Institut De Chimie Moléculaire De Reims

École Doctorale Sciences Technologie Santé

T H È S E

pour obtenir le grade de

Docteur de l'Université de Reims Champagne-Ardenne
Discipline : Chimie Informatique

Par

Bertrand Plainchont

le 18 décembre 2012

Analyse structurale automatique des petites molécules organiques

Directeur de thèse : Dr. Jean-Marc Nuzillard

JURY :

Dr. Marc-André Delsuc	Université de Strasbourg	*Rapporteur*
Dr. Dragos Horvath	Université de Strasbourg	*Rapporteur*
Pr. Éric Henon	Université de Reims Champagne-Ardenne	*Président*
Dr. Jean-Marc Nuzillard	Université de Reims Champagne-Ardenne	*Directeur de thèse*

Remerciements

Je remercie tout particulièrement mon directeur de thèse, le Dr. Jean-Marc Nuzillard, pour sa grande disponibilité, sa pédagogie et ses précieux conseils.

Je tiens à remercier le Dr. Marc-André Delsuc et le Dr. Dragos Horvath pour avoir accepté d'être les rapporteurs de ma thèse. Je remercie également le Pr. Éric Henon d'avoir accepté de faire partie de mon jury.

Je remercie le Pr. Vicente de Paulo Emerenciano pour la mise à disposition de la base de données SISTEMAT, le Dr. Stefan Kuhn pour la mise en œuvre du module nmrshiftdb2 de prédiction des déplacements chimiques, et le Dr. Pavel Kessler pour de stimulantes discussions.

Mes remerciements s'adressent aux membres du groupe Isolement et Structure et aux doctorants de l'ICMR pour leur bonne humeur et leur sympathie.

Je remercie enfin ma famille et mes amis pour leur soutien durant ces années d'études.

Table des matières

Table des figures

Liste des abréviations

ADEQUATE Adequate sensitivity DoublE-QUAntum spEctroscopy

ASV Automatic Structure Verification

CASA Computer-Aided Spectral Assignment

CASE Computer-Assisted Structure Elucidation

CDK Chemistry Development Kit

CISOC-SES Computerized Information System for Organic Chemistry - Structure Elucidation System

COCON COnstitutions from CONnectivities

COSY COrrelation SpectroscopY

DEPT Distortionless Enhancement by Polarization Transfer

H2BC Heteronuclear 2-Bond Correlation

HMBC Heteronuclear Multiple Bond Correlation

HOSE Hierarchically Ordered Spherical description of Environment

HSQC Heteronuclear Single Quantum Correlation

INADEQUATE Incredible Natural Abundance DoublE QUAnTum Experiment

InChI International Chemical Identifier

IUPAC International Union of Pure and Applied Chemistry

LSD Logic for Structure Determination

MCSS Maximum Common Substructure Search

NOESY Nuclear Overhauser Effect SpectroscopY

RMN Résonance Magnétique Nucléaire

ROESY Rotating-frame Overhauser Effect SpectroscopY

SESAMI Systematic Elucidation of Structure Applying Machine Intelligence

SMILES Simplified Molecular-Input Line-Entry System

Introduction générale

La connaissance de la structure des molécules est cruciale dans tous les secteurs de la chimie. En effet, toute synthèse organique s'achève par un contrôle précis du produit obtenu. De même, l'extraction et l'isolement d'une substance d'origine naturelle sont suivis par une étape de détermination de structure.

L'analyse structurale a pour but la connaissance de la nature et de l'enchaînement des atomes ainsi que la détermination de la configuration absolue des éventuels centres asymétriques. Elle peut s'apparenter à une véritable enquête sur l'identité et les liens des constituants d'une molécule. Ainsi, une structure est obtenue à partir de l'assemblage d'informations et du recoupement d'indices fournis par de multiples techniques d'analyse.

La Résonance Magnétique Nucléaire (RMN) est une technique expérimentale devenue incontournable dans le processus d'analyse des molécules organiques car il s'agit d'une méthode très informative à tous niveaux (nature et proximité des atomes). La RMN ne se limite évidemment pas à l'analyse des petites molécules en solution mais possède également un grand potentiel dans l'analyse de systèmes complexes comme les protéines et pour l'analyse de produits en phase solide.

L'étude d'une structure est constituée par un ensemble d'actions répétées à chaque nouvelle analyse. De ce constat naît le challenge de l'automatisation de l'analyse des structures.

Les travaux de ce manuscrit se limitent au cadre de l'analyse structurale de petites molécules organiques. L'étude de systèmes plus complexes comme les protéines nécessite des techniques expérimentales et une méthodologie d'interprétation des données différentes. L'objectif de la thèse est le développement de logiciels d'aide à l'interprétation des spectres RMN.

La première partie du manuscrit situe le sujet de l'étude en rappelant le contexte et en faisant un point sur l'état de l'art dans le domaine. Elle permet ainsi d'introduire les méthodes et les outils d'aide à l'analyse structurale. Les expériences de RMN sont brièvement rappelées avant une présentation successive des outils de prédiction des spectres RMN, puis de l'évolution des logiciels d'élucidation structurale et enfin des logiciels de vérification de structure.

La seconde partie traite des améliorations qui ont été apportées au logiciel d'élucidation structurale LSD durant cette thèse. Tout d'abord le logiciel est présenté en détail afin d'appréhender son principe de fonctionnement et les causes des modifications. Les multiples améliorations visent à faciliter l'utilisation du logiciel et à le rendre plus effi-

cace. Elles concernent le générateur de structure ainsi que la présentation des résultats. Les modifications sont abordées avec des exemples d'applications permettant de valider leur apport.

La troisième partie concerne le développement et la mise au point d'un logiciel d'aide à l'attribution des signaux RMN. Ce travail s'appuie sur des travaux antérieurs du laboratoire où les travaux présentés dans ce manuscrit ont été réalisés. Le logiciel CASA a été réécrit en totalité en gardant l'essence du logiciel de base et en ajoutant un outil de prédiction des déplacements chimiques afin d'affiner l'analyse. Des exemples ont été choisis pour montrer les diverses fonctionnalités du logiciel et permettre de mettre en lumière un effet de bord tout à fait intéressant du logiciel. En effet celui-ci peut être assimilé à un logiciel de vérification de structure. Enfin, nous verrons comment le lien peut être fait entre les deux logiciels mentionnés ci-dessus par l'utilisation de LSD comme générateur de structure alternative en cas de doute lors de la vérification d'une structure avec CASA.

Chapitre 1

Méthodes et outils d'aide à l'analyse structurale par RMN

1.1 L'analyse structurale

1.1.1 Introduction

Le problème de la détermination de structure est un problème aussi vieux que la chimie. De tous temps, les chimistes ont cherché à identifier et décrire l'ensemble des éléments constituants une molécule. Cet ensemble d'éléments peut être présenté de plusieurs façons. La formule brute est une manière très résumée et compacte de représentation. Par exemple la formule brute du benzène est C_6H_6. Par opposition, la formule développée est une représentation détaillée permettant d'avoir accès à la description des liens entre les atomes. Par définition, une molécule ne possède qu'une formule brute mais la réciproque n'est pas vraie. Une formule brute peut être associée à plusieurs structures ayant le même nombre de constituants agencés différemment. Ces structures sont appelées isomères (217 pour la formule brute du benzène en excluant les atomes chargés). Ce principe d'isomérie constitue une des raisons d'être de la détermination de structure.

Les premiers pas réalisés sur le chemin de la détermination structurale ont été guidés par l'idée d'imaginer tous les isomères possibles d'une formule brute et de trouver des indices permettant de faire la différence entre ces isomères. Ce chemin se transforme rapidement en un véritable labyrinthe doté de multiples allées où il est nécessaire de posséder des méthodes et des outils pour se diriger et choisir la direction pour enfin atteindre le but fixé.

Les méthodes et les outils fournissant les indices permettant d'avoir accès à la structure d'une molécule n'ont cessé de s'affiner au cours du temps. L'évolution des techniques expérimentales et des progrès technologiques a permis la naissance de méthodes d'analyse et de raisonnements nouveaux.

Pour un chimiste actuel, qu'il s'agisse d'un chimiste organicien voulant analyser un produit de synthèse ou un chimiste des substances naturelles souhaitant connaître la structure d'une molécule extraite d'une plante, le recours aux méthodes d'analyse modernes est inévitable. La Résonance Magnétique Nucléaire (RMN) a acquis une place prépondérante dans le processus d'analyse structurale des petites molécules en solution car cette technique permet d'avoir accès à des informations sur les atomes pris individuellement et

aussi des informations de proximité de ces atomes.

Lorsqu'il s'agit d'analyser la structure d'une molécule, un chimiste peut se retrouver dans deux cas de figure. Dans la première situation, il a une idée de la structure de la molécule analysée car il s'agit d'un produit de réaction attendu. Dans ce cas, il cherche uniquement à vérifier si le produit obtenu est le bon. Dans la seconde situation, il est en présence d'une substance totalement ou partiellement inconnue. Il peut s'agir d'une substance naturelle extraite d'une plante, d'un sous-produit de réaction inattendu en synthèse organique ou d'un produit issu de la dégradation d'un médicament dans l'industrie pharmaceutique. Il s'agit alors d'établir la structure.

1.1.2 Aide à l'analyse par l'automatisation

Les méthodes et les règles utilisées en analyse structurale constituent un processus répétitif appliqué pour chaque nouvelle structure en cours d'étude. Comme toute action amenée à être répétée, on peut s'interroger sur la pertinence de son automatisation. Dans le cas de l'analyse structurale, l'automatisation a l'avantage de pouvoir réduire le temps nécessaire à l'analyse d'une structure. Le rendement d'analyse est donc augmenté et les chimistes peuvent se concentrer sur les problèmes de synthèse ou de purification.

Toutes les étapes du traitement peuvent être automatisées. L'enregistrement des spectres RMN de manière automatique est aisément réalisable, notamment avec l'utilisation d'un passeur d'échantillon. L'étape limitante dans le processus demeure l'interprétation des spectres. L'accélération de cette étape passe donc par le développement d'outils informatiques d'aide à l'analyse structurale.

Ces logiciels sont développés avec la volonté d'imiter le comportement humain en tirant les avantages de l'utilisation d'un ordinateur : la rapidité de calcul et l'exhaustivité des solutions sans aucun parti pris. Ce dernier argument est très important si l'on considère toute la rigueur dont doit faire preuve un scientifique dans sa démarche. Il devrait convaincre toute personne réfractaire à l'utilisation d'un logiciel d'aide à l'analyse structurale. Un logiciel ne s'arrête pas à la première solution trouvée, il considère toutes les hypothèses. Le raisonnement d'un cerveau humain peut parfois être influencé par une solution à laquelle il s'attend et ne pas chercher à savoir si d'autres solutions existent. Pour preuve, de nombreuses structures décrites dans la littérature se sont révélées inexactes [1, 2].

1.1.3 Les différents outils pour l'analyse structurale automatique

Les outils d'aide à l'analyse structurale par RMN forment 3 familles principales qui possèdent des liens comme nous le verrons dans la suite du manuscrit.

Tout d'abord la prédiction des spectres RMN est certainement l'outil le plus utilisé. La prédiction peut être utilisée soit de manière indépendante par un expérimentateur, soit couplée à un logiciel d'analyse structurale. En effet, elle peut être utilisée afin de classer

les résultats obtenus par un générateur de structure ou afin de vérifier une structure par comparaison des déplacements chimiques.

Cela nous amène aux deux autres familles de programmes informatiques créés pour l'analyse structurale. Ce sont des systèmes dits experts [3–8], d'une part les logiciels d'élucidation structurale regroupés sous le terme anglais de CASE systems (l'acronyme CASE signifiant Computer-Assisted Structure Elucidation) et d'autre part les logiciels de vérification ASV (Automatic Structure Verification).

Les efforts consentis par de nombreux groupes de chercheurs dans le domaine vont être présentés dans la suite du manuscrit avec des réalisations qui ont suivi l'évolution technologique de la technique expérimentale ainsi que l'évolution de l'informatique. Cette description permettra d'introduire des concepts utiles à la compréhension des chapitres suivants.

1.2 Méthodes de prédiction des spectres RMN

1.2.1 Introduction

Pouvoir prévoir les propriétés des molécules, par exemple pour l'évaluation de nouveaux médicaments potentiels, peut constituer un avantage afin d'éviter de longues étapes de synthèse. À l'échelle plus modeste de la RMN, la prédiction des propriétés magnétiques des atomes permet de faire une comparaison entre un spectre simulé et un spectre expérimental pour valider ou non une structure. L'interprétation doit évidemment tenir compte de l'incertitude sur la prédiction.

Cette partie ne concerne que la prédiction des spectres ^{13}C et ^{1}H qui sont les constituants majoritaires des molécules organiques. Les méthodes de prédiction peuvent être appliquées à d'autres noyaux. On peut citer le ^{19}F, le ^{31}P ainsi que d'autres noyaux plus exotiques rencontrés dans les produits de synthèse.

Il existe plusieurs types de prédiction :

- les modèles élaborés à partir de bases de données ;
- les méthodes utilisant directement des bases de données ;
- les calculs quantiques.

1.2.2 Spectres RMN 1D

Les spectres RMN ^{1}H (figure 1.2) fournissent une triple information. Chaque signal est associé à trois paramètres : son déplacement chimique, sa multiplicité et son intégrale.

Le déplacement chimique, mesuré en ppm, représente la fréquence de résonance ν des noyaux. Cette fréquence dépend de l'environnement chimique du ou des protons correspondants. L'intégration permet de savoir combien de protons équivalents sont impliqués dans un signal. Par exemple les 3 protons d'un groupement méthyle CH_3 apparaissent

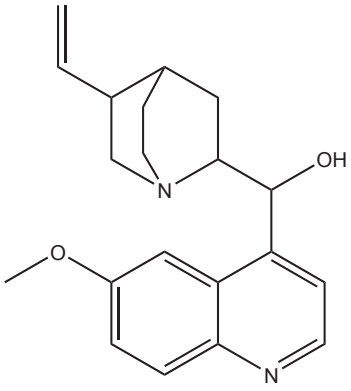


Figure 1.1 – Structure de la quinidine

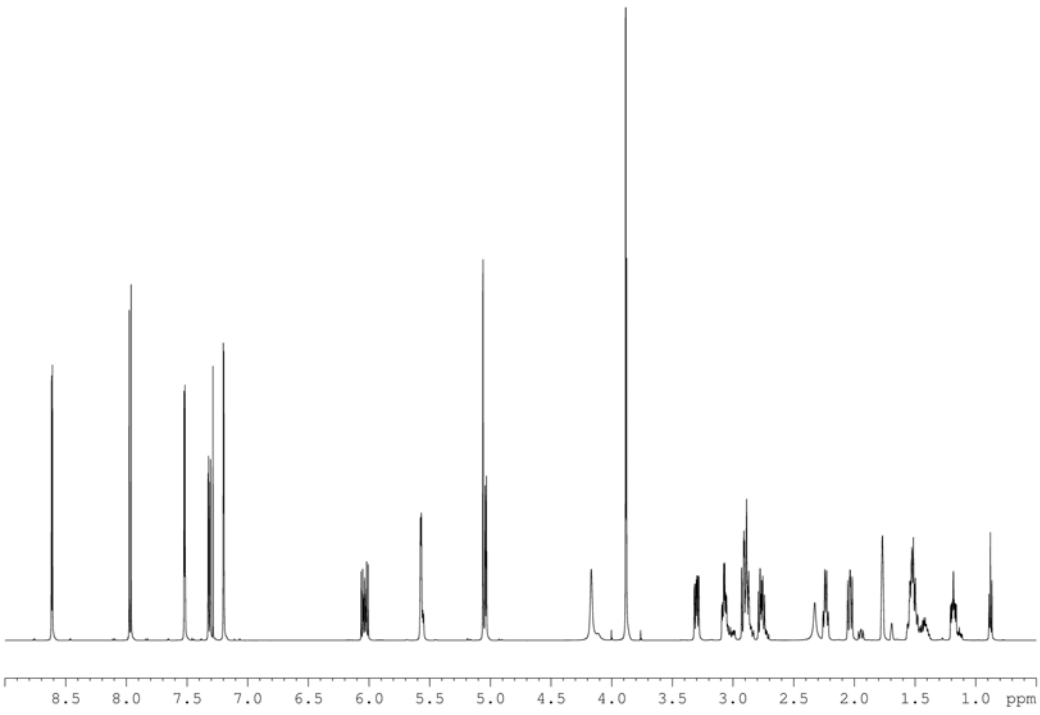


Figure 1.2 – Spectre RMN ^{1}H ($CDCl_3$, 600 MHz) de la quinidine

sur le spectre au même déplacement chimique avec une aire du signal trois fois plus importante que pour un proton CH. La multiplicité des signaux résulte du phénomène de couplage scalaire entre noyaux. Lorsque des noyaux sont couplés, on observe un éclatement du signal de ces noyaux en plusieurs raies d'intensités relatives et d'espacements caractéristiques. L'espacement est égal à la valeur de la constante de couplage J mesurée en Hertz. Lorsque la différence des fréquences des noyaux couplés est forte devant la valeur de la constante de couplage J, on est en présence d'un spectre au premier ordre où les constantes de couplage sont facilement mesurables. Si ce n'est pas le cas, on observe un spectre dit au second ordre dont la forme des multiplets est difficile à interpréter et où il est difficile d'extraire les valeurs des constantes de couplage.

Les spectres ^{13}C (figure 1.3) sont généralement enregistrés de manière à ce qu'on n'observe pas le couplage ^{13}C –^{1}H. Les signaux ne sont donc pas éclatés. Ils sont uniquement caractérisés par leurs déplacements chimiques.

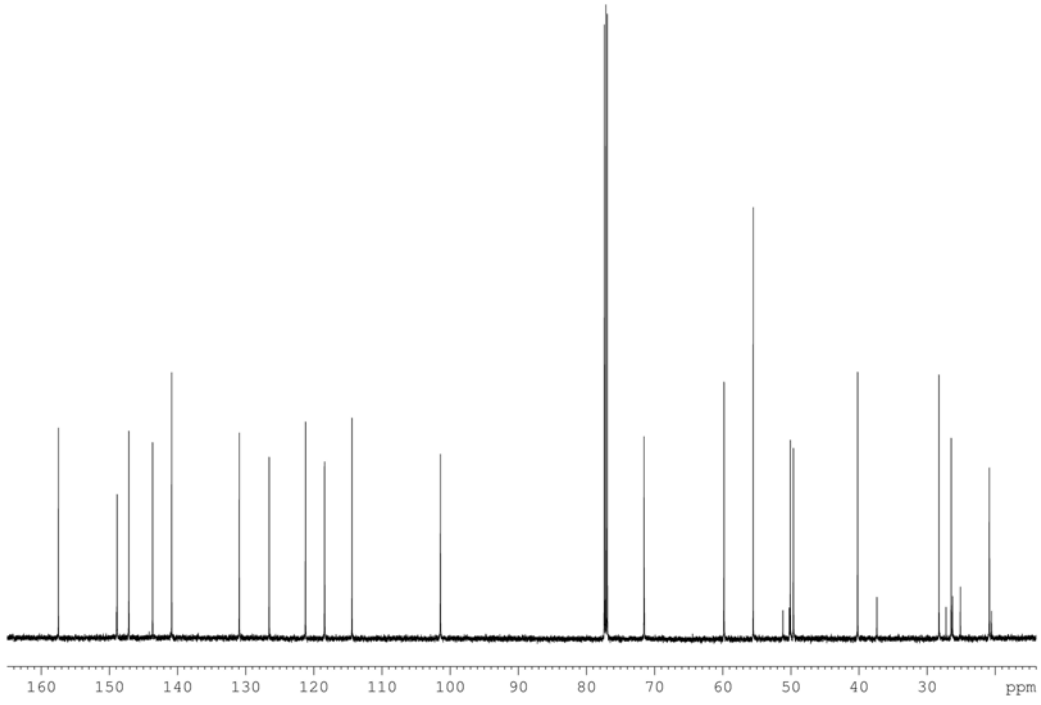


Figure 1.3 – Spectre RMN ^{13}C ($CDCl_3$, 600 MHz) de la quinidine

Les spectres DEPT (Distortionless Enhancement by Polarization Transfer) [9] permettent de déterminer la multiplicité des carbones. Le plus utilisé est le spectre DEPT-135 (figure 1.4) sur lequel les signaux CH et CH_3 sont de signe opposé aux signaux CH_2.

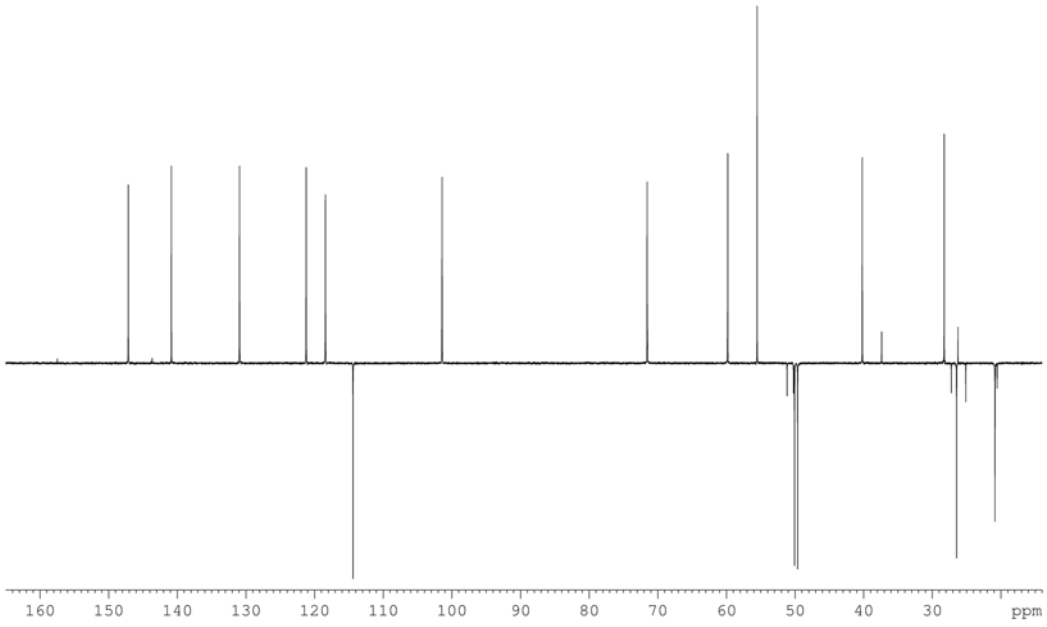


Figure 1.4 – Spectre RMN DEPT ($CDCl_3$, 600 MHz) de la quinidine

1.2.3 Prédiction des déplacements chimiques ^{13}C

Les atomes de carbone constituent le squelette des molécules organiques. L'utilisation de la prédiction des déplacements chimiques ^{13}C est donc une méthode efficace pour l'analyse d'une structure. Un spectre RMN ^{13}C réalisé dans des conditions de découplage du ^{1}H et en l'absence d'autres noyaux ayant des propriétés magnétiques, est constitué de simples résonances. Chaque atome de carbone d'une molécule est associé à un signal ne possédant qu'une propriété : le déplacement chimique.

1.2.3.1 Règles d'additivité

Les pionniers de la prédiction des déplacements chimiques ^{13}C sont Grant et Paul. Leur méthode proposée en 1963-64 repose sur des règles d'additivité déterminées empiriquement [10, 11]. Elle a été appliquée à la famille des alcanes en utilisant des valeurs d'incréments représentant les effets de l'environnement de l'atome de carbone sur son déplacement chimique jusqu'à 5 atomes autour de l'atome central en question. Le déplacement chimique est calculé à l'aide de la relation suivante :

$$\delta_C(i) = B + \sum_j A_j n_{ij}$$

où $\delta_C(i)$ est le déplacement chimique du i-ième atome de carbone, B est une constante égale au déplacement chimique du méthane, n_{ij} est le nombre d'atomes de carbone séparés par j liaisons du i-ième carbone et A_j est le paramètre additionnel attribué aux carbones

en position j. Les valeurs des paramètres A_j sont obtenues par régression linéaire sur des valeurs de déplacements chimiques d'alcanes standards.

D'autres modèles ont ensuite été construits pour d'autres familles de composés prises individuellement. Les résultats de ces modèles incrémentiels étant assez satisfaisants, Clerc et Sommerauer ont développé un programme capable de prédire les déplacements chimiques ^{13}C de molécules appartenant à des familles variées [12]. Le principe étant de choisir le modèle linéaire le plus adapté pour chaque atome de carbone de la molécule afin de calculer la valeur du déplacement chimique. Cette approche était toutefois assez limitée car la prédiction ne pouvait être réalisée que sur des atomes de carbone hybridés sp^3 et les valeurs d'incréments ne pouvaient pas être modifiées par l'utilisateur.

Ces limitations et restrictions ont par la suite été levées par Fürst et Pretsch avec le logiciel C13Shift [13–16]. Leur travail est basé sur de larges bases de données permettant de construire des modèles de prédiction pour des classes de molécules très diverses. Le programme est capable d'effectuer des calculs à partir de paramètres définis pour des atomes de carbone hybridés sp^3 et pour des carbones hybridés sp^2 et sp. Il peut également prendre en compte des paramètres supplémentaires donnés par l'utilisateur pour gagner en précision lors de la prédiction. Les valeurs d'incréments peuvent être modifiées et de nouvelles règles d'additivité ajoutées. L'algorithme de prédiction est appliqué à chaque atome de carbone de la structure. Il analyse et détecte le type de carbone, puis sélectionne et applique automatiquement le modèle de règles d'additivité le plus approprié.

Une autre approche a été décrite par Chen et Robien [17]. Leur logiciel OPSI (Optimized Prediction of ^{13}C - NMR Spectra using Increments) corrige l'inconvénient des autres programmes qui possèdent des tables de règles définies de manière statique. OPSI n'utilise pas de paramètres prédéfinis, ils sont déterminés de manière dynamique à chaque nouvelle prédiction. Le programme divise la structure cible en une série de sous-structures. Une recherche par similarité est ensuite effectuée dans une base de données pour chaque sous-structure. Les résultats de cette recherche sont utilisés pour modéliser des règles utilisant des paramètres additionnels. Les modèles obtenus sont finalement utilisés pour le calcul des déplacements chimiques de la structure donnée en entrée.

1.2.3.2 Fragments centrés sur un atome

Cette autre méthode repose aussi sur de larges bases de données et a été mise au point toujours en se basant sur l'idée que le déplacement chimique dépend de l'environnement de l'atome de carbone.

L'implémentation la plus connue et la plus utilisée de cette idée est le système des codes HOSE (Hierarchically Ordered Spherical description of Environment) présenté par Bremser en 1978 [18, 19]. Les codes HOSE permettent de caractériser l'environnement de chaque atome. La figure 1.5 montre comment le codage est réalisé. Les atomes voisins de l'atome central sont classés par ordre de priorité à l'intérieur de sphères dont le nombre

peut être variable.

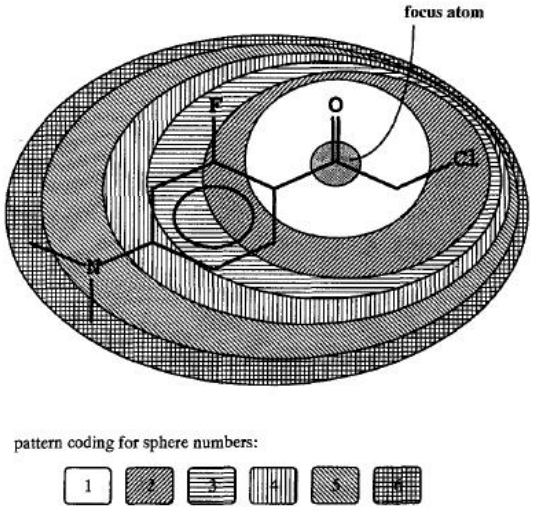


Figure 1.5 – Codage des codes HOSE

À partir d'une base de molécules associées à leur spectre RMN ^{13}C, un algorithme extrait les codes HOSE de tous les atomes de carbone. Ces codes sont regroupés dans un dictionnaire de codes HOSE. Chacun des codes est couplé à une valeur de déplacement chimique. Cette valeur est la moyenne des déplacements chimiques de tous les carbones enregistrés dans la base de données possédant un code HOSE identique. La prédiction des déplacements chimiques se déroule ensuite en plusieurs étapes. Tout d'abord les atomes de carbone de la structure cible sont identifiés et le code HOSE de chacun d'eux est déterminé. Ces codes HOSE sont ensuite recherchés dans le dictionnaire de codes afin d'en extraire une valeur de déplacement chimique. Lorsque que le code est absent du dictionnaire, des codes similaires sont choisis pour calculer une estimation du déplacement chimique.

Le défaut de ce système de codage est l'absence d'information sur la géométrie des molécules, ce qui peut affecter la précision de la prédiction. Schutz *et al* ont tenté de corriger ce défaut par l'introduction de descripteurs 3D pour modifier le codage des codes HOSE [20].

D'autres systèmes de codage incluant la géométrie de la structure dans le codage de l'environnement des atomes, avec des niveaux de complexité différents, ont été proposés par Gray *et al* [21] puis par Gastmans *et al* [22]. Small a décrit un système où l'environnement des carbones est codé sous forme de vecteur tenant compte de la topologie et de la géométrie de la molécule [23]. Si la similarité entre un vecteur de la molécule cible et ceux de la base de données n'est pas assez importante, la prédiction est effectuée à l'aide d'un modèle linéaire. Celui-ci est construit en utilisant des descripteurs pour les propriétés topologiques, géométriques et électroniques de l'environnement de l'atome de carbone.

Les auteurs du programme de la société ACD/Labs ont reporté un système de prédiction des déplacements chimiques basé sur une méthode utilisant un codage des environ-

nements des atomes [8]. Les détails concernant le codage et l'algorithme de prédiction ne sont toutefois pas communiqués.

1.2.3.3 Réseaux de neurones artificiels

L'architecture des réseaux de neurones artificiels peut être comparée à un modèle informatique simplifié du cerveau humain. Il s'agit en effet de plusieurs couches de neurones qui envoient des signaux à d'autres neurones en fonction du signal reçu en entrée. D'une manière générale, les réseaux de neurones artificiels permettent de construire des modèles empiriques de systèmes pour lesquels les règles de dépendance entre une donnée d'entrée et une donnée de sortie sont très complexes. Les modèles sont obtenus par un entrainement du réseau de neurones à l'aide de données de référence. Un réseau de neurones, une fois entrainé, est capable de fournir un signal de sortie pour un signal d'entrée absent des données d'entrainement. Le résultat ne nécessite pas de longs temps de calcul, il est obtenu de manière instantanée.

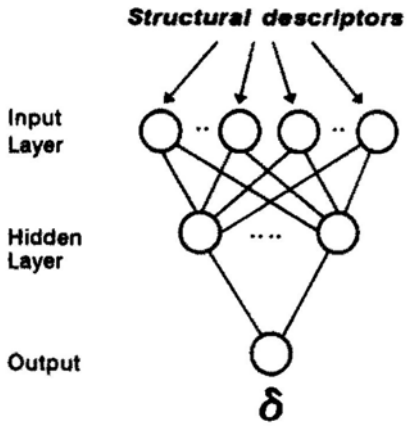


Figure 1.6 – Schéma d'un réseau de neurones artificiels

Afin de prédire des déplacements chimiques en utilisant des réseaux de neurones artificiels, il est nécessaire de décrire le type des atomes et l'environnement chimique de chaque atome pris individuellement. Il faut donc définir et optimiser un certain nombre de descripteurs pour les atomes de carbone. Les descripteurs servent de données d'entrée aux réseaux qui fournissent une valeur de déplacement chimique en sortie (figure 1.6).

La prédiction des spectres RMN ^{13}C à l'aide de réseaux de neurones artificiels a été réalisée pour de nombreuses familles de composés. On peut citer les travaux de Kvasnicka pour les alcanes [24], de Anker et Jurs pour les stéroïdes [25], de Doucet *et al* pour les alcanes [26], de Kvasnicka *et al* pour les benzènes monosubstitués [27], de Panaye *et al* pour les cyclohexanes substitués par des groupements méthyles [28], de Miyashita *et al* pour des composés halogénés [29], de Clouser et Jurs pour les tétrahydropyranes [30], de Mitchell et Jurs pour les monosaccharides [31], de Clouser et Jurs pour les ribonucléosides

[32] et de Ivanciuc *et al* pour les carbones hybridés sp^2 et sp^3 dans les alcènes non cycliques [33, 34].

Meiler *et al* ont développé le programme C_SHIFT permettant une prédiction rapide des déplacements chimiques en utilisant des réseaux de neurones artificiels [35, 36]. L'efficacité du programme a été prouvée avec des tests sur des substances naturelles complexes.

La société ACD/Labs a également testé la performance des réseaux de neurones artificiels et a démontré qu'ils permettent une estimation précise et rapide des déplacements chimiques [37–39].

1.2.3.4 Calculs quantiques

Les calculs quantiques permettent le calcul des déplacements chimiques ^{13}C. Ils ne peuvent être réservés qu'à des petits systèmes car les temps de calcul sont dépendants du nombre d'atomes et peuvent prendre plusieurs heures.

La révision de la structure de l'hexacyclinol par Rychnovsky est un bon exemple d'application de calculs des déplacements chimiques [40]. La figure 1.7 montre la première structure proposée (à gauche) et la structure révisée (à droite).

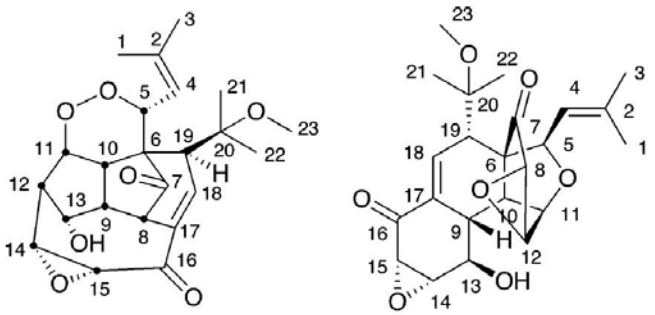


Figure 1.7 – Structures successivement proposées pour l'hexacyclinol

Bagno et Saielli présentent également des résultats très intéressants concernant la prédiction de spectres de substances naturelles complexes [41].

1.2.4 Prédiction des spectres ^{1}H

Les méthodes décrites pour la prédiction des déplacements chimiques ^{13}C ont également été appliquées à la prédiction de spectres ^{1}H. Le principe des méthodes est inchangé, nous passerons donc sur les détails en ne présentant que les applications. La difficulté du problème est cependant accrue car la simulation d'un spectre ^{1}H nécessite une estimation de la valeur des déplacements chimiques et des valeurs des constantes de couplage. De plus les déplacements chimiques ^{1}H sont très dépendants des conditions expérimentales (solvant, température, concentration, pH et effet de sel) et de la géométrie de la molécule.

La complexité de la tâche explique le nombre plus restreint de programmes développés dans ce sens.

1.2.4.1 Règles d'additivité

Cette méthode permet le calcul des déplacements chimiques mais ne permet pas d'avoir accès aux valeurs des constantes de couplage.

Dès 1967, Reed a décrit une série de paramètres pour la prédiction des déplacements chimiques ^{1}H des benzènes substitués [42].

Des modèles concernant de nombreuses familles de molécules ont été mis au point. Les paramètres des règles d'additivité ont été généralisés par Schaller et Pretsch en 1994-95 [43, 44].

1.2.4.2 Fragments centrés sur un atome

Le programme de prédiction de la société ACD/Labs est capable de simuler un spectre ^{1}H complet en estimant les déplacements chimiques et les constantes de couplage [8]. Les auteurs indiquent que la base de données contient des fragments associés à des déplacements chimiques ainsi que des couples de fragments associés à des valeurs de constantes de couplage. Les détails de l'algorithme de prédiction ne sont pas communiqués.

1.2.4.3 Réseaux de neurones artificiels

Les réseaux de neurones artificiels ne sont utilisés que pour la prédiction des déplacements chimiques. Des applications tenant compte des effets de la géométrie de la structure ont été présentées par Aires-de-Sousa *et al* [45] et Binev *et al* [46, 47].

1.2.4.4 Calculs quantiques

On dispose de méthodes permettant de calculer à la fois les déplacements chimiques et les constantes de couplage à partir d'une structure 3D. Bagno a décrit une méthode complète de calcul des spectres de RMN ^{1}H [48]. Après calcul des paramètres, un spectre complet peut être reconstruit à l'aide des valeurs calculées. Les calculs quantiques ont cet avantage de pouvoir tenir compte du rapport $\Delta\nu/J$ et donc de pouvoir simuler des spectres avec des couplages forts (spectres au second ordre).

1.2.5 Conclusion

Une valeur de déplacement chimique prédite ne doit pas être prise comme un résultat auquel on peut accorder une confiance aveugle. Toute valeur de déplacement chimique doit être accompagnée d'une incertitude sur la prédiction. L'utilisateur d'un logiciel de prédiction doit absolument tenir compte de l'erreur de la méthode dans son interprétation du spectre simulé.

Un résumé des différentes méthodes de prédiction des déplacements chimiques, ainsi que leurs avantages et inconvénients respectifs, est présenté en figure 1.8. La comparaison de la précision entre les méthodes est difficile car le résultat est essentiellement dépendant des données initiales utilisées pour mettre en œuvre la prédiction.

Méthode	**Incréments**	**Fragments**	**Réseaux de neurones artificiels**	**Calculs quantiques**
Précision	−	+	+	+
Rapidité	+	+	+	−
Modularité	Ajout possible de nouveaux modèles et modifications des paramètres possibles	Ajouts de fragments possibles dynamiquement	Pas d'ajout dynamique d'informations supplémentaires	Pas de modifications des paramètres
Accès aux détails de la prédiction	Détails visibles	Détails visibles	Détails invisibles	Détails invisibles

Figure 1.8 – Tableau récapitulatif des méthodes de prédiction des spectres RMN

La modularité des logiciels est importante afin de pouvoir gagner en précision. Un chimiste habitué à travailler sur le même type de molécules doit pouvoir créer sa propre base de données et l'utiliser pour effectuer des prédictions. L'ajout de nouvelles données dans un réseau de neurones est possible mais le réseau doit être à nouveau entrainé avec toutes les données déjà utilisées.

L'utilisateur peut juger intéressant d'avoir accès au détail de la prédiction, par exemple pour comprendre les raisons d'une forte incertitude sur une valeur de déplacement chimique. La majorité des programmes indiquent sur quelles règles ou quelles données ils se sont basés pour effectuer l'estimation. Du point de vue de l'utilisateur, les réseaux de neurones artificiels et les calculs quantiques fonctionnent comme une boite noire. Les règles qui gouvernent la prédiction sont donc inaccessibles de l'extérieur.

La prédiction des déplacements chimiques peut être utilisée de manière indépendante ou couplée à un logiciel d'analyse structurale. Ces logiciels, décrits par la suite, utilisent la prédiction à des fins différentes. Le point commun étant une volonté de pouvoir effectuer une prédiction la plus précise possible en un minimum de temps possible.

1.3 Élucidation structurale automatique

1.3.1 Introduction

Établir la structure d'une substance inconnue passe par l'interprétation de données provenant d'un ensemble de techniques d'analyse (spectres de masse, RMN, UV, IR...). L'obtention des solutions d'un problème de détermination de structure peut être facilitée par les logiciels d'élucidation structurale.

Les premiers logiciels ou systèmes experts ont été développés dans les années 1970, puis ont connu un essor parallèlement à l'évolution des techniques d'analyse. Ils étaient d'abord basés sur l'analyse des spectres RMN 1D ^{1}H et ^{13}C combinés aux spectres de masse et IR. Ils manquaient cependant d'efficacité car peu d'informations peuvent être extraites des spectres RMN 1D au-delà d'un certain niveau de complexité. L'arrivée de la RMN à deux dimensions a donc entrainé l'émergence de systèmes plus performants dans les années 1990. Les spectres RMN 2D sont très riches en informations comme nous le verrons par la suite.

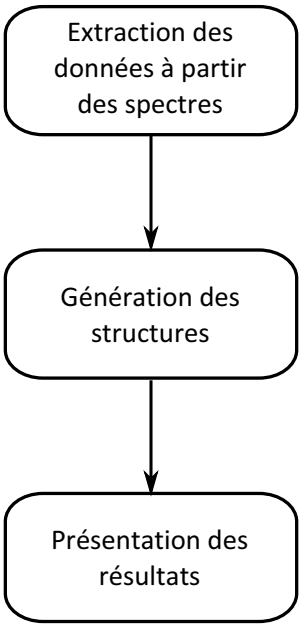


Figure 1.9 – Composition d'un logiciel d'aide à l'élucidation structurale

La composition des logiciels d'élucidation structurale est présentée en figure 1.9. Plusieurs parties successives entrent en jeu. Dans un premier temps l'interface avec l'utilisateur permet de fournir les données au logiciel. L'extraction des données peut être automatique ou manuelle. La seconde partie est l'algorithme de résolution qui est le cœur du logiciel. Cette partie est importante car un logiciel d'élucidation structurale se doit de fournir des solutions de façon exhaustive et sans redondance. Finalement les solutions sont

présentées à l'utilisateur. Un logiciel peut être doté d'un module de prédiction des déplacements chimiques pour faire un classement des solutions par ordre de vraisemblance. Ce classement est obtenu par comparaison du spectre expérimental avec les spectres prédits pour les différentes solutions.

L'ensemble des logiciels présentés par la suite ne forme pas une liste exhaustive des logiciels développés pour l'analyse structurale automatique. Il s'agit de présenter les différentes approches dans la manière d'aborder un problème de génération de structures.

1.3.2 Spectres RMN 2D utilisés en analyse structurale

Les spectres RMN à deux dimensions permettent l'observation de corrélations traduisant un couplage entre deux atomes. Les spectres apportent des informations différentes et complémentaires pour la détermination des structures.

Un résumé des expériences de RMN 2D utilisées par les logiciels d'élucidation structurale est présenté en figure 1.10.

Le spectre COSY (COrrelation SpectroscopY) [49] permet d'observer des corrélations entre déplacements chimiques ^{1}H impliquant des couplages scalaires $^{n}J_{H-H}$. Le nombre n est la longueur du chemin de couplage, c'est-à-dire le nombre de liaisons séparant les protons couplés. Les valeurs de n typiquement observées sont $n = 2$ (protons géminés) et $n = 3$ (protons vicinaux). On observe dans des cas particuliers des couplages à plus longue distance avec $n = 4$, voire $n = 5$.

L'expérience INADEQUATE (Incredible Natural Abundance DoublE QUAnTum Experiment) [50] permet de mettre en évidence des couplages J_{C-C}. Cette technique permet d'établir les connectivités entre carbones et donc d'avoir un accès direct au squelette carboné d'une molécule. C'est une technique très puissante du point de vue de l'analyse structurale mais comme son nom l'indique elle est très peu utilisée car elle comporte un inconvénient majeur. L'atome de ^{13}C a une abondance isotopique d'environ 1%. La chance d'avoir une molécule possédant deux atomes de ^{13}C est donc de 1 sur 10000. Cette très faible probabilité a pour conséquence la très faible sensibilité de l'expérience INADEQUATE.

Les spectres HSQC (Heteronuclear Single Quantum Correlation) [51] mettent en jeu les couplages $^{1}J_{C-H}$. On observe donc des corrélations entre les déplacements chimiques des ^{1}H et des ^{13}C ou des ^{15}N directement liés.

L'expérience HMBC (Heteronuclear Multiple Bond Correlation) [52] corrèle les déplacements chimiques des ^{1}H et des ^{13}C ou des ^{15}N séparés par plus de 1 liaison. Les spectres HMBC ^{1}H –^{15}N sont notamment utilisés pour l'étude des alcaloïdes [53]. Les couplages $^{n}J_{X-H}$ sont observés principalement pour $n = 2$ et $n = 3$, mais n peut être supérieur à 3 si certaines conditions structurales sont réunies (systèmes de liaisons multiples conjuguées). L'analyse de ce type de spectre engendre une ambigüité sur la longueur du chemin de couplage. Lors de la détermination d'une structure, il faut donc faire des hypothèses sur

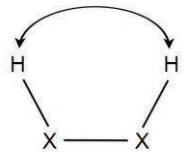
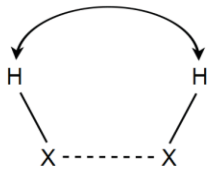
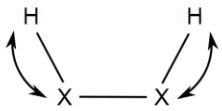
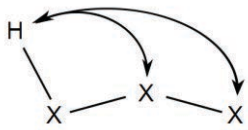
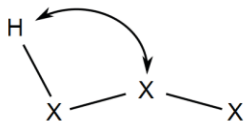
	Expérience 2D	Couplage typiquement observé	
Homonucléaire	COSY ^{1}H-^{1}H		Couplage scalaire ^{1}H-^{1}H à travers 2 ou 3 liaisons
	NOESY ^{1}H-^{1}H		Couplage dipolaire ^{1}H-^{1}H à travers l'espace (distance < 6 Å)
	INADEQUATE ^{13}C-^{13}C		Couplage scalaire ^{13}C-^{13}C direct
Hétéronucléaire	HSQC ^{1}H-^{13}C ou ^{1}H-^{15}N		Couplage scalaire ^{1}H-^{13}C ou ^{1}H-^{15}N à travers 1 liaison
	HMBC ^{1}H-^{13}C ou ^{1}H-^{15}N		Couplage scalaire ^{1}H-^{13}C ou ^{1}H-^{15}N à travers 2 ou 3 liaisons
	H2BC ^{1}H-^{13}C		Pseudo couplage ^{1}H-^{13}C à travers 2 liaisons (combinaisons de couplages scalaires homo- et hétéronucléaires)
	1,1-ADEQUATE ^{1}H-^{13}C		

FIGURE 1.10 – Résumé des expériences de RMN 2D

cette longueur de chemin de couplage. Deux expériences tentent de lever cette ambigüité : l'ADEQUATE et l'H2BC.

L'expérience 1,1-ADEQUATE (Adequate sensitivity DoublE-QUAntum spEctroscopy) [54] permet de corréler sélectivement deux atomes de carbone directement liés avec les protons portés par l'un des deux carbones. Elle permet de distinguer les corrélations dues aux couplages $^2J_{C-H}$ de celles dues aux couplages $^3J_{C-H}$ observées dans les spectres HMBC. Le résultat final de l'expérience montre un pseudo couplage $^2J_{C-H}$ mais en réalité cette expérience repose sur un couplage $^1J_{C-H}$ et un couplage $^1J_{C-C}$.

L'expérience H2BC (Heteronuclear 2-Bond Correlation) [55] a pour but l'observation unique des couplages $^2J_{C-H}$ sur un spectre de type HMBC. Cette expérience est complètement indépendante des couplages $^nJ_{C-H}$ à longue distance car elle repose sur des couplages $^1J_{C-H}$ et $^3J_{H-H}$. Les corrélations observées ne concernent donc que les atomes de carbone protonés. Aucune information sur le voisinage des carbones quaternaires ne peut être obtenue.

Les spectres NOESY (Nuclear Overhauser Effect SpectroscopY) [56] ou ROESY (Rotating-frame Overhauser Effect SpectroscopY) [57] permettent d'observer des couplages dipolaires entre noyaux ^{1}H (couplage à travers l'espace). Le phénomène observé est l'effet Overhauser nucléaire (nOe). Cet effet varie selon une loi en $1/d^6$ et n'est plus observable au-delà d'une distance d entre les noyaux supérieure à 6 Å. Les corrélations fournissent des informations exploitables pour définir des conformations relatives dans des systèmes rigides.

1.3.3 L'élucidation structurale : stratégie générale

Avant d'aborder le fonctionnement d'un logiciel d'élucidation structurale automatique, il est nécessaire de revenir sur la stratégie couramment employée. En effet, la plupart des logiciels d'aide ne sont que des outils qui cherchent à imiter le raisonnement humain pour le processus de résolution. Il n'existe pas que des logiciels déterministes, d'autres logiciels utilisent des méthodes stochastiques. Le but commun étant d'apporter un gain de temps par l'automatisation.

Tout d'abord une formule brute doit être déduite d'une combinaison d'informations (spectre de masse, spectres RMN ^{1}H, ^{13}C et DEPT). Le spectre ^{13}C donne le nombre d'atomes de carbone. Le spectre DEPT donne le nombre d'hydrogènes portés par les carbones car il permet de discriminer les carbones en fonction de leur multiplicité. Les hétéroatomes (oxygènes, azotes...) sont déduits par analyse des déplacements chimiques ^{13}C et ^{1}H.

Les spectres IR et UV peuvent apporter des informations sur la présence de groupements caractéristiques de même que les valeurs de déplacement chimique des spectres ^{13}C et ^{1}H.

Les structures sont assemblées à partir des groupements identifiés et de fragments

de structure construits à l'aide des informations de corrélations des spectres RMN 2D. L'analyse des spectres RMN 2D permet de déduire des distances interatomiques à travers les liaisons dans le but d'établir des connexions entre atomes. L'objectif étant dans un premier temps de construire une structure plane.

Les spectres RMN 2D peuvent parfois être difficiles à interpréter car ils sont accompagnés d'ambigüités liées à la nature des expériences et à la résolution spectrale imposée par l'appareillage. Les spectres HMBC apportent quelquefois des corrélations à très longue distance, de même que les spectres COSY. De plus il peut exister une ambigüité lorsque deux carbones ou deux protons ont des déplacements chimiques très voisins.

Lorsque plusieurs solutions sont possibles, les spectres expérimentaux sont comparés soit à la littérature, soit à une base de données, soit à des spectres prédits pour choisir la plus plausible.

La structure 3D de la molécule est finalement déterminée par l'analyse des valeurs des constantes de couplage et par les données issues des spectres NOESY.

1.3.4 Stratégies des logiciels de détermination structurale

Les logiciels utilisant uniquement les spectres RMN 1D effectuent une analyse des déplacements chimiques pour déduire des sous-structures qui sont assemblées afin de construire des structures.

Les logiciels utilisant les spectres RMN 2D suivent la stratégie décrite dans le paragraphe précédent en tenant compte des ambigüités et en envisageant toutes les hypothèses.

1.3.5 Logiciels basés sur les spectres RMN 1D

1.3.5.1 X-PERT

Le programme X-PERT [58–60] utilise la combinaison des spectres IR, RMN ^{1}H et ^{13}C de la substance inconnue. Il possède trois bases de données de fragments associés à des valeurs de fréquences IR et de déplacements chimiques ^{1}H et ^{13}C.

La formule brute est déterminée à partir des spectres de masse, IR et RMN. La première étape de la génération des structures est une recherche de fragments par comparaison des spectres expérimentaux avec les bases de données. L'étape suivante est la constitution de tous les isomères en combinant les fragments obtenus de toutes les manières possibles avec pour contrainte la formule brute.

Ce programme reste limité à des structures possédant 20 atomes de squelette (hydrogènes non compris).

1.3.5.2 SpecSolv

SpecSolv [61] est un système qui est une extension de SpecInfo [62], un projet de base données et d'interprétation de spectres RMN, IR et de masse. SpecInfo contient plus de

200000 spectres RMN ^{13}C attribués. SpecSolv a la capacité de résoudre des structures en n'utilisant que les informations des spectres ^{13}C et DEPT (extraction des déplacements chimiques et multiplicités des carbones) et sans avoir connaissance de la formule brute ou de la masse moléculaire. Il possède une librairie de fragments associés à leurs sous-spectres. Cette librairie a été créée à partir de la base de données de SpecInfo. Elle contient plus de 400000 fragments codés sous la forme de codes HOSE (3 sphères par rapport à l'atome de carbone central) et plus de 100000 fragments codés sous la forme de codes HOSE (2 sphères).

La résolution commence par une recherche des fragments potentiels. Tous les sous-spectres de la librairie sont comparés au spectre ^{13}C expérimental afin d'extraire les fragments susceptibles de se trouver dans la structure. Les fragments sont classés en fonction de leur degré de correspondance avec le spectre expérimental. Les fragments les mieux classés sont utilisés pour être assemblés afin construire tous les isomères possibles.

Les auteurs de SpecSolv ont annoncé un taux de réussite de 80%.

1.3.5.3 ACD Structure Elucidator 1D

ACD Structure Elucidator 1D [63] intègre la stratégie de reconnaissance des fragments présente dans X-PERT. Il possède une librairie de plus de 1,5 millions de fragments et leurs sous-spectres ^{13}C associés. Cette librairie est construite à partir d'une base de données de plus de 215000 structures associées aux déplacements chimiques ^{13}C.

Le programme possède deux modes d'élucidation : le mode standard et le mode classique. La figure 1.11 présente le fonctionnement du programme.

Le mode standard est basé sur la même stratégie que le système SpecSolv. Les structures sont construites à partir de fragments classés par ordre décroissant d'atomes constituant le fragment et par ordre croissant de déviation moyenne entre le sous-spectre et le spectre expérimental. Un filtrage est possible en utilisant les spectres IR et ^{1}H conjointement à des librairies de fragments caractéristiques. En mode standard, les structures sont assemblées en associant les fragments de manière à superposer les atomes présents dans ces fragments.

En cas d'échec, c'est-à-dire pas de solution, le programme passe en mode classique. L'assemblage des structures est réalisé en ne superposant pas les fragments. Lorsque la formule brute est inconnue, le programme détermine une série de formules brutes probables à partir de la masse moléculaire et des données spectrales.

Si le programme échoue à nouveau, il est possible d'utiliser un mode pas à pas où l'utilisateur guide le programme dans la résolution (choix des fragments, et façon de les assembler).

Les auteurs annoncent un succès du programme dans 90% des cas.

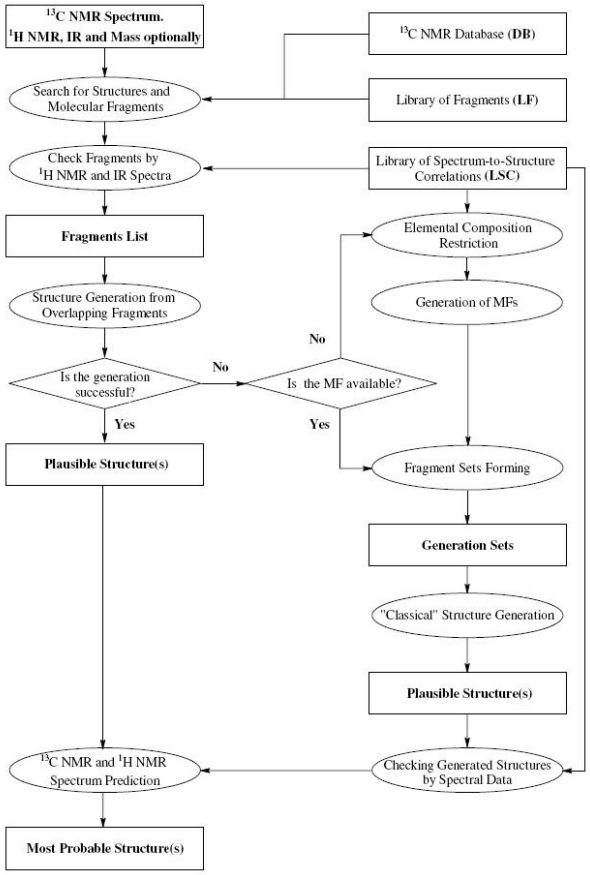


FIGURE 1.11 – Schéma de fonctionnement du programme ACD Structure Elucidator 1D

1.3.5.4 GENIUS

GENIUS [64, 65] se distingue des logiciels présentés précédemment car il utilise un algorithme génétique pour la résolution des structures. Il constitue une alternative à l'approche classique de l'assemblage de fragments.

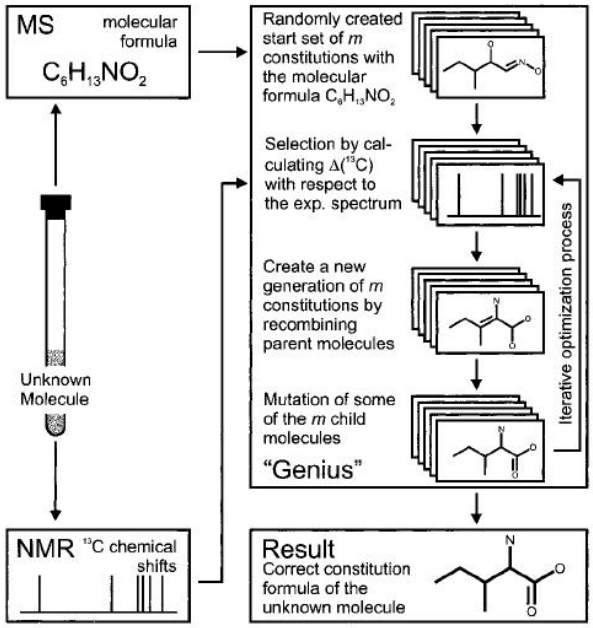


Figure 1.12 – Fonctionnement du programme GENIUS

La figure 1.12 résume le fonctionnement du programme GENIUS. Les données d'entrée sont un spectre RMN ^{13}C et une formule brute obligatoire. La formule brute est utilisée pour générer aléatoirement une série de structures. L'algorithme effectue une évaluation de ces structures par comparaison du spectre ^{13}C expérimental avec un spectre prédit. Les meilleures structures candidates sont sélectionnées pour l'étape de recombinaison. De nouvelles structures sont générées lors de cette étape par des modifications apportées aux structures parentes. Les cycles de recombinaison/évaluation sont répétés jusqu'à ce que la différence des déplacements chimiques soit minimisée.

