Computer-aided Dereplication and Structure Elucidation of Natural Products at the University of Reims

Ali Bakiri, Bertrand Plainchont, Vicente de Paulo Emerenciano, Romain Reynaud, Jane Hubert, Jean-Hugues Renault, Jean-Marc Nuzillard

molecular informatics

DOI: 10.1002/minf.200((full DOI will be filled in by the editorial staff))

Computer-aided dereplication and structure elucidation of natural products at the University of Reims.

Ali Bakiri,[a] Bertrand Plainchont,[a] Vicente de Paulo Emerenciano,[b] Romain Reynaud[c] Jane Hubert,[a] Jean-Hugues Renault,[a] and Jean-Marc Nuzillard*[a]

Abstract: Natural product chemistry began in Reims, France, in a pharmacognosy research laboratory whose main emphasis was the isolation and identification of bioactive molecules, following the guidelines of chemotaxonomy. The structure elucidation of new compounds of steadily increasing complexity favored the emergence of methodological work in nuclear magnetic resonance. As a result, our group was the first to report the use of proton-detected heteronuclear chemical shift correlation spectra for the computer-assisted structure elucidation of small organic molecules driven by atom proximity relationships and without relying on databases. The early detection of known compounds appeared as a necessity in order to deal more efficiently with complex plant extracts. This goal was reached by an original combination of mixture fractionation by centrifugal partition chromatography, analysis by ^{13}C NMR, digital data reduction and alignment, hierarchical data clustering, and computer database search.

Keywords: Artificial Intelligence, Computer-Assisted Structure Elucidation, Dereplication, Liquid Chromatography, NMR Spectroscopy

1 Introduction

The chemistry research teams at the University of Reims have developed analytical skills for the structure identification of small synthetic and natural organic molecules during the last four decades. Natural product chemistry has evolved from the study of indole alkaloid-containing plants,[1] following the discovery of the bis-indole anti-cancer drugs such as vinblastine. In the '80s, the isolation, purification and structure analysis of saponins encouraged the acquisition of higher-performance NMR instrumentation as well as the implementation of the then most recent NMR spectroscopy methods.[2,3] At the same time, the investigation of other natural product classes such as terpenes and phenolic compounds took a growing importance in our laboratory.

The general approach to the structure elucidation of plant secondary metabolites involved the determination of a gross molecular formula using mass spectrometry (MS) and the atom assembly under constraint of NMR, UV, IR spectral features as well as MS fragmentation. Working in compound series such as indolomonoterpenic alkaloids offered the availability of a vast corpus of physico-chemical data, including the coloration of compounds upon spraying of specific reagents on Thin Layer Chromatography (TLC) plates. The quick identification of the compounds that were already reported in the literature was carried out with great reliability by the comparison with archived data of TLC chromatographic R_f values, coloration on TLC plates and low-resolution electron impact mass spectra. A ^{1}H NMR spectrum could help to secure the certainty of an identification, especially when alternative structures existed due to position isomerism. The structure of the few isolated compounds that were not readily identified by these means was determined by *de novo* structure determination.

Structure elucidation by NMR underwent a radical change in the 1980's with the availability of ^{1}H-detected 2D heteronuclear (meaning ^{1}H-^{13}C or ^{1}H-^{15}N) chemical shift correlation spectra, namely HMQC (and later HSQC) for directly bound nuclei pairs and HMBC for remotely bound nuclei pairs.[4] Pulsed static field gradient enhancement of these techniques delivered high-quality 2D NMR spectra that were well suited to the *de novo* computer-assisted structure elucidation.[5]

The use of computers to deal with molecular structures and NMR spectra was initiated at the University of Reims in 1988 by the corresponding author of this article. As an occasional teacher of computer programming at the Reims Institute of Technology, he attended an introductory course on Artificial Intelligence (AI) and on the Prolog (Programmation en Logique) AI language. The newly acquired knowledge was quickly put into action for the design of CASA, an automated system for ^{13}C NMR spectra assignment whose goal was to pair spectral resonances with the atoms of a known or postulated structure.[6]

The conceptual gap between spectral assignment and structure generation was then quickly bridged, still using the Prolog AI language as programming tool, and led to the LSD (Logic for Structure Determination) software.[7]

[a] *Institut de Chimie Moléculaire de Reims, CNRS CPCBAI, Bât. 18, Moulin de la Housse, BP 1039, 51687 REIMS Cedex 2, France. *jm.nuzillard@univ-reims.fr: , phone: 33(0)326918210*

[b] *Instituto de Química, Universidade de São Paulo Av. Prof. Lineu Preste 748, Sao Paulo, CEP 05513-970, Brasil.*

[c] *Soliance-Givaudan Route de Bazancourt, 51110 POMACLE, France*

Supporting Information for this article is available on the WWW under www.molinf.com

The computer code was later rewritten in the C language for speed and portability reasons and is available as free software (www.univ-reims.fr/LSD).

CASA and LSD were designed for the analysis of pure compounds. Purification is a mandatory step for organic chemists. It is nearly impossible to publish the structure of a new compound without its isolation at the highest possible purity level. The thorough purification of the compounds of an extract requires an important investment in equipment and human time. Therefore, the early identification of known compounds, a task also known as dereplication, when carried out in mixtures, is a way to save time and to concentrate on the structure elucidation of new compounds. Recent developments in this field led to the creation of the workflow called CARAMEL (after CARActérisation de MÉLanges, in French, meaning mixture characterization), which combines spectroscopic, data classification, and databasing tools.[8]

2 Dereplication

Natural product (NP) chemists often feel the unpleasant feeling of rediscovering compounds that have already been discovered. However, the study of the chemical content of an extract (from plant, bacteria, mycete, insect...) generally starts with the exact determination of the species of the studied organism and with bibliographic work on what has been already published about the same species or the same genus or even the same family, depending on the number of related species. The exploitation of such chemical studies allows one to define a list of compounds that could be present in the extract under investigation because a taxonomic proximity is presumably correlated with a chemical proximity. In a formal way, bibliography established a relationship between the species space and the molecular structure space (Figure 1). A given species is related to a set of molecular structures (pharmacognosy) and a given molecular structure is related to a set of species (reverse pharmacognosy).[9]

The commonly adopted workflow in NP chemistry involves the isolation of the compounds from organism extracts. Extraction might be more or less selective and/or carried out on particular parts of organisms (such as barks or flowers in plants), when possible, in order to simplify compound isolation and possibly to relate organs and their function. A compound is generally purified by a NP chemist in order to study its biological activity. In the same way that a plant specimen needs to be related to the name of a living species, an isolated compound needs to be related to a molecular structure in order to give value to a biological test.

