Interprétation automatique des spectres de RMN de petites molécules organiques en solution : principe et exemples

Jean-Marc Nuzillard, Bertrand Plainchont

Interprétation automatique des spectres de RMN de petites molécules organiques en solution : principe et exemples

Jean-Marc Nuzillard et Bertrand Plainchont

Institut de Chimie Moléculaire de Reims, CNRS UMR 7312, Moulin de la Housse CPCBAI, Bâtiment 18, BP 1039, 51687 REIMS Cedex 2, France.

Résumé *: L'interprétation automatique des spectres de Résonance Magnétique Nucléaire (RMN) de petites molécules organiques en solution est un moyen puissant d'extraire les valeurs des déplacements chimiques et des constantes de couplage scalaire des spectres proton. C'est donc un outil important dans l'aide à l'analyse spectrale. Dans cet article, les stratégies analytiques permettant l'extraction des données sont présentées, et les relations entre les spectres de RMN et la structure des molécules envisagées pour la validation des structures et l'élucidation structurale assistées par ordinateur (logiciels CASA, CCASA, COCOON, LSD, PERCH, SENECA, …) sont décrites.*

Mots clés : RMN des liquides, Petites molécules organiques, Analyse automatique des spectres, Vérification de structure, Elucidation structurale assistée par ordinateur

Summary **:** *The interpretation of nuclear magnetic resonance (NMR) spectra of small organic molecules in solution is first considered as the method of extracting the chemical shifts and scalar coupling constant spectral parameters from ^1H spectra, and then as the means of establishing relationships between spectra and the structure of small organic molecules. Hence, this approach is an useful tool for symplifying the NMR analysis. In this article, the analytical strategies for the data extraction and the spectral interpretation are discussed, concommitantly with the description of the relationships between the NMR spectra and the molecular structures considered for computer-assisted structure verification and structure elucidation (CASA, CCASA, COCOON, LSD, PERCH, SENECA softwares, …).*

Key words : Liquid-state NMR, Small organic molecules, Automatic spectral analysis, Structure verification, Computer-assisted structure elucidation

Correspondance : jm.nuzillard@univ-reims.fr

I – Préambule

Dans le titre de cet article, l'expression "interprétation automatique" se réfère aux méthodes informatiques de recherche d'information dans les spectres de RMN, dans le cadre de l'étude des molécules organiques de bas poids moléculaire en solution. Le mot « interprétation » sera envisagé de deux manières différentes : d'une part comme l'extraction des paramètres spectraux « encryptés » dans les spectres, et d'autre part comme la validation ou la détermination de la structure d'une molécule à partir des paramètres spectraux expérimentaux.

II – Introduction

Un spectre de RMN ^1H est décrit en termes de deux types de paramètres spectraux : les déplacements chimiques (δ) et les constantes de couplages scalaires (J). La création d'une telle liste de paramètres constitue un premier niveau d'interprétation automatique et peut, en principe, être menée à bien sans rien connaître de la molécule à analyser.

La reconstruction d'un spectre de RMN ^1H à partir d'une liste de paramètres spectraux est donc un problème dont la solution est du ressort de la mécanique quantique appliquée aux états de spin des noyaux atomiques. Des méthodes exactes de résolution de ce problème existent, même si des approximations sont nécessaires pour des systèmes de spins très complexes. L'extraction des paramètres spectraux est le problème inverse du précédent pour lequel des algorithmes spécifiques ont été développés.

Interpréter un spectre de RMN en termes de paramètres spectraux n'est pas un exercice gratuit ; ce ne sont pas les spectres qui importent réellement aux chimistes organiciens mais les structures moléculaires des substances étudiées. Etablir des relations entre les spectres et les structures constitue un second niveau d'interprétation. Une structure peut être vue d'une manière simpliste comme un ensemble d'atomes et de liaisons chimiques entre les atomes. Cette vision exclut toutes les molécules organométalliques qui possèdent des liaisons multicentriques (comme le ferrocène, par exemple), mais inclut la très grande majorité des molécules organiques.

Chaque noyau d'atome actif en RMN possède un déplacement chimique spécifique au sein d'une molécule et chaque paire de tels noyaux est associée à une valeur de constante de couplage scalaire dépendante des liaisons chimiques de la molécule. Le calcul des paramètres spectraux à partir de la structure est lié à la connaissance des orbitales moléculaires de la molécule, obtenue à partir des principes de la

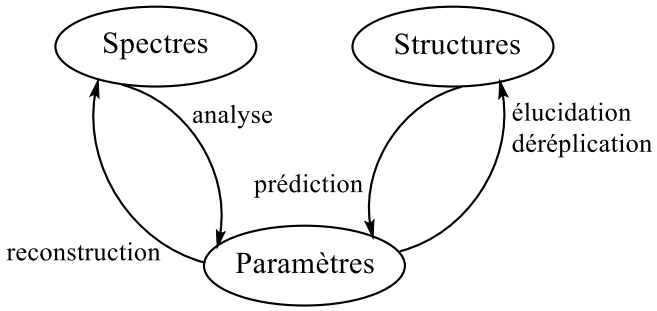


Figure 1. Relations entre spectres, paramètres spectraux et structures moléculaires. Les spectres de RMN 1-D ou 2-D sont ici considérés comme des ensembles de valeurs numériques organisés sous forme de vecteur ou de matrice. Les paramètres spectraux sont les valeurs numériques des déplacements chimiques, des intensités et des formes de pics spectraux ou des constantes de couplage qui reflètent les caractéristiques structurales d'une molécule.

mécanique quantique **[1]**. Dans ce contexte, le calcul des paramètres nécessite d'importantes approximations et les temps de calcul sont de l'ordre de l'heure. Des méthodes alternatives aux calculs quantiques *ab initio* existent depuis longtemps et sont basées sont des corrélations entre des descripteurs moléculaires et les valeurs des paramètres spectraux. La prédiction rapide, de l'ordre de la seconde, est déterminante dans le processus de validation structurale, qui consiste à vérifier qu'une structure hypothétique associée à un composé est compatible avec les données expérimentales, spectroscopiques ou autres, disponibles pour ce composé. L'élucidation structurale est le problème inverse du calcul des paramètres spectraux à partir de la structure. La « déréplication » constitue un cas particulier d'élucidation structurale où la structure recherchée est supposée avoir déjà été caractérisée et publiée. La **Figure 1** présente les différents aspects des relations entre spectres et structures pour lesquels des logiciels ont été développés.

