pieces, for the selection of solution structures that contain the non-rearranged monoterpene structural pattern. (bottom left) The *lsd* commands that code for the search of the non-rearranged monoterpene pattern. All atoms were allowed to be functionalized by transformation of single bonds into double bonds and by hydrogen abstraction. (right) A possible way of matching the structure of camphor with the proposed substructure.

Fig. 9. Global workflow of the *pylsd* algorithm. If a structure elucidation problem involves a fuzzy molecular formula (optionally) and atoms with fuzzy status (optionally), the *pylsd* algorithm generates n input data files that *lsd* can solve. The m solutions of these problems may be then ranked (optionally) in the order of decreasing likelihood.

Data Files

Please follow <http://eos.univ-reims.fr/LSD/MRC2017a/> to access the NMR data and the camphor.lsd file.

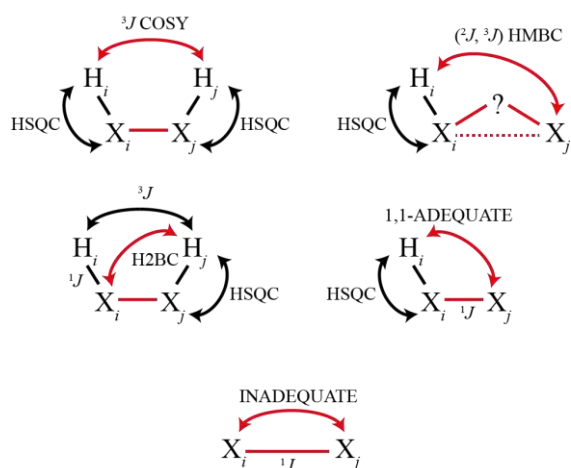


Fig. 1.

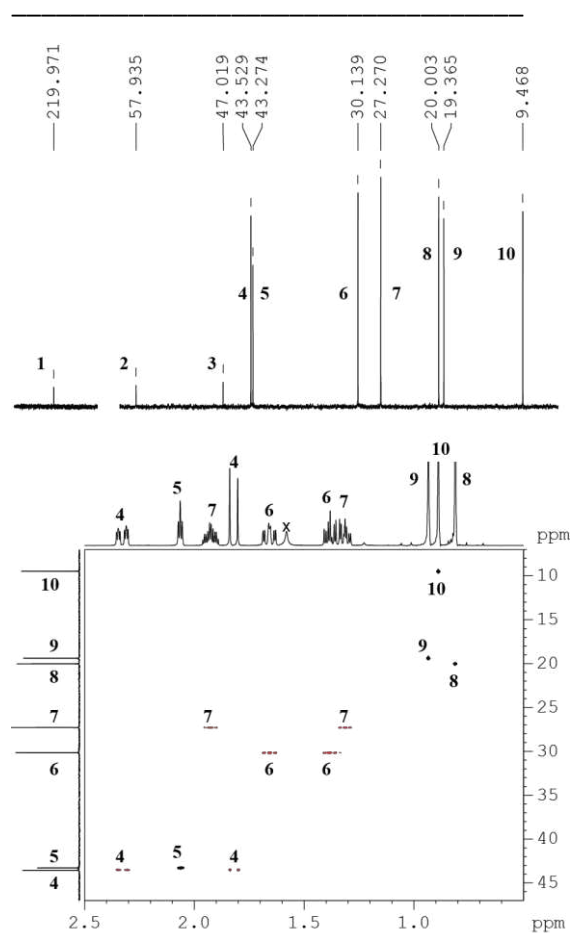


Fig. 2.

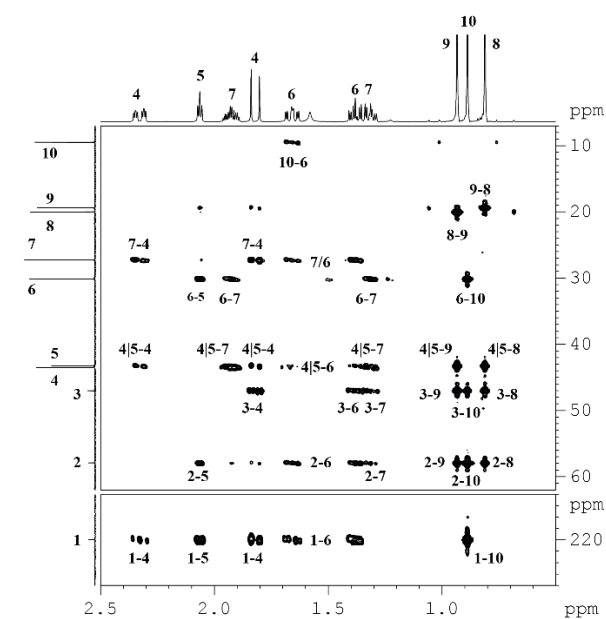
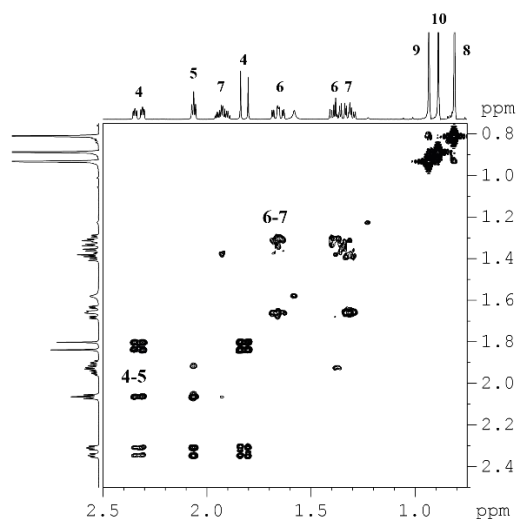


Fig.3.

