# Development of numerical approaches for nuclear magnetic resonance data analysis

Viêt Dung Duong


Numéro National de Thèse : 2017LYSEN010

**THESE de DOCTORAT DE L'UNIVERSITE DE LYON**
opérée par
**l'Ecole Normale Supérieure de Lyon**

**Ecole Doctorale de Chimie de Lyon**

**Spécialité de doctorat : Chimie**

Soutenue publiquement le 04/05/2017, par :
**Viet Dung DUONG**

---

# Development of Numerical Approaches for Nuclear Magnetic Resonance Data Analysis

Développement des méthodes numériques pour analyser les données issues de la résonance magnétique nucléaire

---

Devant le jury composé de :

| | | |
|---|---|---|
| Prof. VRANKEN Wim | Université de Brussel | Rapporteur |
| Mme RINGKJOBING JENSEN Malene | Institut de Biologie Structurale | Rapporteur |
| Prof. ROUMESTAND Christian | Centre de Biochimie Structural | Examinateur |
| M. WALKER Olivier | Institut des Sciences Analytiques | Examinateur |
| M. HERRMANN Torsten | Institut des Sciences Analytiques | Directeur |

# ACKNOWLEDGEMENTS

I am very grateful to the committee members of my Ph.D. defense, Prof. Christian Roumestand, Dr. Olivier Walker, Dr. Torsten Herrmann, and particularly to my two "rapporteurs" Dr. Malene Ringkjobing Jensen and Prof. Wim Vranken for dedicating their time to review this thesis manuscript. It is my honor to have my work assessed by these two experts in the fields of bio-NMR and bio-computation.

The hardest part of these acknowledgements is to avoid forgetting someone. Given my many paths taken, I have had the pleasure of knowing, learning from and sharing experiences with a great number of interesting and magnificent people. Even ideas and talks sometimes seemed anodyne on the spot, with the benefit of hindsight, are essential to understand the evolution and the realization of this thesis. I ask the indulgence of those whose contributions are not nominally thanked below, knowing that it is unintentional.

I would like to express my greatest gratitude to the person who made this thesis happening. I am deeply indebted to my Ph.D. advisor Dr. Torsten Herrmann for his guidance, encouragement and support during my entire Ph.D. study and also for the preceding two years of my Master 1 & 2 studies that ultimately enabled me to conclude my dissertation and to pursue in parallel my study at the medical school in Lyon. There is no way I could have been in my last year of medical school, while being at the same time in the Ph.D. Chemistry program at the ENS Lyon without his indulgence and kindness.

It was Prof. Lyndon Emsley who introduced me to Dr. Torsten Herrmann and who made the simultaneous study of Ph.D. in Chemistry at the ENS Lyon and the medical school within my reach. And for that, I am forever grateful to Prof. Lyndon Emsley.

Speaking of the PhD-med school course, I am thankful to Dr. Benjamin Blaise and Dr. Clement Pontoizeau for their valuable advices in dealing with the double curriculum vitae. My sincere thanks also go to Professor Jerome Etienne, former Dean of

the Medical Faculty Lyon Est for his continuous support in my double-curriculum vitae undertaking.

I am grateful to the NMR metabolomics group members at the CRMN, both current and alumni, with whose I shared enjoyable times and valuable experiences: Benedicte, Sylvie, Anne (Fages), Houda, Aurelien, Tony, Claire, Gilles ... Special thanks are going to Elodie for her kindness and friendship.

I would also like to thank other current and past members of the CRMN: Audrey, Guido, Moreno, Anne, Lenaic, Pierrick, Tanguy, Andreas, Cecile, Jean-Nicolas, Paul, Emmanuel, Emeline, Aaron, Andrew, Camille, Ruben … for the joyful and helpful moments that I had during my time in the lab.

To my ENS alumni friends: Corentin, Fabrice, Mathieu, Alizee, Julien, Anh Thy, Laurianne… To my friends in med-school: Cedric, Raphael, Emeline, Estelle … To my friends at INSA Lyon: Lad, Thach, Thanh, Nam …

My thanks go also to the Ecole Normale Supérieure de Lyon (ENS Lyon) and the Centre National de la Recherche (CNRS) that provided me with an ideal research environment and enough funding for pursuing my research aims.

My thanks to my late grandfather Dang Van Voi. I am grateful to my sibling Hai, whose weekly Skype conversation always provides amusing and inspiring ideas. To my parents who are my eternal cheerleaders and who have provided me through moral and emotional support in my life. To Typhanie, with whom I shared countless joyful moments over the last two years.

I am not a very good, vivid and eloquent writer, but well, you probably know that by now. Looking to the future, I finish this part by quoting Rene Char, “impose ta chance, serre ton bonheur et va vers ton risque. A te regarder, ils s’habitueront.” (Impose your luck, hold tight your happiness, and go towards your risk. To look at you, they will get used to).

Dear reader, you have just finished reading the essential part of this thesis; what follows is somewhat technical.

# RESUME

La résonance magnétique nucléaire (RMN) est devenue une des techniques spectroscopiques les plus puissantes et polyvalentes de la chimie analytique avec des applications multiples dans des différents domaines de la recherche. Cependant, un des inconvénients majeurs de la RMN est le processus fastidieux d'analyse de donnée qui nécessite fréquemment des interventions humaines. Ces dernières font diminuer non seulement l'efficacité et l'objectivité des études mais également renferment les champs d'applications potentielles de la RMN pour les non-initiés. Par conséquent, le développement des méthodes computationnelles non supervisées se trouve nécessaire. Les travaux réalisés ici représentent des nouvelles approches dans le domaine de la métabolomique et de la biologie structurelle.

Le défi ultime de la RMN métabolomique est l'identification complète de l'ensemble des molécules constituant les échantillons biologiques complexes. Cette étape est vitale pour toute interprétation biologique. Dans la première partie de cette thèse, une nouvelle méthode numérique a été développée pour analyser des spectres à deux dimensions HSQC et TOCSY afin d'identifier les métabolites. La performance de cette nouvelle méthode a été démontrée avec succès sur les données synthétiques et expérimentales.

La RMN est une des principales techniques analytiques de la biologie structurale. Le processus conventionnel de détermination structurelle est bien établie avec souvent une attribution explicite des signaux. Dans la seconde partie de cette thèse, une nouvelle approche computationnelle est présentée. Cette nouvelle méthode permet de déterminer les structures RMN sans attributions explicites des signaux. Ces derniers proviennent de données spectrales tridimensionnelles TOCSY et NOESY. Les structures ont été résolues en appliquant cette nouvelle méthode aux données spectrales d'une protéine de 12kDa.

# ABSTRACT

Nuclear Magnetic Resonance (NMR) has become one of the most powerful and versatile spectroscopic techniques in analytical chemistry with applications in many disciplines of scientific research. A downside of NMR is however the laborious data analysis workflow that involves many manual interventions. Interactive data analysis impedes not only on efficiency and objectivity, but also keeps many NMR application fields closed for non-experts. Thus, there is a high demand for the development of unsupervised computational methods. This thesis introduces such unattended approaches in the fields of metabonomics and structural biology.

A foremost challenge to NMR metabolomics is the identification of all molecules present in complex metabolite mixtures that is vital for the subsequent biological interpretation. In this first part of the thesis, a novel numerical method is proposed for the analysis of two-dimensional HSQC and TOCSY spectra that yields automated metabolite identification. Proof-of principle was successfully obtained by evaluating performance characteristics on synthetic data, and on real-world applications of human urine samples, exhibiting high data complexity.

NMR is one of the leading experimental techniques in structural biology. However the conventional process of structure elucidation is quite elaborated. In this second part of the thesis, a novel computational approach is presented to solve the problem of NMR structure determination without explicit resonance assignment based on three-dimensional TOCSY and NOESY spectra. Proof-of principle was successfully obtained by applying the method to an experimental data set of a 12-kilodalton medium-sized protein.

# TABLE OF CONTENT

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 1D | One-dimensional |
| 2D | Two-dimensional |
| 3D | Three-dimensional |
| BMRB | Biological Magnetic Resonance Bank |
| CASD | Critical Assessment of NMR Structure Determination |
| CCPN | Collaborative Computational Project for NMR |
| DINO | Direct NOE Method |
| DNP | Dynamic Nuclear Polarization |
| Gryo-EM | Single-Particle Cryo Electron Microscopy |
| HM | Hybrid Methods |
| HMDB | Human Metabolome Database |
| HSQC | Heteronuclear Single Quantum Spectroscopy |
| Hz | Hertz |
| ISA | Isolated Spin Approximation |
| ITERAMETA | Iterative Metabolite Pattern Recognition |
| kDa | Kilo Dalton |
| MS | Mass Spectrometry |
| MD | Molecular Dynamics |
| NA | Nuclear Acids |
| NMR | Nuclear Magnetic Resonance |
| NOESY | Nuclear Overhauser Effect Spectroscopy |

| | |
|---|---|
| PDB | Protein Data Bank |
| PPM | Parts Per Millions |
| rMD | Restraint Molecular Dynamics |
| RMSD | Root Mean-Square Deviation |
| SA | Simulated Annealing |
| TOCSY | Total Correlation Spectroscopy |
| UNIO | Software suite for NMR data analysis |

# FOREWORD

This thesis is written for obtaining a Ph.D. in Chemistry at the Ecole Normale Supérieure (ENS) de Lyon, France. The aim of my doctoral research, carried out at the "Institut des Sciences Analytiques" (ISA Lyon), was to develop novel numerical approaches for accurate and robust automated Nuclear Magnetic Resonance (NMR) data analysis in the fields of NMR metabonomics and NMR protein structure determination.

The thesis is structured in four parts. Chapter 1 provides a general introduction to NMR spectroscopy by focusing on recent progress and identifying remaining challenges in the fields of NMR metabonomics and NMR structural biology. Chapter 2 describes current barriers for exhaustive NMR metabolite identification and introduces our software solution ITERAMETA that enables automated metabolite profiling from two-dimensional heteronuclear HSQC and homonuclear TOCSY spectra. Chapter 3 gives a general introduction to NMR protein structure determination and then details our novel numerical method that attempts to introduce protein NMR structure determination without prior sequence-specific resonance assignment ("NMR resonance-free structure determination"). The final Chapter 4 concludes the thesis and discusses some future perspectives of the results obtained.

# CHAPTER 1

## 1. General Introduction

The extraordinary progress and witnessed numerous breakthroughs in the fields of Molecular Biology in the last decade are mainly due to discoveries from projects on human and model organisms. A large part of the success of these projects was achieved with the help of Computational Biology allowing reduction in both experimental time and the time needed to analyze the experimental data.

There are about 127000 structures (mainly proteins (93%), nuclear acids (NA) (2%) and protein/NA complexes (5%)) in the Protein Data Bank (PDB) as of end of February 2017 (http://www.rcsb.org/pdb/statistics/holdings.do). That is a considerable advance from about 55 000 deposited experimentally determined structural models of macromolecules as of end of 2008. Nuclear Magnetic Resonance (NMR) spectroscopy was used to determine about 10% of these biomolecular structures, compared to 90% done by X-ray crystallography diffraction and a relatively minor contribution of depositions was using other methods (Electron Microscopy (EM) (1%), Hybrid Methods (HM) (0.1%) and others (0.1%)). Although the number of protein structures determined by X-ray crystallography diffraction is still prevailing, NMR protein structure determination keeps playing an important role in Structural Biology, thanks to a number of underlying experimental benefits that allow the study of the structure-function relation by explicitly including dynamical aspects. Notably, it is also worthy to mention that the above-mentioned percentages of deposited PDB structures divided by different experimental techniques is quite dramatically changing, if one filters the deposited PDB entries by high (90%) sequence identity, in particular by excluding multiple PDB entries of macromolecular structures determined at different X-ray atomic resolution. Regardless of such statistical analysis of deposited experimental three-dimensional (3D) structures in the PDB, automating the NMR protein structure determination process has become a substantial part of ongoing research in order to bridge the gap between the requirements of high-throughput structural genomics/proteomics projects and the tremendously

laborious manual data analysis process involved. Our work in this thesis addresses one of the remaining bottlenecks of automated NMR protein structure determination: the task of sequence-specific resonance assignment that we propose to suppress by the process of so-called "NMR (sequence-specific) assignment-free" structure determination that has the potential to become an attractive alternative to the conventional strategy commonly used in NMR structure determination (Wuthrich, 1986).

In recent years, there is also a growing interest of applying NMR to study human and model metabolism. Metabolomics, or the science of metabolism, is becoming nowadays a regular tool to study the homeostatic responses of biological systems due to intrinsic or extrinsic stimuli and/or perturbations. Metabolites are small biochemical compounds that are found in bio-fluids (urines, blood, plasma or serum). The annotation of mixture composition remains however time-costly, burdensome and frequently subjected to (human) experience bias. With the growing of high-quality public metabolite databases, *e.g.*, such as the Human Metabolite Database (HMDB) or the metabolomics Biological Magnetic Resonance Bank (BMRB), automated metabolite profiling by direct search and matching against catalogued metabolites becomes feasible and even necessary for obtaining a comprehensive compound identification, as it will be demonstrated in the work presented here.

Before going into further details, we feel the need to present some general NMR principles and an overview of computational methods in both fields, metabolomics NMR and protein NMR.

## 1.1 General Principles of Nuclear Magnetic Resonance (NMR)

The physical principle of Nuclear Magnetic Resonance (NMR) is that when commonly detected NMR-active nuclei such as $^{1}H$, $^{13}C$ and $^{15}N$ are placed in a strong magnetic field, these nuclei absorb energy at a characteristic resonance frequency (Larmor frequency). This characteristic frequency is a fundamental function of the nuclei's local chemical and geometric environment; hence, even chemically identical

nuclei usually resonate at different resonance frequencies what ultimately makes NMR spectroscopy such a powerful probe for structural and dynamical studies. The resonance frequency of a nucleus (measured in Hz) can be transformed into a magnetic field-independent variable, called *chemical shift*, via a mathematical formulation called the *Fourier transformation*. The chemical shift parameter is computed as the relative variation of the nucleus's resonance frequency and its standard value (a NMR reference standard). In order to obtain a magnetic field independent parameter, it is commonly indicated in parts per million (ppm) relative to a reference agent.

This very fundamental NMR parameter, the chemical shift values of individual nuclei, is one of the two main types of inputs employed in our numerical methods developed here, the other being the spatial and covalent correlations between nuclei that can be measured by NMR. The design and application of a set of specific pulse sequences can bring very specific structural information about the mixture composition or 3D structure present in the biological sample under investigation. The NMR acquisition for metabolomics and protein structure determination that are explored in our projects rely on routinely used pulse sequences developed to reveal exactly such correlations between atom pairs in small molecules (metabolite) or macromolecules (proteins).

In the first part of this thesis in order to achieve accurate and complete metabolite identification, we propose that for future metabolomics projects, standard two-dimensional homonuclear $\left({}^{1}H-{}^{1}H\right)$ and heteronuclear $\left({}^{13}C-{}^{1}H\right)$ correlation spectroscopy are acquired, such as the TOCSY (TOtal Correlation SpectroscopY) correlating the total (proton) spin system of a compound, and ${}^{13}C-{}^{1}H$ HSQC correlating the covalently connected carbon and proton pairs.

In the second part of this thesis in order to determine protein NMR structures, we propose that for future protein projects, only standard aliphatic and aromatic ${}^{13}C-\left({}^{1}H-{}^{1}H\right)$ -NOESY and ${}^{15}N-\left({}^{1}H-{}^{1}H\right)$ -NOESY and heteronuclear ${}^{13}C-\left({}^{1}H-{}^{1}H\right)$ TOCSY spectra are acquired. We believe that this minimal set of multidimensional 3D NMR spectra may provide sufficiently valuable correlations between NMR-active nuclei (the correlations could be either intra, inter-residue or long-

range) that can be profitably used to determine the fold of a protein without prior sequence-specific resonance assignment, thus replacing the conventional, laborious NMR structure determination process.

## 1.2 NMR Metabolomics

Metabolomics is a steadily growing field of biological sciences measuring the metabolic response of organisms or living systems to internal and external stimulus (Ellinger et al. 2013; Nicholson and Lindon 2008). It is the field of "omics" science concerned with the identification and/or quantification of small molecules called metabolites found in cells, tissues, bio-fluids and organisms (Brown et al. 2005). Metabolomics is often defined as the characterization of the ensemble of resulting metabolites as a phenotyping tool (Goodacre et al. 2004).

Metabolomics covers a wide variety of research fields, such as nutrition science, oncology, disease biomarker discovery and diagnosis, drug development, food quality control, just to name a few. It employs a variety of analytical chemistry techniques to precisely identify metabolites or generate metabolic spectral profiles. Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) are two most popular and standard spectroscopic techniques used for metabolomics studies.

Within recent years thanks to NMR sensitivity-related technology advances - increased magnetic field strengths, gyro-probe detection and very recently dynamic nuclear polarization (DNP) sensitivity enhancements - NMR has developed into a standard, routinely applied technique to identify and quantify metabolites that are present in biogenic samples under or close to physiological conditions. An indispensable prerequisite for any quantitative NMR (equally valid for MS) metabolomics study is the accurate and complete identification of all compounds present in a complex metabolite mixture, with biological sample diversity ranging from bio fluids, intact cell or tissue extracts to whole organisms. The fundamental key for metabolite identification by NMR is its high-resolution fitness that allows the simultaneous detection of a wide range of

different types and hundreds of metabolites in complex biological samples without the need for extensive sample preparation and hyphenation techniques. Notably, NMR spectroscopy as an untargeted, quantitative, reproducible, non-destructive and unbiased spectroscopic technique for metabolomics provides versatile information that is in principle readily amenable to pattern recognition methods for rapid, high-throughput identification of catalogued metabolites. Due to these exclusive benefits in comparison to Mass Spectrometry (MS), NMR plays a vital role in modern medical research in order to detect meaningful disease biomarkers, as well as to explore metabolic pathway within living organisms.

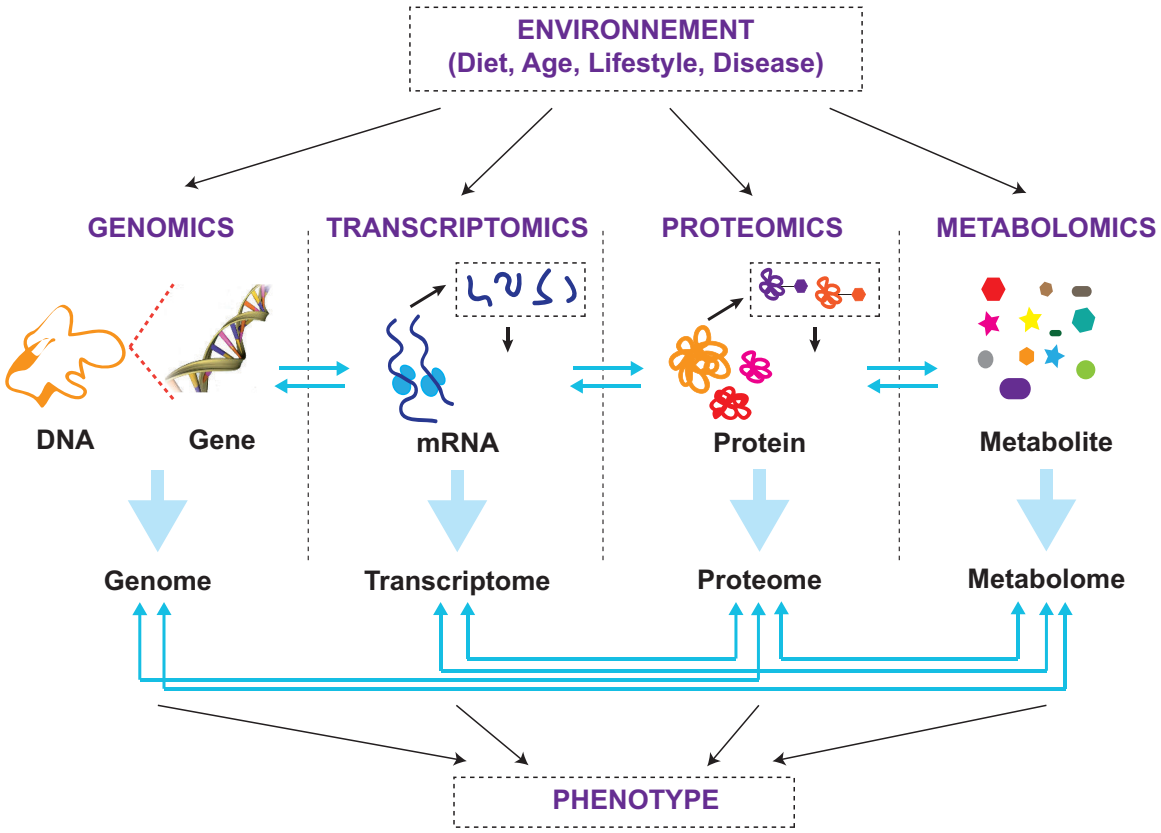



**Figure 1.1.** *The hierarchical classification of "omics" sciences.*

The challenge in metabolite profiling lies in the time and effort needed to identify and quantify compounds in bio-fluid mixtures. The major bottleneck of high-throughput NMR-based metabolomics remains metabolite signal assignment, since manual tasks are

arduous and time-consuming. For example, recent manual intensive efforts to identify metabolites in NMR urine samples yielded over 200 metabolites (Bouatra et al. 2013) that have been assigned by an expert spectroscopist using several men-weeks in interactive spectral analysis. Thus, there is obviously an urgent need and high desire to develop a robust numerical approach for accelerating this ultimate prerequisite for any quantitative NMR metabonomics study.

In principle, NMR is readily amenable to pattern recognition methods for rapid, high-throughput identification of catalogued metabolites. Under the plausible hypothesis that the chemical environment of individual molecules remains virtually unaffected in a complex mixture, and thus corresponds closely to its isolated, pure state, a NMR spectrum of a metabolite blend can be disentangled through its interpretation as a linear combination of the NMR spectra of its pure compounds. Hence, spectral NMR analysis for identifying *a prior* unknown individual constituents in a biological sample is commonly and most efficiently achieved by querying of NMR metabolomics databases (untargeted profiling of metabolites), which for each catalogued metabolite entry provides information about its covalent, chemical structure and a corresponding list of NMR resonance frequency (chemical shifts) of the molecule's atoms. Several public and commercial databases are available for this purpose, *e.g*., such as HMDB (Wishart et al. 2007, 2009, 2013, 2016), Metabolomics Biological Magnetic Resonance Database or BMRB (Markley 2012; Markley et al. 2007).

In daily practice though, the task of metabolite identification in biological complex samples remains still a quite cumbersome, time-consuming process and imposes a frustrating barrier for robust and efficient analysis of endogenous and exogenous metabolites for biomarker discovery. This is so because of the combination of the following five key challenges which may create difficulties for manual and/or automated data analysis based on direct database searching: (i) A simple direct matching between the mixture spectrum and a library of reference spectra recorded for pure metabolites can only be done, if the experimental conditions are closely similar. Even minor and practically unavoidable variations in temperature, pH and buffer conditions (salt concentration, overall dilution and relative concentration of specific ions), or external

magnetic field may change the absolute resonance frequencies, and most troubling alter the relative position in the connectivity pattern of a metabolite in a complex mixture, due to the fact that not all NMR nuclei-active atoms are affected by the same extend. Systematic signal shifts can be theoretically accounted for by applying internal or external re-referencing prior to direct database searching. In practice however, experimental conditions are too diverse in order to achieve exact matching with the controlled conditions used for recording the compounds in the reference libraries. Unsystematic signal shifts are much harder to deal with. Both types of fundamentally unavoidable NMR signal shifts may obscure metabolite identification and commonly demands in turn the use of a priory unknown, thus arbitrarily chosen chemical shift tolerance windows for pattern matching. (ii) Due to experimental imperfections, a key struggle is posed by missing signals in the NMR spectrum. This means that one can usually not expect a perfect, complete matching between a spin pattern of a metabolite in a complex mixture and the one of a single compound in the reference library, recorded under controlled conditions and in high concentration. (iii) In real-world metabolomics applications, severe NMR signal overlap may further complicate the accurate documentation of metabolites present in a complex mixture due to ambiguities in NMR signal assignment. (iv) The NMR detection limit of metabolite concentration (typically measured at natural abundance) is of the order of micromolar and depends on various factors such as external magnetic field strength, chemically equivalent protons of a molecule contributing to a NMR signal and the crowdedness of the spectral region. This implies that the manually or automatically generated input peak list contains NMR signals close to the signal-to-noise ratio of the spectrum, meaning that pattern matching may be further troubled by the presence of erroneous signals in the input data. (v) Numerous public and commercial metabolite databases are available for direct database searching and matching. Unfortunately, a currently foremost problem in the NMR metabolomics field is the fact that the archiving formats exhibit a large, non-universal diversity in metabolite documentation (file format diversity), and also the access to most customized databases is only granted by their own query algorithms.

In the current context of metabolite identification, one-dimensional (1D) $^{1}H$ spectroscopy with multivariate statistical analysis has become one of the main and standardized routine analytical tools in NMR metabolomics. Despite good sensitivity and quick acquisition time, 1D proton spectra have one major disadvantage: the high degree of spectral overlap due to the high complexity of the experimental mixture associated with the low spectral dispersion of proton resonances. Manual spectral deconvolution, *i.e.*, determination of individual proton 1D spectra of known compounds present in the sample, provides a possible way to overcome the signal overlap problem. While 1D automated methods for metabolite profiling are quite provided: proprietary programs like AMIX, dataChord Spectrum Miner and free-of-charge programs like MetaboLab (Ludwig and Günther 2011), Automics (Wang et al. 2009), MetaboAnalyst 2.0 (Xia et al. 2012), 1D spectrum data analysis remains critically subjected to errors due to severe peak overlaps that disfavor significantly automated metabolite identification. Consequently, recent research activities in this field have shifted towards achieving such accurate and complete coverage of metabolite annotation in biological samples by exploring more sophisticated multidimensional NMR experiments.

Increasing spectral dimensionality is probably the best way to efficiently reduce spectral overlap (Aue, Bartholdi, and Ernst 1976) and moreover offers the great prospective for robust and confident metabolite annotation. Hence, the use of two-dimensional (2D) NMR for metabolite identification gained increasing traction during the past recent years, demonstrating an easier and more reliable identification of biomarkers than achievable with 1D spectra. Most widely used two-dimensional NMR experiments in NMR metabolomics are 2D heteronuclear $^{1}$H-$^{13}$C single quantum correlation (HSQC) and 2D homonuclear $^{1}$H-$^{1}$H total correlation spectroscopy (TOCSY). Such 2D NMR experiments provide a two-fold benefit for metabolite identification. The introduction of an additional indirect spectrum dimension leads first to a substantially increased separation of the NMR signals and secondly allows designing NMR experiments that provide covalent connectivity information of the nuclei belonging to the same molecule. These essential advantages for the subsequent data analysis are clearly outweighing the 2-3 folds longer measurement time needed. Obviously, higher dimensional (3D or more)

NMR experiments providing further improved signal dispersion and thus significantly less signal overlap would be best suited for metabolite identification, but the associated restraint of considerably longer acquisition times and foremost sensitivity issues (NMR measurement at natural abundance) limit their routine applications.