III. Détermination des paramètres spectraux

La détermination des paramètres spectraux est une étape clée pour l'interprétation automatique des spectres de RMN. Dans cette section, nous allons aborder l'étude des spectres de RMN ^{1}H, dans lesquels les déplacements chimiques et les constantes de couplage scalaire interagissent de manière complexe dans les situations de couplage fort **[2]**.

Si toutes les paires de noyaux de tous les systèmes de spins étaient faiblement couplés, le calcul des intensités et des positions des pics spectraux serait trivial. On pourrait être tenté de penser que l'hypothèse des couplages faibles est maintenant la règle commune avec la disponibilité de spectromètres à très haut champ pour lesquels les couplages forts sont moins probables. L'analyse des sucres prouve que cette croyance est infondée, même pour des spectres enregistrés à 600 MHz ou à plus haut champ. L'extraction des paramètres dépend de la capacité à reconstruire un spectre à partir des paramètres spectraux. L'état d'un système de $\boldsymbol{N}$ spins ½ est décrit par une combinaison linéaire de 2^N éléments de la base des états de spin. En conséquence, la matrice de l'opérateur Hamiltonien de précession libre contient 4^N éléments, dont un grand nombre de valeurs nulles.

La simulation d'un spectre de RMN, c'est-à-dire la prédiction des positions et des intensités des résonances, requiert la diagonalisation d'une telle matrice, une opération mathématique dont le temps d'exécution croît rapidement avec $\boldsymbol{N}$. Des simplifications exactes sont possibles quand la molécule contient des symétries, des noyaux magnétiquement équivalents ou quand l'approximation des couplages faibles s'applique. D'autres approximations peuvent (ou doivent) être introduites lorsque $\boldsymbol{N}$ s'accroît en considérant que le couplage fort entre deux noyaux, tous deux séparés d'un troisième par un grand nombre de liaisons, n'influencera pas significativement la structure du motif de couplage du troisième. Une application immédiate de la simulation de spectres est la « résurrection » de spectres à partir de listes de paramètres spectraux, tels qu'ils sont transcrits sous forme textuelle (δ en ppm, J en Hz, multplicité (d, t, q, m, ….)) dans les publications. Le but de cette opération est de vérifier par RMN du ^{1}H l'identité d'un composé décrit dans la littérature avec un composé récemment préparé ou isolé, et qui devrait *a priori* lui être identique. La mise en pratique de cette idée est limitée par la qualité médiocre, voire mauvaise, des descriptions spectrales publiées. Des multiplets, même de faible complexité, sont souvent reportés comme *m* (multiplet) sans indication de valeur de constante de couplage *J*, ce qui constitue une perte d'information souvent irréparable et rend l'identification impossible.

L'extraction des paramètres δ et J à partir d'un spectre expérimental est un processus itératif qui doit commencer par une évaluation initiale raisonnablement correcte des valeurs des paramètres recherchés. Ces valeurs initiales sont ensuite ajustées itérativement afin que le spectre simulé devienne aussi similaire que possible au spectre expérimental. La similarité est évaluée par une quantité qui est une fonction des paramètres et dont la valeur doit être aussi faible que possible. Cependant cette valeur peut présenter de nombreux minima locaux dans lesquels l'algorithme peut tomber. La technique des « transformations intégrales » appliquée au couple de spectres expérimental et simulé a été reconnue très tôt comme un moyen d'éviter que le processus itératif ne converge vers un ensemble incorrect de paramètres **[3]**. La création de l'ensemble de paramètres initiaux peut être laissée sous la responsabilité de l'utilisateur du logiciel d'analyse ou assistée par le logiciel si la structure de la molécule est connue ou supposée.

L'existence de conformères pour les molécules flexibles complique l'évaluation initiale des déplacements chimiques et surtout des constantes de couplages en vertu de leur sensibilité à la géométrie moléculaire. Les populations de conformères doivent être déterminées de manière aussi précise que possible, ce qui nécessite une bonne connaissance de leurs énergies relatives, elles-mêmes dépendantes de la manière dont elles sont calculées. La modélisation moléculaire par champ de force peut se révéler insuffisante à cet effet.

La combinaison d'outils de dynamique moléculaire, de prédiction de paramètres spectraux, de simulation de spectres et d'ajustement itératif de paramètres spectraux est intégrée dans le logiciel commercial « **PERCH** », dont l'usage est fortement recommandé pour produire des descriptions de spectres aussi précises que possible **[4]**. Ce logiciel est, au moment où ces lignes sont écrites et à la connaissance de l'auteur, le seul qui fournisse cette fonctionnalité. Ce logiciel est aussi un outil de validation de structures qui fonctionne par comparaison entre paramètres spectraux extraits et prédits de manière itérative.

