UNIVERSIDAD NACIONAL DE COLOMBIA

Robust Automatic Assignment of Nuclear Magnetic Resonance Spectra for Small Molecules

Andrés Mauricio Castillo Robles

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2015

Robust Automatic Assignment of Nuclear Magnetic Resonance Spectra for Small Molecules

Andrés Mauricio Castillo Robles

In fulfillment of the requirements for the degree of:
Doctor en Ingeniería - Ingeniería de Sistemas y Computación

Advisor:
Julien Wist, Ph.D.
Co-advisor:
Fabio Augusto González Osorio, Ph.D.

Research Field:
Chemoinformatics-NMR
Research Group:
Desarrollo y Aplicaciones de Resonancia Magnética Nuclear(DARMN)

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2015

To my little daughters: To show you that everything is possible.

Acknowledgements

I express my thanks to Julien Wist and Luc Patiny for all their fruitful advises, discusions and beers; to the professor Fabio González for his valuable support and advises. I really enjoyed those 4 years working aside of them. I'm especially grateful to Andrés Bernal for all his contributions during this process. I also want to express my thanks to the people from MindLab research group and friends: Jhon, Jorge, Angel, Juan Carlos, Alejandro, Viviana, Oscar for the coffee, the food, the drinks and the trasendental discusions we kept.

Resumen

En este trabajo describimos un sistema completamente automático de asignación de espectros de Resonancia Magnética Nuclear(RMN) para moléculas pequeñas. Este sistema tiene la siguientes características: 1. usa como entrada datos de RMN crudos. Lo que significa que debe ser capaz de extraer de ellos, la información que es útil y dejar de lado el ruido; 2. asigna las señales a átomos en la estructura, y asocia a cada asignación un valor de confianza, que es usado para ordenar todas las posibles soluciones; 3. no depende de predicciones de desplazamientos químicos, de forma que puede usar solo la información de conectividad observada en los espectros de RMN 2D y las integrales(las constantes de acople también son una posibilidad, pero no fueron exploradas en este trabajo). Sin embargo el sistema puede usar los desplazamientos químicos si están disponibles; 4. puede aprender de forma no supervisada, la relación entre configuraciones de átomos y desplazamientos químicos mientras resuelve problemas de asignación, lo que le permite mejorar mientras trabaja, de forma análoga a como lo hace un humano.
Este sistema es completamente de código abierto, al igual que los datos que se usaron en este trabajo.

Palabras clave: RMN, asignación, predicción, similitud.

Abstract

In this document we describe a fully automatic assignment system for Nuclear Magnetic Resonance (NMR) for small molecules. This system has 3 main features: 1. it uses as input raw NMR data. Which means it should be able to extract from them the information that is useful while ignores the noise; 2. it assigns the signals to atoms in the structure, and associates to each assignment a confidence value, which is used to sort all possible solutions; 3. it does not depend on chemical shifts predictions. So it can use the connectivity information observed in 2D NMR spectra and integrals to perform an assignment(coupling constants are also a possibility, but were not explored in this work). However, the system can use chemical shifts if available.; 4. it can learn in an unsupervised fashion, the relation between configurations of atoms and chemical shifts while solving assignment problems, which allows the system to improve while working. Analogous to the way a human works.
This system is completely open source, as well as the data used in this work.

Keywords: NMR, assignment, prediction, similarity

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1. Motivation

The elucidation and validation of structures through Nuclear Magnetic Resonance (NMR) is a crucial step in the process of discovery and synthesis of new compounds. Although manual analysis by human experts is still the most widely spread and reliable method, the large and continuously growing amount of information produced nowadays demands the assistance of computational tools to reduce the time and cost of NMR analysis [**56**, **58**].
According to the fundamentals of NMR spectroscopy, nuclei in a molecule resonate during a NMR experiment, producing signals with properties characteristic of the nuclei's molecular environments. It is this correspondence between NMR signals and molecular structure that allows NMR spectroscopy to be used for the identification of compounds. Human experts resolve structures using NMR spectra by proposing a candidate structure based on previous knowledge and their interpretation of the spectral features. Then, they validate the structure by attempting to assign each signal to each observable nucleus of the proposed molecule in a way that correctly matches the signal's and nucleus' properties. This task is known as *NMR assignment* and is the main topic of this thesis.
Beyond the central place it occupies in human-based NMR analysis, assignment is key to the automatization of NMR analysis as well: most approaches reported in the field of NMR chemical shift prediction [**19**, **27**, **71**, **83**, **85**, **87**, **109**, **110**, **112**, **119**] and automatic elucidation [**33**,**48**,**63**,**84**,**86**,**88**] depend on repositories of well-assigned NMR data, since *ab initio* methods are ruled out because of their heavy computational requirements. Computer algorithms for NMR assignment follow a more or less standard approach: predicting spectra features (chemical shifts, multiplicities, scalar couplings, etc.), synthesizing a spectra from these predictions, and comparing it with its experimental counterpart; by matching the signals (experimental and computed) it is possible to assign them with their respective atoms. However, this process involves dealing with problems of NP-hard complexity [**38**]. It is then not surprising that, although automatic assignment is pivotal both to assist experts in their labor and to implement comprehensive automatic analysis tools, we only found 4 tools that automatically or "almost" automatically perform this task for small molecules: ACD auto-assignment [**72**], Mestre Nova [**103**] , CASA [**94**] and PERCH [**8**]. Furthermore, each of them suffers of significant shortcomings: ACD tool is commercial and the way it works has not yet been disclosed. CASA is a free and open application that works more like a computer assisted tool for assignment given its inability to deal with errors in the input. Mestre auto-assignment makes use of fuzzy logic to assign ^1H-NMR multiplets to predicted multiplets [**43**]. PERCH fits the predicted and observed

values of the ^{1}H chemical shifts and coupling constants by the use of quantum mechanical optimization, nevertheless, the methods has not been disclosed neither; however, there are some aspects in these approaches that call for improvements as they strongly depends on accurate chemical shift and coupling constant prediction, a branch of NMR analysis that poses a considerable challenge on its own and, as noted above, relies heavily on repositories of *already assigned spectra*. We will further discuss about these tools later in the related work section 2.

Due to the limitations of automatic assignment tools, most NMR repositories are created using manually assigned NMR spectra [**111**, **121**]. The huge cost of manually assigning large datasets explains why these repositories are often proprietary. Some efforts are being made to create public NMR repositories which aim to accept raw NMR data submitted by users [**96**, **97**, **114**]; but since the information cannot still be assigned an validated automatically, it has to either be left "as it is" (mylims [**96**] and Chemspider [**97**]) or validated manually, as in the case of BMRB [**114**]. That these repositories are required by the most accurate assignment algorithms leads to the irony that in order to assign an spectra automatically it is first necessary to assign *a lot* of spectra manually.

Another important challenge in the pursuit of fully automatic assignment of experimental NMR spectra consists in dealing with incomplete information (missing or overlapped peaks), spurious or miss-assigned peaks, etc. Most of the efforts in this direction have been focused in curing the information before the analysis [**18**, **42**, **91**, **107**], either manually or by means of robust algorithms for peak-picking, solvent detection, assignment, etc. Nevertheless accurate signal identification is still an issue and automatic NMR analysis a yet unreached ideal.

One of the reasons hindering progress in this enterprise is that it has not yet been fully recognized that the different tasks involved in NMR analysis (depuration, peak-picking, prediction, assignment, elucidation) are strongly related, so that output from one task can provide feedback to be used by another. For instance, peak-picking provides the starting point for assignment, so that the success of the assignment method relies on the quality of peak-picking. Then, in turn, a partial assignment can trigger a new peak-picking routine that takes the information learnt into account to look for expected features in the spectra that allow it to better distinguish meaningful signals from artifacts. Similarly, a partial assignment provides information that can be fed back to a NMR predictor Figure **1-1**. Following this cyclic approach, fully automatic methods based on expert approaches could ideally be build that deal with inconsistencies in the input, taking advantage of redundancy in the information provided by NMR experiments.

Creating an application which does not require human interaction, and that fully embraces the potential of the cyclic relations between the different components of NMR analysis has been the first motivation of this work.

1.2. Research problem

Fully Automatic NMR assignment presents three major challenges:

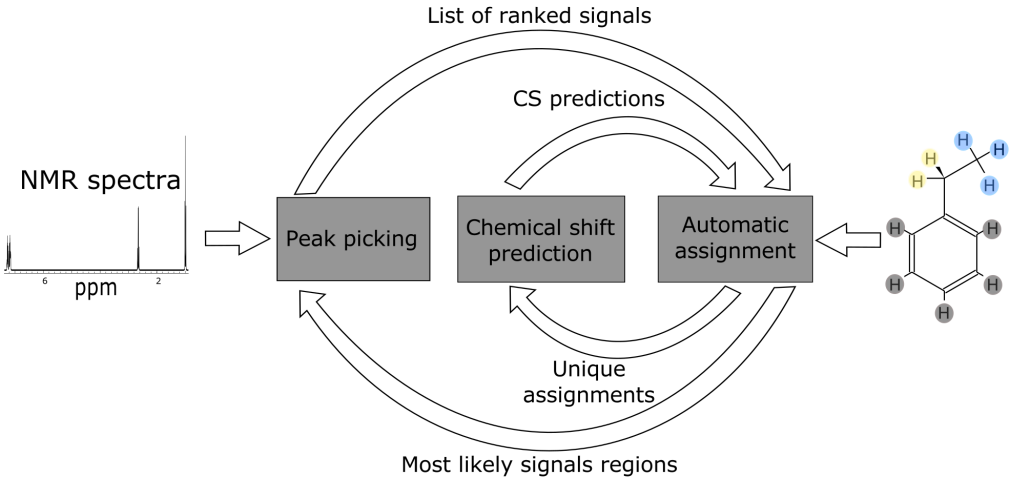


FIGURE **1-1**. Automatic NMR analysis loop. The peak picking routine provides the initial input of the automatic assignment method. The last in turn can produce feedback over the regions that the peak-picking should concentrate on, or the regions that should definitely be avoided.

(1) **The correct extraction of NMR signals from experimental spectra**, which has been approached in many other works [**47**, **66**, **70**, **95**] fails in the fact that it does not use all the available information to validate the reliability of any signal and thus simplifying the task of false peak identification. The issues with this approach arise from experimental noise, i.e. impurities, solvents and artifacts, in the case of 2D spectra, that generate false signals not easily identifiable [**57**, **82**](false positives). On the other hand, in 2D experiments it is well known that spins with very different chemical shifts and small coupling constants, or fast relaxation spins produce low intensity signals that could be easily misinterpreted as noise [**92**](false negatives). A good peak-picking algorithm has to take into account those facts and produce a full list of validated peaks that includes a probability/confidence level that each signal truly represents a structural characteristic.

(2) **The high dimensionality of the search space**. As outlined by Chen et al [**38**] the problem of finding the best match between the graph of the molecule and the partial graph of connectivity defined by NMR spectra is an NP-Hard problem. This means that the brute force approach is not an option, given the combinatorial explosion of the search space. Computational optimization approaches that have been successfully used for the simpler case of assignment of protein backbones [**20**, **21**, **24**, **26**, **32**, **60**, **61**, **62**, **76**, **80**, **89**, **117**, **118**, **125**] could be extended and adapted to the more complex case of any-molecule assignment.

(3) **The high dependency relation between NMR assignment and NMR prediction.** Another big problem with the automatic assignment is that it strongly depends on the capacity to accurately predict NMR chemical shifts. This problem can be partially avoided using 2D-NMR spectra. Nevertheless,

chemical shift information is still important to restrict the search space. It is important to point that this high dependency could be transformed in a source of new knowledge.

1.2.1. Main Objective. The aim of this work is to build a fully automatic system for the NMR analysis of small molecules, that can operate without the need of human assigned datasets.

1.2.2. Specific objectives.

(1) To create a peak picking methodology that makes uses of the redundant information contained in the NMR experiments to validate and score the peaks.
(2) To define a score function for a given assignment based on the match between observed and predicted NMR properties.
(3) To define a search strategy that allows to rank all the feasible solutions for a given assignment problem.
(4) To create a NMR predictor that uses the auto-assignment outputs as sources of knowledge.

1.3. Contributions

The following are the main contributions of the author in the field of NMR and molecular analysis:

- *Castillo, A. M., Patiny, L., and Wist, J. (2010). Fast and Accurate Algorithm for the Simulation of NMR spectra of Large Spin Systems. Journal of Magnetic Resonance, 209, 123-130.*

This paper presents a new approach to simulate ^1H NMR spectra that scales linearly with the size of the spin system by using a divide and conquer approach. This method removes weak coupling interactions in order to split the spin system efficiently and to correct a posteriori for the effect of the neglected couplings. This approach yields accurate spectra when the adequate interactions are removed, i.e., between spins only involved in weak coupling interactions, but fails otherwise. As a result, the computational time for the simulation of 1D spectra grows linearly with the size of the spin system. This paper was a contribution included in my master thesis work.

- *Castillo, A. M., Uribe, L, Patiny, L., and Wist, J. (2013). Fast and shift-insensitive similarity comparisons of NMR using a tree-representation of spectra. Chemometrics and Intelligent Laboratory Systems, 127, 1-6*

In this paper we describe a new approach to represent the NMR spectra as trees and we propose a method to compare them. This approach was shown to perform as good as the best state of the art method for the same task, but allowing a more compact representation of the information and scaling properly with the dimensionality of the spectra. This paper was a contribution included in my master thesis work.

- *Castillo, A. M., Bernal, A., Patiny, L., and Wist, J. (2014). A new method for the comparison of ^1H NMR predictors based on tree-similarity of spectra. Journal of Cheminformatics, 6,*

In this paper we describe a methodology based on spectral similarity that allows to compare NMR predictors without the recourse to assigned experimental spectra, thereby making the task of benchmarking NMR predictors less tedious, faster, and less prone to human error. This approach was used to compare four popular NMR predictors using a dataset of 1000 molecules and their corresponding experimental spectra. The results found were consistent with those obtained by directly comparing deviations between predicted and experimental shifts.

- *Bernal, A., Castillo, A. M., González, F., Patiny, L., and Wist, J. (2015) Improving the efficiency of branch and bound complete-search NMR assignment using the symmetry of molecules and spectra. Journal of Chemical Physics, 142:074103.*

In this paper we analyze 3 strategies that help to reduce the solution space of the assignment problem: 1. To represent the molecules using the condensed symmetry structure, which reduces the number of entities to assign; 2. the use of a branch and bound strategy for the partition of the search space; 3. a criterion of selection of input restrictions that leads to increased gaps between branches and thus faster pruning of non-viable solution.

- *Castillo, A. M., Bernal, A., Patiny, L., and Wist, J. (2015). Fully automatic assignment of small molecules' NMR spectra without relying on chemical shift predictions. Magnetic Resonance in Chemistry, 53:603–611.*

This paper describes a method for the automatic assignment system for small molecules' NMR spectra. The method includes an automatic and novel self-consistent peak-picking routine that validates NMR peaks in each spectrum against peaks in the same or other spectra that are due to the same resonances. The auto-assignment routine used is based on branch-and-bound optimization and relies predominantly on integration and correlation data; chemical shift information may be included when available to fasten the search and shorten the list of viable assignments.

- *Castillo, A. M., Bernal, A., Patiny, L., and Wist, J. (Accepted in April of 2016). "*Ask Ernö*": A self-learning tool for assignment and prediction of Nuclear Magnetic Resonance spectra. Journal of Cheminformatics.*

This paper presents a autonomous learning system for automatic analysis of NMR spectra consisting of coupled chemical shift assignment and prediction tools. The main idea behind this project was to exploit the strong dependency between the prediction of chemical shifts and the assignment of NMR signals to nucleus in a structure. In this system the outputs of the automatic assignment component initializes and improves a database of assigned protons that is used by the chemical shift predictor. In turn, the predictions provided by the latter enable to improve the assignment process. The iteration over

the mentioned task, improves the ability of the whole system to solve both parts of the problem.