Le degré de similarité entre un spectre expérimental, constitué de N signaux, et un spectre prédit est déterminé par la relation :

$$\Delta\left(^{13}C\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\delta^i_{calc}\left(^{13}C\right) - \delta^i_{exp}\left(^{13}C\right)\right)^2}$$

avec δ^i_{calc} et δ^i_{exp} respectivement les déplacements chimiques prédit et expérimental du i-ième atome de carbone.

Si la multiplicité des carbones peut être déterminée expérimentalement (par exemple

à l'aide d'un spectre DEPT), la fonction d'évaluation peut en tenir compte :

$$\Delta\left({}^{13}C\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left|\delta^{i}_{calc}\left({}^{13}C\right) - \delta^{i}_{exp}\left({}^{13}C\right)\right| + MDF\left|M^{i}_{calc} - M^{i}_{exp}\right|\right)^{2}}$$

avec MDF un facteur de correction, M^{i}_{calc} et M^{i}_{exp} respectivement les multiplicités prédite et expérimentale du i-ième atome de carbone.

L'expérience a montré qu'un algorithme génétique ne peut travailler efficacement qu'avec des structures allant jusqu'à 15 atomes lourds (atomes d'hydrogène exclus).

1.3.6 Logiciels basés sur les spectres RMN 2D

D'une manière générale, l'expérience a pu montrer que l'élucidation structurale assistée par ordinateur de structures complexes n'est possible que lorsque des données de RMN 2D sont disponibles. Un spectre RMN ^{13}C ne contient pas assez d'informations pour constituer l'unique source d'entrée d'un logiciel d'élucidation structurale automatique.

1.3.6.1 SESAMI

SESAMI (Systematic Elucidation of Structure Applying Machine Intelligence) [66–69] est le successeur du système CASE [70–73] de Munk. Il est composé d'un module d'interprétation des spectres RMN 2D INTERPRET2D et de générateurs de structures COCOA [74] et HOUDINI [75, 76] (figure 1.13).

SESAMI accepte en entrée des spectres de RMN 1D ^{1}H, ^{13}C et des spectres de RMN 2D COSY, HSQC, HMBC et INADEQUATE. Il a également besoin de la formule brute de la substance inconnue. L'utilisateur doit analyser les spectres et fournir la liste de tous les atomes présents dans la structure. Des éventuelles superpositions de signaux ^{13}C doivent être détectées et l'utilisateur doit déterminer si la superposition est fortuite ou due à une symétrie dans la molécule. En effet si le nombre de signaux ^{13}C est inférieur au nombre de carbones de la formule brute, le programme considère la présence d'une symétrie dans la structure. Les données de corrélation provenant des spectres RMN 2D sont entrées sous la forme de paires de signaux. La longueur du chemin de couplage, c'est-à-dire le nombre de liaisons entre les atomes peut être entrée comme un nombre exacte ou sous la forme d'un intervalle.

La première version de SESAMI (SESAMI-C) utilise le générateur de structure COCOA. Un certain nombre de fragments constituants potentiels de la structure sont déduits des données spectrales. Les fragments servent d'unités de base pour l'assemblage des structures par le générateur COCOA. L'interprétation des spectres est réalisée par le module INTERPRET qui est constitué de deux algorithmes d'interprétation PRUNE et INFER. Les fragments (par exemple : =CH–CH_2–O–) sont centrés sur un atome et sont constitués de groupes définis par un élément, le nombre d'hydrogènes liés et les liaisons libres qui

servent à relier les groupes entre eux (par exemple : $-CH_2-$, $=O$). PRUNE détermine une série de fragments pour la construction des structures en supprimant les fragments incompatibles avec la formule brute et les données provenant des spectres ^1H, ^{13}C et 2D. L'algorithme INFER déduit des sous-structures qui vont servir de contraintes en étant déclarées comme présentes ou absentes lors de la génération des structures.

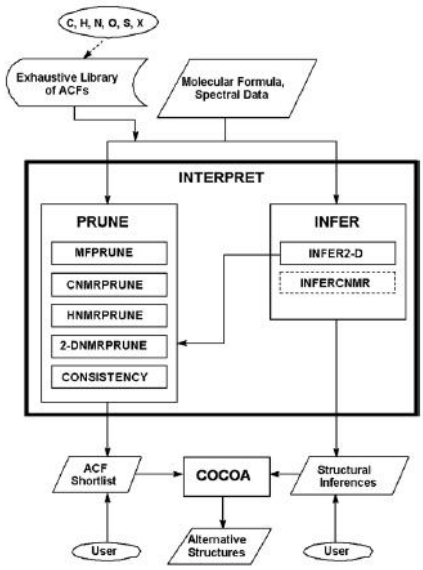


Figure 1.13 – Schéma du fonctionnement de SESAMI

La seconde version de SESAMI (SESAMI-H) utilise le générateur de structure HOUDINI qui remplace COCOA en corrigeant les défauts de ce dernier. HOUDINI possède les avantages suivants : il est plus rapide et l'élimination de certaines structures impossibles est effectuée plus tôt lors de la génération.

1.3.6.2 CISOC-SES

Le système CISOC-SES (Computerized Information System for Organic Chemistry - Structure Elucidation System) [77–80] utilise comme base une formule moléculaire, un spectre ^{13}C et un spectre DEPT pour former un espace de recherche des solutions. L'hybridation des atomes ne doit pas être décrite, ce qui élargit considérablement le nombre de solutions. L'espace de recherche des structures est ensuite affiné à l'aide des données des spectres 2D COSY, HSQC et HMBC. Si le programme ne trouve pas de solution et que les spectres 2D comportent des signaux de faible intensité, l'utilisateur peut élargir les valeurs par défaut des longueurs de chemin de couplage considérées par le programme. Il est également possible de soumettre des fragments de sous-structures (par exemple groupements $C{=}O$ ou $C{\equiv}N$).

1.3.6.3 COCON

Dans le programme COCON (COnstitutions from CONnectivities) [81–85], l'algorithme de génération de structure travaille à partir d'éléments de sous-structure déduits de l'analyse des spectres RMN 2D (COSY, HSQC, HMBC) et/ou fournis par l'utilisateur. Les auteurs ont décrit l'utilisation des spectres 1,1-ADEQUATE afin de lever l'ambigüité des spectres HMBC et distinguer les corrélations correspondant à des $^2J_{C-H}$ de celles correspondant à des $^3J_{C-H}$.

Les données spectrales sont fournies manuellement au programme. Tous les atomes doivent être définis et accompagnés de leur multiplicité (nombre d'hydrogènes liés). L'hybridation des atomes ne doit pas être spécifiée. COCON utilise un certain nombre de règles pour définir automatiquement l'hybridation des carbones à partir des valeurs de déplacement chimique et de multiplicité :

- $\delta_C < 180\ ppm$: l'atome de carbone ne peut pas faire partie d'un groupement allène ;
- $\delta_C > 125\ ppm$: l'atome de carbone ne peut pas être hybridé sp^3 ;
- $\delta_C > 115\ ppm$: l'atome de carbone ne peut pas être hybridé sp ;
- $\delta_C < 75\ ppm$: l'atome de carbone ne peut pas être hybridé sp^2.

L'utilisateur peut choisir d'exclure des sous-structures comme par exemple les cyclopropanes ou les cyclobutanes. Des contraintes additionnelles sont appliquées automatiquement à partir de l'analyse des valeurs de déplacement chimique ^{13}C.

- $\delta_C < 150\ ppm$: l'atome de carbone ne peut pas être lié à un atome d'oxygène ou de soufre hybridé sp^2 ;
- $\delta_C < 130\ ppm$: si l'atome de carbone est hybridé sp^2, alors il ne peut pas être lié à un atome d'oxygène ;
- $\delta_C < 105\ ppm$: si l'atome de carbone est hybridé sp^2, alors il ne peut pas être lié à un atome d'azote ;
- $\delta_C < 45\ ppm$: l'atome de carbone ne peut pas être lié à un atome d'oxygène ;
- $\delta_C > 35\ ppm$: si l'atome de carbone fait partie d'un groupement méthyle, alors il ne peut pas être lié à un autre atome de carbone.

Le générateur de structure examine les sous-structures incomplètes pour les assembler en faisant des vérifications en cours de résolution et une vérification finale lorsqu'une structure est complète.

Si une incohérence est détectée entre la structure en cours et les contraintes alors le programme modifie la structure pour examiner l'hypothèse suivante. Si la structure en cours ne viole aucune contrainte, elle est agrandie en ajoutant une liaison et un atome et vérifiée à nouveau. Lorsqu'une structure complète est obtenue, une vérification finale est effectuée pour valider la structure. Lors de la vérification finale, le programme place les liaisons multiples (doubles et triples liaisons). Toutes les formes possibles d'arrangement des liaisons multiples sont alors envisagées. Les solutions sont filtrées en excluant des sous-structures instables telles que les cyclobutadiènes.

Le programme s'arrête quand toutes les combinaisons possibles de voisins directs pour au moins un atome ont été examinées.

1.3.6.4 LSD

Le programme LSD (Logic for Structure Determination) [86–90] a été développé dans le laboratoire où les travaux présentés dans ce manuscrit ont été réalisés. Les données d'entrée sont extraites manuellement des spectres RMN 1D ^{1}H et ^{13}C et RMN 2D COSY, HSQC, HMBC, H2BC, ADEQUATE et INADEQUATE.

L'ensemble des atomes de la structure doit être défini. Le statut des atomes comprend l'élément chimique, l'hybridation et la multiplicité. Les corrélations sont déclarées comme des paires de signaux qui corrèlent. Lors de l'interprétation d'un spectre HMBC, il est possible de rencontrer une situation où plusieurs carbones ayant des déplacements chimiques très proches corrèlent avec un proton. Dans ce cas la corrélation est déclarée comme un groupe de carbones qui corrèle avec un proton. Au moins un carbone du groupe est obligatoirement impliqué dans la corrélation.

Des contraintes de sous-structure peuvent être définies par l'utilisateur sur la base de ses connaissances et d'une analyse des spectres (déplacements chimiques, forme des multiplets, valeurs de constante de couplage). Ces contraintes concernent le voisinage immédiat des atomes (par exemple une liaison carbone-hétéroatome) ou la présence de cycles.

Lors de la résolution, les structures sont assemblées en formant des liaisons entre les atomes. Les corrélations sont utilisées comme contraintes pour former les liaisons. Du fait des ambigüités sur les corrélations (groupe d'atomes, longueur du chemin de couplage HMBC $^{2}J_{C-H}$ ou $^{3}J_{C-H}$), plusieurs hypothèses sont offertes à chaque nouvel examen d'une corrélation. L'utilisation d'un algorithme récursif permet d'explorer toutes ces hypothèses. On s'assure ainsi de l'exhaustivité des solutions au problème. Le programme peut également prendre en compte l'existence de couplages à très longue distance (par exemple HMBC $^{4}J_{C-H}$). L'utilisateur peut déclarer un certain nombre de corrélations avec une longueur maximale de chemin de couplage.

Les contraintes de voisinage entre atomes sont testées en cours de résolution avant la formation des liaisons.

Lorsqu'une structure est complète, les tests suivants permettent de valider ou de rejeter la solution. Le programme effectue un placement des liaisons multiples. Les contraintes de sous-structures (incluses ou exclues) sont vérifiées. Des contraintes structurales comme la règle de Bredt permettent également de filtrer les solutions. En effet, la règle de Bredt interdit toute double liaison sur les carbones en tête de pont des systèmes bicycliques. La structure 1 de la figure 1.14 est acceptée alors que la structure 2 est rejetée.

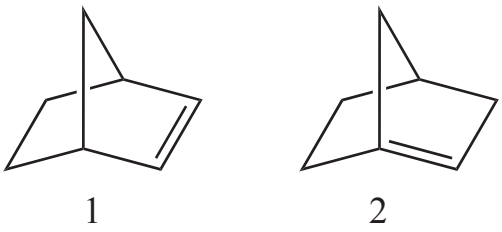


Figure 1.14 – Structures respectant (à gauche) et ne respectant pas (à droite) la règle de Bredt

1.3.6.5 LUCY

Le logiciel LUCY [91] a été développé avec la même philosophie que LSD. Les corrélations HMBC sont utilisées comme contraintes pour former les liaisons entre atomes.

L'algorithme de résolution est basé sur la règle suivante : une corrélation HMBC représente un couplage ${}^2J_{C-H}$ ou un couplage ${}^3J_{C-H}$. Les liaisons entre atomes sont formées de manière récursive pour pouvoir envisager toutes les hypothèses possibles. Les couplages à très longue distance de type ${}^4J_{C-H}$ ne sont pas pris en compte.

Dès qu'une solution est trouvée, elle subit une série de tests (par exemple l'exclusion de cycles de taille donnée) afin de valider la structure.

1.3.6.6 SENECA

SENECA [92, 93], contrairement aux autres systèmes reposant sur les données des spectres RMN 2D n'est pas un système déterministe. Il s'agit d'un système stochastique qui utilise un algorithme génétique ainsi que le recuit simulé. Steinbeck, en reprenant les idées de Faulon [94, 95], a considéré que les limites des systèmes déterministes étaient atteintes au-delà de 30 atomes de squelette. Les données d'entrée du programme sont une formule brute et les spectres ^{1}H, ^{13}C, DEPT, COSY, HSQC et HMBC. L'élucidation structurale est réalisée par un algorithme stochastique en optimisant la structure pour être en accord avec les données spectrales.

La convergence est guidée par un recuit simulé. En utilisant les données d'entrée, le processus commence avec une structure S_1 générée aléatoirement. Une « énergie » E_1 est calculée. Celle-ci correspond à une fonction mesurant l'homogénéité entre la structure et les données spectroscopiques. La structure est ensuite transformée aléatoirement en un isomère S_2 et une nouvelle « énergie » E_2 est calculée. Si $E_2 < E_1$, cela signifie que la structure S_2 est plus homogène avec les données spectroscopiques que la structure S_1. La structure S_2 est donc acceptée. Si $E_2 > E_1$, alors un nombre aléatoire R avec $0 < R < 1$ est généré et la structure est acceptée si :

$$R < e^{-\left(\frac{E_2 - E_1}{kT}\right)}$$

Ce processus, lorsqu'il est réalisé à une température constante T, fait converger le système vers une distribution de Boltzmann. Quand le processus est conduit à des températures successives décroissantes, le système converge vers un état optimal. En effet, seul l'état d'énergie optimal pour la molécule est suffisamment rempli à basse température. Les structures correspondant à cet état optimal sont considérées comme solutions du problème.

L'inconvénient de cette méthode est que l'utilisateur doit optimiser les paramètres de recuit simulé (changement du pas de diminution de température, modifications des fonctions de calcul de score). Le processus doit être réalisé plusieurs fois et les résultats comparés. Si les résultats obtenus ne sont pas les mêmes, le programme de recuit doit être modifié. Les calculs réalisés au cours de plusieurs essais étant indépendants, le processus peut être parallélisé.

Dans une première version du programme, l'étape de modification des structures est réalisée avec la méthode de transformation décrite ci-dessous.

Pour une structure moléculaire, l'opération de mutation ajuste les ordres de liaison (un ordre de liaison égal à 0 signifie une absence de liaison) entre quatre atomes tout en conservant une structure valide chimiquement. La validité de la structure est assurée par une série d'équations de valence. La figure 1.15 illustre l'opération de mutation sur une structure de formule brute $C_{10}H_{16}$.

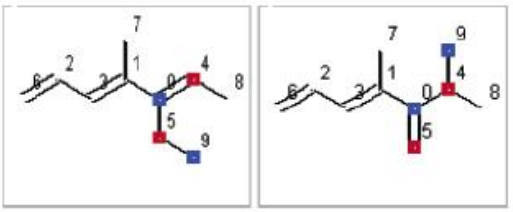


Figure 1.15 – Opérations de mutation dans la première version de SENECA

Dans la structure parente, les atomes numérotés 0, 4, 5 et 9 ont été sélectionnés pour l'opération. Après mutation, la liaison entre les atomes 5 et 9 est supprimée. Les atomes 0 et 5 restent liés mais par une double liaison à la place d'une simple liaison. La modification de cette dernière liaison est compensée par le remplacement d'une double liaison par une liaison simple entre les atomes 0 et 4. Une nouvelle liaison (simple) est formée entre les atomes 4 et 9.

Dans cette version du logiciel, la fonction d'énergie est une somme de différents termes qui tiennent compte de plusieurs paramètres (données spectroscopiques, valeurs de déplacement chimique, symétrie, contraintes de voisinage, contraintes de tailles de cycle)

Dans la seconde version du logiciel, le processus de mutation est différent de celui présenté ci-dessus. En effet, dans la version précédente l'opération de mutation tend à

effectuer un affinage local. Ce processus prend en compte uniquement quatre atomes sur lesquels la mutation est effectuée. Dans ce cas de figure, au plus quatre liaisons entre ces quatre atomes peuvent être modifiées. Les structures filles ont donc très peu de différence par rapport aux structures parentes. Ce processus de mutation est capable de réduire sa portée de destruction, mais son action peut ne pas être pertinente. Pour corriger cette imperfection, des restrictions supplémentaires ont été appliquées pour la sélection des atomes inclus dans le processus de mutation.

Un paramètre appelé rayon de mutation est défini comme le nombre de liaisons maximal entre un atome choisi aléatoirement et le reste des atomes de la structure analysée. Les liaisons présentes à l'intérieur de ce rayon sont susceptibles de changer alors que le reste de la structure reste inchangé dans les structures filles. La mutation sera donc d'autant plus destructive que le rayon de mutation sera élevé. Par exemple, sur la figure 1.16, l'atome marqué d'un point noir est choisi comme centre d'un rayon de mutation égal à 3. Avec cette contrainte, les modifications de structure sont limitées à l'intérieur du cercle indiqué, alors que le reste de la structure reste inchangé.

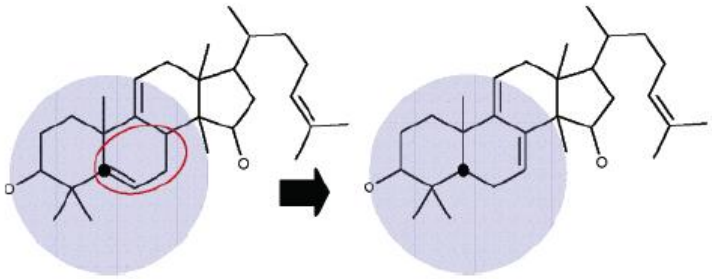


Figure 1.16 – Influence du rayon de mutation sur les modifications

La méthode de mélange qui permet la création de structures filles à partir de structures parentes est présentée avec un exemple. La figure 1.17 montre deux structures sélectionnées comme parentes parmi un ensemble de structures. Les atomes de squelette de chaque structure parente sont aléatoirement séparés en deux groupes (carrés rouges et ronds bleus). Pour le parent 1, le groupe des carrés rouges est constitué des atomes numérotés 1, 2, 3, 6 et 9. Le groupe des ronds bleus regroupe les atomes numérotés 0, 4, 5, 7 et 8. Le parent 2 est divisé suivant la même numérotation. Pour l'opération de mutation, les liaisons entre atomes du même groupe sont préservées alors que les liaisons entre atomes appartenant à des groupes différents sont supprimées (liaisons 3-4 et 6-0 du parent 1).

Les fragments issus de chaque structure parente sont ensuite combinés pour former les structures filles. Les fragments du groupe des carrés rouges du parent 1 sont combinés avec les fragments du groupe des ronds bleus du parent 2. De même, les fragments du groupe des ronds bleus du parent 1 sont combinés avec les fragments du groupe des carrés rouges du parent 2. Deux structures filles sont obtenues en ajoutant les liaisons manquantes (liaisons 6-7, 0-3 et 3-8 de la structure fille 1 et liaisons 2-6 et 3-4 de la structure fille 2).

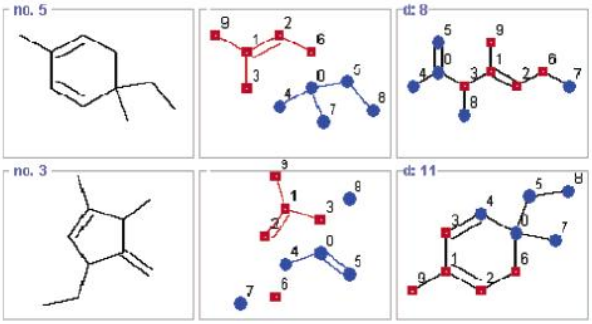


FIGURE 1.17 – Opérations de mutation dans la seconde version de SENECA

La fonction d'énergie de cette seconde version est identique à la précédente avec un terme supplémentaire tenant compte de la règle de Bredt.

La figure 1.18 résume le processus de détermination de structure de SENECA.

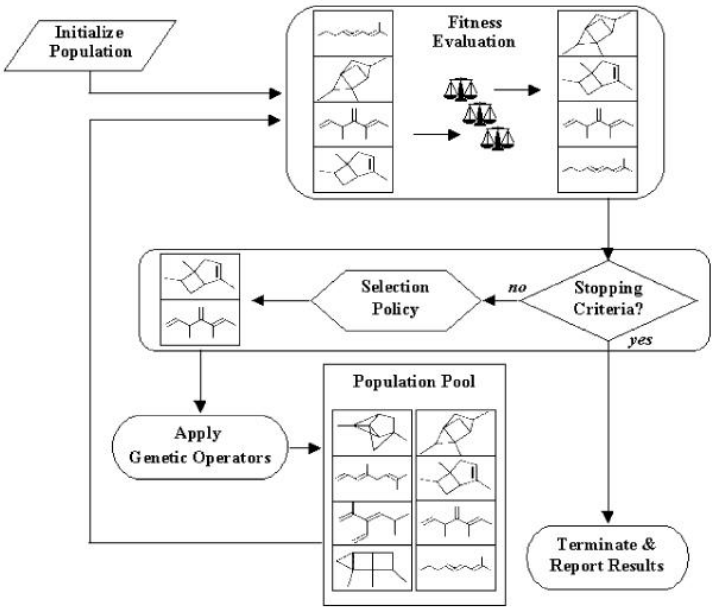


FIGURE 1.18 – Schéma du fonctionnement de SENECA

1.3.6.7 ACD Structure Elucidator 2D

Ce logiciel est une deuxième version de Structure Elucidator qui est capable d'utiliser des informations provenant des spectres RMN 2D en entrée [96–113].

La connaissance de la formule brute est obligatoire. Les données peuvent être entrées manuellement ou extraites des spectres RMN automatiquement. Tous les groupements CH_3, CH_2, CH et C ainsi que tous les hétéroatomes et le nombre d'hydrogènes qu'ils portent sont définis. Les propriétés des carbones (hybridation et éventuel voisin hétéroatome) sont déduites de l'analyse des déplacements chimiques ^{13}C et ^{1}H des groupements

CH_3, CH_2, CH et C. Le programme est également capable de tenir compte de fragments comme contraintes de sous-structure lors de la résolution.

Les informations déduites des spectres (relations de proximité entre atomes) servent à former des liaisons entre les atomes. Une méthode a été présentée pour détecter et enlever automatiquement les corrélations apportant des contradictions (corrélations à très longue distance). Si le programme détecte un atome pour lequel une corrélation semble être une corrélation à très longue distance, il augmente de 1 la longueur par défaut de toutes les corrélations dans lesquelles cet atome est impliqué. Si aucune solution n'est trouvée avec cette méthode, la longueur maximale du chemin de couplage est donc supérieure. Le programme supprime alors une corrélation du jeu de données.

1.3.7 Approches alternatives pour l'élucidation structurale

1.3.7.1 Élucidation structurale 3D directe

Un projet de la société Mestrelab Research vise à réaliser une élucidation structurale 3D sans passer par l'intermédiaire de la construction d'une structure 2D. Le principe est de convertir les données des spectres RMN 2D en distances entre atomes dans l'espace et non plus en nombre de liaisons entre atomes.

La figure 1.19 montre les paramètres géométriques qui peuvent être déduits des spectres RMN 2D.

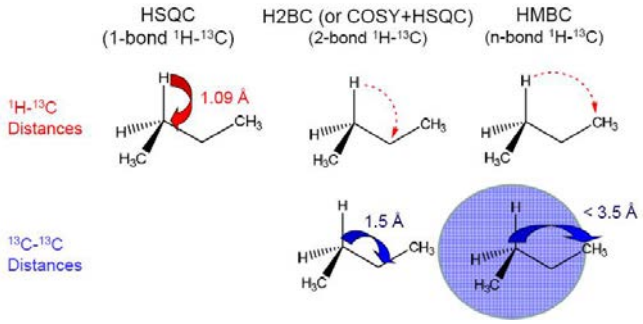


Figure 1.19 – Règles d'élucidation 3D

Une corrélation HSQC est convertie en une distance H–C de 1,09 Å, une corrélation H2BC ou HMBC $^2J_{C-H}$ correspond à une distance C–C de 1,5 Å et une corrélation HMBC $^3J_{C-H}$ à une distance C–C inférieure à 3,5 Å.

La figure 1.20 montre un exemple d'application de cette méthode sur la strychnine.

Il s'agit également de pouvoir utiliser les informations des spectres NOESY [114]. L'intensité des signaux est proportionnelle à r_{ij}^{-6} avec r_{ij} la distance entre les atomes d'hydrogène i et j [115]. D'une manière générale on considère qu'un signal nOe « fort » est observé pour des hydrogènes séparés par 1,8–2,5 Å, un signal « moyen » pour une

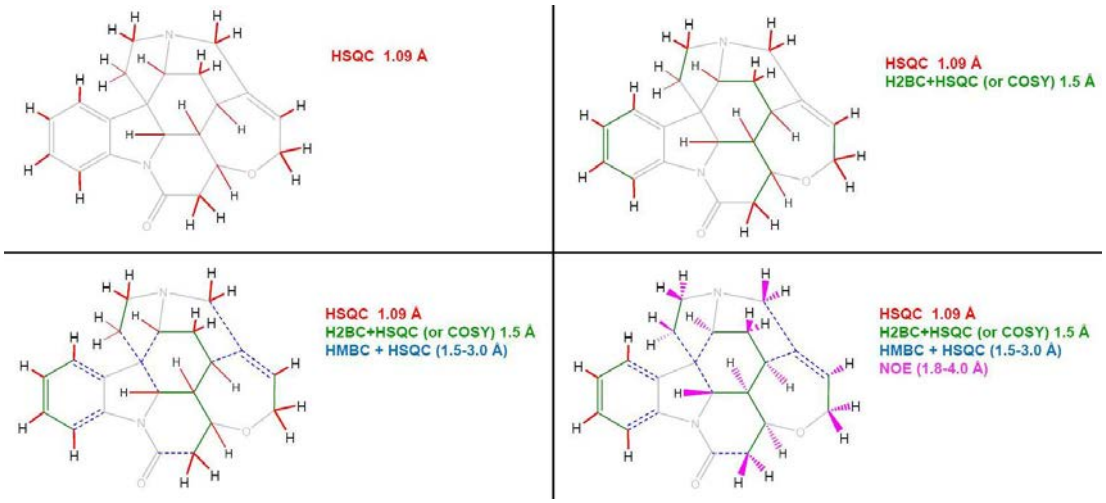


Figure 1.20 – Exemple d'élucidation 3D de la structure de la strychnine

distance de 2,5–4,0 Å et un signal « faible » pour une distance supérieure. On considère que l'on n'observe plus de nOe pour des atomes dont la distance est supérieure à 6,0 Å.

Les valeurs de constantes de couplage permettent d'obtenir des contraintes angulaires entre atomes. Il est possible de déduire d'une valeur de constante de couplage $^3J_{H-H}$ une valeur d'angle dièdre entre les deux atomes d'hydrogène à l'aide de l'équation d'Altona [116]. De même une valeur d'angle dièdre peut être obtenue à l'aide des équations de Bifulco [117] à partir d'une valeur de constante de couplage $^3J_{C-H}$.

1.3.7.2 Déréplication

Un autre outil a été développé afin d'éviter de renouveler des efforts dans l'analyse des données expérimentales : la déréplication. Ce système repose sur de larges bases de données de composés dont l'attribution des déplacements chimiques a déjà été effectuée.

La déréplication est utilisée en élucidation structurale pour éviter de renouveler la procédure complexe d'analyse en comparant directement les données spectrales de la molécule à une base de données. Si la molécule est présente dans la base de données, l'analyse s'achève. Dans le cas contraire, la procédure d'élucidation se poursuit en utilisant les méthodes habituelles. Des méthodes appliquées aux substances naturelles ont été décrites dans la littérature [118].

Bradshaw *et al* ont décrit un outil de recherche dans une base données basé sur la masse molaire et le nombre exact de groupements CH_3, CH_2 et CH contenus dans la structure inconnue [119]. Ces critères pouvant être déduits d'un spectre DEPT ou de la combinaison des spectres ^{1}H et HSQC édité. L'expérience HSQC éditée [120] diffère de l'HSQC classique en modifiant le signe des signaux en fonction de la parité de la multiplicité des carbones. Les signaux des groupements CH_3 et CH ont un signe opposé aux signaux des groupements CH_2.

Lang *et al* ont proposé une méthode de recherche basée sur le nombre exact de groupements CH_3, de groupements CH_2 et CH dont le carbone est hybridé sp^3 et certains groupements chimiques (benzyle, acétal, éther, amide...) facilement déduits d'une analyse manuelle du spectre ^1H (déplacements chimiques et multiplets) [121].

Un article récent de Qiu *et al* présente une méthode de déréplication appliquée à la famille des triterpènes [122]. Les auteurs proposent l'utilisation des valeurs de déplacement chimique des groupements méthyles présents dans ces composés comme critères de reconnaissance.

Pierens *et al* ont développé un algorithme de recherche par similarité dans une base de données de spectres 2D HSQC [123, 124].

Dans l'ensemble des cas, la déréplication est effectuée en recherchant la structure la plus compatible par rapport à des critères provenant de l'analyse des spectres.

1.3.8 Conclusion

Les outils disponibles capables d'aider les chimistes dans le processus de détermination de structure sont de natures différentes. Le choix des outils doit être fait en fonction des données disponibles. On ne peut pas faire sortir du lot un outil plutôt qu'un autre car ils sont complémentaires. Il est plus judicieux d'utiliser plusieurs outils afin de confronter les résultats obtenus.

Le système d'élucidation structurale idéal devrait être constitué des étapes suivantes :

- déréplication ;
- si la structure est absente de la base de données, génération des structures ;
- filtrage des solutions ;
- classement des solutions par une prédiction des déplacements chimiques.

Actuellement, l'offre de logiciels utilisables par la communauté scientifique est assez restreinte. Certains sont disponibles sous forme de produits commerciaux tels que les logiciels payants ACD Structure Elucidator [125], CISOC-SES distribué sous le nom NMR-SAMS par Spectrum Research, LLC [126] et AssembleIt distribué par ScienceSoft, LLC [127]. D'autres comme les logiciels libres et gratuits LSD [128] et SENECA [129] sont accessibles en téléchargement via leur site web. Le programme COCON est utilisable gratuitement avec une version en ligne appelée WebCocon [130].

1.4 Vérification de structure automatique

1.4.1 Introduction

Vérifier la structure d'une molécule est une tâche qui intervient obligatoirement à la suite d'une réaction chimique. Il s'agit de vérifier l'identité et la pureté du produit attendu. Un chimiste peut également disposer de plusieurs hypothèses de produits de réaction et

chercher à savoir quel produit a été obtenu.

Les logiciels de vérification automatique tentent de juger si les données spectrales sont compatibles avec une structure. Le but de l'automatisation est d'effectuer une vérification le plus rapidement possible afin de limiter les temps d'analyse. Les chimistes peuvent ainsi se concentrer sur des problèmes de chimie en passant moins de temps sur la caractérisation de leurs produits.

La finalité de la vérification est de donner une réponse positive ou négative au problème posé. Dans certains cas le logiciel ne prend pas la bonne décision. Il est tout à fait possible que la structure analysée soit très proche de la structure attendue et que le logiciel ne puisse pas faire la différence entre leurs spectres respectifs. Il est également possible qu'une structure correcte ne soit pas acceptée par le programme pour des raisons qui seront discutées par la suite.

Il est impératif de définir quelques notions qui seront utiles pour la suite :

- vrai positif : une paire structure/spectre qui est connue comme correcte et que le logiciel certifie comme correcte ;
- vrai négatif : une paire structure/spectre qui est connue comme incorrecte et que le logiciel certifie comme incorrecte ;
- faux négatif : une paire structure/spectre qui est connue comme correcte mais que le logiciel certifie incorrecte ;
- faux positif : une paire structure/spectre qui est connue comme incorrecte mais que le logiciel considère comme correcte.

1.4.2 Stratégies des logiciels de vérification

La prédiction des spectres RMN joue un rôle central dans ces logiciels. Ils procèdent à une comparaison entre le spectre expérimental et le spectre prédit. Au cours du processus il est également nécessaire de repérer les traces de solvants résiduels et éventuelles impuretés.

Une première famille de logiciel travaille uniquement sur les spectres ^{1}H. Si la réalisation d'un spectre proton est une chose facile et rapide, la prédiction peut devenir difficile quand un haut niveau de précision est requis. La prédiction des spectres ^{13}C est plus facile à réaliser puisque l'on considère uniquement des déplacements chimiques. Une seconde famille de logiciel utilisant les déplacements chimiques ^{13}C a donc vu le jour. Devant la faible sensibilité du noyau de ^{13}C qui peut entrainer des difficultés pour obtenir un spectre exploitable en un temps restreint, une stratégie alternative consiste à enregistrer un spectre HSQC. Cette expérience est plus sensible car l'acquisition est effectuée de manière indirecte par détection des signaux des protons.

1.4.3 Logiciels basés sur les spectres ^{1}H

Ces logiciels proposent une vérification complète du spectre en examinant les déplacements chimiques, les constantes de couplage et l'intégration des signaux. Ils sont basés

sur une comparaison entre un spectre prédit à partir de la structure et un spectre expérimental.

Griffiths et Bright ont décrit une méthode de comparaison des déplacements chimiques expérimentaux avec des déplacements chimiques prédits par le logiciel de prédiction de ACD/Labs [131–133]. Les déplacements chimiques sont classés par valeurs croissantes et directement comparés. Cette approche donne simplement une réponse positive ou négative. L'utilisateur n'a aucune idée sur l'attribution des signaux, à savoir quel signal expérimental correspond à quel signal prédit. Une attribution permet à l'utilisateur de comparer directement les propriétés des signaux. En cas d'échec l'utilisateur dispose de pistes pour poursuivre l'analyse. Une tentative d'attribution peut donc constituer une plus-value dans un processus de vérification. Un autre inconvénient de cette méthode est que la décision dépend uniquement de la précision de la prédiction. Une structure correcte dont la prédiction est mauvaise pourrait être rejetée. Il y a donc un risque d'avoir un fort taux de faux positifs et de faux négatifs.

Golotvin *et al* (ACD/Labs) [134] ont proposé une méthode permettant de corriger les défauts de la précédente en réalisant une attribution du spectre lors de la vérification. Le résultat de la vérification est donné sous la forme d'un score représentant le degré de certitude sur la décision.

L'algorithme de vérification effectue une caractérisation des multiplets du spectre expérimental, puis réalise une prédiction du spectre à partir de la structure et enfin compare les spectres calculés et expérimentaux.

Lors de l'attribution, pour chaque paire signal expérimental/signal prédit, une fonction d'évaluation permet de mesurer les différences entre les différentes propriétés des signaux. Cette fonction a une valeur minimale lorsque la différence des propriétés est minimale. L'état qui correspond au minimum de cette fonction est donc le meilleur cas d'attribution. La valeur de la fonction peut permettre de juger de la correspondance entre la structure et le spectre.

La fonction d'évaluation est calculée de la manière suivante :

$$F = \sum (W_{shift}F_{shift} + W_{Quant}F_{Quant} + W_{Mult}F_{Mult})$$

F_{shift}, F_{Quant} et F_{Mult} représentent respectivement les degrés de similarité entre les déplacements chimiques, intégrales et multiplets. W_{shift}, W_{Quant} et W_{Mult} sont des facteurs de poids pour chacun des termes F_x. La somme est effectuée sur l'ensemble des signaux. Les termes F_x peuvent prendre une valeur comprise entre 0 et 1. Par exemple, F_{shift} prend la valeur 0 si la différence entre les déplacements chimiques prédits et expérimentaux est inférieure à 0,1 ppm et 1 si la différence est supérieure ou égale à 0,5 ppm. Les valeurs de tolérance pour chacun des termes peuvent être modifiées par l'utilisateur.

Un coefficient de correspondance est introduit afin d'évaluer le degré de consistance

entre le spectre expérimental et le spectre prédit. Ce paramètre peut prendre une valeur de 0 à 1, 1 signifiant une correspondance parfaite. Ce cas de figure se présente lorsque la fonction F ne contient aucun terme de pénalité. Cela arrive lorsque les différences, pour toutes les propriétés des signaux, sont comprises dans les bornes de tolérance données par l'utilisateur. Ce coefficient est égal à 0 quand toutes les pénalités atteignent leurs valeurs maximales.

Les auteurs ont choisi de classer les résultats de la vérification en trois catégories en fonction de la valeur du coefficient de correspondance. La figure 1.21 résume les catégories de décision du logiciel.

Coefficient de correspondance	0,0 - X	X - Y	Y - 1,0
Catégorie de correspondance	Incorrecte	Ambigüe	Correcte
Résultat	Vrai Négatif Faux Négatif	Ambigu	Vrai Positif Faux Positif
Comportement à adopter par l'utilisateur	Pas d'action supplémentaire requise	Évaluation manuelle supplémentaire	Pas d'action supplémentaire requise

Figure 1.21 – Catégories de décision pour la vérification

Le choix des valeurs de seuil délimitant ces trois catégories est laissé au soin de l'utilisateur. L'optimisation de ces valeurs détermine quelles structures doivent être évaluées manuellement ou non. L'utilisateur doit définir les valeurs des paramètres de seuil X et Y afin d'obtenir un compromis permettant de réduire le nombre de faux négatifs et de faux positifs tout en ayant un nombre de résultats ambigus assez bas pour garder le bénéfice de l'automatisation.

Un inconvénient des méthodes décrites ci-dessus réside au niveau de la prédiction des valeurs de constantes de couplage et de l'analyse des multiplets. En effet, l'analyse est effectuée en se basant sur l'hypothèse que tous les spectres sont du premier ordre, or il n'est pas rare d'observer des couplages forts qui donnent lieu à des spectres du second ordre.

Le logiciel PERCH [135–137] combat cette imperfection. L'extraction manuelle des paramètres des spectres RMN expérimentaux peut être fastidieuse, difficile voire impossible lorsque des signaux sont superposés ou en cas de couplages forts. Dans ces cas, une analyse spectrale ayant recours à des calculs quantiques permet d'extraire ces paramètres des spectres. PERCH fonctionne de cette manière en optimisant les valeurs de déplacements chimiques et constantes de couplage prédits par une méthode itérative. Cependant les va-

leurs de départ pour ces paramètres doivent être raisonnablement correctes et l'attribution des signaux correcte pour que le processus itératif aboutisse.

L'algorithme ACA (Automated Consistency Analysis) effectue ce type d'analyse de manière totalement automatique en incluant la reconnaissance des signaux des solvants. Les étapes de la procédure sont les suivantes :

- Détection automatique des signaux, intégration, calcul de la forme et de la largeur des raies et calcul du rapport signal sur bruit. Une analyse totale de la forme des raies est utilisée pour effectuer une déconvolution du spectre. Elle permet aussi de détecter des raies larges (par exemple signal de l'eau). Les informations extraites du spectre sont utilisées dans la suite pour effectuer la comparaison avec le spectre prédit au cours du processus itératif d'optimisation ;
- La prédiction du spectre ^{1}H est effectuée en évaluant les déplacements chimiques à partir d'une collection de conformères obtenue par dynamique moléculaire ;
- Un ensemble d'attributions est réalisé sur la base des résultats des deux premières étapes. Ces attributions sont classées par ordre décroissant de correspondance entre les spectres prédits et expérimentaux en tenant compte de l'erreur sur la prédiction ;
- Processus itératif d'optimisation des paramètres RMN pour que le spectre prédit se superpose au maximum au spectre expérimental ;
- Évaluation finale : un score de correspondance est calculé donnant une estimation de la consistance entre la structure et le spectre RMN ^{1}H. Une série de scores supplémentaires permet de juger la divergence entre les paramètres RMN expérimentaux et prédits (déplacements chimiques, constantes de couplage), et aussi de savoir quelles régions du spectres ont une mauvaise consistance.

Lorsque le résultat du mode automatique est ambigu ou que la structure est rejetée, l'utilisateur peut passer en mode interactif pour guider le logiciel dans le processus.

La société Mestrelab Research commercialise un programme appelé Mnova Verify [138] qui utilise un spectre RMN ^{1}H et optionnellement un spectre HSQC pour la vérification des structures. Le programme dispose d'un algorithme de déconvolution et d'un module de prédiction des spectres. Un facteur de qualité permettant d'évaluer la vérification est calculé. Les détails ne sont toutefois pas communiqués.

1.4.4 Logiciels basés sur les spectres ^{1}H et HSQC

Afin d'améliorer leur système, Griffiths *et al* ont envisagé l'utilisation combinée des spectres RMN ^{1}H et HSQC [139, 140]. Ils ont montré qu'avec une approche utilisant uniquement le spectre ^{1}H, le taux de faux positifs est de 4%. L'utilisation des spectres ^{13}C permet de diminuer ce taux à 1%. Les spectres HSQC donnent un taux de 2% et la combinaison des spectres ^{1}H et HSQC réduit le taux à 1%.

Golotvin *et al* ont implémenté la même approche pour améliorer leur logiciel de vérification [141, 142]. Pour évaluer l'attribution des signaux, la même fonction de score est appliquée. Un terme tenant compte de la différence des déplacements chimiques ^{13}C calculés et expérimentaux est ajouté. Pour fournir un résultat, le coefficient de correspondance est calculé. Il est égal à la fonction de score normalisée par rapport au nombre de signaux.

Les auteurs concluent que l'utilisation du spectre HSQC est pertinente car bien qu'augmentant le temps d'analyse, le temps passé par un agent expert pour analyser et évaluer manuellement les spectres en cas d'ambigüité s'en trouve réduit.

Récemment, une étude de Golotvin *et al* a été menée pour tenter de réduire le nombre de faux positifs [143]. Les auteurs ont procédé à des tests de vérifications combinées sur des structures ayant une grande similarité par rapport à la structure correcte. Ils ont montré que la soumission combinée de structures similaires (au moins une) permet d'indiquer des situations où la combinaison des spectres ^{1}H et HSQC n'est pas suffisante pour confirmer la structure. Dans cette étude les structures similaires ont été générées manuellement. On pourrait imaginer d'avoir recours à un générateur de structure automatisé pour fournir des structures alternatives.

1.4.5 Conclusion

Les logiciels de vérification de structure nécessitent des méthodes d'extraction des paramètres des signaux RMN (déplacements chimiques, constantes de couplage et intégrations) perfectionnées et des méthodes de prédiction des déplacements chimiques et des constantes de couplage efficaces et précises.

D'une manière générale l'introduction des spectres HSQC a permis d'améliorer les systèmes d'attribution et de vérification en permettant une réduction du nombre de faux négatifs et faux positifs. L'inconvénient majeur de cette stratégie est la possibilité d'erreur dans la décision finale car il manque l'information des carbones quaternaires.

Pour être efficace, un logiciel de vérification de structure doit posséder un module de prédiction efficace. En effet, la qualité de l'attribution dépend essentiellement de la qualité de la prédiction.

En cas de rejet de structure par un logiciel de vérification, on peut imaginer de soumettre les données spectrales au générateur de structures d'un programme d'élucidation structurale pour obtenir des structures alternatives. Ceci démontre le grand potentiel et la complémentarité de tous les outils d'analyse structurale lorsqu'ils sont utilisés ensemble.

Chapitre 2

Amélioration du logiciel LSD

2.1 Introduction

Le logiciel LSD est né au début des années 1990 et ne cesse d'évoluer depuis sa première version. L'objectif de cette thèse est de supprimer certaines rigidités imposées dès la conception du logiciel LSD et de faciliter son utilisation.

Ce chapitre débute par une description détaillée du fichier d'entrée, des différentes fonctionnalités, des programmes associés et de l'algorithme de résolution. Les manques et limitations du logiciel seront exposés avant de présenter les améliorations apportées au cours de ce travail de thèse. Chaque modification sera traitée par la définition d'un cahier des charges et des outils utilisés. Des exemples de résolution de structure sont proposés afin de valider les modifications.

Enfin pour faire suite à une collaboration avec le Professeur Vicente de Paulo Emerenciano de l'université de São Paulo, des travaux ont été menés pour continuer à exploiter les informations issues de la base de données spectroscopiques SISTEMAT. Le couplage des systèmes SISTEMAT et LSD permet à chacun de tirer profit de l'autre. Les dernières avancées dans ce sens sont dévoilées en fin de chapitre.

2.2 Le logiciel LSD

2.2.1 But du logiciel

Le but du logiciel LSD est de proposer l'ensemble des structures planes compatibles avec des données spectroscopiques. À partir de la formule brute fournie par la spectrométrie de masse et des spectres de RMN 1D ^{1}H et ^{13}C, un statut (état d'hybridation, nombre d'hydrogènes portés) est attribué par l'utilisateur aux atomes de la molécule. Des distances ou des intervalles de distances entre atomes, mesurées en nombre de liaisons, sont déduits des spectres de RMN 2D de corrélation des déplacements chimiques : COSY, HSQC, HMBC, H2BC, 1,1-ADEQUATE et INADEQUATE. Les spectres NOESY ne sont pas exploités pour établir des conformations relatives. L'utilisateur peut fournir au programme des éléments de sous-structure issus de sa propre connaissance des déplacements chimiques et de l'origine naturelle ou synthétique de la substance étudiée. Utiliser le logiciel LSD impose donc un certain nombre de travaux préliminaires à l'utilisateur. Les

informations extraites des spectres sont codées dans un fichier d'entrée pour être traitées par LSD.

2.2.2 Structure du fichier d'entrée

Le fichier d'entrée est constitué de commandes permettant à l'utilisateur de coder les informations issues des spectres mais aussi de contrôler l'exécution du programme (affichage de commentaires en cours d'exécution, exécution en mode pas à pas...). La liste des commandes disponibles dans les fichiers d'entrée de LSD est présentée en annexe A.

Chaque commande est constituée d'un mnémonique suivi de 1 à 5 paramètres. Tous les mnémoniques sont constitués de 4 caractères alphanumériques. Les mnémoniques sont suivis par des paramètres séparés par des blancs. Les paramètres sont classés par types :

- I : un entier positif ou nul ;
- T : -1, 0 ou 1 ;
- V : un seul entier positif ou nul ou un ensemble d'entiers positifs ou nuls entre parenthèses séparés par des blancs ;
- A : un symbole atomique ;
- Ln : la référence à une liste d'atomes, n est strictement positif ;
- B : remplace I ou Ln ;
- S : un ensemble d'entiers positifs séparés par des blancs ;
- Sn : la référence d'un atome dans la description d'une sous-structure ;
- R : un nombre réel ;
- Fn : la référence d'une sous-structure ;
- C : une chaîne de caractère, encadrée par des guillemets (") ;
- H : un signe optionnel : + ou - ;
- O : un entier positif ou nul optionnel ;
- Z : -1, 0, 1 ou 2.

Le codage du jeu de données LSD pour l'α-pinène (figure 2.1) est présenté comme exemple en figure 2.2.

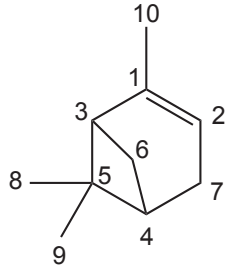


Figure 2.1 – Structure de l'α-pinène

```
ENTR 1
ELIM 3 5

MULT 1 C 2 0      HMBC 1 6       QUAT L1
MULT 2 C 2 1      HMBC 1 9       PROP 8 0 L1
MULT 3 C 3 1      HMBC 2 3       PROP 10 0 L1
MULT 4 C 3 1      HMBC 2 9
MULT 5 C 3 0      HMBC 3 6
MULT 6 C 3 2      HMBC 3 8       SSTR S1 A (2 3) (0 1 2)
MULT 7 C 3 2      HMBC 3 9       SSTR S2 A (2 3) (0 1 2)
MULT 8 C 3 3      HMBC 3 10      SSTR S3 A (2 3) (0 1 2)
MULT 9 C 3 3      HMBC 4 6       SSTR S4 A (2 3) (0 1 2)
MULT 10 C 3 3     HMBC 4 8       LINK S1 S2
                  HMBC 4 10      LINK S2 S3
HSQC 2 2          HMBC 5 6       LINK S3 S4
HSQC 3 3          HMBC 5 8       LINK S1 S4
HSQC 4 4          HMBC 5 10
HSQC 6 6          HMBC 7 6       DEFF F1 "Filters/ring3"
HSQC 7 7          HMBC 8 10
HSQC 8 8          HMBC 9 3       ; skeleton database
HSQC 9 9          HMBC 10 8      PATH "Filters/TERPENES/MONOTERP"
HSQC 10 10        HMBC 10 6
                  HMBC 10 9      ; pinane
BOND 1 2          HMBC 2 8       SKEL F2 "PINANE"
BOND 2 7

                                 FEXP "F0 AND NOT F1 AND F2"
```

Figure 2.2 – Fichier de données LSD pour l'α-pinène

Tous les atomes (hydrogènes exceptés) doivent être définis à l'aide la commande `MULT`. Le statut des atomes comprend le numéro d'atome, l'élément chimique, l'hybridation et la multiplicité. La commande `MULT 1 C 2 0` signifie que l'atome 1 est un carbone (`C`) hybridé sp^2 (`2`) quaternaire (`0` hydrogène). La numérotation des atomes peut être faite de manière arbitraire. Dans l'exemple, les atomes sont numérotés par ordre décroissant de valeurs de déplacement chimique ^{13}C. Les numéros des carbones issus du spectre 1D sont reportés sur les projections des spectres 2D HSQC et HMBC.

Il s'agit ensuite de numéroter les atomes d'hydrogène sur les spectres. La numérotation est arbitraire mais il est plus facile de donner à un hydrogène le même numéro que l'atome auquel il est lié. La correspondance est donc établie sur le spectre HSQC. Les numéros sont reportés sur les spectres 2D COSY et HMBC. Dans le fichier de données, les corrélations des spectres 2D sont définies par le type de spectre, le numéro d'atome dans la première dimension (un carbone pour les spectres HSQC et HMBC, un hydrogène pour les spectres COSY) et le numéro d'atome dans la deuxième dimension (un hydrogène dans tous les cas). La commande `HSQC 2 2` signifie que le carbone 2 et le proton 2 sont liés. Si des carbones ont des déplacements chimiques ^{13}C très proches, la syntaxe de la commande `HMBC` peut être adaptée en définissant un groupe d'atomes entre parenthèses. Cela signifie qu'au moins un des carbones est impliqué dans la corrélation. Par exemple, la commande `HMBC (6 7) 2` signifie que soit le carbone 6, soit le carbone 7 (ou les deux) corrèle(nt) avec l'hydrogène 2.

Les liaisons déduites de façon évidente doivent être déclarées avec la commande `BOND` dont les deux paramètres sont les numéros des deux atomes à lier. Dans l'exemple du pinène, la double liaison entre les deux uniques atomes déclarés hybridés sp^2 (les carbones 1 et 2) est indiquée par la commande `BOND 1 2`.

L'expérience INADEQUATE permet d'observer des corrélations entre carbones directement liés. Ces corrélations peuvent donc être déclarées à l'aide de la commande `BOND`.

Il est possible de donner des propriétés aux atomes. Une propriété d'atome (commande `PROP`) est l'indication pour un atome ou une liste d'atomes d'un nombre de voisins qui doivent appartenir ou non à une autre liste d'atomes. Ces propriétés agissant comme contraintes concernant le voisinage des atomes peuvent être déduites d'une simple analyse des déplacements chimiques et de la multiplicité des signaux. Elles peuvent, par exemple, s'appliquer au voisinage des hétéroatomes. Dans l'exemple du pinène, la liste des carbones quaternaires est déclarée avec la commande `QUAT L1`. Le carbone 10 (δ (^{13}C) = 20 ppm) appartient à un groupement méthyle apparaissant sous la forme d'un singulet vers 1 ppm sur le spectre ^{1}H. La commande `PROP 8 0 L1` impose au carbone 8 d'avoir tous ses voisins quaternaires. Elle force la formation d'une liaison entre le carbone 8 et un carbone quaternaire. Il aurait été possible d'écrire `PROP 8 1 L1` car un groupement méthyle n'a toujours qu'un seul voisin.

L'algorithme de résolution utilise les corrélations HMBC ^nJ avec une valeur de n restreinte à 2 ou 3. L'introduction d'une corrélation HMBC ou COSY ^nJ avec n supérieur

à 3 mène LSD à une situation où il ne fournit aucune solution. Une telle corrélation peut être éliminée afin de permettre au logiciel de trouver au moins une solution. La commande `ELIM` permet de spécifier un nombre maximal de corrélations qu'il est possible d'éliminer ainsi qu'un nombre maximal autorisé de liaisons entre les atomes qui interviennent dans toutes les corrélations éliminées. L'exemple du pinène contient trois corrélations HMBC à très longue distance : entre le carbone 10 et l'hydrogène 6 (^4J), entre 10 et 9 (^5J) et entre 2 et 8 (^5J). Ces corrélations sont imaginaires et ont été ajoutées pour présenter l'utilité de la commande `ELIM`. La commande `ELIM 3 5` signifie que le programme peut éliminer 3 corrélations qui correspondent au maximum à un couplage ^5J entre les atomes. Si cette commande n'est pas présente, le programme est incapable de résoudre la structure du pinène.

La commande `ENTR 1` présente en début du fichier exemple fait partie des commandes permettant de contrôler l'exécution du programme. Elle permet l'affichage et le contrôle des données après lecture du fichier d'entrée.

Les solutions trouvées par LSD peuvent être filtrées pour ne garder que celles satisfaisant à une contrainte de sous-structure. Une sous-structure est un fragment de structure défini par des liaisons entre des sous-atomes que l'utilisateur peut forcer à être présent ou absent des solutions. Une sous-structure peut être définie directement dans le fichier d'entrée ou dans un fichier spécifique. Dans l'exemple, une contrainte de sous-structure est appliquée. Il s'agit d'une combinaison de plusieurs fragments définis de manière indépendante. Un premier fragment est défini dans le fichier d'entrée. Il s'agit d'un cycle de 4 atomes. La commande `SSTR S1 A (2 3) (0 1 2)` permet de déclarer le statut des sous-atomes. Le sous-atome S1 peut correspondre à tout élément (`A` pour Any) hybridé sp^2 ou sp^3 (`(2 3)`) lié à 0, 1 ou 2 hydrogènes (`(0 1 2)`). Les liaisons entre sous-atomes sont définies par la commande `LINK`. `LINK S1 S2` signifie que les sous-atomes S1 et S2 doivent être liés. Une seule définition de sous-structure peut être codée dans le fichier d'entrée de LSD. Il est plus commode de définir les sous-structures de manière externe afin de les réutiliser dans plusieurs problèmes. C'est le cas pour le second fragment. Il s'agit d'un cycle de 3 atomes. La commande `DEFF F1 "Filters/ring3"` permet de spécifier l'emplacement du fichier externe et d'attribuer un numéro au fragment (`F1`). Le troisième fragment correspond au squelette carboné principal de l'α-pinène : le pinane. Il est également défini dans un fichier externe. La commande `PATH "Filters/TERPENES/MONOTERP"` ajoute ce chemin d'accès à la liste des endroits où sont recherchés les squelettes. La commande `SKEL F2 "PINANE"` indique que le fragment `F2` est le squelette du pinane. Le fichier associé à ce squelette sera automatiquement recherché dans le dossier ajouté avec la commande `PATH`. Pour finir, la façon de combiner les fragments pour la recherche de sous-structure est contenue dans la commande `FEXP "F0 AND NOT F1 AND F2"`. Il s'agit d'une expression de logique combinatoire entre les fragments `F0` (sous-structure du fichier d'entrée), `F1` et `F2`. La solution doit contenir un cycle de 4 atomes (`F0`) et ne pas contenir de cycle de trois atomes (`AND NOT F1`) et contenir le squelette du pinane (`AND F2`).