IV. Vérification des structures moléculaires

Un processus de validation structurale est destiné à qualifier l'accord entre une structure supposée pour un composé et ses propriétés physiques et chimiques. Il est en principe possible de vérifier si une structure hypothétique est correcte ou non, mais aussi si ses données spectroscopiques (de toutes natures, les plus diverses de préférence) sont identiques à des données archivées pour la même structure. Des algorithmes ont été développés pour vérifier l'identité entre spectres.

Les variations spectrales dues à des effets de température, de solvant, de concentration, de pH ou d'impureté peuvent conduire à des faux négatifs, c'est-à-dire à des structures déclarées différentes pour de mauvaises raisons. La fiabilité de la vérification dépend de la diversité des données expérimentales disponibles, incluant la spectrométrie de masse, les méthodes optiques (UV, Visible, IR, pouvoir rotatoire, dichroïsme circulaire électronique ou vibrationnel) et toutes sortes de méthodes de RMN multidimensionnelle (RMN *n*D). Le même type de stratégie est applicable au niveau des paramètres spectraux s'ils peuvent être extraits des données expérimentales brutes. La validation de la structure en l'absence de données de référence archivées dépend de la prédiction de paramètres. La comparaison entre expérience et prédiction est possible au niveau des spectres, si ceux-ci sont prédictibles à partir des paramètres correspondants.

La prédiction des déplacements chimiques en RMN par des méthodes non quantiques est fondée sur des relations entre structure et déplacements chimiques établies d'après des attributions déjà effectuées. Le codage de l'environnement d'un atome donné dans une molécule donnée est un aspect important de ces méthodes qui présuppose que des environnements semblables conduisent à des déplacements chimiques semblables. Les codes HOSE (ou de Bremser) sont familiers aux chimistes pour la prédiction des déplacements chimiques en RMN du ^{13}C. Le service web librement accessible « **nmrshiftdb2** » utilise cette méthode pour les noyaux ^{1}H, ^{13}C, ^{15}N, ^{19}F, ^{31}P et d'autres moins communs. L'utilisation de méthodes d'incrément est encore utilisée et implémentée dans le très populaire logiciel de dessin de structures moléculaires « **ChemDraw** ». D'autres moyens de prédiction fondés sur les réseaux de neurones, les forêts aléatoires ou les machines à vecteur de support ont été proposés et fournissent des qualités de prédictions qui sont du même ordre de grandeur. Les prédicteurs proposés par la société « **ACD** » bénéficient d'une très vaste collection de données issues de la littérature et d'algorithmes qui combinent plusieurs méthodes de prédiction. Une comparaison entre méthodes quantiques et non quantiques a fait ressortir la difficulté de traiter les molécules flexibles par les méthodes quantiques et le rôle important du choix d'une de ces méthodes parmi toutes celles qui sont disponibles.

La prédiction des déplacements chimiques, quelle que soit la méthode adoptée, a une fiabilité limitée. Dans une molécule complexe, une valeur de déplacement chimique particulière peut être attribuée à plus d'un seul atome, conduisant ainsi à plusieurs ensembles d'attributions pour l'ensemble des déplacements chimiques de cette molécule. Alors que l'impossibilité de pouvoir proposer une attribution est une manière de conclure à l'invalidité de la structure envisagée, la multiplicité des attributions n'est pas une situation satisfaisante car il n'y a qu'une seule attribution exacte possible.

La prise en compte des couplages scalaires homo- et hétéronucléaires constitue un moyen de limiter le nombre d'attributions possibles d'un spectre en regard d'une structure de molécule. Deux noyaux qui partagent un couplage scalaire non nul sont séparés par un petit nombre de liaisons chimiques. Les atomes correspondants ne peuvent être très éloignés dans la molécule. L'information de couplage peut ainsi invalider des ensembles d'attributions de déplacements chimiques. La détermination expérimentale des paires de déplacements chimiques de noyaux qui sont couplés est donc un problème de première importance, d'autant plus que les valeurs de constante de couplage sont beaucoup moins sensibles aux conditions expérimentales d'enregistrement des spectres (solvant, concentration, température, pH, impuretés, …) que les déplacements chimiques.

En RMN du ^{13}C, la simple connaissance du nombre de noyaux ^{1}H directement couplés (couplage ^{1}J, à travers 1 liaison) ou même de la parité ou de l'imparité de ce nombre fournit un moyen puissant de réduire le

nombre d'ensembles d'attributions. En RMN du ^{1}H, savoir quelles paires de signaux correspondent à des noyaux couplés est du ressort du découplage sélectif, une méthode (ancienne) dite de double résonance. Cependant, l'avènement de la RMN 2D a changé de manière radicale la manière d'explorer les réseaux de couplage scalaire en fournissant un accès pratique (et direct) à l'étude des couplages hétéronucléaires (entre ^{1}H et ^{13}C, le plus souvent). L'enregistrement des spectres 2D COSY, 2D HSQC (Heteronuclear Single Quantum Correlation) et 2D HMBC (Heteronuclear Multiple Bond Correlation) constituent actuellement la base commune pour l'attribution des signaux et l'élucidation structurale.

Un spectre RMN 2D HSQC ^{1}H-^{13}C corrèle les déplacements chimiques de noyaux séparés par une seule liaison exclusivement, tirant avantage des valeurs importantes des constantes ^{1}J(C-H), supérieures à 120 Hz, par rapport aux ^{n}J(C-H) (n > 1) qui sont au moins dix fois plus faibles. Un spectre 2D HSQC impose une cohérence logique entre les attributions des signaux de RMN du ^{1}H et du ^{13}C qui, combinée avec les prédictions de déplacements chimiques de ces deux noyaux, réduit considérablement les erreurs d'attribution. Le spectre HSQC est utilisé pour cette raison comme source d'information dans de nombreux systèmes de validation structurale. Par ailleurs, un spectre 2D HSQC révèle aussi les paires de déplacements chimiques de noyaux ^{1}H différents mais liés au même carbone dans un groupe CH_2. La connaissance de ces paires internucléaires facilite la localisation des pics de corrélation par les couplages géminés ^{2}J(^{1}H-^{1}H) dans les spectres COSY ^{1}H. D'autres pics de corrélation COSY ^{1}H issues de couplages ^{n}J avec n > 2 révèlent une distance de n-2 liaisons entre des atomes lourds (non-hydrogène) qui portent ces noyaux ^{1}H couplés. Toutefois, les valeurs des ^{4}J(^{1}H-^{1}H) sont le plus souvent inférieures à 2 Hz, à l'exception des couplages allyliques ou « en W » respectivement dans les systèmes de liaisons insaturés ou saturés.