- *nmr-auto-assignment*

A JavaScript implementation of the Automatic Assignment project described in chapter 5. Available at: *https://github.com/cheminfo-js/nmr-auto-assignment*

- *savitzky-golay-generalized*
- *global-spectral-deconvolution*
- *curve-fitting*
- *optimize-lorentzian*
- *fft*
- *tree-similarity*

A set of general purpose JavaScript libraries for signal processing used widely in the Cheminfo projects. Available at: *https://github.com/mljs*

- *autolearning*

This JavaScript project has the code and links to the datasets used for the "Ask Ernö" auto-learning project described in chapter 6. Available at: *https://github.com/cheminfo/autolearning*

- spectra-data

A JavaScript library for the manipulation of spectra data. It comprises the 1D and 2D peak-picking routines. Available at: https://github.com/cheminfo-js/spectra-data

- *http://www.nmrdb.org/*

In this website we have the NMR prediction and simulation tools.

- *http://www.cheminfo.org/*

In this website site we have examples of several chemoinformatics tools, including the examples of NMR auto-assignment, NMR peak picking and NMR chemical shifts predictions.

1.4. Document organization

This thesis is divided in 7 chapters. This chapter describes the research motivation an the problem statement. The chapter 2 describes the state of the art and the main contributions of the author in the field of NMR analysis. In order to make this document shorter, we touch in that chapter only the generalities about the main problem and more specific state of the art can be found at the beginning of each chapter. The third chapter describes a methodology to evaluate NMR chemical shift predictors by using raw and unassigned NMR spectra. The fourth chapter describes the preliminary considerations that will allow to create a fully automatic assignment algorithm. In the fifth chapter we describe the NMR automatic assignment program along with our automatic peak picking algorithm. In the sixth chapter we describe "*Ask Ernö*": a fully automatic self-learning assignment and prediction system that progressively improves its capabilities as it solves more instances of

assignment and prediction problems. The seventh chapter contains the general conclusions of the whole research work.

CHAPTER 2

Background and related work

In general, the NMR analysis can be seem as the problem of explaining the structural properties of the molecule **T**, through any kind of NMR experimental information **E,** as shown in figure **2-1.** In practice, one should consider an intermediate kind of information **P,** which is derivable from both, **T** and **E,** representations. The NMR analysis comprises the tasks that allows to move between the **T**, **P** and **E** representations, and they are defined as follows:

(1) Prediction: Is the task of determining the chemical shifts, integrals and couplings constants from the molecular structure. On chapter 6 we tackle this problem.
(2) Simulation: Is the process of applying quantum mechanics to obtain the NMR spectrum from the spin system.
(3) Parameter extraction / peak-picking: Is the task of interpreting the experimental NMR spectra in order to extract the parameters of the spin system(chemical shifts, integrals and coupling constants).
(4) Assignment: It is the problem of mapping the resonating frequencies in the spectrum to the corresponding atoms in the structure. This is is the main topic of this thesis and will be treated in chapters 4, 5 and 6.
(5) Elucidation: It is problem of finding or completing the molecular structure from the NMR parameters.

This work mainly deals with the automation of the parameter extraction, the prediction and the assignment problems through *computational methods.* In this chapter we explain the basic concepts of NMR, focused on a description of the experimental information. Also, we briefly describe the previous works on the NMR automatic assignment field and the principles of the branch and bound strategy used latter in our automatic assignment method

Let's start describing two molecular properties that allow to observe structural information experimentally: The chemical shift and the coupling constant.

2.1. Chemical shifts

Each nucleus in a molecule has a property called intrinsic angular magnetic momentum or spin. The resonance frequency of such spin is affected by other magnetic sources within its neighborhood, nuclear spins, electronic spins and electronic dipolar momentums, as a magnet that is affected by others that are moving in its nearby space. The effect of the neighborhood (electronic density) on a given spin is reflected as a change of its observed resonance frequency, with respect to its Larmor frequency (a frequency associated to each

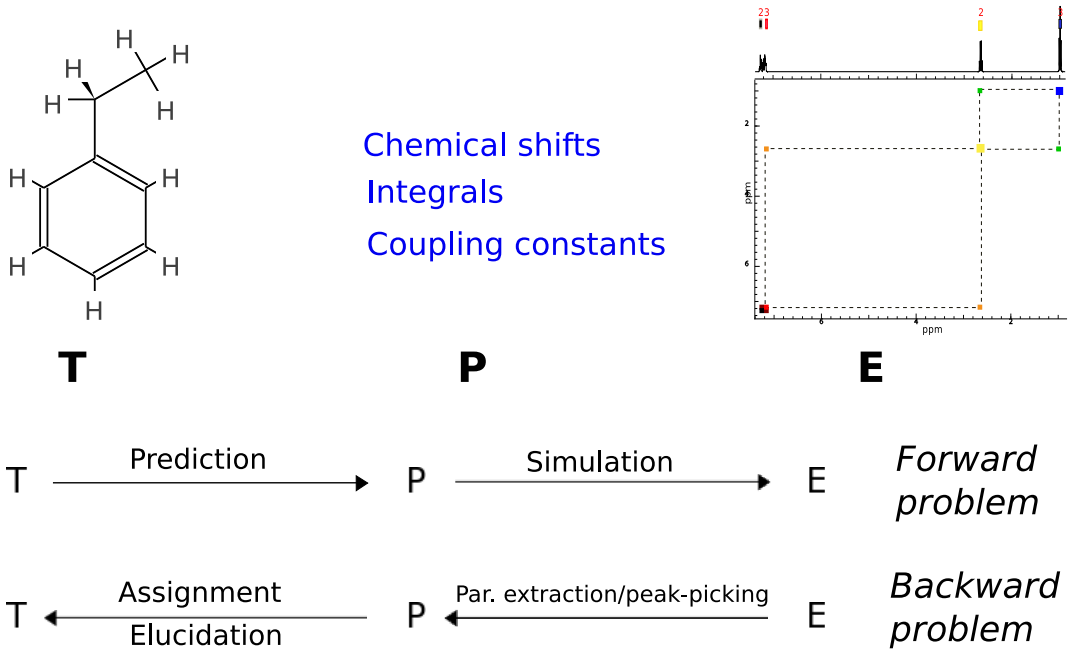


FIGURE **2-1**. The Structural analysis by NMR comprises the tasks(arrows) that allow to move between the molecular structure($\boldsymbol{T}$), the spin system parameters($\boldsymbol{P}$) and the NMR spectra($\boldsymbol{E}$). The problems of going from the molecular structure(T) to the NMR spectra(E) is knows as the forward problem, while the opposite problem of going from E to P is known as the backward problem. In this thesis we approach the prediction, assignment, and parameter extraction problems, thus covering both directions of the problem.

kind of atom). Environments close to a benzene ring or the proximity to an oxygen atom, deshield the nuclear spin and induce higher resonance frequencies observed, while the environments like a linear chain of CH (aliphatic chain), shield the nuclear spin and produce lower observed resonance frequencies [**30**, **52**, **77**]. The difference between the base frequency (Larmor frequency) and the resonance frequency of each spin is referred to as chemical shift, i.e., the shift induced by the chemical environment, but given that the absolute frequencies depend on the magnetic field of the spectrometer, the temperature, and some other experimental variables, this value is reported in parts per million (ppm) with respect to an arbitrary reference, a molecule chosen for that purpose, to make it comparable among experiments.

NMR is unable to distinguish between symmetrically equivalent pairs of nuclei, meaning that observed signals in an NMR experiment do not correspond to individual nuclei, but to subsets of equivalent nuclei. This is an important property that strongly restricts which nuclei might be assigned to each signal. For instance, it is known that all hydrogen from a methyl group have to be assigned to the same NMR signal. In ^1H-NMR spectra, the relative integration of the signals is proportional to the amount of identical spins in the structure and hence the integration of the ^1H-NMR spectrum defines another important

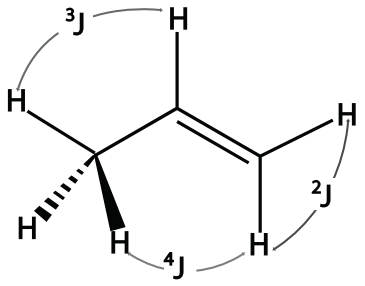


FIGURE **2-2**. Different types of constant couplings between protons. The superscript of J indicates the number of links of the coupling.

restriction over valid assignments: A group of n equivalent atoms could be assigned to an ^1H-NMR signal if and only if the integration of the signal is equal to n, or higher to cover the cases of overlapped signals. The term "peak", in this document, is used to refer the basic shape that conforms a "signal" and it is associated with a single *gaussian* or *lorentzian* function; The term "signal" refers to the set of peaks that conforms the multiplet.
The integration of the ^{1}H NMR signals are set to fit the number of protons from the molecular formula. For the sake of simplicity, the intensities of the spectrum are adjusted such that the total sum of them is exactly the number of protons in the molecule, and once the signals has been identified, the integration is set to the closest integer for each signal. This method could lead to problems in the case of unobserved labile protons or in the case of high contributions to the spectrum integral from sources of noise. An alternative way to calculate the ^{1}H integrations is described by Cobas et al. [**42**]. In this case, the signal integrations are adjusted only after the the noise contributions have been discarded, and instead of adjusting the intensities of the spectrum to fit a given number, all the signal components(peaks) are optimized to fit a set of lorentzian functions which sum of areas determine the integral of the signal.

2.2. J coupling constants

The neighborhood of each spin determines as well the number and the magnitude of the scalar couplings [**23**,**100**]. A scalar coupling, in colloquial terms, measures the interaction between two spins. The magnitude of the scalar coupling is closely linked to the proximity of the spins in terms of the minimum number of atomic bonds that separates them, as shown in Figure **2-2**. It is known that only for certain particular conformations coupling at 4 bounds distance (^4J) exists, while it always exists when the minimum distance is two links (^2J), i.e. when protons are attached to the same carbon. For the ^3J constant coupling, its magnitude depends on the dihedral angle HCCH [**23**] formed by the atoms as shown in Figure **2-3**. There are two problems associated with this: first it is not possible to detect very small coupling constants(smaller than the line-width of the spectrum) and second it is not possible to calculate the dihedral angle correctly from a two-dimensional structure. This results in that sometimes predicted couplings are not observed.

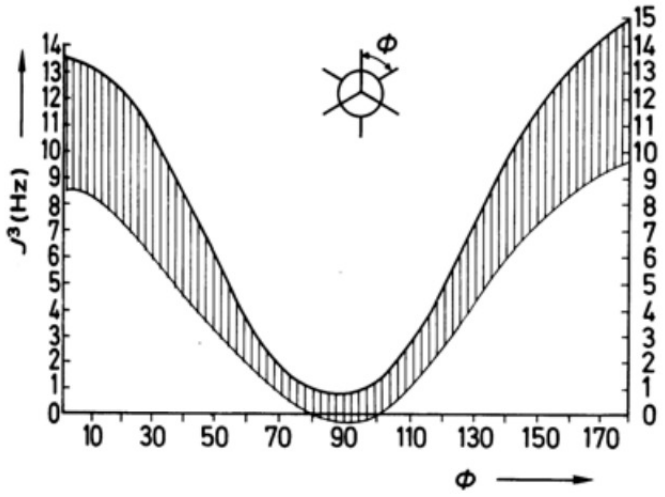


FIGURE **2-3**. Magnitude of scalar coupling ^3J, as function of the dihedral angle HCCH. [**23**]

2.3. 1D NMR experiments

1D NMR spectra allow to observe experimentally the chemical shifts associated with certain nuclei in a molecule. Nuclei observed are specific to each type of NMR experiment. In (arguably) the most widely used experiment, ^1H-NMR, it is possible to observe ^1H nuclei, that is, protons in a given molecule. As already said, each group of symmetrically equivalent protons generates a unique signal whose chemical shift coordinate will depend on its neighborhood. Furthermore, each of these signals "dissociates" in multiple peaks when J coupling is present. The amount of peaks generated in this way, known as the signal's multiplicity, will depend on the number of protons coupled to it.

There is a simple rule that expresses the multiplicities as a function of the number of coupled protons: Each J coupling constant splits the signals in 2 around the original frequency. If all the coupling constants are of the same magnitude, the number of signals and intensities(multiplicity) associated to a signal can be deduced from the Pascal triangle as shown in Fig **2-4**. In Fig **2-5** we show the ^1H-NMR spectrum of ethylbenzene. In this example we can easily identify 3 different signals. The one at the higher resonating frequencies corresponds to the aromatic protons **2-5a**. Although there are three different groups of protons in that region, their chemical shifts are overlapped and thus a massive multiplet that integrates to 5 is observed. Aliphatic protons appear at lower frequencies: the protons of the methyl group generate the signal that integrates to 3 observed at the lowest chemical shift **2-5 c**. 3 peaks with relative intensities 1,2,1 compose its multiplicity pattern, due to coupling with the protons from the methylene group observed in the region around 2.6 ppm (2 identical protons, 2 identical coupling constants) Fig **2-5 b.** This tool http://www.cheminfo.org/Spectra/NMR/Tools/Multiplet_simulator/index.html was designed to better understand the weak coupling problem.

Number of coupled protons	Peak intensities	Multiplicity
0	1	Singulet
1	1 1	Doublet
2	1 2 1	Triplet
3	1 3 3 1	Quartet
4	1 4 6 4 1	Quintet
⋮		

FIGURE **2-4**. Multiplicities explained as function of the number of equally coupled protons throw the Pascal triangle.

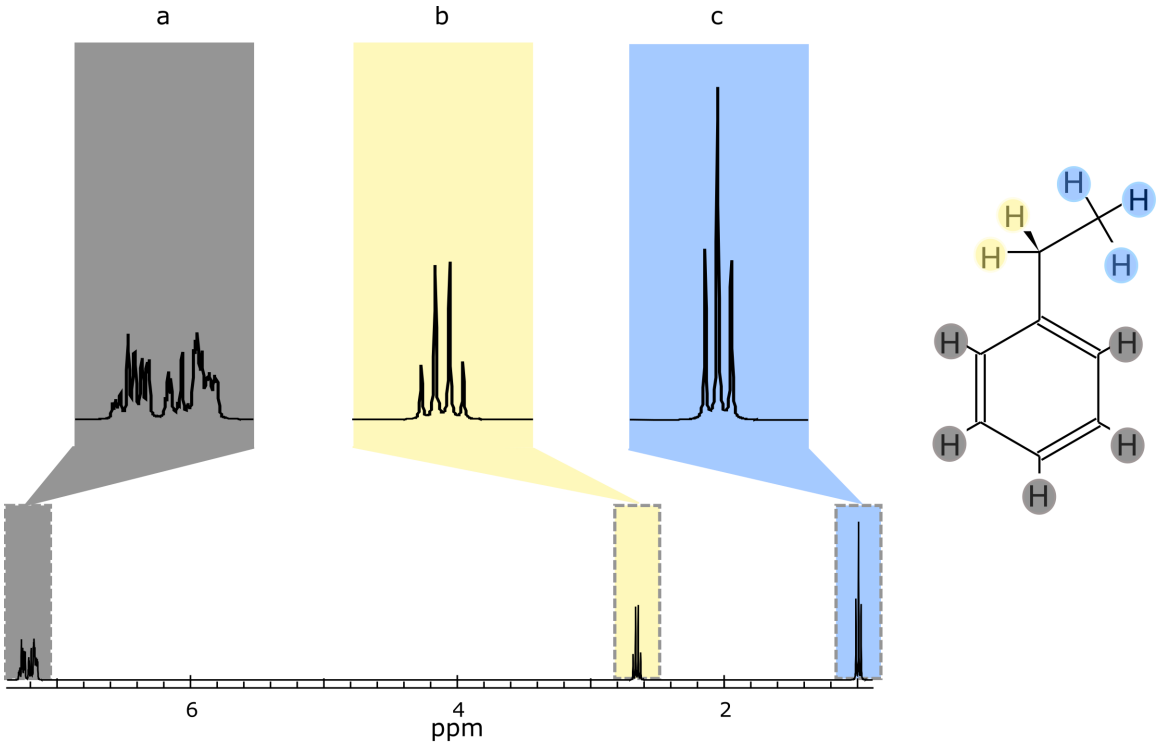


FIGURE **2-5**. ^1H-NMR spectrum at 400 MHz. In this case we can identify 3 different signals: a) A massive multiplet integrating to 5, being the overlap of the aromatic protons; b) A quadruplet integrating to 2, which arises from the 2 yellow protons. The multiplicity indicates that those protons are coupled with the the 3 blue protons; c) A triplet integrating to 3, arising from the blue protons. In this case, the multiplicity indicates the coupling with the 2 yellow protons.