Dereplication, in one of the acceptations of the term, consists in finding the molecular structure of an isolated compound, pure or in mixture, among those already reported. With this meaning, dereplication is the way to assign a chemical structure to a compound using the knowledge that accumulated during decades of NP research. The main difficulty here is to declare that a sample that has just been isolated in the lab next door is made of the same molecules as the one isolated thirty years ago in a lab on another continent. Physico-chemical analysis is the way to do it, by connecting the space of physical samples and that of analytical data. A relationship between the latter and the space of molecular structures is needed in order to connect isolated compounds with their molecular structure (Figure 1).

Two samples with identical analytical data are presumed to be identical, assuming that identical compounds present identical analytical data. The last part of this assertion is subjected to the repeatability of the analytical process and on the identity of the samples, while the first part really depends on the concept of analytical data. Leaving this question aside for the moment, dereplication boils down to acquisition of analytical data and to comparison between actual and published analytical data, a task that seems easy and instantaneously feasible across time and continents in the era of the Internet.[10]

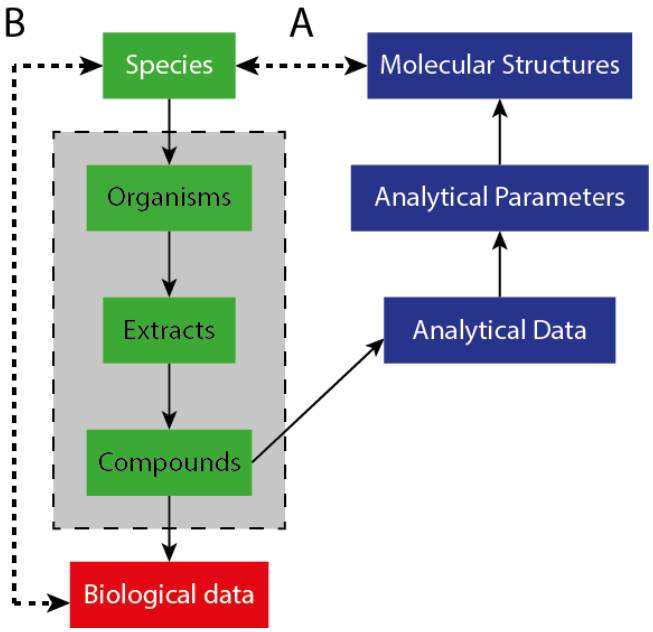

Figure 1. Relationships between different spaces involved in natural product chemistry. The three boxes enclosed by the dotted contour line correspond to physical objects and the others to intellectual objects. Relationship A connects species and molecule structures and is reported in publications whose titles are like “New sesquiterpene lactones from the flowers of *Genus-name species-name*”. Relationship B connects species and biological data as reported in publications with titles like “Biological Activities of Aerial Parts Extracts of *Genus-name species-name*”. Many NP publications belong to the A and to the B type simultaneously, reporting structure and activity of plant parts.

Chemistry journals and publishers require authors to deposit analytical data as Supplementary Information (SI) documents attached to articles, while the visible part of analytical studies most often reduces to lists of analytical parameters. For example, a UV-visible spectrum, acquired through the measurement of hundreds of light absorption values (the analytical data) ends up in a publication as a list of a few maximum absorbance positions (the analytical parameters). Therefore, effective dereplication is carried out by comparison between actual and published analytical parameters. Comparison between spectra is sometimes possible but the plots in most SI documents are of too poor quality to be really useful, a situation that could change with the development

of infrastructures for public storage of spectral data and spectral parameters.

Dereplication clearly depends on analytical methods. Fusion temperature might be useful to declare that two samples are made of different molecules (not considering sample purity and the role of molecular packing in crystals), but is hardly convincing as a validation tool for structure identity. Intuitively, combining analytical methods that rely on unrelated physico-chemical processes increases the probability of being sure that two substances are identical. From a practical point of view, it is not possible for a laboratory to invoke all existing methods, and a single method is often selected.

Mass spectrometry (MS) and Nuclear Magnetic Resonance (NMR) are the two main analytical methods that come into play for dereplication purposes, either in NP chemistry or in metabolomics studies.[11] The pros and cons of each technique are known and, to summarize, NMR is poorly sensitive but is quantitative, reproducible and reveals the finest structural details, while MS is utterly sensitive, but hardly reproducible and gives limited access to structural features. No single technique presently combines the advantages of NMR and MS, even though each of them is still the object of steady improvement efforts in order to reach the advantages of the other.

Dereplication of NPs is commonly carried out all over the world using a combination of High-Performance Liquid Chromatography (HPLC, briefly LC) and MS, thus allowing in one LC-MS sequence the isolation of tiny amounts of pure compounds (nanograms, if not picograms) and measurement of their mass spectra. In particular research fields such as marine products and micro-organism chemistry, specialized databases exist and make it possible to propose molecular structures from mass spectral data. The concept of molecular networking in MS tends to increase its analytical power toward structure elucidation, beyond simple structure recognition.