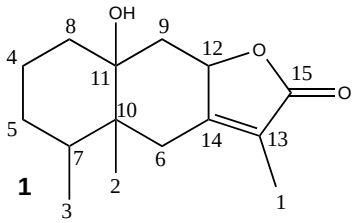
La valeur des constantes de couplage peut être évaluée qualitativement par l'intensité d'un pic de corrélation COSY, voire de manière quantitative dans les spectres COSY phasés. Un couplage ^{n}J(^{1}H-^{13}C) avec n > 1 visualisé par un spectre HMBC révèle une distance de n-1 liaisons entre deux atomes lourds. Ces couplages sont généralement faibles lorsque n > 3 mais des couplages faibles peuvent être observés avec n = 2 ou n = 3. Les couplages ^{3}J(^{1}H-^{13}C), sont faibles (0-3 Hz) lorsque l'angle dièdre central est proche de 90°, ainsi que les ^{2}J(^{1}H-^{13}C) dans les systèmes aromatiques.

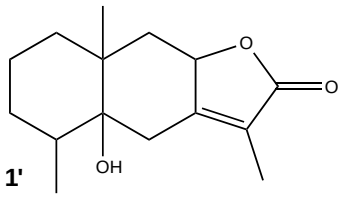
L'expérience 2D H2BC a été conçue pour différencier les couplages hétéronucléaires ^{2}J et ^{3}J. Elle repose sur la présence ou l'absence d'un couplage ^{3}J(^{1}H-^{1}H). A part pour les ^{1}J(^{1}H-^{13}C), il n'existe aucune méthode infaillible pour déduire de la valeur d'une constante de couplage un nombre de liaisons qui séparent les noyaux couplés. Toutefois, les corrélations de forte intensité dans les spectres COSY et HMBC imposent des relations de proximité entre atomes qui sont souvent suffisantes pour conduire à bien une procédure d'attribution des spectres 1D fondée sur le couplage scalaire.

Le logiciel « Computer-Aided Spectral Assignment » (**CASA**) a été initialement écrit pour proposer les attributions les résonances en RMN du ^{13}C d'une molécule de structure connue ou postulée. L'article initial, publié en 1991, et présentant le logiciel contenait déjà les principes de base de la vérification de structure fondée sur le couplage scalaire et de l'élucidation structurale des petites molécules organiques **[5]**. Cet article proposait la réduction des données contenues dans les spectres HMBC, HSQC, et COSY en relations de proximité entre atomes lourds et la visualisation de ce procédé par une table de corrélation en « croix-et-cercles ». Le logiciel **CASA** a été utilisé pour démontrer qu'un sesquiterpène **1** ne pouvait pas être de squelette carboné eudesmane (comme **1'**) mais était compatible avec le squelette érémophilane proposé par ailleurs (cf. **Figure 2**).

En l'absence d'information de couplage, la préférence pour l'une ou l'autre de ces deux possibilités a dû être suggérée avec un argumentaire fondé sur les valeurs de déplacement chimique. Le logiciel **CASA** peut aussi traiter des données de proximité immédiate entre atomes lourds comme celles fournies par les spectres INADEQUATE, comme le montre un exemple d'attribution du spectre RMN du ^{13}C de dérivés de triterpènes pentacycliques. La version de 1991 de CASA était écrite en langage PROLOG, un langage informatique qui a été créé pour des applications en intelligence artificielle. La combinaison des relations de proximité avec la prédiction des déplacements chimiques a été implémentée récemment dans le logiciel **CCASA**, écrit en langage C.

Le logiciel **CCASA** commence par établir pour chaque résonance ^{13}C une liste d'atomes de carbones candidats à l'attribution sur la base de la multiplicité de la résonance (le nombre d'atomes d'hydrogène liés) et, en option, sur l'état d'hybridation des atomes attribué d'après la valeur du déplacement chimique. Les listes peuvent être réduites en comparant chaque déplacement chimique expérimental et ceux prédits pour les atomes de la liste associée. La prédiction des déplacements chimiques dans **CCASA** utilise une version autonome du





Numéro de carbone	Numéro de proton						Multiplicité
	1	2	3	6	9	12	
1	x						3
2		x					3
3			x				3
4							2
5			o				2
6		o		x			2
7		o	o	o			1
8					o		2
9					x		2
10		o	o	o	o		0
11		o		o	o		0
12				o	o	x	1
13	o			o		o	0
14				o	o	o	0
15	o						0

Figure 2. Structure du 10-β-hydroxyeremophilenolide **1** et une structure alternative incorrecte **1'**. Les atomes de carbone sont numérotés dans l'ordre des déplacements chimiques décroissants. La multiplicité est le nombre d'atomes d'hydrogène directement liés.

prédicteur **nmrshiftdb2**. Ce logiciel fournit un intervalle de confiance autour de la valeur prédite qui peut être élargi et seuillé par l'utilisateur. Si un déplacement chimique expérimental se trouve hors de l'intervalle de confiance d'un atome, alors cet atome est éliminé de la liste associée à la valeur expérimentale. Si une résonance est associée à une liste vide, cela signifie que son existence ne peut être justifiée à partir de la structure proposée, ce qui invalide la structure en question. Un faux négatif peut survenir dans le cas d'une mauvaise information de multiplicité ou de degré d'hybridation associée à une résonance ou dans le cas d'une prédiction incorrecte de déplacement chimique et de l'intervalle de confiance associé.