2.4. 2D NMR experiments

A scalar coupling allows the transfer of magnetization from one spin to another, and it has been used by certain pulse sequences (NMR experiments) to detect which atoms are bound together. Therefore, it is possible to establish, at least partially, the connectivity graph of the original molecule.

As an example, consider the 2D-NMR spectra of ethylbenzene shown in Figure **2-6** and Figure **2-7**. In the Figure **2-6**, we observe the COSY (correlation spectroscopy) experiment in which the scalar coupling is seen as a signal at the intersection of the chemical shifts of the coupled protons (^1H). As it can be observed, this spectrum is symmetric with respect to the diagonal. In some cases, it is possible to observe long-range couplings that are not observed in the ^1H-NMR spectrum (orange cross-peak). The spectrum in Figure **2-7** is known as HMBC (Heteronuclear Multiple Bond Correlation spectroscopy). This is a two-dimensional spectrum that shows correlation between carbons and protons separated by 2, 3 and 4 bonds. In red we have marked the 1 bond ^1H-^{13}C correlations. In yellow we have marked the 2 and 3 bonds ^1H-^{13}C correlations. In blue we marked the 4 bound ^1H-^{13}C correlations. These latter are weaker, long-range correlations that are not always observed in the spectrum.

From this it can be noted that two interactions are mainly responsible for the characteristics of NMR spectra, chemical shifts and scalar couplings, and both are affected by the structure of the molecule under study.

Unfortunately, it is only possible to detect isotopes with spin greater than 0, such as: ^1H, ^{19}F, ^{31}P, ^{13}C and ^{15}N (S=1/2) among others. Except for ^1H, ^{19}F, ^{31}P, the other observable isotopes exist in low natural abundance and their detection is less sensitive. However, organic molecules contain hydrogen (^1H isotopes) in most of their structures and therefore, by analyzing their chemical shifts and their interactions with their neighbors it has been possible to elucidate the structure of the majority of molecules that participate in biological and natural processes.

Under ideal conditions, the correct NMR assignment is the one that perfectly fits experimental observations in the NMR spectra with the corresponding expected properties in the proposed molecule. It should be noted that even in ideal conditions, there may be several acceptable solutions for a given assignment problem, given that features of the spectra and molecules can be permuted. For example two singulets of intensity 3 can be attributed to two methyl groups and without additional data both solutions are equally acceptable. Generally, lifting such ambiguities requires additional experiments to differentiate between these solutions, e.g. 2D experiments. Moreover, in real cases, the assignment model should be flexible enough to deal with experimental noise, which makes the number of feasible assignments grow further.

An automatic assignment model should therefore be able to ensure that the proposed solutions are the ones that best satisfy the observations and the model constraints above all other feasible solutions. Due to the complexity of the NMR assignment problem, the only way to ensure optimal solutions, is by giving a score or probability to every single

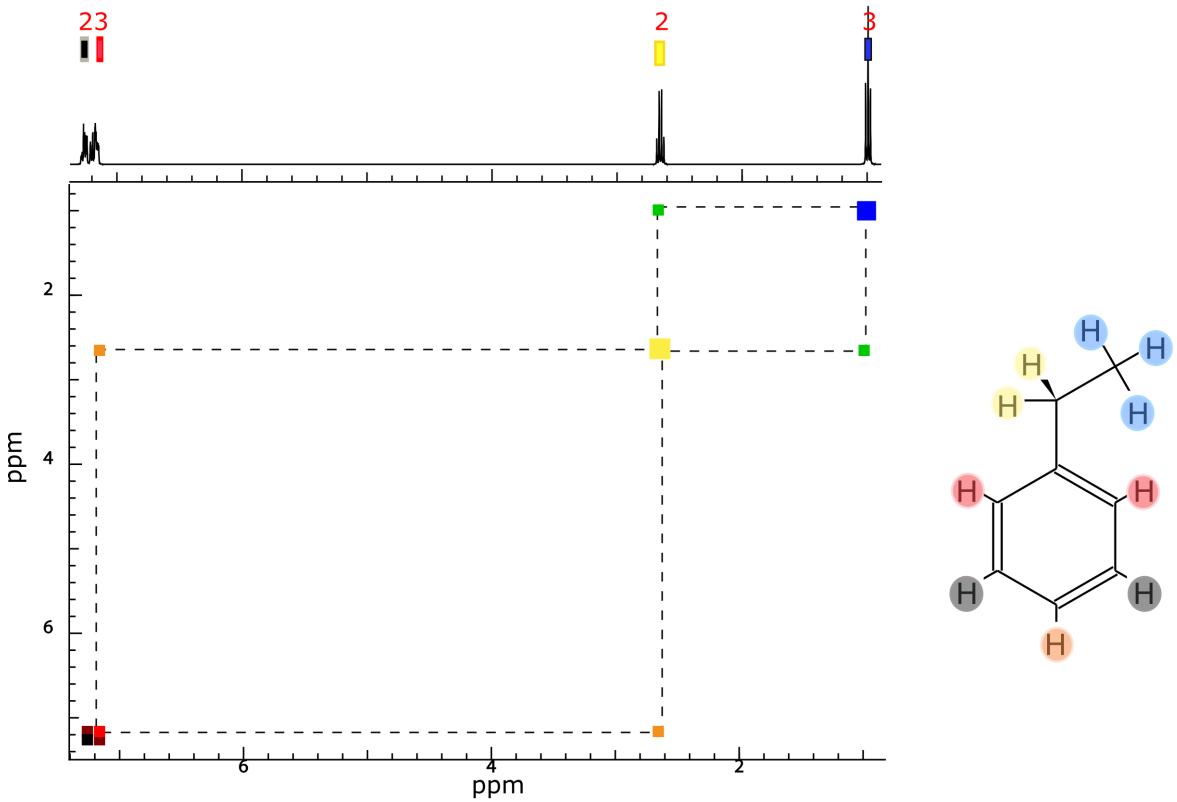


FIGURE **2-6**. COSY spectrum. In the diagonal appears the peaks corresponding to each signal in the 1H-NMR. The off-diagonal peaks indicates the coupling between the given protons, i.e. the green cross-peak reflects the ^{3}J coupling between yellow and blue protons.

possible solution in the search space. This is known as an exhaustive search algorithm for automatic assignment. However, given the combinatorial explosion of the search space in terms of size of the molecule space, it is necessary to develop smart exploration strategies to discard by quick and simple verifications infeasible solutions, and focus instead on a reduced search space.

2.5. Automatic assignment approaches

Many algorithms for automatic assignment have been developed for proteins. These algorithms are generally based on the adjustment of predicted and observed chemical shift information and spatial distances observed in the NOESY (Nuclear Overhauser effect spectroscopy) and ROESY (Rotating frame nuclear Overhauser effect spectroscopy) experiments and assignment is performed over their secondary structures, i.e. the amino acids sequence that conforms the structure. This allows to translate the assignment problem into an optimization problem as mentioned before and several works propose different optimization strategies for that purpose [**24**, **26**, **32**, **61**, **62**, **76**, **80**, **89**, **118**]. A key issue of these optimization methods is the possibility to generate an assignment that is generally

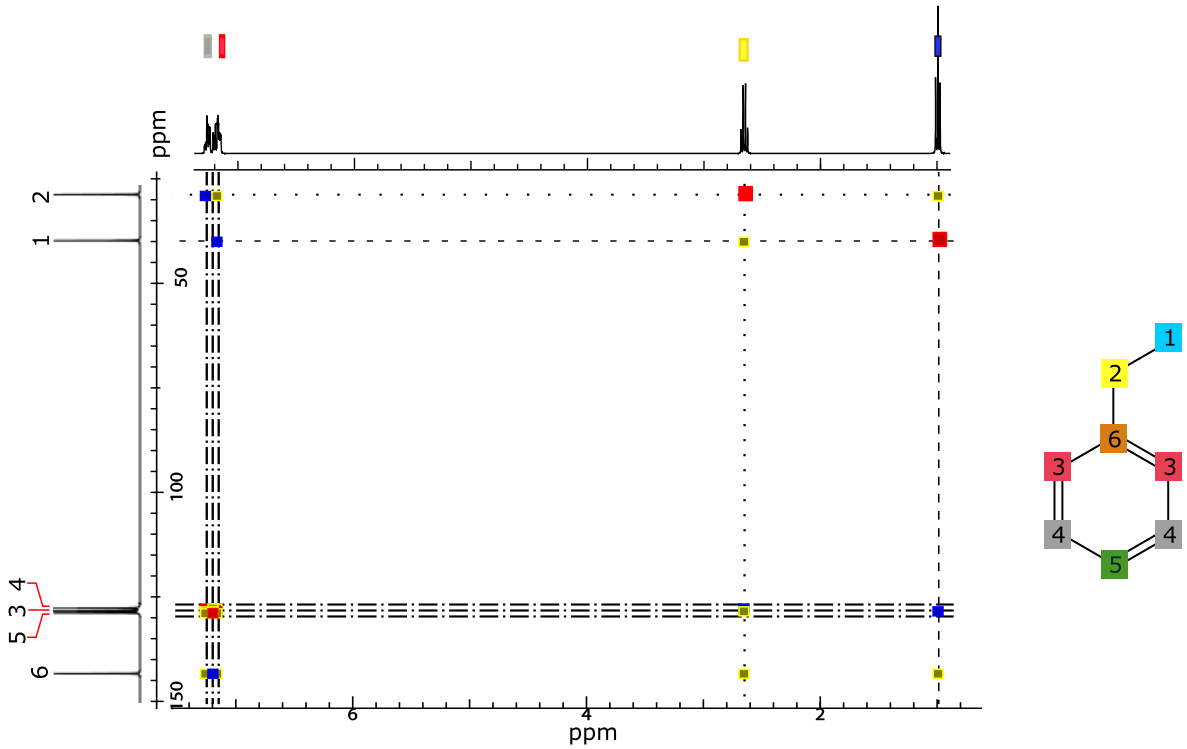


FIGURE **2-7**. HMBC spectrum. ^{13}C spectrum(vertical) is included as reference. In this spectra we can observe the couplings between ^{1}H and ^{13}C. ^{1}J couplings are marked in red, ^{2-3}J couplings are marked in yellow, and ^{4}J couplings are marked in blue.

correct [**24**, **32**, **61**, **118**], but that may be locally ambiguous, which differs from the methods, which give incomplete assignments, but with local high accuracy [**26**, **76**, **80**, **89**]. As well, exhaustive and heuristic search algorithms have been proposed [**20**, **21**, **44**, **60**, **67**]. Other approaches use best-first algorithms, that use ordered trees for the assignment of the signals, such as Auto-Assign by Zimmerman et al. [**125**] and CASA by Wang et al [**117**].

In the field of automatic assignment for small molecules, the best known applications have been found in commercial softwares: The first one is ACD-Lab tools [**72**]. This suite has a complete workshop for chemical data analysis. There is an automatic verifier of structures through NMR experiments and an automatic assignment tool for 1D and 2D spectra. However these algorithms have not been published to our knowledge.

CASA (Computer-Aided Spectral Assignment) [**94**], is the auto assignment tool of LSD. This tool is the only free implementation capable of providing a list of all possible assignments between experimental NMR signals and some proposed structure. However, it requires, as input, a peak-picking list done manually or with the help of another tool. In addition, specific information, such as hybridization of the carbons and their expected chemical shifts should be provided. Therefore, it works more like a computer assisted tool for assignment given that, as described in http://eos.univ-reims.fr/LSD//JmnSoft/CASA/,

the system fails when experimental information is not correctly identified. On the other hand, it is freely available, its algorithm has been described and is inspiring for the community.
The last option in the list has been implemented in the commercial software Mestre as an additional suite for NMR analysis and prediction [**103**]. The auto-assignment procedure is tackled by a fuzzy logic approach [**43**] that matches the observed ^1H-NMR multiplets to the predicted multiplets calculated from the putative structure. This methods requires a refined set of signals as input of the system, which are obtained by a novel peak picking strategy based on the global spectra deconvolution of spectra [**42**] and further impurities remotion by another fuzzy logic function.
The fuzzy logic strategy for the auto-assignment is based on the construction of a matrix containing the scoring between all the observed and predicted patterns, applying a series of tests that match the chemical shifts, signals integration and multiplet shapes. From this matrix, the fuzzy logic system ranks all the possible solutions and finally offers to the user the best ranked assignment.

2.6. Branch and Bound strategy

Branch and Bound is a strategy that has been proposed back in the 60s to solve integer optimization problems. Lets consider the discrete programing problem of maximizing:

(1) $\alpha x_1 + \mu x_2 = \gamma$

subject to the constrains:

(1) $a_1 x_1 + b_1 x_2 \leq c_1$
(2) $a_2 x_1 + b_2 x_2 \leq c_2$
(3) x_i is a nonnegative integer

If the condition of integer values on x does not exist, the solution of the problem is given by the simplex solution showed in the Figure **2-8a** with the dashed line. If x is constrained to take discrete values, then, the feasible solutions are marked by the gray dots within the simplex plane, and it could be observed that the optimal solution to this problem differs from the simplex solution, and additionally, it can be verified that the solution of the integer valued problem is not the pair of integers closest to simplex solution.
The branch and bound strategy starts from the simplex solution of the problem, in which the target function takes its maximum v_0 at $(x1_0, x2_0)$, and pushes down the target function as showed in figure **2-8a** until it reaches an optimal value v_k, that satisfies the integer values criteria for variables $x1$ and $x2$. This value has been tagged with a green dot in all the figures. In the 2 dimensional space it is easy to verify the pair of optimal values that satisfies the criterion. Nevertheless, it is not easy to verify this criterion for dimension higher than 3 because we can't visualize such space, and the numerical methods suited for such purpose work only for punctual values and there is not warranty of obtaining a global optima.
In order to make use of the strategy of pushing down the target function, it is necessary to define a general way to decide until where, the target function has to be pushed further

down. In figure **2-8a**, we showed the 2 new upper-bound candidates values v1 and v2 that should be evaluated. Those 2 candidates are the nearest values from v0, such that the restriction boundaries has integers components in the $x2$ variable around the $x2_0$ initial value. We highlight the intersection between those values and the boundaries with red squares(Figure **2-8a**). Once we determine which of both values produce the higher upper-bound, we redefine the optimization problem, including the new $v1$ upper bound to the set of restrictions and redefining the $x1$ boundaries as showed in Figure **2-8b** . As it can be noted, this new restrictions produces a new convex set of constrains. Once we reached the new simplex problem showed in Figure**2-8b,** we proceed to find the new $v2$ upper-bound value for the maximand function, pushing down the function, but in this case, we use $x1$ as target variable instead of $x2$. In Figure**2-8c**, we show the final simplex problem after adding the new $v2$, and $x1$ restrictions to the problem. Finally, we have to solve this reduced problem to find the optimal solution of the problem.
It is important to point out, that at each step, we solve a convex optimization problem, ensuring that the method will find the global problem optima. It is also important to note that we did not evaluate all the set of feasible solutions but a small subset of it.
In practice, this method is implemented as a search tree, where each node represents a state of the optimization problem, and each path represents one of the feasible values that the active xk variable can take at that stage. The branch and bound method can be used to solve assignment problems. In such case, each discrete variable will represent one of the elements to be assigned, and the values that each of such variables can take, correspond to the elements of the assignment codomain.