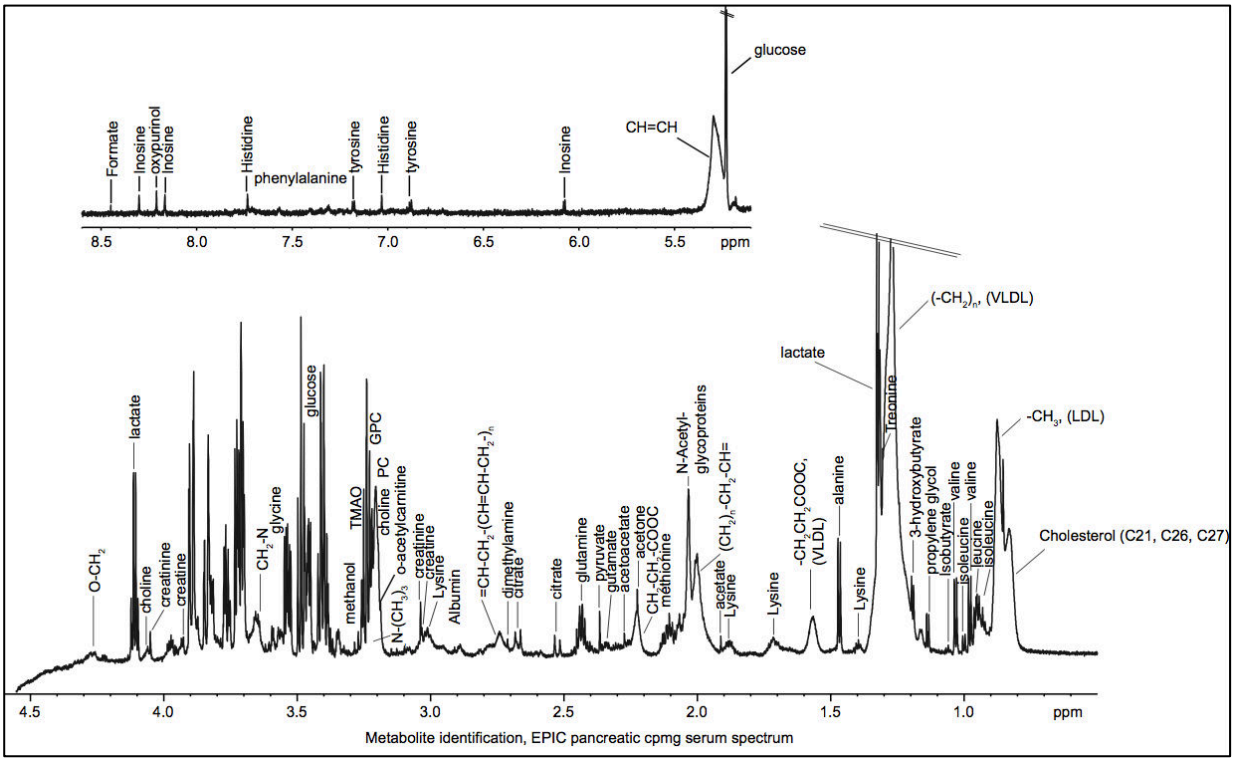



**Figure 1.2.** *One-dimensional spectrum with assigned compounds, issued from EPIC project.*

In 2D $^{1}$H-$^{13}$C HSQC spectra, the experimental cross-peak pattern is formed by all pairs of coupled $^{1}$H-$^{13}$C moieties separated by one covalent bond belonging to the same molecule. Although the resolution of NMR signals is increased by the introduction of the indirect $^{13}$C-dimension, a drawback of this NMR experiment is that connectivity information between the different $^{1}$H-$^{13}$C groups is still absent and complicates unambiguous metabolite identification, much similar to the case of 1D NMR analysis.

More reliable peak annotation and metabolite identification can be achieved using 2D $^{1}$H-$^{1}$H TOCSY. If TOCSY spectra are recorded with sufficiently long mixing times

(isotropic mixing), connectivity information between all nuclei of the spin system becomes accessible. A 2D $^{1}$H-$^{1}$H TOCSY pattern of a metabolite is accordingly formed by all directly and indirectly coupled hydrogen atoms, *i.e.*, that each cross-section of a given proton corresponds to the 1D spectrum of the whole spin system. The advantage of this experiment lies exactly in this inherent redundancy of the connectivity pattern that can be used to obtain accurate cross-peak assignment and metabolite identification even in the presence of strong signal overlap and/or spectral imperfection.

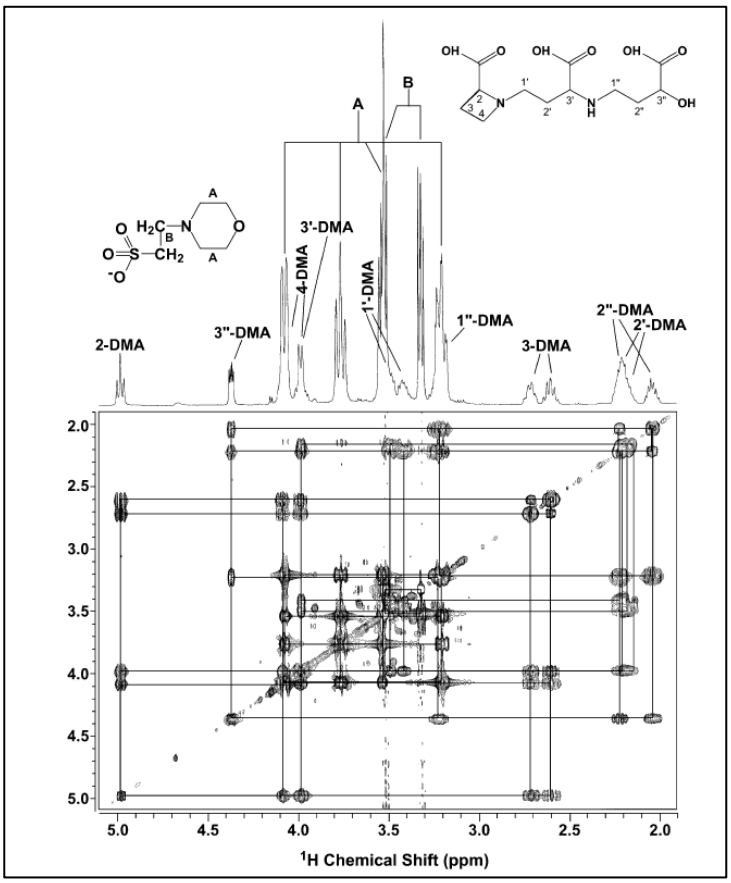



**Figure 1.3.** *Homo-nuclear correlation between 1H NMR signals in the two-dimensional TOCSY spectrum of a wheat exudate (Krishnan, Kruger, and Ratcliffe 2005).*

In the last 10 years or so, several different algorithms for metabolite identification have been developed to identify metabolites using two-dimensional TOCSY/HSQC spectra: semi-automated methods such as MetaboMiner (Xia et al. 2008), Collaborative

Computational Project for NMR (CCPN) Metabolomics Project (Chignola et al. 2011), trace analysis (Bingol and Brüschweiler 2011; Bingol et al. 2014; Robinette et al. 2008), deconvolution methods like Newton (Ellinger et al. 2013), COLMAR based on covariance spectroscopy (Zhang et al. 2008, 2009) and heuristic likelihood search (Xi et al. 2006). These numerical approaches have all their advantages and disadvantages, but notably and most frustrating their use is commonly restricted to their own limited, customized metabolite databases, instead of taking profit of the available complete knowledge that is publically assessable and compiled in complementary repositories, such as HMDB (Wishart et al. 2013) and BMRB (Markley et al. 2007). Despite this general drawback of currently proposed numerical methods imposed by operating on relatively incomplete (home-made) metabolite repositories, further drawbacks are that - generally speaking - all strategies in common are that the manually or automatically detected NMR signals are compiled into a single list of NMR cross-peaks (peak list). This peak list is then subjected to a database searching using direct pattern recognition algorithms which returns a listing of identified compounds. Several reference metabolite databases and querying algorithms have been developed for this purpose over the recent years. The proposed methods are not only showing great variability in their performance and different coverage of catalogued metabolites, but also impose strong demands on the input quality for proper operation that complicate and ultimately limits their practical use. Generally speaking, most strategies share the idea to identify potential metabolites based on the following two criteria: First, a cross-peak in the reference connectivity pattern is labelled as “confirmed”, if it matches with an input cross-peak within a user-given maximal frequency difference (threshold parameter 1). Second, a minimal ratio of the confirmed to the total number of cross peaks in reference pattern is imposed as acceptance criterion (threshold parameter 2). The first parameter is introduced to account for variations of experimental conditions between the complex mixture and the reference spectrum, while prior re-referencing is still required. The second parameter is necessary in order to deal with spectral imperfections. It is obvious that pattern recognition algorithms that rely only on these two parameters can hardly differentiate between true positives and false positives. This is so because the optimal numerical values of the two threshold parameters is *a priori* unknown and sample-dependent, the optimal numerical

values may even differ from metabolite to metabolite resp. may be dependent on the total number of signal in the connectivity pattern. Therefore additional control parameters were introduced that take the "uniqueness" of peak annotation into account. However despite this recent progress, many issues remain to be addressed before numerical methods can become an accurate and robust tool for metabolite identification.

In chapter 2, we address these issues in order to overcome current weaknesses in NMR metabolomics data analysis. We describe a suite of numerical routines implemented in a software suite called ITERAMETA (Iterative Metabolite Pattern Recognition) for metabolite identification in 2D homonuclear and heteronuclear NMR spectra, employing the HMDB and BMRB databases as query reference libraries. Various utility tools for assessment and validation of the results are also presented. Since our numerical method for iterative metabolite (ITERAMETA) pattern recognition is laid out to deal with 2D $^{1}$H-$^{13}$C HSQC and 2D $^{1}$H-$^{1}$H TOCSY spectra, the results of the TOCSY data can be easily used to validate and correct the results obtained for the HSQC data that is inherently prone to be less discriminative concerning true and false positives.

## 1.2 Protein NMR Structure Determination

Proteins are the basic *functional* building blocks of living organisms. They play the role of enzymes and hormones regulating metabolism, creating structures such as muscle and antibodies.

There are commonly four hierarchical levels used to describe or analyze a biological macromolecule. The first level is its primary structure, which is simply the protein sequence. One can see the primary structure of a biomolecule as the linear chain of its successively ordered amino acids, commonly called *residues*; there are 20 of these. The twenty natural amino acids share a common structure with a central alpha carbon, an amino group and a carboxyl group; their distinctive side chain groups differentiate them. The second level is the secondary structure divided into three main classes of secondary structure elements: alpha helix, beta sheet and coil. These sub-structure elements are

stabilized by inter-residue hydrogen bonds, and each individual secondary structure is characterized and defined by their hydrogen bond network in term of residue distance, such as for example the alpha helix is held in place by hydrogen bonds between residue $i$ and $i + 4$, while beta sheet is sustained by hydrogen bonds between the beta strands that are much longer in terms of residue distance. The third level is the tertiary structure that is commonly referred to as three-dimensional (3D) structure or fold of the protein. The fourth level, quaternary structure is the relative orientation or package of individual folded protein molecules into a multi-domain complex formed by multiple either identical or different sub-unit of 3D structures.

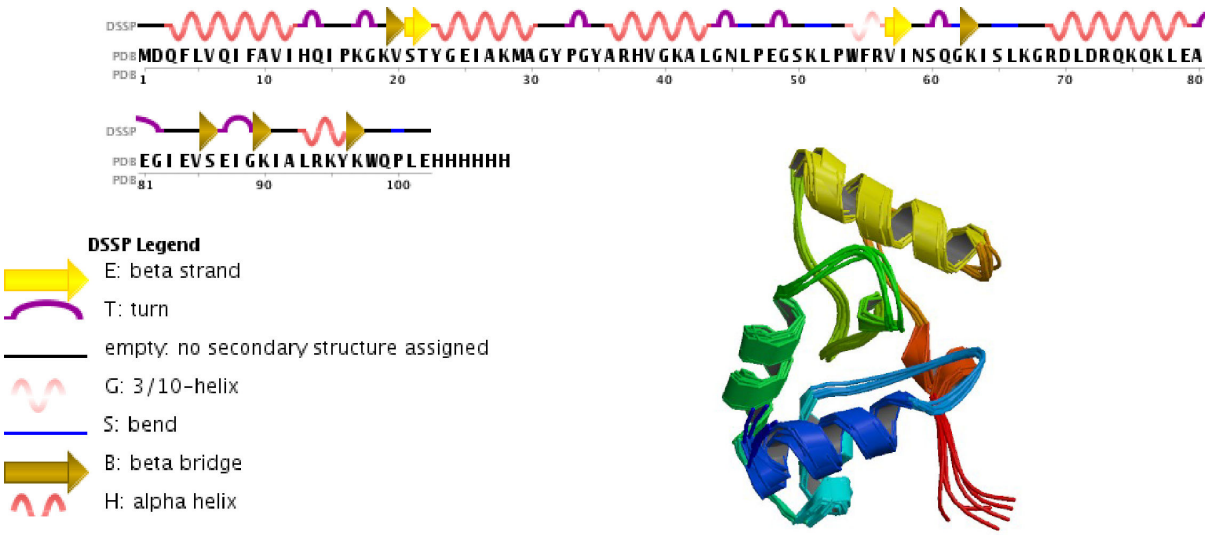



**Figure 1.4.** *The primary structure of our protein model VpR247 (PDB: 2KIM) is presented as a chain of one letter abbreviation for the different amino acids present in the protein sequence. The corresponding secondary structure elements are depicted above the primary sequence. The legend for the different structure elements and the three-dimensional NMR structure represented as bundle of conformers is shown on the left and right hand bottom part of the figure, respectively.*

Despite recent progresses made by other complementary experimental techniques for Structural Biology - notably witnessing the significant progress achieved by near atomic-resolution gryo electron microcopy (gryo-EM) - NMR and X-ray crystallography

still remain the two principal experimental techniques used to study three-dimensional molecular structures of proteins or protein complexes at atomic resolution. The advantage of NMR is that it is the only experimental technique that allows the determination of three-dimensional (3D) structures of protein molecules in aqueous solution, *i.e.*, close to physiological conditions. Another major strength of NMR is its power to complement the static picture of a protein structure with information about kinetic and dynamic properties of a protein or a macromolecular assembly. However, a drawback of NMR spectroscopy is that routine application is typically limited to small and medium-seized proteins with a molecular weight up to 20-25 kilo Dalton (kDa) at the best. Recent progress in solid-state NMR showed the theoretical possibility to study much large complexes, but the intrinsic problem of increased and unavoidable NMR signal overlap can also not be overcome by this alternative technique.

Three-dimensional structure determination of a protein by NMR involves conventionally the preparation of the protein sample, the measurements of a set of two-dimensional (2D) and three-dimensional (3D) NMR experiments, NMR data processing, NMR signal identification (peak picking), sequence-specific resonance assignment, NOE assignment, structure calculation, structure refinement in explicit water, and structure validation (Wüthrich, 1986). A variety of sophisticated automated approaches have been introduced targeting individual parts of the NMR structure determination process, and excellent review articles about automated NMR data analysis have been published (Güntert, P., 2003, Altieri, A. S., et al., 2004, Güntert, P., 2009, Markwick, P. R. L. et al. , 2008, Guerry et al., 2012). While great progress has been achieved for automation of the process of NOE assignment (Guerry et al., 2012), the preceding task of sequence-specific resonance assignment in conventional NMR structure determination still remains a bottleneck, despite the hundreds of approaches proposed (Guerry et al., 2012). Since in this thesis, we intend to introduce "NMR (sequence-specific) assignment-free structure determination" in order to overcome the barrier currently imposed by sequence-specific resonance assignment for efficient NMR structure determination, the following paragraphs are

mainly focusing on the data that can be extracted from Nuclear Overhauser Effect (NOE) spectroscopy.

In conventional *de novo* three-dimensional (3D) structure determinations of proteins by NMR spectroscopy, the key conformational data are obtained from upper distance limits derived from the Nuclear Overhauser effects (NOEs). NOEs result from cross-relaxation due to the dipole–dipole interactions between nearby pairs of nuclear spins in a molecule undergoing Brownian motion, and in two-dimensional (2D) or higher-dimensional heteronuclear-resolved [$^{1}$H,$^{1}$H]-NOESY spectra they are manifested by NOE cross-peaks. These NOEs are translated into a dense network of lower and upper (unambiguous or ambiguous) distance restraints that can be subsequently used to determine the fold of a protein via restrained molecular dynamics (rMD) using a simplified hybrid force field in Cartesian or torsion angle space (Guentert et al., 1997).

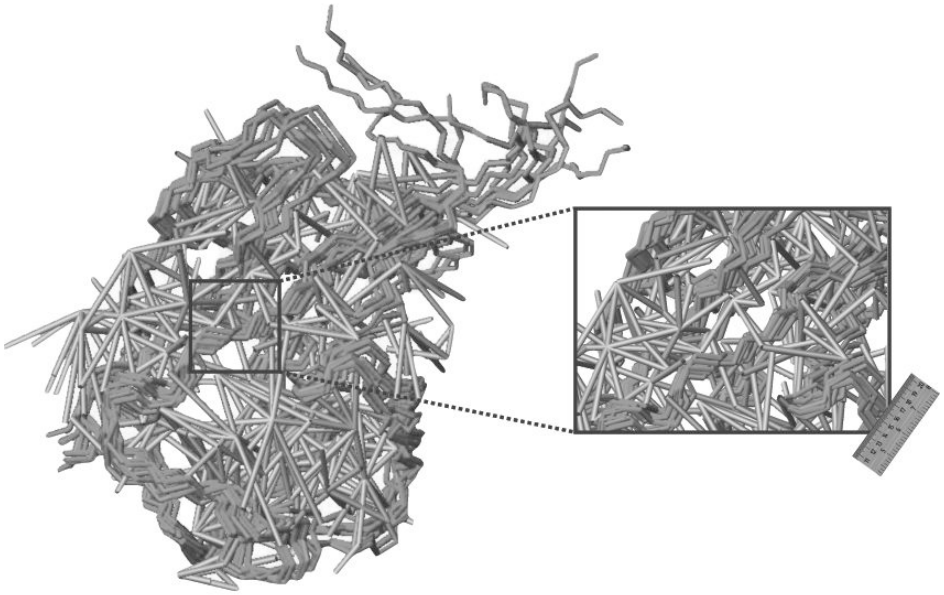


**Figure 1.5.** *NMR experiments only yield a wealth of indirect structural information from which the protein 3D structure can only be uncovered by consistent interpretation of the experimental NMR signals. Protein NMR structure determination thus entails building an atomic resolution model which must simultaneously fulfill all experimentally determined conformational restraints. The principal source for the collection of conformational restraints is derived from the nuclear Overhauser effect that allows for the measurement*

*of interatomic distances between nuclear spins in close proximity. The figure shows a protein NMR structure represented as a bundle of conformers that are all equally well satisfying the dense network of NOE-derived distance restraints.*

The inter-nuclear distance between proton pairs is typically computed from the NOE volume by using the isolated spin approximation (ISA) formula:

$$NOE = \frac{k_{calibration}}{< r >^6}$$

where $k_{calibration}$ is *a prior* unknown calibration constant that for a given protein is often defined by some empirical approximations and is usually calculated for different classes of atoms involved (backbone and side-chain atoms), and $< r >$ is the inter-nuclear distance between proton pairs. Theoretically, $k_{calibration}$ is a clearly defined function $f(\tau_c)$ of the effective correlation time $\tau_c$, but since the isolated spin approximation formula is in principle invalid for a molecular system exhibiting a dense network of proton interactions, NOE volumes or intensities are typically only interpreted as loosely defined lower and upper distance limits.

Proton-proton NOEs relate "through-space" interactions between pairs of protons in close spatial proximity (up to 5 Angstrom or so), i.*e*., being either close in the amino acid sequence (intra- or inter-molecular contacts) or far apart in the polypeptide chain, thus resulting in long-range distance restraints, that are most valuable for determining the fold of the macromolecule under investigation. Because the Brownian motions of large structures in solutions are slow, with long effective correlation times, $\tau_c$, and proteins contain a dense network of hydrogen atoms, "spin diffusion" could partially or fully deteriorate distance measurements based on $^1H - {}^1H$ NOE experiments (Gordon and Wuethrich 1978; Kalk and Berendsen 1976; Wagner and Wüthrich 1979). Spin diffusion arises as a consequence of the dependence of the NOE on the inverse sixth power of the inter-proton distance, since magnetization transfer between two spins through multiple short steps may be more efficient than a one-step transfer over the longer, direct distance.

Computational algorithms that use distances derived from the calibration of NOE spectra are termed in the early times as “distance geometry” approaches, since they are aiming to find numerical exact solutions for a given comprehensive distance matrix between atoms that can however not be provide by NMR measurements (Crippen and Havel 1988; Crippen 1977; Havel and Wüthrich 1985). Therefore, the current most convenient and commonly used approach for NMR structure calculation relies on simulated annealing protocols - restrained molecular dynamics (rMD) - that represents an efficient optimization method typically starting from extended or random structures (Nilges, Clore, and Gronenborn 1988; Nilges, Gronenborn, et al. 1988) are also loosely be classified as “distance geometry” methods since they use only distance data to determine the structure. Generally speaking, distance geometry methods have been used on the basis of underlying resonance assignment; the assigned distances are used as input in molecular dynamics/simulated annealing programs in order to fold the primary sequence into the tertiary structure.

NOE resonance assignment constituted for long time the major hurdle towards high-throughput protein NMR structure determination, due to its highly time-consuming and laborious procedure. Over the last 20 years or so, innovations were successfully made to counter the underlying combinatory problem, such as the introduction of using ambiguous distance restraints (Nilges 1995), network anchoring and restraints combination (Herrmann, Güntert, and Wüthrich 2002a). The remaining bottleneck of efficient NMR structure determination – see the success story for unbiased, automated NOE analysis in the blind test competition in CASD-NMR (Rosato *et al*., 2012, Guerry *et al*., 2015 – is currently imposed by obtaining the sequence-specific backbone and side-chain resonance-assignments.

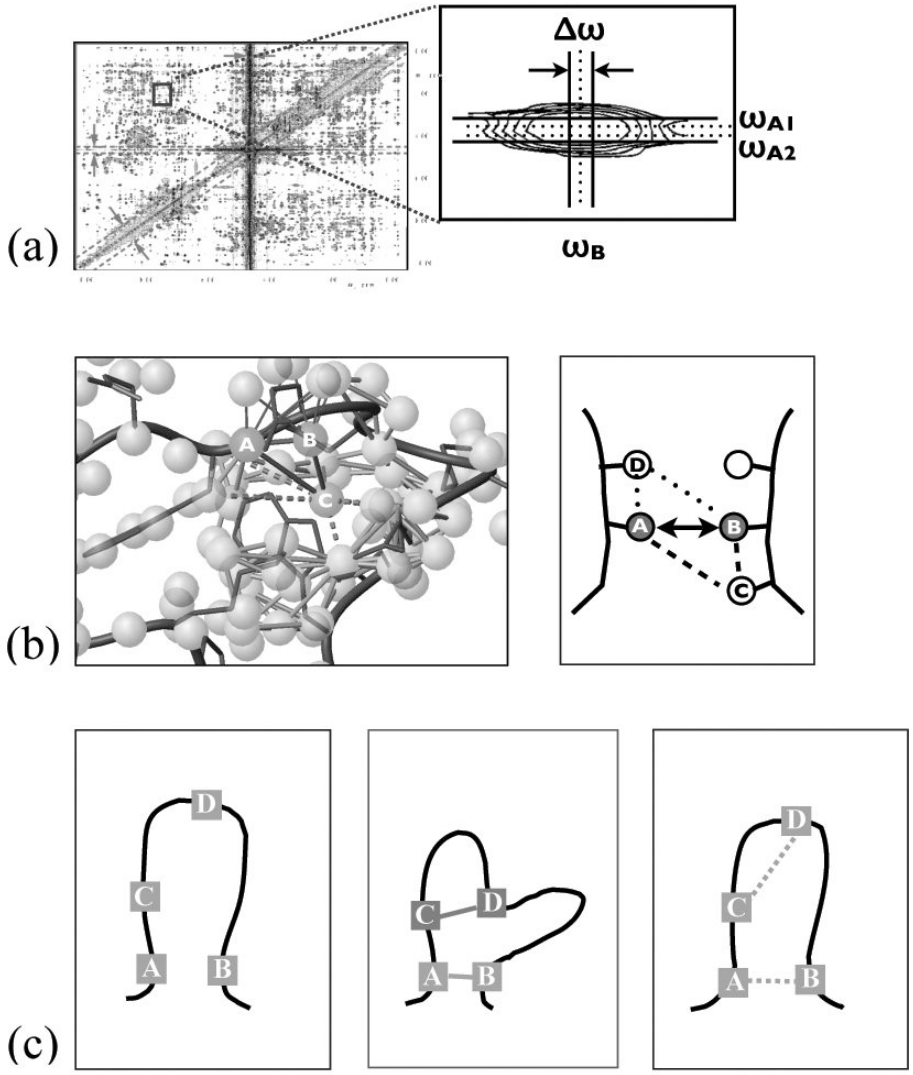



**Figure 1.6.** *Concepts used for automated NOE assignment and structure calculation: (a) Initial chemical shift-based assignment possibilities and ambiguous distance restraint: Because of the limited accuracy with which NOE cross peak positions and chemical shift values of atoms can be determined experimentally, multiple pairs of hydrogen atoms can be in general attributed to a given NOE signal. An ambiguous distance restraint will then be used for the generation of a conformational restraint. (b) Network-anchored assignment exploits the fact that correctly assigned NOE cross peaks form a mutually supportive network of distance restraints. An initial assignment between the two atoms, A and B, is considered as network-anchored, if additional assignments exist that allow to*

*establish one or several triangular connectivities of the atom pair, A and B, to a third atom, C and/or D. (c) The effect of constraint combination during protein structure calculation is depicted. The native protein fold is shown in the left panel. Assuming that the input for a structure calculation comprises one correct distance restraint between the hydrogen atoms, A and B, and one erroneous distance restraint between C and D, then the calculation will result in a distorted protein structure (middle panel), since all conformational restraints needs to be fulfilled simultaneously. However, if constraint combination is applied, i.e., the two original, unrelated distance restraints are merged into a new single, virtual distance restraint, then the correct protein fold can be obtained, since only one of the two distances A and B, or C and D needs to be short in the resulting 3D structure.*

In this thesis, we present an algorithm in order to bypass the conventionally used sequence-specific resonance assignment step and therefore, to directly determine the structure without prior resonance assignment ("NMR resonance-free structure determination"). The algorithm proposed calculates protein structures solely based on spectral peak-lists of 3D TOCSY and NOESY spectra.

## 1.4 Conclusion

Many diseases imply a change in patient's metabolism that causes significant variations to the concentrations of metabolites that appear in bio-fluids. The studies of a person's metabolic profile, *i.e*., the list of concentrations of different metabolites can help detect diseases (Ravanbakhsh et al. 2015). The compound identification process, known as spectral profiling is not yet conveniently automated, making NMR metabolomics a relatively low-throughput science. The automation hurdle is widely recognized and has led to a number of efforts to automate compound identification and/or quantification. While several software suites have been developed to support NMR spectral profiling of 1D $^{1}$H NMR spectra, automated methods to exploit 2D NMR data are still unsolved. The need for manual interventions leads to many issues such as slower throughput, operator accumulated errors hence incoherent or inconsistently interpreted results. In the first part

of the this Ph.D. manuscript, we present the algorithm ITERAMETA in order to address the highly desired need having a software platform that performs robust and automated spectral metabolite profiling and assignment assessment, thus enabling a reliable compound identification process.