Si chaque résonance est attribuable à au moins un atome, l'attribution de l'ensemble des résonance est simplifiée en commençant par celle qui possède le plus petit nombre d'attributions. Toutes les attributions possibles de cette résonance seront considérées séquentiellement. Une attribution est validée si elle ne rentre pas en conflit avec les données de proximité entre atomes fournies par la RMN 2D.

Par exemple, si les résonances R1 et R2 sont attribuées aux atomes A1 et A2, alors la distance d(A1, A2), mesurée en liaisons, doit être compatible avec l'intervalle de distances associé à la corrélation des déplacements chimiques de A1 et A2. L'intervalle de distance de chaque corrélation de type COSY et HMBC est soit spécifiée par défaut soit sur instruction de l'utilisateur. Un nombre maximal de corrélations à très longue distance (plus de 3 liaisons) doit être imposé. Une telle limitation est imposée pour qu'une attribution d'un spectre ne repose pas sur un nombre irréaliste de couplages à très longue distance. Une interprétation erronée des données présente dans les spectres 2D est une source de faux négatifs. Une fois que l'attribution d'une résonance est validée, l'atome concerné est retiré de la liste de candidats de toutes les résonances. Ensuite, une résonance non encore attribuée, possédant un nombre minimal d'attributions possibles, est choisie et attribuée à un atome, jusqu'à ce que toutes les résonances soient attribuées. Si à un moment donné, la liste des candidats associée à une résonance non attribuée devient vide, alors les précédentes hypothèses d'attribution doivent être remises en cause, dans l'ordre inverse de celui où elles ont été proposées. L'appariement systématique des résonances et des atomes est organisé selon un algorithme de recherche « en profondeur d'abord » très proche de celui utilisé dans **CASA**, écrit en langage PROLOG. La catégorie de problème résolu ici est connue sous le nom de « problème à satisfaction de contraintes sur domaines finis », signifiant ainsi qu'une résonance n'est attribuable qu'à un nombre finis d'atomes.

Le logiciel **CCASA** a été testé sur le composé 2, dont la structure triterpénique a été tirée d'une publication (cf. **Figure 3**). Il a fallu autoriser cinq corrélations à longue distance et fournir un rayon de tolérance de 8 ppm sur les déplacements chimiques prédits afin que **CCASA** produise une attribution. Nous en avons conclu que la structure proposée n'était pas correcte et qu'il conviendrait d'en rechercher une plus vraisemblable (voir ci-après la section « Elucidation structurale »). Pour conclure cette section, la vérification de structure fondée sur les connectivités issues de la RMN 2D a été implémentée dans les logiciels **CASA** et **CCASA**. Une autre implémentation, non documentée, existe dans le module **CMC-se** du logiciel « TopSpin » de la société Bruker.

IV. Elucidation structurale

L'élucidation structurale assistée par ordinateur fait régulièrement l'objet d'articles de revue, principalement dans la perspective de la chimie des substances

naturelles car cette discipline est une source inépuisable de problèmes structuraux variés et complexes **[6]**.

Avant de commencer la recherche de la structure d'un composé *a priori* inconnu, la première question à se poser devrait être : « cette substance a-t-elle déjà été isolée et caractérisée ? ». En d'autres termes, la déréplication devrait toujours précéder l'élucidation. La déréplication est possible si les bases de données qui relient les structures et les données spectroscopiques sont accessibles, ainsi que les outils appropriés de recherche dans ces bases. De manière idéale, la déréplication pourrait être effectuée à partir d'un ensemble minimal de données spectroscopiques, comme un spectre de masse et un spectre de RMN ^{1}H, dont l'enregistrement nécessite un minimum de temps et de masse d'échantillon. Formellement, la déréplication peut être considérée comme un cas particulier de vérification de structure, effectuée sur l'ensemble d'une base de données. L'élucidation structurale *de novo* ne devrait être entreprise que si la déréplication n'est pas réalisable ou si elle a été tentée sans succès.

Une récente motivation pour le développement de systèmes « **CASE** » (pour Computer-Assisted Structure Elucidation) a découlé de leur possible intégration dans l'analyse métabolomique, dans le but d'identifier les nouveaux marqueurs biologiques révélés par le traitement statistique des données spectroscopiques. Si l'identification par déréplication échoue, les outils expérimentaux d'analyse structurale doivent être sollicités, principalement la spectrométrie de masse et la RMN. L'interprétation des spectres obtenus est considérée comme un goulet d'étranglement dans les études métabolomiques visant au développement de nouvelles stratégies dans les sciences de la santé.

L'élucidation de structure *de novo* est passée de l'art divinatoire à une technique pratique avec l'avènement des techniques de RMN 2D HMQC et HMBC. Les techniques précédentes, XHCORR et COLOC, fondées sur la détection directe des signaux ^{13}C, étaient fondamentalement moins sensibles et davantage sujettes aux artefacts spectraux. L'utilisation d'expériences inverses (détection du proton) a radicalement amélioré les capacités d'élucidation structurale, même à une époque où ni les sondes multi-noyaux optimisées pour la détection des noyaux ^{1}H ni le découplage large-bande des ^{13}C n'était disponible. L'avènement des sondes dites « inverses » avec des bobines de gradient de champ blindées activement a marqué le début d'une longue période de stabilité des méthodes d'élucidation structurale fondées sur la RMN 2D, période où l'amélioration de la qualité des spectres a été apportée par des aimants de plus en plus puissants et par les sondes RMN ultra-froides.