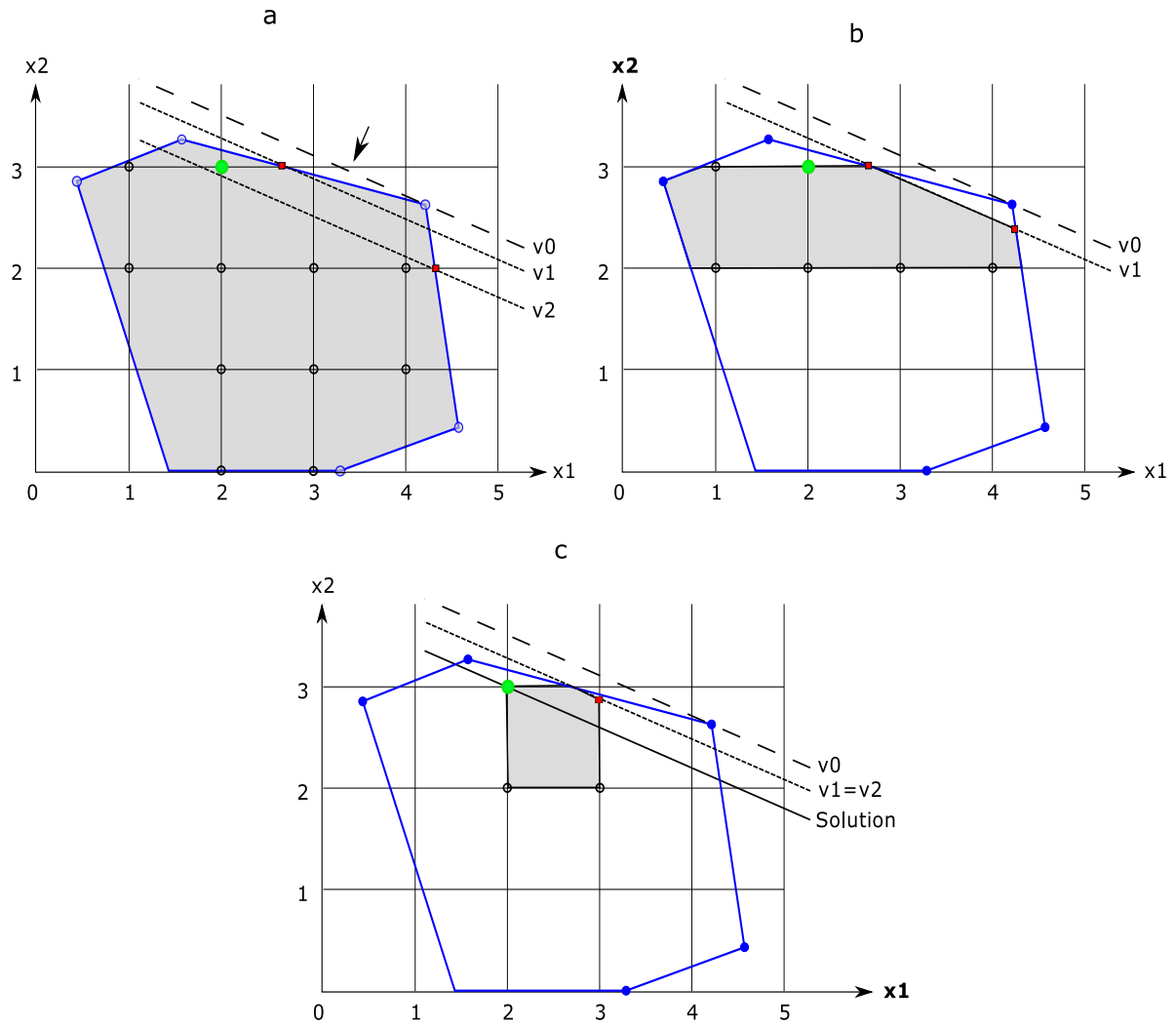


FIGURE **2-8**. The 2 variables discrete programing problem. The boundaries of the simplex problem are shown in blue, and correspond to constrains of the type: $a_k x1 + b_k x2 < c_k$

CHAPTER 3

A new method for the comparison of ^1H NMR predictors based on tree-similarity of spectra

Chemical shift predictions are of great importance to all NMR analysis. As mentioned in the section 1 most automatic methods for assignment and structure elucidation heavily rely on the accuracy of chemical shifts predictions.There are several choices in the market for predicting NMR chemical shifts, all of which are potential targets for the chemical shift prediction component of the automatic NMR analyzer that is the main topic of this work. We thus started by developing a methodology to evaluate and compare the state of the art predictors, in a way that does not require previous assignments, but that works with raw pairs of ^{1}H-NMR spectra and molecules, following our policy of free-of-human-intervention tools. This is important because, otherwise, assigned data that are not included in the predictors' knowledge databases would be required for a fair comparison. Building such test set is time-consuming at best, as data would have to be manually assigned. Another issue with this methodology is that the learning sets of commercial software are undisclosed, hence, a fair comparison is impossible without trusting the commercial software producers and without their collaboration to exclude the molecules from their learning sets. This chapter presents an alternative approach that solves the first issue and partially solves the second, by allowing for the use of unassigned data that are far easier to obtain, for example from open repositories. This method relies on a novel approach for the comparison of NMR spectra (See Section 1.3).
Follows a reproduction of the article published in the Journal of Chemoinformatics [**17**]. The sections' order differs from the published version because we find the current order easier to understand.

3.1. Background

Chemoinformatics plays an increasingly important role in structure validation by NMR spectroscopy, providing methods and algorithms for computer-assisted NMR spectra assignment and structure elucidation [**40**, **50**, **68**, **78**, **81**, **93**, **94**, **115**], as well as prediction and simulation [**11**,**12**,**13**,**27**,**28**,**35**,**55**,**56**,**72**,**79**,**106**] of spectra. Those methods heavily rely on the accuracy of predicted NMR parameters and thus, along with the introduction of such novel methods, comes the need to compare and evaluate available NMR predictors. The established approach for this task consists in comparing the predicted NMR parameters, i.e., chemical shifts and coupling constants, with experimentally determined ones. Such approach comports the need to manually assign experimental data, which demands a huge investment in manpower and introduces the possibility of human error.

To avoid these issues, benchmarking of NMR predictors could be performed using the techniques of chemoinformatics itself. We propose to evaluate the success of a prediction algorithm by measuring the similarity between an observed spectrum and a simulation built upon the predictor's output. As similarity can be computed directly on raw spectra using existing algorithms [**29**, **36**], this novel strategy avoids the requirement of peak-picking and assignment of signals, thus allowing for a simple, cheap and fully automatic approach. This chapter presents a methodology for the evaluation of NMR predictors built around this concept, and validates it against the established approach for four common predictors. In our implementation of the approach we use a tree-based method for measuring similarity between NMR spectra developed previously by our group [**36**]. This method has been shown to produce results comparable to those of the binning method [**29**], with significant improvement in efficiency, by focusing on the regions of the spectrum containing most of the information.

3.2. Methods

Consider the experimental and simulated spectra for each element of a collection of molecules and build the matrix of similarities between each experimental and each simulated spectrum (see Figure **3-1**). An accurate prediction algorithm would ensure that the highest similarity values lay on the diagonal of such matrix, i.e. the experimental spectrum of any given molecule would be more similar to its simulation than to simulated spectra of other molecules.

Now, consider a query of the experimental spectrum of some given molecule to a database of simulated spectrum. The result of the query is a list of database entries sorted in decreasing order of similarity to the experimental spectrum. For each query, the rank of a match is defined as the position of the matched simulated spectrum in this list. The more accurate the prediction, the better the rank of the simulated spectrum corresponding to the target molecule. The average performance of a predictor over a large set of queries can thus be measured by its Mean Reciprocal Rank (MRR) [**45**],

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{rank_i}$$

where n is the number of queries and $rank_i$ is the rank of the correct match in the i-th query.

Note that the MRR ignores the actual similarity values computed. This is intentional, as we are not interested in how exact are similarities between correct matches, but on whether the prediction algorithm is able to generate a spectrum that can be unequivocally distinguished as that of the input molecule. However, a low-ranking correct match may be due not to poor prediction but to poor resolution of the similarity measure, which would lead to large sets of alternatives equally similar to the query spectrum. In such case, it would be the similarity measure, rather than the prediction, that fails at discriminating the correct match. To ensure that results are not biased by issues of the similarity measure, we propose a complementary approach that associates each query with a point in the correct-match-similarity vs. best-match- similarity plane (see Figure **3-2**). In this plane:

- All queries are located on the upper triangle (gray area), as the similarity measure ranges from 0 to 1.
- Points located on the diagonal (dotted line) correspond to those cases where the best match is the correct match.
- The accuracy of the prediction in absolute terms (i.e. in terms of the similarity between the correct match and the experimental spectrum) increases as we move up to the extreme at (1,1), where the correct match and the experimental spectrum are identical. We then refer to the magnitude of the component of the query on this direction as the *absolute prediction accuracy* (see Figure **3-2**).
- The accuracy of the prediction relative to the data set (i.e. the ratio between the experimental spectrum's similarity to the correct match and its similarity to the best match in the data set) decreases as we move away from the identity line. We then refer to the magnitude of the component of the query vector orthogonal to the absolute prediction accuracy as the *relative prediction accuracy* (see Figure **3-2**).

Low relative prediction accuracy means that the correct match is as similar or almost as similar to the queried spectrum as the best match in the whole database. Good predictions can then be associated with low values of relative accuracy. Note that this approach looks into the actual similarity values between the experimental and simulated spectra regardless of the rank of the correct match, which is exactly the opposite of what we achieved with the MRR. Combining the two approaches we can distinguish between low-ranking queries due to poor prediction and low-ranking queries due to an inadequate similarity measure: as long as the same trends result from evaluating performance in terms of the relative accuracy index or in terms of the MRR, we can be certain that the evaluation is not biased by a poorly discriminating similarity measure.
It follows from the previous discussion that the choice of an appropriate similarity measure is key to the success of the methodology proposed. Here we used the tree- based methodology that has been described in detail elsewhere **[36]**. In brief, it consists in building a tree representation of each spectrum that summarizes key information on its signal-rich regions, followed by the computation of a similarity measure between these trees. This similarity measure is defined recursively, so that the similarity between two trees at depth k depends on the similarity between nodes located on that level, and on the similarity between the trees at depth k + 1. This technique is similar to the traditional binning technique **[29]**, but presents the advantage of focusing on regions with high signal intensity, using fewer data points by avoid- ing large blank or merely noisy zones.

3.3. Experimental

A set of 1000 molecules of up to 33 heavy atoms was randomly selected from the Maybridge catalogue **[7]** and the corresponding ^1H NMR spectra kindly provided by Maybridge . The spectra were acquired with a 250 MHz Bruker spectrometer using a standard Bruker pulse sequence (zg30), a relaxation delay of 1 s, a 30° excitation pulse at 27

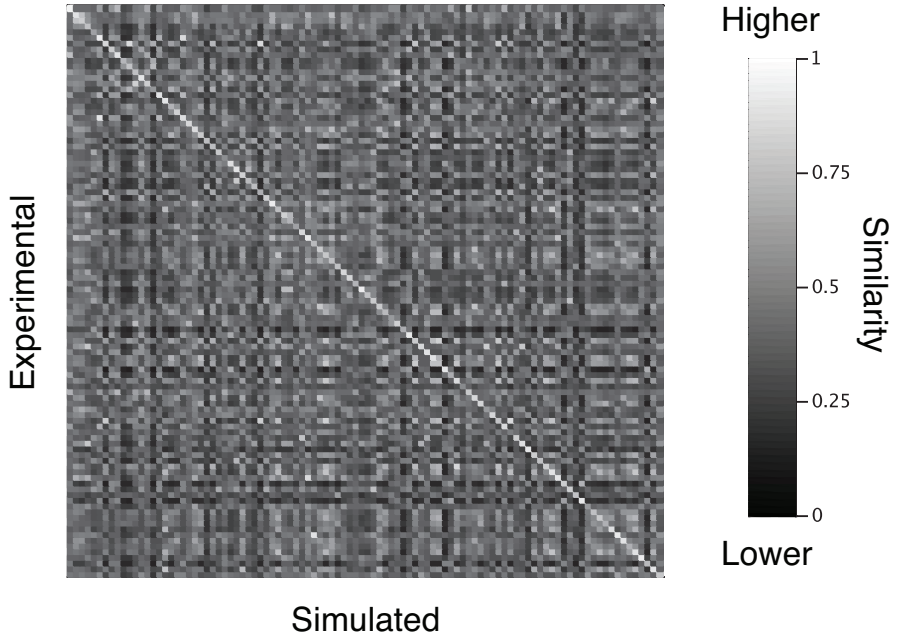


FIGURE **3-1**. Example of a spectra similarity matrix. Rows correspond to experimental spectra and columns to simulated spectra of a 100 molecules data set, matrix elements give the similarity between the corresponding experimental and simulated spectrum. The thin light gray line on the diagonal shows a trend towards higher similarity between the matching spectra, as expected for an accurate NMR predictor. [**17**]

kHz and an observation window of 20.693 ppm centered at 6.175 ppm. Each spectrum was binned and stored as a 1024 real points vector, by averaging the intensities within each of the 1024 bins, between 0 and 11 ppm. For each molecule, the proton chemical shifts were predicted using four different prediction tools, referred to as A, B, C and D. The original spectra (1024 point, JCAMP-DX format), the raw predictions and a matrix of simulated spectra of 1024 points are provided in Additional files 3 and 4. We decided to keep predictors anonymous to maintain the focus of this work on the method to rank predictors, rather than the ranking itself. Each prediction was used to simulate a 1024 point spectrum at a frequency of 250 MHz with an algorithm that we described elsewhere [**35**]. Similarity matrices between simulated and experimental spectra, MRR, and average absolute and relative prediction accuracy were computed for each data set using the methodology described in the previous Section. A subset of 298 randomly chosen molecules were manually assigned in order to perform the evaluation by direct comparison of predicted and observed chemical shifts and compare the results obtained with those produced by our method. A subset was used for this part due to time constraints.

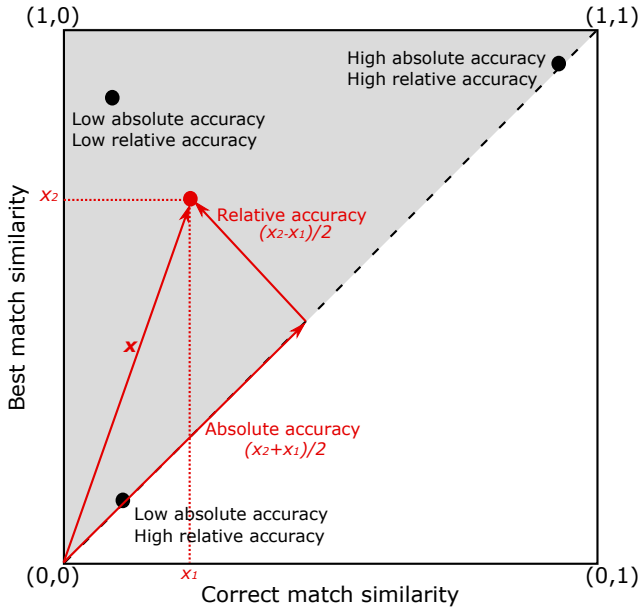


FIGURE **3-2**. Correct-match-similarity vs. best-match-similarity plane. Black dots represent queries of simulated spectra to a database of experimental ones. The position of each query can be described in terms of two orthogonal vectors; one related to the absolute accuracy of the prediction, other to the inaccuracy relative to the dataset. [**17**]

3.4. Results and discussion

Figure **3-3**a shows the distributions of correct matches within the n highest-ranked hits for each prediction algorithm. It can be observed that predictor A performed significantly better than all other algorithms. This result is confirmed by the Mean Reciprocal Rank (MRR) values. To validate our approach, we repeated the rank- ing of the four prediction tools but using a traditional approach: experimental signals were assigned to their corresponding nuclei and the differences between experimental and predicted shifts were computed. These chemical shift errors were partitioned on 0.1 ppm intervals up to 0.35 ppm, a value that already comprises over 90% of the predictions for the best performing method and over 80% for all predictors evaluated.

The resulting histograms are also shown in Figure **3-3**b. Again, the performances of predictor A were found superior, producing around 10% more predictions on the two lower error intervals, while the other systems performed similarly, in agreement with the results obtained using our method.