The emergence of our original approach toward dereplication in Reims is the result of many factors. i) A long tradition of NP studies exists there, in which NMR played a decisive role and was the subject of methodological research in the fields of data acquisition and processing (JMN). ii) LC-MS have become available and open to NP research only recently. iii) Preparative separation methods based on centrifugal partition chromatography (CPC) has undergone tremendous development during the last 25 years (JHR). CPC offers the means to isolate pure compounds or simplified fractions in amounts that are compatible with analysis by ^{13}C NMR spectroscopy and biological activity research (milligrams to tens of grams on a laboratory scale CPC system). iv) An assistant professor (JH) with a background in metabolomics and statistical data analysis was recruited about six years ago. v) Our university and an international company found good reasons to support our research efforts (RR). vi) People there decided to work together.

The CARAMEL dereplication workflow deals with complex mixtures such as plant or microbial extracts (but not limited to these) and starts with extract fractionation by CPC. About 20 simplified fractions are obtained and analysed by ^{13}C NMR (discussion hereafter). The spectra, recorded over a 240 ppm wide chemical shift window, are submitted to an automatic peak search and therefore reduced to a list of peak coordinates. The peak intensity values are stored in 0.2 ppm wide, initially empty, chemical shift intervals called bins, so that each one corresponds to a chemical shift value. The bins that are empty over the whole series of spectra are left aside. The content of the bins is stored in a table: bin index in rows, fraction index in columns, peak intensities for values in table. The table is then converted into a human-readable text format.

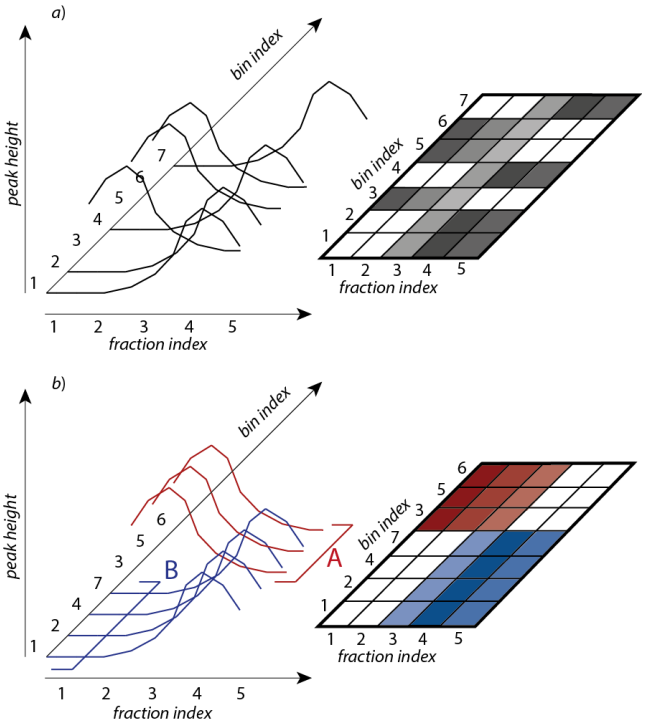

Figure 2. Principle of data handling in the CARAMEL dereplication workflow. The ^{13}C NMR spectra of the fractions contain seven peaks. The evolution of intensity for the different peaks characterized by their bin index (line index in data table, see text) is reported as a function of retention time or fraction index (column index in data table). a) original data, b) rearranged data that permit the appearance of the grouping of the resonances arising from two compounds with different chromatographic behaviours.

Considering two molecules A and B with non-overlapping peaks and for which the CPC emergence profiles are different, for example with A less retained in the CPC column than B. Considering all the bins of compounds A and B, the content of some of them will have non zero values only for low-index fractions, others for high-index fractions only. The former will correspond to compound A and the latter to compound B. In this simple trivial context, the bins that correspond to compounds A and B are easily differentiated (see Figure 2). A way to automate this process in real-life situations is to group the lines of the table according to similarity. The classification of rows by hierarchical clustering achieves this goal and is implemented using the freely available PermutMatrix software.[12] The textual table of peak intensities is used as software input. The resulting rearranged table is organized into contiguous blocks (or clusters) of bins whose associated chemical shift values correspond to the different compounds that constitute the initial complex mixture.

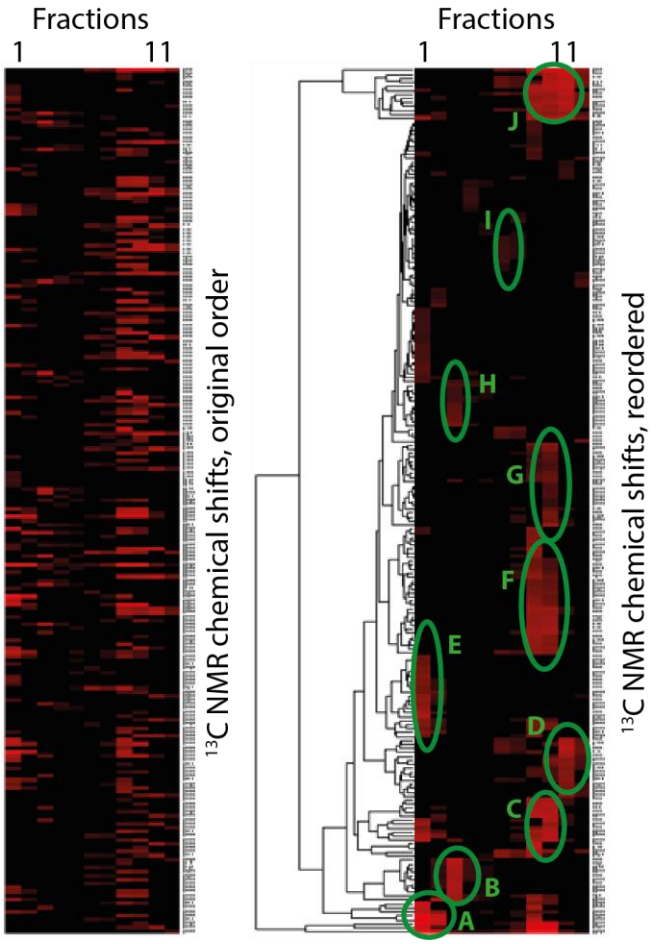

Figure 3. Dereplication of an extract of *Tephrosia purpurea*, as reported in reference 14a. The PermutMatrix software represents data as heat maps, the brightest colors corresponding to the highest spectral peak intensities. (left) Data before hierarchical clustering. (right) Result of peak clustering according to similarity of fractionation profiles. Each cluster (green ellipses) corresponds to a chemical shift list that serves as key for database search. The structure elucidation of compounds D, H and J, unknown to the database, was assisted by the LSD software.