La facilité d'accès à des spectres de corrélation ^{1}H-^{13}C et la fiabilité de leur interprétation ont rapidement fait naître l'idée qu'un logiciel pourrait efficacement assister les chimistes dans la tâche d'élucidation structurale. Cette voie a été encouragée par les premiers succès de l'attribution automatisée par le logiciel **CASA**. A cette époque, la force de proposition des logiciels de **CASE** était l'interprétation des valeurs de déplacement chimique par l'intermédiaires de bases de données, même si l'idée de contraindre la génération de structure par des données de RMN 2D COSY et HMBC avait déjà été mise en œuvre. La proposition de structures fondées sur les couplages révélés par la RMN a constitué un nouveau paradigme pour les logiciels de **CASE**. Cette approche a été publiée en 1991 dans un article intitulé « Logic for Structure Determination » qui présentait les fondements du logiciel **LSD [7].**

Un chimiste considère que le travail d'élucidation structurale a abouti lorsque toutes les liaisons entre tous les atomes ont été placées (structure 2D), lorsque les configurations relatives des différents centres asymétriques ont été déterminées (structure spatiale 3D relative) et, pour les composés énantiomériquement purs, quand les configurations absolues sont connues (structure spatiale 3D absolue). Les informations nécessaires à l'établissement des configurations relatives sont facilitées dans les parties rigides des molécules pour lesquels les équilibres conformationnels n'introduisent pas de variabilité des distances inter-atomiques et des angles dièdres. Des méthodes automatiques de détermination des structures 3D relatives ont été proposées dans ce cadre. La détermination des configurations absolues par RMN reste un problème complexe sans solution générale et dans l'intégration à un protocole automatisé n'existe pas encore **[8]**.

Les logiciels de **CASE** utilisables par les chimistes actuellement (en 2015) incluent **Structure Elucidator** (ACD Labs), **AssembleIt** (ScienceSoft), **CMC-se** (Bruker), **SENECA**, **COCON** et **LSD**. Les trois premiers sont distribués par des sociétés commerciales et les trois derniers sont d'origine universitaire. **COCON** est accessible au travers d'une interface web et les deux derniers sont distribués comme logiciels libres. D'autres logiciels sont éventuellement utilisables dans le cadre de collaborations scientifiques. **Structure Elucidator** a fait l'objet de très nombreux articles qui en discutent les aspects fondamentaux et les applications. **AssembleIt** n'est l'objet que d'une description rapide. **CMC-se** propose deux algorithmes de génération de structure, celui de **LSD** et un autre sur lequel rien n'est publié. **SENECA** utilise une méthode stochastique de génération de structure, alors que **COCON** et **LSD** sont déterministes.

L'algorithme de génération de structure de **SENECA** est fondé sur les travaux originaux de Jean-Loup Faulon. Une structure est tirée initialement au hasard et comporte donc un grand nombre de violations des contraintes imposées par la RMN. **SENECA** essaie d'améliorer la vraisemblance des structures proposées par permutation des atomes, dans un processus de

recuit simulé ou d'optimisation par méthode génétique. La vraisemblance est jugée par divers critères, comme la proximité des valeurs des déplacements chimiques expérimentaux et calculés, le respect des contraintes de distance issues de la RMN 2D, la présence de groupes d'atomes particuliers ou de sous-structures, ou bien la conformité à la règle de Bredt. Un critère de « ressemblance à un composé naturel » a été défini pour guider l'affinement des structures lorsque cela a un sens. Plusieurs recherches sont effectuées en parallèle à partir de structures initiales différentes afin d'augmenter la probabilité d'atteindre une structure optimale. L'analyse stochastique possède l'avantage d'être très tolérante à des données expérimentales apparemment contradictoires ou si la molécule étudiée possède des groupements fonctionnels rares pour lesquels la prédiction des déplacements chimiques est peu fiable. D'autres méthodes d'optimisation stochastique ont été proposées, comme celle fondée sur les « colonies de fourmis ».

Le logiciel **LSD** combine actuellement deux couches logicielles : **pyLSD** **[9]** et la version de **LSD** « historique » dite LSD canonique, ci-après. **PyLSD** prépare un ou plusieurs fichiers d'entrée pour **LSD** canonique, donnant ainsi à l'utilisateur la possibilité de traiter des problèmes que ce dernier ne pourrait pas traiter sans intervention manuelle. La couche **pyLSD** est écrite en langage de haut niveau Python, elle prend en charge les ambiguïtés de formule brute et de statut des atomes (voir ci-dessous), la résolution des problèmes par **LSD** canonique et le classement des solutions.

Un problème de détermination de structure traitable par **LSD** canonique se caractérise par l'absence d'atome de statut ambigu. Le statut d'un atome est défini par son élément chimique, sa multiplicité (nombre d'hydrogènes attachés), son état d'hybridation (*sp*, sp^2 ou sp^3) et sa charge électrique. **LSD** canonique doit connaître le statut de chaque atome car il définit le nombre d'atomes immédiatement voisins, cela lui permet d'appliquer des règles très efficaces de formation de liaisons lorsque tous les voisins d'un atome sont connus (ou postulés). Dans les problèmes réels, la valeur du déplacement chimique en RMN du ^{13}C est parfois insuffisant pour déterminer le degré d'hybridation d'un atome de carbone.