Figure **3-4** displays the queries associated with each of the predictors on the correct match similarity vs. best match similarity plane. Clearly, queries that used predictor A are more closely packed along the identity line, which is associated with better relative accuracy as discussed in the section Methods. This is confirmed by computation of the mean relative prediction accuracy (see Figure **3-4**). The remaining three predictors were found to perform similarly, thus reproducing the ranking given by the MRRs and corroborating

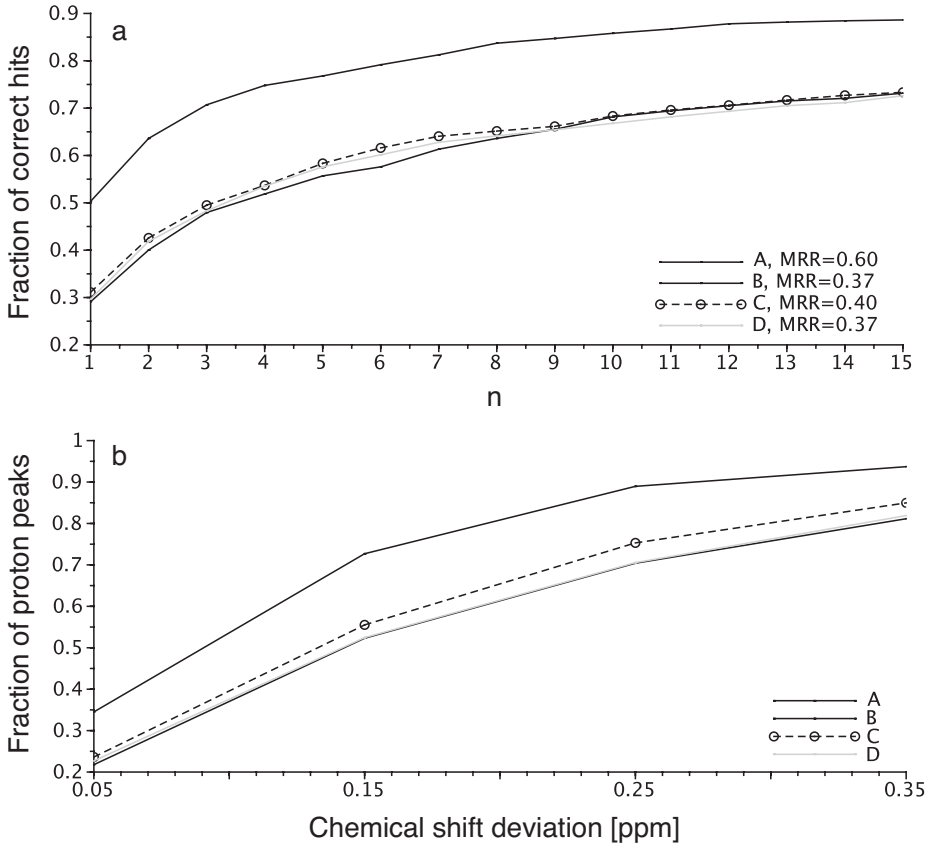


FIGURE **3-3**. Comparison of four commercially available predictors. a) Results of the evaluation of 4 ^1H NMR predictors using the new methodology. Each point in the plot corresponds to the fraction of correct matches within the n highest-ranking hits of a query of 1000 simulated ^1H spectra to the database of the corresponding experimental spectra. For example, using predictor A, around 75% of the correct matches are found within the 4 highest ranked hits. Higher curves then represent better performance. Overall MRRs obtained for each predictor are specified in the legend. b) Results of the evaluation using direct comparison of predicted and observed chemical shifts. Each point in the plot corresponds to the fraction of predicted chemical shifts that fall within the specified deviation from the observed shift. For example, using predictor A, around 75% of the predicted peaks fall within 0.15 ppm of the observed peaks. Higher curves then represent better performance. [**17**]

that these results are not biased by the similarity measure.

3.5. Conclusions

The direct comparison of simulated and experimental spectra using an adequate similarity measure allows for an efficient and fully automatic methodology to evaluate NMR prediction algorithms. Results obtained using this new method are consistent with those

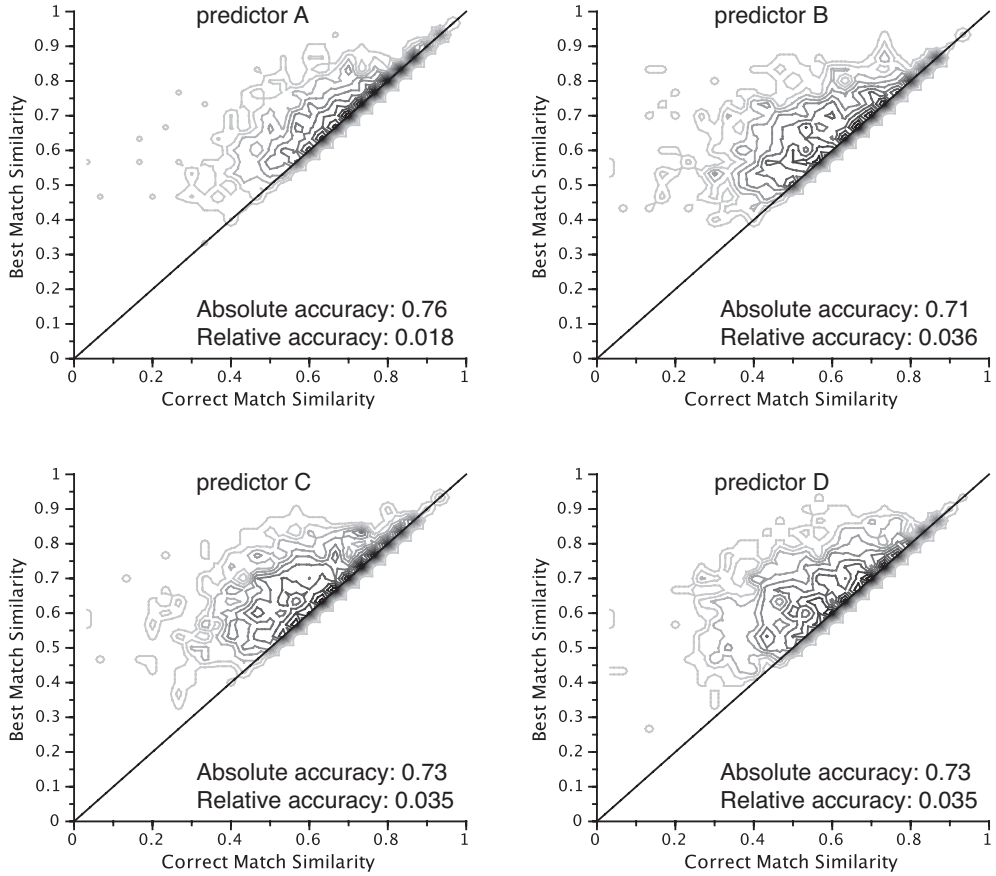


FIGURE **3-4**. Contour plots on the best match vs. correct match similarity plane of the query distributions. An ideal prediction tool would have all the density packed along the diagonal. [**17**]

obtained by the traditional chemical shift comparison method, but without the need for peak-picking and assignment. We therefore provide a method that can help improving NMR predictors in the future by allowing the comparison of predictors using datasets that are too large to be assigned manually. It is important to point out that the spectral resolution after the binning process, gives a minor effect to the coupling constants in the similarity calculation.

CHAPTER 4

Improving the efficiency of Branch-and-Bound complete-search NMR assignment using the symmetry of molecules and spectra

In the previous chapter we developed a methodology for comparing NMR chemical shift predictors. As outlined there, one of the main applications of such predictors is the assignment of NMR signals to nuclei in the structure. Currently available tools for NMR assignment heavily rely on the accuracy of the chemical shift predictors. Unfortunately, this is a property that depends on many variables and that sometimes cannot be predicted with enough accuracy to produce unambiguous assignments by itself. For this reason, we designed another way to assign NMR spectra that doesn't rely on chemical shift predictions but that can make good use of them when available.
Follows a reproduction of the article published in the Journal Of Chemical Physics [**25**].

4.1. Introduction

Several problems in chemical physics deal with optimization in large and complex solution spaces. These include the well-known problem of molecular geometry optimization, but also discrete optimization problems arising in protein design and folding [**120**], combinatorial optimization problems underlying new-compound design [**22**], and optimal path problems in multi-photon dynamics [**65**], to mention a few.
Many methods exist for tackling with them, each with its own limitations. First, naive search attempts to test all possible solutions in order to find the absolute optimum. Such approach is unsuitable in most cases, where the solution space is either infinite or just too large. Gradient methods follow descending paths from a starting seed until a minimum is located; there is no guarantee, however, that this is the absolute minimum of the solution space. In fact, gradient methods tend to get stuck in local minima. Methods such as Simulated Annealing and Genetic Algorithms attempt to address this issue by systematically sampling the solution space, in order to locate the absolute optimum without thoroughly considering all possibilities. Though very powerful and useful, sampling of the solution space in these methods is ultimately stochastic and incomplete, so they are still unable to provide a certificate of optimality for the solution found. Last but not least, complete-search methods such as Branch-and-Bound [**41**] (B&B) aim to achieve a thorough exploration of the solution space without actually testing every single possibility. In the case of B&B this is achieved by partitioning the solution space in clusters (known as branches) with definite bounds for the value of the goal function, rejecting branches with insurmountably suboptimal bounds, and recursively repeating the process with viable branches.

B&B methods are attractive because without traversing every possible solution, as naive search does, they can guarantee absolute optimality in the solution found, a feat unattainable by sampling and gradient methods in most cases. Combining robustness with efficiency, a well-designed B&B algorithm can outperform genetic algorithms in terms of cost-effectiveness [**22**]. Yet, success of a B&B search is not a given; it depends at least on three key factors: the size of the solution space, the size of the branches generated by the algorithm, and the differences between branch bounds. The ideal is that, from the first steps of the iteration, the solutions are split in large branches with large gaps between their bounds, allowing for *en masse* rejection of lots of nonviable candidates.

In this chapter we approach the problem of NMR assignment as a combinatorial optimization problem and propose three strategies that can be implemented to increase the viability and efficiency of automatic NMR assignment by B&B search. The foundation of these strategies lies in the recognition that molecular symmetry imposes restrictions on the goal function and solution space of the problem that can be used to improve the B&B algorithm relative to the three factors mentioned above. Though we will present all our ideas within the context of automatic NMR analysis, given the ubiquity of molecular symmetry in chemical physics we expect that the strategies proposed will be sufficiently general to be applicable to other sub-fields.

Assignment of observed NMR peaks to each nucleus of a candidate structure is a prevalent challenge in NMR spectroscopy and related applications. Indeed, it is the most common procedure for structure validation by NMR, thus playing a key role in structure elucidation and validation by NMR spectroscopy, which in turn is a core component of new compound discovery/synthesis and related fields. Furthermore, most of the research in the field of NMR chemical shift prediction [**19**,**27**,**83**,**85**,**87**,**110**,**111**,**112**,**119**,**122**] and automatic elucidation [**33**,**48**,**49**,**63**,**84**,**86**,**108**] depends on repositories of well-assigned NMR data. Although manual assignment by an expert is the most widely used and most reliable method, the already big and continuously growing amount of information produced nowadays demands computational tools for assisting this task [**56**,**58**].

Small molecules have been the target for automatic assignment strategies ever since the appearance of 2D NMR correlation experiments [**104**]. More recently, much larger molecules of biological interest, such as proteins and DNA strains, became accessible to very high field NMR spectrometers and thus became the focus of bioinformatics [**20**, **21**, **24**, **32**, **44**, **60**, **61**, **62**, **64**, **67**, **76**, **80**, **89**, **102**, **117**, **118**, **125**] to assist with the tedious assignment of hundreds of peaks. Advances in that field are spectacular and take advantage of the sequential motive particular to these macromolecules and of powerful isotope labeling strategies. In contrast, progress on the field of small molecule assignment has been slower. The fact is that small molecules are amazingly challenging due to the enormous variety in their configurations. There are currently 3 leading computational tools for assignment of small molecules: two commercial proprietary applications (ACD auto-assignment [**72**] and Mestre Nova [**103**]), and a free computer-assisted assignment application (CASA [**117**]).

In general terms, the problem of NMR assignment consists in matching each resonating nucleus of a given chemical structure with its corresponding coordinate in a given spectrum, according to the spectroscopic data contained in the spectrum itself or obtained through other NMR experiments (see Figure **4-1**). The degree of agreement with input data may be represented by means of a scoring function. Re-casted in such terms, it becomes clear that NMR assignment is indeed a combinatorial optimization problem, whose goal is to find the assignment with the highest score. The scaling of the solution space of this particular problem can be very intimidating: given a spectrum with n peaks and a candidate molecule with m nuclei, we first have to consider all possible partitions of the m nuclei into n sets, whose number is given by the Stirling numbers of the second kind, $\left\{ {m \atop n} \right\}$. Then, for each partition, we can distribute the subsets of nuclei among the observed peaks in $n!$ different ways, giving a total of $\left\{ {m \atop n} \right\} n!$ possible solutions to the assignment problem. To get an idea of how fast this number scales, note that even a simple molecule such as ethyl-benzene has 10 hydrogen nuclei that give 5 peaks in a fully-resolved ^{1}H-NMR spectrum, meaning that there are in principle over five million possible assignments. For mid-sized molecules, the number of feasible solutions easily scales up to unmanageable magnitudes.

These considerations reveal NMR assignment as a typical example of a combinatorial optimization problem in chemical physics. Developing a viable algorithm for automatic assignment that offers a certificate of optimality on the solution found demands careful consideration of the restrictions involved, in order to avoid exploration of redundant regions of the unmanageably large solution space while still ensuring that no relevant spot is left unvisited. As already mentioned, this awareness translates into understanding the *symmetry* of the problem. The core of this chapter is the assertion that permutations of symmetric peaks in the spectrum and of symmetric nuclei in the molecule generate equally-viable assignments, i.e. solutions with the same value of the goal function. This is an intuitive result that is relatively straightforward to prove; yet it suggests several powerful strategies to tap with the inherent difficulties of automatic NMR assignment.

The first section of the chapter is devoted to a proof of the assertion. To ensure sufficient generality, we put the NMR assignment problem within a Bayesian framework that makes no assumptions regarding the assignment methodology other than *a priori* considering all possible assignments to be equally likely. The likelihood of observed NMR properties is introduced as the goal function or score to be maximized. Afterwards, two groups of permutations $\mathscr{P}$ and $\mathscr{Q}$ are introduced in order to describe the symmetry of the molecule and the spectrum *relative* to these properties. Then, we prove that property likelihoods remain unchanged under the action of these groups, as expected.