The association between chemical shift lists and molecular structures is achieved by database search. We filled our own database (approximatively 2100 entries at manuscript submission time) with molecular structures collected through the bibliographic work carried out at the beginning of each new extract analysis project. Initially, molecular structures were associated in our database with published experimental chemical shift values. It turned out then that it was equally efficient to store the values calculated by the ACD ^{13}C NMR predictor.[13]

The ACD databasing system also allows one to retrieve compounds from an incomplete list of chemical shift values. The discrepancy between ACD-calculated and experimental values and the search for a compound in an incomplete list of values results in a series of candidate structures that has to be checked manually. Checking is carried out by inspection of 1D ^{1}H, ^{1}H-^{13}C, 2D ^{1}H-^{1}H COSY, 2D ^{1}H-^{13}C HSQC and HMBC spectra of the fraction(s) in which the ^{13}C NMR peaks are the most intense and provides a molecular structure that is the most likely. Even though checking is still a mandatory step, the whole process greatly speeds up the search for the structure of known molecules. Additionally, the initial fractionation step delivers samples that may be involved in biological activity tests and further purified when a fraction shows some interesting activity.

The present version of the CARAMEL workflow is based on 1D ^{13}C NMR spectroscopy. This choice results from practical considerations. i) Even extremely hydrogen-poor compounds (such as ellagic acid and its derivatives) contain carbon atoms. Apart from ^{13}C NMR spectroscopy, the only practical way to obtain information on quaternary carbons is the recording of 2D HMBC spectra. ii) The ^{13}C resonances appear in our standard ^{1}H decoupled spectra as sharp singlets (typically 1 Hz width at half height) within a 36 kHz spectral window (240 ppm from a base frequency of 150 MHz), thus minimizing the probability of peak superimposition. iii) Spectral parameter extraction is easily carried out by automatic peak peaking using the spectrometer software and the result is easily exported as a text file. iv) Apart from (rare) symmetrical molecular structures and (rare) coincidences, there is a one-to-one relationship between carbon positions in molecular structures and chemical shift values revealed as peak positions. v) We compensate the commonly mentioned lack of sensitivity of ^{13}C NMR by the use of a high static magnetic field (14.1 T, resonance of ^{1}H nuclei at 600 MHz) and of a Helium-cooled cryo-probe. Spectrum acquisition time is kept under these conditions around one hour for each fraction, during which 1024 transients are co-added, with a 3.6 s repetition time. vi) Prediction of ^{13}C NMR chemical shifts is more reliable than the one of ^{1}H, for which variations of experimental conditions may induce chemical shift variations greater than the expected uncertainty on predicted values.

At this stage, one can see that the CARAMEL workflow has some obvious limitations. Two compounds with identical emergence profiles are of course not distinguishable, but this situation is not likely. Compounds in small amounts for which spectral peaks do not significantly emerge from noise are not visible, so CARAMEL misses the minor components of a mixture. Compounds for which ^{13}C NMR spectra are partly superimposed (collected in the same bin) may be left classified together but not with the other non-superimposed peaks, resulting in incomplete chemical shift lists for the concerned compounds. Figure 3 illustrates the CARAMEL procedure on a published example. [14a]

Possible improvements of the analytical part of the CARAMEL workflow may include the spectral edition of ^{13}C resonances according to the parity of the number of attached ^{1}H nuclei (DEPT spectra), ^{1}H pure shift spectra, 2D HSQC (pure shift or not) and HMBC spectra. Numerous applications of the CARAMEL workflow have already been reported.[14] A manuscript reporting the direct search for the compounds stored in our database within a crude extract, without fractionation, and on the sole basis of ^{13}C NMR data, is currently under review.

3 Structure elucidation by the LSD software

First, and as a rule of thumb: "dereplication precedes structure elucidation". Structure elucidation of new or presumably new compounds is at best performed on pure compounds because

this is the best way to avoid spectral data misinterpretation due to confusion between data arising from different compounds. The following lines will assume compound purity.

Computer-assisted structure elucidation (CASE) has been one of the first playgrounds of artificial intelligence. CASE concepts, software and their application have been reviewed in articles, book chapters and a recent book.[15]

Structure elucidation is the process of establishing bonds between atoms whose number and individual chemical elements are obtained as a gross formula by high resolution mass spectrometry. Hydrogen atoms play a particular role in molecular structures because they are all the same and always share a single bond with any other atom. As a consequence, chemical structure drawings of organic molecules omit them when they are bound to carbon atoms unless some particular stereochemical feature needs to be displayed. The LSD software (in short: LSD) differentiates between light atoms (all the hydrogens) and heavy atoms (all the others) by describing only the latter by an explicit atom status.