Le statut des hétéroatomes n'est pas toujours facile à déterminer, surtout si plusieurs types d'hétéroatomes sont présents dans la molécule étudiée. Le nombre d'hétéroatomes peut aussi ne pas être connu avec

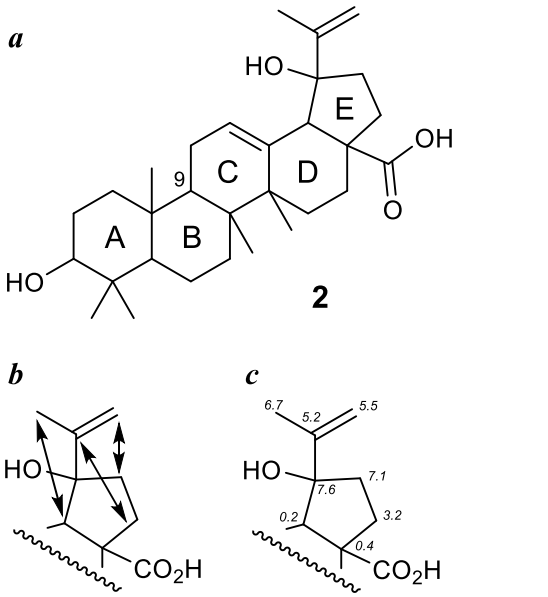


Figure 3. (a) Structure du composé 2, telle que reportée la première fois dans la littérature. (b) Corrélations HMBC à très longue distance dans le cycle E (c) valeur absolue des différences de déplacement chimiques entre valeurs expérimentales et prédites par nmrshiftdb2.

certitude lorsque les données pertinentes de spectrométrie de masse ne sont pas disponibles. Des degrés de libertés sur la formule brute et la charge électrique moléculaire sont ainsi laissés à la charge de **pyLSD**. Une fois que **LSD** canonique a résolu tous les problèmes fournis par **pyLSD**, ce dernier évalue les déplacements chimiques en RMN du ^{13}C de tous les atomes de carbone de toutes les solutions et calcule une déviation pour chaque solution en tenant compte des déplacements chimiques expérimentaux. L'évaluation des déplacements chimiques est effectuée par nmrshiftdb2, dont une version autonome est fournie avec **pyLSD**. Ce prédicteur a été choisi pour son caractère général, ses performances honorables et sa disponibilité comme logiciel libre.

L'implémentation de **pyLSD** n'a été possible que grâce à des modifications récentes (2013) apportées à **LSD** canonique afin d'accepter la plupart des éléments chimiques rencontrés dans les molécules organiques naturelles ou synthétiques. Le statut des atomes inclut maintenant les atomes *sp* et les atomes chargés électriquement. Les différentes valences de l'azote, du soufre et du phosphore sont permises ; l'azote pentavalent a été introduit pour représenter des groupes N-oxy ou nitro- sans introduire deux atomes chargés. D'autres améliorations concernent le point déjà évoqué des couplages COSY et HMBC à très longue distance et à la restriction de leur nombre maximal tolérable. Finalement, la sélection des structures qui satisfont à des critères sub-structuraux a été améliorée et supporte la recherche de multiples fragments et la combinaison logique de ses résultats.

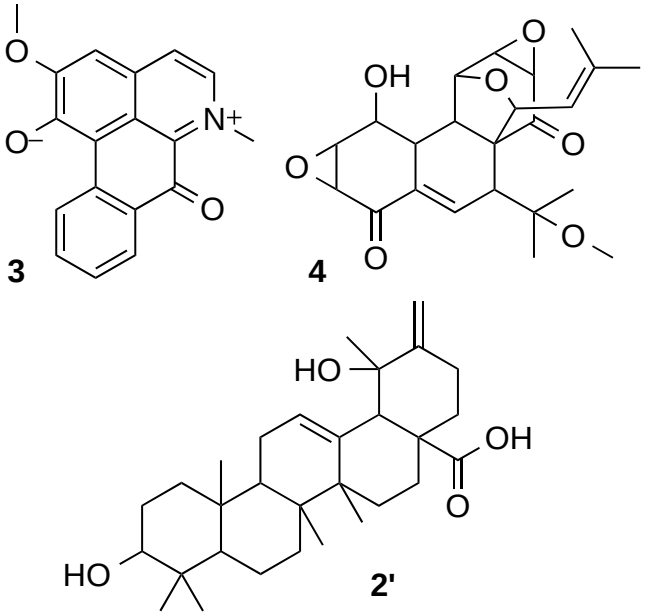


Figure 4. Solutions de problèmes résolus à l'aide du logiciel LSD : 2-O,N-dimethylliriodendronine **3**, hexacyclinol **4**, et une solution alternative **2'** à la structure **2** qui est plus vraisemblable que celle initialement proposée.

De manière semblable à **CASA**, **LSD** opère dans un mécanisme récursif de recherche en profondeur d'abord pendant lequel chaque corrélation est interprétée de toutes les manières possibles afin de construire des liaisons. Quand toutes les données de la RMN 2D ont été exploitées, les atomes auxquels il manque des liaisons sont systématiquement appariés pour obtenir une structure moléculaire complète. Les structures sont enfin filtées par application de la règle de Bredt et les critères de sous-structure. Des structures isomorphes sont susceptibles d'être produites et sont éliminées par comparaison des descripteurs InChI, sur demande de l'utilisateur.