The second section discusses how this result can be used to improve the efficiency of automatic NMR assignment. We open the first subsection recalling that equivalent nuclei (up to enantiotopicity) must be assigned to the same chemical shift in any given NMR experiment, and reinterpreting this well-known property of Nuclear Magnetic Resonance in terms of the language introduced on Section 1. We then show how the molecule's global

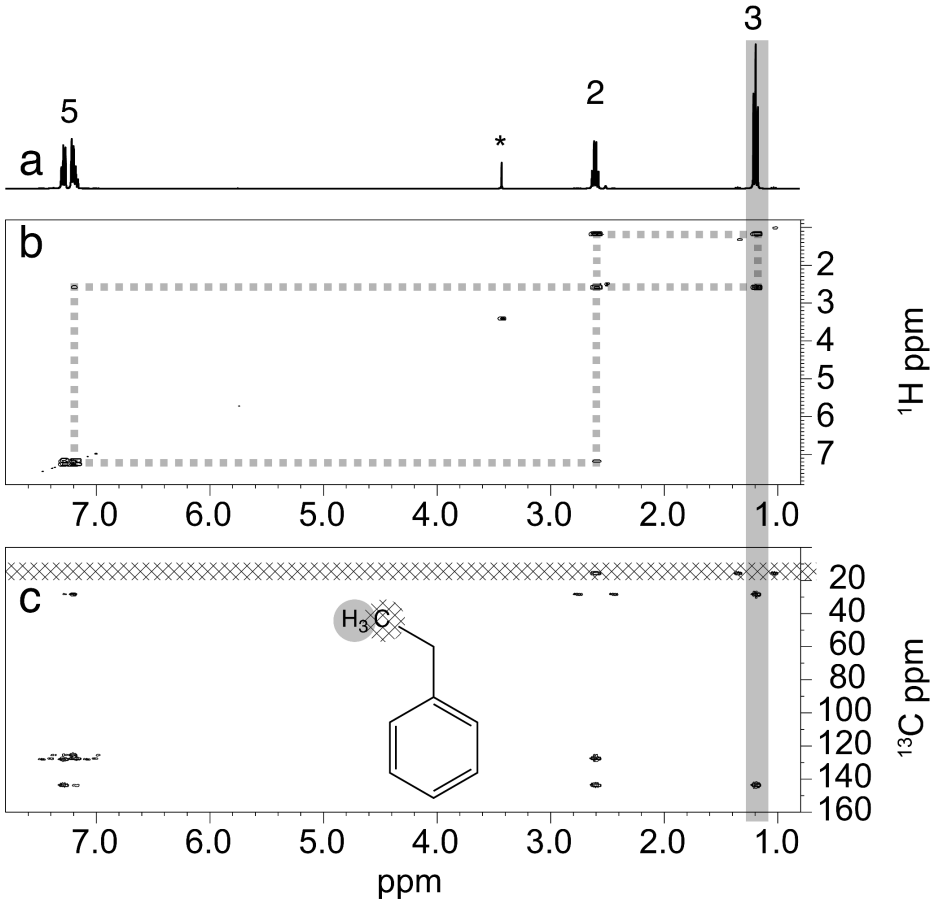


FIGURE **4-1**. Connection between NMR coordinates and nuclei. ^{1}H (a), COSY (b) and HMBC (c) spectra of ethyl-benzene are depicted. Each hydrogen nucleus corresponds to a coordinate on the x axis, while each carbon nucleus corresponds to a coordinate on the y axis. The NMR assignment problem consists in reconstructing this correspondence, guided by the relationship between chemical structure and nuclear magnetic resonance data. The signal marked with an asterisk is attributed to solvent. The gray and crossed area highlights the redundancy present in NMR experiments. [**25**]

group of symmetry can be used to construct a condensed version of structure formulae that implicitly disallows any solution to the assignment problem that does not satisfy the above-mentioned restriction. By performing the assignment of chemical shifts on the condensed structure rather than on the full structure a vast number of sub-optimal candidate assignments are disregarded, which somewhat counters the unfavorable escalation of the solution space and in this way improves the efficiency of optimal assignment search.
The second subsection considers the effects of the symmetry of molecule and spectrum *relative* to the observed NMR properties, which corresponds to the most general case formalized in the first section of the chapter. We discuss how to take advantage of the invariance of the assignment score towards the action of $\mathscr{P}$ and $\mathscr{Q}$ in order to increase branch size and thus accelerate pruning of the search tree in B&B strategies.
The gain achieved by increasing branch size, however, has to be weighted against the cost of computing these groups of symmetry. Furthermore, the speed at which the score gaps between branches increase is of capital importance for optimizing pruning speed as well. Both factors will depend on the choice of the properties to be scored during the assignment process. As we analyze this matters, we present arguments in favor of assignment methodologies that opt for robust properties such as integration, correlation patterns and broad chemical shift restrictions for the greater part of the assignment process, as opposed to methodologies that rely on accurate chemical shift predictions.

4.2. Effect of spectrum and molecular symmetry in assignment scores

An assignment is a function f that maps each resonating nuclei in a molecule G to a chemical shift δ of a spectrum. A vector of chemical shifts $\vec{\delta} = (\delta_1, \ldots, \delta_m)$ defines a coordinate in an m-dimensional NMR spectrum. In the case of 1D experiments $m = 1$ so that the chemical shift vector has but one component; in such case we may just write δ instead of $\vec{\delta}$. A function S that maps the $\vec{\delta}$'s into an appropriate codomain defines an observed NMR property. For example:

- integration in a ^{1}H-NMR spectrum could be defined by a function S such that $S(\delta)$ is always a positive integer.
- a cross-peak pattern in a H,H-COSY could be defined by a function S such that for each $\vec{\delta} = (\delta_1, \delta_2)$, $S(\delta_1, \delta_2) = 1$ if a cross-peak is observed at these chemical shift coordinates, and $S(\delta_1, \delta_2) = 0$ otherwise.

The NMR assignment problem consists in finding an assignment that better fits the observed NMR properties. The suitability of an assignment in light of property S can be expressed as the posterior probability

$$P(f|S) = \frac{P(S|f)P(f)}{P(S)}. \tag{4-1}$$

If we assume that all assignments are equally likely *a priori* we get

$$P(f|S) = KP(S|f), \tag{4-2}$$

where $K = \frac{P(f)}{P(S)}$ is a constant. This shows that the solution to the assignment problem is given by the f's that maximize the likelihood $P(S|f)$. It follows that if a series of independent NMR properties e.g. $S_{Integration}, S_{COSY}$, etc. is to be considered, the posterior probability $P(f|S_{Integration}, S_{COSY}, \ldots)$ is proportional to the product of their likelihoods,

$$P(f|S_{Integration}, S_{COSY}, \ldots) \propto P(S_{Integration}|f)P(S_{COSY}|f)\cdots \tag{4-3}$$

Moreover, for a spectrum with peaks at chemical shift coordinates $\vec{\delta_1}, \ldots, \vec{\delta_n}$ we may write without loss of generality:

$$P(S|f) = \prod_i P(S(\vec{\delta_i})|f) \tag{4-4}$$

Since the likelihood $P(S(\vec{\delta_i}|f))$ depends only on the nuclei assigned to the $\vec{\delta_i}$ chemical shifts we have:

$$P(S|f) = \prod_i P(S(\vec{\delta_i})|f^{-1}(\vec{\delta_i})). \tag{4-5}$$

We shall use this latter as a general expression of assignment score.
Clearly, if we exchange two chemical shifts δ_i, δ_j with the same integral, the spectrum has not changed as far as integration concerns. If the spectrum was fully assigned, swapping these chemical shifts induces a new assignment of the original spectrum that agrees with the observed integration just as much as the previous one did. It is readily proven that this result holds in a more general sense: not just for integration but for any property S, not just for chemical shifts but for nuclei as well, and not just for pairwise-swaps but for any arbitrary permutation that leaves S unchanged.
Formally, let $\mathscr{G}$ be a group of permutations on a set X and f a function with domain in X; $\mathscr{G}$ acts on f according to

$$(g \circ f)(x) = f(g(x)) \text{ for all } x \in X, g \in \mathscr{G}. \tag{4-6}$$

In such terms, spectrum symmetry is determined by $\mathscr{P}$, the largest group of permutations on the set of chemical shifts whose action lets S unchanged, i.e. $p \circ S = S \forall p \in \mathscr{P}$. We state that the action of $\mathscr{P}$ preserves the posterior assignment probability $P(f|S)$: combining equations (4-2) and (4-4) yields

$$P(f|S) = K \prod_i P(S(\vec{\delta_i})|f^{-1}(\vec{\delta_i})). \tag{4-7}$$

Now we consider the effect on $P(f|S)$ of a permutation $p \in \mathscr{P}$ acting on f:

$$P(p \circ f|S) = K \prod_i P(S(\vec{\delta_i})|f^{-1}(p^{-1}(\vec{\delta_i}))). \tag{4-8}$$

Because $\mathscr{P}$ is a group then $p^{-1} \in \mathscr{P}$. Furthermore, p^{-1} is one-to-one and since by construction $S(\vec{\delta_i}) = S(p^{-1}(\vec{\delta_i}))\forall i$ we have:

$$\begin{aligned} P(p \circ f|S) &= K \prod_i P(S(p^{-1}\vec{\delta_i})|f^{-1}(p^{-1}(\vec{\delta_i}))) \\ &= K \prod_i P(S(\vec{\delta_i})|f^{-1}(\vec{\delta_i})) \\ &= P(f|S) \\ P(p \circ f|S) &= P(f|S) \end{aligned} \tag{4-9}$$

which proves that assignment score is invariant towards spectrum symmetry relative to any NMR property.

To write the equivalent statement for nuclei permutations we have to transform (4-5) to express it in terms of the assigned nuclei. NMR specialists estimate the likelihood of observed properties by comparing the experimental values with those predicted from the structure of the molecule. So let v be a resonating nucleus in the molecule, X the set of possible values of property S, and $x \in X$. A specialist's prediction on a property observed at the chemical shift coordinate $\vec{\delta}$ can be expressed through the conditioned probability

$$P(x|f^{-1}(\vec{\delta}) = \vec{v}) \tag{4-10}$$

where $\vec{v}$ matches the dimension of the spectrum. We may use the notation $P(x|\vec{v})$ instead for simplicity. When the peak observed at $\vec{\delta}$ is deemed to be the product of the overlap of signals due to several different $\vec{v_i}$, properties of the observed signal are predicted by "superposing" the predictions for each $\vec{v_i}$. The way this superposition is carried out changes from one property to another; for example, integrals are superposed by adding them, while cross-peaks are superposed through an OR operation: if a cross-peak is expected for any of the $\vec{v_i}$, a cross-peak is expected for their superposition. For the general case, let $P(S(\vec{\delta}|\vec{x})$ be the probability of obtaining $S(\vec{\delta})$ from the superposition of peaks with property values $\vec{x} = (x_1, \ldots, x_k)$; the likelihood $P(S(\vec{\delta}|f^{-1}(\vec{\delta}))$ can be expressed as

$$P(S(\vec{\delta})|f^{-1}(\vec{\delta})) = \int_{\vec{x}} P(S(\vec{\delta})|\vec{x}) \prod_{\vec{v_j} \in f^{-1}(\vec{\delta})} P(x_j|\vec{v_j}) d\vec{x}. \tag{4-11}$$

The integral appears from considering the whole possible range of values that the property may be expected to take, with probability $P(x_i|\vec{v_j})$, for each of the nuclei assigned to $\vec{\delta}$. In other words, we are computing the likelihood by marginalizing over the parameter $\vec{x}$, the vector of property values predicted for the superposed signals. Substituting in (4-4) and (4-2) we get an expression for the posterior probability of assignments in terms of the nuclei of the candidate molecule:

$$P(f|S) = K \prod_i \int_{\vec{x}} P(S(\vec{\delta_i})|\vec{x}) \prod_{\vec{v_j} \in f^{-1}(\vec{\delta_i})} P(x_j|\vec{v_j}) d\vec{x}, \tag{4-12}$$

We shall consider molecular symmetry relative to a specific NMR property S. So let $\mathscr{Q}$ be the largest group of nuclei permutations such that for all $q \in \mathscr{Q}$,

$$P(x|q(\vec{v})) = P(x|\vec{v}) \forall \vec{v}. \tag{4-13}$$

This means that q exchanges nuclei that are indistinguishable regarding S. We are interested in the effect of q on the posterior probability (4-12):

$$\begin{aligned} P(q \circ f|S) &= K \prod_i \int_{\vec{x}} P(S(\vec{\delta_i})|\vec{x}) \prod_{\vec{v_j} \in q^{-1}(f^{-1}(\vec{\delta_i}))} P(x_j|\vec{v_j}) d\vec{x}, \\ &= K \prod_i \int_{\vec{x}} P(S(\vec{\delta_i})|\vec{x}) \prod_{\vec{v_j} \in f^{-1}(\vec{\delta_i})} P(x_j|q(\vec{v_j})) d\vec{x}. \end{aligned} \tag{4-14}$$

Substituting (4-13) we find

$$\begin{aligned} P(q \circ f|S) &= K \prod_i \int_{\vec{x}} P(S(\vec{\delta_i})|\vec{x}) \prod_{\vec{v_j} \in f^{-1}(\vec{\delta_i})} P(x_j|\vec{v_j}) d\vec{x} \\ P(q \circ f|S) &= P(f|S), \end{aligned} \tag{4-15}$$

i.e. that $P(f|S)$ is unchanged by nuclei permutations satisfying (4-5).

4.3. Application to the solution of the assignment problem

4.3.1. Reducing the size of solution space: Condensed Symmetry Structure. A trained chemist does not assign hydrogen nuclei individually, but as ensembles. For instance, he would know that hydrogens from the methyl/methylene group in ethyl-benzene are to be assigned to the same peak and would thus only consider $\left\{ {7 \atop 5} \right\} 5! = 16800$ assignments to begin with. This is a particular instance of a general rule stating that homotopic and enantiotopic nuclei are magnetically equivalent, meaning that they resonate at the same frequencies, thus generating a single peak in a NMR experiment. In other words, homotopicity and enantiopicity are expressions of the *absolute* symmetry of a molecule, that holds *relative* to all NMR properties. Permutations of homotopic/enantiotopic nuclei should then, according to the previous section, preserve the score of all assignments. This equivalence is not limited to nuclei in the same functional group: for instance, on closer inspection a chemist would notice for ethyl-benzene that pairs of hydrogens in the *orto* positions are symmetric and that the same is true for the *meta* hydrogens, which would lead to a total of just $\left\{ {5 \atop 5} \right\} 5! = 120$ possible solutions, a major reduction in the size of the solution space (see Figure **4-2**a).
The restrictions introduced by topicity can be accounted for by computing the quotient graph of the molecule's group of symmetry (up to enantiotopicity) and performing the assignment on the resulting formula rather than on the full chemical structure. Explicitly, let G stand for the graph of some given chemical structure and $\mathscr{P}$ be the group of permutations that exchange homotopic and/or enantiotopic nuclei in G. The natural

action $\circ$ of $\mathscr{P}$ on G is defined by

$$P \circ v = P(v) \tag{4-1}$$

for each permutation $P \in \mathscr{P}$ and for each vertex $v \in G$. The quotient of G under the action of $\mathscr{P}$ is the graph $B(G)$ whose vertices are the orbits $\mathscr{P}v$ of this group action,

$$\mathscr{P}v = \{P \circ v : P \in \mathscr{P}\} \tag{4-2}$$

and whose edge set is $\{(\mathscr{P}u, \mathscr{P}v) : (u, v) \in E_G\}$. We refer to the graph $B(G)$ as the Condensed Symmetry Structure (CSS) of the molecule (see Figure **4-2**). Vertices in $B(G)$ are labeled as X_i^a where X is the chemical symbol of the corresponding element, i is the number of nuclei in the orbit, and a is a unique label used to identify the vertex. When the element and number of nuclei are irrelevant, we may just use the label a to refer to the vertex.

The CSS is readily computable using an algorithm describe elsewhere [**37**] and is useful for NMR spectra interpretation because, unlike what happens with the ordinary structure formula, each vertex in the CSS corresponds exactly to one peak in a fully resolved NMR spectrum and vice-versa. Furthermore, the size of each vertex in $B(G)$, i.e. the number of nuclei in the corresponding orbit, is equal to the integral observed for the corresponding fully resolved ^{1}H-NMR peak, as already noted in [**79**]. All the connectivity information allowing for the prediction of cross-peaks in multidimensional experiments is preserved in the CSS, encoded as n-length paths between ^nJ-coupled nuclei classes. And, even though peak multiplicity is not properly reflected in the graph as defined above, it can be recovered by weighting the edges in a similar manner to what is done for vertices. On the other hand, it is not clear whether it is possible to predict the chemical shifts from the CSS, but this is not such a big downside considering that chemical shift prediction from the full structure is already a convoluted business.

On top of keeping enough information to allow for the prediction of most relevant NMR properties, the CSS neglects all assignments that identify magnetically-equivalent nuclei with different peaks in the spectrum, reducing significantly the size of the solution space of the assignment problem. Indeed, rather than considering all $\left\{{m \atop n}\right\} n!$ assignments, where m is the number of nuclei in the full structure, now we are only considering $\left\{{k \atop n}\right\} n!$ assignments consistent with molecular symmetry, where $k \leq m$ is the number of vertices in the CSS. The reduction achieved in the size of the solution space tends to the asymptotic value n^{m-k} for fixed number of peaks n and is equal to $\left\{{m \atop n}\right\}$ for a fully resolved spectrum. In practical cases this often amounts to a reduction of several orders of magnitude, as shown in Figures **4-2** and **4-4**.