A decade ago or so, protein NMR structure determination required months even years of hard work by well-trained experts. Nowadays, with stunning advances in NMR experiments, instrumentation and computation, structure calculation is feasible within weeks. Sequence-specific and NOE resonance assignment programs are abundant and various: there are ongoing efforts to establish a general and robust protocol for NMR structure determination that could be estimated in hours, not weeks (Billeter, Wagner, and Wüthrich 2008; Gronwald and Kalbitzer 2004; Williamson and Craven 2009).

Here we hope to present a truly real-world proof – without using synthetic data or so - for achieving protein NMR structure determination without prior sequence-specific backbone and side-chain resonance assignment. This second goal of the thesis presents a formidable challenge in terms of algorithmic developments, however with the ultimate promise of possibly fasten and alter the process of NMR structure determination/

A central initiative to promote accurate computational algorithms for NMR structural biology is the worldwide *"Critical assessment of automated structure determination of proteins from NMR data"* (CASD-NMR), providing a survey of unsupervised protein structure determination based on NMR chemical shifts and/or NOESY data (Rosato and Billeter 2015; Rosato et al. 2009, 2012). CASD-NMR offers an ideal source of NMR data in order to test new methods or to assess old ones. To demonstrate the feasibility of NMR assignment-free protein structure determination method, we tested our algorithm on data taken from the CASD-NMR competition.

# CHAPTER 2

## 2. Automated Metabolite Profiling from 2D TOCSY and/or HSQC Spectra

As comprehensively elaborated in Chapter 1 of this thesis, commonly used one-dimensional (1D) proton NMR spectral data analysis in metabonomic studies suffers severely from the unavoidable disadvantage that chemical shift dispersion of NMR signals in complex biological mixtures is relatively small, resulting in a multitude of overlapping signals in most regions of the 1D spectrum. Consequently, the resulting poor spectral resolution of individual nuclei renders the task of unambiguous metabolite identification into a nearly impossible undertaking. This is so because this strong signal overlap compromises the necessary identification of the characteristic underlying spin-coupled resonance patterns of metabolites (fingerprints) for unambiguous, confident metabolite annotation.

Multidimensional (2D, 3D or more) NMR spectroscopy offers a convenient solution to resolve metabolite assignment ambiguities by spreading NMR resonances into distinct spectral dimensions, such as demonstrated by targeted projection spectroscopy, resulting in significantly reduced signal overlap (Pontoizeau et al. 2010). But this comes with a cost, multidimensional NMR spectra require (considerably) more NMR measurement time than needed for the acquisition of 1D NMR spectra, an aspect that is important considering the fact that typically multiple biological samples need to be analyzed in a single NMR metabonomics study. Therefore – as good compromise – two-dimensional (2D) NMR spectroscopy is mainly used in NMR metabonomics. In particular, two-dimensional homonuclear $^{1}H$ - $^{1}H$ total correlation spectroscopy (TOCSY) and two-dimensional heteronuclear single quantum correlation $^{13}C$- and/or $^{15}N$- $^{1}H$ spectroscopy (HSQC) are the most popular NMR experiments used in order to achieve confident metabolite.

## 2.1 Introduction

Recently, a number of numerical strategies have been developed to automatically identify metabolites using the above-mentioned 2D homo- and/or hetero-nuclear NMR spectra: semi-automated methods such as MetaboMiner (Xia et al. 2008) or CCPN Metabolomics Project (Chignola et al. 2011), automated methods such as trace analysis (Bingol and Brüschweiler 2011, 2014; Bingol et al. 2015, 2016; Robinette et al. 2008, 2011), COLMAR based on covariance spectroscopy (Zhang et al. 2008, 2009; Zhang, Bruschweiler-Li, and Brüschweiler 2012) and heuristic likelihood search (Xi et al. 2006). Generally speaking, a common drawback of the currently proposed semi- or fully unsupervised approaches is that they are operating on their own customized, relatively small and thus more or less incomplete metabolite database. As a result, the availability and steadily growing potential of (complementary) public metabolite databases such as the Human Metabolite Database or HMDB (Wishart et al. 2007, 2009, 2013, 2016), Metabolomics Biological Magnetic Resonance Database or BMRB (Markley 2012; Markley et al. 2007), the Madison Metabolomics Consortium Database (Cui et al. 2008) or the Yeast Metabolome Database or YMDB (Jewison et al. 2012) are currently only be partly exploited by direct searching and mapping query algorithms, despite their expanding collection of available experimental two-dimensional NMR peak-lists and/or spectra of hundreds of individual metabolites.

Here we propose a novel and robust numerical method for the individual and/or combined analysis of two-dimensional hetereonuclear $^{1}$H-$^{13}$C HSQC and homonuclear $^{1}$H-$^{1}$H TOCSY spectra that yields automated metabolite identification without any or only minor input restrictions using as reference libraries two of the largest repository archives available, namely the Biological Magnetic Resonance Data Bank (BMRB) and the Human Metabolome Database (HMDB), thus intrinsically guaranteeing a high coverage of currently catalogued metabolites. Key elements of our method designed to overcome current data analysis deficiencies in NMR metabolomics are the combined use of contemporary, sorely mathematically based pattern recognition techniques, such as advanced peak cluster methods and optimal peak assignment methods that enable

confident metabolite determination with high accuracy. These techniques are applied in an iterative, adaptive spectral analysis mode in order to achieve convergence to the best possible matching between experimental and reference spectra, and shows great potential to provide a better identification of known metabolite than existing methods. Proof-of principle was first obtained by evaluating performance characteristics on different synthetic data sets, and more importantly on real-world applications of human urine samples, exhibiting high data complexity. Flawless metabolite assignments were obtained for the synthetic case. Notably for manually prepared input peak lists of experimental urine samples, the unsupervised result delivered a high agreement between automated and in-depth manual, interactive spectral evaluation. These results clearly demonstrate that the numerical method proposed provide a significant advance towards providing a robust tool for rapid identification of metabolites at natural abundance, and most importantly achieves an accurate and high coverage of identified metabolite in complex mixtures.

## 2.2 Description of ITERAMETA approach

The flowchart of the proposed numerical method ITERAMETA is shown as a combination of multiple building blocks with three main elements (Figure 2.1.) that subsequently

(i) initialize the input data

(ii) perform iterative searching and matching between experimental input signals and 2D TOCSY/HSQC metabolite database patterns

(iii) assess the matching quality thanks to various tools implemented along the program.

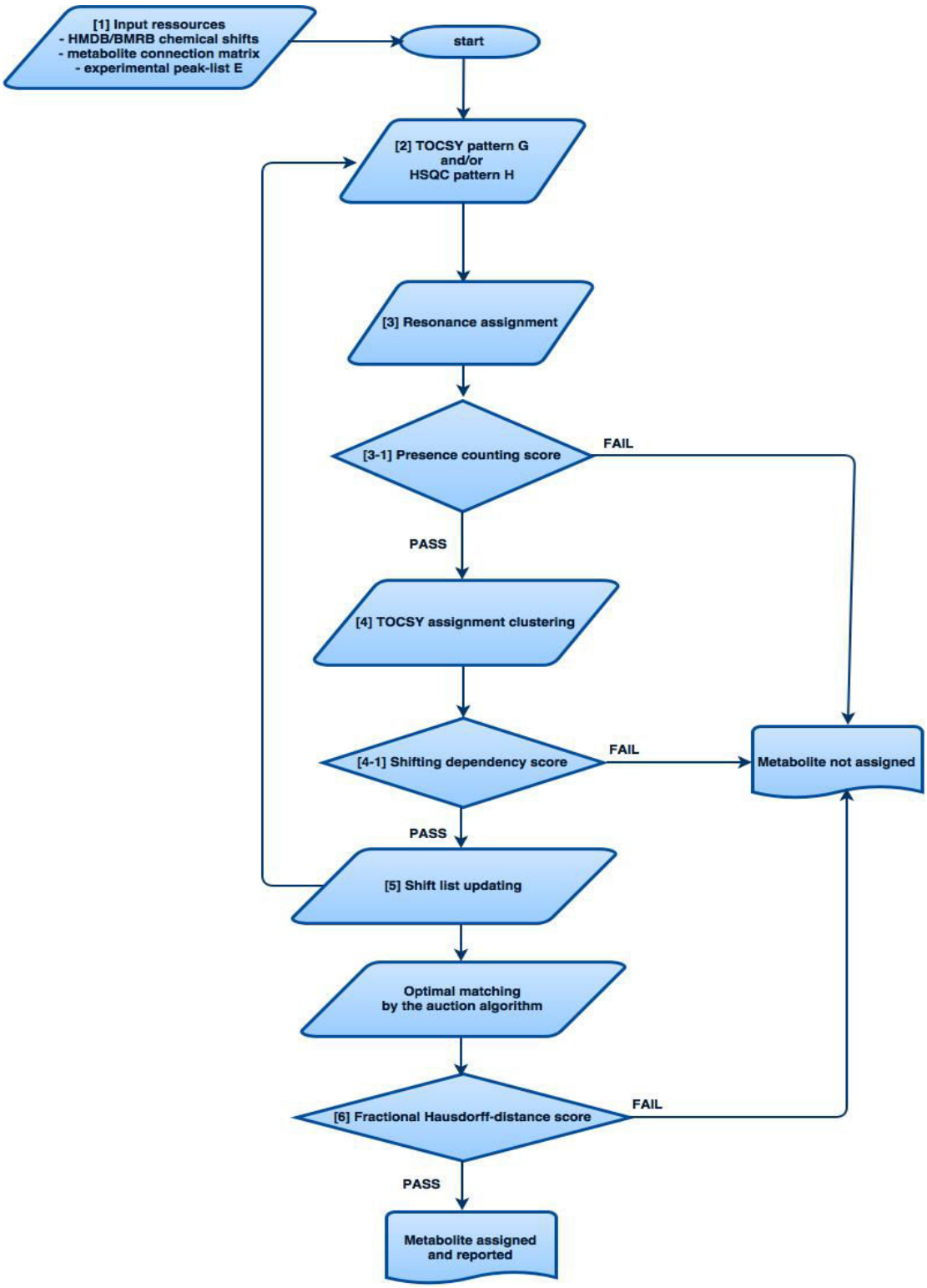



**Figure 2.1.** *Flowchart of NMR metabolite assignment using the ITERAMETA method proposed.*

The input for the ITERAMETA method consists of: (i) one or several experimental two-dimensional NMR peak-lists, HSQC and/or TOCSY; (ii) a set of numerical parameters for controlling applied assignment thresholds in the ITERAMETA method (default parameter values are provided, but can be changed by the user); and (iii) a public metabolite resonance database. The public metabolite database in the ITERAMETA program can be chosen as HMDB (www.hmdb.ca) or BMRB (http://www.bmrb.wisc.edu).

**[Building block 1] Input resources**: Available database-deposited metabolites are processed by their name in alphabetic order: every metabolite will be associated with its corresponding TOCSY and/or HSQC reference peak-lists. The reference peak-list is either already provided by the public reference database or created internally from the list of resonance frequencies of nuclei and the adjacency matrix encoding the covalent connections between atoms (see [Building block 2]).

Each occurring NMR signal between atoms pairs of a metabolite in the reference peak-list is characterized by its two chemical shift coordinates: $\omega_i^k$ with $i \in \{1,2\}$ and $k$ stands for the (arbitrary) peak number in the reference peak-list.

The experimental input peak-list is denoted as a set of NMR signals $\boldsymbol{E}$: $\boldsymbol{E} = \{E_k\}$ with $E_k = \left(\omega_1^k, \omega_2^k\right)$ and its associated peak volume $V_k$. The experimental NMR peak volume is only stored for the final reporting about identified metabolites, no use of it is made during the annotation process. $\omega_1^k$ is the first frequency coordinate of the experimental peak $E_k$: it contains the chemical shift value of a hydrogen or a heavy atom ($^{13}C$ or $^{15}N$) atom, depending if the set of experimental input peaks $\boldsymbol{E}$ originates from a 2D TOCSY or a HSQC spectrum, respectively.

The ITERAMETA user can freely decide to use as query reference library either the HMDB or the BMRB. The extent of available reference metabolite in these two public metabolite databases is reported in Table 2.1.

| | Number of HSQC spectra | Number of TOCSY spectra |
|---|---|---|
| HMDB | 972 | 242 |
| BMRB | 1081 | 1042 |

**Table 2.1.** *The number of available reference metabolite spectra is shown for each of the two databases, HMDB and BMRB, used in the ITERAMETA program.*

**[Building block 2] TOSCY/HSQC pattern**: Each database-deposited metabolite will be internally associated with a chemical shift list $S^i$: $S^i = \{\Omega^i_{M,\alpha}\}$ with $M$ stands for the name of the metabolite, $\alpha$ encodes an individual nucleus of the metabolite, and $i$ assigns the current iteration cycle. In each assignment iteration, the original reference chemical shift values of individual nuclei are adapted in order to find best possible query matching between experimental and reference peak pattern (see [Building block 5]). At the outset of the process, *i.e.*, $i = 0$, the chemical shift values of individual nuclei of a metabolite are taken from the HMDB and/or BMRB data. In later iterations, these values are allowed to change in order to account for different experimental conditions used during the acquisition of the experimental and reference NMR spectra.

Each theoretical HSQC peak is resulting from the coupling of a proton and its covalently bonded carbon/nitrogen atom. The complete expected HSQC pattern is the combination of all these theoretical peaks. The HSQC pattern of each metabolite is denoted $\boldsymbol{H}_M$ that contains individual peaks $H^l_M = (\Omega^l_{Z_\alpha}, \Omega^l_\alpha)$, $l$ is the (arbitrary) peak number in the expected pattern, $\alpha$ the proton and $Z_\alpha$ its corresponding (covalently bonded) heavy atom ($^{13}C$ or $^{15}N$).

A theoretical, expected two-dimensional TOCSY peak is labeled $T$, resulted from two protons $\alpha$ and $\beta$, denoted $(\Omega^{T,i}_{M,\alpha}, \Omega^{T,i}_{M,\beta})$ with $i$ the number of iteration cycle; two protons $\alpha$ and $\beta$ being separated by no more than 4 covalent bonds.

For the a given metabolite *M*, an initial TOCSY pattern $\boldsymbol{G}_M^i$ is generated from the proton chemical shift list and the connectivity matrix between protons: $\boldsymbol{G}_M^i = \{G_M^{T,i}\}$ with $G_M^{T,i} = (\Omega_{M,\alpha}^{T,i}, \Omega_{M,\beta}^{T,i})$ where *i* is the current iteration cycle.

The cardinality $|G_M^i|$ gives the number of elements present in the set $\boldsymbol{G}_M^i$, *i.e.*, the number of NMR signals in the expected peak pattern, and this number remains unchanged throughout the iterations. For the simplification and the readability of the following equations, the superscript *i* for the current iteration cycle, and the subscript *M* encoding the name of a metabolite will be dropped in the following description.

**[Building block 3] Resonance assignment:**

*TOCSY resonance assignment*: For each metabolite and in each iteration cycle *i*, the expected or library query TOCSY pattern $\boldsymbol{G}$ is compared to the set of experimental NMR signals $\boldsymbol{E}$. Each experimental peak *k*, $E_k = (\omega_1^k, \omega_2^k)$ or $E_k(\omega_m^k)$ is assigned to an expected peak, $(G^T)$, if their respective shift coordinates are matched within a user-given chemical shift tolerance range Δδ: $|\omega_m^k - \Omega_\vartheta^T| \leq \Delta\delta$, where $m = \{1, 2\}$ is the frequency dimension and $\vartheta = \{\alpha, \beta\}$ encodes the atoms involved. An assignment possibility is then established between $(G_\vartheta^T)$ and $E_k(\omega_m^k)$ and is denoted as $A^{T,k}$. To each assignment possibility is associated with a Gaussian probability for chemical shift matching that is denoted $P^{T,k}$:

$$P^{T,k} = exp\left(-\frac{1}{2}\sum_{\substack{m=1,2\\ \vartheta=\alpha,\beta}}\left(\frac{\omega_m^k - \Omega_\vartheta^T}{\Delta\delta}\right)^2\right).$$

For each metabolite and in each round of iteration, all possible assignments for a given metabolite are gathered into a comprehensive set of plausible assignment possibilities, denoted $\boldsymbol{A}$:

$$\boldsymbol{A} = \{A^{T,k} = (T,k) |\ (T,k) = (\Omega_\alpha^T, \Omega_\beta^T, \omega_1^k, \omega_2^k, P^{T,k})\}.$$

*HSQC resonance assignment:* Very similarly as computed for the experimental TOCSY input peak-lists, the expected HSQC peak pattern is compared to the set of

experimental NMR signals $\boldsymbol{E}$. The experimental input peak $E_k(\omega_m^k)$ is assigned to the theoretical peak ($H_M^l$), if their respective frequency coordinates are matched within user-defined chemical shift tolerance ranges, $\Delta\delta$ for protons and $\Delta\delta_Z$ for heavy atoms: $\left|\omega_1^k - \Omega_{Z_\alpha}^l\right| \leq \Delta\delta_Z$ and $\left|\omega_2^k - \Omega_\alpha^l\right| \leq \Delta\delta$.

Identically to the task of TOCSY resonance assignment, all plausible assignment possibilities are gathered into a set of assignment possibilities, $\boldsymbol{A}$. But since the HSQC spectrum does not contain any spectral redundancy properties in terms of multiple reoccurring frequency resonances of the same atom that can be profitably exploit for reliable chemical shift adaption between theoretical and experimental peak pattern of a metabolite, the HSQC resonance assignment is applied as simple direct searching and matching method without any iterations. In particular, the ITERAMETA building blocks 4 and 5 of TOCSY *assignment clustering* and *Shift list updating* as described below are not applied for such experimental input data.

**[Building block 3-1] Presence counting score:** The number of theoretical or expected NMR signals that can be found, $\boldsymbol{R}$, in the experimental input peak-list, should be more than $N_1\%$ of the total number of peaks in the reference pattern $\frac{R}{|G|} \geq N_1\%$ or $\frac{R}{|H|} \geq N_1\%$ for TOCSY and HSQC input data, respectively.

**[Building block 4] TOCSY assignment clustering:** From the set of all plausible assignment possibilities $\boldsymbol{A}$, a set of chemical shift deviations $\boldsymbol{D}$ is computed in order to characterize the two-dimensional deviations between the theoretical and the experimental input peaks in each respective assignment:

$$\boldsymbol{D} = \left\{D^{T,k} = \Delta(T,k) \middle| \Delta(T,k) = \left(\omega_i^k - \Omega_\vartheta^T\right)\right\} \text{ with } i \in \{1,2\} \text{ and } \vartheta \in \{\alpha,\beta\}.$$

The quality threshold clustering method (Heyer, Kruglyak, and Yooseph 1999; Jin and Han 2010) is here used for differentiating between most likely and rather unlikely assignment possibilities. The workflow of this clustering algorithm can be summarized as follows:

(i) Build a candidate cluster for each data point $\Delta(T,k)$ by including the closest point, then the next closest point and continuing to do so, until the distance of the cluster is superior to a predefined threshold value $\sigma$.

(ii) Take out the biggest cluster from the data (and from any further consideration) and save it as an outcome cluster $C^j$ with *j=1..n*, and *n* is the number of resulting clusters.

(iii) Repeat the process with the reduced set of points until no more clusters can be formed.

The resulted clusters are denoted $C^j = \{\Delta(T,k)_j\}$ with $j$ being the cluster number. The advantage of the quality threshold clustering method in contrast to other clustering methods proposed is the fact that the number of identified clusters does not need to be specified, *i.e.*, the number of obtained clusters is only a function of the (quality) of the experimental input data.

**[Building block 4-1] Shifting dependency score:** An assignment cluster is considered *eligible* if its cardinality is more than $N_2\%$ of the total number of peaks in the reference pattern: $\frac{|C^j|}{|G|} \geq N_2\%$.

**[Building block 5] Shift list updating:** If an assignment cluster $C^j$ is eligible (see Building block [4-1]), the reference updating process will take place in order to adapt the reference chemical shift list - originally taken from a database - to the experimental input data. Each database chemical shift is updated according to its corresponding eligible assignments:

$$\Omega_\vartheta^{new} = \frac{\sum_{(T,k)\in C^j} \omega_i^k \times P^{T,k}}{\sum_{(T,k)\in C^j} P^{T,k}}, \text{ where } \left|\Omega_\vartheta^{old} - \omega_i^k\right| \leq \Delta\delta.$$

**[Building block 6] Fractional Hausdorff-distance score:** A Hausdorff-distance-based score measures the similarity between the subsequently shifted reference pattern and the experimental input data. The Hausdorff distance between two sets of points ***A*** and ***B*** is defined as: $h(A, B) = \max_{a \in A}\min_{b \in B} d(a,b)$ with d(a,b) holds the distance between two

points depending on the chosen norm. The norm chosen in ITERAMETA is Euclidean norm. A fractional Hausdorff-distance computes the Hausdorff distance only over a pre-defined fraction of $\boldsymbol{A}$, thus removing the unmatched outliers. The user can define the Hausdorff-distance-based fraction denoted $N_3\%$ used in the program.

The matching between the shifted reference pattern $G_M^c$ and the experimental input data $\boldsymbol{E}$ is considered as reliable, good matching, if their $N_3$-Hausdorff distance is lower than a predefined threshold $\gamma$: $h(G_M, E) = \max_{a \in N_3 \times G_M} \min_{b \in E} d(a,b) \leq \gamma$.

The parameter values that are used to control to process of metabolite assignment are listed in Table 2.2., default values are provided, but the user can freely change this values.

| Symbol | Parameter | Value |
|---|---|---|
| $\Delta\boldsymbol{\delta_H}$, $\Delta\boldsymbol{\delta_C}$ | Tolerance range for peak positions in both directions of proton and/or carbon resonance | 0.05-0.1 ppm for proton<br>0.7-1.7 ppm for carbon |
| $\boldsymbol{\sigma}$ | Cluster diameter threshold | 0.03-0.05 |
| $\boldsymbol{N_1}$ | Minimal percentage of pattern found in spectrum | 70% |
| $\boldsymbol{N_2}$ | Minimal percentage of pattern found within a cluster | 70% |
| $\boldsymbol{N_3}$ | Fraction of pattern considered in Hausdorff distance computing | 70% |
| $\boldsymbol{\gamma}$ | Maximum fractional Hausdorff distance between two patterns | 0.05 |
| $\boldsymbol{j}$ | Number size of patterns = number of eligible clusters | |

**Table 2.2.** *Default parameter values used in ITERAMETA.*

**[Building blocks]** ***Optimal matching by the Auction algorithm*** **and** ***Metabolite assigned and reported*****:** In order to select the best assignment for each metabolite among

all the retained assignment possibilities by the previous quality threshold clustering method, an optimization method named *auction algorithm* is implemented in ITERAMETA. In the following we will give a brief description of the auction procedure, a detailed assessment of the auction algorithm and pseudo-code of the algorithm can be found in its dedicated book (Bertsekas 1988) and the original paper (Bertsekas and Castañon 1992). The iterative auction algorithm implemented is found to perform very fast for problems with few elements, which is exactly the case of metabolite assignment optimization.

The problem of reporting selective metabolite assignment, *i.e.,* best-matching pattern for each metabolite among the retained assignment possibilities, is part of a larger numerical research field called *asymmetric assignment problem*. To solve this problem, we have chosen an auction-type algorithm whose general principles are detailed hereafter.

The present assignment optimization problem can be formulated as follows: for each retained metabolite, $N$ theoretical signals are matched against $M$ experimental signals; $v_{ij}$ is the probability that the theoretical signal $M_j$ can be assigned to the theoretical signal $N_i$. We seek a complete assignment $\zeta: N \rightarrow M$ that maximizes the total value $V(\zeta) = \sum_i v_{i,\zeta}(i)$.

When an experimental signal $M_j$ is assigned to the theoretical signal $N_i$, the value function is increased by an amount of $v_{ij}$. However, there is a cost to obtain this value: that is the opportunity cost that $M_j$ cannot be assigned to other theoretical signals or other experimental signals cannot be assigned to $N_i$. This lost cost must be taken into account in the optimization process.

The auction algorithm ((Bertsekas and Castañon 1992) solves the asymmetric assignment problem by simulating an auction session. The "auction" is simultaneously done for all signals, *i.e*., all assignment combinations. In the end, the total value is maximized while minimizing the lost cost function.

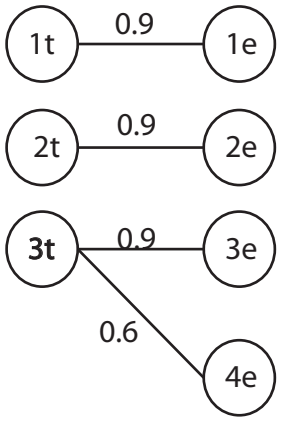



**Figure 2.2.** *Illustration of auction algorithm in order to find the best matching between theoretical peaks, 1t, 2t, 3t, and experimental input peaks, 1e, 2e, 3e and 4e.*

In Figure 2.2., we give a brief explanation how the auction algorithm is used to increase the total optimization value and to simultaneously minimize the lost cost function. The theoretical peaks, 1t, 2t have exactly one corresponding assignment possibility 1e, 2e. The theoretical peaks 3t has however two possible assignments to experimental peaks 3e and 4e. The first assignment possibility is the set $\{1t: 1e, 2t: 2e, 3t: 4e\}$ and has a total value of 0.9 + 0.9 + 0.6 = 2.4, while its lost cost is 0.9 – 0.6 = 0.3. The second assignment possibility is the set $\{1t: 1e, 2t: 2e, 3t: 3e\}$ and has a total value of 0.9 + 0.9 +0.9 = 2.7, while its lost cost is 0. Hence this second set of assignments is the better one and this one would be reported by ITERAMETA for a given metabolite.

## 2.3 Results

The performance characteristics of ITERAMETA is first assessed on synthetic input peak-lists constructed from reference libraries with arbitrarily applied chemical shift variations and using different settings of parameters values (Table 2.2.). Finally, ITERAMETA is evaluated on experimental peak-lists of human urine samples resulting from manually peak-picking by an expert analyst. These experimental input peak-lists

exhibit high data complexity and are used to evaluate ITERAMETA for real-world applications.