2.2.3 LSD et programmes associés

Le programme est divisé en plusieurs programmes indépendants jouant chacun un rôle différent au cours de l'analyse. Tous les constituants d'un logiciel d'élucidation structurale sont présents sauf la partie d'extraction des données des spectres qui reste manuelle. Les informations sont codées dans un fichier texte. Pour utiliser LSD dans une interface graphique, il est possible d'utiliser le module d'élucidation structurale CMC-se (Complete Molecular Consistency-structure elucidation) [144] intégré au logiciel TopSpin de la société Bruker. Les programmes sont utilisés les uns à la suite des autres. Tout d'abord LSD contient l'algorithme de résolution, ensuite OUTLSD permet d'exporter les résultats dans divers formats et enfin M_EDIT et GENPOS permettent de visualiser les solutions.

2.2.3.1 LSD

Le programme LSD est le générateur de structures proprement dit. Il est écrit en langage C. Il reçoit le fichier d'entrée et utilise les données qu'il contient pour générer l'ensemble des solutions. Le détail du fonctionnement interne de LSD est présenté au paragraphe 2.2.4. Les solutions sont placées dans un fichier de sortie dont le format est lisible uniquement par le programme OUTLSD.

2.2.3.2 OUTLSD

Le programme OUTLSD, écrit en langage C, a été conçu afin d'exporter les résultats dans plusieurs formats utilisables par l'utilisateur. Les formats de sortie de OUTLSD sont les suivants :

- des listes de paires d'atomes liés ;
- des chaînes de caractères SMILES [145, 146] ;
- des coordonnées 2D pour les programmes M_EDIT et GENPOS ;
- des coordonnées 2D au format MDL MOL [147] pour certains produits propriétaires et M_EDIT ;
- des coordonnées 3D fantaisistes au format MDL MOL avec H implicites ou explicites.

Les coordonnées 3D sont fantaisistes car le programme ne dispose pas d'information sur la stéréochimie.

Le programme OUTLSD contient un algorithme de génération de chaînes SMILES (Simplified Molecular-Input Line-Entry System). Il s'agit d'un format de codage très compact des structures développé par Weininger. Ce système de représentation est capable de traiter l'aromaticité, la stéréochimie et les différents isotopes d'un élément. Par exemple la chaîne SMILES générée par OUTLSD pour l'α-pinène est C12C(C)(C)C(C1)C(C)=CC2.

La figure 2.3 illustre le principe de codage des chaînes SMILES en prenant comme exemple la structure du pinène (A). Les hydrogènes sont supprimés, ils deviennent implicites (B). Les cycles sont ouverts et des numéros sont placés comme repères pour indiquer les liaisons manquantes. La structure est assimilée à un arbre. Le codage de la chaîne

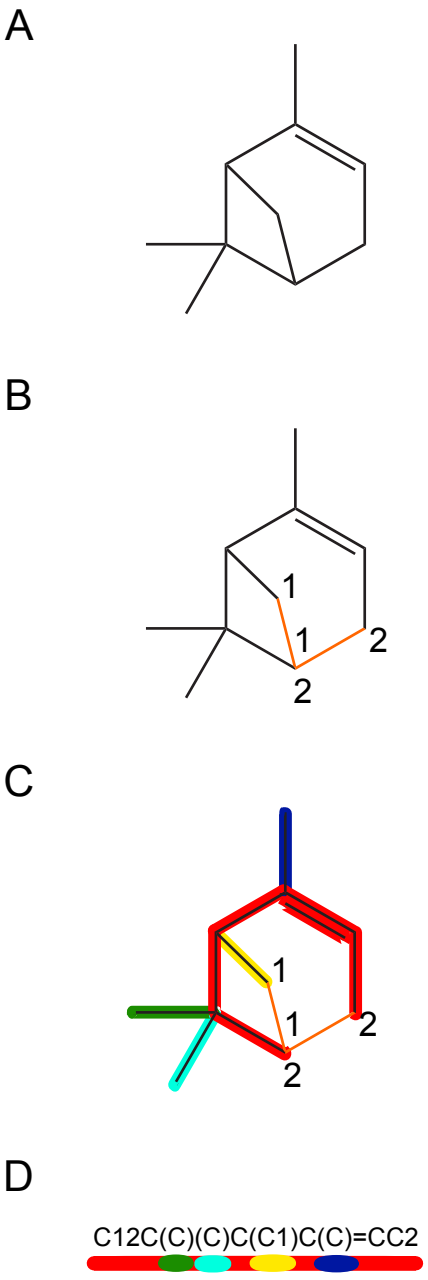


FIGURE 2.3 – Principe de codage d'une chaîne SMILES

SMILES est effectué en parcourant cet arbre (C). Le plus long enchaînement d'atome est repéré pour servir de base au codage. Cet enchaînement est parcouru en symbolisant les branches par des parenthèses pour obtenir la chaîne SMILES (D).

Le programme OUTLSD contient également un algorithme de calcul de coordonnées 2D et 3D à partir d'une table de connectivité entre atomes. Cet algorithme sera présenté en détail au paragraphe 2.3.5.

2.2.3.3 M_EDIT

Le programme M_EDIT est utilisé pour visualiser les structures 2D enregistrées au format MDL MOL ou dans un format de coordonnées spécifique au programme. Il permet d'améliorer la qualité visuelle des dessins des molécules quand les coordonnées produites par OUTLSD engendrent des structures difficiles à interpréter. Le programme M_EDIT est écrit en langage Tcl/Tk, il fournit l'interface de la figure 2.4.

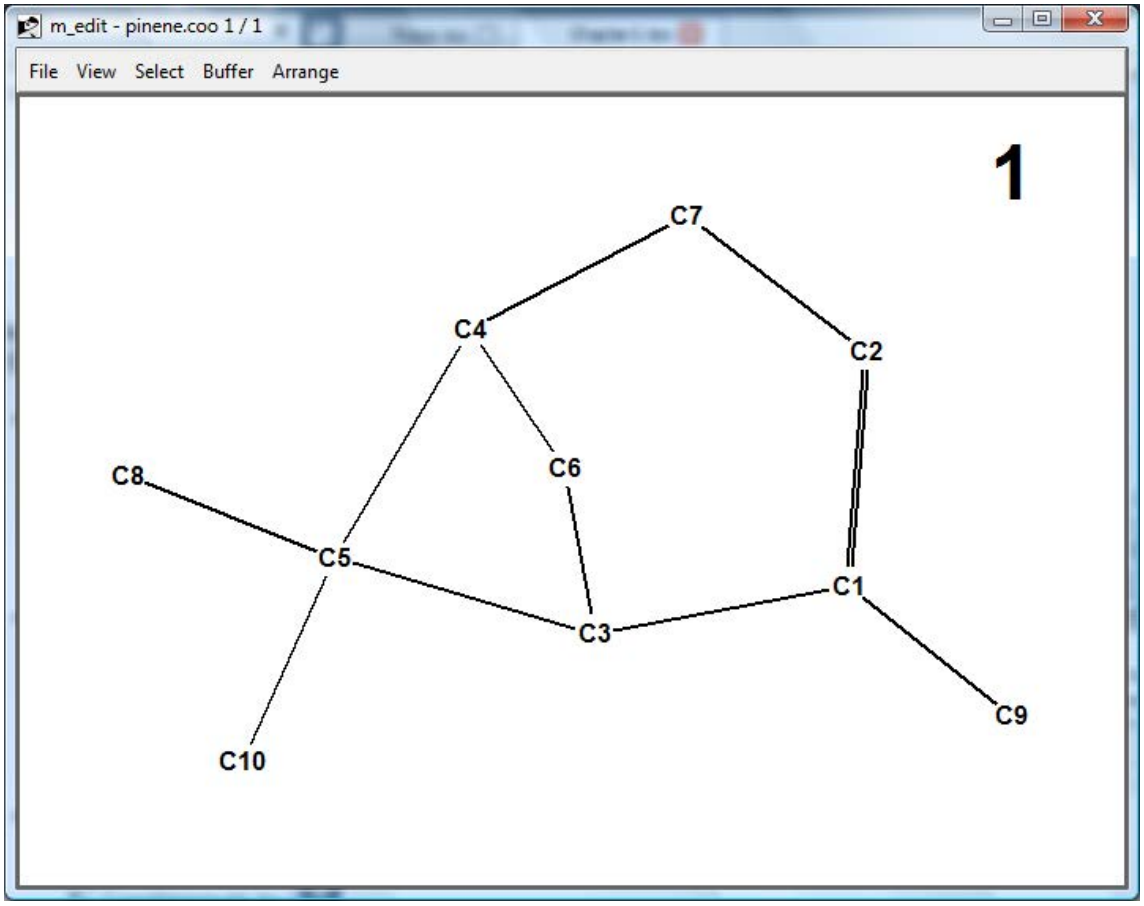


Figure 2.4 – Copie d'écran du programme M_EDIT

Le menu File permet de lire et de sauvegarder des fichiers et de sortir du programme.

Le menu View est utilisé pour naviguer en avant et en arrière dans un ensemble de structures groupées dans un même fichier. L'usage de M_EDIT est très simple : les atomes sont déplacés à l'aide du bouton gauche de la souris.

Les molécules sont toutes sélectionnées par défaut, comme indiqué par leur titre écrit en noir. Le menu Select permet de désélectionner/sélectionner des molécules. Les molécules sélectionnées peuvent être gardées (et les autres supprimées) en utilisant l'item Keep du menu Select.

Les molécules peuvent être inversées horizontalement ou verticalement à l'aide du menu Arrange.

Un buffer (menu Buffer) a été ajouté, de manière à ce qu'une situation antérieure préalablement sauvée (item Save To) puisse être rechargée (item Load From).

2.2.3.4 GENPOS

Le programme GENPOS, écrit en langage C, permet de créer un fichier contenant des instructions en langage PostScript. Le fichier résultant peut être ouvert par des visionneuses PostScript pour être analysé et imprimé.

2.2.4 Structure du logiciel LSD

2.2.4.1 Principe d'utilisation des corrélations

L'algorithme de résolution de LSD utilise les corrélations comme contraintes pour former des liaisons entre les atomes. La figure 2.5 montre comment les distances interatomiques mesurées en nombre de liaisons sont déduites à partir des corrélations des spectres 2D COSY et HMBC.

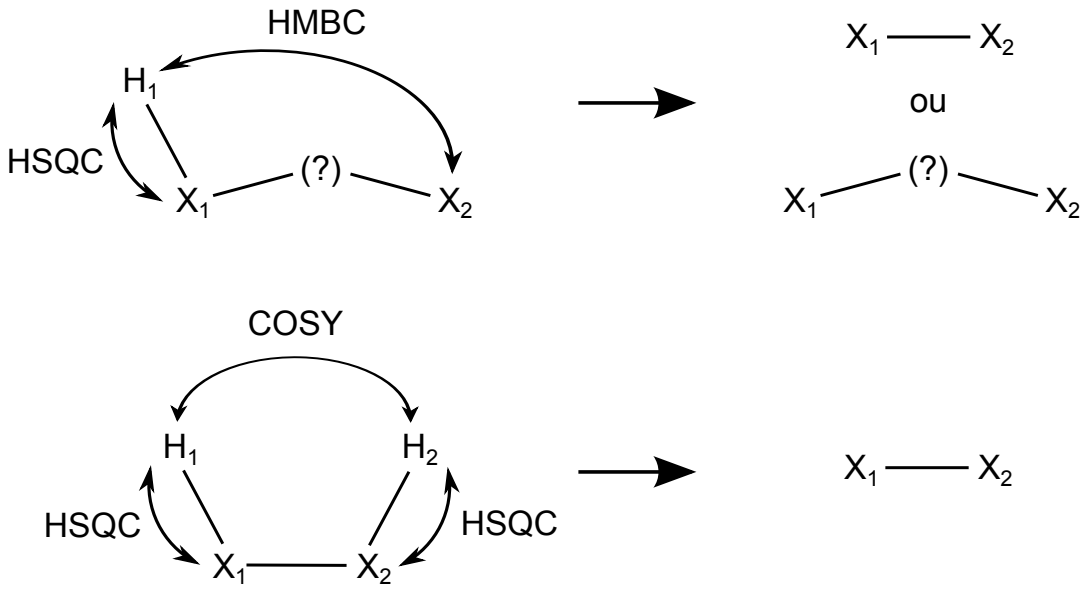


Figure 2.5 – Principe de formation des liaisons

Une corrélation HMBC révèle une relation de proximité entre deux atomes de carbone. Lorsqu'un proton H_1 corrèle avec un atome X_2 (X pouvant être un carbone ou un azote), cela indique qu'ils sont séparés par 2 ou 3 liaisons. Si le spectre HSQC montre que H_1 est directement lié à X_1, alors X_1 et X_2 sont soit liés ensemble soit séparés par un atome commun.

Le même principe est appliqué aux corrélations COSY à travers trois liaisons. Si deux protons H_1 et H_2 corrèlent et qu'ils sont liés respectivement aux atomes X_1 et X_2, alors il existe une liaison entre X_1 et X_2.

Chaque corrélation HMBC possède plusieurs niveaux d'indétermination. La déclaration d'une corrélation HMBC ambigüe entre un groupe de carbones et un proton implique qu'au moins un atome de carbone du groupe corrèle avec l'hydrogène. Il faut envisager tous les choix de carbone parmi la liste lors de l'interprétation d'une corrélation ambigüe. La seconde indétermination vient du fait que la longueur du chemin de couplage est inconnue. Il faut envisager que la corrélation soit le fruit d'un couplage ${}^2J$ ou ${}^3J$ ou alors d'un couplage à plus longue distance. La troisième indétermination intervient lors de l'hypothèse d'une corrélation HMBC ${}^3J$. L'atome voisin commun aux atomes X_1 et X_2 est inconnu. Il faut à nouveau envisager toutes les hypothèses plausibles.

Toutes les hypothèses possibles de ces indéterminations sont envisagées et explorées à l'aide d'un algorithme récursif qui permet l'exhaustivité de la résolution.

2.2.4.2 Principe d'un algorithme récursif

En informatique, un algorithme est dit récursif lorsque sa définition fait appel à l'algorithme lui-même.

La récursivité peut être utilisée pour effectuer des calculs mathématiques. Le calcul de la factorielle d'un nombre n est défini par la fonction suivante :

$$n! = \prod_{i=1}^{n} i = 1 \times 2 \times 3 \times \ldots \times (n-1) \times n$$

Ce calcul présenté sous la forme d'une suite donne pour $n \geq 0$:

$$n! = \begin{cases} 1 & \text{si } n = 0 \\ n \times (n-1)! & \text{sinon.} \end{cases}$$

Dans le deuxième cas, le calcul de $n!$ a besoin de la valeur de $(n-1)!$. La programmation de ce calcul nécessite un appel récursif, l'algorithme va effectuer une tâche identique en s'appelant lui-même plusieurs fois de suite. Pour éviter des appels récursifs à l'infini, il faut déterminer un cas où l'algorithme s'arrête. Ce cas est appelé cas de base, il s'appuie sur une condition de terminaison. Dans l'exemple, il s'agit du premier cas où $n = 0$. Le second cas qui contient l'appel récursif est appelé cas de propagation.

De nombreuses suites mathématiques peuvent être calculées de manière récursive. On peut citer par exemple la suite de Fibonacci qui est liée au nombre d'or $\frac{(1+\sqrt{5})}{2}$.

L'utilisation de la récursivité implique une mémorisation de l'état du problème à chaque appel récursif afin de récupérer un état précédent lors d'un retour en arrière. Cette mémorisation utilise une pile avec le principe LIFO (Last In First Out). Chaque donnée est stockée (empilée) dans son ordre d'arrivée. Lors de la récupération des données dans la pile, le premier élément récupéré (dépilé) est le dernier qui a été enregistré. Les anciennes générations de calculatrice ainsi que les langages PostScript et FORTH fonctionnent sur ce principe.

Le problème des 8 reines est un exemple de problème pouvant être résolu en utilisant la récursivité. Il s'agit de placer 8 reines sur un échiquier sans qu'aucune reine n'en menace une autre, sachant qu'une reine peut attaquer toute pièce se situant sur la même rangée, la même colonne et sur la même diagonale. La résolution du problème simplifié à 4 reines sur un échiquier 4×4 est présenté sur la figure 2.6. Les reines sont placées sur l'échiquier ligne par ligne dans les cases libres. Une case est libre si elle n'est menacée par aucune reine. Pour commencer, une reine est placée sur la première colonne de la première ligne. Ce choix ne laisse que deux possibilités pour le placement d'une reine sur la seconde ligne. Si la seconde reine est placée sur la troisième colonne, toutes les cases de la troisième ligne sont menacées. Suite à cet échec, il faut revenir à l'étape précédente et tester la seconde hypothèse de placement pour la seconde reine. Dans ce cas, la troisième reine peut être placée sur l'échiquier mais pas la quatrième. Les deux possibilités de placement de la seconde reine ayant été testées, il faut revenir en arrière jusqu'au placement de la première reine. Celle-ci est placée sur la deuxième colonne. Le placement successif des autres reines est effectué à chaque fois dans la seule case restant libre. Cette voie permet de déboucher sur une solution au problème. Il existe une seconde solution qui peut être obtenue en effectuant un retour en arrière jusqu'au placement de la première reine et en explorant les autres voies.

2.2.4.3 Phases de résolution

La génération des structures est réalisée au cours de quatre étapes successives (figure 2.7).

La première étape est appelée phase 0. Les informations fournies au programme sont lues dans le fichier d'entrée. Elles sont décodées au cours d'une analyse syntaxique. Après vérification de leur validité par une analyse sémantique, les données sont enregistrées. Certaines incohérences entre les données sont détectées à ce niveau avant de débuter la résolution. Les liaisons évidentes fournies par les commandes `BOND` et `COSY` sont formées.

La génération des structures débute réellement lors de la phase 1. Les corrélations HMBC sont exploitées et des liaisons sont établies en conséquence entre atomes. Les trois niveaux d'indétermination sont explorés par hypothèses successives. Lorsqu'aucun choix ne permet de fabriquer de nouvelles liaisons, c'est qu'un des choix effectués précédemment

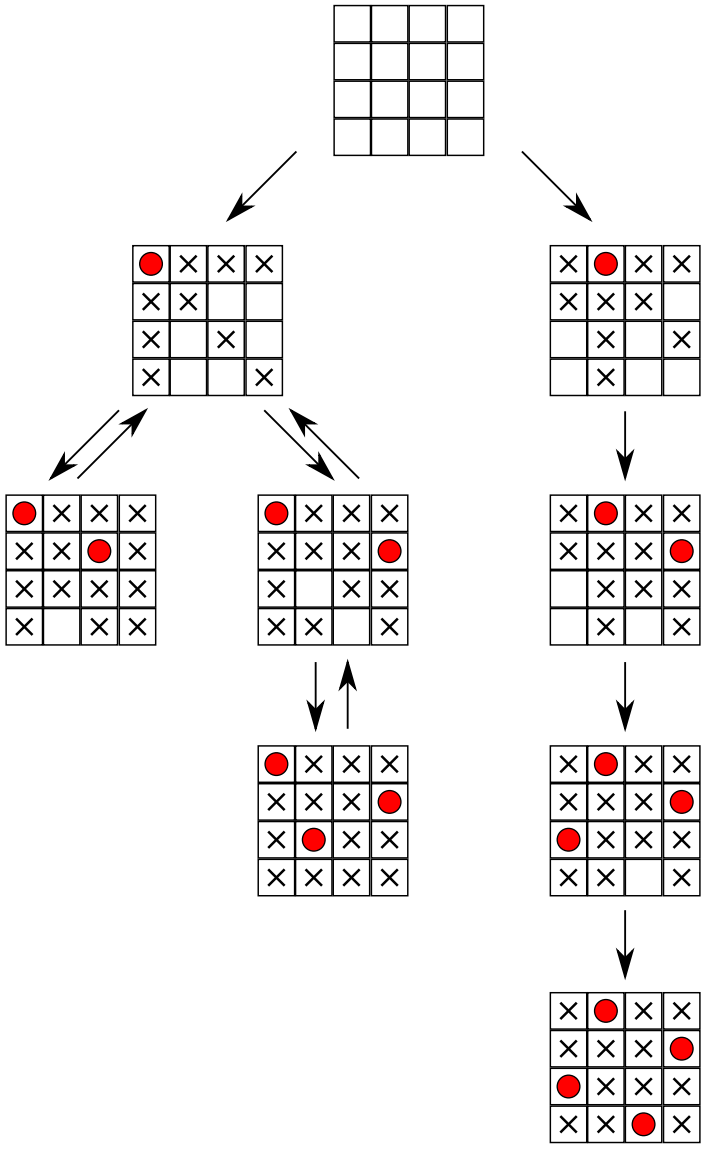


FIGURE 2.6 – Résolution du problème des 4 reines

était faux. Un choix précédent est alors remis en cause par un mécanisme de retour-arrière et l'analyse continue jusqu'à épuisement des corrélations.

La phase 2 consiste à apparier systématiquement des atomes incomplets. Il s'agit des atomes qui n'ont pas tous leurs voisins. La formation des liaisons peut ne pas être effectuée totalement. Le mécanisme de retour-arrière permet d'explorer l'ensemble des combinaisons possibles. Si toutefois le mécanisme ne permet pas d'aboutir à une structure complète, il y a retour à la phase précédente d'analyse des corrélations.

La dernière étape est la phase 3. C'est une phase de test. Si des corrélations ont été éliminées pendant la phase 1, il faut confirmer la validité de l'élimination. La distance entre les atomes impliqués dans les corrélations éliminées doit être en accord avec la longueur de chemin de couplage maximale entrée par l'utilisateur. Les liaisons multiples sont placées avec un mécanisme de retour-arrière. Un algorithme détermine si la structure satisfait à la règle de Bredt [148]. Les éventuelles contraintes de sous-structures sont vérifiées à l'aide d'un algorithme de recherche de sous-structure [149]. L'ultime point est la vérification de l'originalité de la structure, en la confrontant aux solutions précédemment obtenues. En cas de succès, la structure est écrite dans le fichier de sortie.

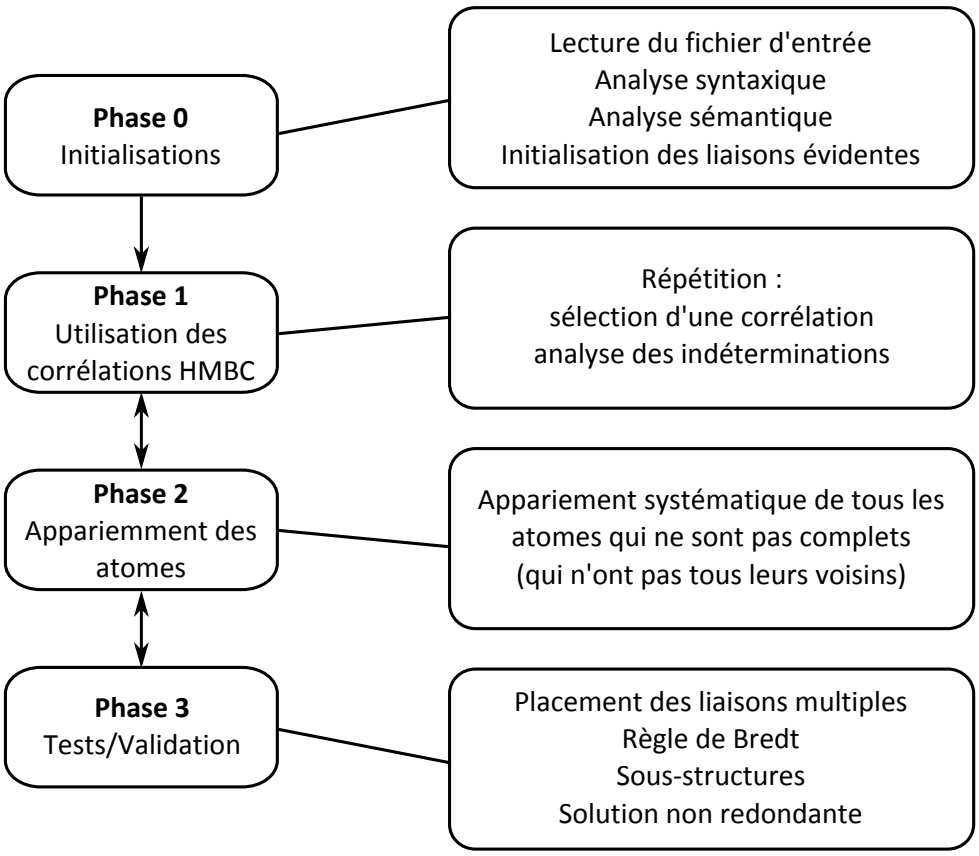


FIGURE 2.7 – Méthode de résolution des structures

2.2.5 Exemples d'applications de LSD

Le logiciel LSD a été utilisé principalement dans l'élucidation structurale de substances naturelles. La figure 2.8 montre quelques exemples de structures étudiées avec LSD [150–157].

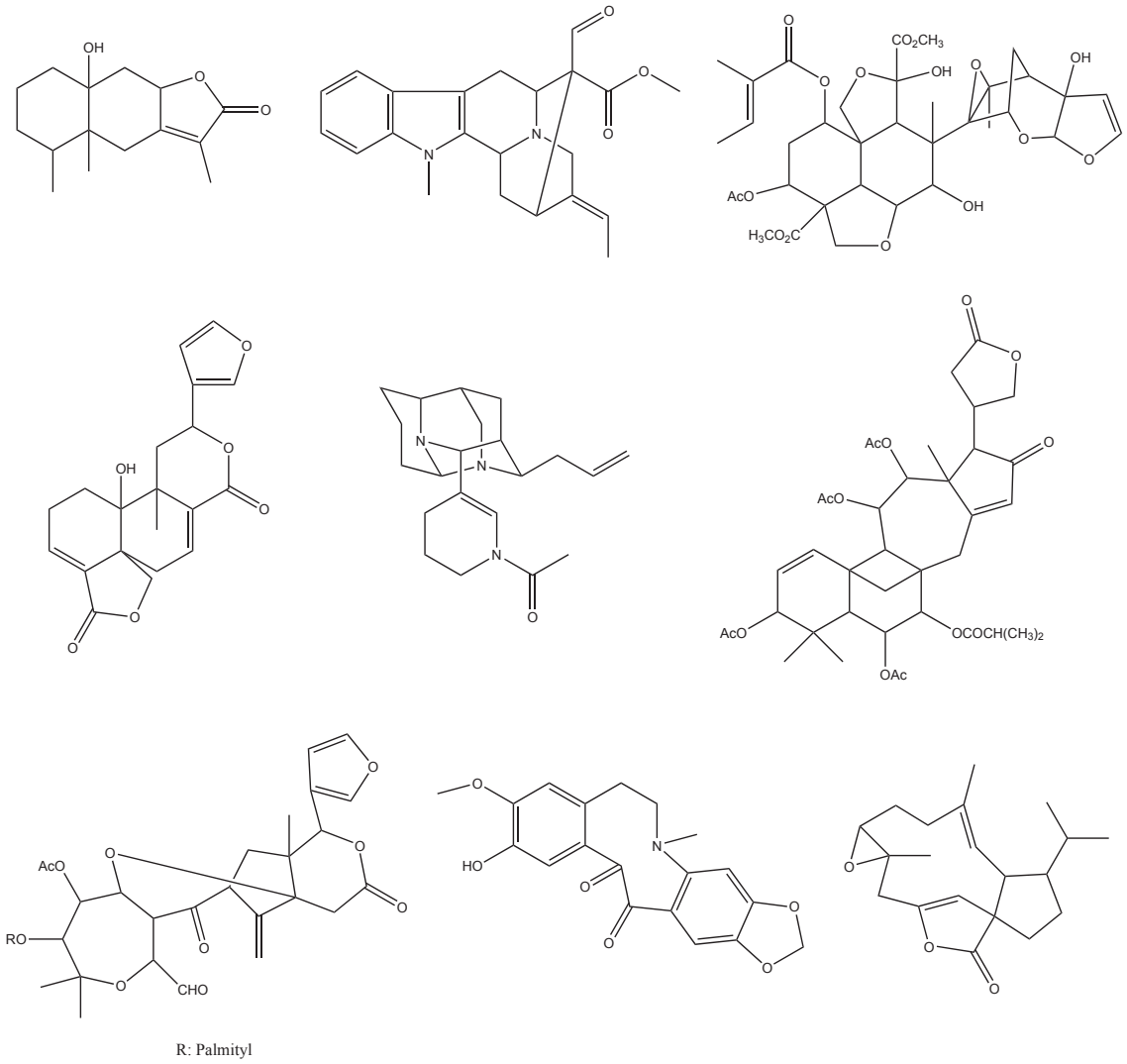


Figure 2.8 – Structures étudiées à l'aide de LSD

2.3 Développements récents

2.3.1 Introduction

Les développements récents visent à améliorer plusieurs points comme la diversité des molécules analysables, la souplesse d'utilisation, l'utilisation des corrélations ou la présentation des résultats.

Les développements touchant directement le générateur de structure ont été effectués en modifiant le code existant. La structure du logiciel et l'organisation interne des données ont été gardées en aménageant toutes les parties nécessaires. La programmation étant l'art de corriger des bugs sans en ajouter de nouveaux, une attention particulière a été portée aux dommages collatéraux que chaque modification pouvait apporter. De plus, les nouveautés ont été ajoutées avec le souci de garder une compatibilité avec les anciens jeux de données.

2.3.2 Traitement amélioré des corrélations HMBC et COSY

2.3.2.1 Objectif

Le but de cette modification est de permettre à l'utilisateur d'introduire des valeurs de longueur de chemin de couplage de manière optionnelle lors de la déclaration des corrélations HMBC et COSY. La finalité est de permettre d'affiner le traitement lors de la résolution en réduisant le nombre d'hypothèses et donc le temps de résolution. Potentiellement, la probabilité de présence de corrélations COSY ambigües est forte du fait de la faible gamme de déplacements chimiques ^{1}H. Nous avons profité de ces aménagements pour ajouter la possibilité de déclarer des corrélations COSY comportant des groupes d'atomes ayant des déplacements chimiques proches.

2.3.2.2 État des lieux avant modification

Avant cette modification, suivant la nomenclature du fichier d'entrée de LSD, la commande HMBC possédait uniquement deux paramètres. Après changements, ces deux paramètres restent toujours obligatoires. Le premier peut être un entier correspondant à un numéro de carbone ou une liste de carbones écrite entre parenthèses. Le second est un entier correspondant à un numéro d'hydrogène. S'il existe une incertitude car des atomes de carbone ont des déplacements chimiques proches, l'utilisateur peut définir une corrélation ambigüe ou variable. Par exemple, `HMBC (2 9) 5` signifie que soit le carbone 2, soit le 9 (ou les deux) corrèle(nt) avec l'hydrogène 5.

D'autre part, la commande COSY possède aussi deux paramètres obligatoires. Ce sont deux entiers correspondant chacun à un numéro d'hydrogène. Les corrélations COSY étaient jusqu'ici considérées comme provenant de couplages $^3J_{H-H}$, c'est-à-dire des corrélations entre deux atomes d'hydrogènes H_1 et H_2 portés par deux atomes de squelette voisins X_1 et X_2. Ces corrélations intervenaient en tout début de la résolution (phase 0) pour former directement la liaison X_1–X_2. Il est cependant courant que des informations de couplage à plus longue distance apparaissent sur ces spectres. Il ne fallait alors considérer que les signaux les plus intenses des spectres COSY pour ne pas aboutir à un échec de la résolution. Ignorer certaines corrélations à longue distance peut constituer un facteur limitant dans le processus de résolution. En effet certaines de ces informations peuvent

jouer un rôle dans le processus de résolution. Afin de palier à ce manque, il a été décidé de modifier le programme afin de permettre à l'utilisateur d'introduire des corrélations COSY à longueur de chemin de couplage variable pour qu'elles puissent intervenir lors de l'analyse des corrélations en phase 1.

2.3.2.3 Mise en œuvre des modifications

L'intervention de LSD sur les corrélations (du fichier d'entrée à leur utilisation pendant la résolution) va être détaillée dans la suite en décomposant les différents traitements étape par étape.

Syntaxe du fichier d'entrée L'utilisateur a la possibilité d'ajouter un ou deux paramètres optionnels à la suite de la déclaration de la corrélation. Un paramètre optionnel unique représente un nombre de liaisons entre deux atomes qui corrèlent. Si deux paramètres optionnels sont présents, ils s'agit de bornes inférieure et supérieure pour le nombre de liaisons. Les exemples suivants illustrent plusieurs cas pouvant se présenter :

- `HMBC 15 5 2` : l'atome 15 corrèle avec l'hydrogène 5 à travers 2 liaisons ;
- `HMBC 15 5 4` : l'atome 15 corrèle avec l'hydrogène 5 à travers 4 liaisons (n'est possible que si une commande `ELIM` est présente) ;
- `HMBC 15 5 2 3` : l'atome 15 corrèle avec l'hydrogène 5 à travers 2 ou 3 liaisons.

Dans le second cas, si la commande `ELIM` est absente, le programme s'arrête et signale le problème avec un message d'erreur. Dans le dernier cas, le programme ne tentera jamais d'éliminer la corrélation, même si une commande `ELIM` est présente.

Les corrélations COSY sont aussi concernées par l'ajout de paramètres optionnels. Par conséquent, une corrélation COSY ${}^{n}J_{H-H}$ peut être éliminée si les valeurs de n le permettent. Par exemple, la commande `ELIM 3 4` signifie que trois corrélations correspondant au maximum à des couplages ${}^{4}J_{C-H}$ ou ${}^{5}J_{H-H}$ peuvent être éliminées. Il est à noter que, par défaut, LSD considère une corrélation COSY sans paramètre optionnel comme une corrélation ${}^{3}J_{H-H}$.

La figure 2.9 résume l'ensemble des corrélations types qu'il est possible de déclarer. Les liaisons pouvant être déduites sont équivalentes. En effet l'information apportée par une corrélation HMBC ${}^{n}J$ est équivalente à une corrélation COSY ${}^{n+1}J$.

On remarque le comportement différent des cas de corrélations par défaut, sans paramètre optionnel, vis-à-vis de l'élimination. Une corrélation HMBC peut être éliminée, alors qu'une corrélation COSY ne peut pas l'être.

Les corrélations de la quatrième ligne ayant un seul paramètre optionnel n sont obligatoirement supprimées. En tant que corrélations à très longue distance, elles ne sont pas analysées pour la formation des liaisons. À première vue ces corrélations peuvent paraître inutiles. Toutefois lorsqu'elles sont éliminées, le programme vérifie que la distance entre les atomes est conforme à la longueur du chemin de couplage n.

Cas	Liaisons déduites	Élim.	Cas	Liaisons déduites	Élim.
HMBC X Y	X−Y ou X−(?)−Y	✓	COSY X Y	X−Y	✗
HMBC X Y 2	X−Y	✗	COSY X Y 3	X−Y	✗
HMBC X Y 3	X−(?)−Y	✗	COSY X Y 4	X−(?)−Y	✗
HMBC X Y n ($n > 3$)	X−$(?)_{n-2}$−Y	✓	COSY X Y n ($n > 4$)	X−$(?)_{n-3}$−Y	✓
HMBC X Y 2 3	X−Y ou X−(?)−Y	✗	COSY X Y 3 4	X−Y ou X−(?)−Y	✗
HMBC X Y 2 n ($n > 3$)	X−Y ou X−$(?)_{n-2}$−Y	✓	COSY X Y 3 n ($n > 4$)	X−Y ou X−$(?)_{n-3}$−Y	✓

FIGURE 2.9 – Tableau récapitulatif des corrélations types : les différents cas pouvant être déclarés dans un fichier d'entrée sont recensés (première colonne) avec les liaisons pouvant être déduites des corrélations (deuxième colonne). La troisième colonne indique si la corrélation peut être éliminée lors de la résolution. (Remarque : les atomes de carbone et d'hydrogène directement liés possèdent des numéros X et Y identiques)

Un cas particulier n'est pas présenté dans le tableau. La borne supérieure optionnelle peut être déclarée comme égale à zéro, ce qui signifie une valeur infinie. Dans ce cas, la borne supérieure prend la valeur maximale de liaisons entre atomes donnée par la commande `ELIM`. On se retrouve alors dans le cas de la dernière ligne du tableau.

Analyse sémantique des corrélations (phase 0) Suite à la lecture du fichier d'entrée, l'analyse sémantique des commandes permet de vérifier le bon respect de la nomenclature de LSD. À l'issue de cette analyse, les corrélations sont enregistrées et subissent les mêmes traitements par la suite. Afin de faciliter la résolution, les corrélations HMBC et COSY sont transformées en corrélations entre atomes de squelette. Il s'agit le plus souvent de carbones. Les hydrogènes deviennent donc totalement implicites pendant la résolution. Pour ce faire, le programme utilise l'information de la commande `HSQC`.

Les règles suivantes sont appliquées aux corrélations non ambigües pendant l'analyse sémantique :

- Les atomes d'hydrogène doivent être définis par ailleurs avec la commande `HSQC` ;
- Les atomes de carbone doivent être définis par ailleurs avec la commande `MULT` ;
- Après conversion des numéros d'hydrogène en numéros d'atomes de squelette, il faut vérifier si les deux atomes de squelette provenant d'une corrélation sont bien différents car un atome ne peut pas corréler avec lui-même.

Si une des règles n'est pas respectée, l'utilisateur est averti par un message d'erreur décrivant la cause du problème.

Ces règles permettent de vérifier les erreurs de frappe dans le fichier d'entrée. La troisième règle possède une exception, les corrélations COSY entre protons géminés (groupements CH_2) sont écartées silencieusement du jeu de données. En effet les corrélations COSY ^2J n'apportent aucune information pour la construction des structures.

En ce qui concernent les corrélations ambigües entre une liste d'atomes et un hydrogène, les deux premières règles restent valables. Dans le cas d'une liste d'hydrogènes, il faut vérifier s'ils sont tous définis par ailleurs avec la commande `HSQC`. Dans le cas d'une liste de carbones, il faut vérifier s'ils sont tous définis par ailleurs avec la commande `MULT`. La troisième règle ne provoque pas de message d'erreur. Si un atome corrèle avec lui-même, il est enlevé de la liste. Si la liste contenait deux éléments, la corrélation devient une corrélation non variable. Sinon une nouvelle liste est créée.

Il s'agit ensuite de vérifier si les longueurs de chemin de couplage sont compatibles avec le type de corrélation (COSY ou HMBC).

Lorsqu'une des règles suivantes n'est pas respectée, l'utilisateur est informé par un message d'erreur :

- Les deux paramètres doivent être différents de 1 ;
- S'il s'agit d'une corrélation COSY, les paramètres doivent être différents de 2 ;
- La borne inférieure ne doit pas être nulle ;
- Si la borne supérieure n'est pas nulle ou absente, elle doit être supérieure ou égale à la borne inférieure ;
- Si une commande `ELIM` est présente, la borne supérieure doit être inférieure ou égale à la longueur autorisée.

À partir de ce moment, les corrélations sont transformées. Les bornes inférieures et supérieures sont ramenées à des distances entre atomes de squelette. Elles sont alors diminuées de 2 dans le cas d'une corrélation COSY et de 1 dans le cas d'une corrélation HMBC.

À partir des valeurs de bornes inférieures et supérieures, les corrélations sont marquées comme pouvant ou non être interprétées par la suite comme des corrélations fictives ^1J, ^2J ou ^nJ (éliminable) entre atomes de squelette. En réalité, ce sont des nombres de liaisons séparant les atomes de squelette. La figure 2.10 recense les 6 types de corrélations en terme de distance entre atomes de squelette.

Une corrélation ne peut évidemment pas être uniquement éliminable en l'absence d'une commande `ELIM`. Ce cas engendre un message d'erreur demandant à l'utilisateur d'ajouter une commande `ELIM` au fichier d'entrée.

Prétraitement des corrélations (phase 0) Le but du prétraitement est de vérifier, avant la génération des structures, que les données sont cohérentes et qu'elles ne comportent pas de conflits. Ceci permet d'éviter de lancer le processus de la phase 1 inutilement.

Les corrélations sont comparées les unes aux autres et confrontées aux liaisons déjà

Nombre de liaisons entre atomes lourds		
1	**2**	$n > 2$
✓		
	✓	
		✓
✓	✓	✓
	✓	✓
✓	✓	

Figure 2.10 – Tableau récapitulatif des 6 corrélations types en terme de distances potentiellement déduites entre atomes de squelette

établies par la commande `BOND` pour gérer les conflits d'information et supprimer les corrélations qui n'apportent aucune information supplémentaire. Un conflit peut aboutir soit à un message d'erreur, soit à la modification d'une corrélation. Les corrélations inutiles sont marquées comme non valides, le programme ne cherchera pas à les utiliser pendant la phase 1.

Le prétraitement des corrélations est différencié avec d'abord les corrélations non variables et ensuite les corrélations variables.

La figure 2.11 présente les tests (a-g) subis par les corrélations non variables.

Pour commencer, les corrélations non variables sont comparées entre elles (a et b). Le but est d'éviter d'avoir des corrélations redondantes. Le programme recherche les corrélations dupliquées entre deux atomes X et Y et vérifie si les intervalles de nombre de liaisons entre atomes sont compatibles. Pour ce faire, l'intersection des intervalles est calculée. Si l'intersection n'est pas vide, il ne subsiste qu'une corrélation valide avec comme intervalle l'intersection des deux corrélations (a). Si l'intersection est un ensemble vide, il y a incompatibilité entre les deux corrélations et le programme s'arrête en prévenant l'utilisateur par un message d'erreur (b).

La partie suivante est la formation de liaisons lorsqu'il n'y a aucune ambigüité possible. Le programme construit des liaisons entre atomes qui corrèlent et qui ne peuvent être séparés que par une liaison uniquement (c). Par exemple, c'est le cas des corrélations COSY entrées sans aucun paramètre optionnel.

Enfin, le programme compare toutes les corrélations non variables encore valides avec les liaisons déjà établies par ailleurs avec la commande `BOND` (d-g). Il s'agit de supprimer les corrélations qui n'apportent pas d'information. S'il existe une corrélation qui autorise une distance de 1 liaison entre deux atomes liés, celle-ci est supprimée (d). Au contraire s'il existe une corrélation qui n'autorise pas deux atomes liés à l'être, celle-ci entre en conflit avec la liaison (e). De même, s'il existe une corrélation qui autorise une distance

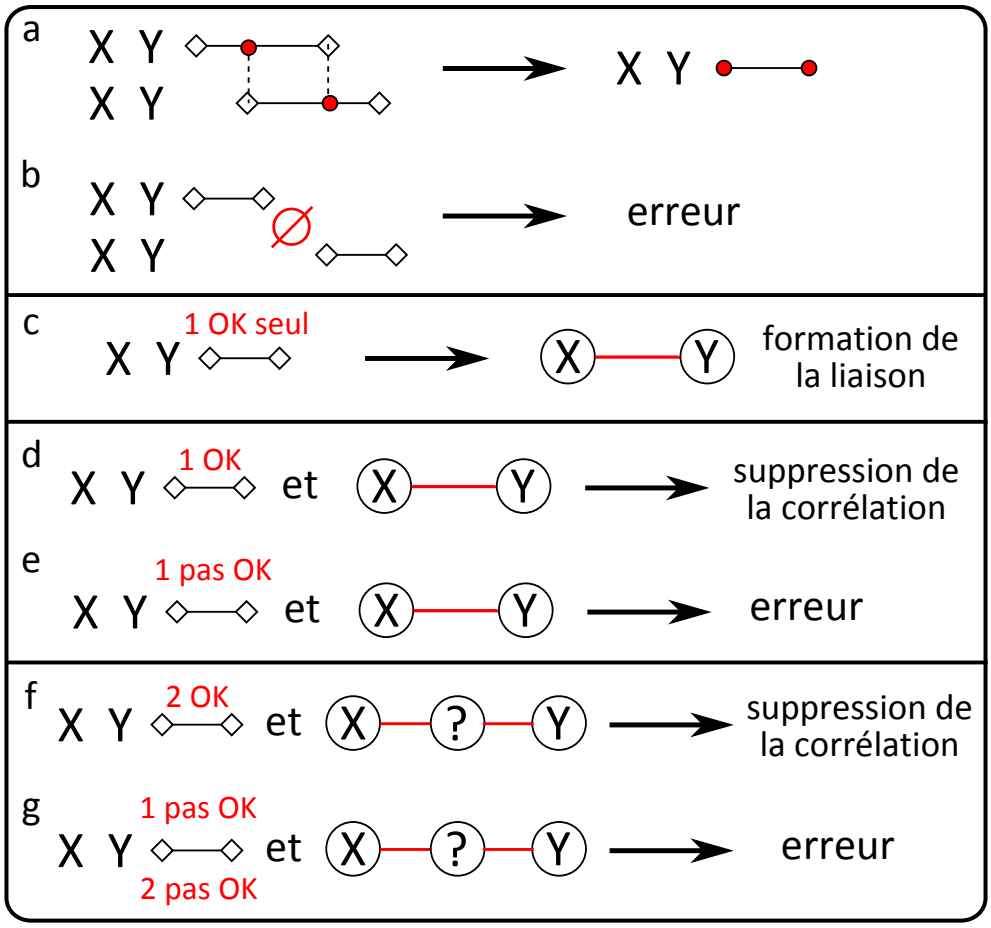


FIGURE 2.11 – **Prétraitement des corrélations non variables :** le comportement du programme en présence de différents cas est résumé de manière synthétique. Pour le détail, voir texte. (Remarque : les atomes de carbone et d'hydrogène directement liés possèdent des numéros X et Y identiques)

de 2 liaisons entre deux atomes séparés par deux liaisons, elle est rendue non valide (f). Si une corrélation n'autorise pas une distance de 1 ou 2 liaisons, alors il y a un conflit avec les liaisons établies (g). Une corrélation qui n'autorise pas une distance de 2 liaisons entre 2 atomes X et Y ayant un voisin commun mais accepte une distance de 1 liaison est gardée car il est possible de lier X et Y pour former un cycle de trois atomes. Un désaccord entre une corrélation et des liaisons existantes entraine l'arrêt du programme et l'affichage d'un message d'erreur.

La figure 2.12 présente les tests (a-h) subis par les corrélations variables. Ces corrélations concernent des listes d'atomes ayant des déplacements chimiques proches.

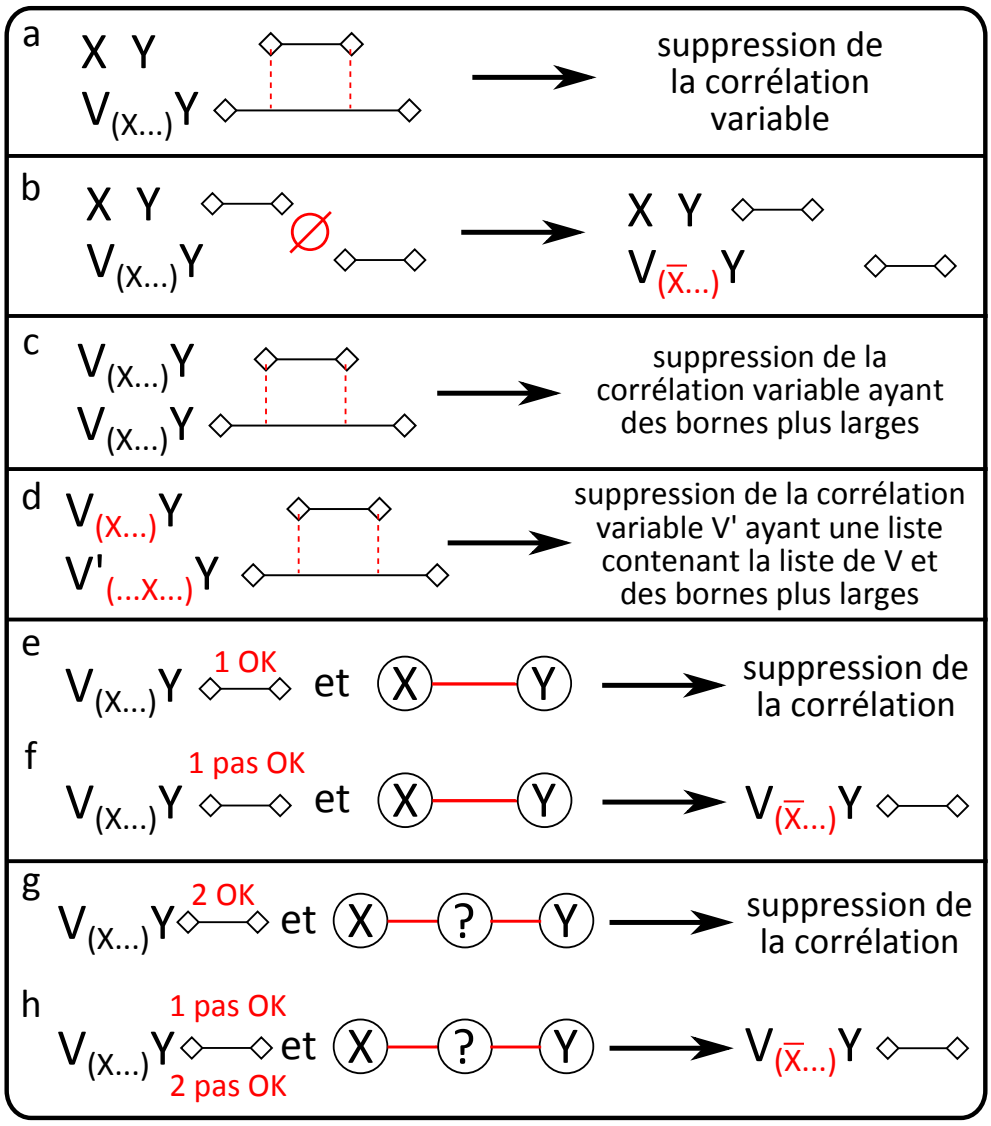


FIGURE 2.12 – Prétraitement des corrélations variables : le comportement du programme en présence de différents cas est résumé de manière synthétique. Pour le détail, voir texte. (Remarque : les atomes de carbone et d'hydrogène directement liés possèdent des numéros X et Y identiques)

Le programme commence par vérifier si l'indétermination sur une corrélation variable entre Y et une liste d'atomes V qui contient X peut être expliquée (a et b). Une corrélation hypothétique entre X et Y peut avoir été déclarée par ailleurs sous la forme d'une corrélation non variable. Les distances minimale et maximale autorisées pour la corrélation non variable doivent être égales ou incluses dans celles de la corrélation variable (a). Si tel est le cas, la corrélation variable est supprimée. Par contre si l'intersection des distances autorisées est nulle, il y a contradiction entre les corrélations (b). L'information de la corrélation non variable est considérée comme prioritaire par rapport à celle de la corrélation variable. L'atome X apportant la contradiction est supprimé de la liste V de la corrélation variable. Lorsqu'une liste est réduite à un unique atome, la corrélation variable devient une corrélation non variable. En retirant une alternative, l'indétermination sur la corrélation est simplifiée voire levée. Cela permet au programme d'aborder la phase 1 avec un jeu de données ayant le moins d'ambigüité possible.

Dans la partie suivante, les corrélations variables sont comparées entre elles (c et d). Dans un premier temps, il faut éliminer les corrélations variables dupliquées, c'est-à-dire dont les listes d'atomes V sont identiques. Il faut également que les distances autorisées soient identiques ou qu'elles soient incluses. Dans ce dernier cas, la corrélation ayant les bornes les plus larges est éliminée. Dans un second temps, on procède à l'élimination des corrélations variables expliquées par d'autres corrélations variables. Pour qu'une corrélation soit expliquée, une seconde corrélation doit être totalement incluse dans celle-ci (d). La liste d'atomes de la première (V') doit contenir la totalité de la seconde (V) et les bornes de distance de la première doivent être identiques ou plus larges que celles de la seconde.

Pour la dernière partie, de la même manière qu'avec les corrélations non variables, le programme supprime les corrélations variables qui n'apportent pas de nouvelle information par rapport aux liaisons déjà existantes (e-h). En cas de conflit d'une corrélation hypothétique entre X et Y avec les liaisons, l'atome X est supprimé de la liste (f et h).

Choix et élimination des corrélations pendant la résolution (phase 1) La vérification étant terminée, le programme va travailler avec un jeu de données non redondantes et cohérentes.

Le générateur de structure choisit les corrélations et établit des liaisons. Lorsqu'une corrélation est utilisée, elle est rendue non valide. Quand une ou plusieurs liaisons sont formées suite à l'utilisation d'une corrélation, le programme détecte et rend également invalides toutes les corrélations qui deviennent inutiles car n'apportant plus d'informations supplémentaires. Ces corrélations expliquées doivent être éliminées sinon le programme peut se retrouver dans une situation où il essaie de former une ou des liaison(s) déjà formée(s). Une corrélation variable peut être éliminée si au moins un atome du groupe permet d'expliquer la corrélation.

La présence de longueurs de chemin de couplage dans la définition des corrélations

donne également un autre rôle aux corrélations. Celles-ci ont la possibilité d'interdire la formation d'une ou plusieurs liaisons si les longueurs de couplage sont en désaccord avec les distances entre atomes. La façon dont le programme envisage toutes les hypothèses de façon exhaustive va être décrite dans la suite.

Tout d'abord, lorsque la corrélation choisie est variable, le programme considère successivement toutes les corrélations hypothétiques entre chacun des atomes de la liste et l'atome d'hydrogène. Ensuite, si les propriétés d'une corrélation entre X et Y le lui permettent, il est fait l'hypothèse qu'elle correspond à une distance de 1 puis de 2 entre ces deux atomes de squelette. Enfin, si une commande `ELIM` est présente, la corrélation est éliminée. Elle ne participe pas à l'élucidation mais jouera un rôle dans la dernière phase de la résolution.

La première hypothèse a pour conséquence la formation de la liaison entre les deux atomes X et Y (figure 2.13 a). Le programme repère alors toute corrélation devenue inutile entre un atome Z, déjà lié à X, avec Y et inversement. Il faut également rendre invalide toute corrélation qui subsisterait entre les atomes X et Y.

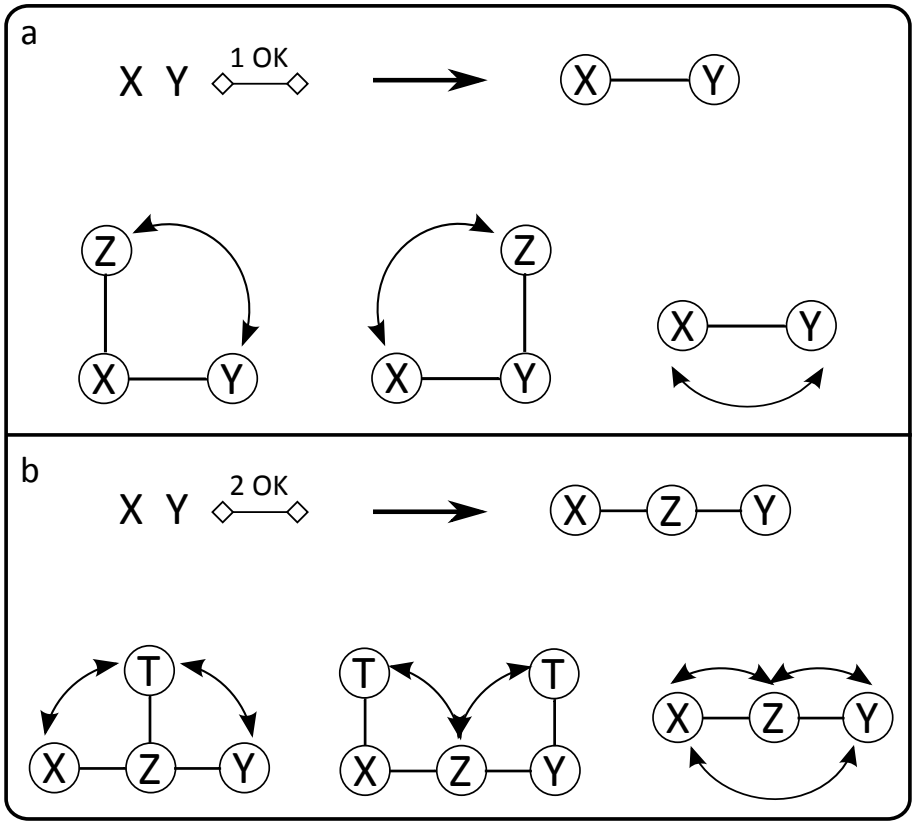


Figure 2.13 – Schéma représentant tous les cas d'éliminations de corrélations devenant inutiles pendant la résolution

La seconde hypothèse conduit à séparer X et Y par un atome commun Z (figure 2.13 b). Le choix de cet atome Z constitue une indétermination qu'il faut envisager de manière successive et exhaustive. Le programme repère toute corrélation devenue inutile entre

un atome T lié à Z avec les atomes X et Y. Il faut aussi supprimer toute corrélation entre l'atome Z avec un atome T lié à X ou à Y. Il faut également rendre invalide toute corrélation qui subsisterait entre les atomes X et Y.