An atom status consists of an atom index, a chemical element symbol, a multiplicity, a hybridization state, and an electric charge. Carbon indexes are, by tradition, assigned to atoms in the decreasing order of the corresponding ^{13}C chemical shift values. Atoms other than hydrogens and carbons, referred to as heteroatoms, receive arbitrarily given atom indexes. All parts of the status of all atoms must be completely defined at the beginning of the resolution of a problem by LSD. The status defines for each heavy atom the number of directly attached heavy atoms and this number must be accurately known in order to efficiently carry out structure generation. A chemical element symbol may include a valence suffix for non-usual valence values such as S4 for tetravalent sulfur, while S is sufficient for divalent sulfur. The multiplicity of a heavy atom is the number of directly attached hydrogen atoms; this is easily determined for carbons and for heteroatoms when only a single kind of the latter is present. Hybridization states are only sp^3, sp^2 and sp. For commodity, the sulfur atom of a sulfoxide is declared as sp^2, even though this does not correspond to the actual bond electronic structure. This example touches the border of the usual "atom and bond" representation of organic molecules that excludes bonds over more than two atoms. Hybridization state of carbons may be unambiguously deduced as being sp^2 or sp^3 from ^{13}C chemical shift values when they fall in high-value or low-value ranges.

In some cases, different atom status sets are compatible with the available data and more than one data set must be prepared for LSD in order to solve a single problem. This inconvenience is eliminated in pyLSD, but computer system compatibility issues still must be solved in order to make its use of general interest. PyLSD is a driver software for LSD written in the Python programming language that expands all alternatives on individual atom status and retains the status sets compatible with the molecular formula. PyLSD also allows the user to introduce some flexibility in the gross molecular formula of the unknown molecule.[16]

Ideally, ^{13}C-^{13}C INADEQUATE spectra should be able to delineate the carbon skeleton of an organic molecule, thus simplifying a big part of structure elucidation problems.[17] This approach was considered in the first Computer-Aided Structure Elucidation programs in which 2D NMR was taken into account.[18] The insensitivity of INADEQUATE, that requires neighbouring pairs of ^{13}C nuclei while the natural abundance of the nuclei is only about 1%, makes it rarely applicable, even though this is the only way to directly prove by NMR that two quaternary carbon atoms are bound together. LSD accepts INADEQUATE data as predefined bonds between atoms. All other forms of atom connectivity involve hydrogen atoms and their ^{1}H NMR resonances.

Hydrogen atom indexing is preferentially related to heavy atom indexing, according to the HSQC spectrum that correlates the chemical shifts of ^{1}H and ^{13}C (or ^{15}N) nuclei that are directly bound, thus defining with a single index what chemists call a chemical position (C-1 is bound to H-1 at position 1). Inequivalent hydrogens in methylene groups are therefore identically indexed. The COSY and HMBC spectra provide proximity relationships that LSD exploits in order to set bonds between atoms. In the simplest vision, a ^{1}H-^{1}H COSY spectrum reveals magnetic couplings between ^{1}H nuclei through 2 or 3 bonds. Two-bond (or ^{2}J) couplings only occur within methylene groups with inequivalent hydrogen atoms. The ^{2}J COSY couplings are quickly identified because they happen between identically indexed atoms; they do not bring any information about heavy atom connectivity. A ^{3}J H-a/H-b coupling arises through the H-a/C-a/C-b/H-b coupling path and therefore proves the existence of the C-a/C-b bond that is translated by LSD into the setting of a firmly established, predefined bond (Figure 4).

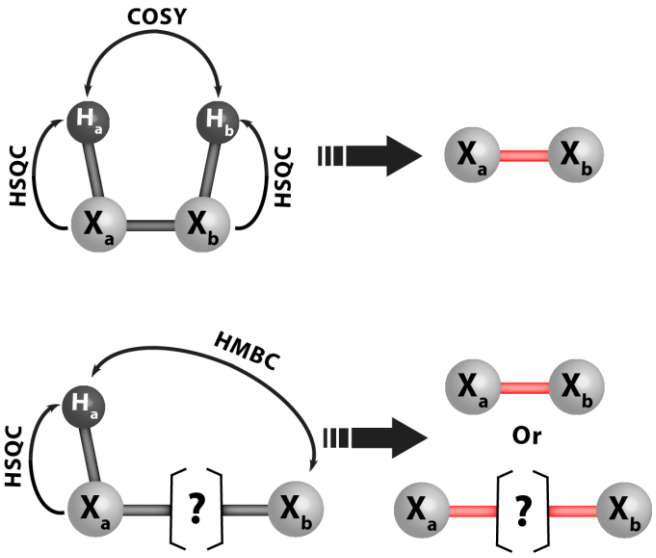

Figure 4. Deduction of bonds and of alternative proximity relationships from 2D HSQC, COSY, and HMBC NMR data. In the case of a ^{3}J (H-X) coupling, the intermediate atom noted as a question mark may be any atom of the studied molecule.

A HMBC spectrum reveals ^{2}J or ^{3}J couplings between ^{1}H and ^{13}C (or ^{15}N) nuclei, thus proving that the corresponding heavy atoms are either directly bound or two bonds away, as if a ^{1}J or ^{2}J ^{13}C-^{13}C INADEQUATE spectrum were recorded.[7] There is no general and reliable method to decide upon which alternative is correct.[19] The intermediate heavy atom, see Figure 4, in

the ^3J hypothesis can be any atom in the molecule, either a carbon or a heteroatom.

Low resolution along the ^{13}C chemical shift axis adds a supplementary indeterminacy level in the interpretation of an HMBC spectral peak because more than a single carbon atom may be responsible for it. LSD offers the possibility of handling such alternatives on close resonance groups. Resolution enhancement in HMBC may be obtained without increasing recording times by means of strategies such as folding or non-uniform sampling. [20] Groups of poorly distinguishable ^1H resonances may also be exploited by LSD.