4.3.2. Improving the efficiency of solution search.

Branch-and-Bound strategy for NMR assignment. Several strategies have been proposed to tackle the problem of NMR assignment, particularly in the field of protein NMR assignment. This includes the use of neural networks [**61**], simulated annealing [**89**, **102**],

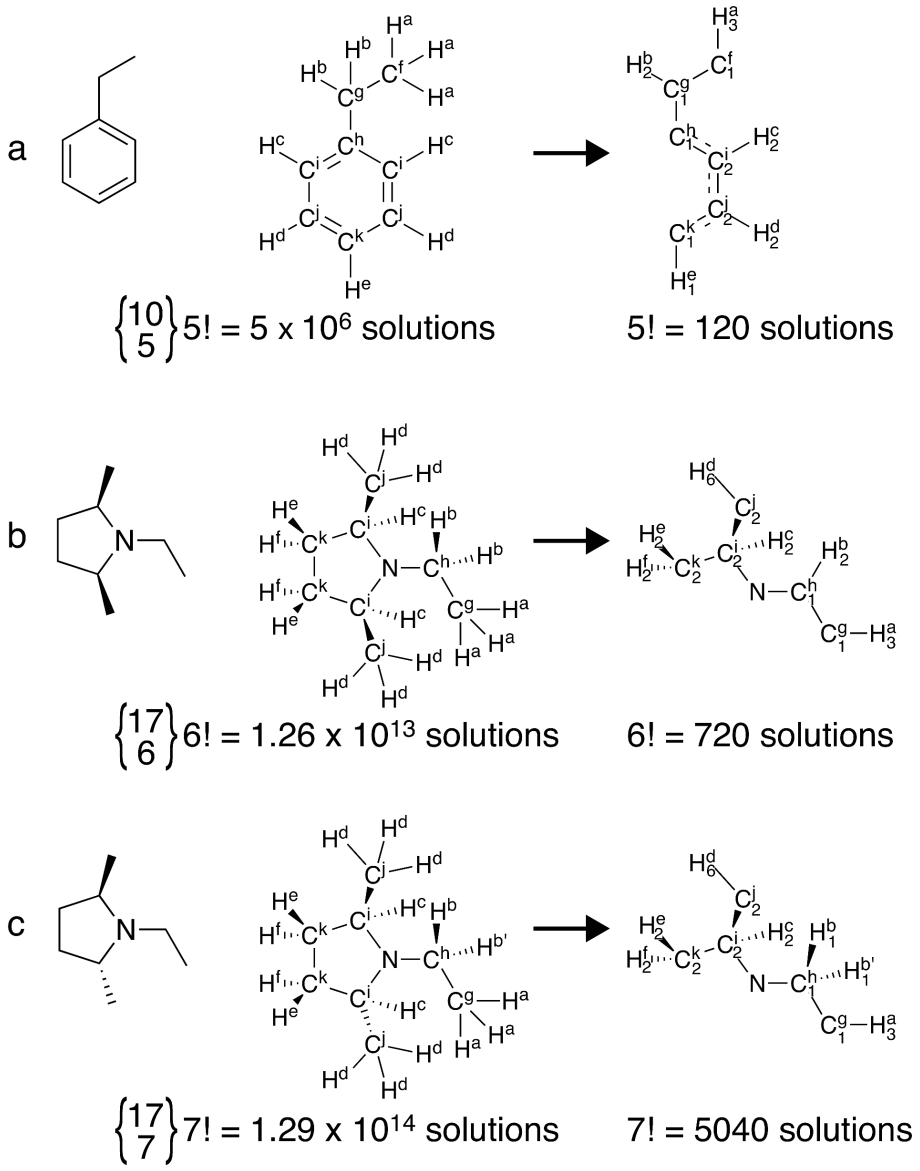


FIGURE **4-2**. Three small molecules along with their corresponding Condensed Symmetry Structures (quotient graphs). Superindices identify families of symmetric nuclei with their corresponding vertex in the CSS, subindices correspond to the number of equivalent nuclei. The size of the corresponding solution spaces of the associated assignment problems for fully resolved spectra are also depicted. [**25**]

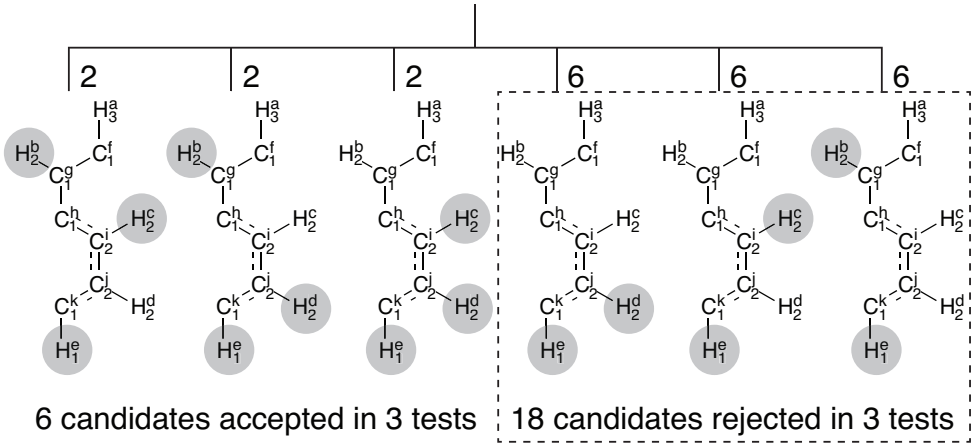
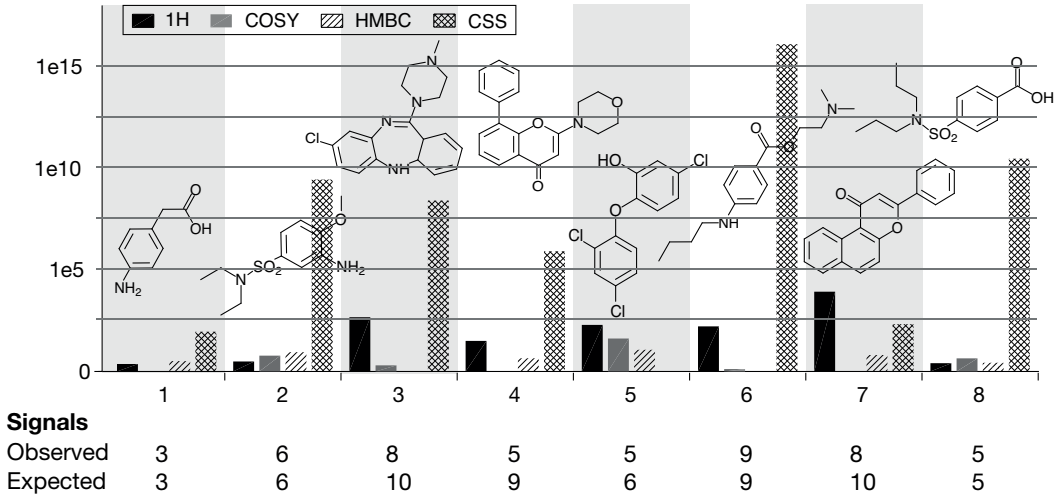


FIGURE **4-3**. Fragment of the first level of a B&B search for the assignment of the ^{1}H spectrum of ethylbenzene (see Figures **4-2**a and **4-3**a) based on integration. A per-peak branching procedure is used: each branch corresponds to a set of assignments that assign the same given CSS vertices (circled in green) to the first peak of the spectrum (Integral=5). For instance, the leftmost branch comprises 2 assignments that map vertices b, c, and e to the first peak of the spectrum; one of these assignment maps vertex a to the second peak and vertex d to the third peak, while the other does exactly the opposite. As all assignments in the same branch predict the same integral for the first peak, they are simultaneously accepted or rejected by comparison with the observed value. Here each of the three branches to the left is accepted, as they predict the observed integral of 5, while the three to the right are rejected, as they predict an integral of 3, amounting to a reduction in the number of assignment score computations by a factor of 4. [**25**]

mean-field simulated annealing [**32**], genetic algorithms [**24**, **118**], Monte Carlo optimization [**62**, **76**, **80**] full and heuristic search [**20**, **21**, **44**, **60**, **67**] and best-first approaches as proposed with Auto-Assign by Zimmerman et al [**125**] and CASA by Wang et al [**117**]. Among all these strategies, the Branch-and-Bound method (B&B) intends to allow a complete exploration of the solution space, thus ensuring that solutions found are optimum. Branch and Bound (B&B) search [**41**] consists in a systematic enumeration of all candidate solutions, grouped in clusters or 'branches' whose scores have been bound according to the evaluated data. This allows for *en masse* rejection, while building the tree, of non-viable candidates whose bounds are found inferior to those of alternative branches. In the case of NMR assignment, an intuitive branching procedure consists in assigning one chemical shift at a time, successively evaluating the relevant NMR properties. For example, the assignment of the ^{1}H-NMR spectrum of ethyl-benzene would start by selecting a peak and scoring all possible assignments of nuclei to that peak according e.g. to the peak's integral, chemical shift, and/or multiplicity. Branches are defined as sets of assignments that map the same nuclei to such peak (see Figure **4-3**). Further steps would sequentially

consider remaining unassigned peaks. The success of a B&B strategy depends on the size of the branches generated and on the differences between their lower and upper bounds. An ideal algorithm would generate large branches on the early steps of the search with quickly increasing gaps in their score bounds. In this way, non-viable branches with low scores can be pruned early during the search, promptly narrowing the number of feasible solutions within a few steps.

Symmetry restrictions increase branch size. The intuitive per-chemical shift branching procedure previously described groups together solutions that are known to have the same score up to a certain point by having assigned the same vertices to some given chemical shifts. We can improve on this strategy by clustering together branches that are known to have the same scores by having assigned *symmetric* vertices to *symmetric* chemical shifts. The scores of all clustered branches can then be determined by computing the scoring function on a single assignment, improving the speed of the search provided symmetry can be readily computed. For example, note that the three leftmost branches in Figure **4-3** differ only by permutations of vertices of size 2, which from the results of Section II means that they are equally fitting, and thus we may cluster them together in a single branch that is accepted after one single score test. The same is true of the three rightmost branches; so by exploiting these symmetries of the assignment problem, we may accept/reject the set of assignments of Figure **4-3** after just two rather than six calls to the score function. In order to apply this strategy it is necessary to compute the relevant group of symmetry. The complexity of this task varies greatly from one property to another. Furthermore, recall that the group itself is contingent on the property with respect to which the assignment is being scored. Choosing the adequate properties then has a great impact in the speed of the search, specially in the early stages where the number of viable solutions is largest and we want to partition this huge set of possibilities in a few large branches.
Integration is an interesting property on this regard. Orbits of the relevant group of symmetry are trivially determined by clustering together peaks that have the same integral, and CSS vertices of the same size. Furthermore, groups of peaks with the same proton integral are common, anticipating that a large improvement in branching is often possible. On the opposite extreme we have chemical shift, which is quite characteristic of each class of resonating nuclei, represented by a vertex in the CSS. No symmetries are expected then, so no improvement would be achieved. An alternative strategy that may prove much more useful consists in considering broad *intervals* of chemical shifts, where e.g. aromatic or aliphatic protons are expected to be present. This technique partitions the set of resonating nuclei in large classes of equivalences that should lay within a given zone of the spectrum, effectively creating large branches that can be pruned or accepted in a single step.
Cross-peak patterns in 2D experiments present a different case. Here the limiting step is the computation of the group of symmetry of the spectrum regarded as a matrix of couplings. This problem is equivalent to that of computing the group of automorphisms of a graph, which is NP-complete. Furthermore, the assignment of one chemical shift



Signals							
Observed	3	6	8	5	5	9	8	5
Expected	3	6	10	9	6	9	10	5

FIGURE **4-4**. Reduction in the size of the search space achieved by the CSS and reduction in the number of B&B search steps achieved by taking symmetry restrictions into account for the assignment of ^{1}H, COSY and HMBC spectra of 8 chosen molecules. Assignment of ^{1}H is based on integration, that of HMBC is based on the number of couplings observed for each shift in the H coordinate, and that of COSY is based on the full H-H coupling pattern. Values in the y axis correspond to the ratio between the number of solutions or search steps taken when symmetry restrictions are not taken into account, and the number of solutions/search steps when symmetry is taken into account. Small molecules were chosen with significant variety in symmetry, size, complexity, and peak overlap across this small sample. On this regard, note the comparison between the number of observed and expected signals (i.e. chemical shifts) reported at the bottom of the image. [**25**]

may break the symmetry of the graph, forcing re-computation of the group. The burden of repeatedly carrying this calculation may overshadow the gain of improved branching, making the strategy non-viable.

Figure **4-4** compares the reduction in the number of nodes visited by a B&B procedure during the assignment of the hydrogen coordinate on a set of 8 molecules. The basic per-chemical shift strategy was compared with symmetry branching, using three different properties: ^{1}H-NMR proton integration, H,H-COSY cross-peak patterns, and number of HMBC cross-peaks. Spectra used in this exercise were simulated used the method described in [**35**], only molecular symmetry was taken into account, and only perfect matches were accepted. It is immediately noted that the gain achieved by the improved B&B strategy often pale in comparison with the terrific reduction in the size of the solution space achieved by using the CSS, which goes up to 16 orders of magnitude in the examples considered. The value of symmetry branching is not to be neglected, however. First note that it achieves an increase in speed of up to 4 orders of magnitude for assignment

based on proton integration. Proton integrals are the cheapest deal when it comes to NMR assignment: data comes from a cheap experiment, score calculation involves simple arithmetic and logic operations with integers. It can then be used for a fast sweep of unviable candidate assignments. On top of that, as noted above, integral symmetries are trivial to compute and need only be computed once at the start of the assignment. A reduction of 4 orders of magnitude in the number of B&B nodes visited at such a cheap price is not to be dismissed.
COSY and HMBC present a less promising scenario. Here the reduction is of 2 orders of magnitude at most. Furthermore, the relevant groups of symmetry here are those of the 1-3 length H-H paths and 1-4 length C-H paths graphs respectively, which as noted before are more costly to compute. Plus, they need to be recomputed after each chemical shift is assigned. Yet, symmetry B&B assignment on COSY and HMBC data can still provide some value. On this regard Molecule 5, triclosan, provides an interesting example: the structure graph presents no symmetries whatsoever, so no reduction in the size of the solution space is achieved by the CSS. On the other hand, triclosan presents high symmetry on its spin couplings, which allows for a 100 times decrease in the number of nodes visited during B&B search. Last, note that the gain given by taking symmetry into account during B&B assignment scales up with the degree of overlap of the spectrum, so it still may be worth it in the case of highly overlapped 2D spectra.

Increasing the gap between branch bounds. In Section II we proved that permutation of symmetric chemical shifts or nuclei generates new assignments with the same score. In the previous subsection we discussed how to take advantage of this fact by creating branches of assignments that differ only by such permutations. However, we do not know how large the score gap would be between two different branches. This is a key factor in B&B search, as a branch can only be rejected after the gap between alternative branches has grown sufficiently large.
The gap created by a poor match relative to a good match is determined by the $P(x_i|\vec{v_j})$ parameters in equation (4-15), that describe our prediction of the values of the property expected from the assigned nuclei. Sharper distributions lead to larger gaps (Figure **4-5**). As success of a B&B strategy highly depends on the early detection of large branches of clearly sub-optimal solutions, the previous analysis points that sharply predictable properties with biased $P(x_i|\vec{v_j})$ distributions are most valuable when evaluated at early steps of the search, thereby achieving a greater gap between alternative branches and prompting the *en masse* rejection of unsuitable solutions. Once more, this points to the value of properties such as integration, that can be trivially predicted from examination of the CSS with a high degree of confidence, and to the risk of over-relaying on chemical shift, which comports a significant degree of incertitude even by the high standards of state-of-the-art predictors.