Toutes les corrélations devenues inutiles suite à la formation de liaisons sont considérées comme des corrélations de confirmation.

Cependant, l'apparition des paramètres optionnels a donné naissance à une nouvelle notion inexistante dans les versions antérieures de LSD : la notion de corrélation d'infirmation (en opposition aux corrélations de confirmation). Dans la nouvelle implémentation, le programme est capable de réfuter une hypothèse si les longueurs précisées pour les chemins de couplage sont en désaccord avec le nombre de liaisons entre les atomes. Le programme effectue alors un retour arrière et continue avec d'autres hypothèses. Le retour arrière a lieu dans deux cas :

- Si deux atomes X et Y sont liés et qu'une corrélation entre X et Y n'autorise pas une distance de une liaison ;
- Si deux atomes X et Y ont un voisin commun et qu'une corrélation entre X et Y n'autorise pas une distance de 1 ou 2 liaison.

Dans le second cas, une distance de 1 liaison est envisageable car en liant X et Y il y a formation d'un cycle de trois atomes. La corrélation ne confirme et n'infirme pas l'hypothèse en cours. Elle reste valide pour être interprétée par la suite.

On notera que seule une corrélation non variable est capable d'infirmer une hypothèse. En effet, il est impossible à ce niveau de faire une hypothèse sur l'atome responsable de l'observation d'une corrélation variable.

Vérification des distances entre atomes pour les éventuelles corrélations éliminées (phase 3) Ce point concerne les corrélations éliminées dans la phase 1 par la présence de la commande `ELIM`. La vérification a lieu pendant la dernière phase de tests lorsqu'une structure complète est obtenue. Lorsqu'une structure est complète, c'est-à-dire lorsque tous les atomes sont complets, il faut vérifier si les distances entre les atomes sont bien conformes aux longueurs autorisées pour les corrélations éliminées.

Le programme calcule la matrice des distances (nombres de liaisons) entre les atomes et effectue une comparaison par rapport aux propriétés des corrélations. La distance entre les atomes doit être comprise entre les bornes des distances autorisées. Si les contraintes de distances ne sont pas respectées, la structure est rejetée, le programme effectue un retour en arrière.

2.3.2.4 Exemple d'application

L'exemple est un sesterterpène de formule brute $C_{25}H_{36}O_3$ (figure 2.14). Il s'agit d'une structure originale isolée par l'équipe du Pr. Dulcie A. Mulholland de l'Université de Surrey et élucidée avec l'aide de LSD. Le fichier d'entrée de LSD est présenté en annexe B. L'analyse des spectres ^{13}C et DEPT nous indique la présence de 5 groupements CH_3, 9

groupements CH_2, 3 groupements CH et 8 carbones quaternaires. Les signaux ^{13}C sont numérotés par ordre de déplacements chimiques décroissants. Les protons portent les mêmes numéros que les carbones auxquels ils sont directement liés. Une première analyse des déplacements chimiques ^{13}C permet de soupçonner la présence d'une cétone (C1 à δ 213,1) et d'un groupement ester (C2 à δ 177,1). Une analyse plus fine des déplacements chimiques nous indique la présence d'une double liaison (entre les carbones quaternaires 3 et 4 à respectivement δ 135,6 et δ 132,9). Toutes les liaisons facilement déduites ont été déclarées dans le fichier d'entrée de LSD (liaisons C=O 1–26 et 2–27, liaison C–O 2–28 et liaison C=C 3–4). Le déplacement chimique du carbone quaternaire 5 est 98,2 ppm. Cette valeur est caractéristique d'un carbone lié à deux oxygènes mais aucun autre signal ^{13}C ne permet de suspecter la présence d'un quatrième atome d'oxygène. Ce carbone a été déclaré lié à l'unique oxygène ayant encore une liaison libre (liaison C–O 5–28). D'après la structure obtenue, le déplacement chimique du carbone 5 est expliqué par son caractère spiranique.

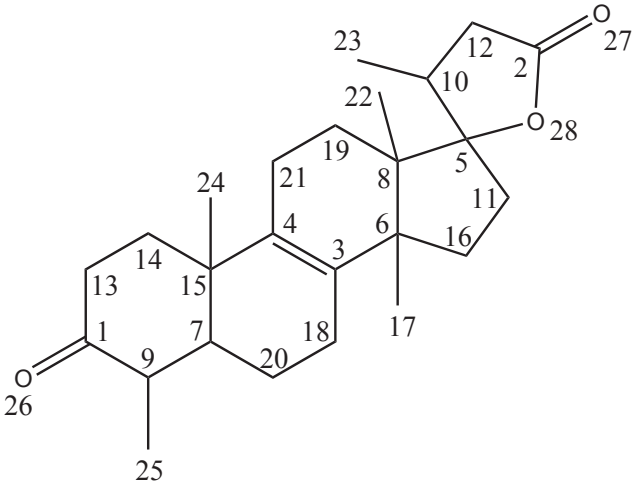


Figure 2.14 – Structure du sesterterpène

Cet exemple contient des corrélations à très longue distance. Le spectre HMBC contient deux corrélations liées à des couplages ^4J (H25→C20 et H13→C25). Le fichier d'entrée contient une commande `ELIM` qui permet au programme d'éliminer 2 corrélations $^4J_{C-H}$ ou $^5J_{H-H}$. Sans cette commande aucune solution n'est produite.

Le spectre COSY contient également des corrélations à longue distance. Les protons des groupements méthyles 17 (δ (^{13}C) 25,7 et δ (^1H) 1,13) et 24 (δ (^{13}C) 17,5 et δ (^1H) 1,23) apparaissent sous forme de singulets. Les carbones 17 et 24 sont donc liés à des carbones quaternaires. L'observation de corrélations COSY impliquant les protons 17 et 24 provient nécessairement de couplages à longue distance. Dans le fichier d'entrée, les corrélations COSY H24→H14 et H17→H16 ont donc été déclarées comme provenant de couplages $^4J_{H-H}$ au minimum (borne supérieure indéfinie). Les carbones 24 et 14 d'une

part et 17 et 16 d'autre part doivent donc être séparés par au minimum 2 liaisons. Le programme fixe automatiquement la limite supérieure à 3 liaisons en tenant compte de la valeur maximum prévue par la commande `ELIM`. Des corrélations HMBC impliquant ces atomes sont également dans le jeu de données. Ces corrélations sont déclarées sans spécifier de longueurs de chemin de couplage. Elles imposent donc des distances de 1 à 3 liaisons entre les carbones. Lors du prétraitement des corrélations, le recoupement des informations des corrélations COSY et HMBC enlève l'hypothèse d'une distance de 1 liaison entre les carbones. Lors de la formation des liaisons par l'analyse des corrélations, le programme évitera donc cette hypothèse. Dans ce cas, la déclaration des corrélations COSY ${}^4J_{H-H}$ permet au programme de ne pas perdre de temps à traiter une hypothèse n'aboutissant à aucune solution.

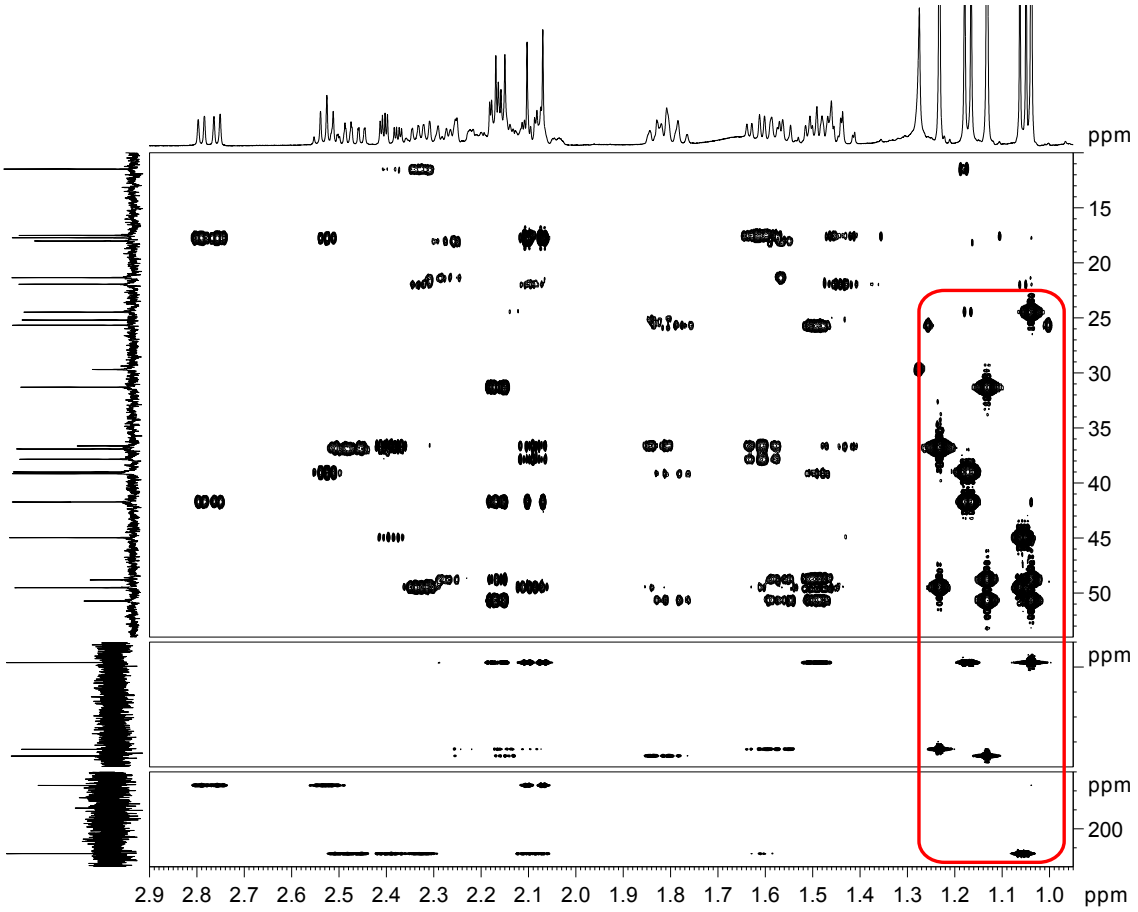


Figure 2.15 – Spectre HMBC du sesterterpène

Dans cet exemple, les corrélations HMBC intenses des groupements méthyles (entourées en rouge sur la figure 2.15) sont forcées à être interprétées comme des couplages ${}^2J_{C-H}$ ou ${}^3J_{C-H}$. Cela signifie que le programme ne tente pas de les éliminer lors de la résolution. Pour évaluer le gain en performance dû à cette contrainte supplémentaire, un compteur a été introduit dans l'algorithme pour mesurer le nombre d'étapes récursives pendant la résolution. On observe une réduction significative du nombre d'étapes récursives par 1/3.

En l'absence de corrélations COSY et de liaisons pré-établies, les contraintes sont réduites. Lorsque les corrélations HMBC sont uniquement utilisées, le nombre d'étapes récursives est réduit de 3/4.

2.3.2.5 Bilan des modifications

Les bénéfices de ces modifications sur l'efficacité de l'algorithme de génération de structure sont multiples. L'utilisateur peut apporter une précision supplémentaire sur les informations de corrélation extraites des spectres 2D. Il est par exemple possible d'éviter à certaines corrélations d'être éliminées lors de la résolution. Des corrélations COSY à longue distance ${}^{n}J_{H-H}$ avec $n > 3$ peuvent être exploitées sans entrainer un échec du programme. De plus des corrélations COSY ambigües peuvent être déclarées. Il était préférable auparavant de laisser de côté les corrélations peu intenses et les corrélations où il existait un doute sur l'un des deux hydrogènes impliqué (déplacements chimiques très proches). Il est intéressant de disposer de ces informations car elles peuvent s'avérer utiles pour l'élucidation.

Les expériences 1,1-ADEQUATE et H2BC permettent d'observer des corrélations interprétables comme des corrélations HMBC issues de couplage ${}^{2}J_{C-H}$. L'utilisateur peut donc déclarer ces corrélations facilement en utilisant la commande `HMBC` associée à une longueur de chemin de couplage fixée à 2.

2.3.3 Hybridation des atomes

2.3.3.1 Objectif

Dans le fichier d'entrée, la définition du statut des atomes contient l'état d'hybridation. Dans un premier temps, LSD a été conçu pour ne traiter que deux alternatives : les atomes hybridés sp^3 ou sp^2. Les molécules analysées par LSD ne pouvaient donc contenir que des simples ou des doubles liaisons. Pour élargir la diversité des molécules analysables, il a été décidé d'intégrer les atomes hybridés sp. Cela ouvre la porte à la production de solutions possédant des triples liaisons ou des allènes.

2.3.3.2 État des lieux avant modification

Avant de commencer la résolution, pendant la phase 0, l'un des premiers tests concerne le nombre d'atomes hybridés sp^2. Ce nombre doit absolument être pair de sorte qu'un atome hybridé sp^2 puisse toujours être lié au moins à un autre atome hybridé sp^2. Si cette condition n'est pas respectée, le programme s'arrête et informe l'utilisateur de son erreur.

Dans un premier temps, la génération des structures est effectuée sans placer les doubles liaisons. Au moment de la formation des liaisons, une simple vérification est effectuée afin d'éviter les structures invalides. Un atome hybridé sp^2 doit avoir au moins un voisin sp^2 lorsqu'il devient complet. Un atome complet est un atome qui ne peut plus

former de liaisons. Le nombre de liaisons que peut former un atome, ou plus exactement son nombre de voisins (hydrogènes exceptés) est connu. Il est calculé à l'aide de la relation suivante : $n_i = Val_i - (3 - Hyb_i) - Mult_i$ pour des atomes neutres, ou si la charge a été incluse dans la valence. Pour un atome i, n_i est son nombre de voisins, Val_i est sa valence, Hyb_i est son état d'hybridation (3 pour sp^3 et 2 pour sp^2) et $Mult_i$ est sa multiplicité.

Lorsqu'une solution est trouvée à la fin de la phase 2, après appariement des atomes incomplets, le programme tente de placer les doubles liaisons par un processus récursif. En cas d'échec la solution n'est pas validée, le programme retourne à la phase précédente.

2.3.3.3 Mise en œuvre des modifications

L'ajout des atomes hybridés *sp* a nécessité quelques aménagements au niveau de l'analyse du fichier d'entrée. L'interdiction de déclarer des atomes hybridés *sp* a été levée. Aucune contrainte supplémentaire concernant la parité du nombre d'atomes *sp* n'a été ajoutée car un atome hybridé *sp* est équivalent à deux atomes hybridés sp^2. Le nombre d'atomes hybridés *sp* peut donc être pair ou impair.

En cours de résolution, la vérification pendant la formation des liaisons a été aménagée. Un atome sp^2 complet doit avoir au moins un voisin sp^2 ou *sp* et un atome *sp* complet doit avoir au moins un voisin *sp* (future triple liaison) ou deux voisins sp^2 (deux doubles liaisons).

Le placement des doubles et triples liaisons, regroupées sous le terme de liaisons multiples, est toujours reporté à la fin de la résolution comme test de validation de structure. Si le placement des liaisons multiples échoue, la solution est rejetée. La modification de l'algorithme de placement des liaisons multiples pour incorporer les atomes hybridés *sp* est détaillée ci-après.

Par définition, un atome hybridé *sp* peut être impliqué dans une triple liaison ou dans deux doubles liaisons (allène). Ceci entraine une indétermination lors du placement des liaisons multiples sur les atomes *sp*. Toutes les hypothèses sont envisagées par une procédure récursive.

Le programme sélectionne un atome hybridé *sp* ou sp^2 ne possédant pas encore l'ensemble des liaisons double ou triple qu'il est capable de former puis recherche parmi ses voisins un atome disponible pour établir la liaison multiple. En cas d'échec, l'ensemble des voisins est successivement testé.

L'ordre de priorité pour la recherche d'atomes est le suivant :

1. un atome *sp* portant déjà une double liaison (pour la découverte des systèmes de type allène) ;
2. un atome sp^2 ne portant pas encore de double liaison et ayant parmi ses voisins un atome sp^2 porteur d'une double liaison (pour la découverte de systèmes de doubles liaisons conjuguées) ;
3. un atome sp^2 non porteur d'une double liaison ;

4. un atome *sp* ne portant pas encore de triple liaison et ayant parmi ses voisins un atome *sp* porteur d'une triple liaison (pour la découverte de systèmes de triples liaisons conjuguées) ;
5. un atome *sp* non porteur d'une triple liaison.

Lorsqu'aucun atome n'est trouvé, toutes les liaisons multiples ont été placées.

2.3.3.4 Exemple d'application

L'exemple est la thiarubrine B (figure 2.16), un polyyne naturel. Le fichier d'entrée utilisé contient toutes les liaisons de la structure pré-établies à l'aide de la commande `BOND`. Ceci permet de tester directement l'algorithme de placement des liaisons multiples. La procédure récursive du placement des liaisons multiples sur la structure de la thiarubrine B est décrite sur la figure 2.17.

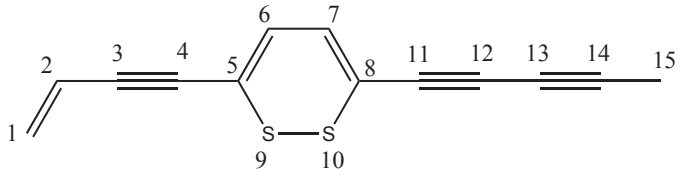


Figure 2.16 – Structure de la thiarubrine B

L'algorithme est prévu pour donner priorité au placement des doubles liaisons. D'après les conditions données dans l'ordre de priorité du choix des atomes, le premier atome sélectionné est nécessairement un atome sp^2 (cas 3). Le carbone 1 est choisi et une double liaison est placée entre C1 et C2 son unique voisin sp^2. Le candidat sp^2 suivant est le carbone 5. Une double liaison est placée entre C5 et C4 qui est hybridé *sp*. On se retrouve alors dans le cas ayant la plus haute priorité pour le choix du candidat suivant (cas 1). Le carbone 4 possède un voisin *sp* libre. Une double liaison est placée entre C4 et C3. Le carbone 3 est ensuite choisi pour la même raison que C4 juste avant. C3 possède un voisin sp^2, le carbone 2, qui aurait pu être l'extrémité du cumulène. Cependant C2 est déjà impliqué dans une double liaison. C3 n'ayant pas d'autre voisin disponible, la situation est bloquée. Le programme fait donc un retour-arrière. La double liaison C4–C3 est ôtée, le programme cherche un autre voisin de C4 libre. C4 n'a pas d'autre voisin sp^2 ou *sp*, un retour-arrière est à nouveau effectué. Après suppression de la double liaison 5–4, on retourne à la situation où C5 est candidat. Le voisin suivant de C5 pour former une double liaison est le carbone 6. Pour le choix du candidat suivant, le programme détecte un cas de doubles liaisons conjuguées (cas 2). En effet le carbone 7 (sp^2) a un voisin hybridé sp^2 portant une double liaison (C6) et un autre voisin sp^2 encore libre (C8). Après placement de la double liaison C7–C8, tous les atomes hybridés sp^2 possèdent des doubles liaisons.

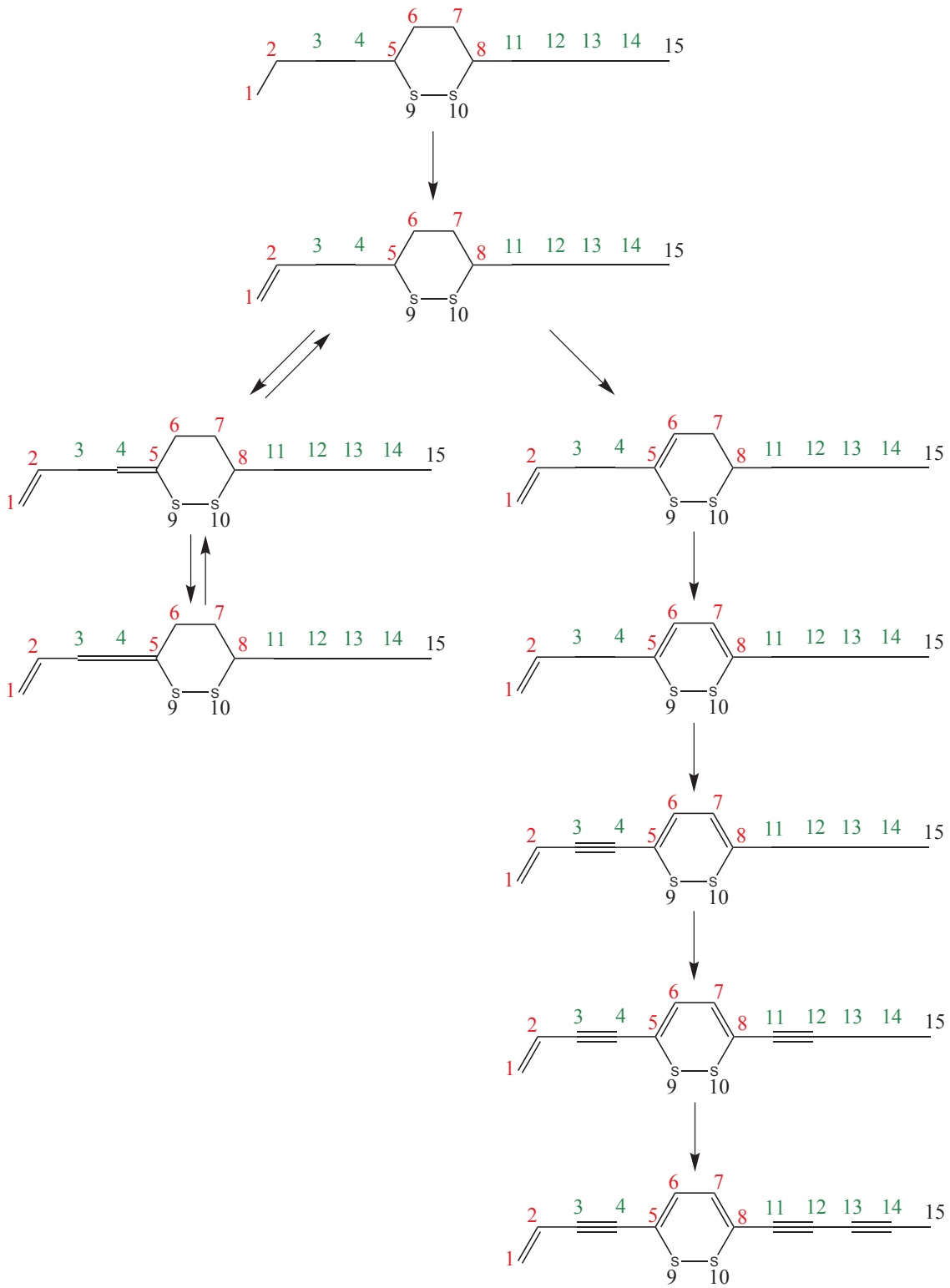


FIGURE 2.17 – Placement des liaisons multiples sur la structure de la thiarubrine B (atomes en rouge : atomes hybridés sp^2 et atomes en vert : atomes hybridés sp)

L'algorithme recherche ensuite des atomes hybridés *sp* pour placer des triples liaisons. Il s'agit du cas ayant la plus faible priorité pour la recherche de candidats (cas 5). Le carbone 3 est choisi et une triple liaison est placée entre C3 et C4 son unique voisin *sp*. Puis une triple liaison est placée entre les carbones 11 et 12. Enfin, un système de triples liaisons conjuguées est détecté (cas 4). Le carbone *sp* 13 est sélectionné car il possède un voisin portant une triple liaison (C12) et un second voisin *sp* encore disponible (C14). Le placement de la triple liaison entre les carbones 13 et 14 termine le processus. Toutes les liaisons multiples ont été placées, la structure est validée.

2.3.3.5 Bilan des modifications

Suite aux modifications apportées, LSD peut produire des solutions contenant des atomes hybridés *sp*. Le programme valide la structure dès qu'il trouve une manière de placer toutes les liaisons multiples. Cependant, un point est encore à améliorer au niveau de l'algorithme de placement des liaisons multiples. En prenant comme exemple la recherche de tous les isomères de formule brute C_6H_6, PyLSD (voir paragraphe 2.3.6) permet d'obtenir 212 solutions. En réalité, à cette formule brute correspondent 217 isomères (atomes chargés exclus). L'ensemble de ces 217 isomères peut être obtenu à l'aide de MOLGEN [158–161], un programme de génération automatique d'isomères à partir d'une formule brute. Après comparaison des résultats, les 5 structures « oubliées » par LSD ont été identifiées. La figure 2.18 montre que ces 5 structures correspondent à des paires de structures avec des atomes ayant la même connectivité mais avec un placement différent des liaisons multiples. LSD ne produit que l'une des deux structures pour chaque paire.

Afin de corriger ce manque, il faudrait étendre le processus récursif pour tenter de placer les liaisons multiples sans s'arrêter à la première solution.

2.3.4 Atomes chargés et atomes à valence multiple

2.3.4.1 Objectif

Dans la conception initiale de LSD, les règles de déclaration du statut des atomes ne permettaient pas de travailler avec d'éventuels atomes chargés. Il était nécessaire de créer un nouveau type d'atome avec ses propriétés propres. Par exemple, l'atome d'azote d'un ammonium quaternaire devait être déclaré en définissant un nouveau type d'atome de valence 4 et ayant la masse molaire de l'azote.

Les atomes possédaient des valences fixes. Par exemple l'atome de soufre avait une valence de 2. En réalité le soufre peut avoir une valence de 2, 4 ou 6. Les modifications qui suivent ont pour but d'étendre la gamme de valence des atomes et de pouvoir déclarer directement des atomes chargés.

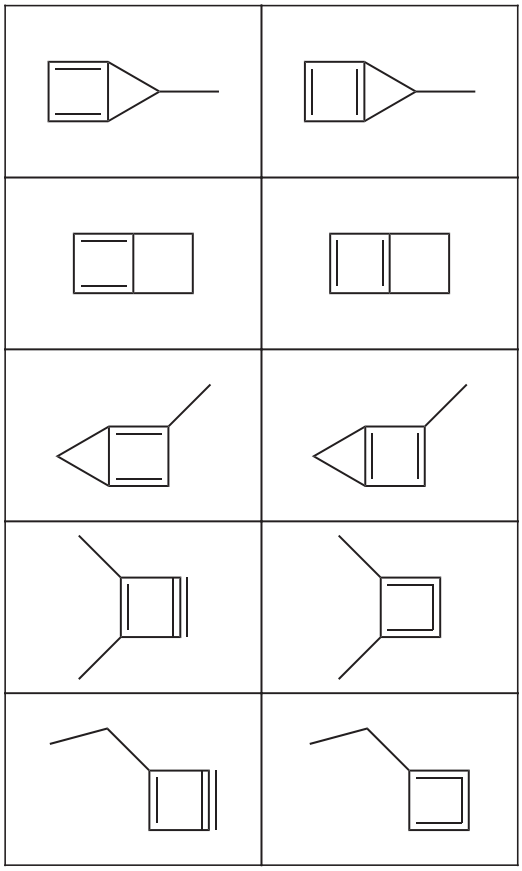


FIGURE 2.18 – Quelques isomères de formule brute C_6H_6

2.3.4.2 État des lieux avant modification

Jusqu'ici, chaque type d'atome ne possédait qu'une seule valence définie. La liste des éléments chimiques acceptés par LSD est la suivante : C, N, O, S, F, Cl, Br, I, P, Si et B. LSD ne pouvait prendre en charge que des molécules contenant des atomes d'azote trivalents, de soufre divalents et de phosphore trivalents. La connaissance de la valence des atomes est importante car comme vu au paragraphe 2.3.3.2, elle permet de calculer le nombre de voisins (hydrogènes exceptés) en fonction de l'état d'hybridation et de la multiplicité des atomes.

2.3.4.3 Mise en œuvre des modifications

Un paramètre supplémentaire a été ajouté à la déclaration du statut des atomes pour spécifier la charge des atomes. Dans le fichier d'entrée, la commande `MULT` a maintenant un cinquième paramètre. Ce paramètre de charge est optionnel afin de garder la compatibilité avec les anciens jeux de données. La liste des charges acceptées est restreinte de -1 à +2.

L'ajout des charges a été réalisé en même temps que les atomes à valence multiple car les valeurs de valence et de charge sont liées. Les nouveaux atomes sont l'azote pentavalent, le soufre tétravalent, le soufre hexavalent et le phosphore pentavalent. La liste étendue des éléments chimiques acceptés par LSD est la suivante : C, N, N5, O, S, S4, S6, F, Cl, Br, I, P, P5, Si et B.

La valence est maintenant déduite à partir de l'élément chimique et de la charge. Les valeurs sont regroupées dans un tableau à double entrée (figure 2.19).

Il faut noter que la définition de l'hybridation dans LSD n'est pas la définition de l'hybridation au sens de la théorie de l'hybridation des orbitales atomiques. Selon la nomenclature de LSD, un atome qui porte un double liaison est un atome hybridé sp^2 et un atome qui porte deux doubles liaisons est un atome hybridé sp. Par exemple un carbocation possède une structure plane avec des angles de valence de 120°, il est hybridé sp^2. Pour LSD, cet atome ne porte pas de double liaison, il est donc hybridé sp^3.

Suite aux nouveautés apportées à LSD, les programmes de sortie et d'affichage des structures (OUTLSD, M_EDIT et GENPOS) ont été adaptés en conséquence pour permettre le traitement des atomes chargés. La génération de chaînes SMILES a été modifiée en accord avec les règles de représentation des atomes chargés. La visualisation de structures chargées a été rendue possible avec quelques aménagements dans le code de M_EDIT et de GENPOS.

2.3.4.4 Exemple d'application

La possibilité de définir des atomes chargés a permis la détermination de la structure de la 2-O,N-dimethylliriodendronine (figure 2.20), un alcaloïde aporphinique de formule brute $C_{18}H_{13}NO_3$.

Élément \ Charge	-1	0	+1	+2
C	3	4	3	-
N	2	3	4	3
N5	-	5	-	-
O	1	2	3	-
S	1	2	3	4
S4	-	4	-	-
S6	-	6	-	-
P	2	3	4	-
P5	-	5	-	-
F	-	1	-	-
Cl	-	1	-	-
Br	-	1	-	-
I	-	1	-	-
Si	3	4	3	-
B	4	3	-	-

FIGURE 2.19 – Tableau récapitulatif de la valence des atomes

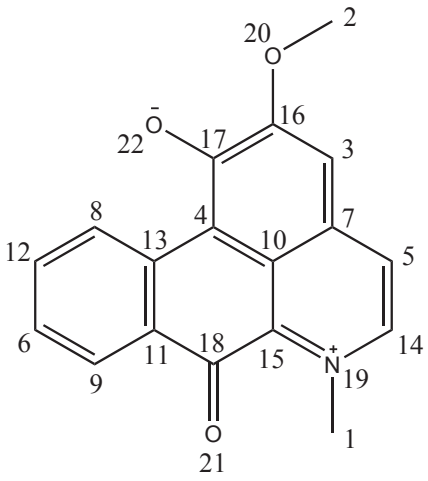


FIGURE 2.20 – Structure de la 2-O,N-dimethylliriodendronine

Le fichier d'entrée de LSD est présenté en annexe B. L'analyse des spectres ^{13}C et DEPT nous indique la présence de 2 groupements CH_3, 7 groupements CH aromatiques et 9 carbones quaternaires. Les carbones sont numérotés par ordre de déplacements chimiques croissants. L'analyse des déplacements chimiques ^{13}C nous indique la présence de 3 oxygènes liés aux carbones C16 (δ 162,2), C17 (δ 171,8) et C18 (δ 176,3). Les déplacements chimiques du méthyle 2 (δ (^{13}C) 56,5 et δ (^{1}H) 4,00) sont caractéristiques d'un groupement O–CH_3. De même les déplacements chimiques du méthyle 1 (δ (^{13}C) 50,4 et δ (^{1}H) 4,75) nous indiquent la présence d'un groupement N^+–CH_3. Les liaisons 1–19 et 2–20 ont été déclarées dans le fichier d'entrée de LSD. Ce type de molécule est connu pour se trouver sous forme zwitterionique. Pour équilibrer la charge de la molécule, un oxygène est donc porteur d'une charge négative. Des corrélations COSY sont observées entre les protons 5 et 14, 8 et 12, 6 et 12 et enfin 6 et 9. La liaison 5–14 et le système 8–12–6–9 sont donc facilement déduits par LSD. Le proton 3 apparait sous la forme d'un singulet, il est donc isolé des autres protons. Le programme trouve 2 solutions (figure 2.21) après élimination de 3 corrélations HMBC à très longue distance ^{4}J.

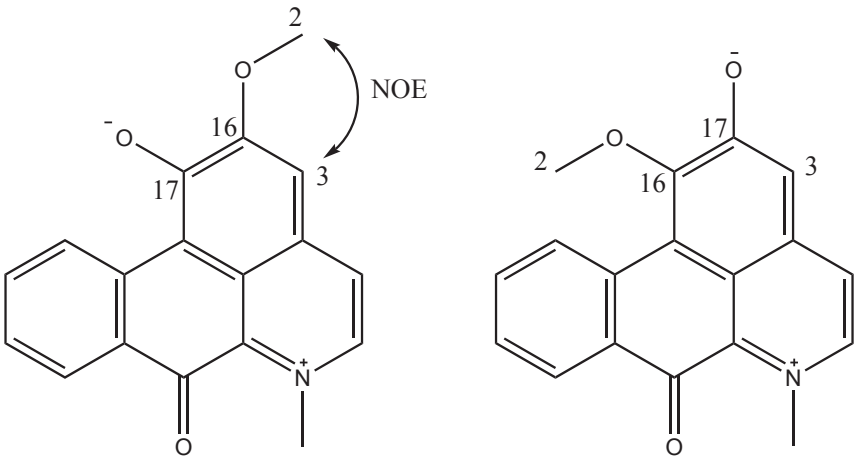


Figure 2.21 – Solutions obtenues par LSD

La différence entre ces 2 solutions est l'inversion des carbones 16 et 17 accompagnés de leurs substituants respectifs. La première solution est préférée à la seconde car on observe une corrélation NOE entre les protons 2 et 3. La comparaison des déplacements chimiques ^{1}H avec des données de la littérature [162] nous confirme la structure proposée par LSD.

La procédure de codage de la chaîne SMILES par OUTLSD est présentée sur la figure 2.22. Les atomes chargés sont placés entre crochets suivis du signe de la charge. Ce n'est pas le cas dans cet exemple mais si un atome chargé est lié à un hydrogène, celui-ci doit être explicitement inscrit à l'intérieur des crochets (exemple : ion hydroxyde [OH-]).

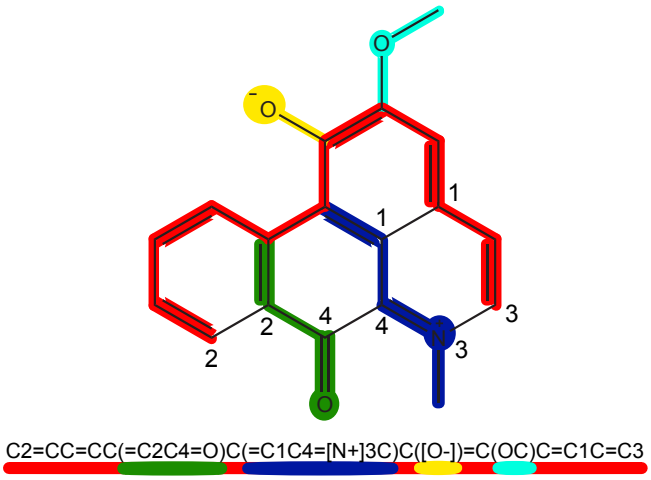


Figure 2.22 – Codage de la chaîne SMILES par OUTLSD

2.3.4.5 Bilan des modifications

Ces modifications permettent de produire des solutions contenant des atomes chargés ainsi que de nouvelles fonctions chimiques (sulfoxide, sulfine, sulfone, nitro, isocyano, phosphate...). Les programmes LSD, OUTLSD, M_EDIT et GENPOS peuvent traiter des molécules avec une plus large diversité chimique.

2.3.5 Présentation des structures

2.3.5.1 Objectif

Le programme OUTLSD contient un algorithme de calcul de coordonnées 2D à partir de tables de connectivité entre atomes fournies par le générateur de structure. Les coordonnées produites fournissent parfois des dessins de structures difficiles à interpréter. Le but des tests suivants est d'améliorer la qualité visuelle des dessins des molécules.

2.3.5.2 Calcul des coordonnées par OUTLSD

L'algorithme de calcul des coordonnées est inspiré par l'algorithme proposé par Carhart en 1976 [163]. Les atomes sont initialement distribués aléatoirement dans l'espace. La position des atomes est ensuite améliorée de manière itérative par un processus de modélisation moléculaire simplifié. À chaque cycle, des vecteurs de déplacement $d(i)$ sont calculés pour chaque atome à l'aide de la relation suivante :

$$d(i) = A \sum_j \frac{(X_j - X_i)}{\left(r_{ij}^2 + B\right)^2} + C \sum_k (X_k - X_i) \left[1 - \frac{1}{(r_{ik}^2 + B)}\right]$$

où $X_i = (X_i, Y_i, Z_i)$ sont les coordonnées de l'atome i, r_{ij} est la distance entre les atomes i et j, $A = -0{,}3$, $B = 10^{-6}$ et $C = 0{,}2$. Les constantes A, B et C ont été déterminées

empiriquement. La constante B est présente aux dénominateurs des deux termes pour éviter qu'ils ne s'approchent trop de zéro et ainsi prévenir des problèmes de division par zéro. Dans le premier terme, la somme s'applique à tous les atomes j non liés à i. Dans le second terme, la somme s'applique à tous atomes k liés à i. Globalement le déplacement des atomes a pour but de séparer les atomes non-liés et à rapprocher les atomes liés pour atteindre une distance d'équilibre de 1. Les déplacements calculés sont finalement ajoutés aux positions des atomes et un nouveau cycle débute. Le processus s'arrête lorsque la moyenne quadratique des déplacements devient inférieure à 0,001 ou lorsque 80 cycles ont été effectués. Les coordonnées 3D ainsi calculées servent de base pour l'obtention des coordonnées 2D. Celles-ci sont obtenues par projection des coordonnées 3D dans un plan choisi de manière à ce qu'il soit parallèle au côté de la molécule offrant l'aspect le plus plat. De plus, lors de la projection, la structure est réorientée pour occuper le maximum d'espace suivant l'axe horizontal du dessin. Les coordonnées 2D obtenues subissent le même processus itératif que précédemment. Les distances interatomiques sont ainsi réajustées car la projection a tendance à les raccourcir.

2.3.5.3 Solutions alternatives

En cherchant à améliorer le côté esthétique de la présentation des molécules, il s'agit surtout de faciliter la lecture et la compréhension des structures. Les défauts à éviter sont principalement les superpositions d'atomes ou les superpositions des atomes avec les liaisons. Il faut également éviter des liaisons qui se chevauchent trop.

Les méthodes et algorithmes utilisés pour la génération de dessins de molécules ont été rapportés par Helson en 1999 [164]. Les algorithmes les plus récents divisent les structures en morceaux et utilisent des templates, notamment pour les cycles, en association avec des règles concernant les contraintes de distances et d'angles de liaisons.

Afin de corriger les défauts de OUTLSD, nous nous sommes tournés vers des outils existants. Parmi les outils à disposition, il existe les librairies libres OpenBabel [165], CDK (Chemistry Development Kit) [166, 167] ou RDKit [168]. Des tests ont été effectués à l'aide de ces librairies afin de comparer les performances de génération de coordonnées 2D. Nous nous sommes concentrés sur la librairie OEChem [169], fournie par l'éditeur OpenEye Scientific Software, dans le but de l'utiliser comme méthode alternative à l'algorithme de OUTLSD.

2.3.5.4 Résultats

Les tests menés à l'aide des librairies citées précédemment visent à trouver le meilleur compromis dans la représentation des molécules. La figure 2.23 permet de comparer les résultats obtenus avec les différentes méthodes. Les structures sont celles de la santonine (**1**) et de deux de ses isomères (**2 et 3**).

Pour des structures comme la santonine, ne possédant que des cycles accolés un à

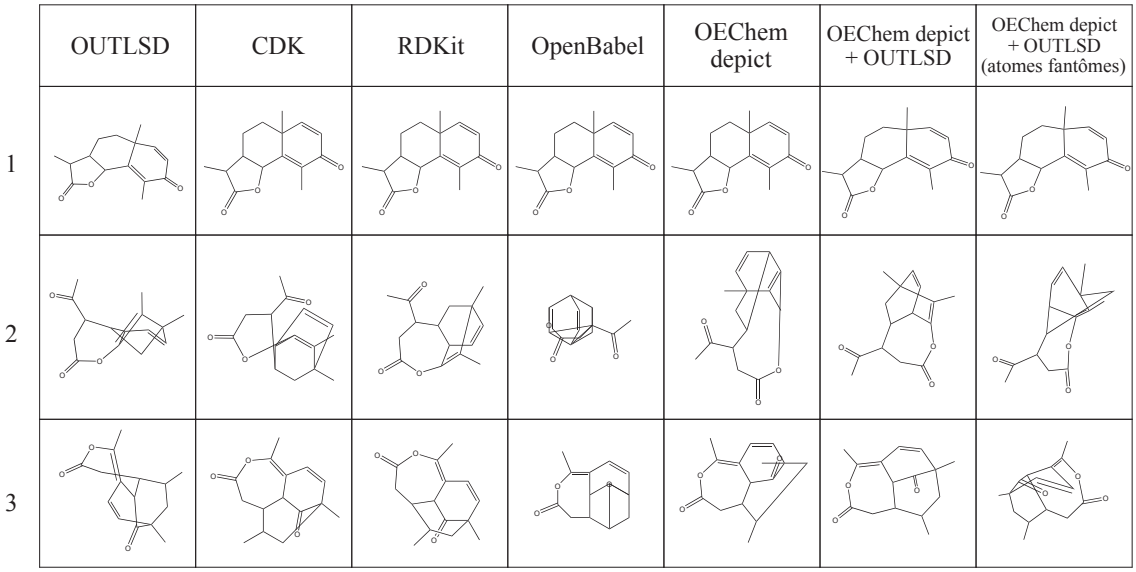


Figure 2.23 – Tableau récapitulatif des tests de génération de coordonnées 2D : dessins obtenus pour 3 structures avec les 7 méthodes décrites dans le texte

un, toutes les méthodes donnent des résultats très satisfaisants. Les composés 2 et 3 ont des systèmes cycliques plus complexes. Dans ces deux cas, OUTLSD a produit des coordonnées qui ne facilitent pas l'interprétation de la structure. Les résultats produits par CDK et RDKit sont plutôt satisfaisants malgré quelques superpositions de liaisons. OpenBabel propose des structures avec des atomes superposés. Il s'agit d'un très mauvais défaut car les structures sont rendues illisibles.

OEChem fournit des structures dont la qualité est perfectible. Les atomes et les liaisons sont parfois superposés. Les liaisons peuvent être très étendues ce qui déforme la structure. On observe qu'il suffirait de déplacer quelques atomes pour parvenir à un bon résultat. Les coordonnées 2D obtenues avec OEChem constituent une bonne base pour une amélioration par un algorithme tel que celui de OUTLSD. La sortie de OEChem a donc été couplée à l'algorithme de OUTLSD. Le résultat de cette manipulation des coordonnées 2D de OEChem est présenté dans la colonne 6 de la figure 2.23. La structure de la santonine est très peu modifiée et reste lisible. Les structures des composés 2 et 3 sont nettement corrigées. La qualité des dessins produits par cette méthode est très bonne. Il n'y a pas de superpositions de liaisons inutiles et peu agréables à la lecture.

La dernière colonne de la figure 2.23 montre les résultats obtenus en utilisant des atomes fictifs au centre des liaisons pendant les cycles de calcul de OUTLSD. Cette méthode avait pour but de supprimer d'éventuelles superpositions entre atomes et liaisons mais n'a pas donné de résultats intéressants.

2.3.5.5 Bilan des tests

L'utilisation de la librairie OEChem couplée à l'algorithme existant de OUTLSD a permis une nette amélioration de la production automatique de dessins 2D de molécules organiques. Une version de OUTLSD a été entièrement réécrite en tirant parti au maximum de toutes les possibilités de la librairie OEChem. Les différents modules utilisés sont le calcul de coordonnées 2D, le calcul de coordonnées 3D en utilisant un champ de force MMFF94 et la génération de chaînes SMILES. Le programme OUTLSD_OE peut être utilisé comme alternative à la version classique de OUTLSD. L'inconvénient de la librairie OEChem est d'être un produit « propriétaire ». Son accès est donc limité malgré la possibilité d'obtenir des licences gratuites pour les universitaires. Il pourrait être intéressant d'écrire des versions de OUTLSD faisant l'interface avec les différentes librairies libres. L'utilisateur pourrait ainsi disposer de plusieurs outils pour visualiser les résultats du générateur de structure. Malgré le manque d'interfaces pour le moment, ces librairies peuvent être utilisées de manière indépendante. Elles permettent en effet la conversion entre différents formats de fichier. Il est possible de convertir une représentation linéaire (par exemple une chaîne SMILES provenant de OUTLSD) en un dessin de la molécule.

2.3.6 Conclusion

Les gains et les avantages des changements présents dans les dernières versions de LSD sont multiples.

Tout d'abord, le programme peut procéder à une utilisation plus fine des corrélations avec l'utilisation de la définition des longueurs de chemin de couplage. Cela permet un gain de temps d'analyse et une réduction du nombre de solutions si on autorise l'élimination de corrélations.

Ensuite la présentation des résultats a été nettement améliorée par l'emploi d'outils existants combinés avec le programme OUTLSD gérant les solutions produites par le générateur de structures de LSD.

Enfin, les modifications ont également permis d'augmenter la diversité des molécules analysables avec l'introduction des atomes hybridés *sp*, des atomes chargés et des atomes à valence multiple.

Ces aménagements récents de LSD ont permis l'écriture d'une couche supplémentaire à LSD appelée PyLSD (car écrite en Python) qui permet la gestion des atomes à statut indéfini. Il arrive parfois que l'hybridation des atomes ne puisse pas être déterminée avec certitude. Par exemple un atome de carbone ayant un déplacement chimique aux alentours de 100 ppm peut être hybridé sp^2 ou sp^3. Les hétéroatomes peuvent être concernés par l'indétermination sur l'hybridation et la multiplicité, notamment si on est en présence d'atomes d'hydrogène mobiles. Jusqu'ici il était nécessaire de créer un fichier de données différent pour chaque hypothèse. PyLSD permet de définir des atomes dont le statut n'est pas clairement défini. De même lorsque la formule brute n'est pas connue avec exactitude,

toutes les hypothèses peuvent être envisagées. Cette version possède également un module de prédiction des déplacements chimiques ^{13}C permettant de classer les solutions trouvées par ordre de vraisemblance par rapport aux déplacements chimiques expérimentaux.

2.4 Utilisation associée des systèmes LSD et SISTEMAT

2.4.1 Introduction

Cette partie s'inscrit dans le cadre d'une collaboration avec le Professeur Vicente de Paulo Emerenciano de l'université de São Paulo.

Le programme LSD repose essentiellement sur l'utilisation des données de corrélation issues de spectres RMN 2D. L'information des déplacements chimiques n'est pas prise en compte pour l'élucidation des structures chimiques. Cette méthode constitue sa force car les valeurs de déplacements chimiques peuvent parfois fausser la résolution. Au contraire, SISTEMAT s'appuie sur une base de données de composés associés, entre autres, à leurs spectres de RMN ^{13}C. La combinaison des approches des deux systèmes permet de bénéficier des avantages de chacun pour résoudre des problèmes d'élucidation structurale de petites molécules organiques.

SISTEMAT peut, en utilisant les connaissances de la base de données, apporter des contraintes de sous-structures qui peuvent être incluses dans les problèmes d'élucidation traités par LSD.

2.4.2 Description de SISTEMAT

La base de données de SISTEMAT [170–174] a été développée au Brésil dans le but d'assister l'élucidation structurale de produits d'origine naturelle. Chaque structure est associée à plusieurs propriétés : des données spectroscopiques (spectres RMN ^{13}C et ^{1}H, spectre de masse), la classe de produit naturel, le squelette carboné principal et des sources bibliographiques. La propriété qui fait son originalité est l'information sur l'origine botanique (famille, espèce, genre) de la plante d'où provient la substance. La base de données contient environ 20000 molécules faisant partie de plusieurs classes de produits naturels (alcaloïdes, iridoïdes, monoterpènes, sesquiterpènes, diterpènes, etc...).

2.4.2.1 Outils de recherche

SISTEMAT est un système fonctionnant avec une base de données et des outils pour effectuer des recherches dans cette base. La figure 2.24 présente les différents outils utilisés par SISTEMAT.

Ces outils permettent d'extraire des fragments de molécules et des squelettes carbonés susceptibles d'être présents dans la structure analysée :

- PESQUISA fournit une série de noms de squelettes moléculaires et leurs probabilités de présence associées, à partir des informations fournies par l'utilisateur ;
- MACRONO recherche des substituants caractéristiques (appelés macronos) comme des groupements acétates, angélates, tiglates qui peuvent être facilement déduits à partir des données spectrales ;
- C13MACH et H1MACH sont des outils de comparaison des déplacements chimiques ^{13}C et ^{1}H ;
- SISOCBOT exploite les données botaniques (famille, genre, espèce) de la plante d'où le composé analysé est extrait ;
- REGRAS détermine la composition de la molécule défonctionnalisée (nombre de carbones et multiplicités) et la compare avec celle des squelettes présents dans la base de données ;
- SISCONST trouve des fragments (par comparaison avec les molécules de la base de données) dont les déplacements chimiques et les multiplicités correspondent à certains de ceux de la molécule étudiée.

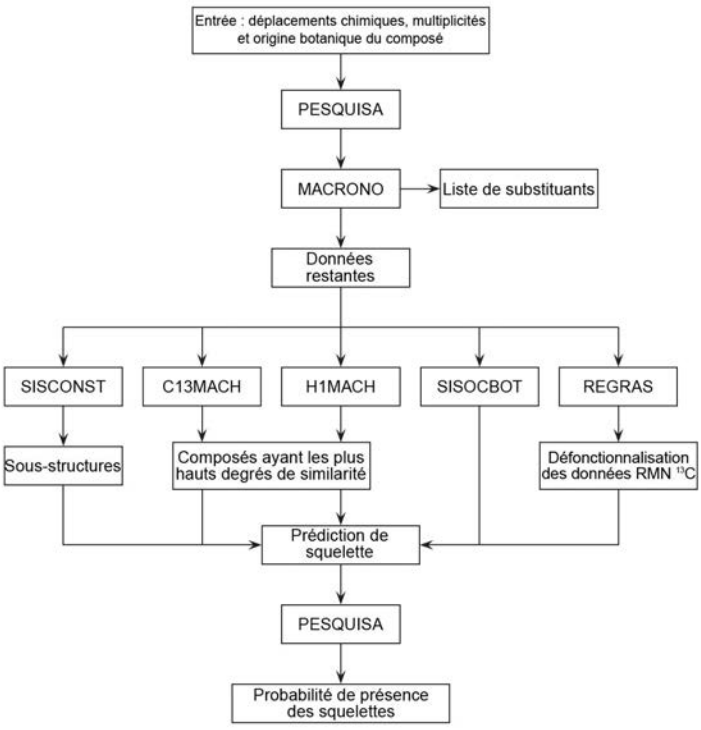


FIGURE 2.24 – Outils de recherche dans la base de données de SISTEMAT

2.4.2.2 Codage des molécules

SISTEMAT a été développé à une époque où les ordinateurs ne disposaient que de très peu de mémoire. Les auteurs ont donc accordé une grande importance à la mise au point d'une méthode de codage compacte permettant la création d'une grande base de données. Le système de codage des composés a été publié en 1989 [22]. Les molécules sont codées sous forme de chaînes de caractères alphanumériques qui sont appelées vecteurs.

Le vecteur débute par la description des chaînes principales (numéros initiaux et finaux). Une chaîne est une succession d'atomes liés et de numérotation croissante. Des balises servent ensuite à indiquer de multiples informations :

- –1 : numéros des atomes des liaisons non présentes dans les chaînes ;
- –2 : numéros atomiques des hétéroatomes et de leurs positions ;
- –3 : numéros des atomes impliqués dans des doubles liaisons ;
- –4 : numéros des atomes aromatiques ;
- –5 : numéros des atomes impliqués dans des triples liaisons ;
- –6–1 : liaisons beta ;
- –6–2 : liaisons alpha ;
- –6–3 : atomes chargés positivement ;
- –6–4 : atomes chargés négativement ;
- –7 : liaisons cis entre cycles ;
- –8 : positions et description des macronos ;
- –9 : fin du vecteur.

Le codage du vecteur pour le squelette labdane est utilisé comme exemple sur la figure 2.25.

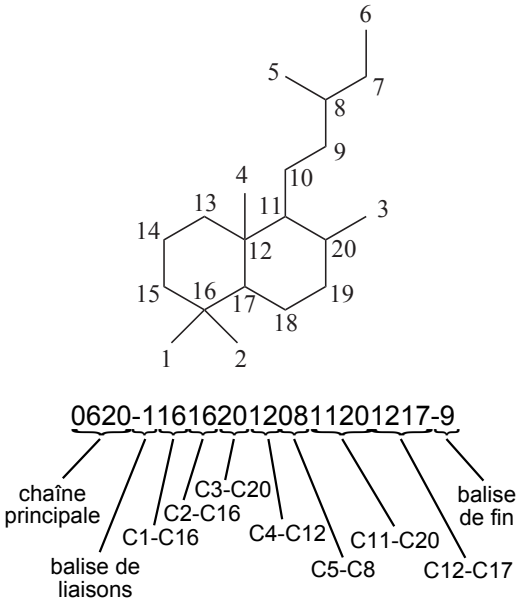


Figure 2.25 – Codage du vecteur SISTEMAT pour le labdane

Les groupements méthyles (C1 à C5) sont numérotés en priorité pour que les atomes restants forment une longue chaîne (C6 à C20). Les liaisons supplémentaires C11–C20 et C12–C17 permettent de fermer les deux cycles.

2.4.2.3 Complémentarité des systèmes LSD et SISTEMAT

Le potentiel de l'usage complémentaire des deux systèmes a été montré dans deux articles [175, 176]. SISTEMAT est capable de prédire le squelette carboné principal et/ou des éléments de sous-structures (fonctions chimiques, groupements caractéristiques) à partir des données spectroscopiques et botaniques. Lorsque les données utilisées par LSD (corrélations provenant des spectres de RMN 2D) produisent un nombre trop élevé de solutions, il est possible de filtrer ces solutions en ne retenant que celles qui contiennent ou ne contiennent pas un certain nombre d'éléments de sous-structure. Des exemples extraits de la littérature ont été utilisés afin de montrer l'efficacité de l'utilisation conjointe des deux systèmes. L'élucidation structurale de produits naturels de complexité variable a été réalisée à partir des données spectrales publiées. Un des exemples montre la résolution de la structure de la xylocarpine A (figure 2.26) isolée de *Xylocarpus granatum* (Meliaceae) [177].

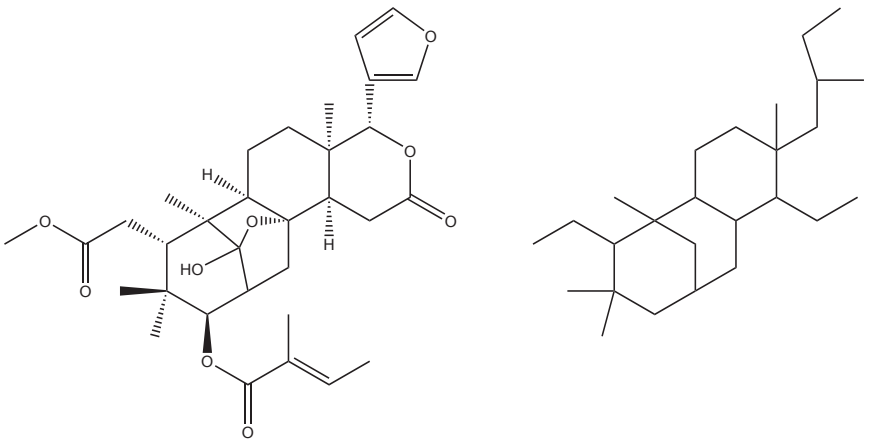


Figure 2.26 – Structure de la xylocarpine A (à gauche) et squelette mexicanolide (à droite)

Les groupements méthoxy et angelate ont été détectés par MACRONO. L'analyse des données restantes a été effectuée par PESQUISA, par l'intermédiaire des modules SISCONST, C13MACH, SISOCBOT et REGRAS. PESQUISA a proposé une liste de 16 squelettes ainsi que leur probabilité de présence. Le squelette mexicanolide (figure 2.26) a été classé premier de cette liste.