Writing datasets for LSD implies that 2D spectra may adequately be reduced, a problem that remains difficult to automate. Results, as chemical shift pair lists, still need to be edited by hand in order to avoid the introduction of contradictory input data in LSD input files that would lead to empty solution lists. Efforts for effective reduction of 2D NMR data have been carried out and were integrated into the CMC-se module of the commercial Bruker TopSpin software. CMC-se integrates a company-developed structure elucidation algorithm as well as that of LSD.[21]

The length of the coupling paths revealed by COSY and HMBC spectra is not limited to 2 and 3, even though the initial algorithm was designed as if it were the case. Longer length COSY and HMBC data caused LSD to fail. They originate from coupling constants of low magnitude and give rise to low intensity peaks. A good practice consists in considering first only the most intense 2D peaks. LSD was later modified so that a limited user-defined number of long range couplings may be handled. Additionally, each individual HMBC may be associated with a minimum and a maximum coupling path length.

LSD was designed to mimic the way chemists build structures from spectra. Chemical shift values and ^1H NMR line splitting patterns are related to immediate environments of atoms and therefore LSD accepts constraints, called "atom properties", on the nature of the neighbouring atoms a given heavy atom may have. Atom properties are taken into account by the LSD algorithm during structure generation. Of course, invalid constraints may result in incorrect structures or in no structure at all. User-provided substructure elements may also be given and are used for the solution filtering step that takes place after structure generation.

Natural products may be categorized in classes that correspond to various skeletons, another word for substructures. Their identification was one of the motivations for starting a collection that is included in LSD. The present collection was drawn from the SISTEMAT knowledge base and was established for mono-, sesqui-, and diterpenes. SISTEMAT contains databases (50000 compounds, overall) for various natural product classes and software dedicated to their management and exploitation. A SISTEMAT database contains records for natural products as reported in literature. A record connects the molecular structure of a compound with its spectroscopic data, including those of ^{13}C NMR, and with botanical data on the plant from which it originates. In this way a species may be immediately related to the probability of the presence of the skeletons of the compounds that were isolated in plants of this species.[22] SISTEMAT may be used for dereplication from ^{13}C NMR data of mixture.[23] It can also assist elucidation by proposing fragments of known molecules as possible fragments of an unknown molecule.[24] Structure generation by LSD has been successfully guided by SISTEMAT proposals for natural compounds from various classes.[25]

The structure generation process in LSD starts (phase 1) with a set of well-defined status atoms connected by explicitly user-supplied bonds and by those arising for ^3J COSY data (^2J COSY data is automatically discarded). Atom proximity information from HMBC is then selected and exploited in order to build bonds. The selection-exploitation process (phase 2) is repeated until all the proximity information is processed. The order in which proximity data are taken into account greatly impacts the computation times. A heuristic criterion was implemented, stating that the heavy atoms for which the number of missing bonds is the lowest must have their proximity data exploited first. Atoms for which bonds are missing at the end of phase 2 are systematically paired to produce a meaningful molecular structure during phase 3. Phase 4 includes a set of acceptance tests comprising the determination of the order of bonds, the compliance of structures to Bredt's rule,[26] and the validation of substructure constraints, if any.

The exploitation of proximity data in step 2 involves the selection of a hypothesis in the case of close resonance groups, the choice between a one-bond and a two-bond distance between heavy atoms, and, in the last case, the choice of the index of the intermediate atom. The solution search process has therefore been organized as a depth-first tree exploration algorithm (Figure 5).[27]

The resulting solutions are stored in a text file, from which conversion to various standard formats is carried out by a program named outlsd. The depictions in 2D SD files are most often very crude and can be manually improved using the m_edit utility program or any other multi-structure 2D structure editor, or regenerated from scratch using any available structure drawing generator and depictors, like those from the RDKit,[28] OpenBabel,[29] and CDK[30] chemoinformatic toolkits.

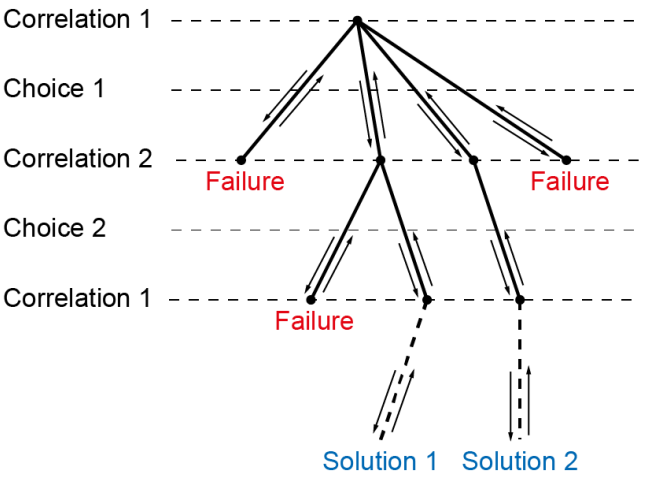

Figure 5. Solution search by depth-first tree exploration. The analysis of a correlation potentially opens a range of interpretation choices. An invalid interpretation may lead to the impossibility of interpreting a still-unexploited correlation (failure).

The final goal of the structure elucidation process is to find the structure of a molecule. The generation of more than one solution, all compatible with input data, sets the problem of solution ranking. Ranking is provided in pyLSD by a score calculated as a distance between the set of the experimental ^{13}C NMR chemical shifts and those predicted for each solution structure. The reliability of this approach naturally depends on the quality of the prediction algorithm. The only standalone predictor we had at hand was the one provided by the nmrshiftdb2 free software project.[31]

The LSD software played an important role in the structure of natural and sometimes synthetic compounds, as exemplified by published articles (see Figure 6).[32]