### 2.3.1. ITERAMETA applied to TOCSY synthetic input data

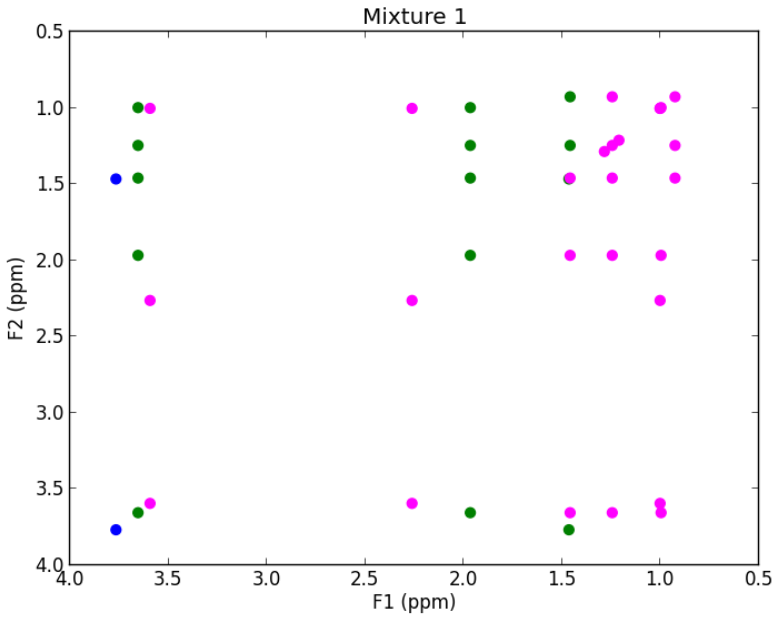
Three model mixtures extracted from the BMRB reference library are evaluated by ITERAMETA. The compositions of the three synthetic model mixtures is listed in Table 2.3. below and shows different complexity in the input peak pattern.

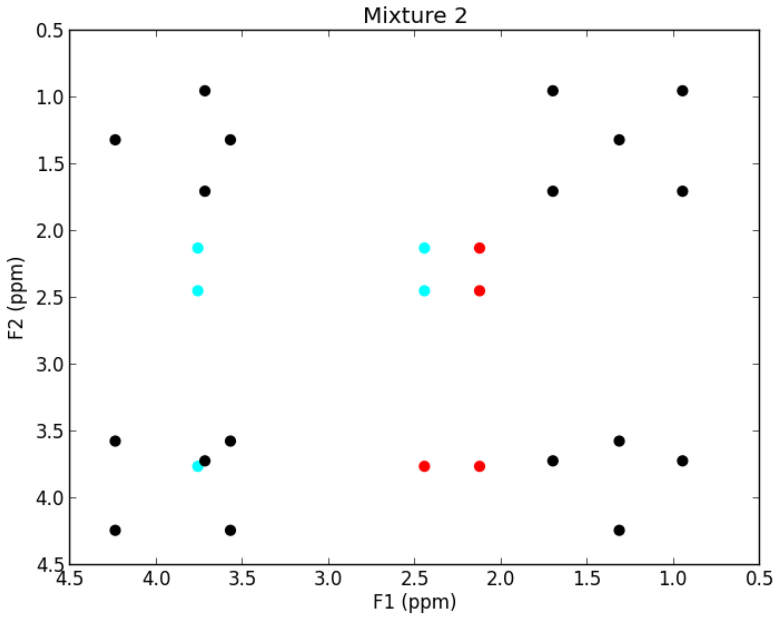
| Mixture | TOCSY synthetic amino acid composition |
|---|---|
| 1 | Alanine (Ala), Isoleucine (Ile), Valine (Val) |
| 2 | Glutamine (Glu), Leucine (Leu), Threonine (Thr) |
| 3 (mixture 1 + 2) | Alanine (Ala), Isoleucine (Ile), Valine (Val), Glutamine (Glu), Leucine (Leu), Threonine (Thr) |

**Table 2.3.** *Three TOCSY synthetic model mixtures with their metabolite composition are shown.*

The three model mixtures are generated from the BMRB reference database by adding an arbitrary variation of chemical shifts to each metabolite (Figure 2.3.). The quality threshold $\sigma$ used for obtaining different assignment clusters is a specific measure that is equivalent to alignment tolerance range, but while alignment tolerance range is usually used to align peaks along one spectral dimension, the quality threshold $\sigma$ is applicable to quantify the assignments over the full pattern of the metabolite (over all spectral dimensions). The larger the value of $\sigma$ is set, the more matching possibilities (assignments) will be included into one cluster, thus more assigned patterns can exist. For the specific case of synthetic data here, the quality threshold can be fixed at 0.00 ppm to quantify the perfect match between initial and shifted patterns. For practical cases, as

discussed below, a numerical value for quality threshold $\sigma$ of 0.03-0.04 ppm is recommended.



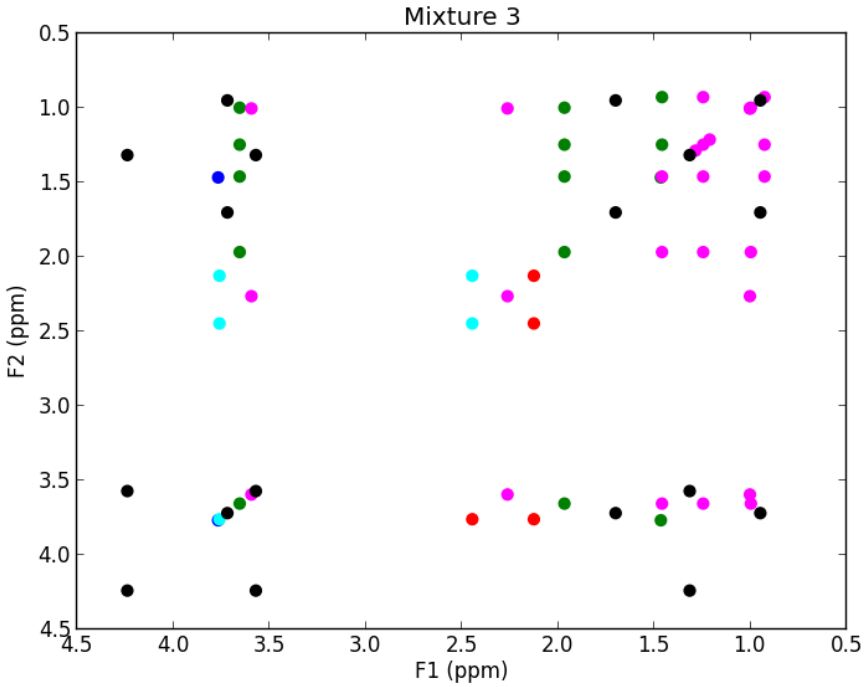



**Figure 2.3.** *The synthetic spectra resulted from mixture 1, mixture 2 and mixture 3 (mixture 1 + mixture 2) are displayed (see Table 2.3.). In color, alanine: blue, isoleucine: green, valine: magenta, glutamine: cyan, leucine: red, threonine: black.*

For the first evaluation of the programs performance, the algorithm ITERAMETA is applied to these three mixtures using the following parameter values: $N_1$= 70%, $N_2$ = 70%, $N_3$= 70%, $\gamma$ = 0.05 (see Table 2.2.), huge chemical shift matching tolerance $\Delta\delta$ of 0.1 ppm in order to challenge the robustness of the algorithm, and the quality threshold $\sigma$ is set to 0.00 and 0.04 ppm, respectively.

Using $\sigma$ at 0.00 ppm, ITERAMETA assigned correctly the compounds targeted in each mixture. When increasing $\sigma$ to its recommended value of 0.04 ppm, we expect new compounds to be listed since we increase the scope of assignment possibilities; the corresponding result is listed in the Table 2.4.

| Mixture | Metabolite profiling by ITERAMETA |
|---|---|
| 1 | Alanine, Isoleucine, Valine |
| 2 | Glutamine, Leucine, Threonine, monoethyl-malonate |
| 3 | Alanine, Glutamine, Isoleucine, Leucine, Threonine, Valine, Monoethyl-malonate, nonadecane |

**Table 2.4.** *Metabolite profiling by ITERAMETA setting the value* of the quality threshold σ to 0.04 ppm.

As expected setting the clustering threshold value $\sigma$ to 0.04 ppm, the ITERAMETA algorithm is listing all the original compounds, but in addition new compounds are assigned whose patterns are sub-pattern of the combinations of the original compounds. For example, *Monoethyl malonate* was reported "found" with 4 over 5 peaks, and its observed pattern is a sub-pattern of *threonine*. The found peak pattern is shown in Figure 2.4 below.

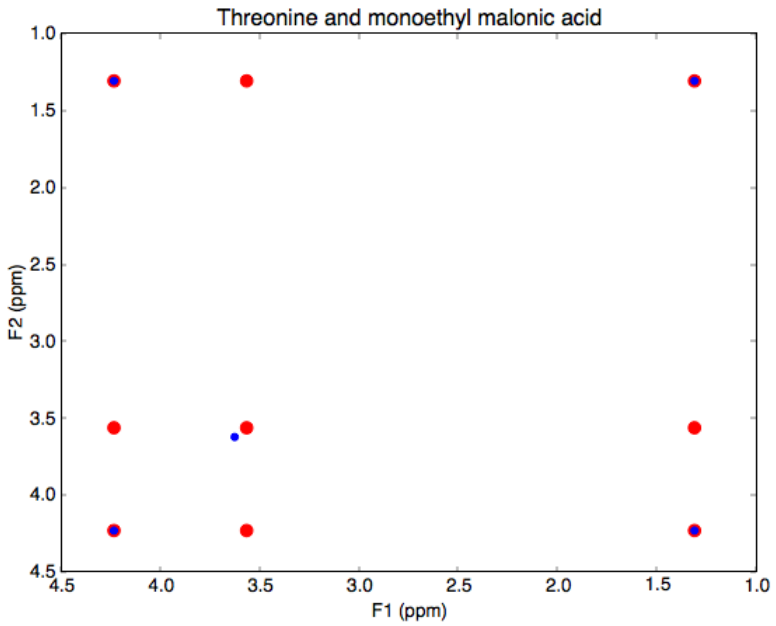


**Figure 2.4.** *The pattern of threonine and monoethyl malonic acid are shown in red and bleu, respectively. The two patterns are superposed over 4 out of 5 peaks of the pattern of monoethyl malonic acid. Note that the patterns are shifted towards their respective matches.*

By varying the tolerance value used for initial chemical shift matching $\Delta\delta$, we observed its effect on the quality of assignment obtained. In the present case, the assignment quality is defined as the number of compounds found in the mixture (see Table 2.5 below)

| $\Delta\delta$ (ppm) | $\sigma$ (ppm) | Mixture 1 | Mixture 2 | Mixture 3 |
|---|---|---|---|---|
| 0.05 | 0.00 | 3 | 3 | 6 |
| 0.1 | 0.00 | 3 | 3 | 6 |

**Table 2.5.** *The effect of the variation of the tolerance matching $\Delta\delta$ on the number of assigned metabolites.*

Despite using a large matching tolerance $\Delta\delta$, the ITERAMETA algorithm correctly assigns metabolites in the input peak-lists, therefore demonstrates the robustness of the approach.

In the following tests, without noted otherwise, the values of the ITERAMETA parameters are the following: $\Delta\delta = 0.07$ ppm and $\sigma = 0.04$ ppm, $N_1$ = 70%, $N_2$ = 70%, $N_3$ = 70% and $\gamma$ = 0.05 and the reference database HMDB is used to construct the theoretical, expected peak patterns.

## 2.3.2. ITERAMETA applied to TOCSY synthetic human urine sample

In order to further validate the ITERAMETA approach, the algorithm is applied to a synthetic human urine sample containing the 16 most frequently occurring metabolites (Saude and Sykes 2007). The composition of this complex model mixture is the

following: alanine, aspartate, citrate, creatine, creatinine, cytidine, formate, glutamate, hippurate, hypoxanthine, lactate, phenylalanine, threonine, tryptophan, tyrosine, uridine.

The ITERAMETA parameters were set to its default values (see above). Similar to the results obtained for the previous three small model mixtures, all 16 initial compounds are identified, but again also additional metabolites are found. ITERAMETA assigns a total of 33 metabolites that are listed in Table 2.6 below. As intermediate conclusion using synthetic input data, one can state that ITERAMETA is exactly performing as originally designed, *i.e.*, all original compounds could be reliable identified, and additional metabolites that form sub-patterns are also identified, thus a exhaustive set of metabolite identification is obtained.

| Pattern size | Assigned metabolites |
|---|---|
| 1 | Dimethylamine<br>Glycolic acid<br>Glycine<br>Guanidoacetic acid<br>**Formic acid**<br>Oxalacetic acid<br>Pyruvic acid<br>Trimethylamine |
| 2 | Syringic acid<br>**Hypoxanthine** |
| Between 2 and 10 | **Citric acid**<br>**Creatine**<br>**Creatinine**<br>Glyceric acid<br>**Alanine**<br>**Lactic acid**<br>Serine<br>Taurine<br>Acetyl-alanine (4/5)<br>Pyridoxamine (7/9)<br>**Hippuric acid** |
| Between 10 and 20 | Fructose (12/13)<br>Guanosine (17/21)<br>**Glutamic acid**<br>**Phenylalanine**<br>**Threonine**<br>**Aspartic acid** |

| | |
|---|---|
| | Myoinositol (13/17)<br>**Tyrosine**<br>Glucaric acid<br>Indoxyl sulfate (14/15)<br>Aminobenzoic acid (9/12), |
| >20 | **Cytidine**<br>**Uridine**<br>Xanthosine (17/21)<br>**Tryptophan**<br>Phosphogluconic acid (19/23) |

**Table 2.6.** *ITERAMETA assigned metabolites for a synthetic human urine model mixture, in bold character are the initial compounds; for incomplete matched peak patterns, the number of experimental peaks found and the number of expected peaks are noted between the brackets.*

### 2.3.3. ITERAMETA applied to an experimental human urine sample

To demonstrate the robustness and the reliability of the ITERAMETA approach for real-world applications, ITERAMETA was applied to experimental data measured at 800 MHz proton resonance frequency that were issued from human urine samples prepared under physiological condition. The typical computing time of the program is considerably less than a minute on a single CPU unit of contemporary desktop computers, and usually ranges from 10-20 seconds for an experimental input peak-list of about 1000 NMR signals.

For this current performance evaluation of ITERAMETA, the input data consists of manually prepared 2D TOCSY and HSQC peak-lists. The input peak-lists were previously used to identify metabolites in these human urine samples. The total time to achieve an exhaustive manual, interactive metabolite assignment was estimated to have taken several man-weeks of an experienced analyst. A comparison of the manually and automatically identified metabolites is shown in Table 2.7 below.

| | Number of peaks | Automatically assigned metabolites | | Number of automatically assigned peaks | | Manually assigned metabolites | Number of manually assigned peaks |
|---|---|---|---|---|---|---|---|
| HSQC | 1484 | HMDB | BMRB | HMDB | BMRB | 281 | 324 |
| | | 355 | 271 | 810 | 612 | | |
| TOCSY | 824 | HMDB | BMRB | HMDB | BMRB | 109 | 749 |
| | | 120 | 336 | 502 | 577 | | |

**Table 2.7.** *Comparison of identified metabolite resulting from the automatic ITERAMETA approach and manual, interactive analysis for both HSQC and TOCSY spectra of human urine measured at 800 MHz proton resonance frequency.*

ITERAMETA employs a *non-mutually-exclusive approach*: one experimental peak can be used for the assignment of several metabolite patterns, a criterion that is also followed by an interactive, manual expert analysis. This assignment strategy is designed so to list any plausible assignment possibility. The independently performed assignment of both HSQC and TOCSY spectra and the intersection between the two assignments found yield a number of overlapping metabolites (Table 2.8). The combined analysis of HSQC and TOCSY spectra is obviously the best procedure to reliably achieve metabolite annotation in a complex biological sample. All over, the automated method shows comparable results to the intensive manual assignment process in terms of both the total number of identified metabolites and identical metabolites found by the two approaches (Table 2.9). The agreement between identical metabolites assigned by both methods for the combined TOCSY/HSQC analysis is in the range of 76-87%, respectively.

| | Automated assignment by ITERAMETA using HMDB | Manual assignment |
|---|---|---|
| HSQC | 355 | 281 |
| TOCSY | 120 | 109 |
| HSQC ∩ TOCSY | 72 | 63 |

**Table 2.8.** *Comparison of the number of manually or automatically identified metabolite using 2D HSQC and TOCSY spectra.*

| | Number of identical metabolites identified by both approaches | Number of metabolites identified by ITERAMETA | Number of metabolites identified by interactive analysis |
|---|---|---|---|
| HSQC | 213 | 355 | 281 |
| TOCSY | 87 | 120 | 109 |
| HSQC ∩ TOCSY | 55 | 72 | 63 |

**Table 2.9.** *Comparison of identical metabolites identified by automated and manual approach.*

The automated ITERAMETA approach has however significantly reduced the time amount needed for finding metabolites within TOCSY/HSQC peak-lists. In minutes, the program provides plausible assignments and could therefore leave the user the softer task of validating and/or completing metabolite annotation in the NMR spectra. In conclusion, the automated ITERAMETA approach shows very satisfying assignment robustness through different input peak-lists and yields results in reliable agreement with manual procedures.

## 2.4 Discussions

In the following, some novel concepts introduced by ITERAMETA are discussed and their importance for robust metabolite annotation is illustrated with some hopefully intriguing example cases.

### 2.4.1. About the importance of assignment-clustering in ITERAMETA

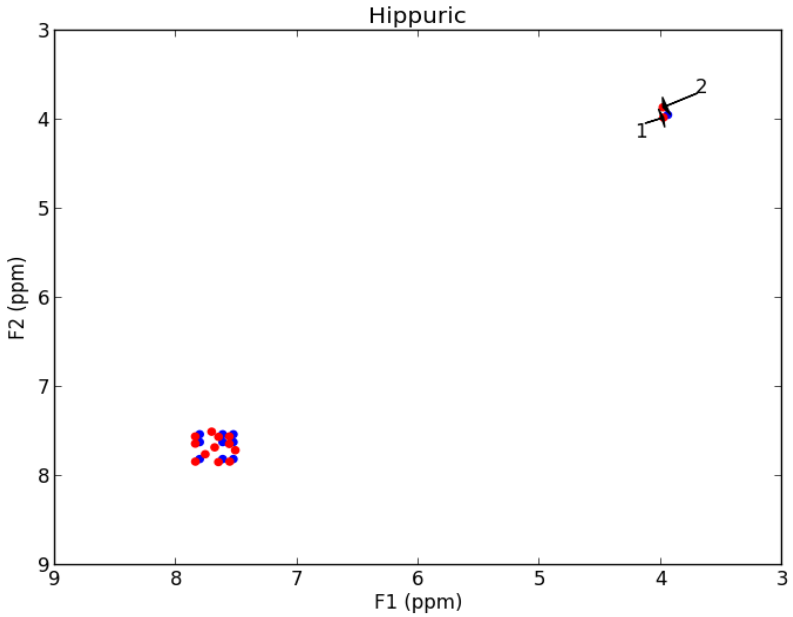
Due to the high sensibility of metabolite resonances to pH as well as to other physiological conditions, the chemical shift matching tolerance value should be sufficiently large for obtaining a nearly complete matching between the experimental input and reference patterns. It is straightforward that the larger the value of the chemical shift matching tolerance, the more assignment possibilities can be found in the input peak-list for a given metabolite. As the number of assignment possibilities increases, an assignment assessment process is needed to identify plausible assignments and to discard the improbable ones. The algorithm denoted *assignment clustering* is introduced for this purpose.

The assignment-clustering step explores the symmetrical property of TOCSY/COSY spectrum as well as the correlation between systematic or unsystematic chemical shift deviations between reference and experimental data. ITERAMETA uses a centroid-based algorithm, called quality-threshold clustering algorithm (Heyer et al. 1999) to differentiate between correct/wrong assignments. For a quality-threshold clustering algorithm, it's not necessary to specify the number of clusters. The quality of each cluster is defined by its diameter and size and is controlled by the threshold of largest allowed cluster radius that is defined by the user. The algorithm is time-consuming by the building of clusters for each data point, but as the number of ambiguous assignments is limited, the computing time is negligible.

To demonstrate the usefulness of this assessment step, we use the case of the assignment of *hippuric* for whose expected peak pattern multiple matching possibilities are possible in the experimental input data. In Figure 2.5, the expected 2D TOCSY

pattern of *hippuric* and the experimental input peak-list is shown in blue and red, respectively. There is obvious a systematic shift deviation between the two peak sets, as best seen in the bottom panel of Figure 2.5. The isolated diagonal peak at 3.8 ppm presents two assignment possibilities, labeled *1* and *2* in the figure. Using a value for the quality threshold $\sigma$ of 0.03 ppm, the algorithm is able to differentiate between the correct and wrong assignment.

In Figure 2.6, the chemical shift deviation in each spectral dimension between expected and experimental peaks is depicted. The correct assignments (in orange) are gathered around a centroid while all wrong assignments (in green) are dispersed, and also the number of elements found in the green clusters would be too low for confident *hippuric* identification.



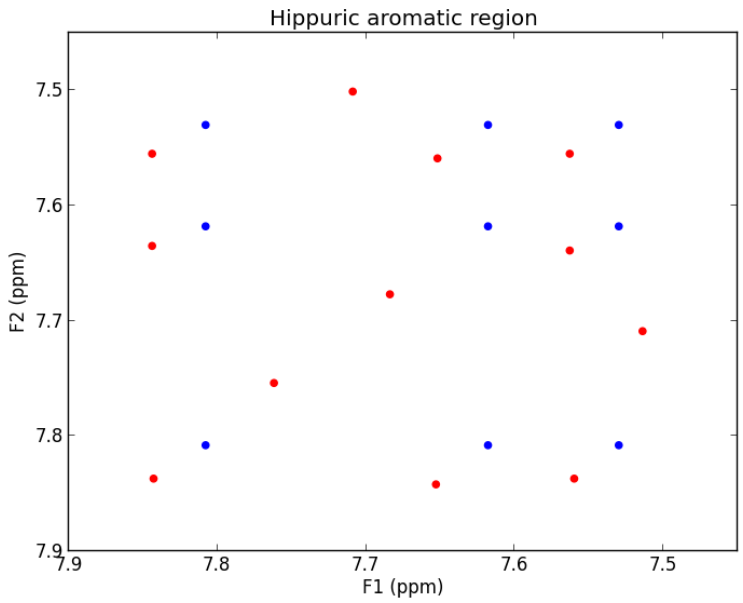



**Figure 2.5.** *The experimental and theoretical patterns of hippuric, and zooming-in of the aromatic region. The isolated diagonal peak at 3.8 ppm presents two assignment possibilities that will be differentiated using a clustering algorithm as employed in ITERAMETA.*

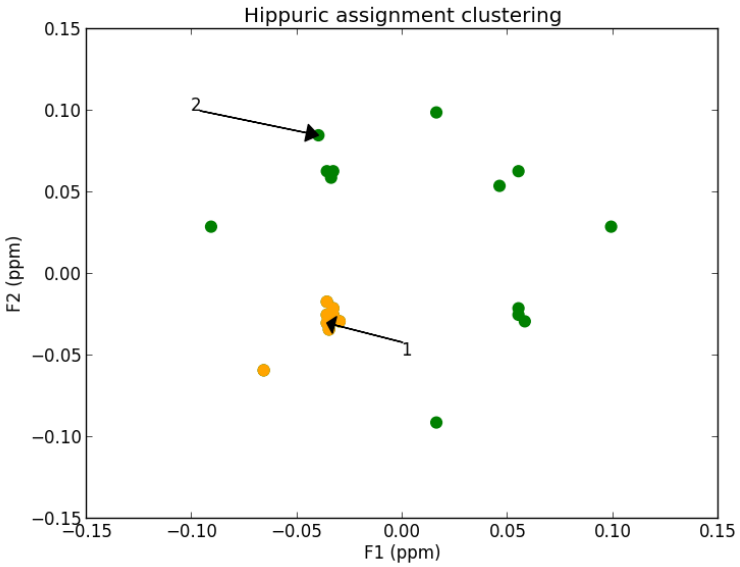



**Figure 2.6.** *Assignment clustering of hippuric resonance assignment. The best cluster resulted from the quality threshold algorithm is colored in orange, the rest is in blue.*

The choice of the value for the quality threshold $\sigma$ is crucial: since peaks in the same pattern do not shift in exactly the same pace altogether, one should expect $\sigma$ larger than 0.00. A too small $\sigma$ will not probably include enough good peaks for the iteration; on the other hand, a too large $\sigma$ will include too many matching possibilities, thus slow down the process of chemical shift reference database updating. In the specific case of the hippuric acid assignment here, the value threshold $\sigma$ of 0.03 allows the finding of the whole correct pattern while the value of 0.02 allows the finding of only 9 over 10 peaks.

The assignment-clustering step provides an independent way to verify the consistency of the assigned pattern. The resulting shift-dependency-score objectives the number of peaks in each cluster over the number of peaks expected in the reference pattern: only clusters score over the threshold $N_2$ are retained for database updating. This score is complementary to the simple presence-score $N_1$ which is the number of peaks found in the input peak-list over the number of peaks expected in the reference pattern.

### 2.4.2. About the importance of fractional Hausdorff distance-based assignment assessment in ITERAMETA

To add to the robustness of ITERAMETA, a fractional Hausdorff distance-based score is introduced to deal with imperfect matching between expected and experimental peak pattern, potentially caused by missing signals, strongly shifted signals, *etc*.. The fractional Hausdorff distance score introduced in ITERAMETA is based on the practical aspect of NMR resonance assignment: *if a pattern is only partially found in the experimental peak-list but the found part is of good-quality assignment, should one report this pattern?* In the case of perfect superposition between two patterns, their Hausdorff-distance will be zero, as well as any of their fractional Hausdorff-distance. The Hausdorff distance quantifies the similarity between two sets of peaks, and the fractional Hausdorff distance does the same but only on a fraction of the pattern.

The fractional Hausdorff distance score takes into account the fact that even if the presence of a metabolite is confirmed, its 2D TOCSY theoretical pattern can hardly be superposed perfectly onto its found experimental pattern. To illustrate this idea, we take the hypothetical case of *threonine* for which the reference pattern can be perfectly assigned and moved toward its experimental pattern except for one pair of peaks denoted ***A*** for the theoretical peak and ***1*** for its closest experimental counterpart (see these peaks in Figure 2.7).

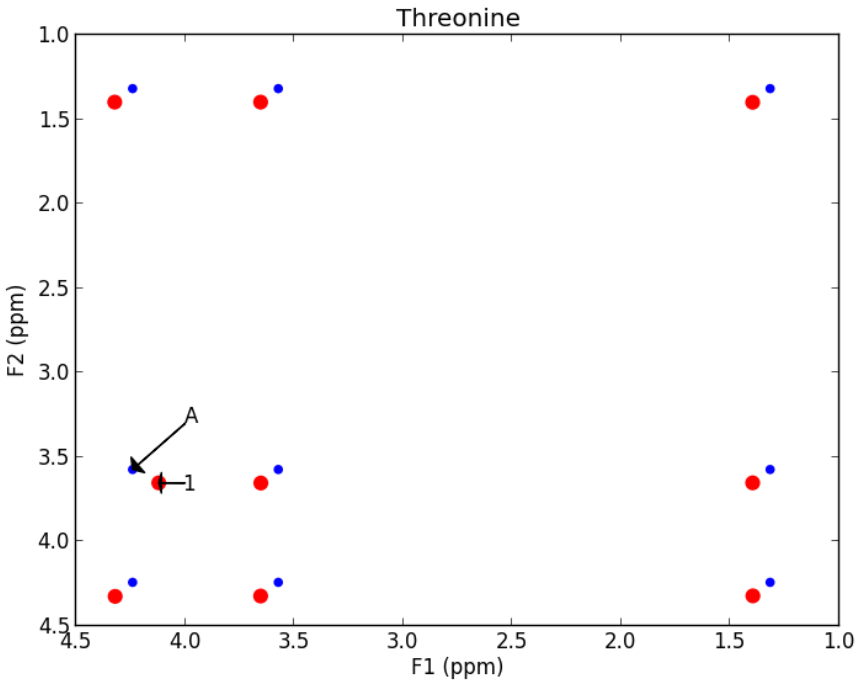



**Figure 2.7.** *The theoretical and experimental peak pattern of threonine with an initially introduced systematic reference offset.*

We have introduced a systematic reference offset between the two sets of peaks, this shift deviation will be overcome by the fact that ITERAMETA successive adapts the reference shift list to the experimentally found peaks. However the unsystematic shift deviation between peaks ***A*** and ***1*** in Figure 2.7 is harder to deal with, and leads to the fact

that the input and the reference peak pattern could never be perfectly superposed, despite the perfect matching and clustering of 8 out of 9 peaks within the pattern.