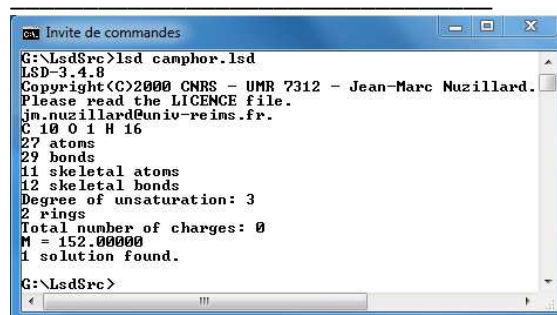


Fig. 5.

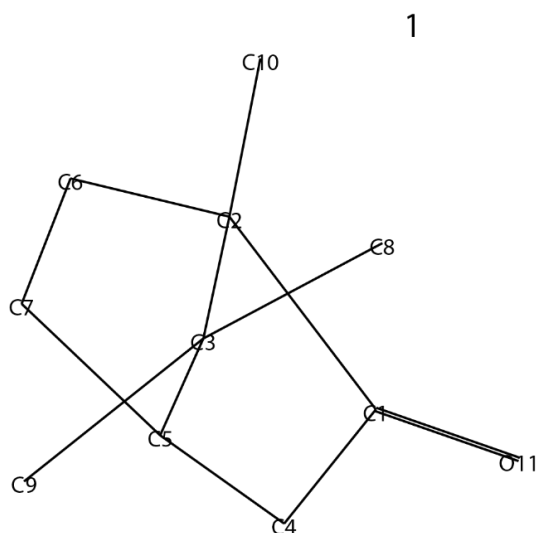


Fig. 6

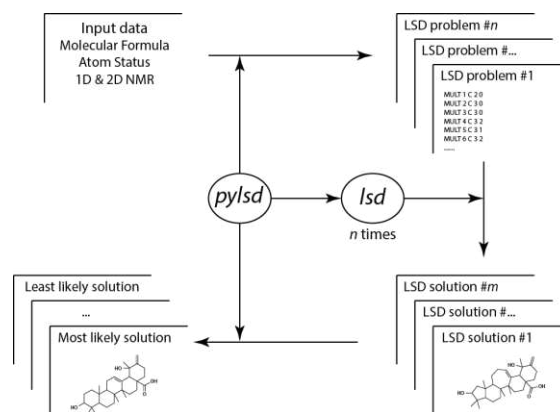


Fig. 9

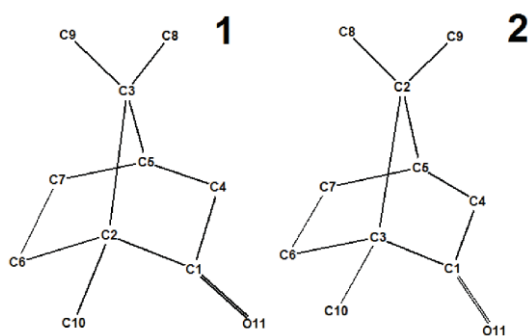


Fig. 7.

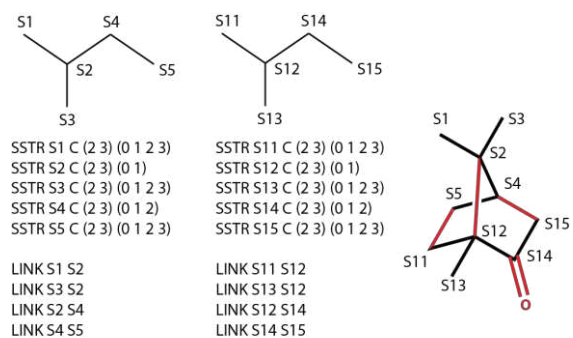


Fig. 8

MULT 1 C 2 0	SHIX 1 219.97	BOND 1 11	SHIH 4 2.328	HMBC 1 4	HMBC 3 10
MULT 2 C 3 0	SHIX 2 57.93		SHIH 4 1.819	HMBC 1 5	HMBC (4 5) 4
MULT 3 C 3 0	SHIX 3 47.02	HSQC 4 4	SHIH 5 2.062	HMBC 1 6	HMBC (4 5) 7
MULT 4 C 3 2	SHIX 4 43.53	HSQC 5 5	SHIH 6 1.657	HMBC 1 10	HMBC (4 5) 6
MULT 5 C 3 1	SHIX 5 43.27	HSQC 6 6	SHIH 6 1.381	HMBC 2 5	HMBC (4 5) 8
MULT 6 C 3 2	SHIX 6 30.14	HSQC 7 7	SHIH 7 1.930	HMBC 2 6	HMBC (4 5) 9
MULT 7 C 3 2	SHIX 7 27.27	HSQC 8 8	SHIH 7 1.314	HMBC 2 7	HMBC 6 5
MULT 8 C 3 3	SHIX 8 20.00	HSQC 9 9	SHIH 8 0.811	HMBC 2 8	HMBC 6 7
MULT 9 C 3 3	SHIX 9 19.37	HSQC 10 10	SHIH 9 0.934	HMBC 2 9	HMBC 6 10
MULT 10 C 3 3	SHIX 10 9.47		SHIH 10 0.888	HMBC 2 10	HMBC 7 4
MULT 11 O 2 0		COSY 4 5		HMBC 3 4	HMBC 7 6
		COSY 6 7	LIST L1 8 9 10	HMBC 3 6	HMBC 8 9
			QUAT L2	HMBC 3 7	HMBC 9 8
			PROP L1 1 L2	HMBC 3 8	HMBC 10 6
				HMBC 3 9	

Fig. 4.