Le fichier d'entrée de LSD contenant des propriétés de voisinage pour certains atomes (liaisons carbone–oxygène et liaisons carbone quaternaire–groupement méthyle) donne

6 solutions. Lorsque le squelette mexicanolide prédit par SISTEMAT est utilisé comme contrainte de sous-structure, LSD produit une seule solution.

2.4.3 Derniers progrès dans le lien entre les systèmes LSD et SISTEMAT

2.4.3.1 Objectifs

Les squelettes utilisés comme contraintes de sous-structures dans LSD doivent être codés manuellement par l'utilisateur. Nous avons jugé utile de pouvoir proposer à l'utilisateur un certain nombre de squelettes « prêts à l'emploi » afin de simplifier l'utilisation du programme. L'objectif est d'extraire les squelettes carbonés principaux de la base de données. Il s'agit de rendre les squelettes disponibles sous la forme de fichiers MDL MOL visualisables à l'aide de logiciels courants. Il faut également prévoir un programme de conversion automatique des fichiers MDL MOL en fichiers de sous-structure pour LSD.

2.4.3.2 Choix des outils

La base de données de SISTEMAT est divisée en plusieurs sous-groupes pour chaque classe de produit naturel. Les données sont disponibles sous la forme de fichiers textes contenant entre autres les informations suivantes : nom de la molécule, vecteur, classe, squelette, famille, genre, espèce, référence bibliographique, numérotation biogénétique, données spectroscopiques (essentiellement déplacements chimiques ^{13}C). L'interprétation des vecteurs permet de récupérer les informations structurales. Un programme permet de convertir les vecteurs en listes d'atomes et listes de liaisons.

Le processus d'extraction des squelettes de la base de données nécessite plusieurs étapes :

- lecture du fichier texte et extraction des données utiles ;
- reconstruction des structures ;
- extraction des squelettes ;
- écriture des fichiers MDL MOL.

La première étape est réalisée par un script qui lit les fichiers textes et récupère uniquement les informations nécessaires à la suite du processus. Pour réaliser la suite du programme, la librairie OEChem a été utilisée. La seconde étape consiste en la construction des molécules à partir des propriétés des atomes et des listes de liaisons. La troisième étape est la plus importante, il s'agit d'extraire les squelettes carbonés principaux. La stratégie choisie est l'utilisation de l'algorithme de recherche de plus grande sous-structure commune MCSS (Maximum Common Substructure Search) [178] de la librairie OEChem. Les molécules correspondant à un nom de squelette donné sont isolées et l'algorithme MCSS leur est appliqué. Le squelette ainsi obtenu est finalement enregistré au format MDL MOL. Les coordonnées 2D sont générées par la librairie OEChem.

Un exemple d'application d'un algorithme MCSS est présenté sur la figure 2.27. Quelques molécules associées au squelette labdane permettent d'obtenir un squelette commun qui dans le meilleur des cas est le squelette labdane. Il peut arriver que le squelette obtenu se rapproche fortement du squelette attendu avec quelques atomes en plus. Dans ce cas le squelette doit être retouché manuellement pour supprimer les atomes superflus.

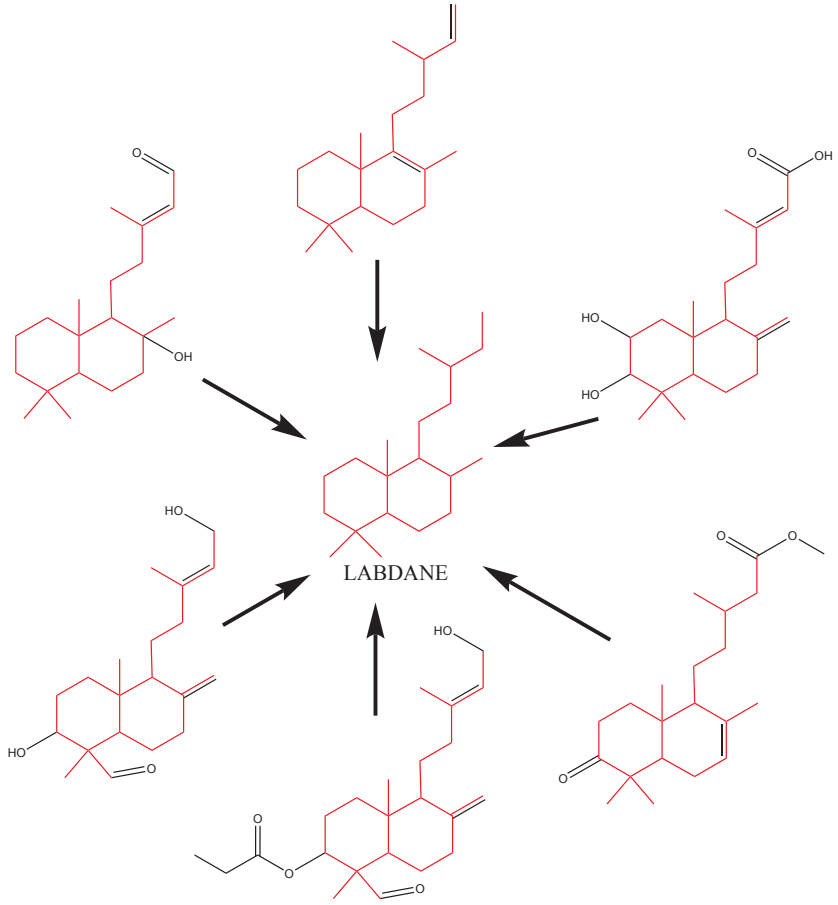


Figure 2.27 – Recherche de plus grande sous-structure commune

2.4.3.3 Résultats

Extraction des squelettes Un total de 448 squelettes ont été extraits de SISTEMAT pour être mis à disposition des utilisateurs de LSD. L'ensemble des squelettes est regroupé en annexe C. On dispose ainsi d'une collection de squelettes au format MDL MOL qui peuvent être traduits au format de sous-structure pour LSD. La collection compte 77 squelettes de monoterpènes, 226 squelettes de sesquiterpènes et 145 squelettes de diterpènes. Un programme de conversion est distribué avec LSD pour coder automatiquement les sous-structures.

Programme de conversion MOL2AB Le programme MOL2AB a été écrit dans le but de convertir automatiquement une structure au format MDL MOL en fichier de

sous-structure utilisable dans LSD.

Comme vu dans le paragraphe 2.2.2, il est nécessaire de laisser une certaine liberté sur l'hybridation et la multiplicité des atomes. En effet, par définition, une sous-structure est substituée et la nature des liaisons chimiques n'est pas fixée. Un certain nombre de règles est donc suivi durant la transcription :

- tout atome d'hydrogène de la structure doit pouvoir être substitué ;
- toute liaison simple peut être transformée en double liaison ;
- toute double liaison peut être transformée en triple liaison ;
- une triple liaison reste une triple liaison ;
- la nature chimique des atomes est préservée.

Il est possible de laisser une liberté sur l'élément chimique. Dans ce cas, le symbole atomique est A (pour Any), ce qui est la convention adoptée dans le format MDL MOL.

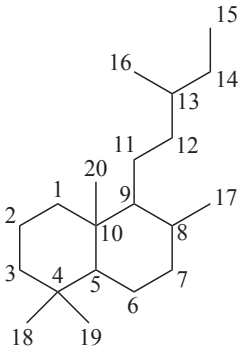
En résumé, les contraintes sur la nature des liaisons ne peuvent évoluer que dans le sens de « l'oxydation » des liaisons. Un atome hybridé *sp* reste hybridé *sp*, cela veut dire qu'une triple liaison ne peut pas être « réduite ». De même, un système type allène ne peut être modifié. Au contraire, une liaison peut être « oxydée » par élimination d'un hydrogène porté par un atome hybridé sp^2 ou sp^3. Une liaison simple peut être transformée en double liaison qui elle-même peut être transformée en triple liaison à condition que les atomes formant la liaison soient porteurs d'hydrogènes. Par conséquent, un atome hybridé sp^2 ne portant pas d'atome d'hydrogène reste sp^2. De la même manière, un atome hybridé sp^3 ne peut pas devenir sp^2 s'il ne porte pas au moins un atome d'hydrogène. Par extension, un atome hybridé sp^3 ne peut pas devenir *sp* s'il ne porte pas au moins deux atomes d'hydrogène.

Un exemple de codage du squelette labdane par MOL2AB est présenté sur la figure 2.28.

Les atomes C4 et C10 sont deux carbones sp^3 quaternaires. Dans les structures où le squelette labdane est présent, ils ne peuvent devenir sp^2 et les liaisons dans lesquelles ils sont impliqués demeurent des liaisons simples. Ils ne portent pas d'atome d'hydrogène qui pourrait être substitué. Leur statut est donc fixe, il n'y a pas de liberté sur l'hybridation et la multiplicité.

Les atomes C5, C8, C9 et C13 sont des carbones CH sp^3. Ils peuvent devenir sp^2 par élimination de l'hydrogène. L'hybridation *sp* leur est interdite car aucun autre atome d'hydrogène ne peut être éliminé.

Code de LSD Les travaux précédents ont motivé des modifications au code source de LSD pour faciliter la façon dont les contraintes de sous-structure sont imposées par l'utilisateur dans le fichier d'entrée. Précédemment, les noms de squelette devaient être associés à des numéros de squelette par l'intermédiaire de la commande `SKEL`. Ces numéros étaient ensuite utilisés dans la commande `FEXP` qui permet d'imposer ou d'interdire la présence d'une ou plusieurs sous-structure(s) à l'aide d'une expression logique. Dans la commande




```
SSTR S1 C (1 2 3) (0 1 2)     LINK S1 S2
SSTR S2 C (1 2 3) (0 1 2)     LINK S1 S10
SSTR S3 C (1 2 3) (0 1 2)     LINK S2 S3
SSTR S4 C 3 0                 LINK S3 S4
SSTR S5 C (2 3) (0 1)         LINK S4 S5
SSTR S6 C (1 2 3) (0 1 2)     LINK S4 S18
SSTR S7 C (1 2 3) (0 1 2)     LINK S4 S19
SSTR S8 C (2 3) (0 1)         LINK S5 S6
SSTR S9 C (2 3) (0 1)         LINK S5 S10
SSTR S10 C 3 0                LINK S6 S7
SSTR S11 C (1 2 3) (0 1 2)    LINK S7 S8
SSTR S12 C (1 2 3) (0 1 2)    LINK S8 S9
SSTR S13 C (2 3) (0 1)        LINK S8 S17
SSTR S14 C (1 2 3) (0 1 2)    LINK S9 S10
SSTR S15 C (1 2 3) (0 1 2 3)  LINK S9 S11
SSTR S16 C (1 2 3) (0 1 2 3)  LINK S10 S20
SSTR S17 C (1 2 3) (0 1 2 3)  LINK S11 S12
SSTR S18 C (1 2 3) (0 1 2 3)  LINK S12 S13
SSTR S19 C (1 2 3) (0 1 2 3)  LINK S13 S14
SSTR S20 C (1 2 3) (0 1 2 3)  LINK S13 S16
                              LINK S14 S15
```

FIGURE 2.28 – Codage du squelette labdane en sous-structure LSD

`FEXP`, les définitions de fragments sont combinées par les opérateurs `NOT`, `AND` et `OR`. Les noms de squelette peuvent maintenant être directement nommés dans la commande `FEXP` sans passer par la commande `SKEL` qui devient non obligatoire. Par exemple pour imposer le squelette du pinane, l'utilisateur devait utiliser les deux commandes suivantes : `SKEL F2 "PINANE"` (définition du fragment F2) et `FEXP "F2"` (expression des contraintes de sous-structure). Cette dernière peut maintenant contenir directement le nom du squelette : `FEXP "‘PINANE‘"`. La partie du code qui effectue la lecture du fichier d'entrée et celle qui interprète l'expression logique pour construire la stratégie de recherche de sous-structure ont été modifiées en conséquence.

2.4.4 Conclusion

L'extraction des squelettes permet de franchir un palier supplémentaire dans l'intégration de LSD avec la base de données SISTEMAT. Les noms de squelettes prédits par SISTEMAT peuvent directement être entrés dans le fichier d'entrée de LSD sans aucune intervention supplémentaire de l'utilisateur. Le codage des sous-structures au format de LSD est réalisé automatiquement par le programme MOL2AB. Ce programme n'est pas seulement conçu pour être réservé aux squelettes distribués avec LSD, il est utilisable avec toute sous-structure dont l'utilisateur pourrait avoir besoin.

L'extraction des squelettes a été effectuée pour les monoterpènes, les sesquiterpènes et les diterpènes. La base de données SISTEMAT contient d'autres classes de produits naturels. Par conséquent de nombreux autres squelettes restent encore à extraire.

Le programme MOL2AB souffre d'un petit défaut à corriger. Le codage des sous-structures n'a pas été adapté à l'éventualité d'une molécule chargée. En utilisant l'exemple

de l'alcaloïde aporphinique (figure 2.29), avec le squelette aporphine appliqué comme contrainte de sous-structure, LSD n'aboutit à aucune solution. D'après les règles en place dans MOL2AB, l'atome d'azote ne peut pas changer d'hybridation (passer de sp^3 à sp^2) car il ne possède pas d'atome d'hydrogène éliminable. Il faudrait adapter les règles de codage des sous-structures en tenant compte des atomes potentiellement porteurs de charge.

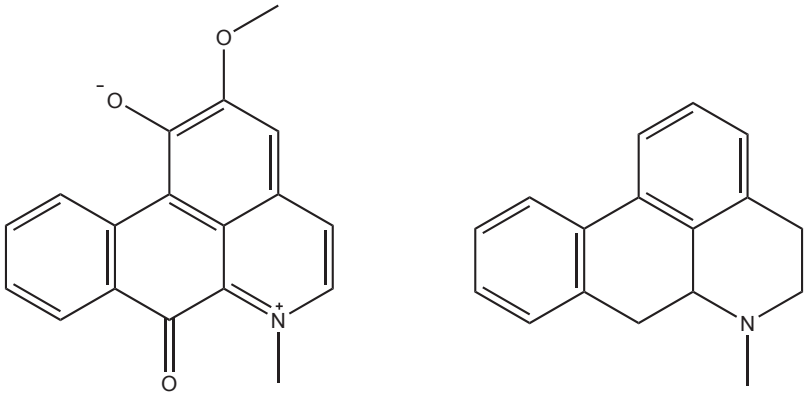


Figure 2.29 – Structure de la 2-O,N-dimethylliriodendronine (à gauche) et squelette aporphine (à droite)

La complémentarité démontrée des deux systèmes LSD et SISTEMAT nous invite à poursuivre l'amélioration de la gestion des contraintes de sous-structures. Il pourrait être judicieux d'utiliser ces contraintes pendant la résolution et non à la fin comme test de validation. Cela permettrait un gain de temps car de nombreuses structures incompatibles peuvent être générées alors qu'il serait possible de les éviter. Une autre piste envisagée serait de permettre un filtrage des solutions après génération de toutes les structures. Un filtrage dynamique pourrait être mis en place, par exemple au niveau de M_EDIT.

2.5 Conclusion

Ces travaux avaient pour but de faire progresser un programme existant en facilitant son utilisation et en augmentant son efficacité.

Le logiciel d'élucidation structurale automatique LSD a été amélioré en supprimant certaines rigidités notamment sur l'interprétation des corrélations et le statut des atomes. Cette dernière amélioration a permis le développement d'une version de LSD gérant les atomes à statut variable.

Des progrès ont été accomplis dans la présentation des résultats. Les tests ont abouti à une version alternative de OUTLSD par l'utilisation de la librairie OEChem.

L'élucidation structurale automatique de produits naturels peut être rendue plus aisée par l'intégration des systèmes LSD et SISTEMAT. L'utilisation des squelettes de

SISTEMAT comme contraintes de sous-structures dans LSD est nettement facilitée par l'extraction et la mise à disposition de squelettes issus de la base de données.

Chapitre 3

Développement d'un outil d'attribution automatique des résonances, le logiciel CASA

3.1 Introduction

Ce chapitre traite de l'écriture d'un programme d'attribution automatique des signaux RMN. La conception de ce programme repose sur la base de travaux anciens du groupe où les travaux présentés dans ce manuscrit ont été réalisés.

Le programme CASA (Computer-Aided Spectral Assignment) a été réécrit totalement avec comme principale nouveauté l'utilisation d'un module de prédiction des déplacements chimiques ^{13}C.

Le chapitre commence par une présentation du cahier des charges exposant les données sur lesquelles le programme s'appuie et la stratégie employée pour l'attribution. Le fonctionnement du programme sera ensuite décrit pour finir avec des exemples d'application. La validation de CASA a été effectuée à l'aide de structures décrites dans la littérature récente.

L'attribution des signaux RMN est un moyen d'aborder le problème de la vérification de structure. En cas d'échec de la vérification ou de doute sur la validité d'une structure, le logiciel LSD peut être utilisé pour proposer des structures alternatives.

3.2 Analyse du problème

3.2.1 But du logiciel

Le but du logiciel est de réaliser l'attribution automatique des signaux RMN ^{13}C et ^{1}H des petites molécules organiques. Pour ce faire, il s'agit de vérifier la cohérence d'une structure chimique avec les données des spectres RMN 2D.

Des travaux précédents du groupe ont montré un effet intéressant lors de la résolution de ce type de problème [179]. Il a été possible de déterminer la bonne structure entre deux structures proches en éliminant celle dont les données n'étaient pas compatibles. En effet si l'attribution des signaux n'est pas possible, la structure est rejetée. La figure 3.1 montre les structures proposées à CASA pour l'attribution. Elles ne diffèrent que par l'inversion d'un

groupement hydroxyle et d'un groupement méthyle. La structure possédant un squelette eremophilane a été acceptée par CASA alors que la structure comportant un squelette eudesmane a été rejetée.

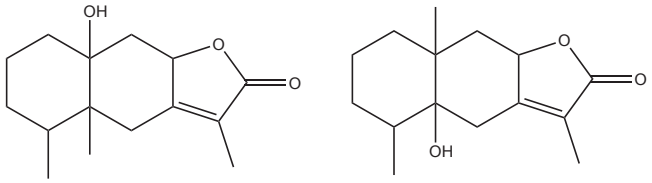


FIGURE 3.1 – Structure acceptée (à gauche) et structure rejetée (à droite) par CASA

L'attribution des signaux peut être vue comme une méthode de vérification de structure car une proposition non cohérente peut être refusée. La finalité du programme CASA est de répondre à la question : la structure est-elle compatible avec les données spectrales ?

3.2.2 Données d'entrée

Les données spectrales utilisées par le programme sont issues de spectres RMN 1D et 2D. Les signaux des spectres ^{13}C et ^{1}H sont attribués en se basant, entre autres, sur la multiplicité des carbones. L'information sur la multiplicité est accessible sur un spectre 1D DEPT ou par une expérience 2D HSQC éditée. Une expérience HSQC éditée fournit un spectre similaire à un spectre HSQC classique, avec un signe opposé pour les groupements CH et CH_3 et les groupements CH_2. La cohérence de l'attribution avec la structure est vérifiée à l'aide des corrélations des spectres HSQC, COSY, HMBC et INADEQUATE. Toute autre information provenant de spectres H2BC ou 1,1-ADEQUATE (voir paragraphe 1.3.2) peut également être utilisée.

Le principe d'interprétation des corrélations est identique à celui de LSD. Des relations de proximité entre atomes sont déduites des corrélations et servent de contraintes pour l'attribution. Les valeurs de longueur de chemin de couplage par défaut sont $^{2}J_{C-H}$ ou $^{3}J_{C-H}$ pour une corrélation HMBC, $^{3}J_{H-H}$ pour une COSY et $^{1}J_{C-C}$ pour une INADEQUATE.

Afin d'éviter un maximum de faux négatifs, c'est-à-dire de ne pas rejeter des structures correctes, il faut prendre en compte différents niveaux d'incertitude. L'existence de couplages à très longue distance peut amener le programme à rejeter une structure. Il est donc nécessaire de ne pas être trop restrictif et de permettre au programme de traiter un certain nombre de corrélations à très longue distance. D'autre part, en cas d'utilisation des valeurs de déplacements chimiques ^{13}C comme contraintes pour l'attribution, il faut tenir compte de l'erreur sur la prédiction.

3.2.3 Problème de satisfaction de contraintes

Pour mener à bien l'exercice d'attribution, le programme est restreint dans ses choix par diverses contraintes. L'exécution de CASA peut donc être assimilée à la résolution d'un problème de satisfaction de contraintes.

Un problème de satisfaction de contraintes peut être défini par une série de n variables possédant chacune un domaine de valeurs et d'une série de contraintes impliquant plusieurs de ces variables. La résolution d'un problème de satisfaction de contraintes revient à trouver toutes les combinaisons possibles de valeurs pour les variables de manière que toutes les contraintes soient respectées. Quand une valeur est attribuée à une variable, la restriction de l'espace de recherche est effectuée par un processus de propagation de contraintes. Le domaine de recherche des valeurs pour les autres variables est ainsi réduit.

Dans le cas de l'attribution, chaque résonance en RMN ^{13}C est assimilée à une variable pour laquelle les valeurs possibles sont les atomes de carbone de la structure. Chaque signal est associé à un domaine de recherche comportant un nombre réduit d'atomes et établi en fonction de critères présentés par la suite. Les contraintes proviennent des corrélations des spectres RMN 2D et, de manière optionnelle, des valeurs de déplacements chimiques. Au cours de l'attribution, le programme effectue une réduction de l'espace de recherche en enlevant les atomes déjà attribués du domaine de recherche pour chaque signal. Pour résoudre le problème de l'attribution, l'algorithme utilisé est un algorithme récursif.

3.2.4 Structure du logiciel

Les différentes tâches successives accomplies par le programme ont été séparées dans trois phases. Ces trois phases du processus d'attribution des signaux par le logiciel CASA sont présentées en figure 3.2.

La phase initiale (phase 0) permet d'initialiser le problème avant d'entamer la résolution. Elle débute par la lecture et l'analyse des données d'entrée afin de vérifier la cohérence des informations. Un prétraitement des corrélations est effectué pour éliminer les informations dupliquées. Le programme effectue une détection des atomes équivalents. La raison de cette détection et la méthode utilisée seront exposées par la suite. Enfin, pour chaque signal RMN ^{13}C, une liste d'atomes candidats à l'attribution est créée. En effet, le programme ne peut tenter d'apparier signal et atome que si un certain nombre de critères sont respectés. Ces critères d'admissibilité des atomes à l'attribution sont détaillés par la suite.

La phase suivante (phase 1) est la résolution du problème par un algorithme récursif qui permet d'envisager et de tester toutes les solutions possibles de manière exhaustive. À chaque étape, un signal non encore attribué est choisi. Puis un atome est choisi dans la liste des candidats à l'attribution et enfin le programme vérifie l'hypothèse d'attribution en cours. La vérification est réalisée en examinant si les corrélations des spectres 2D ne contredisent pas les distances à travers les liaisons entre les atomes de la structure. Si

aucune corrélation ne contredit l'hypothèse, le programme passe à une étape suivante, avec un autre signal. Dans le cas contraire, le programme effectue un retour-arrière et remet en cause les attributions précédentes.

Pour finir, la phase 2 concerne la présentation des résultats à l'utilisateur. Lorsqu'une solution est produite, c'est-à-dire quand tous les signaux ont été attribués et toutes les corrélations inspectées, le programme l'ajoute à un fichier de sortie. Il s'agit d'un fichier texte qui pourra être lu par l'utilisateur. Chaque solution consiste en une liste de paires signal/atome attribués.

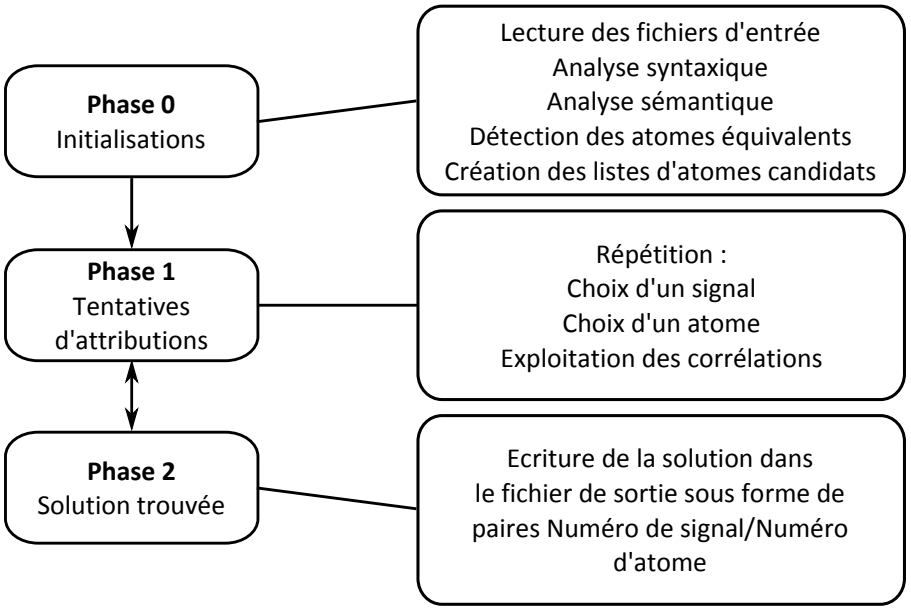


Figure 3.2 – Phases d'exécution de CASA

3.3 Mise en œuvre du programme

3.3.1 Introduction

L'écriture et le fonctionnement détaillé de CASA vont être présentés en mettant l'accent sur les points importants du processus d'attribution.

Le programme a été écrit en langage C en utilisant certains éléments du code source de LSD. L'organisation interne des données est assez similaires à celle de LSD. La syntaxe du fichier d'entrée principal a pour modèle celle d'un fichier d'entrée de LSD. De cette manière, les modifications à apporter pour utiliser un jeu de données avec les deux programmes LSD et CASA sont mineures.

3.3.2 Fichiers d'entrée

Les données et informations utiles sont lues par le programme dans plusieurs fichiers. Au total, deux fichiers d'entrée sont obligatoires et un troisième fichier facultatif peut être utilisé lorsque l'utilisateur souhaite recourir à une prédiction des déplacements chimiques ^{13}C pour l'attribution. Le fichier principal contient les informations des spectres RMN et permet de gérer le contrôle de l'exécution. La structure chimique soumise à l'attribution est enregistrée dans un fichier au format MDL MOL.

3.3.2.1 Fichier de données RMN

Le fichier principal contient des commandes constituées par des mnémoniques suivis par un ou plusieurs paramètres séparés par des blancs. Toutes les commandes utilisables dans les fichiers d'entrée de CASA sont présentées en annexe D. Les paramètres sont classés par types :

- I : un entier positif ou nul ;
- F : un nombre réel ;
- R : un nombre réel optionnel ;
- V : un seul entier positif ou nul ou un ensemble d'entiers positifs ou nuls entre parenthèses séparés par des blancs ;
- O : un entier positif ou nul optionnel ;
- Ln : la référence à une liste de signaux ou d'atomes, n est strictement positif ;
- B : remplace I ou Ln ;
- H : un signe optionnel : + ou - ;
- P : un chemin d'accès à un fichier ;
- S : un ensemble d'entiers positifs séparés par des blancs ;

Le codage des données pour l'attribution des signaux RMN du camphre (figure 3.3) est donné en figure 3.4.

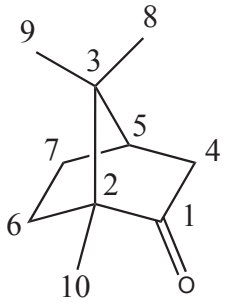


Figure 3.3 – Structure du camphre

L'ensemble des spectres utilisés est présenté en annexe E. Tous les signaux observés sur le spectre ^{13}C sont désignés à l'aide de la commande `DEPT`. Un signal est décrit par un

```
ENTR 1
AWCS 1

CCSF "camphor_cstable.txt"

DEPT 1 2 0 219.33      HMBC 1 4       HMBC (4 5) 6
DEPT 2 3 0 57.65       HMBC 1 5       HMBC (4 5) 7
DEPT 3 3 0 46.76       HMBC 1 6       HMBC (4 5) 8
DEPT 4 3 2 43.29       HMBC 1 10      HMBC (4 5) 9
DEPT 5 3 1 43.09       HMBC 2 5       HMBC 5 4
DEPT 6 3 2 29.95       HMBC 2 6       HMBC 6 5
DEPT 7 3 2 27.08       HMBC 2 7       HMBC 6 7
DEPT 8 3 3 19.77       HMBC 2 8       HMBC 6 10
DEPT 9 3 3 19.15       HMBC 2 9       HMBC 7 4
DEPT 10 3 3 9.25       HMBC 2 10      HMBC 7 6
                       HMBC 3 4       HMBC 8 9
HSQC 4 4               HMBC 3 6       HMBC 9 5
HSQC 5 5               HMBC 3 7       HMBC 9 8
HSQC 6 6               HMBC 3 8       HMBC 10 6
HSQC 7 7               HMBC 3 9
HSQC 8 8               HMBC 3 10
HSQC 9 9
HSQC 10 10

COSY 4 5
COSY 6 7
```

Figure 3.4 – Fichier de données CASA pour l'attribution des signaux du camphre

numéro, une hybridation, une multiplicité et un déplacement chimique. La numérotation des signaux peut être faite de manière arbitraire. Dans cet exemple, les signaux sont numérotés par ordre décroissant de déplacements chimiques ^{13}C. L'hybridation peut être fixée ou non. Dans le cas d'une hybridation variable, les différentes hybridations possibles sont placées entre parenthèses. La multiplicité est définie à l'aide d'un spectre DEPT ou d'un spectre HSQC édité. Le déplacement chimique est optionnel. Un des aspects originaux de CASA est de pouvoir fournir une attribution sans utiliser les valeurs de déplacement chimique ^{13}C.

Les numéros des carbones issus du spectre 1D sont reportés sur les projections des spectres 2D HSQC et HMBC. Le protocole d'interprétation manuelle des spectres est identique à celui de LSD (voir paragraphe 2.2.2). Le codage des données des spectres 2D est accompli à l'aide des commandes `HSQC`, `COSY` et `HMBC`. Les commandes `COSY` et `HMBC` permettent la définition de corrélations impliquant des groupes d'atomes ayant des déplacements chimiques proches. De plus, des paramètres optionnels peuvent être ajoutés pour définir un intervalle de longueur de chemin de couplage entre atomes. Les spectres H2BC et 1,1-ADEQUATE révèlent des corrélations interprétables comme des corrélations HMBC issues de couplages $^2J_{C-H}$. Il n'existe pas de commande spécifique pour la déclaration des corrélations H2BC ou 1,1-ADEQUATE. Il est cependant possible d'utiliser une commande `HMBC` suivie du paramètre optionnel 2. Par exemple, une corrélation H2BC entre le carbone 3 et le proton 5 peut être déclarée à l'aide de la commande `HMBC 3 5 2`.

La commande `ENTR 1` présente en début du fichier exemple fait partie des commandes permettant de contrôler l'exécution du programme. Elle permet l'affichage et le contrôle des données après lecture du fichier d'entrée.

La commande `AWCS 1` permet de spécifier le mode de fonctionnement du programme. Dans l'exemple, CASA utilise conjointement les données des spectres 2D et les valeurs de déplacements chimiques. Il est possible de faire fonctionner le programme en utilisant uniquement les corrélations des spectres 2D ou uniquement les valeurs de déplacements chimiques ^{13}C.

Étant donné le mode de fonctionnement choisi, il faut fournir au programme des valeurs de déplacements chimiques prédites. Celles-ci sont regroupées dans un fichier spécifique dont le chemin d'accès est donné par la commande `CCSF`.

Une attribution facilement déduite lors de l'analyse des spectres peut être fournie comme point de départ au processus d'attribution avec la commande `ASGN`. Par exemple, la molécule de camphre ne possède qu'un groupement cétone et le spectre ^{13}C ne comporte qu'un signal à δ 219,33 qui correspond à un atome de carbone hybridé sp^2. Ce signal peut donc directement être attribué à l'atome de carbone C1 du groupement cétone.

Il est possible d'ajouter des informations supplémentaires issues de l'interprétation des déplacements chimiques et de la multiplicité des signaux des spectres 1D ^{1}H et ^{13}C. La commande `PROP` permet d'associer à un signal des propriétés d'environnement pour l'atome auquel il sera attribué. Dans le cas du camphre, les trois signaux ^{1}H correspondant

à des groupements méthyles apparaissent sous forme de singulets entre 0 et 1 ppm et les signaux ^{13}C S8, S9 et S10 se situent entre 9 et 20 ppm. Cela signifie que les groupements méthyles sont liés à des atomes de carbone quaternaires. Il s'agit donc de forcer l'attribution des signaux S8, S9 et S10 à des atomes de carbone liés à des carbones quaternaires. Pour coder cette contrainte, il faut d'abord créer une liste des signaux S8, S9 et S10 avec la commande `LNMR L1 8 9 10` et une liste des atomes de carbone quaternaires C2 et C3 avec la commande `LMOL L2 2 3`. Enfin, la commande `PROP L1 0 L2` oblige tous les signaux de la liste `L1` à être attribués à des atomes de carbone qui ont tous leurs voisins parmi la liste `L2`. Dans la commande `PROP`, le 0 qui signifie « tous » peut être remplacé par un 1 car un groupement méthyle ne possède qu'un voisin.

Lors de la phase d'initialisation, le programme procède à une analyse syntaxique et sémantique du contenu du fichier principal. Le respect de la nomenclature des commandes de CASA est ainsi vérifié. Des éventuelles fautes de frappe peuvent être détectées et la cohérence des données inspectée. Par exemple, le programme examine les numéros de signaux utilisés dans les corrélations et repère s'ils ont bien été déclarés préalablement.

3.3.2.2 Fichier de structure

La structure chimique dont l'utilisateur souhaite faire l'attribution est donnée sous la forme d'un fichier MDL MOL qui peut être généré à l'aide d'un logiciel de dessin de structure chimique. Le logiciel utilisé pour les tests de CASA est ACD/ChemSketch [180].

Après lecture et extraction des informations du fichier, le programme dispose de la table de connectivité de la structure. À partir de cette table de connectivité, le programme calcule la matrice des distances entre atomes (nombre de liaisons séparant les atomes). Cette dernière intervient lors de la résolution au moment de l'interprétation des corrélations. En effet, les contraintes de distance entre atomes déduites des corrélations doivent satisfaire les distances entre atomes de la structure.

L'interprétation des informations présentes dans le fichier MDL MOL (nature des éléments chimiques, charges et liaisons) permet de déterminer, pour chaque atome de carbone, son hybridation et sa multiplicité. Ces deux propriétés ne sont pas décrites explicitement dans le fichier mais sont essentielles pour la résolution car elles comptent parmi les critères d'attribution.

La valeur de déplacement chimique ^{13}C est un autre critère d'attribution. La prédiction des déplacements chimiques ^{13}C de la structure est transcrite dans un fichier différent.

3.3.2.3 Fichier de déplacements chimiques prédits

Un programme basé sur le module de prédiction libre et gratuit nmrshiftdb2 [181–183] permet de générer automatiquement le fichier de déplacements chimiques ^{13}C à partir du fichier de structure. Il s'agit d'une version autonome de l'outil de prédiction disponible sur

le site web du projet nmrshiftdb2. Toute autre méthode de prédiction des déplacements chimiques est envisageable. L'utilisateur doit alors écrire les valeurs dans un fichier texte selon une syntaxe simple.

```
 1   216.90    5.00
 2    57.30    5.00
 3    46.60    5.00
 4    43.10    5.00
 5    43.13    5.00
 6    30.07    5.00
 7    27.20    5.00
 8    19.55    5.00
 9    19.55    5.00
10     9.40    5.00
```

FIGURE 3.5 – Fichier contenant la prédiction des déplacements chimiques ^{13}C du camphre

La composition du fichier de déplacements chimiques prédits pour le camphre est donnée comme exemple en figure 3.5. Chaque ligne du fichier est associée à un atome de carbone de la structure. Les trois champs correspondent :

- au numéro de l'atome dans le fichier de structure ;
- à la valeur du déplacement chimique ^{13}C estimé ;
- à l'incertitude sur la valeur.

Par exemple, le déplacement chimique du carbone numéro 1 de la structure est estimé à (216,90 ± 5,00) ppm.

3.3.3 Algorithme de résolution

3.3.3.1 Prétraitement des corrélations

Les corrélations constituent l'essence même du principe de CASA. Un soin tout particulier a donc été apporté à leur vérification, avant de commencer la résolution, pendant la phase 0. De même que dans LSD, le prétraitement des corrélations a pour but le contrôle de la cohérence et de la non redondance des données.

Les corrélations sont comparées les unes aux autres et confrontées aux attributions manuelles faites par la commande `ASGN` pour gérer les conflits d'information et supprimer les corrélations qui n'apportent aucune information supplémentaire. Un conflit peut aboutir soit à un message d'erreur, soit à la modification d'une corrélation. Les corrélations inutiles sont marquées comme non valides, le programme ne cherchera pas à les exploiter pendant l'attribution.

Le prétraitement des corrélations est différencié avec d'abord les corrélations non variables et ensuite les corrélations variables.

La figure 3.6 présente les tests (a-d) subis par les corrélations non variables.

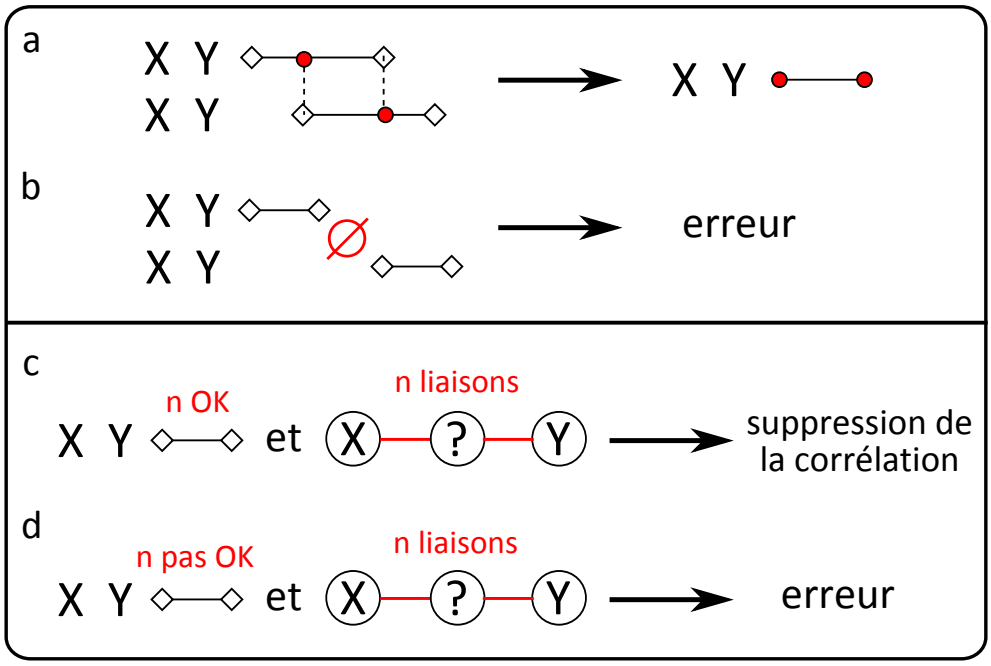


FIGURE 3.6 – Prétraitement des corrélations non variables : le comportement du programme en présence de différents cas est résumé de manière synthétique. Pour le détail, voir texte. (Remarque : les atomes de carbone et d'hydrogène directement liés possèdent des numéros X et Y identiques)

Les deux premiers cas (a et b) concernent la comparaison des corrélations non variables entre elles pour éviter une redondance dans les informations. Le programme recherche les corrélations dupliquées entre deux atomes X et Y et vérifie si les intervalles de nombre de liaisons entre atomes sont compatibles. Pour ce faire, l'intersection des intervalles est calculée. Si l'intersection n'est pas vide, il ne subsiste qu'une corrélation valide avec comme intervalle l'intersection des deux corrélations (a). Si l'intersection est un ensemble vide, il y a incompatibilité entre les deux corrélations et le programme s'arrête en prévenant l'utilisateur par un message d'erreur (b).

Ensuite, le programme compare toutes les corrélations non variables encore valides avec les attributions déjà effectuées avec la commande `ASGN` (c et d). Il s'agit de supprimer les corrélations qui n'apportent pas d'information. Le nombre de liaisons entre deux atomes qui corrèlent et qui sont déjà attribués doit être compris dans l'intervalle de couplage associé à la corrélation (c). Dans le cas contraire, le programme s'arrête avec l'affichage d'un message d'erreur (d). Il faut noter le cas particulier des cycles de trois atomes qui n'amène pas de message d'erreur. Chaque atome peut corréler avec l'un de ses voisins et la distance minimale autorisée peut être de 2 liaisons.

La figure 3.7 présente les tests (a-f) subis par les corrélations variables. Ces corrélations concernent des listes de carbones ayant des déplacements chimiques proches.

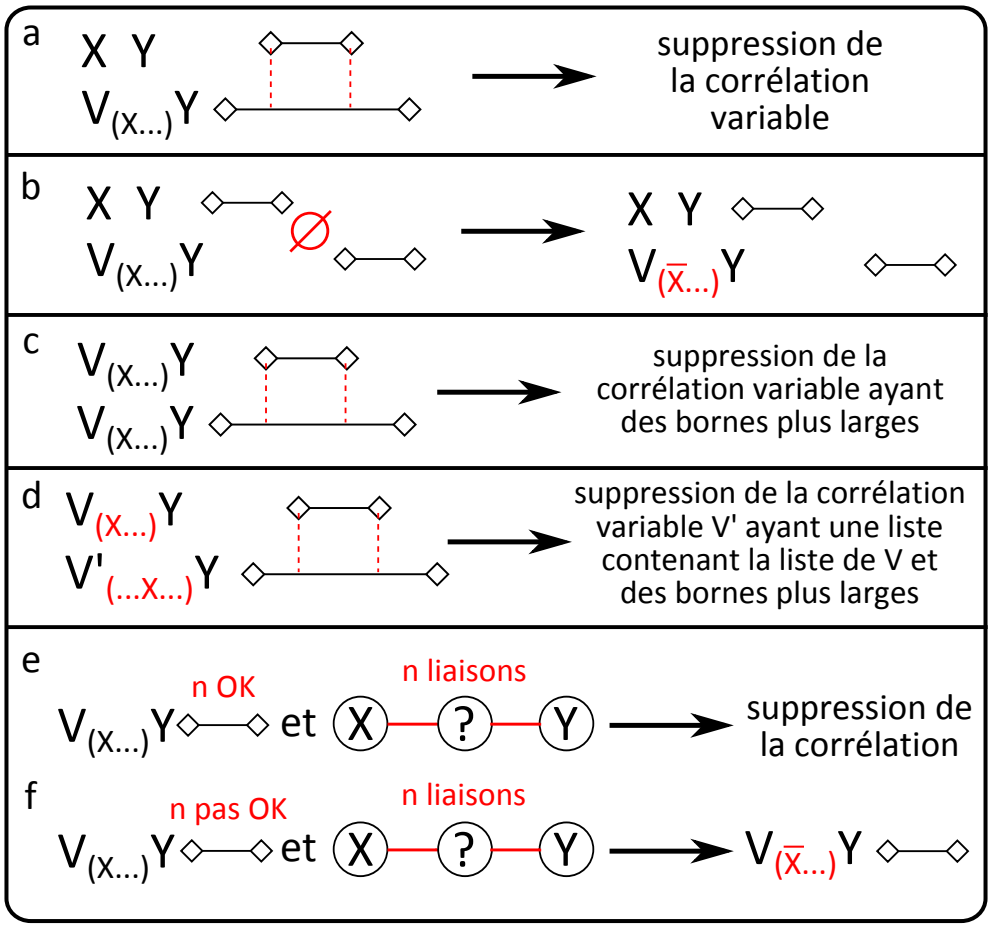


FIGURE 3.7 – Prétraitement des corrélations variables : le comportement du programme en présence de différents cas est résumé de manière synthétique. Pour le détail, voir texte. (Remarque : les atomes de carbone et d'hydrogène directement liés possèdent des numéros X et Y identiques)

Le programme commence par vérifier si l'indétermination sur une corrélation variable entre Y et une liste d'atomes V qui contient X peut être expliquée (a et b). Une corrélation hypothétique entre X et Y peut avoir été déclarée par ailleurs sous la forme d'une corrélation non variable. Les distances minimale et maximale autorisées pour la corrélation non variable doivent être égales ou incluses dans celles de la corrélation variable (a). Si tel est le cas, la corrélation variable est supprimée. Par contre si l'intersection des distances autorisées est nulle, il y a contradiction entre les corrélations (b). L'information de la corrélation non variable est considérée comme prioritaire par rapport à celle de la corrélation variable. L'atome X apportant la contradiction est supprimé de la liste V de la corrélation variable. Lorsqu'une liste est réduite à un unique atome, la corrélation variable devient une corrélation non variable. En retirant une alternative, l'indétermination sur la corrélation est simplifiée voire levée. Cela permet au programme d'aborder la phase 1 avec

un jeu de données ayant le moins d'ambigüité possible.

Dans la partie suivante, les corrélations variables sont comparées entre elles (c et d). Dans un premier temps, il faut éliminer les corrélations variables dupliquées, c'est-à-dire dont les listes d'atomes V sont identiques. Il faut également que les distances autorisées soient identiques ou qu'elles soient incluses. Dans ce dernier cas, la corrélation ayant les bornes les plus larges est éliminée. Dans un second temps, on procède à l'élimination des corrélations variables expliquées par d'autres corrélations variables. Pour qu'une corrélation soit expliquée, une seconde corrélation doit être totalement incluse dans celle-ci (d). La liste d'atomes de la première (V') doit contenir la totalité de la seconde (V) et les bornes de distance de la première doivent être identiques ou plus larges que celles de la seconde.

Dans la dernière partie, de la même manière qu'avec les corrélations non variables, le programme supprime les corrélations variables qui n'apportent pas de nouvelle information par rapport aux attributions déjà effectuées (e). En cas de conflit d'une corrélation hypothétique entre X et Y avec les attributions, l'atome X est supprimé de la liste (f).

Le prétraitement des corrélations du camphre est décrit sur la figure 3.8.

Les chiffres entre crochets indiquent les bornes inférieures et supérieures de longueurs de chemin de couplage retranscrites en nombre de liaisons entre les deux atomes impliqués dans la corrélation.

Dans le fichier de données, toutes les corrélations ont été déclarées avec des longueurs de chemin de couplage par défaut. Les deux premières corrélations sont des COSY car les distances minimale et maximale sont égales à 1. Les atomes correspondants (4 et 5 d'une part, 6 et 7 d'autre part) doivent donc être liés. Les autres corrélations sont des HMBC avec des bornes inférieure et supérieure respectivement égales à 1 et 2. Les atomes correspondants doivent donc être séparés par une ou deux liaisons.

Le prétraitement a supprimé quelques corrélations. Les corrélations HMBC C6/H7 et C7/H6 sont supprimées car rendues inutiles par l'existence d'une corrélation COSY entre les protons 6 et 7. La corrélation C5/H4 est supprimée pour la même raison. Les corrélations HMBC C9/H8 et C8/H9 sont redondantes, une seule subsiste après le prétraitement. Il en est de même pour les corrélations C10/H6 et C6/H10. On remarque que plusieurs corrélations variables ont été supprimées. Celles-ci sont expliquées par au moins une corrélation non variable. Par exemple, la corrélation HMBC entre le groupe de carbones C4-C5 et le proton H6 peut être expliquée par la corrélation HMBC C6/H5. Les corrélations HMBC entre ce même groupe de carbone et les protons H7 et H9 sont aussi expliquées par d'autres corrélations.

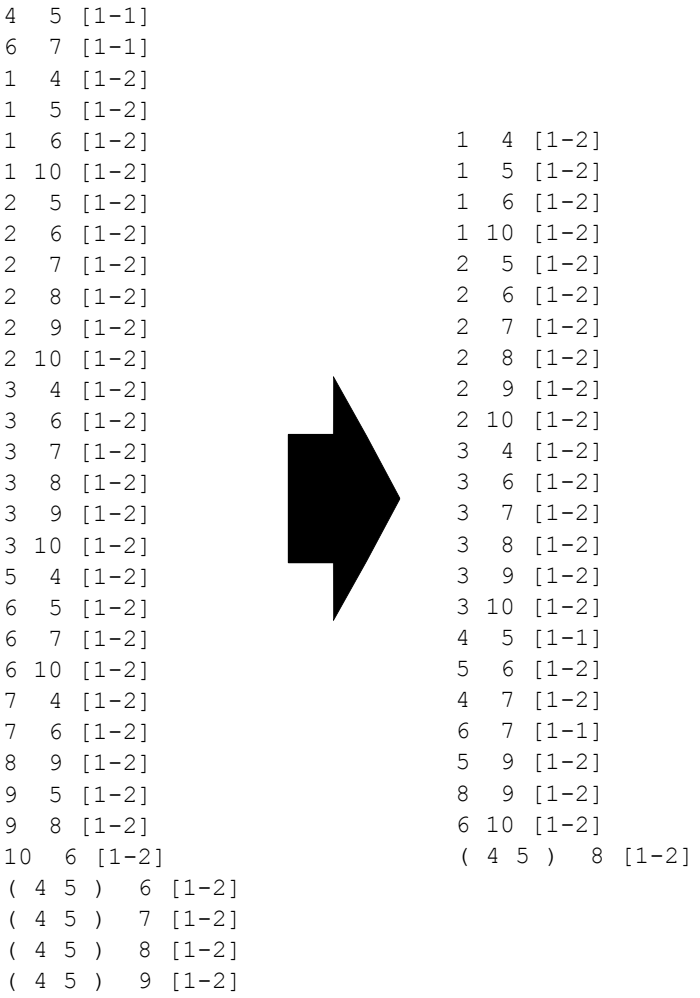


Figure 3.8 – Corrélations du camphre avant (à gauche) et après prétraitement (à droite)

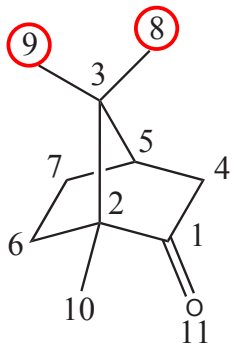
3.3.3.2 Gestion des atomes équivalents topologiquement

Des atomes équivalents topologiquement ont une distance (en nombre de liaisons) identique avec tous les autres atomes de la molécule.

Lorsqu'une structure candidate à l'attribution possède deux atomes équivalents topologiquement C_i et C_j, l'appariement du signal S_i avec l'atome C_i et de S_j avec C_j entraine également l'appariement de S_i avec C_j et de S_j avec C_i. Cet effet est gênant car il a pour conséquence une augmentation du nombre de solutions. Afin de résoudre ce problème, un système de détection des atomes équivalents a été mis en place.

L'algorithme proposé par Morgan en 1965 [184] peut produire des résultats erronés [185]. Nous avons choisi de ne pas implémenter cet algorithme dans CASA car nous disposions d'une alternative plus commode à mettre en place. Il s'agit du système InChI (International Chemical Identifier) [186] présent dans LSD pour la vérification de l'originalité des structures et qui détecte les atomes équivalents. Ce système d'identification est distribué par l'IUPAC (International Union of Pure and Applied Chemistry). Il fournit un codage des structures sous la forme de chaînes de caractère. En complément de la numérotation canonique des atomes de la molécule, des données auxiliaires comme la liste des atomes équivalents sont également produites.

Par exemple, le code InChI du camphre est donné sur la figure 3.9. Les informations auxiliaires contiennent deux suites de caractères particulièrement intéressantes : la numérotation canonique débutant par « /N » et la liste des atomes équivalents débutant par « /E ». Les atomes équivalents sont regroupés entre parenthèses.



InChI=1/C10H16O/c1-9(2)7-4-5-10(9,3)8(11)6-7/h7H,4-6H2,1-3H3
AuxInfo=1/0/N:8,9,10,7,6,4,5,1,3,2,11/E:(1,2)
/rA:11CCCCCCCCCCO/rB:s1;s2;s1;s3s4;s2;s5s6;s3;s3;s2;d1;/rC:;;;;;;;;;;

Figure 3.9 – Structure et code InChI du camphre

La récupération des informations auxiliaires et leur interprétation permet de construire des classes d'atomes équivalents. Dans l'exemple, les atomes 1 et 2 (en numérotation canonique) sont repérés comme étant équivalents. La table de correspondance entre la

numérotation canonique et la numérotation initiale permet de déduire l'équivalence des atomes de carbone 8 et 9. Pendant la phase d'initialisation de CASA, ces atomes sont placés dans une classe d'équivalence. Lors d'une tentative d'attribution, un seul atome de la classe pourra être choisi. La détection des équivalences topologiques réduit ainsi le nombre de solutions produites par CASA.

L'emploi des classes d'équivalence ne cause pas d'incompatibilité avec la prédiction des déplacements chimiques ^{13}C si celle-ci ne tient pas compte des informations de stéréochimie et donne des valeurs identiques pour des carbones équivalents. C'est le cas avec nmrshiftdb2 mais si l'utilisateur dispose d'une autre méthode de prédiction donnant des déplacements chimiques différents, cela peut conduire à l'échec de l'attribution. La commande `CCLA` permet de contrôler l'usage des classes d'atomes équivalents. L'utilisateur peut donc abandonner leur utilisation en cas de besoin.

3.3.3.3 Critères d'attribution

La création de listes d'atomes candidats à l'attribution permet de simplifier le problème combinatoire. Dans le cas général, un exercice d'appariement accepte un nombre N de solutions égal à :

$$N = \frac{n!}{(n-k)!}$$

Avec, dans le cas présent, n le nombre d'atomes de carbone de la molécule et k le nombre de signaux RMN ^{13}C à attribuer. Dans le cadre de CASA, le nombre de solutions potentielles à examiner par le programme est réduit du fait du statut des signaux et des atomes. Avant de commencer la résolution, le problème combinatoire est réduit par une analyse préliminaire des données. Le programme construit, pour chaque signal, une liste d'atomes potentiellement candidats à l'attribution en fonction des critères suivantes : hybridation, multiplicité, déplacement chimique et propriétés de voisinage. Pendant la résolution, le programme inspectera de manière récursive toutes les combinaisons d'attribution possibles. À ce stade, si au moins un signal possède une liste vide, on sait qu'aucune solution ne pourra être produite. Le programme s'arrête alors automatiquement.

L'utilisation de la valeur des déplacements chimiques comme critère d'attribution est optionnelle. Par défaut, CASA se base uniquement sur l'hybridation et la multiplicité pour la construction des listes d'atomes.

L'utilisateur peut ajouter des informations supplémentaires à partir d'une analyse des spectres 1D ^{1}H et ^{13}C. L'interprétation des déplacements chimiques et de la multiplicité des signaux permet d'associer à un signal des propriétés d'environnement pour l'atome auquel il sera attribué. Ces propriétés de voisinage ou d'environnement sont utilisées lors de la création des listes d'atomes. Par exemple, l'attribution d'un signal ^{13}C aux environs de 70 ppm peut être forcée à un atome de carbone possédant exactement un atome d'oxygène parmi ses voisins.

La création des listes d'atomes candidats à l'attribution pour les signaux du camphre est illustrée sur la figure 3.10.