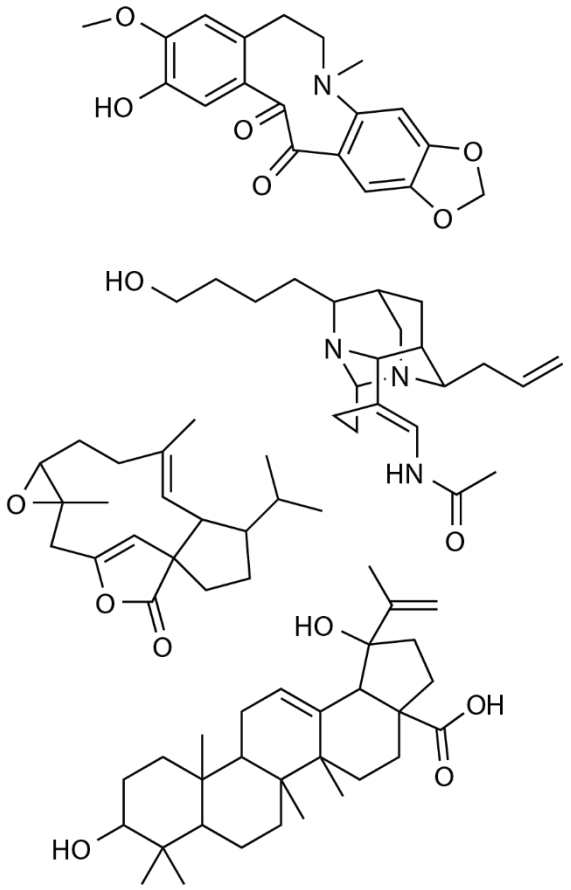

Figure 6. Examples of molecular structures, whose determination was assisted by the LSD software.

Future developments of LSD include a better separation between structure generation and validation/filtering/ranking tasks. The structure generator should be modified so that atoms with alternative status are handled during resolution and not before it, as presently achieved in pyLSD for the combinatorial enumeration of status indetermination-free LSD problems. The application of substructure constraints should also be integrated in the bond-building processes, if desired. Molecules that contain parts with symmetry or even symmetrical molecules are not presently easily handled by LSD and solving this problem would enlarge the capabilities of the software. Tasks such as substructure search and depiction generation should be carried out by thoroughly tested free chemoinformatic toolkits. The efficiency of dereplication and of automatically generated structure ranking should improve through the public availability of an increasing amount of experimental chemical shift data, in a format that may facilitate their transfer from laboratory notebooks to free public databases.[10]

Acknowledgements

Financial support by CNRS, Conseil Regional Champagne Ardenne, Conseil Général de la Marne, French Ministry of Higher Education and Research (MESR), FAPESP, CNPq, the Soliance-Givaudan Company and EU-programme FEDER to the PlAneT CPER project is gratefully acknowledged. We thank Dr. Mark Andrew Gannon for his stylistic and grammatical improvement of the manuscript.

References

[1] J. Sapi, G. Massiot, in *Monoterpenoid Indole Alkaloids (Ed.:* J. E. Saxton) in *The Chemistry of Heterocyclic Compounds, Supplement to Vol. 25, Part 4 (Ed.:* E. C.Taylor*)*, Wiley, New York **1994**, pp. 279-355.

[2] G. Massiot, C. Lavaud, D. Guillaume, L. Le Men-Olivier, G. Van Binst, *J. Chem. Soc., Chem. Commun.* **1986**, 1485-1487.

[3] G. Massiot, C. Lavaud, and J.-M. Nuzillard, in *Structure elucidation of plant secondary products. Chemical from plants. Perspectives on plant secondary products*, (Eds: N. J. Walton, D. E. Brown) Imperial College Press, London **1999**, pp. 187-214.

[4] L. Mueller, *J. Am. Chem. Soc.* **1979**, *101*, 4481-4484.

[5] a) R. E. Hurd, *J. Magn. Reson.* **1990**, *87*, 422-428; b) R. E. Hurd, B. K. John, *J. Magn. Reson.* **1991**, *91*, 648-653.

[6] a) J.-M. Nuzillard, G. Massiot, *Anal. Chim. Acta* **1991**, *242*, 37-41; b) B. Plainchont, J.-M. Nuzillard, *Magn. Reson. Chem.* **2013**, *51*, 54-59.

[7] J.-M. Nuzillard, G. Massiot, *Tetrahedron* **1991**, *47*, 3655-3664.

[8] a) J. Hubert, J.-M. Nuzillard, J.-H. Renault, *Phytochem. Rev.* **2015**, DOI: 10.1007/s11101-015-9448-7; b) J. Hubert, J.-M. Nuzillard, S. Purson, M. Hamzaoui, N. Borie, R. Reynaud, J.-H. Renault, *Anal. Chem.* **2014**, *86*, 2955-2962.

[9] S. Blondeau, Q. T. Do, T. Scior, P. Bernard, L. Morin-Allory, *Curr. Pharm. Des.* **2010**, *16*, 1682-1696.

[10] D. Jeannerat, *Magn. Reson. Chem.* **2017**, *55*, 7-14.

[11] I. Pérez-Victoria, J. Martín, F. Reyes, *Planta Med.* **2016**, *82*, 857-871.

[12] http://www.atgc-montpellier.fr/permutmatrix/; accessed January 2017.

[13] http://www.acdlabs.com/; accessed January 2017.

[14] a) J. Hubert, S. Chollet, S. Purson, R. Reynaud, D. Harakat, A. Martinez, J.-M. Nuzillard, J.-H. Renault, *J. Nat. Prod.* **2015**, *78*, 1609-1617; b) P. Sientzoff, J. Hubert, C. Janin, L. Voutquenne-Nazabadioko, J.-H. Renault, J.-M. Nuzillard, D. Harakat, A. Alabdul Magid, *Molecules* **2015**, *20*, 14970-14984; c) L.-P. Tisserant, J. Hubert, M. Lequart, N. Borie, N. Maurin, S. Pilard, P. Jeandet, A. Aziz, J.-H. Renault, J.-M. Nuzillard, C. Clément, M. Boitel-Conti, E. Courot, *J. Nat. Prod.* **2016**, *79*, 2846-2855; d) S. K. Öttl, J. Hubert, J.-M. Nuzillard, H. Stuppner, J.-H. Renault, J. M. Rollinger, *Anal. Chim. Acta* **2014**, *846*, 60-67; e) A. Abedini, S. Chollet, A. Angelis, N. Borie, J.-M. Nuzillard, A.-L. Skaltsounis, R. Reynaud, S. C. Gangloff, J.-H. Renault, J. Hubert, *J. Chromatogr. B* **2016**, 1029–1030, 121–127; f) A. Angelis, J. Hubert, N. Aligiannis, R. Michalea, A. Abedini, J.-M. Nuzillard, S. C. Gangloff, A.-L. Skaltsounis, J.-H. Renault, *Molecules* **2016**, *21*, 1586.