By using a simple matching probability strategy and assignment assessment solely based on the counting of theoretical peaks found in experimental peak-list, in the example shown here, threonine would have a nice matching of 9/9 peaks within a user-given chemical shift tolerance value. Using a fractional Hausdorff-distance score, we quantify differently the similarity between two patterns. Considering only a pattern of 8 peaks, the two patterns are perfectly matched, *i.e*. by using a fraction $N_3$= 0.7 or 70%, the Hausdorff distance between the two patterns is zero, well under the standard threshold $\gamma$ = 0.05.

By employing the fractional Hausdorff distance, we provide a score attesting the assignment quality that is not only based on the number of peaks found within the patterns, but accounts for experimental errors like spectral alignment, signal missing or unsystematic shift deviations. Therefore, the fractional Hausdorff distance score is an important reporting and assessment feature, valuable both for the correct performance of ITERAMETA and for providing quantitative response to the user.

### 2.4.3. The ITERAMETA user interface

We design the ITERAMETA interface to be user-friendly (GUI programming using the standard library Tkinter). ITERAMETA allows user to assign multiple TOCSY and/or HSQC peak-lists at the same time (opened in different windows), allowing them to compare between different mixtures. Once selecting the choice of TOCSY or HSQC spectrum, the user can adjust the algorithm parameters in order to restrict or widen one's searching field. Furthermore, the user disposes of an image tool following the assignment in order to visually assess the assignment of each metabolite.

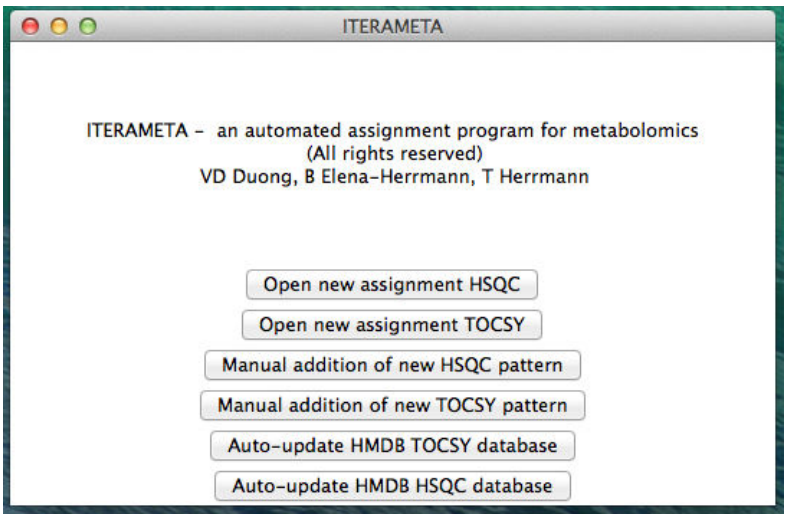



**Figure 2.8.** *ITERAMETA main menu presenting the user's multiple options: peak-list automated assignment, manual addition of new peak patterns and auto-update HMDB database by pre-downloading the updated database version.*

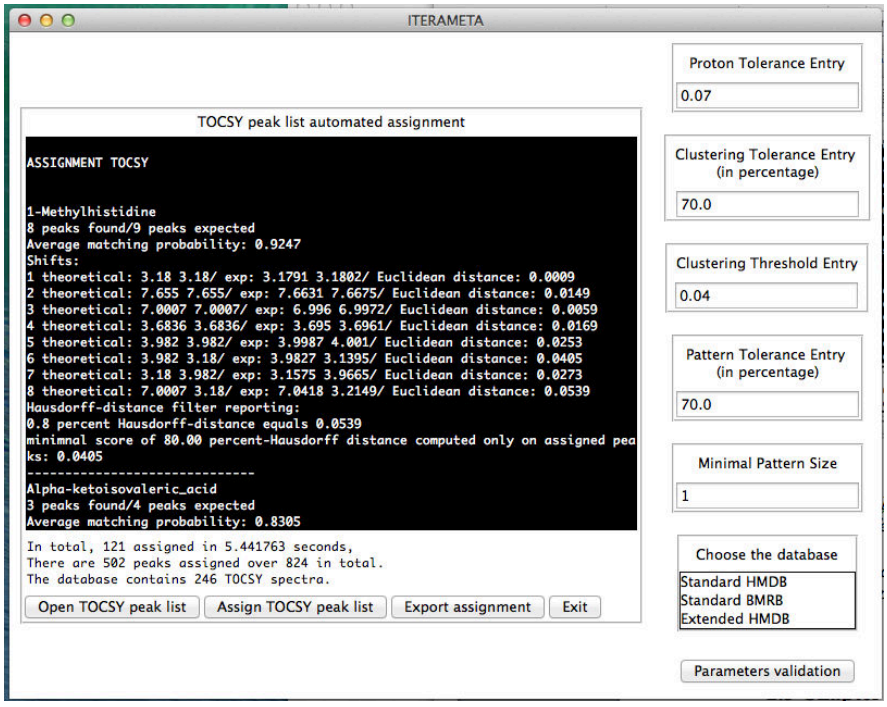



**Figure 2.9.** *ITERAMETA user-friendly interface. Here we showed an example of TOCSY matching. Different matching parameters can be adjusted on the right panel & the main window shows the automated assignment result. For each assigned signal, ITERAMETA*

*also shows its contribution to fractional Hausdorff distance score (Euclidean distance), in order to help the user in assessing the results.*

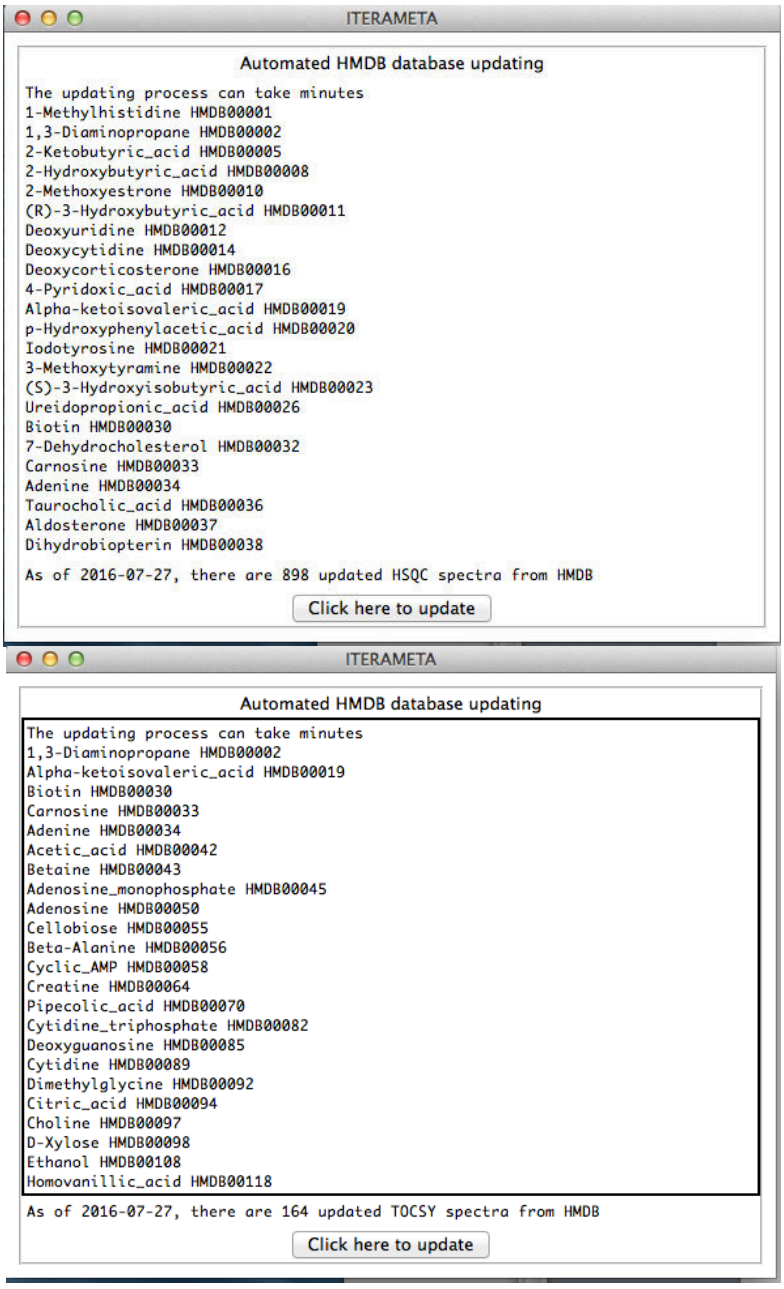



**Figure 2.10.** *In ITERAMETA, the user has also the option to auto-update the HMDB database by pre-downloading the frequently updated database from the homepage of HMDB project.*

The TOCSY standard database in ITERAMETA contains 246 high-quality patterns compared to 164 TOCSY patterns from HMDB, and 972 HSQC patterns compared to 898 from HMDB. Therefore, our *standard* database outnumbers the experimental HMDB. This is the consequence of the fact that we have not only built our database based on the initial experimental patterns issued from HMDB but also manually added (and verified) new patterns from metabolite structure connectivity and individual chemical shift assignment. We qualify our manually created database as *standard*, and the downloaded version of HMDB as *extended*. We equally leave the user the possibility to add new metabolite patterns themselves, in order to broaden his/her specific query library.

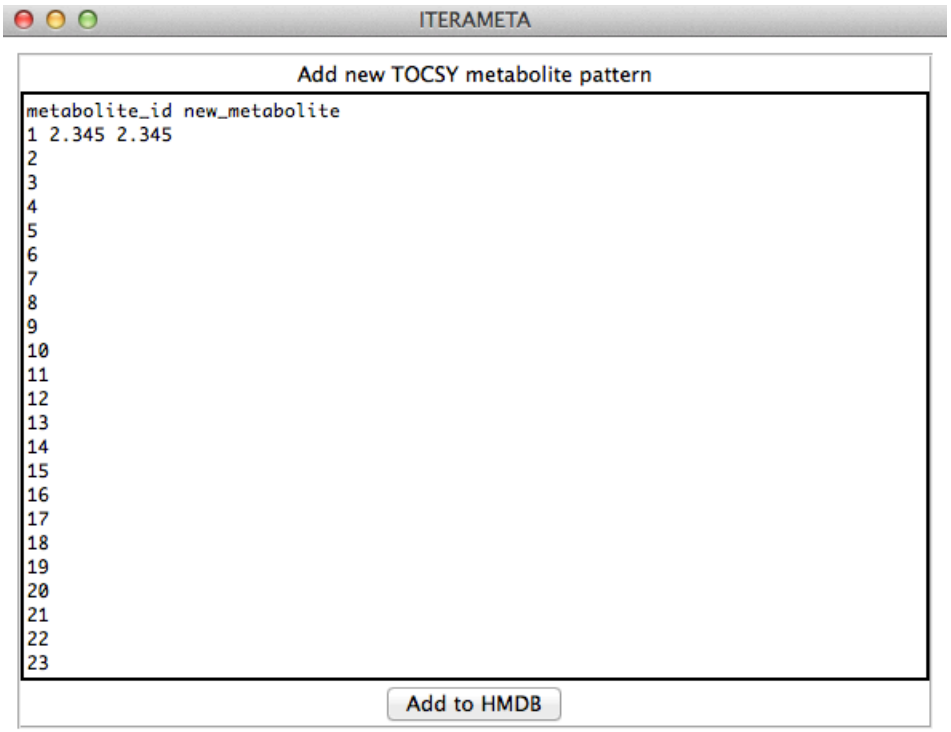



**Figure 2.11.** *The user has the possibility to add new metabolite patterns to their own database for user-specific querying purposes.*

## 2.5 Conclusions

The imperfect matching between a given reference chemical shift library and the input experimental signals, as well as the large variation of chemical shift deviations due to different physiological conditions make automated metabolite profiling a formidable challenge. Here we propose the algorithm ITERAMETA that automatically identifies a significant number of compounds in complex biofluids. Using various quality filters, the assignments are archived with high degree of reliability and good agreement with manual procedures. The algorithm was tested on synthetic peak-lists and on experimental peak-lists issued from either automated peak-picking program or manual peak-picking step. We have also shown on selected examples the advantage of assignment clustering strategy and assignment assessment using Hausdorff-distance concept.

Our results show that automated 2D homo- and heteronuclear TOCSY/HSQC NMR metabolic profiling by ITERAMETA can be a powerful and reliable approach providing an exhaustive listing of metabolites. An automated approach does not suffer from the effects of a hypothesis-driven metabolic research *i.e*. the search for metabolites is not biased by the prior knowledge of the user about metabolic pathways. Our approach allows the user to potentially find new metabolites and therefore new biomarkers and new metabolic pathway.

### 2.5.1. ITERAMETA software availability

ITERAMETA is an open source software tool with a friendly user interface, developed to solve practical problems encountered in metabolomics. The program is written in Python 3.0 with standard libraries, and it is distributed free-of-charge to the academic community.

# CHAPTER 3

## 3. NMR protein structure determination

Beside the numerical developments for NMR metabonomics as presented in the previous chapter, a second major aim of this PhD program was to establish a new automated method for NMR protein structure determination. The desired method follows the stream of "direct methods" rather than "indirect methods" (Guntert 2003), *i.e.*, the desired development of the here proposed numerical approach aims at obtaining protein structures directly from NMR data, without *prior* sequence-specific resonance assignments ("NMR assignment-free method). Before diving into the findings, I believe, however, a detailed review of protein structures and the conventional process of structure determination by NMR spectroscopy is necessary.

## 3.1 NMR in structural biology

NMR has nowadays determined about 12% of protein structures that are deposited in Protein Data Bank (PDB). The overwhelming remaining 87% are done with X-ray diffraction (a small part with neutron diffraction or other techniques, see details given in the Introduction Chapter 1). In X-ray crystallography, a measured diffraction pattern is directly converted into an *electron densities* and therefore into a three-dimensional structure. However the principal experimental quantity in NMR, the chemical shift parameter, is in the conventional workflow not readily converted into a three-dimensional structure.

In conventional NMR, the main source of structural information is taken from inter-proton distances provided by the Nuclear Overhauser effect (NOE), backbone torsion angle restraints determined from scalar coupling constants and orientation restraints (especially residual dipolar couplings). These experimentally determined structural parameters are then used as input for a computational structure calculation

procedure in order to find a set or bundle of structures (conformers) that is in simultaneous agreement with all the experimental input conformational restraints.

The conventional process of the NMR protein structure determination usually consists of the following steps (Wuthrich, 1986): (1) Sample preparation and data acquisition, (2) resonance assignment, (3) extraction of structural restraints from spectra, (4) structure computation, and (5) structure validation (see Figure 3.1).

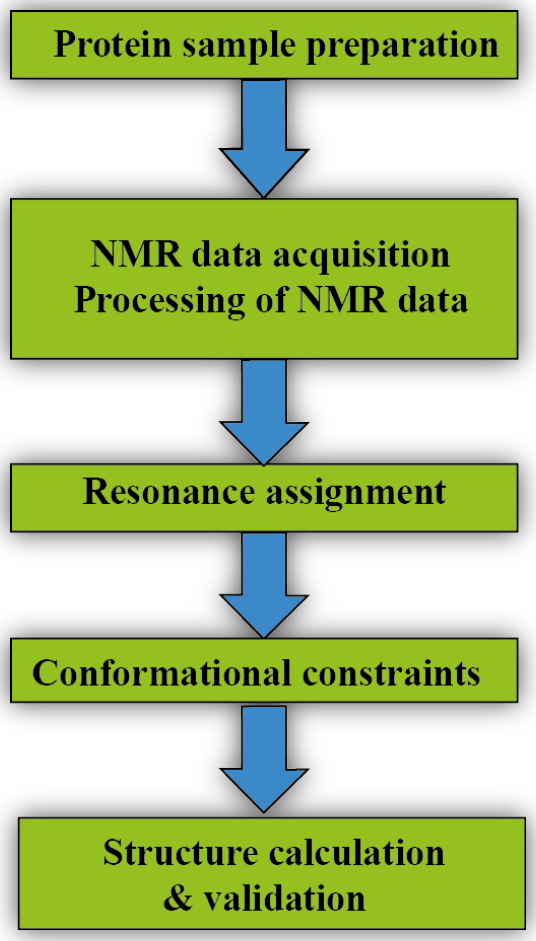



**Figure 3.1.** *Depicted is the conventional workflow of NMR structure determination.*

In this conventional workflow, sequence-specific backbone and side-chain resonance assignment is an *intermediate step* to ultimately translate all available NMR

spectral information into a set of meaningful distance restraints between pairs of hydrogen atoms present in the protein sequence. This intermediate step of sequence-specific resonance assignment forms also the basis for characterizing regular secondary structure elements (helix, beta-strands, coil regions) and three-dimensional structures of proteins (Moseley and Montelione 1999). Conventional automated sequence-specific resonance assignment methods use the same general analysis schema as follows:

1. Grouping of chemical shifts into spin systems that are related to a single amino acid or dipeptide. Many methods gather resonances via common "root" resonances found in all or most of the spectra. Other methods use bond pattern templates to group resonances into spin systems.
2. Amino acid typing. Most programs identify amino acid spin systems with respect to the reference bond-pattern templates.
3. Linking sequential spin systems into segments. There are two major linking methods: deterministic best-first methods and energy optimization algorithms, such as simulated annealing.
4. Mapping spin-system segments onto the protein primary sequence.

An exhaustive search algorithm could map the NMR-identified signals to their most probable positions in the primary sequence. However, the inevitable presence of spectral artifacts as well as spectral overlap in the experimental data is unavoidably inducing ambiguities into the available sequential information. Therefore, the computational time needed for an exhaustive search of the corresponding configuration space is exponentially growing with increasing protein size. This combinatorial explosion makes the development of highly sophisticated assignment algorithms necessary, exhaustive search algorithms being typically only successfully applicable for small to medium-sized biological systems with optimal, high quality data, in order to achieve reliable and robust results.

## 3.2 Structure-oriented methods for protein NMR structure determination

A successfully applied project used for conventional protein NMR structure determination relies on nearly complete sequence-specific resonance assignment of the resonance frequency of atoms in the protein sequence (Wuthrich 1986). A 2011-2012 review by Guerry and Herrmann (Guerry and Herrmann 2011, 2012) lists 44 publications of automated or semi-automated programs performing automated chemical shift assignment; 19 of which work exclusively on (manually prepared) peak-lists and others require additional input information such as grouping of resonances into spin systems, partial assignments, residual dipolar coupling or available information about an initial three-dimensional fold.

Backbone sequence-specific resonance assignment can generally be obtained in a reasonable amount of time. A number of algorithms (non exhaustive list is given next) are available for this purpose, such as AUTOASSIGN (Zimmerman et al. 1997), MATCH (Volk, Herrmann, and Wüthrich 2008) or MARS (Jung and Zweckstetter 2004), PASA (Xu et al. 2006), RASP (MacRaild and Norton 2014). However, much fewer computational developments have been seen for the subsequent step for obtaining side-chain resonance assignment with numerical algorithms, one of the very few examples frequently used is ATNOS/ASCAN (Fiorito et al. 2008). Most popular NMR programs address only the NOE assignment process: ARIA (Fossi et al. 2005; Linge et al. 2003; Linge, O'Donoghue, and Nilges 2001; Mareuil et al. 2015; Nilges et al. 1997) and AutoStructure (Huang et al. 2006). Automation of the entire process of conventional NMR structure determination was so far only proposed by the FLYA procedure (Schmidt and Güntert 2012) and the J-UNIO protocol (Guerry, Duong, and Herrmann 2015; Herrmann et al. 2002a, 2002b; Serrano et al. 2012).

Apart from obtaining sequence-specific resonance-assignment, a further bottleneck for indirect methods consists in the final step of NOE resonance assignment. It is indeed difficult and time-consuming to get near-complete NOE chemical shift-based

assignments, due to missing signals and/or artifacts and noises. As NOE chemical shift-based assignment is a NMR procedure with no biological equivalent (Guntert 2003), it is in theory possible to generate structures without explicit NOE assignment. Such a strategy attempts to model a three-dimensional biomolecular structure as a spatial distribution of covalently unconnected atoms ("gas" of atoms) (Bermejo and Llinás 2008). Gronwald and Kalbitzer refer to this direct structure-oriented methods as "top-down" protocols, compared to the "bottom-up" approach of assignment-oriented procedure (Gronwald and Kalbitzer 2004).

The aim of *direct methods* is to obtain protein structure directly from NMR data (usually NOE data), without prior sequence-specific resonance assignment. Direct methods seek to bypass the time-consuming sequence-specific backbone, side-chain and NOE resonance assignment process by translating directly the spectral information into distance restraints between (unassigned) pairs of atoms. These atoms are unassigned (hence the classification of *assignment-free* methods) and only labeled by their chemical shifts (or resonance frequency). No *prior* covalent connectivity is known.

The fundamental idea behind direct methods is rather simple: the presence of a NOE signal (not artifact or noise) implies the presence of two nuclear spins (or group of chemical equivalent atoms) and the spatial proximity between them. Four distance restraints are necessary and sufficient to define a unique spin position in real space (4n+1 distances for a system of 3n-6 degrees of liberty). The resonance assignment as a by-product in itself brings along two additional sources of information:

1. The specific position of each nuclear spin in the primary sequence of the protein.
2. A supplementary list of covalent distance restraints due to molecular structure constraint of the system of investigation.

### 3.2.1. Nuages

The direct approach was first mentioned in 1992 as *nuages* (clouds in French) using simulated NOE data from the X-ray crystallographic structure of lyzosyme (129 residues) and only $^{1}H^{N}$-$^{1}H^{N}$ NOE were considered (Malliavin et al. 1992). Using a cut-off distance of 4.5 Angstrom, simulated NOE data yielded 302 distance restraints, considered determined with 5% precision. Relative to the crystallographic coordinates, the overall accuracy of the $^{1}H^{N}$ -only clouds was poor (11.47 angstrom RMSD); however, regions corresponding to elements of secondary structure were more accurately determined. An $^{1}H^{N}$ chain was then threaded in each cloud by assuming likely sequential $^{1}H^{N}$ -$^{1}H^{N}$ distances. Importantly, the directionality of the primary sequence was not mentioned (Malliavin et al. 1992).

### 3.2.2. ANSRS

Kraulis reported the ANSRS (Assignment of NOESY Spectra in Real Space) algorithm in 1994 (Kraulis 1994). The algorithm needs two sets of data inputs: the first data set is a list of all detectable $^{1}H$ spins of the protein under consideration, with chemical shifts of the proton nuclei and their covalently attached heavy atoms ($^{13}C$ or $^{15}N$); the second data set is a list of the distance restraints derived from all observable NOEs in the protein (assumed to be derived from 3D or 4D NOESY).

The algorithm hence generates proton clouds or three-dimensional real-space structures of the unassigned $^{1}H$ spins from the NOE distance restraints via a restraint molecular dynamics or simulated annealing (rMD/SA) protocol (Nilges et al., 1988a, Nilges et al., 1988b, Nilges et al., 1988c) with combined restraint molecular dynamics (Kaptein et al., 1985; Clore et al., 1985).

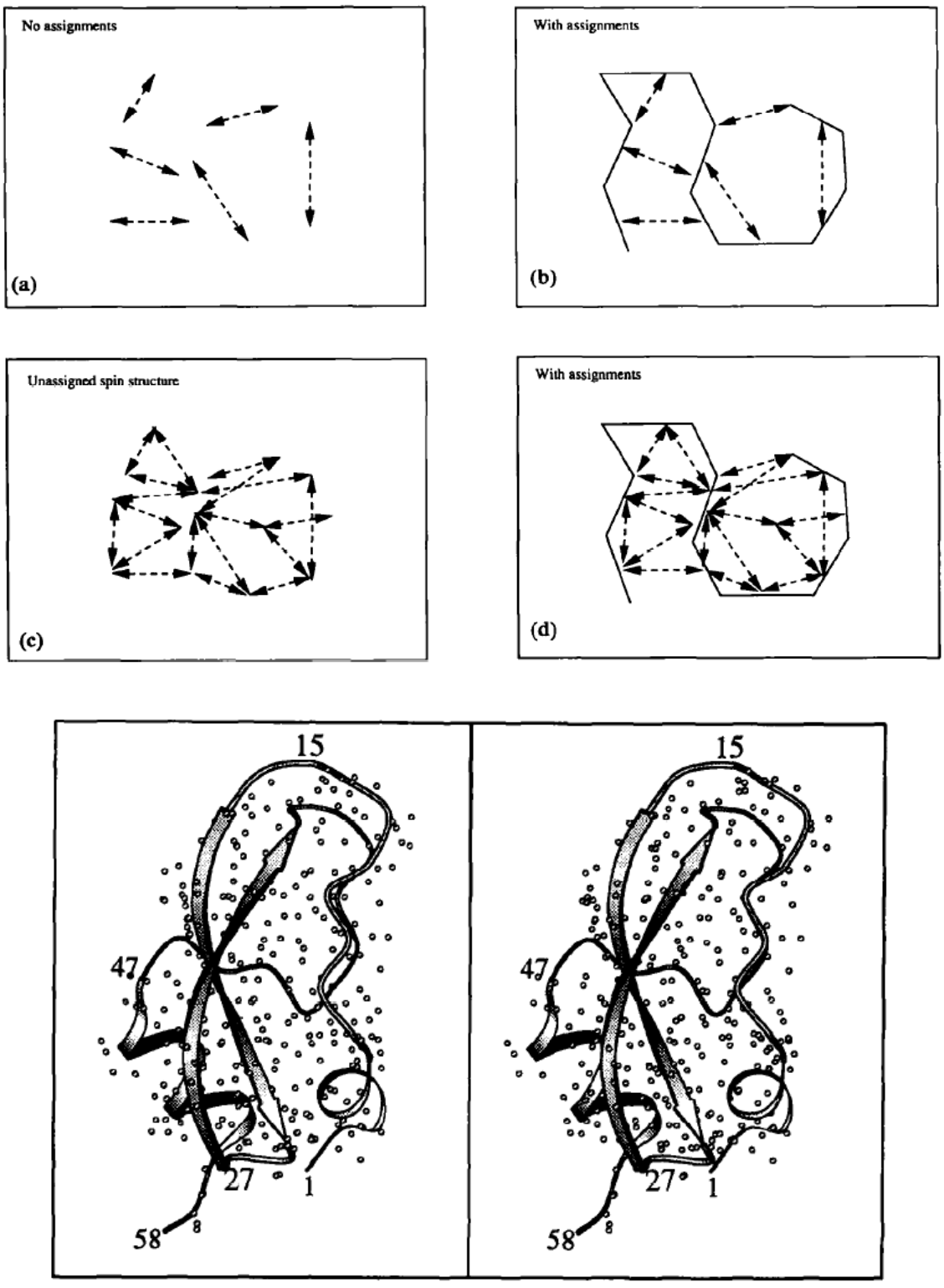



**Figure 3.2.** *Illustration of the workflow for direct structure determination by ANSRS.*

In the ANSRS approach, the molecular system is only roughly mimicked, *i.e.*, the simulated annealing (SA) procedure is only applied to one single type of atom that represents the unassigned $^{1}H$ spins, and also the force field comprises only terms for the *van der Waals* repulsive potential between the atoms and attractive potential terms for the NOE distance restraints. There is no explicit term for imposing the covalent peptide structure. The *van der Waals* repulsion force constant is set to a low value at the beginning of the SA process when the initial temperature is high and is steadily increased during the simulation when the temperature is reduced towards zero.