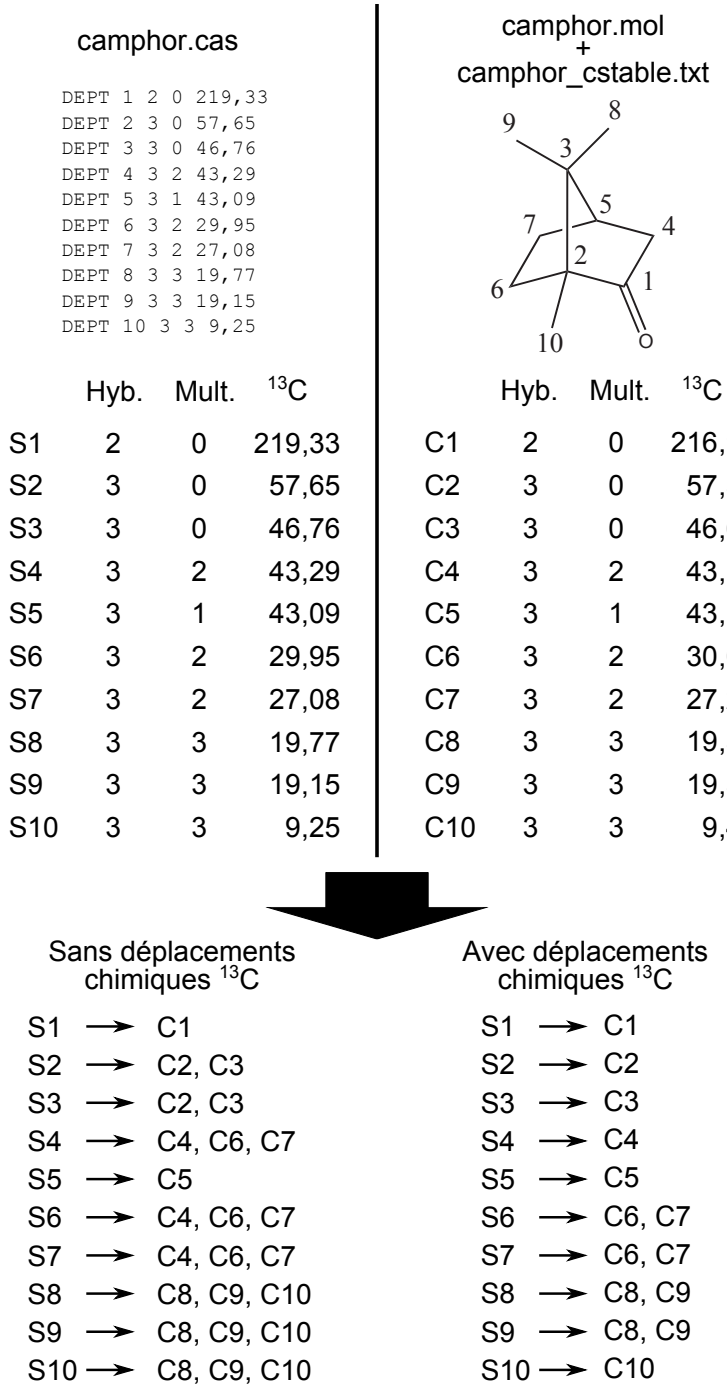


Figure 3.10 – Construction des listes d'atomes pour l'attribution des signaux du camphre

Après lecture du fichier de données RMN et interprétation du fichier de structure, le programme dispose de toutes les informations sur les critères d'attribution. En fonction du mode de fonctionnement choisi, les listes peuvent être construites soit en se basant sur

l'hybridation et la multiplicité uniquement, soit avec le déplacement chimique en complément. Les éventuelles propriétés d'environnement interviennent quel que soit le mode de fonctionnement choisi. Dans l'exemple, le signal S1 n'a qu'un candidat. Il s'agit de l'atome C1 qui est le seul atome de carbone hybridé sp^2. En revanche, le signal S8 possède trois candidats lorsque la prédiction n'est pas utilisée. Ce sont les atomes C8, C9 et C10 qui sont les trois atomes de carbone appartenant à des groupements méthyles (hybridation sp^3 et multiplicité égale à trois). Après ajout de la prédiction des déplacements chimiques ^{13}C comme critère supplémentaire, on observe une réduction des possibilités d'attribution. L'utilisation des valeurs de déplacement chimique tient compte de l'erreur sur la prédiction. Dans le cas du camphre, l'erreur sur la prédiction est de 5 ppm pour toutes les valeurs. Il s'agit de l'erreur minimum choisie par défaut pour la prédiction fournie par nmrshiftdb2. Le signal S8 ne possède plus que deux atomes candidats (les atomes C8 et C9). On peut remarquer que les atomes équivalents C8 et C9 ont la même valeur de déplacement chimique ^{13}C estimée avec nmrshiftdb2. Ces deux atomes étant équivalents, un seul sera testé pour l'attribution du signal S8. On constate que l'unique indétermination concerne les signaux S6 et S7 qui ont deux atomes candidats (C6 et C7). Tous les autres signaux n'ont qu'un seul atome candidat pour l'attribution.

La phase 0 s'achève avec la création des listes. Le processus d'attribution et de vérification des attributions à l'aide des corrélations peut commencer.

3.3.3.4 Attribution des signaux

Lors de la phase 1, le programme utilise les listes et les corrélations comme contraintes pour l'attribution. Le processus d'attribution est accompli par la répétition des actions suivantes :

- choix d'un signal ;
- choix d'un atome ;
- exploitation des corrélations.

Un procédé récursif permet d'explorer tout le domaine de recherche jusqu'à épuisement des hypothèses.

À chaque nouvelle tentative d'attribution, un signal est choisi en fonction du nombre d'atomes présents dans sa liste de candidats. La règle étant que plus le nombre d'atomes candidats pour un signal est petit, plus vite il sera attribué. Un signal ne possédant qu'un seul atome candidat sera donc attribué dès le début de la résolution. De plus, à taille de liste égale, les signaux sont choisis en priorité dans une liste de signaux précédemment impliqués dans les corrélations des signaux venant d'être attribués. Le but de ces règles est de savoir le plus rapidement possible si une solution existe ou non.

Après le choix d'un signal, un atome est sélectionné dans la liste des candidats. Tous les atomes de la liste seront successivement testés. Lors du choix des atomes, le programme

tient compte des classes d'atomes équivalents topologiquement. Uniquement le premier atome de la classe peut être sélectionné pour l'attribution.

À chaque nouvelle hypothèse d'attribution, un test est effectué pour vérifier si celle-ci ne mène pas à une situation sans solution. Les listes d'atomes candidats pour chaque signal sont mises à jour dynamiquement quand un atome est sélectionné. Celui-ci est supprimé des listes des signaux non encore attribués. Si une liste devient vide, cela signifie que l'hypothèse courante n'est pas valable.

Enfin, toutes les corrélations impliquant tous les signaux déjà attribués (incluant le signal actuellement analysé) sont vérifiées. L'hypothèse est confirmée si aucune corrélation ne vient contredire le nombre de liaisons entre les atomes. Les corrélations sont ainsi expliquées au fur et à mesure de la résolution. Dans le cas d'une corrélation variable, au moins un signal du groupe doit satisfaire les contraintes de distance pour permettre à la corrélation d'être expliquée.

Lors de la vérification d'une corrélation non variable, si un seul des deux signaux a été attribué, le second est placé dans une liste de signaux prioritaires pour l'attribution.

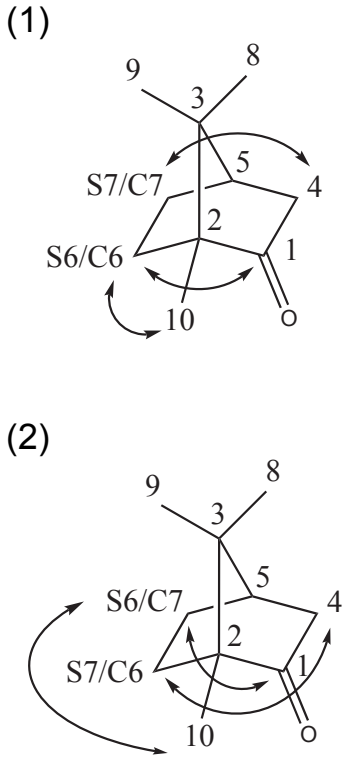
Avec le principe adopté, l'existence de couplages à très longue distance peut mener à un rejet de la structure. En effet, une corrélation non standard est la cause d'une violation de contrainte. Pour éviter un échec de l'attribution qui constituerait un faux négatif, le programme dispose d'un mécanisme permettant d'accepter un certain nombre de violations spécifié par l'utilisateur. La commande `ELIM` possède deux paramètres : le nombre maximum de corrélations entrainant des violations et une longueur de chemin de couplage maximale autorisée n. Cette dernière correspond à un couplage ${}^{n}J$ pour une corrélation HMBC et ${}^{n+1}J$ pour une corrélation COSY.

Pour qu'une solution soit produite, deux conditions doivent être remplies. Tous les signaux doivent être attribués et toutes les corrélations doivent être vérifiées et expliquées. Dans ce cas, le programme passe à la phase 2 : l'écriture de la solution dans un fichier de sortie. Une solution est présentée sous la forme d'une liste de lignes composées d'un numéro de signal RMN ^{13}C suivi par le numéro de l'atome correspondant.

L'exécution du programme se termine quand toutes les combinaisons de paires signal/atome possibles ont été considérées.

Pour finir l'attribution des signaux du camphre, les listes d'atomes et les corrélations présentées dans les paragraphes précédents vont être utilisées. Lorsque les déplacements chimiques sont utilisés comme critères complémentaires pour l'attribution, les signaux S1, S2, S3, S4, S5 et S10 n'ont qu'un seul atome candidat. Ils sont donc attribués dès le début de la résolution. Les signaux S8 et S9 possèdent deux candidats mais il s'agit des atomes C8 et C9 qui sont équivalents. Une seule combinaison d'attribution est donc effectuée. Il ne reste que les signaux S6 et S7 qui ont deux candidats : les atomes C6 et C7. La figure 3.11 montre les deux possibilités d'attribution pour ces deux signaux. Toutes les corrélations des autres signaux précédemment attribués peuvent être expliquées. La

solution 1 est jugée valable car les corrélations des signaux S6 et S7 sont en accord avec les distances entre atomes. Par contre, la solution 2 comporte trois violations. Ce sont les corrélations entre les signaux 4 et 7, 1 et 6, et 6 et 10. Les atomes associés à ces signaux sont trop éloignés. Ils sont séparés par trois liaisons alors que le nombre maximum de liaisons autorisé est deux. Le fichier d'entrée ne comportant pas de commande `ELIM`, cette solution est donc refusée. L'exemple du camphre n'accepte donc qu'une solution : la solution 1.




```
1  4 [1-2]
1  5 [1-2]
1  6 [1-2]
1 10 [1-2]
2  5 [1-2]
2  6 [1-2]
2  7 [1-2]
2  8 [1-2]
2  9 [1-2]
2 10 [1-2]
3  4 [1-2]
3  6 [1-2]
3  7 [1-2]
3  8 [1-2]
3  9 [1-2]
3 10 [1-2]
4  5 [1-1]
5  6 [1-2]
4  7 [1-2]
6  7 [1-1]
5  9 [1-2]
8  9 [1-2]
6 10 [1-2]
( 4 5 )  8 [1-2]
```

Figure 3.11 – Hypothèses d'attribution des signaux du camphre et corrélations

3.3.4 Prédiction des déplacements chimiques ^{13}C

L'utilisation d'une méthode de prédiction des valeurs de déplacements chimiques ^{13}C constitue la principale différence avec l'ancien système CASA. Le module de prédiction nmrshiftdb2 a été choisi car il s'agit d'un outil libre et gratuit. La prédiction repose sur une base de données de codes HOSE associés à des déplacements chimiques.

Un programme est mis à disposition de l'utilisateur pour réaliser la prédiction des déplacements chimiques à partir du fichier de structure à l'aide de nmrshiftdb2 ainsi que l'écriture automatique du fichier contenant les valeurs prédites et l'estimation de l'erreur sur la prédiction.

Le déplacement chimique prédit est la moyenne de toutes les valeurs présentes dans la base de données pour le code HOSE du carbone considéré. Pour estimer l'incertitude sur la valeur prédite, les valeurs de déplacement chimique minimum et maximum présentes dans la base de données pour le code HOSE sont récupérées.

L'incertitude Δ est définie de la façon suivante :

$$\Delta = \max\left(\left|\delta_{moy} - \delta_{max}\right|, \left|\delta_{moy} - \delta_{min}\right|\right)$$

Elle est égale au maximum des valeurs absolues des différences de la moyenne δ_{moy} avec le minimum δ_{min} et le maximum δ_{max}. Si la valeur tombe en dessous du seuil par défaut placé à 5,0 ppm, l'incertitude est automatiquement ramenée à cette valeur seuil.

Cette méthode d'estimation de l'incertitude manque de précision et peut se révéler comme une autre cause d'échec de l'attribution et donc de production de faux négatifs. Lors de la création des listes d'atomes candidats pour chaque signal, le programme CASA doit tenir compte de l'erreur sur l'estimation de l'incertitude. Un facteur multiplicatif défini par l'utilisateur permet de disposer d'une sécurité sur l'incertitude. De plus, toute valeur d'incertitude se situant en dessous d'une valeur seuil définie par l'utilisateur est automatiquement ramenée à cette valeur. Ces deux paramètres supplémentaires sont donnés dans le fichier principal à l'aide des commandes `SCLF` et `TOLE`.

Pour améliorer l'évaluation de l'incertitude sur la prédiction des déplacements chimiques, il serait envisageable de prendre en considération le nombre d'occurrences des codes HOSE dans la base de données. En effet, une prédiction de déplacement chimique est d'autant plus fiable que la valeur est présente en grand nombre dans la base de données.

3.3.5 Conclusion

La réécriture du programme CASA a été associée à une méthode de prédiction des déplacements chimiques ^{13}C. Ceci permet de profiter des informations structurales contenues dans les valeurs de déplacements chimiques.

L'originalité de CASA est l'utilisation des données des spectres RMN 2D. Les signaux RMN peuvent être attribués uniquement si les corrélations sont expliquées. La liberté sur les contraintes peut être adaptée à chaque problème de manière à limiter le nombre de faux négatifs et de faux positifs. Cette liberté est soumise au jugement de l'utilisateur.

CASA est distribué sous la forme d'un logiciel libre sous licence GPL [187]. Tout comme pour LSD, l'accès au code source est garanti et des versions précompilées pour Windows/DOS et MacOS X sont également disponibles [188].

3.4 De l'attribution à la vérification de structure

Afin de démontrer l'efficacité du logiciel, des exemples de structures de produits naturels publiées dans la littérature récente ont été choisis. La complexité des exemples est

variable.

L'utilisation de CASA comme logiciel de vérification de structure sera démontrée, de même que l'utilisation du logiciel LSD comme générateur de structures aléatoires en cas d'échec de l'attribution.

Dans un soucis de clarté et de simplicité, les signaux ^{13}C dans le fichier de données de CASA et les atomes de carbone dans le fichier MDL MOL possèdent la même numérotation, de telle manière qu'un signal S_i soit supposé être originaire de l'atome C_i.

3.4.1 Une alternative à l'utilisation de la prédiction des déplacements chimiques : les propriétés d'environnement

Le composé utilisé comme exemple est la Guttiferone H, de formule moléculaire $C_{38}H_{50}O_6$ (figure 3.12). Il s'agit d'une benzophénone extraite de *Garcinia xanthochymus* Hook. f. (Clusiaceae) [189].

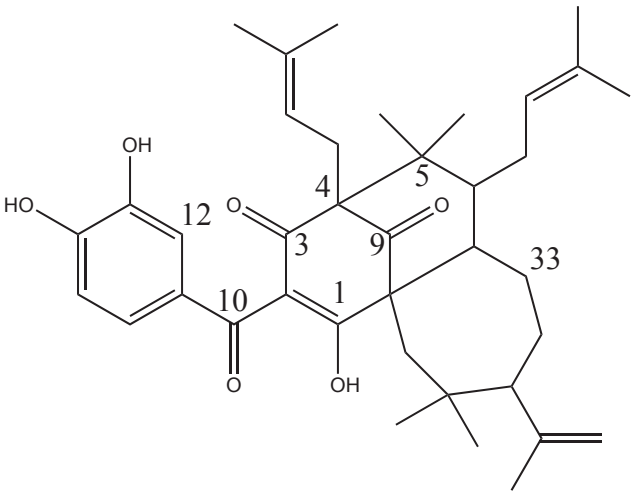


Figure 3.12 – Structure de la guttiferone H

Un premier essai d'attribution sans utiliser la prédiction des déplacements chimiques ^{13}C donne 48 solutions. Ce résultat est obtenu sans autoriser de corrélations à très longue distance. Ce succès signifie que les données des spectres RMN 2D sont en accord avec la structure proposée et qu'il n'existe pas de corrélations à très longue distance ou que les auteurs de l'article ont éliminé ces corrélations des données présentées.

Un second essai a été réalisé en utilisant des propriétés d'atome déduites des valeurs de déplacement chimique ^{13}C et de la multiplicité des signaux ^{1}H. Toutes les résonances des groupements méthyles singulets doivent être attribuées à des carbones de groupements méthyles liés à des carbones quaternaires. Tous les signaux RMN ^{13}C entre 145 et 220 ppm doivent être attribués à des atomes de carbone liés à des oxygènes et plus précisément à des oxygènes hybridés sp^2 pour les deux signaux ^{13}C les plus déblindés S9 (δ 198,0) et S10

(δ 212,7). Le carbone hybridé sp^2 produisant le signal S12 (δ 115,5) est lié à un proton dont le signal à 7,36 ppm apparait sous la forme d'un doublet fin avec une constante de couplage J = 1,8 Hz. Ce carbone est donc entouré par deux carbones quaternaires. Avec ces contraintes supplémentaires, le nombre de solutions diminue de 48 à 2 solutions. La seule ambigüité restante concerne les signaux S4 et S5, dont les attributions peuvent être échangées.

Pour effectuer une attribution en utilisant une prédiction des déplacements chimiques, un seuil d'incertitude de 10,0 ppm a été utilisé (aucun facteur multiplicatif n'a été appliqué). Cet essai produit une unique solution correspondant à l'attribution proposée dans l'article. Le fichier d'entrée de CASA correspondant est présenté en annexe F. La nécessité d'un seuil d'incertitude minimum aussi élevé est due au signal S33 mesuré à 32,9 ppm et prédit à 24,55 ppm. Le rejet de structures dans lesquelles seulement un déplacement chimique n'est pas prédit avec une assez bonne précision, possiblement parce que l'environnement moléculaire correspondant n'est pas bien représenté dans la base de données, est une source potentielle de faux négatifs.

3.4.2 Génération de structures alternatives par LSD

Le composé utilisé pour valider le fonctionnement de CASA est un triterpène (figure 3.13) de formule moléculaire $C_{30}H_{46}O_4$ qui a été isolé de *Paragonia pyramidata* (Bignoniaceae) [190]. Un premier essai d'attribution spectrale a été réalisé sans le recours à une prédiction des déplacements chimiques ^{13}C. Les listes d'atomes candidats ont été construites seulement sur la base de l'hybridation et de la multiplicité des atomes.

Les signaux RMN ^{13}C S2 et S30 sont mesurés à δ 26,6 et 26,7. Cette faible différence de déplacement chimique nous a incité à les considérer comme un groupe. Cela signifie que chaque corrélation HMBC du signal S2 a été reportée dans le fichier de données de CASA comme une corrélation de S2 ou S30. Le même traitement a été appliqué aux corrélations HMBC du signal S30. Deux autres groupes de résonances ^{13}C proches ont également été considérés : S1/S4/S22 et S9/S17.

Uniquement les corrélations des spectres HSQC, HMBC et COSY ont été utilisées comme contraintes et la prédiction des déplacements chimiques ^{13}C n'a pas été utilisée. Le premier essai d'attribution n'a donné aucune solution. Cet échec peut être interprété de deux façons. D'une part, il est possible que la structure publiée soit fausse. D'autre part, si on considère que la structure publiée est correcte, il s'agit de la méthode de vérification qui doit être jugée comme mauvaise. L'échec à produire une attribution constituerait alors un cas de faux négatif.

Un second essai d'attribution a été réalisé en donnant plus de souplesse aux contraintes. L'autorisation de 4 violations parmi les corrélations est nécessaire à l'obtention d'une solution. Une violation consiste à permettre l'interprétation d'une corrélation HMBC ^4J ou COSY ^5J. Ces deux cas sont équivalents à une distance autorisée de 3 liaisons entre les

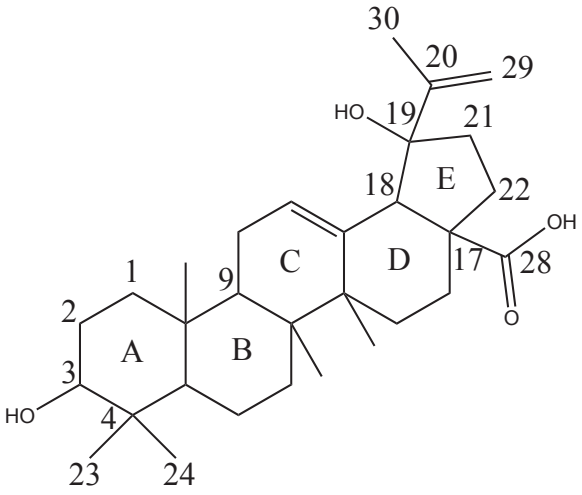


Figure 3.13 – Structure de l'article soumise à l'attribution

atomes de carbone. Une inspection détaillée des données révèle la présence de 3 corrélations HMBC et 1 corrélation COSY à très longue distance dans ou autour du cycle E de la structure publiée (figure 3.14).

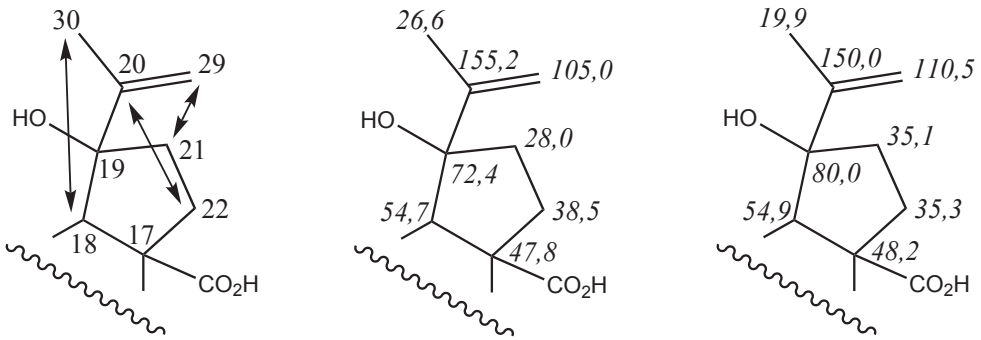


Figure 3.14 – Cycle E de la structure publiée

La comparaison entre les déplacements chimiques prédits et expérimentaux montre une grande différence pour C19 (7,6 ppm), C20 (5,2 ppm), C29 (5,5 ppm) et C30 (6,7 ppm), tous dans ou autour du cycle E. En faisant agir CASA avec les valeurs de déplacement chimique ^{13}C comme critères de d'attribution, une solution est produite en plaçant l'erreur minimum sur la prédiction à 8 ppm. Cette valeur correspond à la valeur minimum par défaut multipliée par un facteur de sécurité de 1,6. Le fichier d'entrée de CASA correspondant est présenté en annexe F. En supposant que la structure publiée est fausse

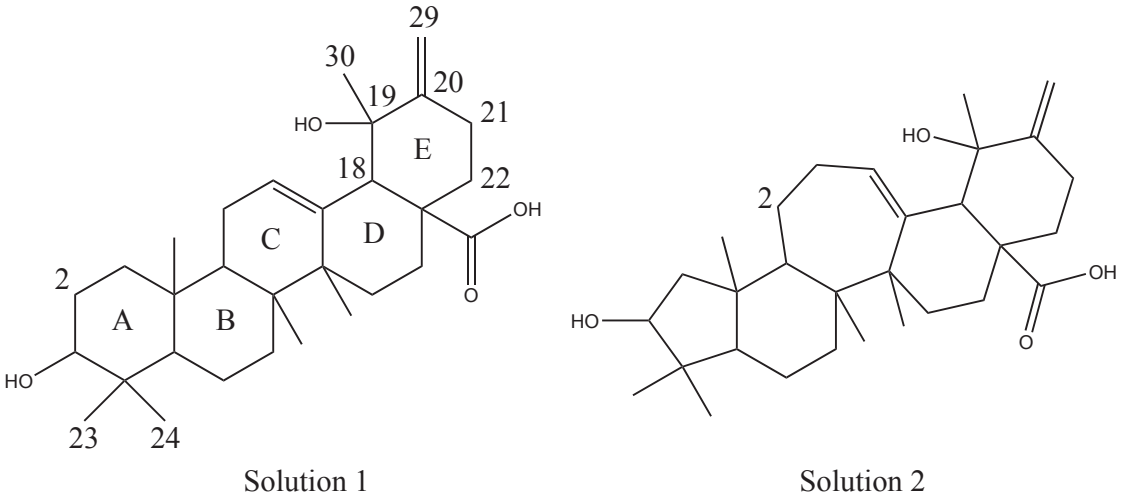
et qu'elle constitue alors un faux positif, nous avons décidé de chercher des structures alternatives en utilisant le logiciel LSD. Cette approche de la validation de structure a déjà été présentée dans le premier article qui concernait LSD [86].

La détermination de structure assistée par LSD a été entreprise avec les mêmes données de RMN 2D que celles utilisées par CASA. Les atomes de carbone des groupements méthyles qui apparaissent sous la forme d'un singulet sur le spectre RMN ^{1}H sont contraints à être liés à des carbones quaternaires. Les carbones C3 et C19 à δ 78,6 et 83,47 ont été forcés à être liés à un atome d'oxygène et C28 à δ 180,0 a été forcé à être lié à deux atomes d'oxygène. Jusqu'à 4 corrélations à très longue distance ont été autorisées et toutes les corrélations HMBC ont été autorisées à être considérées comme des corrélations à très longue distance. La longueur du chemin de couplage des corrélations COSY H1/H3 et H21/H29 a été étendue à 4 et 5 liaisons respectivement, alors que les autres corrélations COSY ont été considérées comme issues de couplages ^3J.

LSD a trouvé 74 solutions qui ont été classées par ordre de vraisemblance croissante entre les déplacements chimiques expérimentaux et les déplacements chimiques prédits par nmrshiftdb2. La variable Δ permettant d'estimer la vraisemblance a été calculée pour une série de N signaux par la somme des valeurs absolues des différences entre les valeurs expérimentales et prédites.

$$\Delta = \sum_{i=1}^{N} |\delta_i^{\text{pred}} - \delta_i^{\text{exp}}|$$

Les solutions de LSD classées aux deux premières places sont représentées sur la figure 3.15.



Solution 1 Solution 2

Figure 3.15 – Structures alternatives fournies par LSD

La troisième solution correspond à la structure publiée dans l'article. Les valeurs de Δ de ces solutions sont respectivement égale à 29,35, 54,0 et 59,3 ppm. La structure 1 peut être considérée comme une meilleure solution que la structure publiée pour deux

raisons (figure 3.16). Premièrement, Δ est environ deux fois plus petit pour la structure 1. Les plus fortes différences entre la prédiction et l'expérience ont été trouvées pour C23 et C24, avec une valeur absolue des différences égale à 5,7 ppm et 6,7 ppm. Les prédictions de nmrshiftdb2 ne différencient pas les groupements méthyles géminés axiaux et équatoriaux. La prédiction produit le même déplacement chimique de 21,9 ppm pour les deux. Cette valeur est comparée avec les valeurs expérimentales δ 27,6 et 15,2. Les valeurs de déplacement chimique minimum et maximum pour de tels groupements méthyles dans la base de données de nmrshiftdb2 sont 15,3 et 28,2 ppm. Dans ce cas, il y a donc une grande variabilité dans les déplacements chimiques des groupements méthyles équivalents topologiquement. Cette variabilité est due à la présence d'un groupement méthyle axial et d'un groupement méthyle équatorial. Le second argument en faveur de la structure 1 est son nombre réduit de corrélations à très longue distance. C'est un argument qui repose sur l'interprétation des données des spectres 2D par l'utilisateur et non sur des bases de données externes. La seule corrélation à très longue distance dans la structure 1 est la corrélation C29/H30, une corrélation ^{4}J avec une double liaison dans le chemin de couplage. Les corrélations HMBC ^4J C29/H21, C20/H22 and C18/H30 dans la structure publiée sont des ^3J dans la structure 1. La corrélation COSY ^5J H29/H21 dans la structure publiée est une corrélation ^4J dans la structure 1, avec une double liaison dans le chemin de couplage.

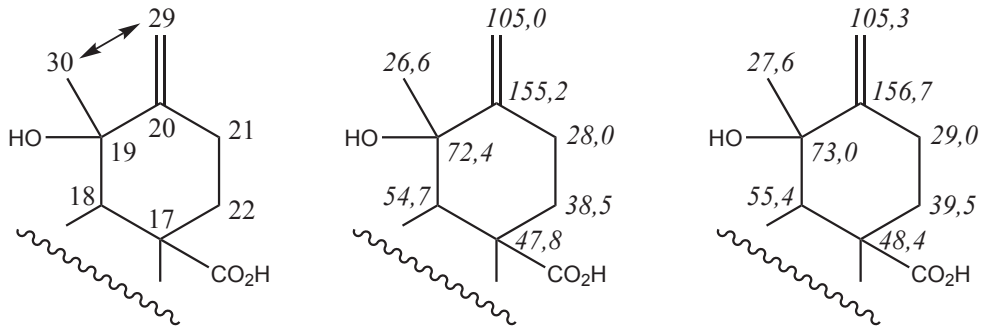


Figure 3.16 – Cycle E de la structure 1 proposée par LSD

Le structure 2 de la figure 3.15 a les mêmes cycles D et E que la structure 1, mais C2 n'est pas placé par LSD dans le cycle A mais dans le cycle C. Le squelette résultant est plutôt inhabituel et n'est pas vraiment pertinent dans le domaine de la chimie des triterpènes.

La structure publiée appartient à la famille des triterpènes de type lupane. Cependant, la structure 1, de la famille des ursanes, est préférée d'après les données de RMN 1D et

2D.

3.5 Conclusion

Ces travaux avaient pour but l'écriture d'un programme d'attribution des signaux RMN ^{13}C afin de permettre la vérification de structures. L'écriture du programme CASA a été menée à bien et a abouti à une version opérationnelle. La prédiction des déplacements chimiques ^{13}C est assurée par le module nmrshiftdb2. Un mode d'emploi a été rédigé pour permettre une diffusion du programme dans la communauté scientifique.

L'efficacité du programme a pu être prouvée en utilisant des exemples de structures connues. En cas d'échec de la vérification ou de doute sur la validité d'une structure, le logiciel LSD peut être utilisé pour proposer des structures alternatives. L'ensemble des faux positifs de CASA pourrait être généré avec LSD si les solutions produites par ce dernier pouvaient être retenues ou éliminées en utilisant le même critère sur l'incertitude des déplacements chimiques prédits.

Les perspectives d'amélioration et d'évolution de CASA sont multiples. Afin de réduire l'intervention de l'utilisateur au niveau de la souplesse accordée aux contraintes, il serait possible de laisser plus de liberté sur les corrélations lorsqu'il y a une double ou une triple liaison dans le chemin de couplage. CASA pourrait être intégré dans une interface graphique permettant une extraction automatique des données des spectres et une présentation visuelle des résultats. L'exercice d'attribution pourrait être étendu à des structures 3D. Dans ce cas, il serait nécessaire d'utiliser des contraintes géométriques supplémentaires provenant des spectres NOESY et des valeurs de constante de couplage.

Conclusion et perspectives

Les travaux présentés dans ce manuscrit portent sur plusieurs aspects de l'analyse structurale des petites molécules organiques. La finalité générale est de fournir une aide aux chimistes par l'automatisation, d'une part du processus d'élucidation structurale et d'autre part de la vérification de structure et de l'attribution des signaux RMN. Tous ces points ont été abordés à travers l'amélioration du logiciel LSD et l'écriture du programme CASA.

Le premier objectif de cette thèse était de supprimer certaines rigidités imposées dès la conception du logiciel de détermination de structure LSD et de faciliter son utilisation. Les améliorations ont été réalisées en apportant des modifications au code source du logiciel existant.

Le cœur de LSD, c'est-à-dire le générateur de structure, a subi plusieurs aménagements. L'utilisation des données des spectres RMN 2D a été affinée avec un traitement amélioré des corrélations. LSD accepte désormais une plus grande diversité dans les molécules analysées avec des nouveautés dans la définition du statut des atomes. L'utilisateur a maintenant la possibilité d'introduire des atomes hybridés *sp*, des atomes chargés et des atomes à valence multiple.

La présentation des solutions et plus précisément la qualité visuelle des dessins des molécules ont fait l'objet de nettes améliorations dans le but de faciliter l'interprétation des résultats. Différentes librairies permettant la génération de dessins de structures en 2D ont été employées et évaluées. Les tests ont débouché sur l'écriture d'une version alternative du programme d'exportation des résultats.

La complémentarité de LSD avec la base de données de produits naturels SISTEMAT a été démontrée depuis quelques années. Les dernières avancées destinées à faciliter l'usage combiné des deux systèmes ont abouti à la création d'une banque de squelettes de produits naturels. Des contraintes de sous-structure prédites par SISTEMAT en utilisant l'information des déplacements chimiques ou l'origine botanique de la substance analysée peuvent être directement choisies dans cette banque de squelettes.

Les modifications récentes de LSD concernant le statut des atomes ont permis l'écriture du programme PyLSD permettant la gestion d'atomes à statut indéfini. D'autres points comme la gestion de la symétrie dans les structures ou la définition automatique du statut des atomes à partir des valeurs de déplacements chimiques restent à mettre en place. Une structure n'étant pas totalement élucidée sans la connaissance de sa géométrie, il faudrait faire suivre le générateur de structures planes par une détermination de structure 3D.

Le second objectif de cette thèse était de mettre au point un outil de vérification de structure par une méthode originale utilisant les données des spectres de RMN 2D. Le logiciel CASA a été développé dans ce sens. Il réalise une vérification par l'attribution automatique des résonances en utilisant notamment un outil de prédiction des déplacements chimiques ^{13}C reposant sur la base de données nmrshiftdb2. Les tests de validation du programme nous ont amené à utiliser des structures publiées dans la littérature. Après avoir émis des doutes sur la validité d'une structure prise comme exemple, nous avons envisagé l'existence de structures alternatives à l'aide de LSD. Ceci nous a permis de proposer une structure jugée plus compatible avec les données spectroscopiques à disposition.

En ce qui concerne les évolutions futures de CASA, on pourrait entre autres imaginer une intégration dans une interface graphique pour faciliter son utilisation. On pourrait également compléter la vérification par l'utilisation de données permettant d'avoir accès à des informations sur la géométrie des molécules. Le but ultime étant, comme pour LSD, la prise en compte de la nature tridimensionnelle des molécules.

L'ensemble des travaux de cette thèse s'inscrit dans le contexte global de l'analyse structurale des molécules organiques. Il s'agit d'un domaine en perpétuel progrès, avec notamment la mise au point de techniques expérimentales de plus en plus fines et précises qui amènent parfois à remettre en cause des structures déjà étudiées. En effet, une structure n'est qu'une proposition faite à partir de données spectroscopiques disponibles à un instant donné. De leur côté, les outils informatiques d'analyse servent à aider les chimistes en leur offrant un gain de temps et en leur procurant un niveau de certitude supplémentaire. On peut présager que le développement des logiciels d'interprétation automatique des données va donner un nouveau rôle aux instruments d'analyse. En effet, ils fourniront non plus des spectres mais directement ce que les chimistes recherchent, c'est-à-dire des structures...

Les travaux présentés dans ce manuscrit font l'objet de plusieurs articles :

- B. Plainchont, J.-M. Nuzillard, G. V. Rodrigues, M. J. P. Ferreira, M. T. Scotti, V. P. Emerenciano, **New Improvements in Automatic Structure Elucidation Using the LSD (Logic for Structure Determination) and the SISTEMAT Expert Systems**, Natural Product Communications 2010, 5, 763-770.
- B. Plainchont, J.-M. Nuzillard, **Structure verification through computer-assisted spectral assignment of NMR spectra**, Magnetic Resonance in Chemistry 2013, 51, 54-59.
- B. Plainchont, V. P. Emerenciano, J.-M. Nuzillard, **Recent advances in the structure elucidation of small organic molecules by the LSD software**, Phytochemical Analysis, article soumis.

Bibliographie

[1] K. C. Nicolaou and S. A. Snyder. Chasing molecules that were never there : Misassigned natural products and the role of chemical synthesis in modern structure elucidation. *Angewandte Chemie International Edition*, 44(7) :1012–1044, 2005.

[2] M. E. Elyashberg, A. J. Williams, and K. A. Blinov. Structural revisions of natural products by computer-assisted structure elucidation (CASE) systems. *Natural Product Reports*, 27(9) :1296–1328, 2010.

[3] M. E. Munk. Computer-based structure determination : Then and now. *Journal of Chemical Information and Computer Sciences*, 38(6) :997–1009, 1998.

[4] M. Jaspars. Computer assisted structure elucidation of natural products using two-dimensional NMR spectroscopy. *Natural Product Reports*, 16(2) :241–248, 1999.

[5] C. Steinbeck. Recent developments in automated structure elucidation of natural products. *Natural Product Reports*, 21(4) :512–518, 2004.

[6] W. L. Chen. Chemoinformatics : Past, present, and future. *Journal of Chemical Information and Modeling*, 46(6) :2230–2255, 2006.

[7] R. Stefani, P. G. B. D. Nascimento, and F. B. D. Costa. Elucidação estrutural de substâncias orgânicas com auxílio de computador : evoluções recentes. *Quimica Nova*, 30(5) :1347–1356, 2007.

[8] M. E. Elyashberg, A. J. Williams, and G. E. Martin. Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 53(1-2) :1–104, 2008.

[9] D. M. Doddrell, D. T. Pegg, and M. R. Bendall. Distortionless enhancement of NMR signals by polarization transfer. *Journal of Magnetic Resonance (1969)*, 48 (2) :323–327, 1982.

[10] E. G. Paul and D. M. Grant. Additivity relationships in carbon-13 chemical shift data for the linear alkanes. *Journal of the American Chemical Society*, 85(11) : 1701–1702, 1963.

[11] D. M. Grant and E. G. Paul. Carbon-13 magnetic resonance. II. chemical shift data for the alkanes. *Journal of the American Chemical Society*, 86(15) :2984–2990, 1964.

[12] J. T. Clerc and H. Sommerauer. A minicomputer program based on additivity rules for the estimation of ^{13}C NMR chemical shifts. *Analytica Chimica Acta*, 95(1) : 33–40, 1977.

[13] A. Fürst and E. Pretsch. A computer program for the prediction of ^{13}C NMR chemical shifts of organic compounds. *Analytica Chimica Acta*, 229(1) :17–25, 1990.

[14] A. Fürst, E. Pretsch, and W. Robien. Comprehensive parameter set for the prediction of the ^{13}C NMR chemical shifts of sp^3-hybridized carbon atoms in organic compounds. *Analytica Chimica Acta*, 233(1) :213–222, 1990.

[15] E. Pretsch, A. Fürst, and W. Robien. Parameter set for the prediction of the ^{13}C NMR chemical shifts of sp^2- and sp-hybridized carbon atoms in organic compounds. *Analytica Chimica Acta*, 248(2) :415–428, 1991.

[16] E. Pretsch, A. Fürst, M. Badertscher, R. Buergin, and M. E. Munk. C13Shift : a computer program for the prediction of carbon-13 NMR spectra based on an open set of additivity rules. *Journal of Chemical Information and Computer Sciences*, 32 (4) :291–295, 1992.

[17] L. Chen and W. Robien. OPSI : a universal method for prediction of carbon-13 NMR spectra based on optimized additivity models. *Analytical Chemistry*, 65(17) : 2282–2287, 1993.

[18] W. Bremser. HOSE - a novel substructure code. *Analytica Chimica Acta*, 103(4) : 355–365, 1978.

[19] W. Bremser. Expectation ranges of ^{13}C NMR chemical shifts. *Magnetic Resonance in Chemistry*, 23(4) :271–275, 1985.

[20] V. Schütz, V. Purtuc, S. Felsinger, and W. Robien. CSEARCH-STEREO : A new generation of NMR database systems allowing three-dimensional spectrum prediction. *Fresenius' Journal of Analytical Chemistry*, 359(1) :33–41, 1997.

[21] N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith, and C. Djerassi. Stereochemical substructure codes for ^{13}C spectral analysis. *Organic Magnetic Resonance*, 15(4) :375–389, 1981.

[22] J.-P. Gastmans, J. C. Zurita, J. Sahao Jr, and V. de P. Emerenciano. Prévision des spectres de résonance magnétique nucléaire de ^{13}C par intelligence artificielle : le problème de la codification. *Analytica Chimica Acta*, 217(1) :85–100, 1989.

[23] G. W. Small. Database retrieval techniques for carbon-13 nuclear magnetic resonance spectrum simulation. *Journal of Chemical Information and Computer Sciences*, 32(4) :279–285, 1992.

[24] V. Kvasnicka. An application of neural networks in chemistry. prediction of ^{13}C NMR chemical shifts. *Journal of Mathematical Chemistry*, 6(1) :63–76, 1991.

[25] L. S. Anker and P. C. Jurs. Prediction of carbon-13 nuclear magnetic resonance chemical shifts by artificial neural networks. *Analytical Chemistry*, 64(10) :1157–1164, 1992.

[26] J. P. Doucet, A. Panaye, E. Feuilleaubois, and P. Ladd. Neural networks and carbon-13 NMR shift prediction. *Journal of Chemical Information and Computer Sciences*, 33(3) :320–324, 1993.

[27] V. Kvasnicka, S. Sklenak, and J. Pospichal. Application of recurrent neural networks in chemistry. prediction and classification of carbon-13 NMR chemical shifts in a series of monosubstituted benzenes. *Journal of Chemical Information and Computer Sciences*, 32(6) :742–747, 1992.

[28] A. Panaye, J. P. Doucet, B. T. Fan, E. Feuilleaubois, and S. R. El Azzouzi. Artificial neural network simulation of ^{13}C NMR shifts for methyl substituted cyclohexanes. *Chemometrics and Intelligent Laboratory Systems*, 24(2) :129–135, 1994.

[29] Y. Miyashita, H. Yoshida, O. Yaegashi, T. Kimura, H. Nishiyama, and S. Sasaki. Non-linear modelling of ^{13}C NMR chemical shift data using artificial neural networks and partial least squares method. *Journal of Molecular Structure : THEOCHEM*, 311(1) :241–245, 1994.

[30] D. L. Clouser and P. C. Jurs. Simulation of ^{13}C nuclear magnetic resonance spectra of tetrahydropyrans using regression analysis and neural networks. *Analytica Chimica Acta*, 295(3) :221–231, 1994.

[31] B. E. Mitchell and P. C. Jurs. Computer assisted simulation of ^{13}C nuclear magnetic spectra of monosaccharides. *Journal of Chemical Information and Computer Sciences*, 36(1) :58–64, 1996.

[32] D. L. Clouser and P. C. Jurs. Simulation of the ^{13}C nuclear magnetic resonance spectra of ribonucleosides using multiple linear regression analysis and neural networks. *Journal of Chemical Information and Computer Sciences*, 36(2) :168–172, 1996.

[33] O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye, and J. P. Doucet. ^{13}C NMR chemical shift prediction of sp^2 carbon atoms in acyclic alkenes using neural networks. *Journal of Chemical Information and Computer Sciences*, 36(4) :644–653, 1996.

[34] O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye, and J. P. Doucet. ^{13}C NMR chemical shift prediction of the sp^3 carbon atoms in the α position relative to the

double bond in acyclic alkenes. *Journal of Chemical Information and Computer Sciences*, 37(3) :587–598, 1997.

[35] J. Meiler, R. Meusinger, and M. Will. Fast determination of ^{13}C NMR chemical shifts using artificial neural networks. *Journal of Chemical Information and Computer Sciences*, 40(5) :1169–1176, 2000.

[36] J. Meiler, W. Maier, M. Will, and R. Meusinger. Using neural networks for ^{13}C NMR chemical shift prediction-comparison with traditional methods. *Journal of Magnetic Resonance*, 157(2) :242–252, 2002.

[37] K. A. Blinov, Y. D. Smurnyy, M. E. Elyashberg, T. S. Churanova, M. Kvasha, C. Steinbeck, B. A. Lefebvre, and A. J. Williams. Performance validation of neural network based ^{13}C NMR prediction using a publicly available data source. *Journal of Chemical Information and Modeling*, 48(3) :550–555, 2008.

[38] Y. D. Smurnyy, K. A. Blinov, T. S. Churanova, M. E. Elyashberg, and A. J. Williams. Toward more reliable ^{13}C and ^{1}H chemical shift prediction : A systematic comparison of neural-network and least-squares regression based approaches. *Journal of Chemical Information and Modeling*, 48(1) :128–134, 2008.

[39] K. A. Blinov, Y. D. Smurnyy, T. S. Churanova, M. E. Elyashberg, and A. J. Williams. Development of a fast and accurate method of ^{13}C NMR chemical shift prediction. *Chemometrics and Intelligent Laboratory Systems*, 97(1) :91–97, 2009.

[40] S. D. Rychnovsky. Predicting NMR spectra by computational methods : Structure revision of Hexacyclinol. *Organic Letters*, 8(13) :2895–2898, 2006.

[41] A. Bagno and G. Saielli. Computational NMR spectroscopy : reversing the information flow. *Theoretical Chemistry Accounts : Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 117(5) :603–619, 2007.

[42] J. J. R. Reed. Structure determination of substituted benzenes by proton magnetic resonance. empirical treatment of substituent effects and their utility in predicting chemical shifts. *Analytical Chemistry*, 39(13) :1586–1593, 1967.

[43] R. B. Schaller and E. Pretsch. A computer program for the automatic estimation of ^{1}H NMR chemical shifts. *Analytica Chimica Acta*, 290(3) :295–302, 1994.

[44] R. B. Schaller, C. Arnold, and E. Pretsch. New parameters for predicting ^{1}H NMR chemical shifts of protons attached to carbon atoms. *Analytica Chimica Acta*, 312 (1) :95–105, 1995.

[45] J. Aires-de Sousa, M. C. Hemmer, and J. Gasteiger. Prediction of ^{1}H NMR chemical shifts using neural networks. *Analytical Chemistry*, 74(1) :80–90, 2002.

[46] Y. Binev and J. Aires-de Sousa. Structure-based predictions of ^{1}H NMR chemical shifts using feed-forward neural networks. *Journal of Chemical Information and Computer Sciences*, 44(3) :940–945, 2004.

[47] Y. Binev, M. Corvo, and J. Aires-de Sousa. The impact of available experimental data on the prediction of ^{1}H NMR chemical shifts by neural networks. *Journal of Chemical Information and Computer Sciences*, 44(3) :946–949, 2004.

[48] A. Bagno. Complete prediction of the ^{1}H NMR spectrum of organic molecules by DFT calculations of chemical shifts and spin-spin coupling constants. *Chemistry – A European Journal*, 7(8) :1652–1661, 2001.

[49] W. P. Aue, E. Bartholdi, and R. R. Ernst. Two-dimensional spectroscopy. application to nuclear magnetic resonance. *The Journal of Chemical Physics*, 64(5) : 2229–2246, 1976.

[50] A. Bax, R. Freeman, and T. A. Frenkiel. An NMR technique for tracing out the carbon skeleton of an organic molecule. *Journal of the American Chemical Society*, 103(8) :2102–2104, 1981.

[51] G. Bodenhausen and D. J. Ruben. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1) :185–189, 1980.

[52] A. Bax and M. F. Summers. Proton and carbon-13 assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. *Journal of the American Chemical Society*, 108(8) :2093–2094, 1986.

[53] M. Köck, J. Junker, and T. Lindel. Impact of the ^{1}H,^{15}N-HMBC experiment on the constitutional analysis of alkaloids. *Organic Letters*, 1(13) :2041–2044, 1999.

[54] B. Reif, M. Köck, R. Kerssebaum, H. Kang, W. Fenical, and C. Griesinger. ADEQUATE, a new set of experiments to determine the constitution of small molecules at natural abundance. *Journal of Magnetic Resonance, Series A*, 118(2) :282–285, 1996.

[55] N. T. Nyberg, J. Ø. Duus, and O. W. Sørensen. Heteronuclear two-bond correlation : Suppressing heteronuclear three-bond or higher NMR correlations while enhancing two-bond correlations even for vanishing $^2J_{CH}$. *Journal of the American Chemical Society*, 127(17) :6154–6155, 2005.

[56] J. Jeener, B. H. Meier, P. Bachmann, and R. R. Ernst. Investigation of exchange processes by two-dimensional NMR spectroscopy. *The Journal of Chemical Physics*, 71(11) :4546–4553, 1979.

[57] A. A. Bothner-By, R. L. Stephens, J. Lee, C. D. Warren, and R. W. Jeanloz. Structure determination of a tetrasaccharide : transient nuclear Overhauser effects in the rotating frame. *Journal of the American Chemical Society*, 106(3) :811–813, 1984.

[58] M. E. Elyashberg, Y. Z. Karasev, E. R. Martirosian, H. Thiele, and H. Somberg. Expert systems as a tool for the molecular structure elucidation by spectral methods. strategies of solution to the problems. *Analytica Chimica Acta*, 348(1-3) :443–463, 1997.

[59] M. E. Elyashberg, E. R. Martirosian, Y. Z. Karasev, H. Thiele, and H. Somberg. X-PERT : a user-friendly expert system for molecular structure elucidation by spectral methods. *Analytica Chimica Acta*, 337(3) :265–286, 1997.

[60] M. E. Elyashberg, Y. Z. Karasev, and E. R. Martirosian. Spectroscopic determination of elemental composition of organic compounds with the aid of the X-PERT system. *Analytica Chimica Acta*, 388(3) :353–363, 1999.

[61] M. Will, W. Fachinger, and J. R. Richert. Fully automated structure elucidation - a spectroscopist's dream comes true. *Journal of Chemical Information and Computer Sciences*, 36(2) :221–227, 1996.

[62] W. Bremser and M. Grzonka. Specinfo–a multidimensional spectroscopic interpretation system. *Microchimica Acta*, 104(1) :483–491, 1991.

[63] M. E. Elyashberg, K. A. Blinov, and E. R. Martirosian. A new approach to computer-aided molecular structure elucidation : the expert system *Structure Elucidator*. *Laboratory Automation and Information Management*, 34(1) :15–30, 1999.

[64] J. Meiler and M. Will. Automated structure elucidation of organic molecules from ^{13}C NMR spectra using genetic algorithms and neural networks. *Journal of Chemical Information and Computer Sciences*, 41(6) :1535–1546, 2001.

[65] J. Meiler and M. Will. Genius : A genetic algorithm for automated structure elucidation from ^{13}C NMR spectra. *Journal of the American Chemical Society*, 124(9) : 1868–1870, 2002.

[66] B. D. Christie and M. E. Munk. The role of two-dimensional nuclear magnetic resonance spectroscopy in computer-enhanced structure elucidation. *Journal of the American Chemical Society*, 113(10) :3750–3757, 1991.

[67] R. B. Schaller, M. E. Munk, and E. Pretsch. Spectra estimation for computer-aided structure determination. *Journal of Chemical Information and Computer Sciences*, 36(2) :239–243, 1996.

[68] M. Badertscher, A. Korytko, K.-P. Schulz, M. S. Madison, M. E. Munk, P. Portmann, M. Junghans, P. Fontana, and E. Pretsch. Assemble 2.0 : a structure generator. *Chemometrics and Intelligent Laboratory Systems*, 51(1) :73–79, 2000.

[69] P. N. Penchev, K.-P. Schulz, and M. E. Munk. INFERCNMR : A ^{13}C NMR interpretive library search system. *Journal of Chemical Information and Modeling*, 52 (6) :1513–1528, 2012.

[70] C. A. Shelley and M. E. Munk. Case, a computer model of the structure elucidation process. *Analytica Chimica Acta*, 133(4) :507–516, 1981.

[71] M. E. Munk, C. A. Shelley, H. B. Woodruff, and M. O. Trulson. Computer-assisted structure elucidation. *Fresenius' Journal of Analytical Chemistry*, 313(6) :473–479, 1982.

[72] B. D. Christie and M. E. Munk. The application of two-dimensional nuclear magnetic resonance spectroscopy in computer-assisted structure elucidation. *Analytica Chimica Acta*, 200(1) :347–361, 1987.

[73] M. E. Munk and B. D. Christie. The characterization of structure by computer. *Analytica Chimica Acta*, 216(1) :57–68, 1989.

[74] B. D. Christie and M. E. Munk. Structure generation by reduction : a new strategy for computer-assisted structure elucidation. *Journal of Chemical Information and Computer Sciences*, 28(2) :87–93, 1988.

[75] A. Korytko, K. P. Schulz, M. S. Madison, and M. E. Munk. HOUDINI : A new approach to computer-based structure generation. *Journal of Chemical Information and Computer Sciences*, 43(5) :1434–1446, 2003.

[76] K. P. Schulz, A. Korytko, and M. E. Munk. Applications of a HOUDINI-based structure elucidation system. *Journal of Chemical Information and Computer Sciences*, 43(5) :1447–1456, 2003.

[77] C. Peng, S. Yuan, C. Zheng, and Y. Hui. Efficient application of 2D NMR correlation information in computer-assisted structure elucidation of complex natural products. *Journal of Chemical Information and Computer Sciences*, 34(4) :805–813, 1994.

[78] C. Peng, S. Yuan, C. Zheng, Y. Hui, H. Wu, K. Ma, and X. Han. Application of expert system CISOC-SES to the structure elucidation of complex natural products. *Journal of Chemical Information and Computer Sciences*, 34(4) :814–819, 1994.

[79] C. Peng, S. Yuan, C. Zheng, Z. Shi, and H. Wu. Practical computer-assisted structure elucidation for complex natural products : Efficient use of ambiguous 2D NMR correlation information. *Journal of Chemical Information and Computer Sciences*, 35(3) :539–546, 1995.

[80] C. Peng, G. Bodenhausen, S. Qiu, H. H. S. Fong, N. R. Farnsworth, S. Yuan, and C. Zheng. Computer-assisted structure elucidation : application of CISOC-SES to the resonance assignment and structure generation of betulinic acid. *Magnetic Resonance in Chemistry*, 36(4) :267–278, 1998.

[81] T. Lindel, J. Junker, and M. Köck. Cocon : From NMR correlation data to molecular constitutions. *Journal of Molecular Modeling*, 3(8) :364–368, 1997.

[82] J. Junker, W. Maier, T. Lindel, and M. Köck. Computer-assisted constitutional assignment of large molecules : Cocon analysis of ascomycin. *Organic Letters*, 1(5) : 737–740, 1999.

[83] M. Köck, J. Junker, W. Maier, M. Will, and T. Lindel. A COCON analysis of proton-poor heterocycles – application of carbon chemical shift predictions for the evaluation of structural proposals. *European Journal of Organic Chemistry*, 1999 (3) :579–586, 1999.

[84] T. Lindel, J. Junker, and M. Köck. 2D-NMR-guided constitutional analysis of organic compounds employing the computer program COCON. *European Journal of Organic Chemistry*, 1999(3) :573–577, 1999.

[85] J. Meiler, E. Sanli, J. Junker, R. Meusinger, T. Lindel, M. Will, W. Maier, and M. Köck. Validation of structural proposals by substructure analysis and ^{13}C NMR chemical shift prediction. *Journal of Chemical Information and Computer Sciences*, 42(2) :241–248, 2002.

[86] J.-M. Nuzillard and G. Massiot. Logic for structure determination. *Tetrahedron*, 47 (22) :3655–3664, 1991.

[87] J.-M. Nuzillard, W. Naanaa, and S. Pimont. Applying the constraint satisfaction problem paradigm to structure generation. *Journal of Chemical Information and Computer Sciences*, 35(6) :1068–1073, 1995.

[88] J.-M. Nuzillard. Détermination assistée par ordinateur de la structure des molécules organiques. *J. Chim. Phys.*, 95(2) :169–177, 1998.

[89] G. Massiot, C. Lavaud, and J.-M. Nuzillard. Structure elucidation of plant secondary products. In *Chemical from plants. Perspectives on plant secondary products*, pages 187–214. Imperial College Press, 1999.

[90] J.-M. Nuzillard. Automatic structure determination of organic molecules : Principle and implementation of the LSD program. *Chinese Journal of Chemistry*, 21(1) : 1263–1267, 2003.

[91] C. Steinbeck. LUCY–a program for structure elucidation from NMR correlation experiments. *Angewandte Chemie International Edition in English*, 35(17) :1984–1986, 1996.

[92] C. Steinbeck. SENECA : A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *Journal of Chemical Information and Computer Sciences*, 41(6) :1500–1507, 2001.

[93] Y. Han and C. Steinbeck. Evolutionary-algorithm-based strategy for computer-assisted structure elucidation. *Journal of Chemical Information and Computer Sciences*, 44(2) :489–498, 2004.

[94] J.-L. Faulon. Stochastic generator of chemical structure. 1. application to the structure elucidation of large molecules. *Journal of Chemical Information and Computer Sciences*, 34(5) :1204–1218, 1994.

[95] J.-L. Faulon. Stochastic generator of chemical structure. 2. using simulated annealing to search the space of constitutional isomers. *Journal of Chemical Information and Computer Sciences*, 36(4) :731–740, 1996.