Un point commun de toutes les approches de type COSY-HSQC-HMBC est la nécessité pour la molécule recherchée d'avoir des atomes d'hydrogène en nombre suffisant, si possible liés à des atomes actifs en RMN. Les spectres HSQC et HMBC ^{1}H-^{15}N, lorsque cela est pertinent, fournissent des informations dont le rôle peut s'avérer capital pour le succès de l'assemblage d'une structure inconnue. L'observation de couplages ^{1}H-^{13}C à très longue distance par des spectres dédiés constitue une autre source de données utiles. Si le nombre de données est faible eu égard au nombre de liaisons à établir, le problème peut posséder un très grand nombre de solutions dont le classement peut se révéler fructueux s'il ne nécessite pas un temps de calcul déraisonnablement long. L'enregistrement d'un spectre 2D INADEQUATE ^{13}C-^{13}C reste la dernière arme disponible pour peu que quelques dizaines de mg du composé inconnu soient disponibles, ainsi qu'un « week-end » de temps de spectromètre à très haut champ. Ce spectre indique directement les paires de déplacements chimiques de ^{13}C des carbones directement liés. **LSD** est ensuite capable d'intégrer les données de RMN 2D de l'azote 15 et du spectre INADEQUATE ^{13}C-^{13}C.

La structure du 2-O,N-diméthylliriodendronine **3** a été trouvée par **LSD** en laissant l'atome d'azote avoir une charge nulle ou +e et deux oxygènes avoir une charge nulle ou –e. La structure de l'hexacyclinol **4** a été trouvée à partir de données de la littérature. Cette molécule ne présente pas de problème d'ambiguïté de statut d'atomes mais la diversité des fonctions chimiques oxygénées favorise la production par **LSD** d'un grand nombre d'isomères. La prédiction des déplacements chimiques a permis de classer comme la plus vraisemblable la structure communément considérée comme correcte. Finalement, la structure **2'** a été proposée comme plus vraisemblable que la structure **2** (cf. **Figure 4**).

V. Conclusion

Malgré les efforts de quelques groupes de recherche à travers le monde depuis la fin des années 1960, les outils d'analyse structurale automatique des petites molécules ne sont pas encore un outil d'usage commun dans les laboratoires de chimie organique. Ce déficit d'impact n'est certainement pas lié à un seul facteur ; les raisons en sont probablement autant humaines que techniques. L'importance des raisons techniques devraient pouvoir diminuer avec la recherche de méthodes performantes d'analyse des images que sont les spectres de RMN 2D, dont il faut extraire de manière la plus fiable possible les informations pertinentes. Des améliorations algorithmiques pour résoudre les problèmes combinatoires d'assemblage des structures et pour affiner la qualité des prédictions des déplacements chimiques devraient contribuer à améliorer la performance de l'élucidation structurale.

Gageons que ces améliorations pourront venir à bout des réticences bien humaines à voir des machines mener à bien des tâches complexes que les chimistes sont toujours enclins à croire être exclusivement du ressort de leurs capacités intellectuelles.

VI – Remerciements

J.-M. N. et B. P. remercient le CNRS, le Conseil Régional de Champagne-Ardenne, le Conseil Général de la Marne, le Ministère de l'Enseignement Supérieur et de la Recherche et le programme européen FEDER pour leur soutien financier en général, et au projet CPER PlAneT en particulier.

VII – Acronymes utilisés

CASE : Computer-Assisted Structure Elucidation
CASA : Computer-Aided Spectral Assignment

CCASA : CASA written in C language
CMC-se : Complete Molecular Confidence – structure elucidation
COCON : from COrrelation data to CONstitution
COSY : Correlation SpectroscopY
HMBC: Heteronuclear Multiple Bond Correlation
HSQC : Heteronuclear Single Quantum Correlation
INADEQUATE : Incredible Natural Abundance DoublE QUAntum Transfer Experiment
LSD: Logic for Structure Determination
PERCH : PEak reseaRCH
RMN : Résonance Magnétique Nucléaire
SENECA : Structure Elucidation by NMR Edited generation of Constitutional isomers based on simulated Annealing

VIII – Eléments de bibliographie

[1] BARONE V., CIMINO P., CRESCENZI, O., PAVONE M., Ab initio computation of spectroscopic parameters as a tool for the structural elucidation of organic systems. J. Mol. Struct. (Theochem.), 2007, 811, 323–335.
[2] ERNST R.R., BODENHAUSEN G., WOKAUN A., "Principles of Nuclear Magnetic Resonance in One and Two Dimensions", Oxford University Press, Oxford, 1987.
[3] DIEHL P., SYKORA A., VOGT J., Automatic Analysis of NMR Spectra: An Alternative Approach. J. Magn. Reson., 1975, 19, 67-82.
[4] LAATIAKINEN R., TIAINEN M., KORHONEN S.-P., NIEMITZ M., Computerized Analysis of High-resolution Solution-state Spectra.eMagRes DOI: 10.1002/9780470034590.emrstm1226, 2013.
[5] MASSIOT G., NUZILLARD J.-M., Computer-aided spectral assignment in nuclear magnetic resonance spectroscopy. Anal. Chim. Acta, 1991, 242. 37-41.
[6] JASPARS M., Computer assisted structure elucidation of natural products using two-dimensional NMR spectroscopy. Nat. Prod. Rep., 1999, 16, 241-248.
[7] NUZILLARD J.-M., MASSIOT G., Logic for Structure Determination. Tetrahedron, 1991, 47, 3655-3664.
[8] BERGER R., COURTIEU J., Gil R. R., GRIESINGER C., KÖCK M., LESOT P., LUY B., MERLET D., NAVARRO-VASQUEZ A., REGGELIN M., REINSCHEID U.M., THIELE C.M., ZWECKSTETTER M., Is Enantiomeric Assignment Possibly by NMR Using Residual Dipolar Couplings from Chiral Non-Racemic Alignment Media? - A Critical Assessment. Angew. Chem. Int. Ed. 2012, 51, 8388-8391.
[9] PLAINCHONT B., EMERENCIANO, V.dP., NUZILLARD, J.-M., Recent advances in the structure elucidation of small organic molecules by the LSD software. Magn. Reson. Chem., 2013, 51, 447-453.