ANSRS was tested on a segment of the DNA binding domain of GAL4 (residues 9-41) and the bovine pancreatic trypsin inhibitor (BPTI, 58 residues), on the basis of experimental chemical shifts and *simulated* NOE distance restraints corresponding to a cut-off inter-proton distance of 4 Angstrom in the reference structures. In both cases, average proton clouds exhibited less than 2 Angstrom root mean-square deviation (RMSD) from the reference structures.

### 3.2.3. CLOUDS

Grishaev and Llínas reported in 2002 a complete direct approach using experimental NOE data (Grishaev and Llinas 2002). CLOUD relies on precise and abundant inter-proton distance restraints calculated via a relaxation matrix analysis of sets of experimental NOE cross-peaks. The protocol was tested on the col 2 domain of human matrix metalloproteinase-2 (60 residues) and the kringle 2 domain of human plasminogen (83 residues), starting from a list of unassigned, *unambiguous* experimental NOESY data available from the *previously assigned structures* (Briknarová et al. 1999; Marti, Schaller, and Llinás 1999).

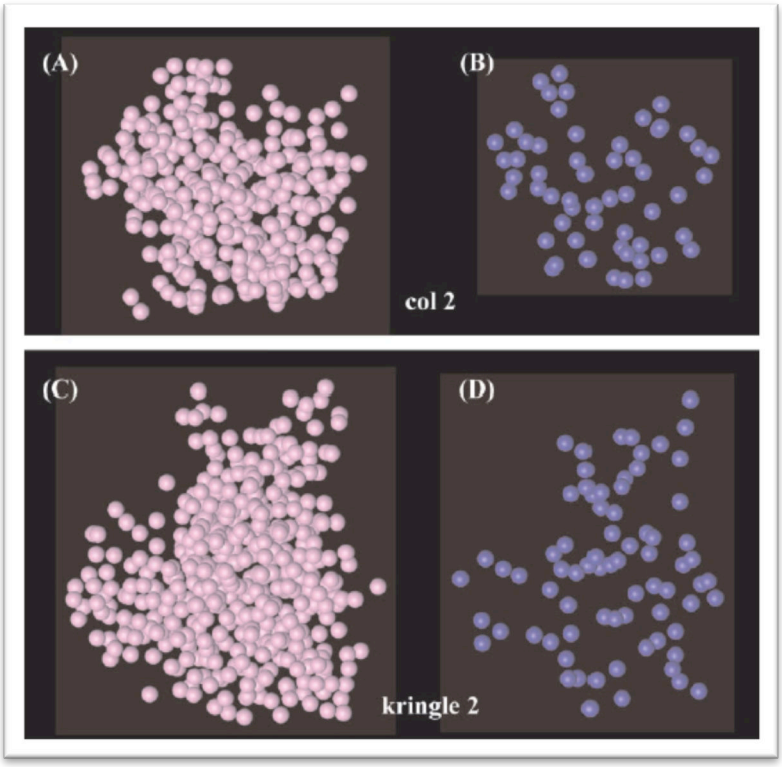



**Figure 3.3.** *Individual clouds for col 2 (*A *and* B*) and kringle 2 (*C *and* D*). All* $^{1}H$ *atoms are included in* A *and* C*;* $H^{N}$ *atoms only are shown in* B *and* D*. The illustrated clouds are those closest to the average minimal RMSD.*

Unassigned hydrogen atoms (labeled solely by their chemical shifts) are extracted from standard multidimensional experiments and listed. NOE inter-proton distances are obtained by relaxation matrix analysis. A force field consisting only of NMR-derived distance restraints and a repulsive *van der Waals* term was applied to an initial gas of randomly distributed proton atoms. In the applied process, average proton clouds exhibited more than 0.8 Angstrom $^{1}H^{N}$ RMSD from the known NMR structures are discarded.

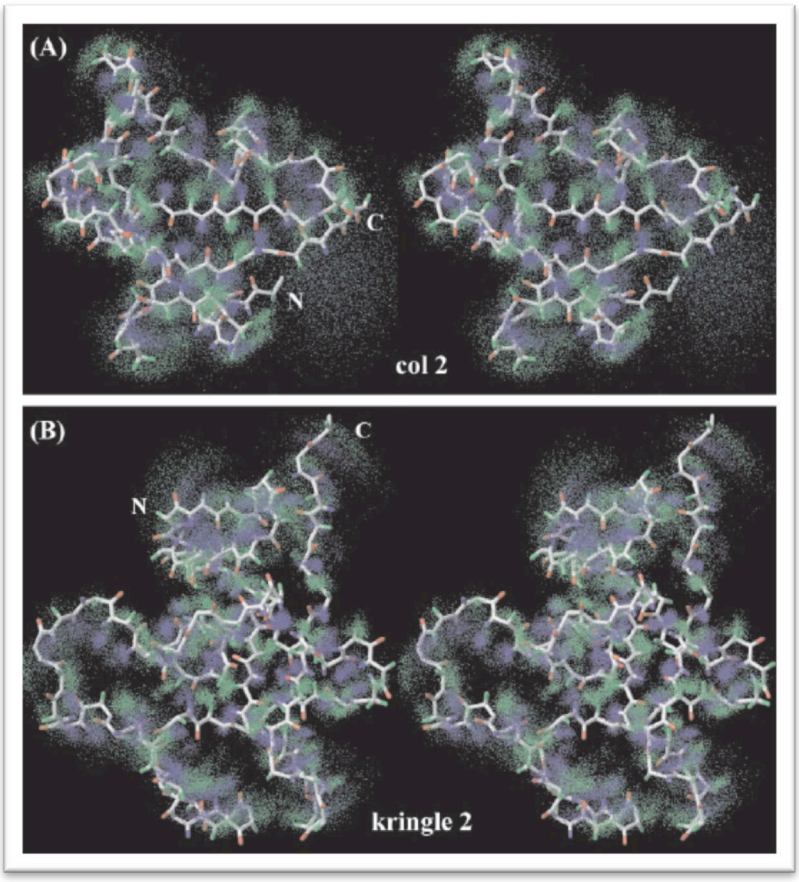



**Figure 3.4.** *Stereo views of molecular focs (backbone H atoms only). (A) col 2; (B) kringle 2. $H^N$ and $H^\alpha$ atoms are shown in blue and green, respectively. The illustrated focs are all-cloud overlaps by reference to the cloud closest to the average.*

The polypeptide backbone is traced through the HN and HA atoms in the clouds via a Bayesian approach where the probabilities of sequential connectivity hypotheses are inferred from likelihoods of HN-HN, HN-HA and HA-HA distances, as well as chemical shifts, derived from public databases. Once the polypeptide sequence of (HN, HA) atoms becomes identified, a similar procedure is followed to link the side chain protons to the main chain.

In 2008, Bermejo and Llínas proposed the sparse-constraint CLOUDS (SC-CLOUDS) (Bermejo and Llinás 2008). These sparse distance constraints were obtained

from a highly deuterated protein. While NMR spectra of highly deuterated proteins give rise to less signal overlap hence less assignment ambiguity, but on the downside they also include a smaller number of distance restraints available for modeling the protein fold.

In order to compensate the inherent loss of distance restraint information, the authors proposed to include a number of "anti-distance constraints" (ADC) in the structure calculation or simulated annealing process. The effect of ADCs and chemical shift degeneracy on the accuracy of proton cloud calculations is already detailed (Atkinson and Saudek 2002) with the protein model BPTI (58 residues). ADCs assume that when an NOE signal is not observed between a pair of proton nuclei, the protons are likely to be separated by a longer distance than an usual NOE cutoff of 5-6 Angstrom; thus in the molecular dynamics calculation, their non-proximity is well-kept by the application of an additional repulsive atom-atom potential (Bruschweiler et al., 1991; Rejante and Llinas, 1994). In SC-CLOUDS, ADCs are based on NOE intensities simulated from a structural database of known proteins (Bermejo and Llinas, 2008).

The proposed SC-CLOUDS approach was tested with experimental three-dimensional $^{15}$N- and $^{13}$C-edited NOESY data on the Z domain of *Staphylococcal* protein A (58 residues). A total of 234 NOEs and 4483 ADCs were included in the calculation. One can observe that the used ADCs outnumber the NOEs by more than 19:1. While ADCs had a significant effect in preventing the collapse of the cloud under the attractive NOE forces, there is much uncertainty to call for the existence of an arbitrary ADC: signal missing in NOESY spectra is common and even more common in highly deuterated samples. The higher quality of NOE data comes at a cost: lesser quantity.

The chain-tracing (cloud interpretation) algorithm used in SC-CLOUDS is noteworthy. In the CLOUDS version, backbone and side-chain proton clouds were identified via a Bayesian protocol based on proton-proton distance distributions derived from high quality public database (CLOUDS 2002). The protocol assumed that distances within a cloud comply with the distance distribution derived from the database, hence assumes a high-quality (highly accurate) computed cloud. Clouds generated from sparse

data are of lower accuracy hence the reliance on the database distribution is not warranted.

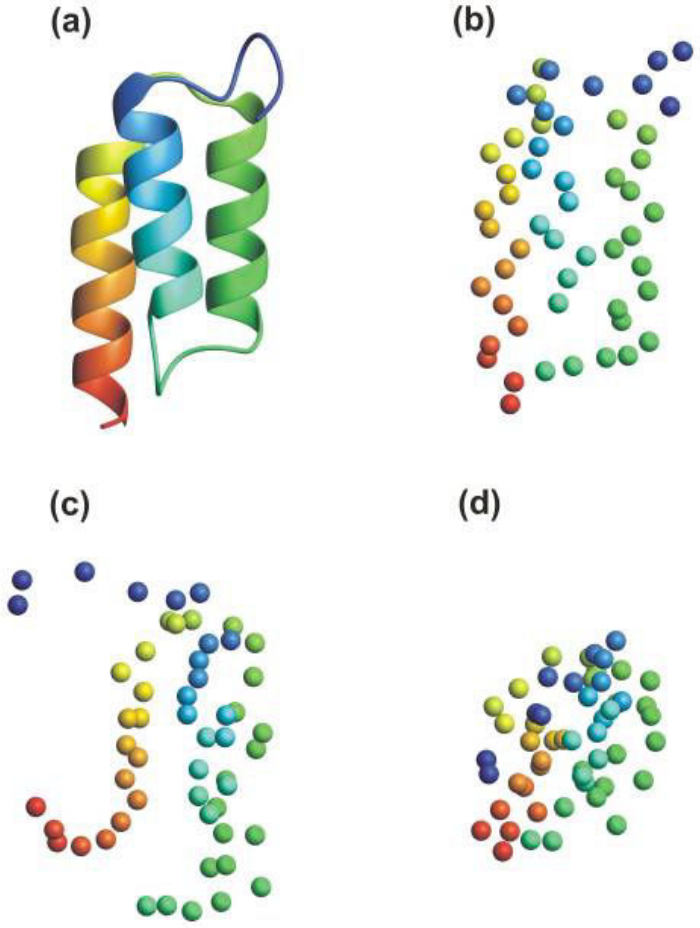



**Figure 3.5.** *SC-CLOUDS on the Z domain of Staphylococcal protein A. (a) Backbone trace of reference NMR structure generated via an assignment-oriented method using a fully protonated sample (PDB code: 2spz). (b) Backbone HN atoms from 2szp. (c) SC-CLOUDS-derived backbone HN atoms using ADCs (lowest-energy cloud). (d) SC-CLOUDS-derived backbone HN atoms without using ADCs (lowest-energy cloud). The coloring represents a blue (N-terminus) to red (C-terminus) gradient.*

The chain-tracing (cloud interpretation) algorithm used in SC-CLOUDS is noteworthy. In the proposed CLOUDS method, backbone and side-chain proton clouds

were identified via a Bayesian protocol based on proton-proton distance distributions derived from high quality public database (Grishaev and Llinas 2002). The protocol assumes that proton-proton distances within a cloud comply with the distance distribution derived from their customized database, hence assumes a high-quality (highly accurate) computed cloud. Clouds generated from sparse data are of lower atomic precision and accuracy hence the reliance on the database distribution is not warranted.

The chain-tracing algorithm is based on the ARP/wARP method for building a CA-chain from an X-ray electron density map (Morris, Perrakis, and Lamzin 2002, 2003). ARP/wARP was the first automatic interpretation tool successfully used to establish protein models from X-ray electron density map and remains one of the most used tools in the crystallographic community for 3D map interpretation. It focuses on the best placement of individual atoms in the map and requires in general an atomic resolution of 2-3 Angstrom or higher in order to produce an accurate, reliable trace. Given a map, it can form a backbone trace by looking for pairs of atoms that are separated by a proper distance. The algorithm verifies candidate pairs by overlaying them with a peptide template; if there is a match between the template and the map, the algorithm saves the candidate pairs. Given a chain of candidate CA pairs, ARP/wARP considers all possible connections between those pairs in order to extend the main-chain connections.

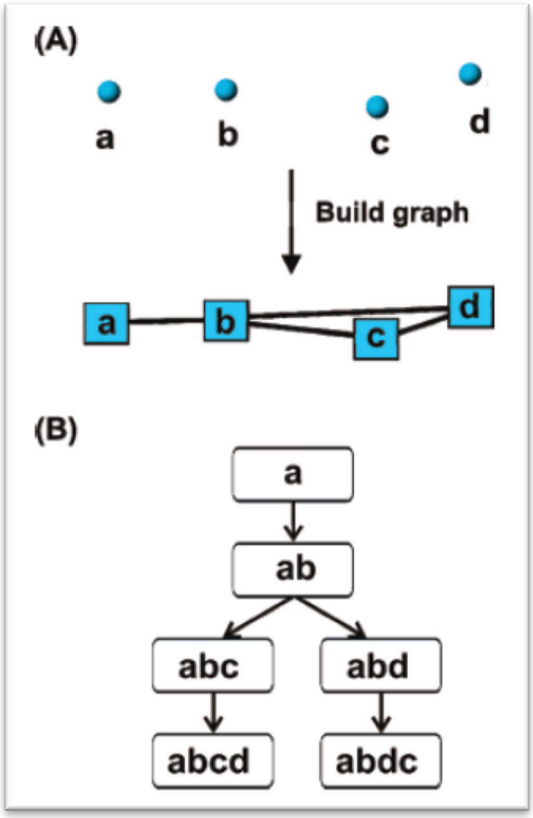



**Figure 3.6.** *The chain-tracing algorithm in SC-CLOUDS is based on a graph search for a chain of connected backbone amide* $^{1}H^{N}$ *hydrogen atoms. Exhaustive search being computationally impossible are excluded, the search is limited with candidates within a distance cutoff of 5 Angstrom.*

## 3.3 Description of the DINO approach

Throughout the brief description of the very few currently existing "NMR resonance assignment-free" or direct methods mentioned above, one can state that for all so far approaches proposed the required input information consists either of simulated, synthetic data or experimental NMR data exhibiting nearly *perfect* quality of NMR data. The most successful, yet still pure theoretical, non-practical NMR assignment-free method to date, CLOUDS or its updated version SC-CLOUDS, uses abundant and unambiguous NOE data to obtain three-dimensional structures of biomolecules. In

general, such (unrealistic) input data is very likely difficult to obtain for real-world applications, due to the presence of unavoidable artifacts and spectral overlap or others.

Most current manual or automatic sequence-specific resonance assignment approaches rely on a large suite of triple resonance NMR spectra, *e.g.*, HNCA, HNCACB, HN(CO)CACB in order to establish sequential connectivities of adjacent spin systems in the polypeptide backbone chain of $^{13}C$ and $^{15}N$ doubled labeled protein samples. Isotopic labeling of a protein required by most automated sequence-specific assignment algorithms is expensive, but more importantly also the required NMR acquisition time is very lengthy, and the following NMR data analysis might - despite the use of unattended approaches – remain cumbersome. On the other side, determining biomolecular structures using solely a small set of NMR spectra raises a number of very challenging algorithmic pattern matching and combinatorial issues, so one needs to find a reasonable balance for the development of a novel NMR structure determination procedure between the desired limited number of input NMR experiments and the possibility for achieving reliable and robust NMR structure determination. In the following, we introduce a new method for direct NMR structure determination without explicit *prior* sequence-specific resonance assignment that is dubbed as DINO (Direct NOE structure determination).

The DINO algorithm utilizes the small set of experimental NMR spectra as follows:

1. NOESY data: The 3D $^{15}N$-edited NOESY correlates an amide proton $^{1}H^{N}$ and its covalently bound $^{15}N$ heavy atom with another hydrogen atoms that are in spatial proximity of less than 5.5Å. The aliphatic and/or aromatic $^{13}C$-edited NOESY correlates a proton and its directly bound carbon atom with a second proton that gives rise to dipolar interactions within a distance of less than 5.5Å.
2. HCCH-TOCSY data: The 3D $^{13}C$-edited TOCSY correlates a carbon-proton root with another proton that is covalently bonded. Within the DINO method, the TOCSY spectrum is used to extract amino acid spin systems through a partitioning algorithm.

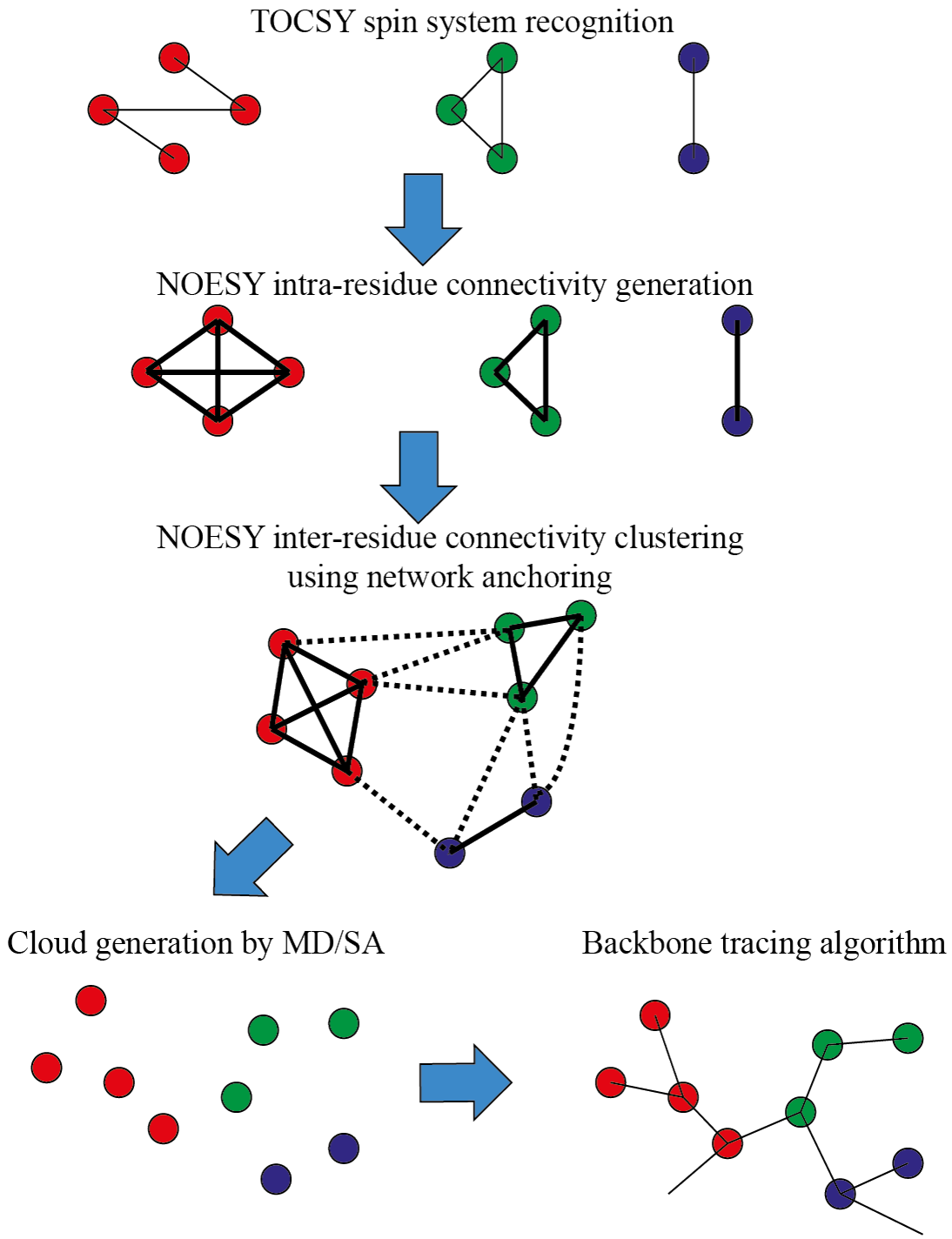



**Figure 3.7.** *Schematic illustration of the DINO workflow for direct NMR structure determination.*

### 3.3.1. TOCSY and NOESY spin system recognition

During the spin system forming process, the unassigned resonance frequencies or chemical shifts from the input NMR signals are gathered together in order to build spin systems of each amino acid in the protein sequence. The set of unassigned spin systems is a major building block used for the following processes of defining "interatomic connectivities" and "backbone tracing". A novel clustering-based method is here developed so to group the resonance frequencies of different spectra into individual spin systems. The input for the DINO spin system identification module is given by 3D HCCH-TOCSY, $^{15}$N-and $^{13}$C-resolved ($^{1}$H, $^{1}$H) NOESY spectra.

**Spin system identification by analyzing 3D HCCH-TOCSY data:**

To extract spin systems present in the protein sequence from 3D HCCH-TOCSY data, we use a numerical method that has previously been introduced (Li and Sanctuary, 1996). However, the originally proposed workflow of the algorithm by Li and Sanctuary was modified so to fit out input data:

1. Search the HCCH-TOCSY cross peak list for pairs of $(H_i, C_l, H_j)$ and $(H_{i'}, C_{l'}, H_k)$ where $H_i$ and $H_{i'}$ are within the $^{1}H$ chemical shift range ($tol_H$), and $C_l$ and $C_{l'}$ are within the $^{13}C$ chemical shift range ($tol_C$).
2. If a HCCH-TOCSY $(H_j, C_m, H_k)$ is found and a HCCH-TOCSY $(H_j, C_m, H_i)$ is found; then add $H_i$, $C_l$, $H_j$, $H_k$ and $C_m$ to a spin system.
3. Else if a HCCH-TOCSY $(H_k, C_m, H_j)$ is found and a HCCH-TOCSY $(H_k, C_m, H_i)$ is found; then add $H_i$, $C_l$, $H_j$, $H_k$ and $C_m$ to a spin system.
4. Back to step 1, until no more TOCSY cross peak pair fulfilled the condition of step 1 remain in the data set.

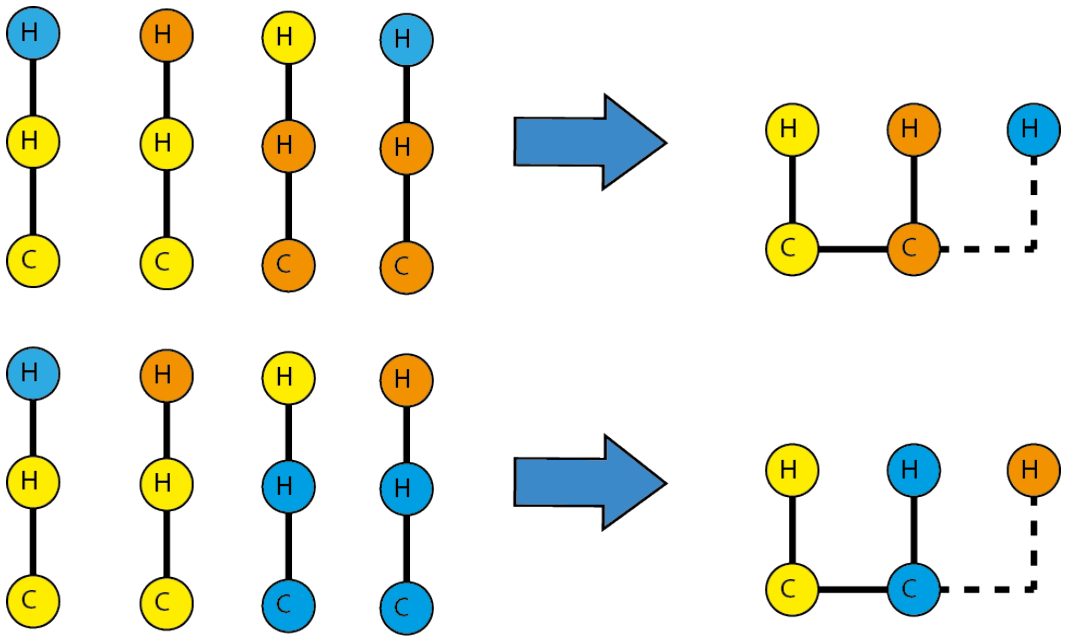



**Figure 3.8.** *Spin system identification using 3D TOCSY data.*

**Root identification by analyzing 3D NOESY data:**

Experimental NOE data from standard $^{13}$C- and $^{15}$N-edited NOESY experiments consist of a list of spectral NMR singlas with their spectral coordinates $(\delta_i^H, \delta_j^H, \delta_k^X)$ where $\delta_i^H$ and $\delta_j^H$ are the chemical shifts of protons $i$ and $j$, respectively, and $\delta_k^X$ (X = $^{13}$C, $^{15}$N) is the chemical shift of the heavy heteroatom $k$, directly bonded to proton $j$. In addition to the spectral coordinates, each NMR signal can be via its peak volume associated to an upper distance restraint value (in Ångstrom). The process of peak intensity or volume calibration individually performed for each input experimental peak-list yields its corresponding upper distance restraint value list. Conventionally, each peak-list is calibrated using the isolated-spin-approximation (IPA) (embedded as a routine in UNIO) (Herrmann et al. 2002b) with lower and upper bounds set to 2.4 and 5.5 Å., respectively.

The connection between a proton and its directly bound heavy atom is defined as a root. Each spin system is a vector of chemical shift values. Therefore, each spin system

issued from HCCH-TOCSY includes a set of roots that are likely to belong to the same residue in the sequence.