[96] K. A. Blinov, M. E. Elyashberg, S. G. Molodtsov, A. J. Williams, and E. R. Martirosian. An expert system for automated structure elucidation utilizing ^{1}H–^{1}H, ^{13}C–^{1}H and ^{15}N–^{1}H 2D NMR correlations. *Fresenius' Journal of Analytical Chemistry*, 369 (7) :709–714, 2001.

[97] M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. R. Martirosian, and S. G. Molodtsov. Application of a new expert system for the structure elucidation of natural products from their 1D and 2D NMR data. *Journal of Natural Products*, 65(5) :693–703, 2002.

[98] G. E. Martin, C. E. Hadden, D. J. Russell, B. D. Kaluzny, J. E. Guido, W. K. Duholke, B. A. Stiemsma, T. J. Thamann, R. C. Crouch, K. A. Blinov, M. E. Elyashberg, E. R. Martirosian, S. G. Molodtsov, A. J. Williams, and P. L. Schiff. Identification of degradants of a complex alkaloid using NMR cryoprobe technology and ACD/structure elucidator. *Journal of Heterocyclic Chemistry*, 39(6) :1241–1250, 2002.

[99] K. A. Blinov, M. E. Elyashberg, E. R. Martirosian, S. G. Molodtsov, A. J. Williams, A. N. Tackie, M. M. H. Sharaf, P. L. Schiff, R. C. Crouch, G. E. Martin, C. E. Hadden, J. E. Guido, and K. A. Mills. Quindolinocryptotackieine : the elucidation of a novel indoloquinoline alkaloid structure through the use of computer-assisted structure elucidation and 2D NMR. *Magnetic Resonance in Chemistry*, 41(8) : 577–584, 2003.

[100] K. A. Blinov, D. Carlson, M. E. Elyashberg, G. E. Martin, E. R. Martirosian, S. Molodtsov, and A. J. Williams. Computer-assisted structure elucidation of natural products with limited 2D NMR data : application of the struceluc system. *Magnetic Resonance in Chemistry*, 41(5) :359–372, 2003.

[101] M. E. Elyashberg, K. A. Blinov, E. R. Martirosian, S. G. Molodtsov, A. J. Williams, and G. E. Martin. Automated structure elucidation – the benefits of a symbiotic relationship between the spectroscopist and the expert system. *Journal of Heterocyclic Chemistry*, 40(6) :1017–1029, 2003.

[102] M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, G. E. Martin, and E. R. Martirosian. Structure elucidator : A versatile expert system for molecular structure elucidation from 1D and 2D NMR data and molecular fragments. *Journal of Chemical Information and Computer Sciences*, 44(3) :771–792, 2004.

[103] S. G. Molodtsov, M. E. Elyashberg, K. A. Blinov, A. J. Williams, E. E. Martirosian, G. E. Martin, and B. Lefebvre. Structure elucidation from 2D NMR spectra using the *StrucEluc* expert system : Detection and removal of contradictions in the data. *Journal of Chemical Information and Computer Sciences*, 44(5) :1737–1751, 2004.

[104] G. J. Sharman, I. C. Jones, M. P. Parnell, M. C. Willis, M. F. Mahon, D. V. Carlson, A. J. Williams, M. E. Elyashberg, K. A. Blinov, and S. G. Molodtsov. Automated structure elucidation of two unexpected products in a reaction of an α,β-unsaturated pyruvate. *Magnetic Resonance in Chemistry*, 42(7) :567–572, 2004.

[105] Y. D. Smurnyy, M. E. Elyashberg, K. A. Blinov, B. A. Lefebvre, G. E. Martin, and A. J. Williams. Computer-aided determination of relative stereochemistry and 3D models of complex organic molecules from 2D NMR spectra. *Tetrahedron*, 61(42) : 9980–9989, 2005.

[106] M. E. Elyashberg, K. A. Blinov, A. J. Williams, S. G. Molodtsov, and G. E. Martin. Are deterministic expert systems for computer-assisted structure elucidation obsolete ? *Journal of Chemical Information and Modeling*, 46(4) :1643–1656, 2006.

[107] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, A. J. Williams, and G. E. Martin. Fuzzy structure generation : A new efficient tool for computer-aided structure elucidation (CASE). *Journal of Chemical Information and Modeling*, 47(3) :1053–1066, 2007.

[108] A. J. Williams, M. E. Elyashberg, K. A. Blinov, D. C. Lankin, G. E. Martin, W. F. Reynolds, J. A. Porco, C. A. Singleton, and S. Su. Applying computer-assisted structure elucidation algorithms for the purpose of structure validation : Revisiting the NMR assignments of Hexacyclinol. *Journal of Natural Products*, 71(4) :581–588, 2008.

[109] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, Y. D. Smurnyy, A. J. Williams, and T. S. Churanova. Computer-assisted methods for molecular structure elucidation : realizing a spectroscopist's dream. *Journal of Cheminformatics*, 1(3), 2009.

[110] M. Elyashberg, K. Blinov, and A. Williams. A systematic approach for the generation and verification of structural hypotheses. *Magnetic Resonance in Chemistry*, 47(5) :371–389, 2009.

[111] S. F. Cheatham, M. Kline, R. R. Sasaki, K. A. Blinov, M. E. Elyashberg, and S. G. Molodtsov. Enhanced automated structure elucidation by inclusion of two-bond specific data. *Magnetic Resonance in Chemistry*, 48(8) :571–574, 2010.

[112] M. E. Elyashberg, K. A. Blinov, S. G. Molodtsov, and A. J. Williams. Elucidating 'undecipherable' chemical structures using computer-assisted structure elucidation approaches. *Magnetic Resonance in Chemistry*, 50(1) :22–27, 2012.

[113] A. Moser, M. E. Elyashberg, A. J. Williams, K. A. Blinov, and J. DiMartino. Blind trials of computer-assisted structure elucidation software. *Journal of Cheminformatics*, 4(5), 2012.

[114] C. P. Butts, C. R. Jones, E. C. Towers, J. L. Flynn, L. Appleby, and N. J. Barron. Interproton distance determinations by NOE - surprising accuracy and precision in a rigid organic molecule. *Organic & Biomolecular Chemistry*, 9(1) :177–184, 2011.

[115] K. Wüthrich. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angewandte Chemie International Edition*, 42(29) :3340–3363, 2003.

[116] C. A. G. Haasnoot, F. A. A. M. de Leeuw, and C. Altona. The relationship between proton-proton NMR coupling constants and substituent electronegativities : An empirical generalization of the Karplus equation. *Tetrahedron*, 36(19) :2783–2792, 1980.

[117] G. Palermo, R. Riccio, and G. Bifulco. Effect of electronegative substituents and angular dependence on the heteronuclear spin–spin coupling constant $^3J_{C-H}$: An empirical prediction equation derived by density functional theory calculations. *The Journal of Organic Chemistry*, 75(6) :1982–1991, 2010.

[118] D. G. Corley and R. C. Durley. Strategies for database dereplication of natural products. *Journal of Natural Products*, 57(11) :1484–1490, 1994.

[119] J. Bradshaw, D. Butina, A. J. Dunn, R. H. Green, M. Hajek, M. M. Jones, J. C. Lindon, and P. J. Sidebottom. A rapid and facile method for the dereplication of purified natural products. *Journal of Natural Products*, 64(12) :1541–1544, 2001.

[120] W. Willker, D. Leibfritz, R. Kerssebaum, and W. Bermel. Gradient selection in inverse heteronuclear correlation spectroscopy. *Magnetic Resonance in Chemistry*, 31(3) :287–292, 1993.

[121] G. Lang, N. A. Mayhudin, M. I. Mitova, L. Sun, S. van der Sar, J. W. Blunt, A. L. J. Cole, G. Ellis, H. Laatsch, and M. H. G. Munro. Evolving trends in the dereplication of natural product extracts : New methodology for rapid, small-scale investigation of natural product extracts. *Journal of Natural Products*, 71(9) :1595–1599, 2008.

[122] F. Qiu, A. Imai, J. B. McAlpine, D. C. Lankin, I. Burton, T. Karakach, N. R. Farnsworth, S.-N. Chen, and G. F. Pauli. Dereplication, residual complexity, and rational naming : The case of the *Actaea* triterpenes. *Journal of Natural Products*, 75(3) :432–443, 2012.

[123] G. K. Pierens, M. Mobli, and V. Vegh. Effective protocol for database similarity searching of heteronuclear single quantum coherence spectra. *Analytical Chemistry*, 81(22) :9329–9335, 2009.

[124] G. Pierens, S. Brossi, Z. Yang, D. Reutens, and V. Vegh. HSQC spectral based similarity matching of compounds using nearest neighbours and a fast discrete genetic algorithm. *Journal of Cheminformatics*, 4(25), 2012.

[125] ACD/Structure Elucidator. `http://www.acdlabs.com/products/com_iden/elucidation/struc_eluc/`, . (Consulté le 24/10/2012).

[126] NMR-SAMS - Spectrum Research, LLC. `http://www.specres.com/nmrsams.asp`, . (Consulté le 24/10/2012).

[127] AssembleIt - ScienceSoft, LLC. `http://www.sciencesoft.net/AssembleIt.html`, . (Consulté le 24/10/2012).

[128] Le logiciel LSD. `http://www.univ-reims.fr/LSD`, . (Consulté le 24/10/2012).

[129] SENECA. `http://sourceforge.net/projects/seneca/`, . (Consulté le 24/10/2012).

[130] WebCocon. `http://cocon.nmr.de/`, . (Consulté le 24/10/2012).

[131] L. Griffiths. Towards the automatic analysis of ^{1}H NMR spectra. *Magnetic Resonance in Chemistry*, 38(6) :444–451, 2000.

[132] L. Griffiths. Towards the automatic analysis of ^{1}H NMR spectra : Part 2. accurate integrals and stoichiometry. *Magnetic Resonance in Chemistry*, 39(4) :194–202, 2001.

[133] L. Griffiths and J. D. Bright. Towards the automatic analysis of ^{1}H NMR spectra : Part 3. confirmation of postulated chemical structure. *Magnetic Resonance in Chemistry*, 40(10) :623–634, 2002.

[134] S. S. Golotvin, E. Vodopianov, B. A. Lefebvre, A. J. Williams, and T. D. Spitzer. Automated structure verification based on ^{1}H NMR prediction. *Magnetic Resonance in Chemistry*, 44(5) :524–538, 2006.

[135] H. Thiele, G. McLeod, M. Niemitz, and T. Kühn. Structure verification of small molecules using mass spectrometry and NMR spectroscopy. *Monatshefte für Chemie / Chemical Monthly*, 142 :717–730, 2011.

[136] E. Haapaniemi and M. Mesilaakso. ^{1}H and ^{13}C{^{1}H} NMR spectral parameters of sulfur mustards, nitrogen mustards, and lewisites : Computing and predicting of reference spectra for chemical identification. *Magnetic Resonance in Chemistry*, 50 (3) :196–207, 2012.

[137] PERCH Solutions. `http://www.perchsolutions.com/`, . (Consulté le 24/10/2012).

[138] Mnova Verify. `http://mestrelab.com/software/mnova-verify/`, . (Consulté le 24/10/2012).

[139] L. Griffiths and R. Horton. Towards the automatic analysis of NMR spectra : Part 6. confirmation of chemical structure employing both ^{1}H and ^{13}C NMR spectra. *Magnetic Resonance in Chemistry*, 44(2) :139–145, 2006.

[140] L. Griffiths, H. H. Beeley, and R. Horton. Towards the automatic analysis of NMR spectra : Part 7. assignment of ^{1}H by employing ^{1}H and ^{1}H/^{13}C correlation spectra. *Magnetic Resonance in Chemistry*, 46(9) :818–827, 2008.

[141] S. S. Golotvin, E. Vodopianov, R. Pol, B. A. Lefebvre, A. J. Williams, R. D. Rutkowske, and T. D. Spitzer. Automated structure verification based on a combination of 1D ^{1}H NMR and 2D ^{1}H–^{13}C HSQC spectra. *Magnetic Resonance in Chemistry*, 45(10) :803–813, 2007.

[142] P. Keyes, G. Hernandez, G. Cianchetta, J. Robinson, and B. Lefebvre. Automated compound verification using 2D-NMR HSQC data in an open-access environment. *Magnetic Resonance in Chemistry*, 47(1) :38–52, 2009.

[143] S. S. Golotvin, R. Pol, R. R. Sasaki, A. Nikitina, and P. Keyes. Concurrent combined verification : reducing false positives in automated NMR structure verification through the evaluation of multiple challenge control structures. *Magnetic Resonance in Chemistry*, 50(6) :429–435, 2012.

[144] Bruker CMC-se. `http://www.bruker-biospin.com/topspin3-cmc-se.html`, . (Consulté le 24/10/2012).

[145] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1) :31–36, 1988.

[146] D. Weininger, A. Weininger, and J. L. Weininger. SMILES. 2. algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences*, 29(2) :97–101, 1989.

[147] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *Journal of Chemical Information and Computer Sciences*, 32(3) :244–255, 1992.

[148] J.-M. Nuzillard. Quick method for anti-Bredt structure detection. *Journal of Chemical Information and Computer Sciences*, 34(4) :723–724, 1994.

[149] H.-C. Ehrlich and M. Rarey. Systematic benchmark of substructure search in molecular graphs - from Ullmann to VF2. *Journal of Cheminformatics*, 4(13), 2012.

[150] S. V. Ley, K. Doherty, G. Massiot, and J.-M. Nuzillard. "Connectivist" approach to organic structure determination LSD-program assisted NMR analysis of the insect antifeedant Azadirachtin. *Tetrahedron*, 50(42) :12267–12280, 1994.

[151] G. Almanza, L. Balderrama, C. Labbé, C. Lavaud, G. Massiot, J.-M. Nuzillard, J. D. Connolly, L. J. Farrugia, and D. S. Rycroft. Clerodane diterpenoids and an ursane triterpenoid from *Salvia haenkei*. Computer-assisted structural elucidation. *Tetrahedron*, 53(43) :14719–14728, 1997.

[152] J.-M. Nuzillard, J. D. Connolly, C. Delaude, B. Richard, M. Zèches-Hanrot, and L. Le Men-Olivier. Computer-assisted structural elucidation. Alkaloids with a novel diaza-adamantane skeleton from the seeds of *Acosmium panamense* (Fabaceae). *Tetrahedron*, 55(38) :11511–11518, 1999.

[153] D. A. Mulholland, M. Randrianarivelojosia, C. Lavaud, J.-M. Nuzillard, and S. L. Schwikkard. Limonoid derivatives from *Astrotrichilia voamatata*. *Phytochemistry*, 53(1) :115–118, 2000.

[154] D. A. Mulholland, S. L. Schwikkard, P. Sandor, and J.-M. Nuzillard. Delevoyin C, a tetranortriterpenoid from *Entandrophragma delevoyi*. *Phytochemistry*, 53(4) : 465–468, 2000.

[155] D. A. Mulholland, A. Langlois, M. Randrianarivelojosia, E. Derat, and J.-M. Nuzillard. The structural elucidation of a novel iridoid derivative from *Tachiadenus longiflorus* (Gentianaceae) using the LSD programme and quantum chemical computations. *Phytochemical Analysis*, 17(2) :87–90, 2006.

[156] A. Toribio, A. Bonfils, E. Delannay, E. Prost, D. Harakat, E. Henon, B. Richard, M. Litaudon, J.-M. Nuzillard, and J.-H. Renault. Novel seco-dibenzopyrrocoline alkaloid from *Cryptocarya oubatchensis*. *Organic Letters*, 8(17) :3825–3828, 2006.

[157] D. A. Mulholland, M. K. Langat, N. R. Crouch, H. M. Coley, E. M. Mutambi, and J.-M. Nuzillard. Cembranolides from the stem bark of the southern African medicinal plant, *Croton gratissimus* (Euphorbiaceae). *Phytochemistry*, 71(11-12) : 1381–1386, 2010.

[158] C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland. MOLGEN, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Analytica Chimica Acta*, 314(3) :141–147, 1995.

[159] T. Wieland, A. Kerber, and R. Laue. Principles of the generation of constitutional and configurational isomers. *Journal of Chemical Information and Computer Sciences*, 36(3) :413–419, 1996.

[160] C. Benecke, T. Grüner, A. Kerber, R. Laue, and T. Wieland. MOLecular structure GENeration with MOLGEN, new features and future developments. *Fresenius' Journal of Analytical Chemistry*, 359(1) :23–32, 1997.

[161] MOLGEN - Molecular Structure Generation. `http://molgen.de/?src=documents/molgenonline`, . (Consulté le 24/10/2012).

[162] P. D. Senter and C.-L. Chen. Liriodendronine, an oxoaporphine pigment from discolored sapwood of *Liriodendron tulipifera*. *Phytochemistry*, 16(12) :2015–2017, 1977.

[163] R. E. Carhart. A model-based approach to the teletype printing of chemical structures. *Journal of Chemical Information and Computer Sciences*, 16(2) :82–88, 1976.

[164] H. E. Helson. Structure diagram generation. In *Reviews in Computational Chemistry*, pages 313–398. WILEY-VCH, 1999.

[165] Open Babel. `http://openbabel.org/wiki/Main_Page`, . (Consulté le 24/10/2012).

[166] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK) : An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43 (2) :493–500, 2003.

[167] CDK : The Chemistry Development Kit. `http://sourceforge.net/projects/cdk/`, . (Consulté le 24/10/2012).

[168] RDKit : Cheminformatics and Machine Learning Software. `http://www.rdkit.org/`, . (Consulté le 24/10/2012).

[169] OEChem Toolkit. `http://www.eyesopen.com/products`, . (Consulté le 24/10/2012).

[170] M. J. P. Ferreira, V. de P. Emerenciano, G. A. R. Linia, P. Romoff, P. A. T. Macari, and G. V. Rodrigues. ^{13}C NMR spectroscopy of monoterpenoids. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 33(3-4) :153–206, 1998.

[171] S. A. Vestri Alvarenga, J.-P. Gastmans, G. do Vale Rodrigues, P. Roberto H. Moreno, and V. de P. Emerenciano. A computer-assisted approach for chemotaxonomic studies – diterpenes in lamiaceae. *Phytochemistry*, 56(6) :583–595, 2001.

[172] M. J. P. Ferreira, A. J. C. Brant, G. V. Rodrigues, and V. de P. Emerenciano. Automatic identification of terpenoid skeletons through ^{13}C nuclear magnetic resonance data disfunctionalization. *Analytica Chimica Acta*, 429(1) :151–170, 2001.

[173] M. J. P. Ferreira, F. C. Oliveira, S. A. V. Alvarenga, P. A. T. Macari, G. V. Rodrigues, and V. de P. Emerenciano. Automatic identification by ^{13}C NMR of substituent groups bonded in natural product skeletons. *Computers & Chemistry*, 26 (6) :601–632, 2002.

[174] M. J. P. Ferreira, S. A. V. Alvarenga, P. A. T. Macari, G. V. Rodrigues, and V. de P. Emerenciano. A program for terpenoid skeleton prediction based on botanical information. *Biochemical Systematics and Ecology*, 31(1) :25–43, 2003.

[175] J.-M. Nuzillard and V. de P. Emerenciano. Automatic structure elucidation through data base search and 2D NMR spectral analysis. *Natural Product Communications*, 1(1) :57–64, 2006.

[176] B. Plainchont, J.-M. Nuzillard, G. V. Rodrigues, M. J. P. Ferreira, M. T. Scotti, and V. de P. Emerenciano. New improvements in automatic structure elucidation using the LSD (logic for structure determination) and the SISTEMAT expert systems. *Natural Product Communications*, 5(5) :763–770, 2010.

[177] M. Li, J. Wu, S. Zhang, Q. Xiao, and Q. Li. Xylocarpins A and B, two new mexicanolides from the seeds of a Chinese mangrove *Xylocarpus granatum* : NMR investigation in mixture. *Magnetic Resonance in Chemistry*, 45(8) :705–709, 2007.

[178] H. Bunke, P. Foggia, C. Guidobaldi, C. Sansone, and M. Vento. A comparison of algorithms for maximum common subgraph on randomly connected graphs. In

Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition. Springer-Verlag, 2002.

[179] J.-M. Nuzillard and G. Massiot. Computer-aided spectral assignment in nuclear magnetic resonance spectroscopy. *Analytica Chimica Acta*, 242(1) :37–41, 1991.

[180] ACD/Chemsketch Freeware. `http://www.acdlabs.com/resources/freeware/chemsketch/`, . (Consulté le 24/10/2012).

[181] C. Steinbeck, S. Krause, and S. Kuhn. NMRShiftDB – constructing a free chemical information system with open-source components. *Journal of Chemical Information and Computer Sciences*, 43(6) :1733–1739, 2003.

[182] C. Steinbeck and S. Kuhn. NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*, 65(19) :2711–2717, 2004.

[183] nmrshiftdb2 – open nmr database on the web. `http://nmrshiftdb.nmr.uni-koeln.de/`, . (Consulté le 24/10/2012).

[184] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2) :107–113, 1965.

[185] Z. Ouyang, S. Yuan, J. Brandt, and C. Zheng. An effective topological symmetry perception and unique numbering algorithm. *Journal of Chemical Information and Computer Sciences*, 39(2) :299–303, 1999.

[186] S. E. Stein, S. R. Heller, and D. Tchekhovski. An open standard for chemical structure representation - the IUPAC Chemical Identifier. In *Proceedings of the Nîmes International Chemical Information Conference*, pages 131–143, 2003.

[187] The GNU General Public License (GPL). `http://www.gnu.org/licenses/gpl.html`, . (Consulté le 24/10/2012).

[188] Le logiciel CASA. `http://www.univ-reims.fr/LSD/JmnSoft/CASA`, . (Consulté le 24/10/2012).

[189] S. Baggett, P. Protiva, E. P. Mazzola, H. Yang, E. T. Ressler, M. J. Basile, I. B. Weinstein, and E. J. Kennelly. Bioactive benzophenones from *Garcinia xanthochymus* fruits. *Journal of Natural Products*, 68(3) :354–360, 2005.

[190] X.-L. Wang, A.-E. Hay, A. Matheeussen, M. P. Gupta, and K. Hostettmann. Structure elucidation and NMR assignments of two new triterpenoids from the stems of *Paragonia pyramidata* (bignoniaceae). *Magnetic Resonance in Chemistry*, 49(4) : 184–189, 2011.

Annexe A

Liste des commandes pour le fichier d'entrée de LSD

Commandes de base

Mnémonique de commande	Application	Description des paramètres
`MULT` I A I I Z	Définition du statut des atomes	– P1 : Numéro d'atome. – P2 : Symbole de l'atome. – P3 : État d'hybridation (1, 2 ou 3 pour les atomes *sp*, sp^2 et sp^3). – P4 : Nombre d'atomes d'hydrogène portés par l'atome. – P5 : Charge de l'atome (facultative).
`BOND` I I	Liaison	– P1 : Numéro de l'atome 1. – P2 : Numéro de l'atome 2.
`HSQC` I I (ou `HMQC`)	Corrélation hétéronucléaire à travers 1 liaison	– P1 : Numéro de l'atome (non hydrogène). – P2 : Numéro de l'atome d'hydrogène.

Mnémonique de commande	Application	Description des paramètres
`COSY` V I O O	Corrélation COSY	– P1 : Numéro d'hydrogène ou liste d'hydrogènes entre parenthèses. – P2 : Numéro d'hydrogène. – P3 : Longueur de chemin de couplage optionnelle (limite inférieure). – P4 : Longueur de chemin de couplage optionnelle (limite supérieure).
`HMBC` V I O O	Corrélation HMBC	– P1 : Numéro d'atome non hydrogène ou liste d'atomes entre parenthèses. – P2 : Numéro d'hydrogène. – P3 : Longueur de chemin de couplage optionnelle (limite inférieure). – P4 : Longueur de chemin de couplage optionnelle (limite supérieure).
`LIST` Ln S	Définition d'une liste d'atomes	– P1 : Référence de la liste (n est compris entre 1 et 20). – P2 : Numéro d'atomes non hydrogènes.
`PROP` B I Ln H	Définition de l'environnement des atomes	– P1 : Numéro d'atome ou référence d'une liste d'atomes. – P2 : Nombre de voisins. La valeur 0 signifie tous les voisins. – P3 : Référence de la liste des voisins. – P4 : Signe optionnel (+ ou −).
`SHIX` I R	Déplacement chimique des atomes non hydrogènes	– P1 : Numéro d'atome. – P2 : Valeur du déplacement chimique.

Mnémonique de commande	Application	Description des paramètres
SHIH I R	Déplacement chimique des atomes d'hydrogène	– P1 : Numéro d'hydrogène. – P2 : Valeur du déplacement chimique.

Définition des listes d'atomes

Mnémonique de commande	Application	Description des paramètres
CARB Ln	Liste des atomes de carbone	P1 : Référence de la liste créée.
HETE Ln	Liste des atomes qui ne sont pas des carbones (et pas des hydrogènes)	P1 : Référence de la liste créée.
SP3 Ln	Liste des atomes qui n'ont que des liaisons simples	P1 : Référence de la liste créée.
SP2 Ln	Liste des atomes qui ont exactement une double liaison	P1 : Référence de la liste créée.
SP Ln	Liste des atomes qui ont une triple liaison ou deux doubles liaisons	P1 : Référence de la liste créée.
FULL Ln	Liste de tous les atomes	P1 : Référence de la liste créée.
QUAT Ln	Liste des carbones liés à 0 hydrogène	P1 : Référence de la liste créée.
CH Ln	Liste des carbones liés à 1 hydrogène	P1 : Référence de la liste créée.
CH2 Ln	Liste des carbones liés à 2 hydrogènes	P1 : Référence de la liste créée.
CH3 Ln	Liste des carbones liés à 3 hydrogènes	P1 : Référence de la liste créée.
CHAR Ln	Liste des atomes chargés	P1 : Référence de la liste créée.
CPOS Ln	Liste des atomes portant une charge positive	P1 : Référence de la liste créée.
CNEG Ln	Liste des atomes portant une charge négative	P1 : Référence de la liste créée.

Mnémonique de commande	Application	Description des paramètres
`ELEM` Ln A	Liste de tous les atomes de symbole atomique A	– P1 : Référence de la liste créée. – P2 : Symbole atomique.
`GREQ` Ln I	Liste des atomes dont le numéro est supérieur ou égal à P2	– P1 : Référence de la liste créée. – P2 : Entier qui sert de valeur de comparaison.
`LEEQ` Ln I	Liste des atomes dont le numéro est inférieur ou égal à P2	– P1 : Référence de la liste créée. – P2 : Entier qui sert de valeur de comparaison.
`GRTH` Ln I	Liste des atomes dont le numéro est strictement supérieur à P2	– P1 : Référence de la liste créée. – P2 : Entier qui sert de valeur de comparaison.
`LETH` Ln I	Liste des atomes dont le numéro est strictement inférieur à P2	– P1 : Référence de la liste créée. – P2 : Entier qui sert de valeur de comparaison.

Mnémonique de commande	Application	Description des paramètres
`UNIO` B B Ln	P3 est l'union de P1 et P2	– P1 : Numéro d'atome ou référence d'une liste d'atomes. – P2 : Numéro d'atome ou référence d'une liste d'atomes. – P3 : Référence de la liste créée.
`INTE` B B Ln	P3 est l'intersection de P1 et P2	– P1 : Numéro d'atome ou référence d'une liste d'atomes. – P2 : Numéro d'atome ou référence d'une liste d'atomes. – P3 : Référence de la liste créée.
`DIFF` B B Ln	P3 contient les numéros des atomes de P1 qui ne sont pas dans P2	– P1 : Numéro d'atome ou référence d'une liste d'atomes. – P2 : Numéro d'atome ou référence d'une liste d'atomes. – P3 : Référence de la liste créée.

Contrôle de l'exécution

Mnémonique de commande	Application	Description des paramètres
`ENTR` I	Affichage de l'état du problème avant sa résolution	– P1 = 0 : Pas d'affichage (défaut). – P1 = 1 : Affichage actif.
`HIST` I	Affichage du détail des étapes de la résolution	– P1 = 0 : Pas d'affichage (défaut). – P1 = 1 : Affichage actif.
`DISP` I	Format de sortie	– P1 = 0 : Imprime des listes de liaisons. – P1 = 1 : Résultat formaté pour OUTLSD (défaut).
`VERB` I	Verbosité	– P1 = 0 : Programme muet (défaut). – P1 = 1 : Verbosité active. – P1 = 2 : Programme très verbeux.
`PART` I	Production des solutions incomplètes	– P1 = 0 : Non (défaut). – P1 = 1 : Oui.
`STEP` I	Exécution pas à pas	– P1 = 0 : Exécution normale (défaut). – P1 = 1 : État actif.
`WORK` I	Autorisation de la recherche des solutions	– P1 = 0 : Uniquement lecture et interprétation du fichier de données. – P1 = 1 : Résolution active (défaut).

Mnémonique de commande	Application	Description des paramètres
`MLEV` I	Arrêt de l'analyse à l'étape P1	– P1 = 0 : Rien (défaut). – P1 = Numéro d'étape.
`DUPL` I	Élimination des solutions dupliquées	– P1 = 0 : Des solutions dupliquées peuvent être produites. – P1 = 1 : Les solutions dupliquées sont éliminées. – P1 = 2 : Les structures dupliquées sont éliminées (défaut).
`SUBS` I	Validation de sous-structure	– P1 = 0 : Pas de recherche de sous-structure. – P1 = 1 : Acceptation des solutions qui valident la contrainte de sous-structure (défaut). – P1 = -1 : Acceptation des solutions qui ne valident pas la contrainte de sous-structure.
`ELIM` I I	Élimination des corrélations HMBC et/ou COSY invalides	– P1 : Nombre maximum de corrélations à éliminer. – P2 : Nombre maximum de liaisons admissibles pour une corrélation éliminée.
`FILT` I	Mode filtrage par sous-structure	– P1 = 0 : Usage normal de LSD (défaut). – P1 = 1 : Activation du mode filtre.
`CNTD` I	Élimination des solutions non connexes (en plusieurs morceaux)	– P1 = 0 : Aucune vérification n'est effectuée. – P1 = 1 : Seules les solutions connexes (en un seul morceau) sont retenues (défaut).

Mnémonique de commande	Application	Description des paramètres
`MAXS` I	Limitation du nombre de structures produites	– P1 = 0 : Aucune limitation (défaut). – P1 > 0 : LSD s'arrête après avoir produit P1 solutions.
`MAXT` I	Limitation du temps de résolution	– P1 = 0 : Aucune limitation (défaut). – P1 > 0 : LSD s'arrête après P1 secondes.
`CCLA` I	Contrôle de l'usage des classes d'atomes de carbone équivalents	– P1 = 0 : Pas de classes. Toutes les attributions des ^{13}C sont produites (défaut). – P1 = 1 : Les atomes de carbone initialement non liés en non corrélés sont équivalents.
`COUF` C	Définition du nom du fichier de comptage des solutions	– P1 : Nom du fichier de comptage des solutions (défaut : solncounter).
`STOF` C	Définition du nom du fichier d'arrêt de LSD	– P1 : Nom du fichier d'arrêt de LSD (défaut : stoplsd).
`BRUL` I	Test pour les structures anti-Bredt	– P1 = 0 : Pas de vérification. – P1 = 1 : Les structures anti-Bredt sont exclues (défaut).

Informations de sous-structure

Mnémonique de commande	Application	Description des paramètres
`SSTR` Sn A V V	Statut des sous-atomes	– P1 : Numéro de sous-atome. – P2 : Symbole du sous-atome. – P3 : État d'hybridation, 1 (sp), 2 (sp^2), 3 (sp^3), (2 3) (l'un des deux) ou (1 2 3) (l'un des trois). – P4 : Multiplicité.
`LINK` Sn Sn	Sous-liaisons	– P1 : Numéro de sous-atome 1. – P2 : Numéro de sous-atome 2.
`ASGN` Sn I	Attribution d'un sous-atome	– P1 : Numéro de sous-atome. – P2 : Numéro de l'atome correspondant.
`DEFF` Fn C	Définition d'un fragment externe à partir d'un nom de fichier	– P1 : Numéro de fragment, supérieur ou égal à 1. – P2 : Nom (chemin d'accès) du fichier où le fragment est défini.
`SKEL` Fn C	Définition d'un fragment externe à partir d'un nom de squelette	– P1 : Numéro de fragment, supérieur ou égal à 1. – P2 : Nom de squelette.
`PATH` C	Définition des dossiers où sont stockés les squelettes	– P1 : Chemin d'accès à un dossier de squelettes.
`FEXP` C	Expression logique entre fragments	– P1 : Expression logique qui combine les résultats de recherche des fragments.

Annexe B

Fichiers d'entrée de LSD

Figure B.1 – Fichier LSD du sesterterpène

```
ENTR 1
ELIM 2 4

MULT 1 C 2 0
MULT 2 C 2 0
MULT 3 C 2 0
MULT 4 C 2 0
MULT 5 C 3 0
MULT 6 C 3 0
MULT 7 C 3 1
MULT 8 C 3 0
MULT 9 C 3 1
MULT 10 C 3 1
MULT 11 C 3 2
MULT 12 C 3 2
MULT 13 C 3 2
MULT 14 C 3 2
MULT 15 C 3 0
MULT 16 C 3 2
MULT 17 C 3 3
MULT 18 C 3 2
MULT 19 C 3 2
MULT 20 C 3 2
MULT 21 C 3 2
MULT 22 C 3 3
MULT 23 C 3 3
MULT 24 C 3 3
MULT 25 C 3 3
MULT 26 O 2 0
MULT 27 O 2 0
MULT 28 O 3 0
```

```
HSQC 7 7
HSQC 9 9
HSQC 10 10
HSQC 11 11
HSQC 12 12
HSQC 13 13
HSQC 14 14
HSQC 16 16
HSQC 17 17
HSQC 18 18
HSQC 19 19
HSQC 20 20
HSQC 21 21
HSQC 22 22
HSQC 23 23
HSQC 24 24
HSQC 25 25

BOND 1 26
BOND 2 27
BOND 2 28
BOND 5 28
BOND 3 4

COSY 7 9
COSY 7 20
COSY 10 12
COSY 11 16
COSY 13 14
COSY 20 18
COSY 23 10
COSY 25 9
COSY 23 12 3 4
COSY 13 9 3 4
COSY 17 16 4 0
COSY 24 14 4 0
```

```
HMBC 24 7
HMBC 20 7
HMBC 15 7
HMBC 9 7
HMBC 25 9
HMBC 20 9
HMBC 7 9
HMBC 1 9
HMBC 2 10
HMBC 11 10
HMBC 12 10
HMBC 23 10
HMBC 16 11
HMBC 10 11
HMBC 8 11
HMBC 6 11
HMBC 5 11
HMBC 23 12
HMBC 10 12
HMBC 2 12
HMBC 5 12
HMBC 25 13
HMBC 15 13
HMBC 14 13
HMBC 9 13
HMBC 1 13
HMBC 24 14
HMBC 15 14
HMBC 13 14
HMBC 7 14
HMBC 4 14
HMBC 1 14
```

```
HMBC 17 16
HMBC 11 16
HMBC 6 16
HMBC 3 16
HMBC 8 16
HMBC 5 16
HMBC 16 17 2 3
HMBC 8 17 2 3
HMBC 6 17 2 3
HMBC 3 17 2 3
HMBC 22 19
HMBC 21 19
HMBC 8 19
HMBC 6 19
HMBC 4 19
HMBC 18 20
HMBC 15 20
HMBC 7 20
HMBC 3 20
HMBC 19 22 2 3
HMBC 8 22 2 3
HMBC 6 22 2 3
HMBC 5 22 2 3
HMBC 10 23 2 3
HMBC 12 23 2 3
HMBC 5 23 2 3
HMBC 7 24 2 3
HMBC 4 24 2 3
HMBC 14 24 2 3
HMBC 15 24 2 3
HMBC 20 25
HMBC 9 25 2 3
HMBC 7 25 2 3
HMBC 1 25 2 3
```

Figure B.2 – Fichier LSD de l'alcaloïde aporphinique

```
ENTR 1
ELIM 3 4

MULT 1 C 3 3
MULT 2 C 3 3
MULT 3 C 2 1
MULT 4 C 2 0
MULT 5 C 2 1
MULT 6 C 2 1
MULT 7 C 2 0
MULT 8 C 2 1
MULT 9 C 2 1
MULT 10 C 2 0
MULT 11 C 2 0
MULT 12 C 2 1
MULT 13 C 2 0
MULT 14 C 2 1
MULT 15 C 2 0
MULT 16 C 2 0
MULT 17 C 2 0
MULT 18 C 2 0
MULT 19 N 2 0 1
MULT 20 O 3 0
MULT 21 O 2 0
MULT 22 O 3 0 -1

HSQC 1 1
HSQC 2 2
HSQC 3 3
HSQC 5 5
HSQC 6 6
HSQC 8 8
HSQC 9 9
HSQC 12 12
HSQC 14 14

BOND 19 1
BOND 2 20

COSY 5 14
COSY 8 12
COSY 6 12
COSY 6 9

HMBC 18 9
HMBC 17 3
HMBC 16 2
HMBC 16 3
HMBC 15 14
HMBC 15 1
HMBC 13 9
HMBC 13 12
HMBC 11 8
HMBC 11 6
HMBC 10 1
HMBC 10 14
HMBC 10 3
HMBC 7 5
HMBC 7 3
HMBC 4 8
HMBC 4 3
HMBC 14 1
HMBC 14 5
HMBC 12 9
HMBC 12 6
HMBC 9 12
HMBC 8 6
HMBC 6 8
HMBC 5 14
HMBC 5 3
HMBC 3 5
HMBC 1 14

DEFF F1 "Filters/ring3"
DEFF F2 "Filters/ring4"
DEFF F3 "Filters/ring7"
DEFF F4 "Filters/ring8"
DEFF F5 "Filters/ring9"

FEXP "NOT F1 AND NOT F2 and NOT F3 and NOT F4 and NOT F5"
```

Annexe C

Squelettes extraits de la base de données SISTEMAT

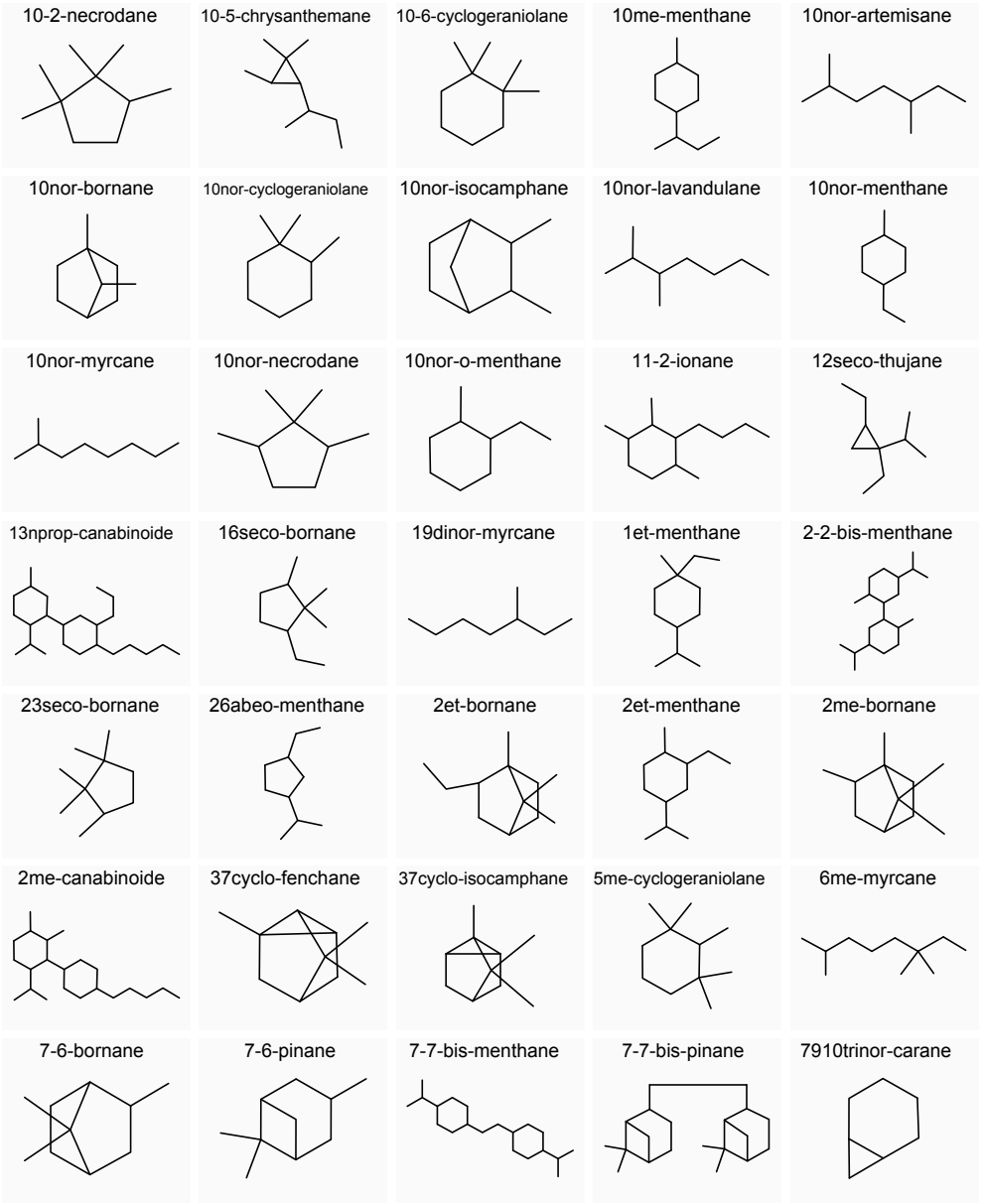


Figure C.1 – Squelettes de monoterpènes

FIGURE C.2 – Squelettes de monoterpènes (suite)

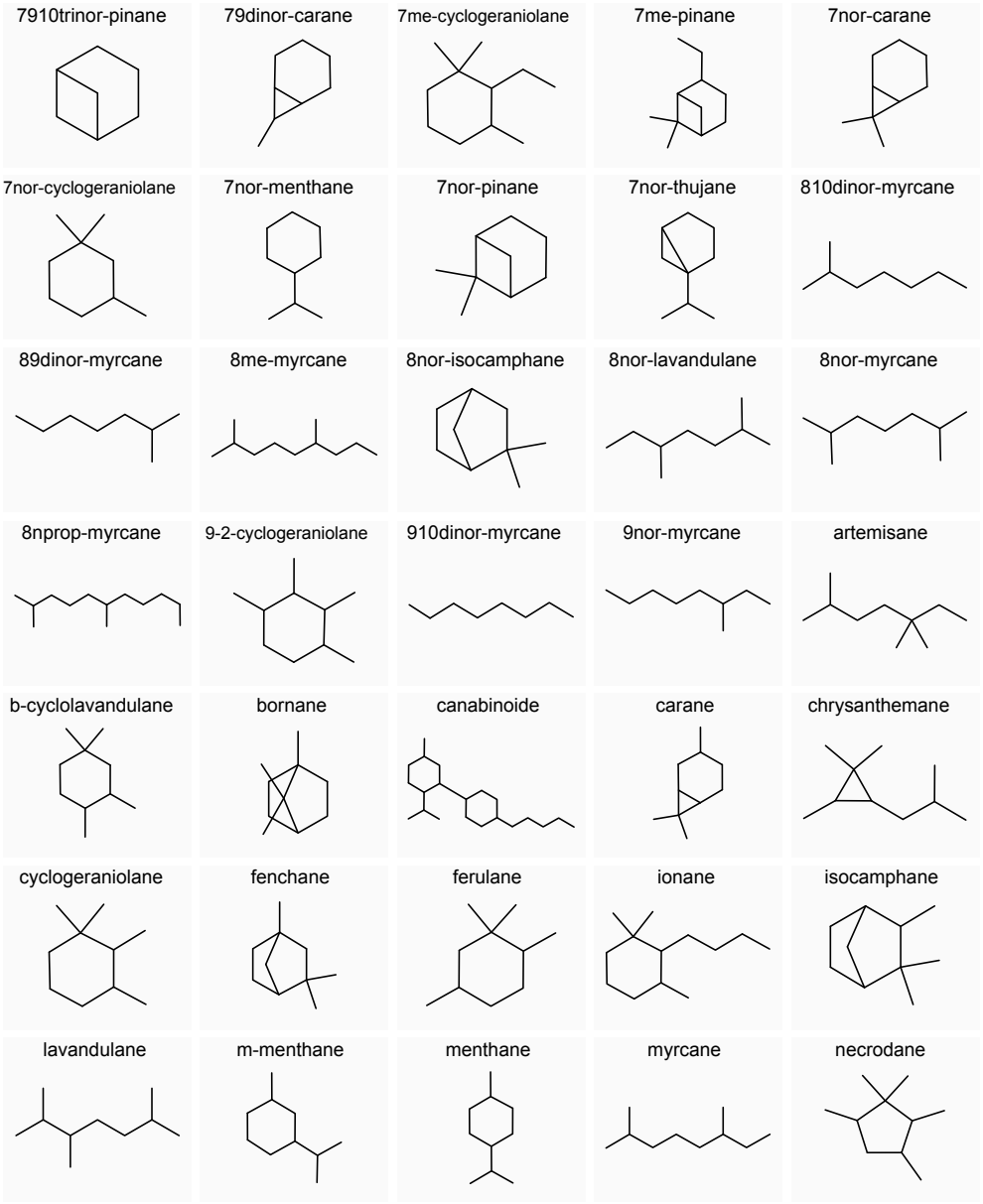


FIGURE C.3 – Squelettes de monoterpènes (suite)

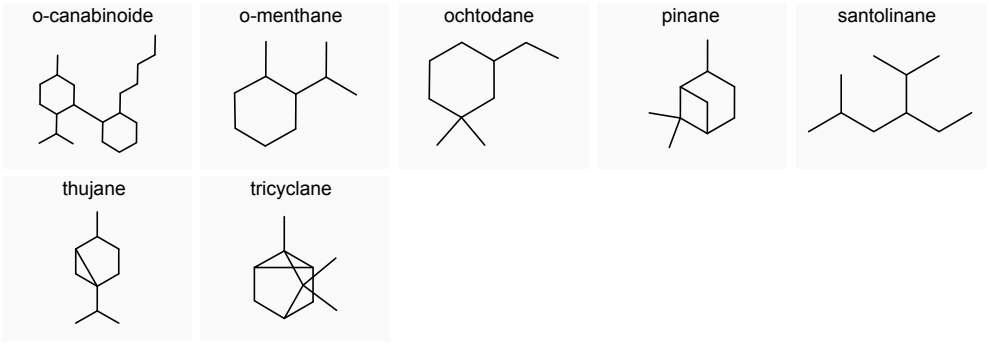


FIGURE C.4 – Squelettes de sesquiterpènes

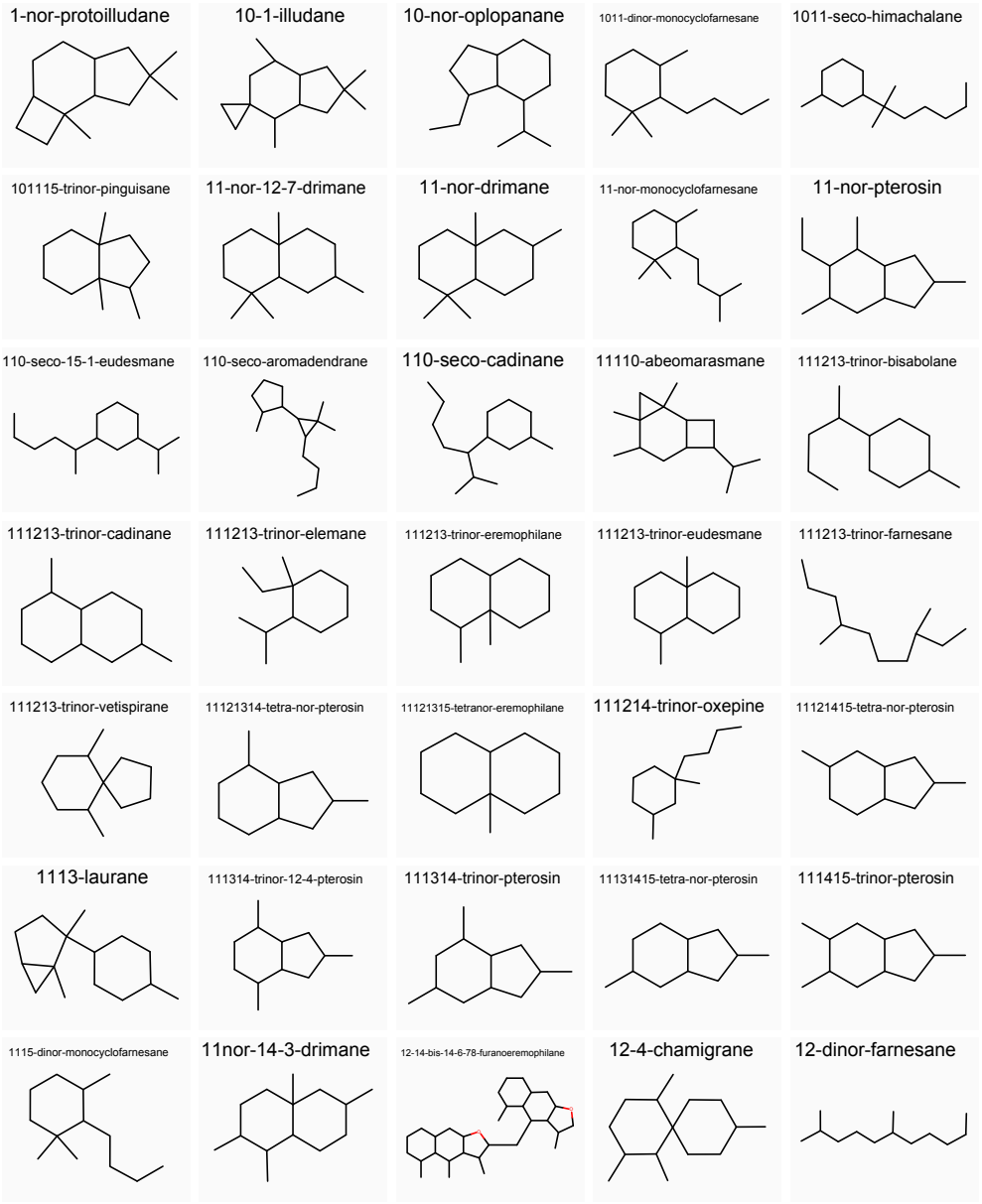


FIGURE C.5 – Squelettes de sesquiterpènes (suite)

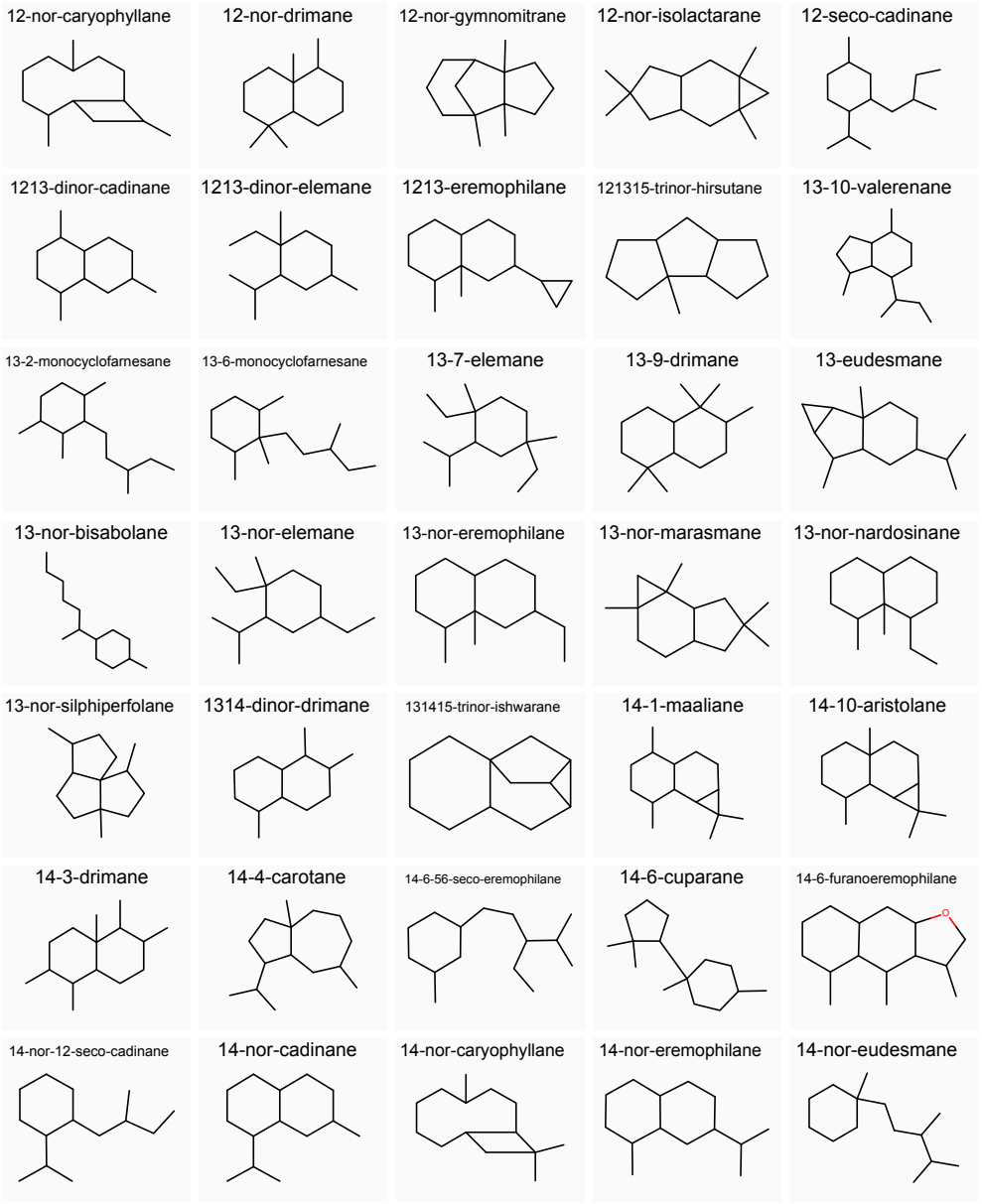


FIGURE C.6 – Squelettes de sesquiterpènes (suite)

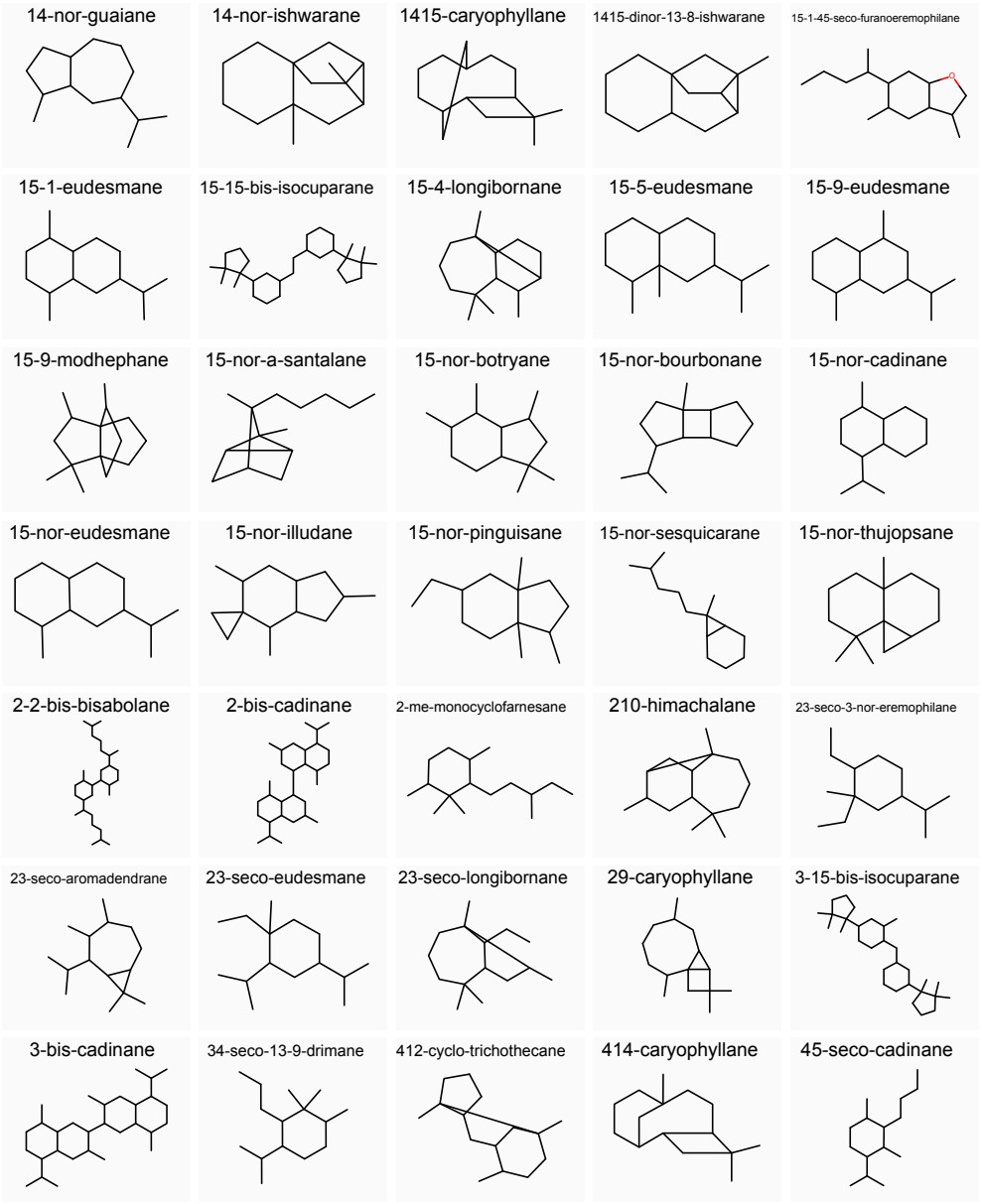


FIGURE C.7 – Squelettes de sesquiterpènes (suite)