[15] a) M. Elyashberg, A. Williams, K. Blinov, in *Contemporary Computer-Assisted Approaches to Molecular Structure*

Elucidation, RSC Publishing, Cambridge (UK) **2012**; b) M. E. Elyashberg, A. J. Williams, G. E. Martin, *Prog. Nucl. Magn. Reson. Spectrosc.* **2008**, *53*, 1-104; c) C. Steinbeck, *Nat. Prod. Rep.* **2004**, *21*, 512-518; d) M. Jaspars, *Nat. Prod. Rep.* **1999**, *16*, 241-248.

[16]B. Plainchont, V. de P. Emerenciano, J.-M. Nuzillard, *Magn. Reson. Chem.* **2013**, *51*, 447-453.

[17]R. Freeman, T. Frenkiel, M. B. Rubin, *J. Am. Chem. Soc.* **1982**, *104*, 5545-5547.

[18]a) K. Funatsu, Y. Susuta, S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6–11; b) B. D. Christie, M. E. Munk, *Anal. Chim. Acta* **1987**, *200*, 347-362.

[19]N. T. Nyberg, J. O. Duus, O. W. Sørensen, *J. Am. Chem. Soc.* **2005**, *127*, 6154-6155.

[20]a) 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; b) K. Kazimierczuk, V. Orekhov, *Magn. Reson. Chem*. **2015**, *53*, 921-926.

[21]https://www.bruker.com/products/mr/nmr/nmr-software/software/complete-molecular-confidence.html; accessed January 2017.

[22]M. B. Costantin, M. J. Ferreira, G. V. Rodrigues, V. de P. Emerenciano, *Nat. Prod. Commun.* **2010**, *5*, 755-762.

[23]a) M. J. P. Ferreira, M. B. Costantin, P. Sartorelli, G. V. Rodrigues, R. Limberger, A. T. Henriques, M. J. Kato, V. de P. Emerenciano, *Anal. Chim. Acta* **2001**, *447*, 125-134; b) M. B. Costantin, P. Sartorelli, R. Limberger, A. T. Henriques, M. Steppe, M. J. P. Ferreira, M. T. Ohara, V. de P. Emerenciano, M. J. Kato, *Planta Med.* **2001**, *67*, 771-773.

[24]D. L. G. Fromanteau, J. P. Gastmans, S. A. Vestri, V. de P. Emerenciano, J. H. G. Borges, *Comput. Chem.* **1993**, *17*, 369-378.

[25]a) J.-M. Nuzillard, V. de P. Emerenciano, *Nat. Prod. Commun.* **2006**, *1*, 57-64; b) B. Plainchont, J.-M. Nuzillard, G. V. Rodrigues, M. J. P. Ferreira, M. T. Scotti, V. de P. Emerenciano, Nat. Prod. Commun. **2010**, *5*, 763-770.

[26]J.-M. Nuzillard, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 723-724.

[27]J.-M. Nuzillard, *Chin. J. Chem.* **2003**, *21*, 1263-1267.

[28]*RDKit: Cheminformatics and Machine Learning Software*, http://www.rdkit.org; accessed January 2017.

[29]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminform.* **2011**, *3*, 33.

[30]C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E. L. Willighagen, *Curr. Pharm. Des.* **2006**, *12*, 2111-2120.

[31]C. Steinbeck, S. Kuhn, *Phytochemistry* **2004**, *65*, 2711-2717.

[32]a) D. A. Mulholland, M. K. Langat, N. R. Crouch, H. M. Coley, E. M. Mutambi, J.-M. Nuzillard, *Phytochemistry* **2010**, *71*, 1381-1386; b) D. A. Mulholland, A. Langlois, M. Randrianarivelojosia, E. Derat, J.-M. Nuzillard, *Phytochem. Anal.* **2006**, *17*, 87-90; c) A. Toribio, A. Bonfils, E. Delannay, E. Prost, D. Harakat, E. Hénon, B. Richard, M. Litaudon, J.-M. Nuzillard, J.-H. Renault, *Org. Lett.* **2006**, *8*, 3825-3828; d) J.-P. Bouillon, B. Tinant, J.-M. Nuzillard, C. Portella, *Synthesis* **2004**, 711-721; e) D. Mulholland, M. Randrianarivelojosia, C. Lavaud, J.-M. Nuzillard, S. L. Schwikkard, *Phytochemistry* **2000**, *53*, 115-118; f) D. Mulholland, S. L. Schwikkard, P. Sandor, J.-M. Nuzillard, *Phytochemistry* **2000**, *53*, 465-468; g) J.-M. Nuzillard, J. D. Connolly, C. Delaude, B. Richard, M. Zèches-Hanrot, L. Le Men-Olivier, *Tetrahedron* **1999**, *55*, 11511-11518; h) G. Almanza, L. Balderama, C. Labbé, C. Lavaud, G. Massiot, J.-M. Nuzillard, J. D. Connolly, L. J. Farrugia, D. S. Rycroft, Computer-assisted structural elucidation. *Tetrahedron* **1997**, *53*, 14719-14728 ; i) S. V. Ley, K. Doherty, G. Massiot, J.-M. Nuzillard, *Tetrahedron* **1994**, *50*, 12267-12280.

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))