### 3.3.2. NOESY inter-residue connectivity clustering using network anchoring

The network-anchoring approach was proposed (Herrmann et al., 2002a) in order to reduce the number of initial chemical-shift based NOE assignment possibilities during an iterative NMR structure determination process. The approach exploits the fact that any network of correctly assigned constraints forms a self-consistent subset within the initial network of constraints. Each initial assignment is weighted by the extent to which it can be embedded into the network formed by all other NOE peak assignments.

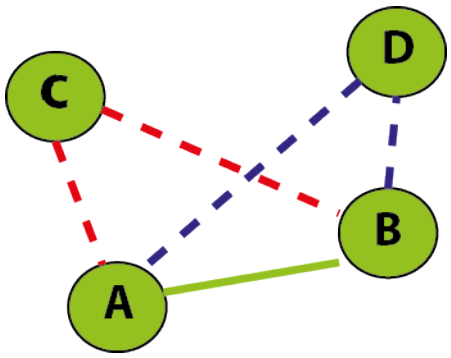



**Figure 3.8.** *NOE assignment attributed to the interaction between atom A and B is reinforced by the possible assignments of related constraints.*

Network anchoring evaluates the self-consistency of NOE assignments independent of knowledge on the three-dimensional structure, thus compensates for the absence of 3D structural knowledge at the outset of a *de novo* structure calculation.

Notably, this process of network-anchoring guided determination of interatomic proximities – originally developed for the conventional procedure of assigned chemical-shift based NOE assignment - does not at all rely in theory on *prior* knowledge of sequence-specific resonance assignment, but only on the mutual NOE connectivity support between individual atoms or in our case between unassigned resonance frequencies. This fact is exactly exploit in the present DINO approach.

**Mathematical formulation of network-anchoring:**

Consider two unassigned spin-systems $i$ and $j$. We denote $A_{ij}$ as the interaction probability between two spin-systems $i$ and $j$. If no resonance assignment ambiguity exists, *i.e.*, the interaction between $i$ and $j$ is known, then:

$$A_{ij} = \begin{cases} 0\ if\ no\ interaction \\ 1\ if\ interaction \end{cases}$$

However, since the interaction is not known in advance, one can only quantify the interaction by a probability. We construct a model to assess this interaction probability based on the network anchoring algorithm.

Consider an experimental peak $p_k$ issued from a NOESY peak list. If $p_k$ can be assigned to $A_{ij}$ then we have the following Bayesian probability formulation:

$$P(A_{ij}|p_k)\times P(p_k) = \frac{P(p_k|A_{ij})\times P(A_{ij})}{\sum_{A_{ij}} P(p_k|A_{ij})\times P(A_{ij})}$$

with $P(A_{ij}|p_k)$ the probability that the signal $p_k$ emerges from the interaction $A_{ij}$; $P(p_k|A_{ij})$ the probability there is a signal close to the coordinates of $p_k$ given the interaction $A_{ij}$; $P(A_{ij})$ is the probability of interaction between two spin-systems $i$ and $j$; $P(p_k)$ is the probability of existence of the signal $p_k$ which is equal 1, $P(p_k) = 1.0$.

One observes that the probability $P(p_k|A_{ij})$ is proportional to $P(A_{ij})$. If there is no interaction between two spin-systems $i$ and $j$, *i.e.*, $P(A_{ij}) = 0$, then $P(p_k|A_{ij})$ will be also zero.

$$P(p_k|A_{ij}) = k_{emission} \times P(A_{ij})$$

By supposing the same emission probability $k_{emission}$ for each spin-system, we can simplify the above original Bayesian probability formulation to

$$P(A_{ij}|p_k) = \frac{P(A_{ij})^2}{\sum_{A_{ij}} P(A_{ij})^2}$$

The probability of interaction between two spin-systems $i$ and $j$, $P(A_{ij})$ is computed as the number of signals that can be attributed to the interaction, weighted by their mean probability or a given maximal threshold number:

$$P(A_{ij}) = \frac{\sum_{p_k} P(A_{ij}|p_k)}{\max(threshold, \sum_{p_k} 1)}$$

### 3.3.3. Clouds generation by rMD/SA

Molecular dynamics (MD) algorithm solves Newton's equation of motion in order to obtain a trajectory for the molecular system. Standard MD tries to simulate the behavior of a real physical system as close as possible. MD used for NMR structure calculation searches the conformational space of a given protein for the 3D structure or bundle of conformers that simultaneously fulfills all the experimental restraints and simplified physical forces (hybrid force field) using simulated annealing (SA) with a hybrid target energy function that controls this process.

Multiple independent simulating annealing runs are performed, and a subset of *converged* conformers is selected by the values of their target energy function. A single starting structure (randomly generated) is heated to a high temperature in this simulation. During many discrete cooling steps, this starting structure can evolve towards the energetically most favorable final structure under the influence of the hybrid force field derived from the experimental restraints, and typically only using the *van de Waals* repulsive potential (Guntert 2002).

In order to generate a cloud of unassigned atoms in close spatial proximity from the given list of most likely connected intra- and inter-residual atoms, we implemented a simulated annealing procedure in Cartesian space using the program UNIO (Guerry et al. 2015; Serrano et al. 2012) to perform such rMD optimization. It is worth mentioning that such optimization runs product due to the limited information available – covalent connectivities are *a prior* unknown - also mirror image structures that can however be easily removed by simple geometric analysis using Ramachandran statistics about allowed combinations of backbone torsion angles (applied after the process of "backbone tracing").

### 3.3.3. Dynamic backbone tracing algorithm

The intermediate result of the previous step is a set of clouds of atoms. The challenging problem is now to fit the polypeptide chain (protein primary sequence) into this cloud of more or less accurately defined atoms. At the outset of the calculation, this initial cloud, obtained by molecular dynamics/simulated annealing, does not generally provide exact distances between covalently connected atoms. The fundamental idea of the algorithm is to recursively constructing a search space having the structure of a tree, and by verifying the feasibility of any forward connection, either to move back then try another possibility, or to move forward then try the next possibility. The sequence fitting algorithm has recently been employed to enumerate all possible protein conformations that verify a set of distance constraints, using an interval Branch-and-Prune algorithm

(Cassioli et al. 2015). This approach requires however a complete backbone and side-chain assignment as well as the corresponding assigned distances.

Given the cloud structures from the previous molecular dynamics step, a downstream analysis with known residue assignment yields an rmsd of a range of 3-4 Å, removing mirror image structures (see Results section). This quality range of structures does provide an initial good estimation of the target structures, but the quality is not high enough to execute the backbone tracing used in SC-CLOUDS. The latter algorithm uses a cut-off search of 5 Å. Notably consider that real experimental data is used here (see the Results section).

Therefore we designed an algorithm to generate an initial primary sequence fitting independently of the knowledge of the intermediate atomic low resolution cloud structure. The *backbone tracing* algorithm takes as inputs the previously determined spin systems and the spatial interaction network between them. The spatial interaction network is the result of the precedent step that includes both correct and incorrect constraints. By further analyzing the outcome between the sequence mapping, amino acid typing and the spatial interaction network, the algorithm performs a restraint violation analysis to further remove incompatible interactions between atoms. In the following iteration step, this refined atomic distance restraint list together with the available covalent information from the backbone tracing analysis will lead to more precise three-dimensional model and so on.

The DINO sequence fitting algorithm is based on a previously published algorithm GANA (Lin et al. 2005) for obtaining sequence-specific backbone assignment. Taking spin systems as input data and using two data structures, GANA uses a genetic algorithm to automatically perform backbone resonance assignment. Two data structures are the candidate lists and adjacency lists to assign the spin systems to each amino acid of a target protein. To assign chemical shifts and make sequential resonance assignment on backbone structures, GANA uses the data from 2D HSQC and 3D NMR experiments CBCANH and CBCA(CO)NH. We have correspondingly adapted the GANA algorithm to serve our purpose in order to achieve backbone tracing of unassigned spin systems.

**Amino acid typing:**

A spin system contains (all) the chemical shifts of atoms within a residue. Two consecutive or sequentially adjacent residues, the $(i-1)$ and $(i)$ residues, have better chances of issuing NOE signals than two residues that are far from each other in the protein sequence. Mutually, given two spin systems whose interaction produces NOE signals, amino acid typing allows the placing of these spin systems in consecutive positions in the protein sequence with a matching probability to assess how well the couple is placed in those specific positions.

Given two spin-systems $SS_i$ and $SS_j$ whose interaction expresses NOE signals and two consecutive positions $A_n$ and $A_{n+1}$ in the primary sequence, the matching probability to assess how well the couple of spin-systems can be fitted in the position of the protein sequence is:

$$P_{(i\to n, j\to n+1)} = \sum_{(state_n, state_{n+1})} P(A_n, state_n | SS_i) \times P\left(A_{n+1}, state_{n+1} \middle| SS_j\right) \times P_{transition}(state_n | state_{n+1})$$

with $P(A_n, state_n|SS_i)$ and $P\left(A_{n+1}, state_{n+1}\middle|SS_j\right)$ the probabilities that the spin-systems $SS_i$ and $SS_j$ can be fitted in the positions $A_n$ and $A_{n+1}$, respectively, using the statistical distribution of their respective states (alpha helix, beta strand, coil). The spin-system-position fitting probability is computed by local chemical shift assignment, as the number of peaks found in the locally assigned spin-system divided by the number of signals that are expected for the amino acid position.

The state transition probability is known in advance by computing a secondary structure database. The construction of the probability formula is based on the statistical observation that the chemical shift assignment has shown the correlation between the

chemical shifts and the amino acid types and underlying secondary structures, as well as some underlying secondary structures are rarely consecutive found next to each other in the sequence. Thereby, an alpha-helix-chemical-shift spin-system has less chance to be found immediately next to a beta-sheet-chemical-shift spin-system.

Chemical shift value of a nucleus depends on the local chemical and local geometric environment. Hence, a structured protein has a dispersed chemical shift distribution. The goal of spin system typing is to reduce the number of candidate spin systems for each position in the sequence.

Chemical shift statistical analysis has shown the correlation between the chemical shifts and the amino acid types and underlying secondary structures. These statistics are used to build a probabilistic model to estimate how likely a spin system can be matched/mapped to a certain position in the sequence. For this purpose, the re-referenced chemical shift database or RefDB (Zhang, Neal, and Wishart 2003) is used for the computation of the spin-system-position fitting probability. The database is assembled from comparing predicted shifts using the program SHIFTX (Neal et al. 2003) to predict protein backbone chemical shifts from X-ray and NMR coordinate data of previously assigned proteins, and the corresponding assigned shifts. The side-chain chemical shift database is completed with experimentally assigned chemical shift values as available in the BMRB database.

Considering only three secondary structure elements: alpha helix, beta sheet and coil, the spin-system-position fitting probability is computed as follows:

$$P(A_n|SS_i) = \frac{E(SS_i)}{E(A_n)}$$

with $E(SS_i)$ the number of matched peaks and $E(A_n)$ the number of expected peaks. The local chemical shift assignment uses a simple matching and counting algorithm within a predefined tolerance range (as default, 0.03 ppm for proton and 0.3 ppm for heavy atoms are used).

**Data structures:**

By setting a user-given threshold value for the probability that two spin-systems $SS_i$ and $SS_j$ are connected, two data structures are created: the candidate list and the adjacency list.

The candidate list is used to record potential spin systems for each residue in the target protein. For each residue $(n)$ in the target protein, the candidate list $CL_n$ records all the spin-systems $\{SS_k\}$ that match residue $(n)$.

Adjacency lists are used to express the consecutive connectivity relations between spin-systems. Each adjacency list $AL_i$ contains two kinds of lists: an L-list (adjacent to the left), denoted by $ALL_i$ and an R-list (adjacent to the right), denoted by $ALR_i$. The L-list records the spin-systems that can be connected to the left and the R-list records those that can be connected to the right of the sequence direction.

If the probability

$$\sum P(A_n, state_n|SS_i) \times P\left(A_{n+1}, state_{n+1}\middle|SS_j\right) \times P_{transition}(state_n|state_{n+1})$$

exceeds the matching probability threshold, then the algorithm adds $SS_i$ to $CL_n$, $SS_j$ to $CL_{n+1}$, $SS_i$ to $ALL_j$ and finally, $SS_j$ to $ALR_i$.

**Genetic algorithm model:**

Genetic algorithms, first proposed by (Holland and Reitman 1977) mimic the Darwinist biological evolution to solve efficiently optimization problem with large search space. Genetic algorithms usually begin with an initial population of chromosomes and a metric to measure the fitness of each chromosome. At each generation, only the top-ranking chromosomes in the population survive. The top half mate with each other, and their offspring constitute the population for the next iteration. When two chromosomes mate, the newborn of the population inherits a new sequence, half of which randomly

comes from the father and from the mother. Mutations are also introduced to allow the algorithm to escape local optimal. The algorithm stops when a maximum number of iterations are reached or a chromosome of the population with maximum fitness is found.

**Chromosome initialization:** Each chromosome is a string of spin-systems and represents a candidate solution for sequence fitting. A chromosome $ch$ has $N$ components corresponding to the size of the target sequence. Each position of $ch$ is denoted by $ch[n]$ that is assigned to either a spin-system or is empty.

Initially, all positions of $ch$ are set as being empty. The algorithm performs iteratively the following steps:

1. Randomly select a position $n$ that is empty.
2. Given the position $n$, randomly select a spin-system $SS_i$ from $CL_n$ that has not been assigned to any other position and assigns $SS_i$ to $ch[n]$.
3. Extend the fragment first to the left by examining $ALL_i$. Sequentially and randomly select a spin-system $SS_{i-1}$ from $ALL_i$ that is also in $CL_{n-1}$, then assign $SS_{i-1}$ to $ch[n-1]$. Repeat the process for the next left positions $(n-1)$, $(n-2)$… until no further extension is possible. Similarly proceed the extension to the right by examining $ALR_i$
4. When performing step 2, if no spin-system from $CL_n$ can be found for position $n$, label $ch[i]$ as empty.

In out approach, the chromosome initialization is repeated 100 times to create the initial population.

**Fitness score:** The fitness score determines the direction of the population evolution, therefore the fitness score is critical for the outcome. For a chromosome $ch$, we proceed from the first position to the last one. The fitness score is initially set to be zero.

1. If $ch[n]$ and $ch[n+1]$ both are not empty i.e. $ch[n] \neq \emptyset$ and $ch[n+1] \neq \emptyset$ and a spatial interaction is found between $ch[n]$ and $ch[n+1]$, then increase the fitness score by the spin-system-position probability previously computed. If no spatial interaction is found, then decrease the fitness score by the same amount.
2. If one of the two positions is empty or both are empty, the fitness score is unchanged.

**Reproduction operations** After ranking chromosomes according to their fitness scores, the top half of the population is kept for the next iteration. We use the top half as parent candidates to generate child chromosomes. The crossover operation between two randomly selected parents produces an offspring that has inherited as many connected fragments as possible from its parents (Lin et al. 2005). Initially, all positions of the child chromosome are set to be empty. The procedure of the crossover operation is as follows:

1. Randomly select a position $n$ of $child$ that is empty.
2. Randomly select a parent $p$ ($p_1$ or $p_2$). If $p[n]$ is empty, then label $child[n]$ as empty. Otherwise, proceed as follows: if $p[i]$ has not been assigned to any other position in $child$, then assign $p[n]$ to $child[n]$.
3. Extend the connected fragment from $child[n]$ by referencing $p$ and return to step 1.
4. If $p[n]$ is already assigned to another position then label $child[n]$ to be empty.

The mutation operation provides the population with reasonable diversity and prevents the solutions from falling into a local optimal. Single-point or multiple-point mutation operations can delete the continuity of the connected fragment, therefore whenever a position is muted; we consecutively modified its subsequent neighbors. A mutation frequency variable is introduced to control how often a mutation can occur.

Let $mch$ denote a new mutated chromosome to be generated and $ch$ the template chromosome for the mutation. All positions of $mch$ are set to be empty. The mutation procedure is as follows:

1. Start with the first position i.e. $n = 1$.
2. If the mutation frequency variable exceeds the mutation threshold, the current position will be muted. Randomly select a spin-system $SS_i$ from $CL_n$ that has not been assigned to any other position of $mch$, and attribute $mch[n] = SS_i$. If no spin-system is qualified, set $mch[n]$ to be empty i.e. $mch[n] = \emptyset$. Then perform only the right extension from $SS_i$ by following the same procedure as described in the chromosome initialization, until no further extension is possible.
3. If position $n$ is not muted, if $ch[n]$ has not been assigned to any position in $mch$ then set $mch[n] = ch[n]$.
4. Proceed to the next position until all positions are processed.

**Structure calculation reiteration:**

After ranking chromosomes in the current generation according to their fitness scores, the top half of the population as candidate solutions are selected for the structure/sequence mapping process, *i.e.,* the most likely position in the protein sequence are getting to be known (in an iterative process). As consequence, a more reliable evaluation of possible interatomic (unassigned) atom proximities can be performed. But equally important, sequential covalent knowledge (polypeptide) can be used as additional interatomic distance restraints in order to successively obtain more and more accurate atomic model of the protein under investigation. Especially the use of covalent polypeptide bond-derived (non-experimental) distance restraints between two consecutive amino acids are decisive for obtaining in an iterative fashion more and more precise and accurate structural models.

## 3.4 Results

Rapid protein structure determination has become attractive in recent years because it can yield structural information for proteins in minimal amount of time providing quick insights to biological functions. The variety of NMR data structure is valuable in this area and offers various (computational) methods to define protein folds. We have described an algorithm dubbed DINO combining a set of TOCSY derived spin-systems and NOE data to provide "acceptable-accuracy" or "medium-accuracy" protein structures. The structure calculation takes less than ten minutes of computation time to obtain protein structures within a range of 2-3 Å RMSD compared to the mean coordinates of the reference structure bundle of our target protein. Although, we will present in the following only the results obtained for one protein data set, this first proof-of-principle shows that the DINO algorithm, has great potential to provide automated assignment-free protein structure determination in a robust way.

### 3.4.1. Collection of distance restraints between spin systems and structure calculation

The performance of the DINO algorithm was tested with the target protein VpR247 (PDB deposition code: 2KIF) that is a 11.5kDa monomeric protein consisting of 102 residues. Experimental (refined) NOESY peak lists were obtained from the CASD-NMR website (www.wenmr.eu/wenmr/casd-nmr-data-sets). Since only NOESY data are provided by CASD-NMR, the 3D TOCSY peak lists were synthetically generated from the deposited chemical shift list (BMRB deposition entry: 16272).

Setting the proton and heavy atom tolerance range to 0.03 ppm and 0.3 ppm, respectively, used to establish spin systems based on the TOCSY input data, the NOESY peak lists are used to extract intra-and inter-residual NOE-derived distance restraints as descripted in the former section. At the outset of the calculation, a total of 2041 upper

distance restraints were computed (see Table 3.1) containing 359 long-range distance restraints between atoms. This list of distance restraints contains a large number of correct restraints between atoms, but also atom contacts that are not compatible with the reference structure of our target protein. This is not further surprising, since the inclusion of artifactual restraints is hard to avoid at the outset of a calculation, since no restraint violation based on a preliminary structural model can be performed.

| | |
|---|---|
| Number of distance restraints | 2041 |
| Number of wrong distance restraints | 92 |
| Number of long-range distance restraints | 359 |
| Number of wrong long-range distance restraints | 88 |

**Table 3.1.** *Statistics about the collected NOE-derived distance restraints in DINO iteration cycle 1. Long-range distance restraints are defined as contact between two residues separated by at least 4 residues.*

This list of distance restraint served then as input for the following structure calculation using a simulated annealing protocol. 20 structures were calculated and the best 5 structures were selected according to lowest residual target function value. The RMSD of the bundle of conformers is 1.77 Angstrom, and the RMSD deviation between the mean coordinates of the structure bundle to the reference protein is 3.34 Angstrom. The clouds of heavy atoms for the best conformer with the lowest residual target function value is shown in Figure 3.9.

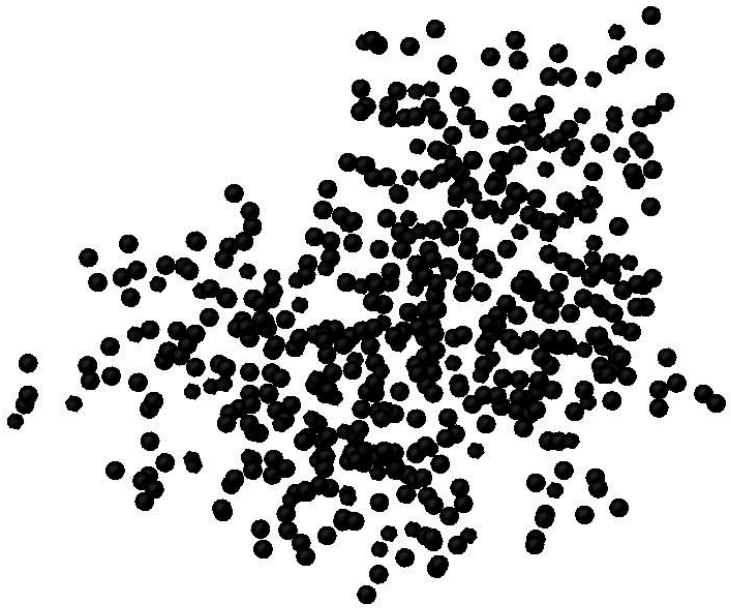


**Figure 3.9.** *Clouds of heavy atom for the best conformer with lowest residual target function value.*

### 3.4.2. Mapping spin systems into the primary sequence

Our mapping algorithm used the following parameters: the number of chromosomes in each generation = 100, the number of generations (iterations) = 100, the mutation rate = 30%. In order to prevent the genetic algorithm to fall into local maximum, we perform multiple rounds to select the chromosome with highest fitness scores.

The GANA algorithm (Lin et al. 2005) from which our algorithm is inspired, reported up to 97-100% of correct matches, on simulated datasets generated from BMRB library.

Without prior structures, our mapping algorithm is capable of creating consistent and correct connected fragments. The mapping showed an average correct positioning of 60-70%. A population of 100 chromosomes is diversified enough to cover the entire sequence.

With the updated distance restraint set obtained after iteration 1 with violated-restraints consensus analysis applied, the mapping results showed up to 90-95% of average correct positioning.

### 3.4.3. Structure calculation quality

**Violation analysis**: The violation analysis algorithm is computed as follows:

1. Start with the first position $i = 1$. Consider another position $j$ that is separated from $i$ more than 6 residues.
2. If $ch[i]$ shows NOE contact with $ch[j]$, consider the NOE contacts between the group $(ch[i-1], ch[i], ch[i+1])$ and the group $(ch[j-1], ch[j], ch[j+1])$, if a number of contacts are detected then confirm the contact between $ch[i]$ and $ch[j]$.
3. If no contact detected other than the one between $ch[i]$ and $ch[j]$, set the contact between $ch[i]$ and $ch[j]$ to be non-existent (i.e. violated restraint).

The basic idea is that there should be at least one contact pair in the neighbor other than the pair itself to support the pair contact. The positions are chosen to be far enough to be consider "long-range" distance restraints.

By repeating the violation analysis over the population of 100 chromosomes, the algorithm builds a consensus of violated distance restraints *i.e.*, the restraints that are weakly supported by other restraints. By reiterating the restraint assessment step with the consensus, we create a better set of restraints between spin-systems.

### 3.4.3. DINO iterative structure calculation

The idea behind the iteratively applied DINO protocol is that (a) the knowledge about intermediate structural models can be successfully exploit for obtaining successively more accurate distance restraint list via violation analysis, and (b) backbone tracing becomes also more and more accurate, and thus covalent bond information can be successively introduces as additional covalent distance restraints during the simulated annealing protocol. This is of special importance for forming the correct geometry of the peptide bonds between adjacent residues. After iteration 3, a total of 2034 distance restraints were computed. This restraint list is virtually artifact free (see Table 3.2).

| | |
|---|---|
| Number of distance restraints | 2034 |
| Number of wrong distance restraints | 3 |
| Number of long-range distance restraints | 357 |
| Number of wrong long-range distance restraints | 3 |

**Table 3.2.** *Statistics about the collected NOE-derived distance restraints in DINO iteration cycle 3. Long-range distance restraints are defined as contact between two residues separated by at least 4 residues.*

In DINO iteration 3, this updated list of distance restraint served then as input for the following structure calculation using a simulated annealing protocol. As in the previous cycles, 20 structures were calculated and the best 5 structures were selected according to lowest residual target function value. The RMSD of the bundle of conformers is 0.99 Angstrom, and the RMSD deviation between the mean coordinates of

the structure bundle to the reference protein is 2.23 Angstrom. The bundle of NMR conformers superimposed onto the reference structure is shown in Figure 3.10.

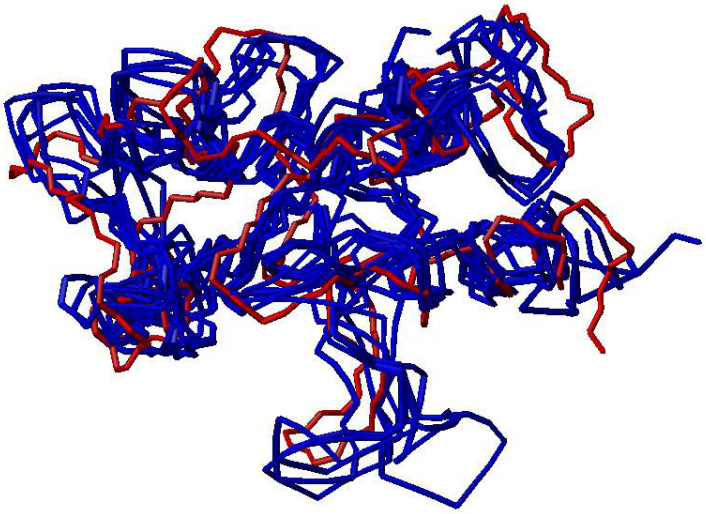


**Figure 3.10.** *Bundle of the 5 best NMR conformers is shown in blue, the mean atomic coordinates of the reference structure is presented in red.*

## 3.5 Conclusions

In this chapter, we have presented an algorithm performing assignment-free protein structure determination. The dubbed DINO algorithm takes spin-systems and standard NOE data set as input. The algorithm includes two main modules that interact iteratively with each other: (1) restraint assessment to assess interactions between spin-systems and (2) spin-system-position mapping to place spin-systems onto the sequence based on the restraints previously identified; in its turn, the mapping result provides the list of potential violated restraints that are integrated and reiterated in the restraint assessment so to obtain successively more accurate distance restraints, and also to include more accurate the available knowledge about the covalent polypeptide structure.

The performance of the DINO algorithm was so far tested only on our target small-sized protein VpR247, but these first results are promising and provided a NMR structure bundle of acceptable quality in terms of precision and accuracy of the atomic coordinates.

# CHAPTER 4

## 4. General conclusions and perspectives

**Synopsis:** After 3 and half years of my PhD work (and almost 6 years for me at the CRMN including my Master studies), we have reached the targets originally described in the PhD proposal: to create algorithms to automatically perform NMR data analysis. In this dissertation, we propose two fully automatic algorithms, one to generate automated metabolite assignment for the fields of NMR metabolomics and one to determine protein structure without performing the time-costly sequence-specific resonance assignment step.