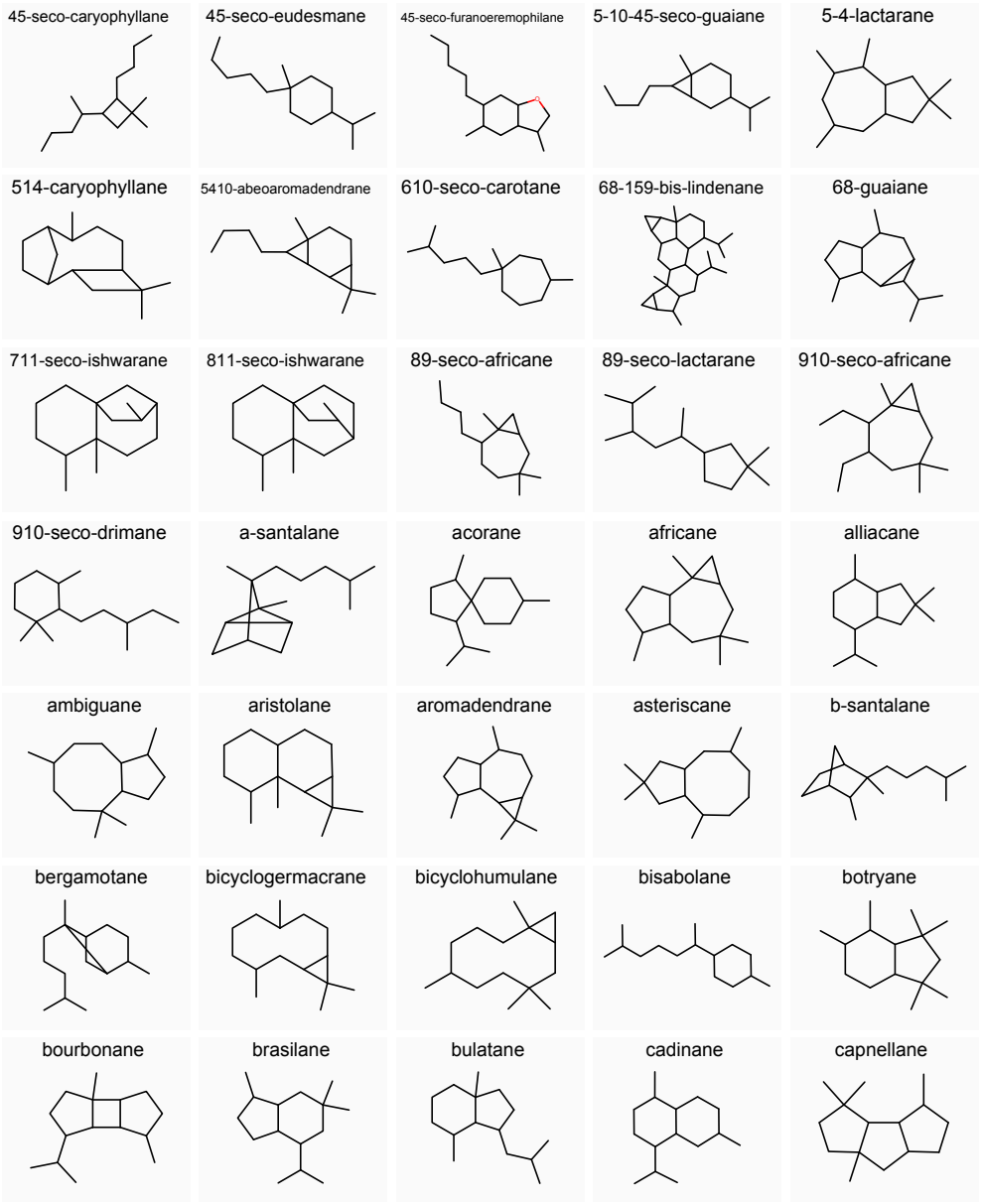


FIGURE C.8 – Squelettes de sesquiterpènes (suite)

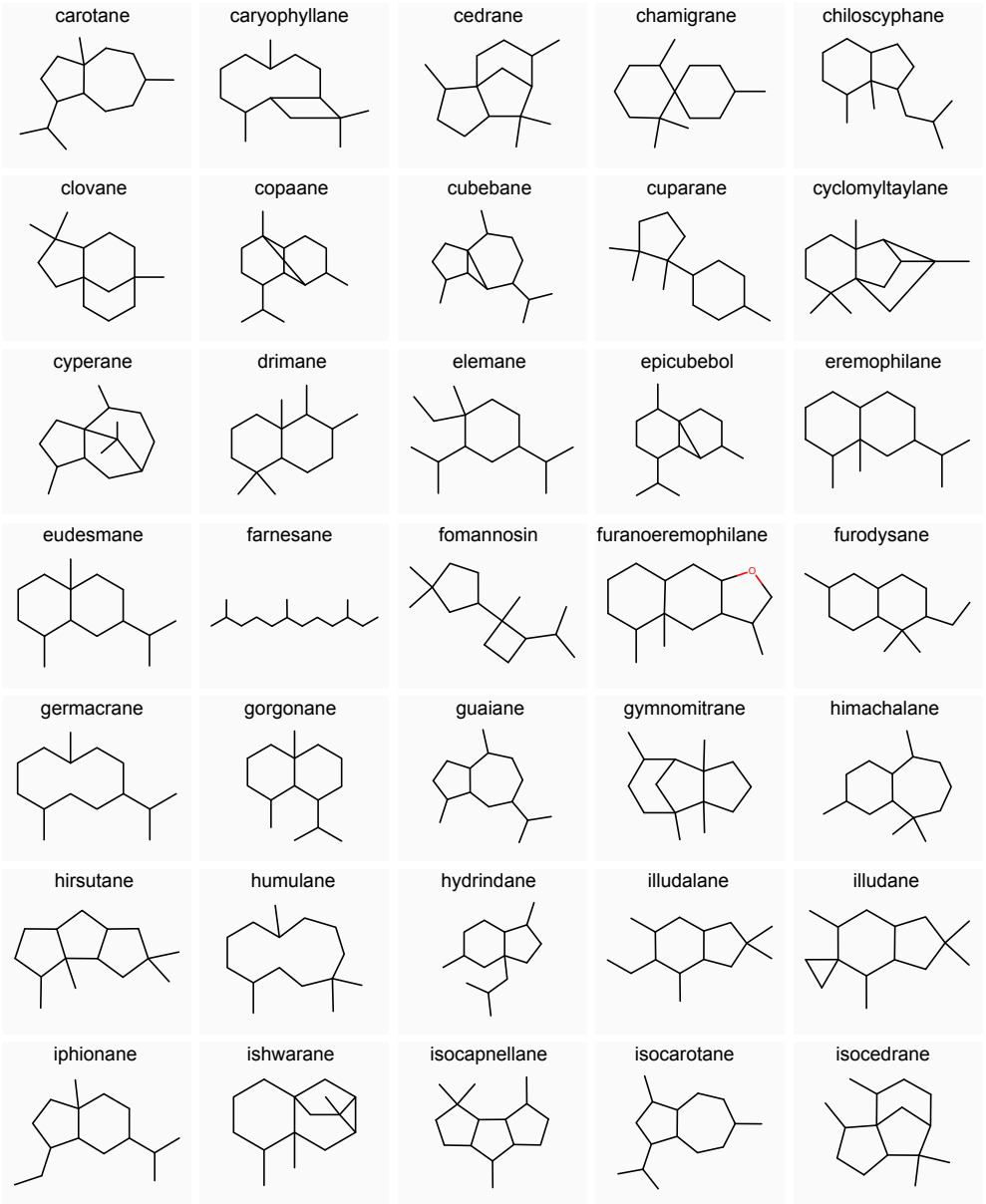


FIGURE C.9 – Squelettes de sesquiterpènes (suite)

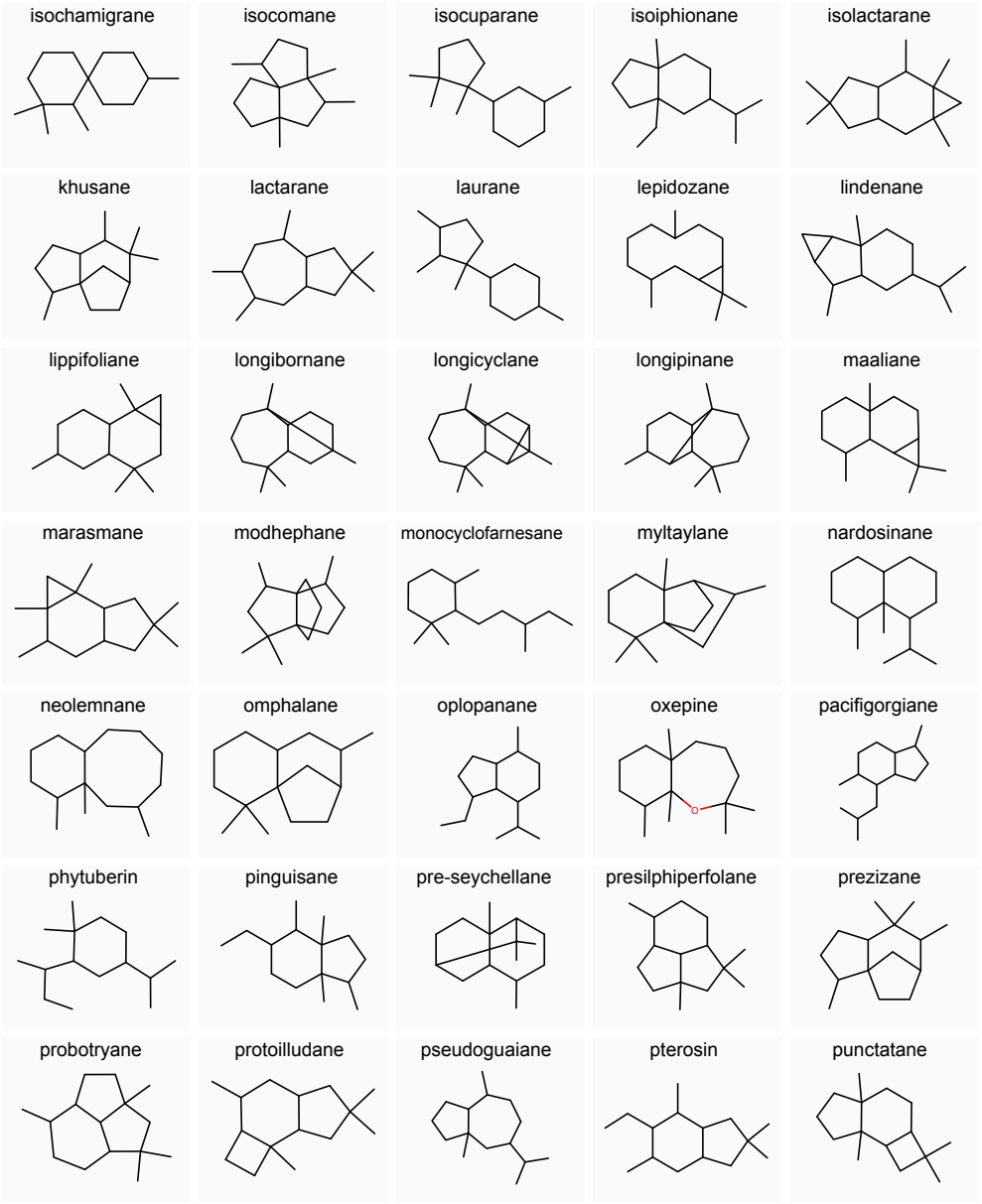


FIGURE C.10 – Squelettes de sesquiterpènes (suite)

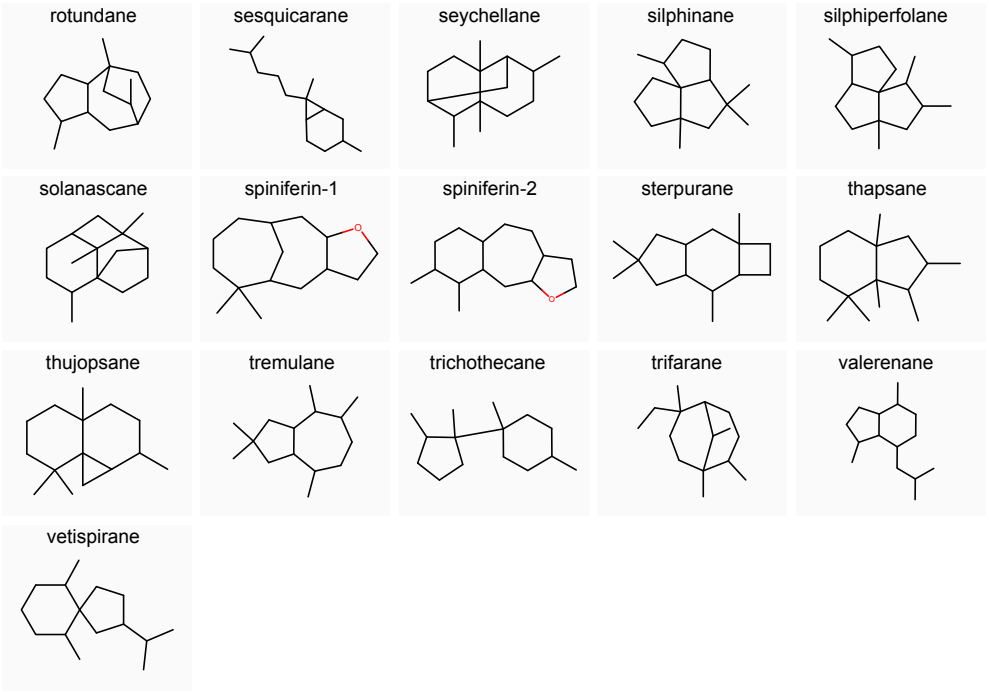


FIGURE C.11 – Squelettes de diterpènes

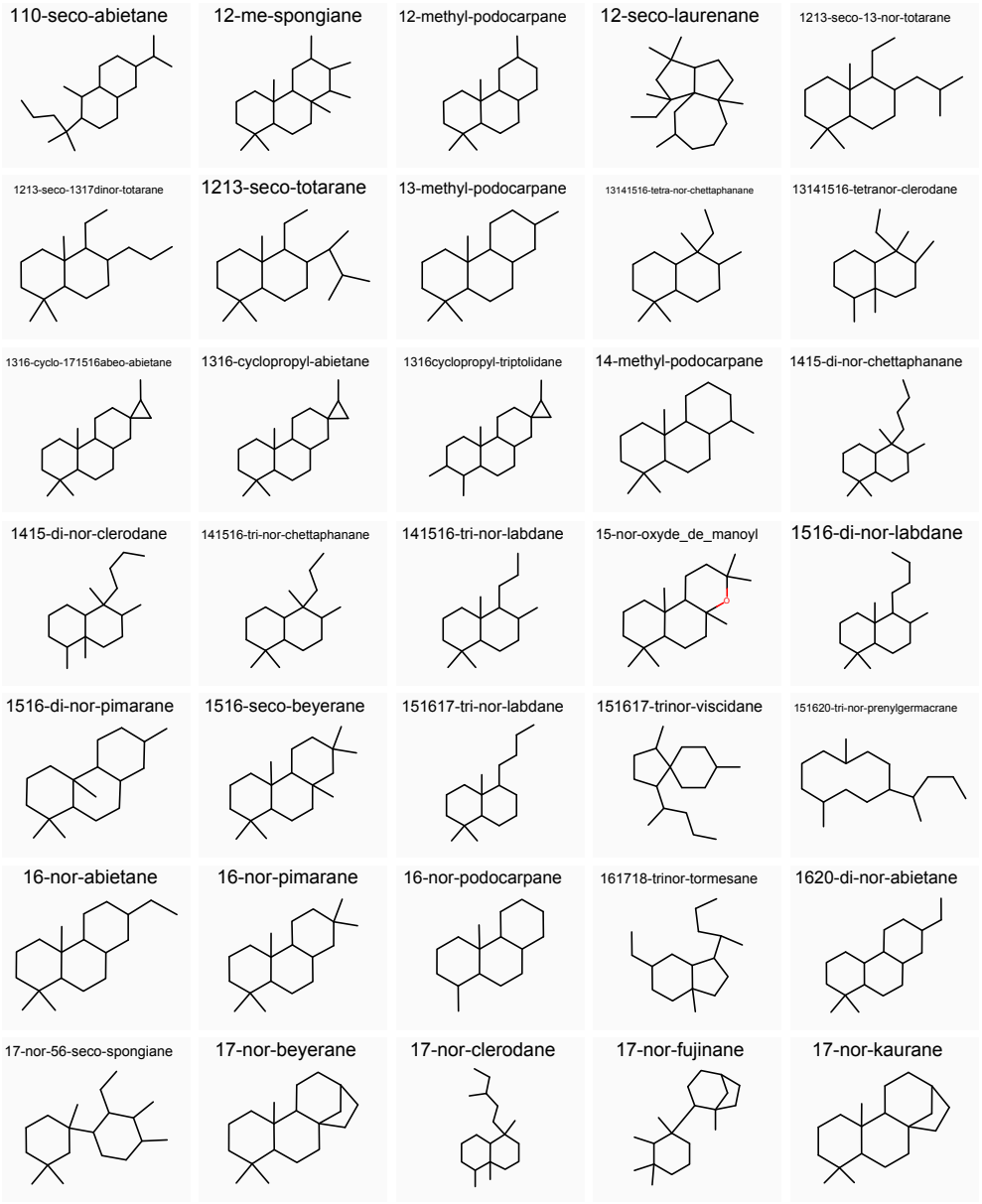


FIGURE C.12 – Squelettes de diterpènes (suite)

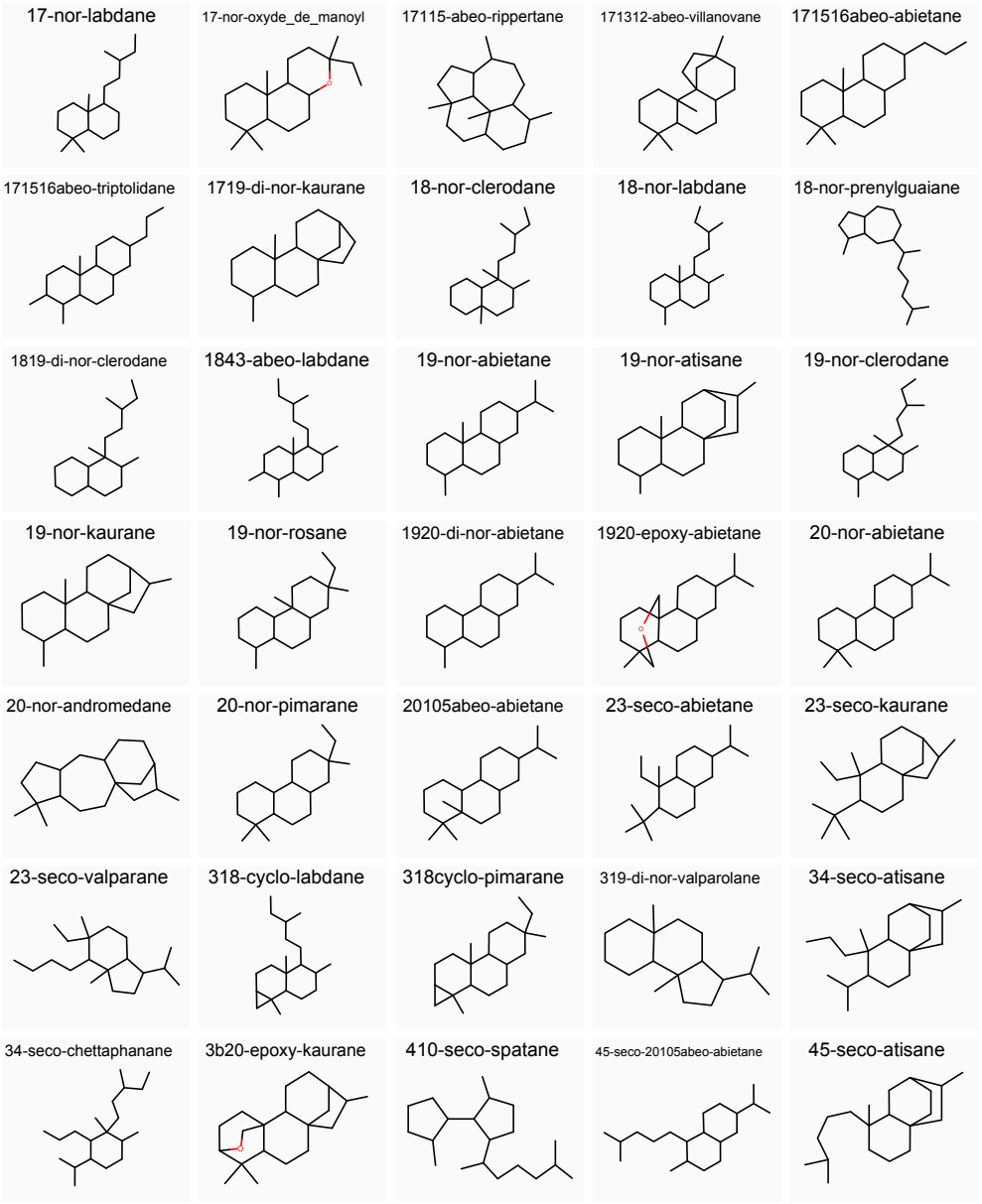


FIGURE C.13 – Squelettes de diterpènes (suite)

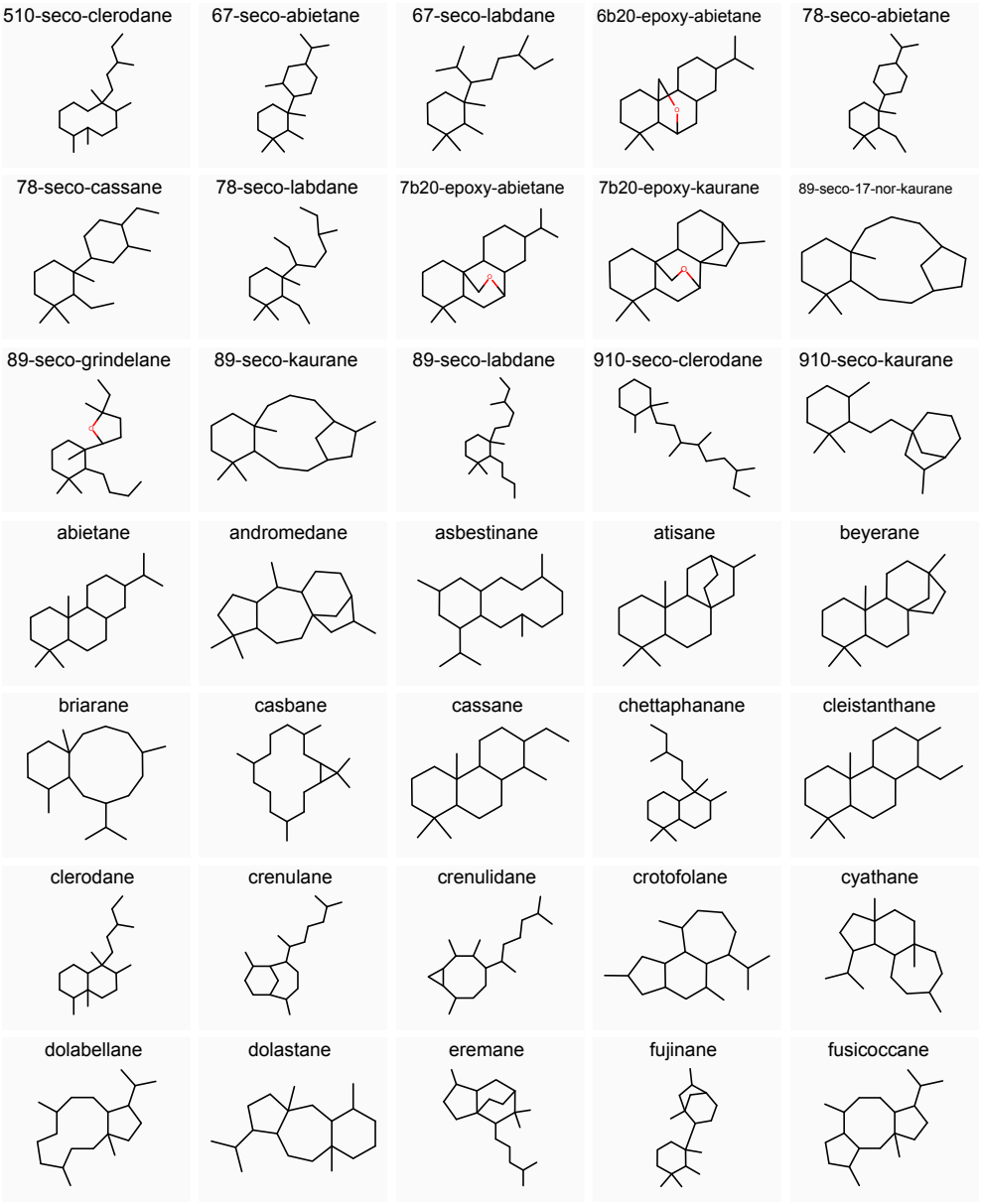


FIGURE C.14 – Squelettes de diterpènes (suite)

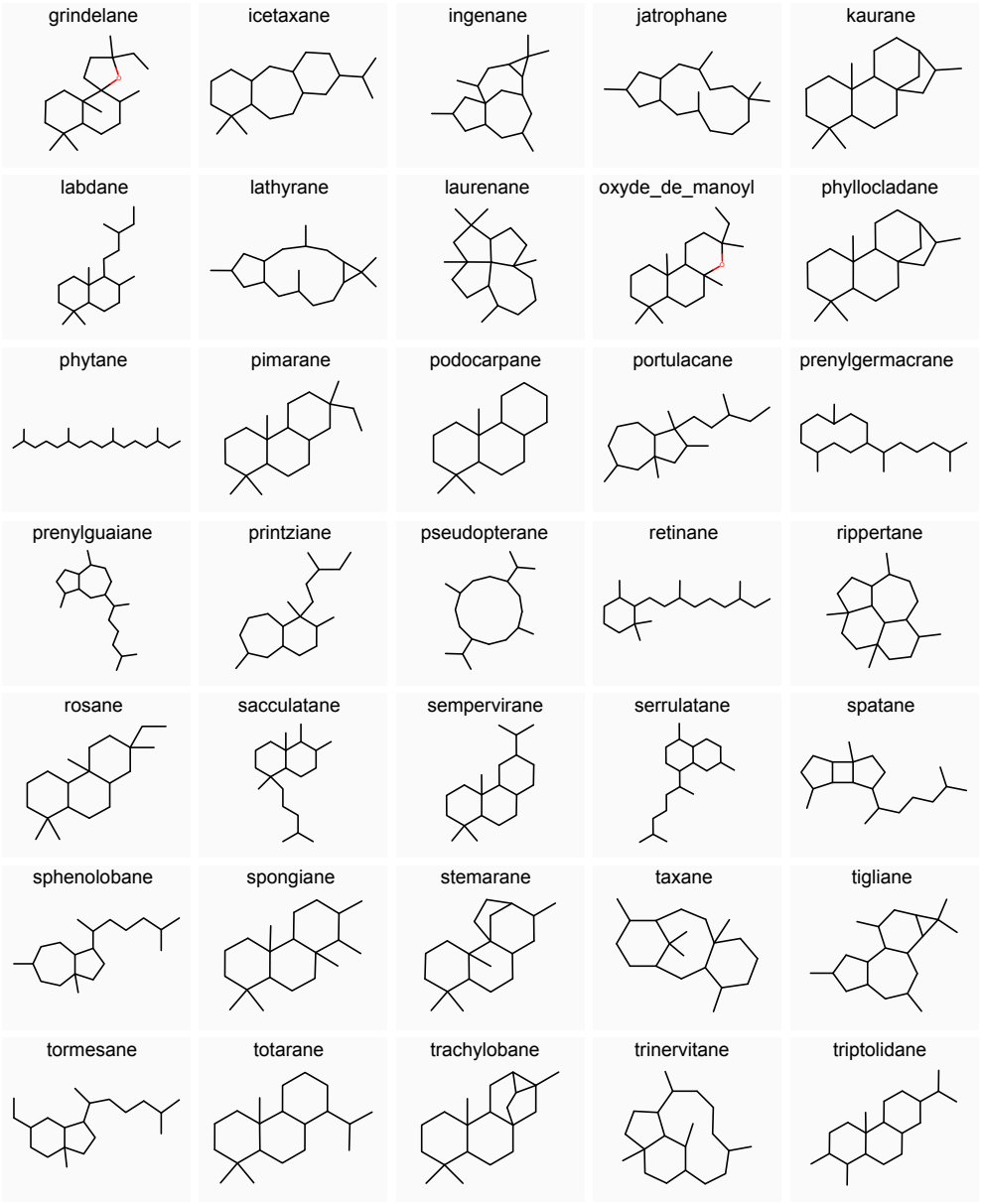
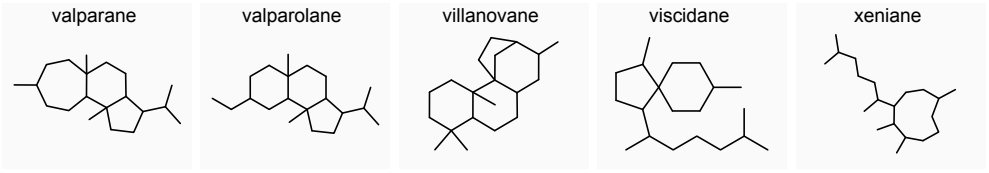


FIGURE C.15 – Squelettes de diterpènes (suite)



Annexe D

Liste des commandes pour le fichier d'entrée de CASA

Commandes basiques

Mnémonique de commande	Application	Description des paramètres
`DEPT` I V I R	Définition du statut des signaux RMN ^{13}C	– P1 : Numéro de signal RMN ^{13}C. – P2 : État d'hybridation ou liste d'hybridations entre parenthèses (1, 2 ou 3 pour les atomes *sp*, sp^2 et sp^3). – P3 : Multiplicité (0, 1, 2 ou 3 pour les atomes C, CH, CH_2 ou CH_3). – P4 : Déplacement chimique ^{13}C (optionnel).
`HSQC` I I	Corrélation hétéronucléaire à travers 1 liaison	– P1 : Numéro du carbone. – P2 : Numéro de l'hydrogène.
`COSY` V I O O	Corrélation COSY	– P1 : Numéro d'hydrogène ou liste d'hydrogènes entre parenthèses. – P2 : Numéro d'hydrogène. – P3 : Longueur de chemin de couplage optionnelle (limite inférieure). – P4 : Longueur de chemin de couplage optionnelle (limite supérieure).

Mnémonique de commande	Application	Description des paramètres
`HMBC` V I O O	Corrélation HMBC	– P1 : Numéro de carbone ou liste de carbones entre parenthèses. – P2 : Numéro d'hydrogène. – P3 : Longueur de chemin de couplage optionnelle (limite inférieure). – P4 : Longueur de chemin de couplage optionnelle (limite supérieure).
`INAD` V I	Corrélation INADE-QUATE	– P1 : Numéro de carbone ou liste de carbones entre parenthèses. – P2 : Numéro de carbone.
`ASGN` I I	Attribution d'un signal RMN ^{13}C à un atome de la molécule	– P1 : Numéro de signal RMN ^{13}C. – P2 : Numéro de l'atome de carbone correspondant.

Définition des propriétés des signaux RMN ^{13}C

Mnémonique de commande	Application	Description des paramètres
`LNMR` Ln S	Définition d'une liste de signaux RMN ^{13}C	– P1 : Numéro de la liste (n est compris entre 1 et 20). – P2 : Numéros des signaux RMN ^{13}C.
`LMOL` I I	Définition d'une liste d'atomes de la molécule	– P1 : Numéro de la liste (n est compris entre 1 et 20). – P2 : Numéro des atomes de la molécule.

Mnémonique de commande	Application	Description des paramètres
`PROP` B I Ln H	Définition de l'environnement des atomes	– P1 : Numéro d'un signal RMN ^{13}C ou numéro d'une liste de signaux RMN ^{13}C. – P2 : Nombre de voisins. La valeur 0 signifie tous les voisins. – P3 : Numéro de la liste d'atomes. – P4 : Signe optionnel (+ ou −).

Contrôle de l'exécution

Mnémonique de commande	Application	Description des paramètres
`ENTR` I	Affichage de l'interprétation des données du fichier d'entrée	– P1 = 0 : Pas d'affichage (défaut). – P1 = 1 : Affichage actif.
`WORK` I	Autorisation de la production des solutions	– P1 = 0 : Uniquement lecture et interprétation des fichiers d'entrée. – P1 = 1 : Résolution active (défaut).
`VERB` I	Verbosité	– P1 = 0 : Programme muet (défaut). – P1 = 1 : Verbosité active. – P1 = 2 : Programme très verbeux.
`STEP` I	Exécution pas à pas	– P1 = 0 : Exécution normale (défaut). – P1 = 1 : État actif.

Mnémonique de commande	Application	Description des paramètres
`CCLA` I	Contrôle de l'utilisation des classes d'équivalence des carbones	– P1 = 0 : Pas de classes. Toutes les attributions sont produites. – P1 = 1 : Les carbones symétriques sont équivalents (défaut).
`AWCS` I	Attribution avec prédiction des déplacements chimiques ^{13}C	– P1 = 0 : Attribution sans prédiction des déplacements chimiques (défaut). – P1 = 1 : Ajout de la prédiction des déplacements chimiques aux contraintes des spectres RMN 2D. – P1 = 2 : Utilisation uniquement des déplacements chimiques prédits pour l'attribution.
`CCSF` P	Fichier qui contient les valeurs de déplacements chimiques prédits	– P1 : Chemin d'accès au fichier contenant les déplacements chimiques prédits.
`SCLF` F	Facteur multiplicatif de l'erreur sur la prédiction	– P1 = 0 : Valeur du facteur multiplicatif.
`TOLE` F	Seuil de tolérance pour l'erreur sur la prédiction	– P1 : Valeur de la tolérance (5,0 ppm par défaut).
`ELIM` I I	Élimination des corrélations HMBC et/ou COSY invalides	– P1 : Nombre maximum de corrélations à éliminer. – P2 : Nombre maximum de liaisons admissibles pour une corrélation éliminée.

Annexe E

Spectres RMN du camphre

Figure E.1 – Spectre RMN ^{1}H et ^{13}C ($CDCl_3$, 500 MHz) du camphre

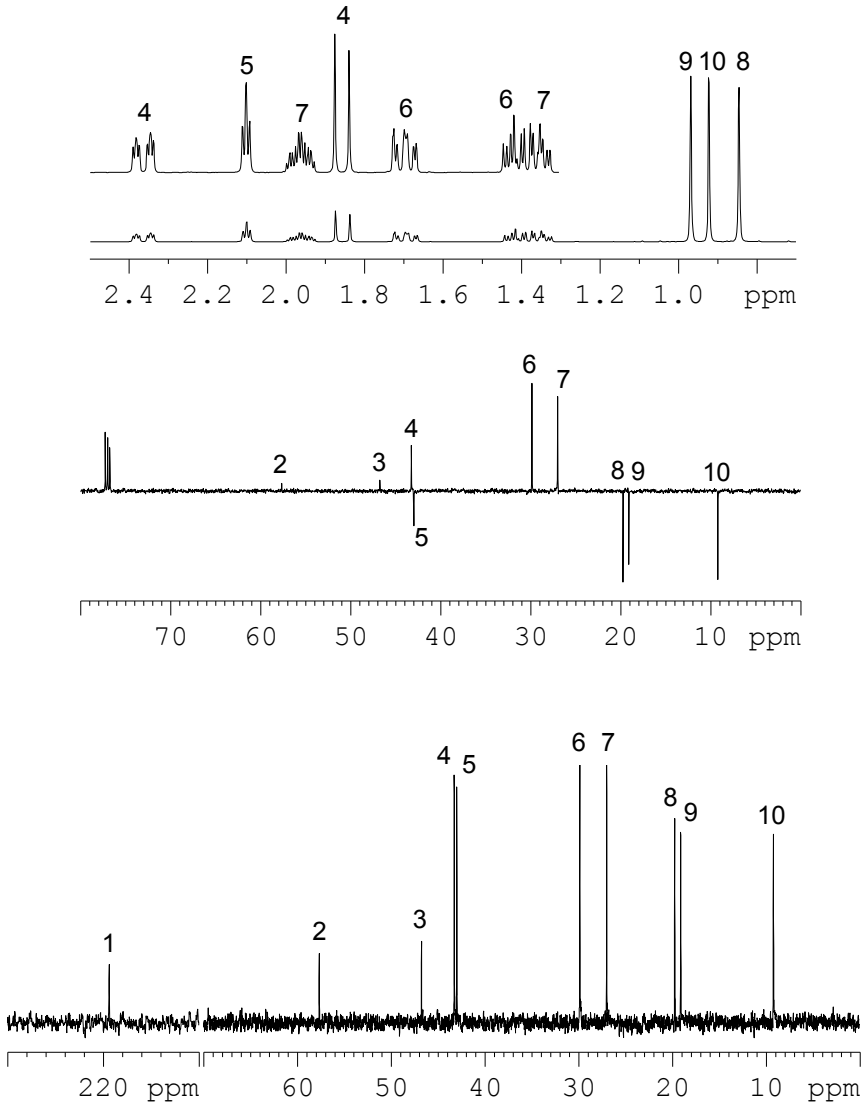


FIGURE E.2 – Spectre RMN COSY ($CDCl_3$, 500 MHz) du camphre

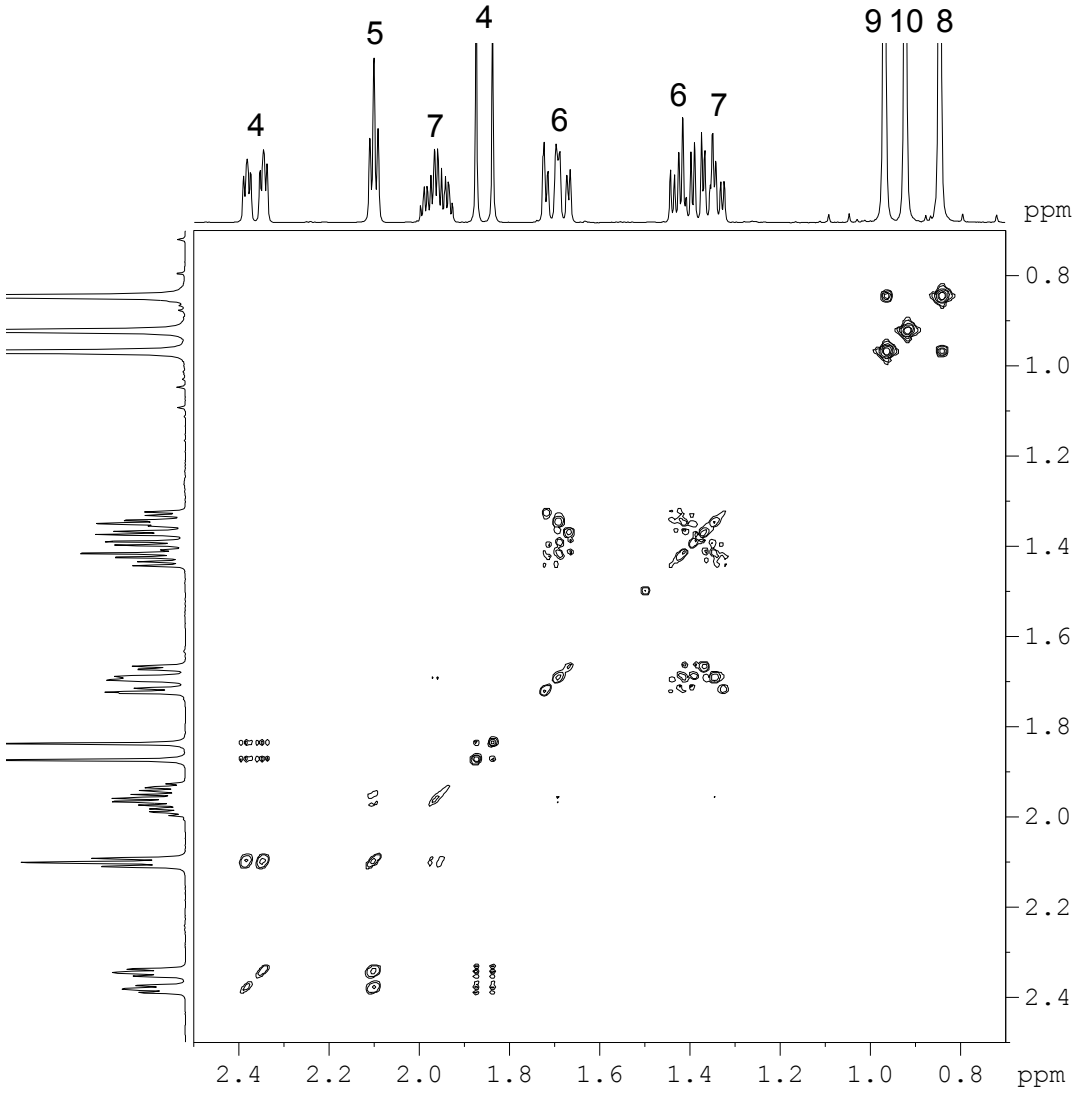
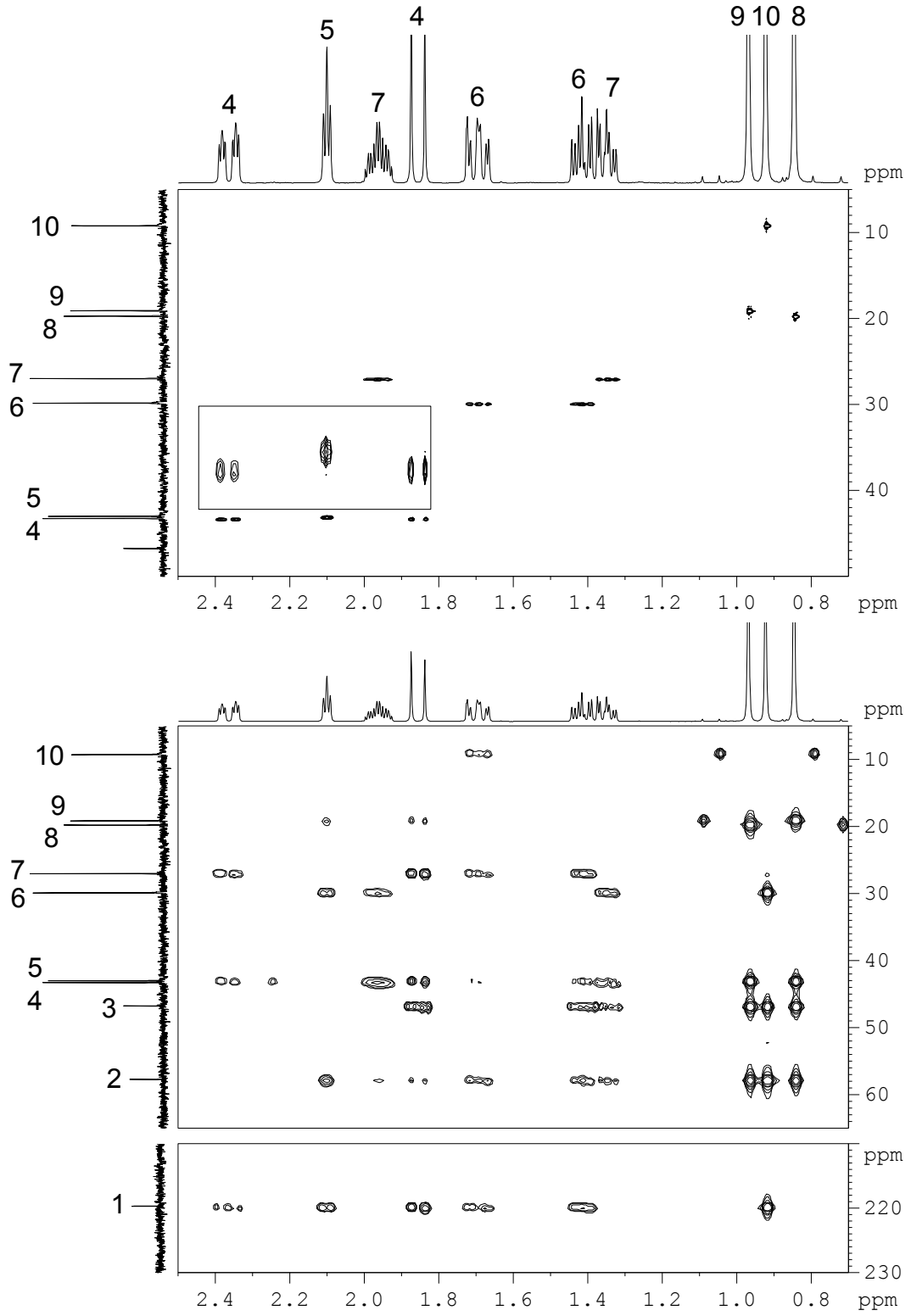


Figure E.3 – Spectre RMN HSQC (en haut) et HMBC (en bas) ($CDCl_3$, 500 MHz) du camphre



Annexe F

Fichiers d'entrée de CASA

Figure F.1 – Fichier CASA de la Guttiferone H

```
AWCS 1
CCSF "guttiferoneH_cstable.txt"
TOLE 10.0

DEPT 1 2 0 190.9       HSQC 6 6       HMBC 1 7       HMBC 24 6
DEPT 2 2 0 119.1       HSQC 7 7       HMBC 1 29      HMBC 24 7
DEPT 3 2 0 190.6       HSQC 12 12     HMBC 3 17      HMBC 25 27
DEPT 4 3 0 67.1        HSQC 15 15     HMBC 4 6       HMBC 25 28
DEPT 5 3 0 48.0        HSQC 16 16     HMBC 4 17      HMBC 26 27
DEPT 6 3 1 47.6        HSQC 17 17     HMBC 4 22      HMBC 27 28
DEPT 7 3 1 38.6        HSQC 18 18     HMBC 4 23      HMBC 28 25
DEPT 8 3 0 60.2        HSQC 20 20     HMBC 5 22      HMBC 28 27
DEPT 9 2 0 212.7       HSQC 21 21     HMBC 5 23      HMBC 29 37
DEPT 10 2 0 198.0      HSQC 22 22     HMBC 6 22      HMBC 29 38
DEPT 11 2 0 131.6      HSQC 23 23     HMBC 6 23      HMBC 30 29
DEPT 12 2 1 115.5      HSQC 24 24     HMBC 7 32      HMBC 30 31
DEPT 13 2 0 145.1      HSQC 25 25     HMBC 8 6       HMBC 30 32
DEPT 14 2 0 150.4      HSQC 27 27     HMBC 8 7       HMBC 30 37
DEPT 15 2 1 114.1      HSQC 28 28     HMBC 8 29      HMBC 30 38
DEPT 16 2 1 124.1      HSQC 29 29     HMBC 8 33      HMBC 31 29
DEPT 17 3 2 25.8       HSQC 31 31     HMBC 9 17      HMBC 31 32
DEPT 18 2 1 121.8      HSQC 32 32     HMBC 9 7       HMBC 31 35
DEPT 19 2 0 131.9      HSQC 33 33     HMBC 9 29      HMBC 31 36
DEPT 20 3 3 25.4       HSQC 35 35     HMBC 10 12     HMBC 31 37
DEPT 21 3 3 17.4       HSQC 36 36     HMBC 10 16     HMBC 31 38
DEPT 22 3 3 22.6       HSQC 37 37     HMBC 11 15     HMBC 32 33
DEPT 23 3 3 26.6       HSQC 38 38     HMBC 12 16     HMBC 33 32
DEPT 24 3 2 29.7                      HMBC 13 15     HMBC 34 31
DEPT 25 2 1 126.0                     HMBC 14 12     HMBC 34 32
DEPT 26 2 0 132.5      COSY 18 17     HMBC 14 16     HMBC 34 36
DEPT 27 3 3 24.5       COSY 24 25     HMBC 16 12     HMBC 35 31
DEPT 28 3 3 17.4       COSY 24 6      HMBC 18 17     HMBC 35 36
DEPT 29 3 2 40.2       COSY 6 7       HMBC 18 20     HMBC 36 31
DEPT 30 3 0 40.9       COSY 7 33      HMBC 18 21     HMBC 36 35
DEPT 31 3 1 49.4       COSY 33 32     HMBC 19 20     HMBC 37 31
DEPT 32 3 2 28.8       COSY 32 31     HMBC 19 21     HMBC 37 38
DEPT 33 3 2 32.9                      HMBC 20 21     HMBC 38 31
DEPT 34 2 0 146.7                     HMBC 21 20     HMBC 38 37
DEPT 35 2 2 108.3                     HMBC 22 23
DEPT 36 3 3 23.5                      HMBC 23 22
DEPT 37 3 3 25.1
DEPT 38 3 3 23.6


LMOL L1 1 2 3 4 5 8 9 10 11 13 14 19 26 30 34 ; quaternary carbons
LMOL L2 41 42 44 ; sp2 oxygens
LMOL L3 39 40 41 42 43 44 ; all oxygens

LNMR L6 20 21 22 23 27 28 36 37 38
LNMR L7 9 10
LNMR L8 1 3 13 14

PROP L6 1 L1 ; methyl groups bound to quaternary carbons
PROP 12 2 L1 ; H-12 is a doublet J = 1.8 Hz, has 2 quaternary
neighbors
PROP L7 1 L2 ; carbon atoms bound to sp2 oxygens
PROP L8 1 L3 ; carbon atoms bound to oxygens
```

Figure F.2 – Fichier CASA du triterpène

```
AWCS 1
CCSF "triterp_cstable.txt"
ELIM 4 4
SCLF 1.6

DEPT 1 3 2 38.7
DEPT 2 3 2 26.7
DEPT 3 3 1 78.6
DEPT 4 3 0 38.7
DEPT 5 3 1 55.6
DEPT 6 3 2 18.4
DEPT 7 3 2 33.3
DEPT 8 3 0 39.7
DEPT 9 3 1 47.6
DEPT 10 3 0 37
DEPT 11 3 2 23.5
DEPT 12 2 1 128.4
DEPT 13 2 0 138.6
DEPT 14 3 0 41.7
DEPT 15 3 2 28.3
DEPT 16 3 2 25.8
DEPT 17 3 0 47.8
DEPT 18 3 1 54.7
DEPT 19 3 0 72.4
DEPT 20 2 0 155.2
DEPT 21 3 2 28
DEPT 22 3 2 38.5
DEPT 23 3 3 27.6
DEPT 24 3 3 15.2
DEPT 25 3 3 14.8
DEPT 26 3 3 16.5
DEPT 27 3 3 23
DEPT 28 2 0 180
DEPT 29 2 2 105
DEPT 30 3 3 26.6

HSQC 1 1
HSQC 2 2
HSQC 3 3
HSQC 5 5
HSQC 6 6
HSQC 7 7
HSQC 9 9
HSQC 11 11
HSQC 12 12
HSQC 15 15
HSQC 16 16
HSQC 18 18
HSQC 21 21
HSQC 22 22
HSQC 23 23
HSQC 24 24
HSQC 25 25
HSQC 26 26
HSQC 27 27
HSQC 29 29
HSQC 30 30

COSY 1 3
COSY 5 6
COSY 11 12
COSY 15 16
COSY 21 29

HMBC 3 1
HMBC 3 1
HMBC 5 1
HMBC (1 4 22) 2
HMBC 10 2
HMBC (1 4 22) 3
HMBC (1 4 22) 3
HMBC 23 3
HMBC 24 3
HMBC (1 4 22) 5
HMBC 6 5
HMBC 7 5
HMBC (9 17) 5
HMBC 10 5
HMBC 23 5
HMBC 24 5
HMBC 7 6
HMBC (9 17) 7
HMBC 26 7
HMBC (1 4 22) 9
HMBC 8 9
HMBC 10 9
HMBC 11 9
HMBC 14 9
HMBC 25 9
HMBC 26 9
HMBC 8 11
HMBC (9 17) 11
HMBC 12 11
HMBC 13 11
HMBC (9 17) 12
HMBC 11 12
HMBC 14 12
HMBC 15 16
HMBC 12 18
HMBC 13 18
HMBC 16 18
HMBC (9 17) 18
HMBC 19 18
HMBC (9 17) 21
HMBC 20 21
HMBC (1 4 22) 21
HMBC 29 21
HMBC 16 22
HMBC (9 17) 22
HMBC 18 22
HMBC 20 22
HMBC 3 23
HMBC (1 4 22) 23
HMBC 5 23
HMBC 24 23
HMBC 3 24
HMBC (1 4 22) 24
HMBC 5 24
HMBC 23 24
HMBC (1 4 22) 25
HMBC 5 25
HMBC (9 17) 25
HMBC 10 25
HMBC 7 26
HMBC 8 26
HMBC (9 17) 26
HMBC 14 26
HMBC 8 27
HMBC 13 27
HMBC 14 27
HMBC 15 27
HMBC 19 29
HMBC 20 29
HMBC 19 30
HMBC 20 30
HMBC 29 30
HMBC 18 30
```

Résumé : Ce mémoire traite du développement d'outils informatiques d'aide à l'analyse structurale des petites molécules organiques par Résonance Magnétique Nucléaire (RMN). Il comprend deux volets axés sur l'automatisation de tâches auxquelles les chimistes des laboratoires de synthèse organique ou d'isolement de substances naturelles sont confrontés au quotidien, à savoir l'élucidation et la vérification de structure. Le premier volet concerne des améliorations apportées au logiciel de génération de structure LSD (Logic for Structure Determination). Ce logiciel est basé sur l'interprétation des corrélations des spectres de RMN 2D pour la détermination de structures complètement ou partiellement inconnues. Les progrès récents ont pour but d'augmenter la diversité des molécules analysables et d'améliorer le traitement des corrélations ainsi que la présentation des résultats. L'intégration de LSD avec la base de données SISTEMAT permet de bénéficier de la source d'information supplémentaire que constitue la valeur des déplacements chimiques. Cet avantage se traduit par un filtrage des solutions en fonction d'éléments de sous-structure sélectionnés dans une collection de squelettes de produits naturels. Le second volet présente le développement du logiciel CASA (Computer-Aided Spectral Assignment) dont le rôle est de réaliser une vérification automatique de structure par l'attribution des résonances. Il s'appuie sur des contraintes issues des spectres de RMN 2D et sur un module de prédiction des déplacements chimiques ^{13}C.
Mots clés : RMN 2D, élucidation structurale assistée par ordinateur, LSD, vérification automatique de structure, CASA

Abstract : This thesis deals with the development of computational tools for structural analysis of small organic molecules by Nuclear Magnetic Resonance (NMR). It consists of two parts that focus on the automation of tasks that chemists working in the fields of organic synthesis or natural substance isolation daily face, namely structure elucidation and verification. The first part reports improvements of the structure generation software LSD (Logic for Structure Determination). This software is designed for the determination of completely or partially unknown structures from the interpretation of 2D NMR correlation spectra. The present work aims to increase the diversity of the molecules that can be analyzed and to improve the processing of correlation data as well as the presentation of results. The integration of LSD with the SISTEMAT database introduces chemical shifts values as an additional information source. It allows the chemist to sort the possible solutions of a problem according to the presence of known natural product skeletons. The second part presents the development of the CASA (Computer-Aided Spectral Assignment) software as a tool for automatic structure verification through resonance assignment. It jointly makes use of constraints from 2D NMR spectra and from the matching between experimental and predicted ^{13}C chemical shifts.
Keywords : 2D NMR, computer assisted structure elucidation, LSD, automatic structure verification, CASA