Metabolomics is defined as the science studying the metabolic response of organisms to internal/external stimuli. NMR metabolomics has a number of advantages in the race with mass spectroscopy, due to its minimal sample preparation, non-sample destruction and highly reproducible qualities. While one-dimensional NMR is still popular to study biological mixture composition, two-dimensional NMR offers more convenience in term of metabolite assignment quality as well as of the discovery of new, unknown metabolites. The NMR signals can be assigned based on two-dimensional homo- and hetero-nuclear correlation NMR experiments and by database querying, such as HMDB or BMRB.

Our proposed algorithm ITERAMETA performs automatically metabolite profiling on two-dimensional NMR, both on TOCSY or HSQC spectra. The assignment results are robust, reliable and comparable to exhaustive manual spectral analysis. The database employed is publicly available and can be regularly updated with new metabolites. The first year of the PhD was spent to develop the core algorithm while half the third year was used to write the stand-alone version and incorporate various useful tools such as the possibility to auto-update the reference databases used.

In the second year and the forth year of the thesis, we investigated the potential of algorithms performing assignment-free NMR protein structure determination from minimal set of standard data. In NMR protein structure determination field, the idea of bypassing the tedious sequence-specific resonance assignment step is obviously very attractive. While the idea of NMR assignment-free protein structure determination is not new and a number of novel algorithms have already been developed (Grishaev and Llinas 2002; Kraulis 1994), these algorithms used selective inputs that are frequently simulated and/or high-quality *unambiguous* data. We propose an algorithm, dubbed DINO, taking spin-systems and standard NOE data as inputs, to perform NMR assignment-free structure elucidation.

The DINO algorithm includes two parts that mutually support each other. The first part called "restraint assessment" evaluates the contacts between spin-systems based on the resulted NOE signals, using a Bayesian scoring scheme. The second part called "spin-system-position mapping" puts spin-systems in the sequence, using a genetic algorithm that maximizes the contacts between consecutive residues. The outcome is used to remove violated restraints, by building a consensus of violated restraints through the population of possible solutions. The restraint set is recomputed according to the latter outcome, in order to take into account the incorrect restraints.

The iteration loop is repeated until a self-consistent restraint set and a self-consistent mapping result are found. The derived final protein structures are of 2-3 Å rmsd to the reference structure. While the structures are of "medium-accuracy" and can certainly be improved, the computing time to obtain such outcome is considerably

reduced. For our target protein, the total computation time was well below one hour on a single CPU unit.

**Perspectives:** The computational NMR group at ISA develops analytical tools to solve recurrent and manually time-consuming problems for scientists in the field of NMR metabolomics and protein structure determination. In collaboration with the NMR metabolomics group at ISA, the ITERAMETA algorithm was developed. The daily exchange with the NMR metabonomics group members provided very valuable feedback and also the possibility to steadily test and improve the underlying numerical concepts. Therefore ITERAMETA is now ready to real-world applications in NMR metabonomics and has and will continue to provide a valuable aid to metabolomics scientists in their quest in finding biomarkers.

As for the protein project, we have demonstrated the feasibility of obtaining the global fold of the target protein VpR247 using only the spin-systems and unassigned NOE peak list obtained from 3D $^{13}C$- and $^{15}N$-edited NOESY experiments as input. The generated structures are within 2-3 Å rmsd to those of the reference structure that has been determined following the conventional structure determination process. Yet, this was only the very first proof-of-principle of the DINO method proposed. Applications to other protein data sets are needed to fully assess the potential of the DINO approach, yet this first result is very promising and documents the feasibility of robust and accurate NMR assignment-free structure elucidation.

# 5. References


Atkinson, RA and V. Saudek. 2002. “The Direct Determination of Protein Structure by NMR without Assignment.” FEBS 510:1–4.

Aue, W. P., E. Bartholdi, and R. R. Ernst. 1976. “Two-Dimensional Spectroscopy. Application to Nuclear Magnetic Resonance.” Journal of Chemical Physics 64:2229–46.

Bermejo, Guíllermo A. and Míguel Llinás. 2008. “Deuterated Protein Folds Obtained Directly from Unassigned Nuclear Overhauser Effect Data.” Journal of the American Chemical Society 130(12):3797–3805.

Bertsekas, Dimitri P. and David A. Castañon. 1992. “A Forward/reverse Auction Algorithm for Asymmetric Assignment Problems.” Computational Optimization and Applications 1(3):277–97. Retrieved (http://dx.doi.org/10.1007/BF00249638).

Bertsekas, DP. 1988. “The Auction Algorithm: A Distributed Relaxation Method for the Assignment Problem.” Annals of operations research.

Billeter, Martin, Gerhard Wagner, and Kurt Wüthrich. 2008. “Solution NMR Structure Determination of Proteins Revisited.” Journal of Biomolecular NMR 42(3):155–58. Retrieved (http://dx.doi.org/10.1007/s10858-008-9277-8).

Bingol, Kerem et al. 2015. “Unified and Isomer-Specific NMR Metabolomics Database for the Accurate Analysis of 13C–1H HSQC Spectra.” ACS Chemical Biology 10(2):452–59. Retrieved (http://dx.doi.org/10.1021/cb5006382).

Bingol, Kerem et al. 2016. “Emerging New Strategies for Successful Metabolite Identification in Metabolomics.” Bioanalysis 8(6):557–73. Retrieved (http://dx.doi.org/10.4155/bio-2015-0004).

Bingol, Kerem, Lei Bruschweiler-Li, Da-Wei Li, and Rafael Brüschweiler. 2014. “Customized Metabolomics Database for the Analysis of NMR 1H–1H TOCSY and

13C–1H HSQC-TOCSY Spectra of Complex Mixtures." Analytical Chemistry 86(11):5494–5501. Retrieved (http://dx.doi.org/10.1021/ac500979g).

Bingol, Kerem and Rafael Brüschweiler. 2011. "Deconvolution of Chemical Mixtures with High Complexity by NMR Consensus Trace Clustering." Analytical Chemistry 83(19):7412–17. Retrieved (http://dx.doi.org/10.1021/ac201464y).

Bingol, Kerem and Rafael Brüschweiler. 2014. "Multidimensional Approaches to NMR-Based Metabolomics." Analytical Chemistry 86(1):47–57. Retrieved (http://dx.doi.org/10.1021/ac403520j).

Bouatra, Souhaila et al. 2013. "The Human Urine Metabolome." PLoS ONE 8(9):e73076. Retrieved (http://dx.doi.org/10.1371%2Fjournal.pone.0073076).

Briknarová, Klára et al. 1999. "The Second Type II Module from Human Matrix Metalloproteinase 2: Structure, Function and Dynamics." Structure 7(10):S1–2. Retrieved (http://dx.doi.org/10.1016/S0969-2126(00)80057-X).

Brown, Marie et al. 2005. "A Metabolome Pipeline: From Concept to Data to Knowledge." Metabolomics 1(1):39–51. Retrieved (http://dx.doi.org/10.1007/s11306-005-1106-4).

Cassioli, Andrea et al. 2015. "An Algorithm to Enumerate All Possible Protein Conformations Verifying a Set of Distance Constraints." BMC Bioinformatics 16(1):1–15. Retrieved (http://dx.doi.org/10.1186/s12859-015-0451-1).

Chignola, Francesca et al. 2011. "The CCPN Metabolomics Project: A Fast Protocol for Metabolite Identification by 2D-NMR." Bioinformatics 27(6):885–86. Retrieved (http://bioinformatics.oxfordjournals.org/content/27/6/885.abstract).

Crippen, G. and T. Havel. 1988. Distance Geometry and Molecular Conformation. New York: Wiley.

Crippen, Gordon M. 1977. "A Novel Approach to Calculation of Conformation: Distance Geometry." Journal of Computational Physics 24(1):96–107. Retrieved

(http://www.sciencedirect.com/science/article/pii/0021999177901127).

Cui, Qiu et al. 2008. "Metabolite Identification via the Madison Metabolomics Consortium Database." Nat Biotech 26(2):162–64. Retrieved (http://dx.doi.org/10.1038/nbt0208-162).

Ellinger, James J., Roger A. Chylla, Eldon L. Ulrich, and John L. Markley. 2013. "Databases and Software for NMR-Based Metabolomics." Current Metabolomics 1(1):10.2174/2213235X11301010028. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3832261/).

Fiorito, Francesco, Torsten Herrmann, Fred F. Damberger, and Kurt Wüthrich. 2008. "Automated Amino Acid Side-Chain NMR Assignment of Proteins Using 13C- and 15N-Resolved 3D [1H,1H]-NOESY." Journal of Biomolecular NMR 42(1):23–33. Retrieved (http://dx.doi.org/10.1007/s10858-008-9259-x).

Fossi, Michele et al. 2005. "Influence of Chemical Shift Tolerances on NMR Structure Calculations Using ARIA Protocols for Assigning NOE Data." Journal of Biomolecular NMR 31(1):21–34. Retrieved (http://dx.doi.org/10.1007/s10858-004-5359-4).

Goodacre, Royston, Seetharaman Vaidyanathan, Warwick B. Dunn, George G. Harrigan, and Douglas B. Kell. 2004. "Metabolomics by Numbers: Acquiring and Understanding Global Metabolite Data." Trends in Biotechnology 22(5):245–52. Retrieved (http://dx.doi.org/10.1016/j.tibtech.2004.03.007).

Gordon, Sidney L. and Kurt Wuethrich. 1978. "Transient Proton-Proton Overhauser Effects in Horse Ferrocytochrome c." Journal of the American Chemical Society 100(22):7094–96. Retrieved (http://dx.doi.org/10.1021/ja00490a068).

Grishaev, Alexander and Miguel Llinas. 2002. "CLOUDS, a Protocol for Deriving a Molecular Proton Density via NMR." Proceedings of the National Academy of Sciences of the United States of America 99(10):6707–12.

Gronwald, Wolfram and Hans Robert Kalbitzer. 2004. "Automated Structure

Determination of Proteins by NMR Spectroscopy." Progress in Nuclear Magnetic Resonance Spectroscopy 44(1-2):33–96.

Guerry, Paul, Viet Dung Duong, and Torsten Herrmann. 2015. "CASD-NMR 2: Robust and Accurate Unsupervised Analysis of Raw NOESY Spectra and Protein Structure Determination with UNIO." Journal of Biomolecular NMR 62(4):473–80. Retrieved (http://dx.doi.org/10.1007/s10858-015-9934-7).

Guerry, Paul and Torsten Herrmann. 2011. "Advances in Automated NMR Protein Structure Determination." Quarterly Reviews of Biophysics 44(03):257–309. Retrieved (http://journals.cambridge.org/article_S0033583510000326).

Guerry, Paul and Torsten Herrmann. 2012. "Comprehensive Automation for NMR Structure Determination of Proteins." Pp. 429–51 in Protein NMR Techniques, edited by A. Shekhtman and S. D. Burz. Totowa, NJ: Humana Press. Retrieved (http://dx.doi.org/10.1007/978-1-61779-480-3_22).

Guntert, Peter. 2003. "Automated NMR Protein Structure Calculation." Progress in Nuclear Magnetic Resonance Spectroscopy 43(3-4):105–25.

Havel, Timothy F. and Kurt Wüthrich. 1985. "An Evaluation of the Combined Use of Nuclear Magnetic Resonance and Distance Geometry for the Determination of Protein Conformations in Solution." Journal of Molecular Biology 182(2):281–94. Retrieved (http://www.sciencedirect.com/science/article/pii/0022283685903468).

Herrmann, Torsten, Peter Güntert, and Kurt Wüthrich. 2002a. "Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA." Journal of Molecular Biology 319(1):209–27. Retrieved (http://www.sciencedirect.com/science/article/pii/S0022283602002413).

Herrmann, Torsten, Peter Güntert, and Kurt Wüthrich. 2002b. "Protein NMR Structure Determination with Automated NOE-Identification in the NOESY Spectra Using the New Software ATNOS." Journal of Biomolecular NMR 24(3):171–89. Retrieved

(http://dx.doi.org/10.1023/A:1021614115432).

Heyer, Laurie J., Semyon Kruglyak, and Shibu Yooseph. 1999. "Exploring Expression Data: Identification and Analysis of Coexpressed Genes." Genome Research 9 (11 ):1106–15. Retrieved (http://genome.cshlp.org/content/9/11/1106.abstract).

Holland, John H. and Judith S. Reitman. 1977. "Cognitive Systems Based on Adaptive Algorithms." SIGART Bull. (63):49. Retrieved (http://doi.acm.org/10.1145/1045343.1045373).

Huang, Y. J., R. Tejero, R. Powers, and G. T. Montelione. 2006. "A Topology-Constrained Distance Network Algorithm for Protein Structure Determination from NOESY Data." Proteins 62:587–603.

Jewison, Timothy et al. 2012. "YMDB: The Yeast Metabolome Database." Nucleic Acids Research 40 (D1 ):D815–20. Retrieved (http://nar.oxfordjournals.org/content/40/D1/D815.abstract).

Jin, Xin and Jiawei Han. 2010. "Quality Threshold Clustering." P. 820 in Encyclopedia of Machine Learning, edited by C. Sammut and G. I. Webb. Boston, MA: Springer US. Retrieved (http://dx.doi.org/10.1007/978-0-387-30164-8_686).

Jung, Young-Sang and Markus Zweckstetter. 2004. "Mars - Robust Automatic Backbone Assignment of Proteins." Journal of Biomolecular NMR 30(1):11–23. Retrieved (http://dx.doi.org/10.1023/B:JNMR.0000042954.99056.ad).

Kalk, A. and H. J. C. Berendsen. 1976. "Proton Magnetic Relaxation and Spin Diffusion in Proteins." Journal of Magnetic Resonance (1969) 24(3):343–66. Retrieved (http://www.sciencedirect.com/science/article/pii/0022236476901153).

Kraulis, Per J. 1994. "Article." Journal of Molecular Biology 243(4):696–718. Retrieved (http://www.sciencedirect.com/science/article/pii/0022283694900426).

Krishnan, P., N. J. Kruger, and R. G. Ratcliffe. 2005. "Metabolite Fingerprinting and Profiling in Plants Using NMR." Journal of Experimental Botany 56 (410 ):255–65.

Retrieved (http://jxb.oxfordjournals.org/content/56/410/255.abstract).

Lin, Hsin-Nan, Kun-Pin Wu, Jia-Ming Chang, Ting-Yi Sung, and Wen-Lian Hsu. 2005. "GANA—a Genetic Algorithm for NMR Backbone Resonance Assignment." Nucleic Acids Research 33(14):4593–4601. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1184223/).

Linge, J. P., S. I. O'Donoghue, and Michael Nilges. 2001. "[5] - Automated Assignment of Ambiguous Nuclear Overhauser Effects with ARIA." Pp. 71–90 in Nuclear Magnetic Resonance of Biological Macromolecules - Part B, vol. Volume 339, edited by V. D. and U. S. B. T.-M. in E. Thomas L. James. Academic Press. Retrieved (http://www.sciencedirect.com/science/article/pii/S0076687901393102).

Linge, Jens P., Michael Habeck, Wolfgang Rieping, and Michael Nilges. 2003. "ARIA: Automated NOE Assignment and NMR Structure Calculation." Bioinformatics 19(2):315–16.

Ludwig, Christian and Ulrich L. Günther. 2011. "MetaboLab - Advanced NMR Data Processing and Analysis for Metabolomics." BMC Bioinformatics 12(1):1–6. Retrieved (http://dx.doi.org/10.1186/1471-2105-12-366).

MacRaild, Christopher A. and Raymond S. Norton. 2014. "RASP: Rapid and Robust Backbone Chemical Shift Assignments from Protein Structure." Journal of Biomolecular NMR 58(3):155–63. Retrieved (http://dx.doi.org/10.1007/s10858-014-9813-7).

Malliavin, TE, A. Routh, M. Delsuc, and JY Lallemand. 1992. "Approche Directe de La Determination de Structures Moleculaires a Partir de L'effet Overhauser Nucleaire." CR Acad Sci Paris 315(2):653–59.

Mareuil, Fabien, Th??r??se E. Malliavin, Michael Nilges, and Benjamin Bardiaux. 2015. "Improved Reliability, Accuracy and Quality in Automated NMR Structure Calculation with ARIA." Journal of Biomolecular NMR 62(4):425–38.

Markley, JL et al. 2007. "New Bioinformatics Resources For Metabolomics." Pac Symp

Biocomput. 157–68.

Markley, JL. 2012. “In Support of the BMRB.” Nat Struct Mol Biol 19(9):854–60. Retrieved (http://dx.doi.org/10.1038/nsmb.2371).

Marti, Daniel N., Johann Schaller, and Miguel Llinás. 1999. “Solution Structure and Dynamics of the Plasminogen Kringle 2−AMCHA Complex: 31-Helix in Homologous Domains,.” Biochemistry 38(48):15741–55. Retrieved (http://dx.doi.org/10.1021/bi9917378).

Morris, Richard J., Anastassis Perrakis, and Victor S. Lamzin. 2002. “ARP/wARP’s Model-Building Algorithms. I. The Main Chain.” Acta Crystallographica Section D 58(6 Part 2):968–75. Retrieved (http://dx.doi.org/10.1107/S0907444902005462).

Morris, Richard J., Anastassis Perrakis, and Victor S. Lamzin. 2003. “ARP/wARP and Automatic Interpretation of Protein Electron Density Maps.” Pp. 229–44 in Macromolecular Crystallography, Part D, vol. Volume 374, edited by B. T.-M. in Enzymology. Academic Press. Retrieved (http://www.sciencedirect.com/science/article/pii/S0076687903740117).

Moseley, Hunter N. B. and Gaetano T. Montelione. 1999. “Automated Analysis of NMR Assignments and Structures for Proteins.” Current Opinion in Structural Biology 9(5):635–42. Retrieved (http://www.sciencedirect.com/science/article/pii/S0959440X99000196).

Neal, Stephen, Alex M. Nip, Haiyan Zhang, and David S. Wishart. 2003. “Rapid and Accurate Calculation of Protein 1H, 13C and 15N Chemical Shifts.” Journal of Biomolecular NMR 26(3):215–40. Retrieved (http://dx.doi.org/10.1023/A:1023812930288).

Nicholson, Jeremy K. and John C. Lindon. 2008. “Systems Biology: Metabonomics.” Nature 455(7216):1054–56. Retrieved (http://dx.doi.org/10.1038/4551054a).

Nilges, Michael. 1995. “Calculation of Protein Structures with Ambiguous Distance

Restraints. Automated Assignment of Ambiguous NOE Crosspeaks and Disulphide Connectivities." Journal of Molecular Biology 245(5):645–60. Retrieved (http://www.sciencedirect.com/science/article/pii/S0022283684700532).

Nilges, Michael, G. Marius Clore, and Angela M. Gronenborn. 1988. "Determination of Three-Dimensional Structures of Proteins from Interproton Distance Data by Hybrid Distance Geometry-Dynamical Simulated Annealing Calculations." FEBS Letters 229(2).

Nilges, Michael, Angela M. Gronenborn, Axel T. Brünger, and G. Marius Clore. 1988. "Determination of Three-Dimensional Structures of Proteins by Simulated Annealing with Interproton Distance Restraints. Application to Crambin, Potato Carboxypeptidase Inhibitor and Barley Serine Proteinase Inhibitor 2." Protein Engineering 2 (1 ):27–38. Retrieved (http://peds.oxfordjournals.org/content/2/1/27.abstract).

Nilges, Michael, Maria J. Macias, Séan I. O'Donoghue, and Hartmut Oschkinat. 1997. "Automated NOESY Interpretation with Ambiguous Distance Restraints: The Refined NMR Solution Structure of the Pleckstrin Homology Domain from β-spectrin1." Journal of Molecular Biology 269(3):408–22. Retrieved (http://www.sciencedirect.com/science/article/pii/S0022283697910445).

Pontoizeau, Clément, Torsten Herrmann, Pierre Toulhoat, Bénédicte Elena-Herrmann, and Lyndon Emsley. 2010. "Targeted Projection NMR Spectroscopy for Unambiguous Metabolic Profiling of Complex Mixtures." Magnetic Resonance in Chemistry 48(9):727–33. Retrieved (http://dx.doi.org/10.1002/mrc.2661).

Ravanbakhsh, Siamak et al. 2015. "Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics." PLoS ONE 10(5):e0124219. Retrieved (http://dx.doi.org/10.1371%2Fjournal.pone.0124219).

Robinette, Steven L. et al. 2011. "Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology." Analytical Chemistry 83(5):1649–57. Retrieved

(http://dx.doi.org/10.1021/ac102724x).

Robinette, Steven L., Fengli Zhang, Lei Brüschweiler-Li, and Rafael Brüschweiler. 2008. "Web Server Based Complex Mixture Analysis by NMR." Analytical Chemistry 80(10):3606–11. Retrieved (http://dx.doi.org/10.1021/ac702530t).

Rosato, Antonio et al. 2009. "CASD-NMR: Critical Assessment of Automated Structure Determination by NMR." Nature methods 6(9):625–26. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841015/).

Rosato, Antonio et al. 2012. "Blind Testing of Routine, Fully Automated Determination of Protein Structures from NMR Data." Structure (London, England : 1993) 20(2):227–36. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3609704/).

Rosato, Antonio and Martin Billeter. 2015. "Automated Protein Structure Determination by NMR." Journal of Biomolecular NMR 62(4):411–12. Retrieved (http://dx.doi.org/10.1007/s10858-015-9966-z).

Saude, Erik J. and Brian D. Sykes. 2007. "Urine Stability for Metabolomic Studies: Effects of Preparation and Storage." Metabolomics 3(1):19–27. Retrieved (http://dx.doi.org/10.1007/s11306-006-0042-2).

Schmidt, Elena and Peter Güntert. 2012. "A New Algorithm for Reliable and General NMR Resonance Assignment." Journal of the American Chemical Society 134(30):12817–29. Retrieved (http://dx.doi.org/10.1021/ja305091n).

Serrano, Pedro et al. 2012. "The J-UNIO Protocol for Automated Protein Structure Determination by NMR in Solution." Journal of Biomolecular NMR 53(4):341–54. Retrieved (http://dx.doi.org/10.1007/s10858-012-9645-2).

Volk, Jochen, Torsten Herrmann, and Kurt Wüthrich. 2008. "Automated Sequence-Specific Protein NMR Assignment Using the Memetic Algorithm MATCH." Journal of Biomolecular NMR 41(3):127–38. Retrieved (http://dx.doi.org/10.1007/s10858-008-9243-5).

Wagner, G. and K. Wüthrich. 1979. “Truncated Driven Nuclear Overhauser Effect (TOE). A New Technique for Studies of Selective 1H-1H Overhauser Effects in the Presence of Spin Diffusion.” J Magn Reson 33:675–80.

Wang, Tao et al. 2009. “Automics: An Integrated Platform for NMR-Based Metabonomics Spectral Processing and Data Analysis.” BMC Bioinformatics 10(1):1–15. Retrieved (http://dx.doi.org/10.1186/1471-2105-10-83).

Williamson, Mike P. and C. Jeremy Craven. 2009. “Automated Protein Structure Calculation from NMR Data.” Journal of Biomolecular NMR 43(3):131–43. Retrieved (http://dx.doi.org/10.1007/s10858-008-9295-6).

Wishart, David S. et al. 2007. “HMDB: The Human Metabolome Database.” Nucleic Acids Research 35 (suppl 1 ):D521–26. Retrieved (http://nar.oxfordjournals.org/content/35/suppl_1/D521.abstract).

Wishart, David S. et al. 2009. “HMDB: A Knowledgebase for the Human Metabolome.” Nucleic Acids Research 37 (suppl 1 ):D603–10. Retrieved (http://nar.oxfordjournals.org/content/37/suppl_1/D603.abstract).

Wishart, David S. et al. 2013. “HMDB 3.0—The Human Metabolome Database in 2013.” Nucleic Acids Research 41 (D1 ):D801–7. Retrieved (http://nar.oxfordjournals.org/content/41/D1/D801.abstract).

Wishart, David S., Rupasri Mandal, Avalyn Stanislaus, and Miguel Ramirez-Gaona. 2016. “Cancer Metabolomics and the Human Metabolome Database.” Metabolites 6(1):10. Retrieved (http://www.mdpi.com/2218-1989/6/1/10).

Wuthrich, K. 1986. NMR of Proteins and Nucleic Acids. Wiley New York.

Xi, Yuanxin, Jeffrey S. de Ropp, Mark R. Viant, David L. Woodruff, and Ping Yu. 2006. “Automated Screening for Metabolites in Complex Mixtures Using 2D COSY NMR Spectroscopy.” Metabolomics 2(4):221–33. Retrieved (http://dx.doi.org/10.1007/s11306-006-0036-0).

Xia, Jianguo, Trent C. Bjorndahl, Peter Tang, and David S. Wishart. 2008. "MetaboMiner--Semi-Automated Identification of Metabolites from 2D NMR Spectra of Complex Biofluids." BMC bioinformatics 9:507. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2612014&tool=pmcentrez&rendertype=abstract).

Xia, Jianguo, Rupasri Mandal, Igor V Sinelnikov, David Broadhurst, and David S. Wishart. 2012. "MetaboAnalyst 2.0—a Comprehensive Server for Metabolomic Data Analysis." Nucleic Acids Research 40(Web Server issue):W127–33. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3394314/).

Xu, Yizhuang, Xiaoxia Wang, Jun Yang, Julia Vaynberg, and Jun Qin. 2006. "PASA -- A Program for Automated Protein NMR Backbone Signal Assignment by Pattern-Filtering Approach." Journal of Biomolecular NMR 34(1):41–56. Retrieved (http://dx.doi.org/10.1007/s10858-005-5358-0).

Zhang, Fengli, Lei Bruschweiler-Li, and Rafael Brüschweiler. 2012. "High-Resolution Homonuclear 2D {NMR} of Carbon-13 Enriched Metabolites and Their Mixtures." Journal of Magnetic Resonance 225:10–13. Retrieved (http://www.sciencedirect.com/science/article/pii/S1090780712003059).

Zhang, Fengli, Lei Bruschweiler-Li, Steven. L. Robinette, and Rafael Brüschweiler. 2008. "Self-Consistent Metabolic Mixture Analysis by Heteronuclear NMR. Application to a Human Cancer Cell Line." Analytical Chemistry 80(19):7549–53. Retrieved (http://dx.doi.org/10.1021/ac801116u).

Zhang, Fengli, Steven L. Robinette, Lei Bruschweiler-Li, and Rafael Brüschweiler. 2009. "Web Server Suite for Complex Mixture Analysis by Covariance NMR." Magnetic Resonance in Chemistry 47(S1):S118–22. Retrieved (http://dx.doi.org/10.1002/mrc.2486).

Zhang, Haiyan, Stephen Neal, and David S. Wishart. 2003. "RefDB: A Database of Uniformly Referenced Protein Chemical Shifts." Journal of Biomolecular NMR 25(3):173–95. Retrieved (http://dx.doi.org/10.1023/A:1022836027055).

Zimmerman, Diane E. et al. 1997. “Automated Analysis of Protein {NMR} Assignments Using Methods from Artificial intelligence1.” Journal of Molecular Biology 269(4):592–610. Retrieved (http://www.sciencedirect.com/science/article/pii/S0022283697910524).