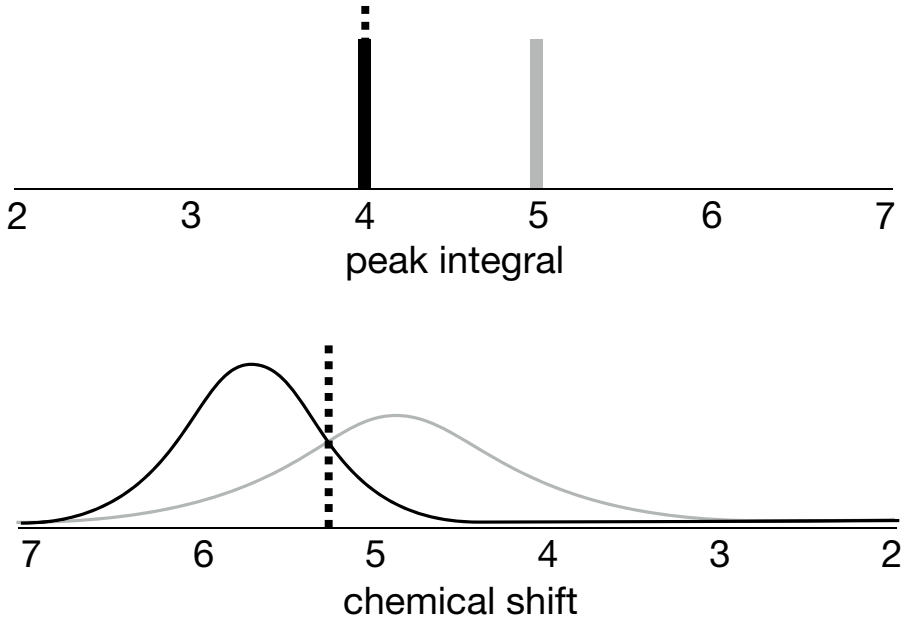


FIGURE **4-5**. Properties such as proton integration (top) can be precisely predicted, leading to sharp $P(x_i|\vec{v_j})$ distributions and immediate rejection of unfitting branches (gray distribution). Prediction of properties such as chemical shift (bottom) is more difficult and always involves a significant degree of error, leading to broader $P(x_i|\vec{v_j})$ distributions that may not allow for rejection of any branches: in this case, the branch corresponding to the gray distribution is just as good as that corresponding to the black one). The observed value of the property is represented by the dotted vertical line, while black and gray lines represent possible predicted values or their distributions. [**25**]

4.4. Conclusions

Assignment of NMR spectra of small molecules is a difficult combinatorial optimization problem associated to an intractable solution space. Solving it requires flexible and efficient algorithms that are able to explore this space efficiently. While methods that approximate the solution by using non-linear programming approaches like Monte Carlo methods are the most efficient, they present the disadvantage of not achieving a thorough exploration, increasing the risk of producing sub-optimal solutions. Complete-search methods such as B&B, on the other hand, carry a thorough exploration of the solution space, but demand greater care in the design of the search strategy in order to be viable. Awareness of the symmetries involved in the problem plays a key role in the design of a general purpose B&B algorithm that solves the NMR assignment problem for small molecules.

The most general elements of symmetry involved in NMR assignment are materialized in the CSS. This simple tool goes a long way towards the reduction of the combinatorial complexity of the assignment problem. It is intuitive for chemists, formalizing a key

step known to experts in manual assignment, keeps all the information relevant for the assignment process, and can be implemented efficiently in a computer program. Further and more specific symmetries that can be used to increase the efficiency of branching procedures are given by nuclei and chemical shift permutations that leave specific NMR properties unaffected. Recognition of even a single such transformation can half the number of branches appearing on one step of the B&B search, which in the long run may drastically reduce the number of operations performed. Such symmetries can be identified through examination of the CSS.

Awareness of this factors provides useful rules and heuristics that can be used to improve the speed of automatic NMR assignment. For instance, we argued that on the early steps of the assignment process reliability of property prediction is to be valued well above discriminating power, thereby giving properties such as integration a more prominent role in NMR assignment than they have been given previously.

.

CHAPTER 5

Fully automatic assignment of small molecules' NMR spectra without relying on chemical shift predictions

In this chapter we use the concepts developed in chapter **4**, to build a NMR automatic-assigner. As it should be known by now, the success of a automatic assignment system depends on the input data, i.e. the correct identification of the chemical shifts, coupling constants, and correlations observed in 1D and 2D NMR spectra. One of the main topics of this chapter is the description of the peak-picking algorithm used as pre-processing step for the NMR auto-assignment system. The idea behind this peak picking approach is to take advantage of the redundancy in the NMR information in order to validate the peak-picking to produce a more consistent set input for the assignment process. The second part of this chapter describes the NMR auto-assignment itself: The scoring strategy for each NMR property and the solution search strategy.
Follows a reproduction of the article published in Magnetic Resonance In Chemistry [**34**].

5.1. Introduction

For several decades, NMR has been used to elucidate the structure of molecules. Advances in parallel synthesis, continuous flow organic synthesis, microwave catalysis as well as solid phase organic reactions have considerably increased the number of new compounds produced, putting pressure on analytical labs to match the pace. While the experimental time for NMR acquisition has been substantially reduced with the introduction of pulse field gradients, high-field spectrometers, cryoprobes, and non-uniform sampling among others, the posterior analysis, i.e. the interpretation of the data, is still mainly done by hand. Despite several tools available to make the expert's life easier and steady and encouraging advances in the field [**43**,**50**,**56**,**75**,**94**,**99**], the challenge of automatic spectral analysis has not been yet definitely addressed.
Algorithms for computer-assisted assignment follow one of two strategies: one centered on analysis of the match between predicted and observed chemical shifts [**43**,**50**,**56**], and one centered on analysis of spin coupling data [**75**, **94**, **99**]. The first one plays an important role in protein backbone NMR assignment [**51**, **67**, **90**], where connectivity always follows a sequential motive imposed by protein primary structure and accurate chemical shift predictions are available. The latter strategy, on the other hand, is attractive for small molecules, where structural diversity leads to more varied spin coupling patterns and higher uncertainty on ^{1}H chemical shift predictions. Mixed strategies for small molecules NMR assignment should be able to take advantage of chemical shift restrictions without

depending on their availability and accuracy to produce quality results. This contribution presents such a system, designed around an integration and connectivity-centered approach.
Automatic assignment necessarily implies automatic peak-picking. Just as experts use their knowledge to pick peaks and assign their resonances at the same time, computational assignment systems should be able to discriminate meaningful signals from noise and assign them in an integrated process [**9**]. The first section of this paper presents a novel method for fully automatic peak-picking based on self-consistency among signals observed in different spectra.
The second section will board the problem of NMR assignment, first by introducing a reduced representation of structure formula used as a mediator in the assignment process. This significantly reduces the size of the associated search space. After presenting our methodology for scoring assignments, we consider the central problem of efficient exploration of the search space. The strategy chosen is a branch-and-bound algorithm that clusters assignments with the same score at each given step in the assignment process and rejects those whose upper bound has fallen below a certain threshold. This allows for *en masse* rejection of unsuitable assignments on the early steps of the search, warranting an efficient search that converges to the best candidates very quickly.
In the last section we show the results of an automatic assignment test using experimental ^{1}H, COSY, HSQC and HMBC spectra of 74 molecules and a web-based tool based on the methodology exposed is used to visualize the results.

5.2. Self-consistent peak-picking in 2D experiments

Peak-picking is a key feature for any automated task relying on spectral information [**53**] and a plethora of solutions have been proposed. Peak-picking consists first in distinguishing meaningful signals from noise, and second, in distinguishing relevant peaks from impurities and solvents. The first task has been achieved by several means, well reviewed by Garrett [**54**]: by defining a threshold value, by multiplet analysis, by line-shape analysis, or more recently by combining several such approaches. Once the noise has been detected and removed, if necessary, the remaining components of the spectrum are converted to one dimensional traces for further analysis by sophisticated algorithms [**15**, **16**, **70**]. Two bottlenecks remain: first the false positive rates and second the highly overlapped areas. To address the first issue methods have been proposed using wavelet to improve the initial detection [**123**], using complex filtering to improve the final selection [**10**, **90**], or using bayesian posterior probability and a bivariate gaussian model [**39**]. Attempts to tackle the second issue reported the use of three-way decomposition [**95**] or non-negative matrix factorization to decompose the spectra into components (peaks) [**113**], making use of the internal structure of the data. To our knowledge however, the redundancy of the information gathered from different NMR experiments has not been fully capitalized, indeed each spectrum is still analyzed individually. The novelty of the method described below is to improve the quality of the peak-picking by consistently comparing peaks detected

in several NMR spectra. For instance, experts would certainly expect signals in a COSY spectrum at positions where peaks have been detected in the corresponding 1D proton spectrum. Detecting these expected signals would in turn reinforce our confidence in the previously detected maxima. This iterative procedure closely resembles that of human experts: peaks are detected and scored according to their consistency with other signals observed (or not) in the same or in another related spectrum.
Thus the first step of our method consists in detecting meaningful signals. For 1D spectra we use a peak-picking algorithm based on the peak identification approach as proposed by Dietrich, Rüdel and Neumann **[46]**. The method is based on constructing a pseudo power spectrum of the data series which is the square of its first derivative. Signals with low intensity, attributable to noise, experimental artifacts, and in certain cases impurities and solvents, are eliminated by a recursive algorithm as follows:

(1) initialize a boolean mask for the data series with all values equal to true
(2) from the data series, calculate the mean and standard deviation for all points which have a true value in the bit-mask
(3) eliminate all the points with values below the mean plus x times the standard deviation, where x is a parameter determined by noise level
(4) go back to 2 until the mask is no longer changed

For 2D spectra (COSY, HSQC, HMBC), the second derivative method proposed for 1D spectra by Massart and collaborators **[116]** is applied using the LoG ("Laplacian of Gaussian") kernel approximation to efficiently approximate the second derivative of a smoothed 2D spectrum. In this methodology, peaks correspond to the minima of the second derivative curve. Minima of the discrete data series do not necessarily coincide with minima of the continuous surface, however. A better approximation to the 'true' peak coordinates is given by the center of mass of an area s around the minimum found.
For simple, fully-resolved singlet peaks, the coordinates of the center of mass are resilient with respect to the choice of s. In contrast, the center of mass of overlapped and multiplet signals may change dramatically depending on this area. To increase the robustness of the method, we propose a technique based on the boosting principle: repeat the peak-picking procedure using different values for s and then use a clustering method on all peaks generated to obtain the final set of peaks. For example, this allows to correctly position COSY cross-peaks, with multiplet structure, whose coordinates may not coincide with any of its local maxima.
On top of this peak-picking algorithm we add a second layer, intended to take advantage of data redundancy to improve the quality of automatic peak-picking. For instance, self-consistency demands that a peak detected at 1.2 ppm in the ^{1}H spectrum of a molecule is accompanied by a peak around the (1.2, 1.2) chemical shift coordinate of the corresponding COSY spectrum. In a similar way, a cross-peak at the (1.2, 6) coordinate should be accompanied by the corresponding diagonal peaks at (1.2, 1.2) and (6, 6), and by the symmetric cross-peak at (6, 1.2). One may then say that our posterior belief that a picked peak is a "true" peak depends on the faithful detection of other peaks.

Following this line of thought, we are building a self-consistent peak-picking method where picked peaks are scored with a probability related to their consistency with other picked peaks. For example, in the case of COSY spectra the steps are as follows (See Figure **5-1**): after each detected peak in the ^{1}H spectra has been given a unit score,

(1) set the score of each peak detected in the 2D spectra to 1
(2) increment the scores of diagonal peaks if they are aligned with peaks in the 1D trace (direct dimension)
(3) repeat this operation for the indirect dimension

Furthermore, COSY cross-peaks are validated against the expected symmetry of the spectrum, that is:

(4) increase the score of cross-peaks that are found aligned with a diagonal peak in the direct dimension
(5) repeat this operation for the indirect dimension
(6) increase the score of cross-peaks if its symmetric counterpart is detected
(7) the scores of symmetric COSY cross-peaks are made identical
(8) Scores are then normalized and the low-scoring peaks are filtered out:
(9) each score is normalized with respect to the maximum score expected for that peak (See Figure **5-1**)
(10) peaks with a score lower than a desired value are rejected

HSQC and HMBC spectra were also included into the self-consistent iterative peak-picking procedure. Similar to the steps described above, the scores of peaks in both heteronuclear spectra start at 1 upon detection then, increase the scores of

(1) any peak aligned with a peak in the 1D experiment (direct dimension)
(2) any HMBC peak aligned with a HSQC peak in the direct dimension
(3) any HMBC peak aligned with a HSQC peak in the indirect dimension
(4) any HSQC peak aligned with a HMBC peak in the direct dimension
(5) any HSQC peak aligned with at least 2 HMBC peaks in the indirect dimension (alignment with two peaks is required to account for ^{13}C satellites)
(6) any pair of HMBC peaks aligned in the indirect dimension.

Just as in the case of COSY, these scores are normalized to the highest expected score for each type of cross-peak. Having scored the peaks in all input spectra and selected a set of accepted peaks, both the peak set and their scores are used as input for the automatic assignment procedure. Figure **5-1** best illustrates the application of these peak scoring rules: correlated peaks will receive higher scores than uncorrelated ones, as desired.
To apply this peak-picking procedure it is necessary to define a criterion for saying that two peaks are aligned. Two parameters, ε ppm and $\varepsilon^{'}$ ppm, are introduced for this purpose: we say that two peaks are aligned in the proton dimension when the difference in their chemical shifts is less than ε ppm, and that they are aligned in the carbon dimension when the difference in their chemical shifts if less than $\varepsilon^{'}$ ppm.

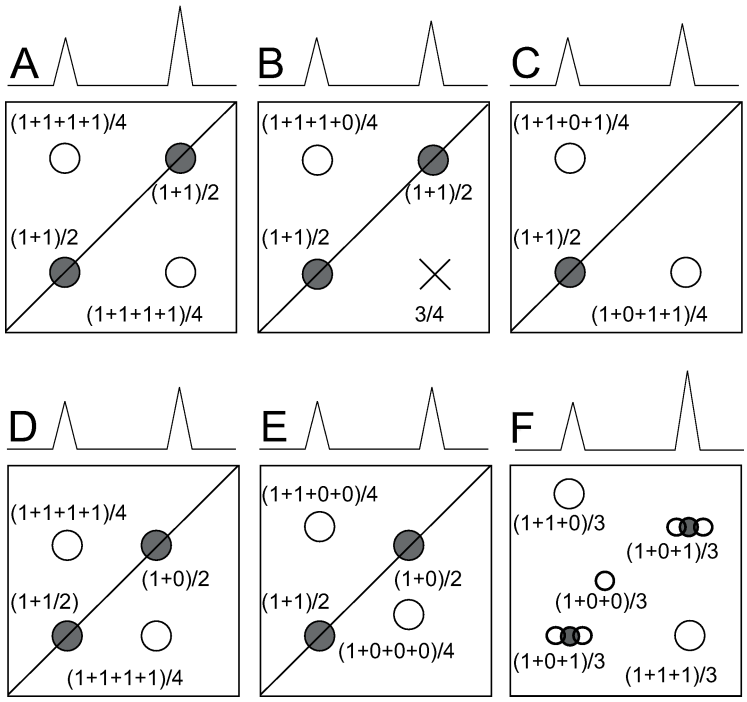


FIGURE **5-1**. Examples of peak scoring during automatic peak-picking. A) to E) COSY spectra: In the perfect case (A) all the peaks are detected and granted a unit score. If one of the cross-peaks is missing (B) both cross-peaks will receive a final score of $\frac{3}{4}$. If the diagonal peak of a coupling group of protons is not detected (C), each of its cross-peaks will be granted a score of $\frac{3}{4}$. A diagonal peak not aligned with a signal in the 1D spectrum receives a score of $\frac{1}{2}$ (D, E). Symmetric off-diagonal peaks that are properly aligned with diagonal peaks get the maximum score, even if the diagonal peaks themselves are not properly aligned with the 1D spectrum (D). On the other hand, non-symmetric off-diagonal peaks that are not aligned neither with diagonal peaks nor with 1D peaks receive a score as low as 1/4 (E), acknowledging that the likelihood of such signals being artifacts is high. F) HMBC spectra: peak scores are increased if they are vertically aligned with ^{1}H peaks or if they are horizontally aligned with HSQC peaks (grey circles). Misaligned peaks get a score between 1/3 and 2/3. For the purpose of assignment, 13-carbon satellites may be discarded on the basis of their score (the HSQC peaks give the same and more information) or detected by their disposition around an HSQC signal. All peak scores are normalized to the maximum possible score of 3. [**34**]

The last step of the peak-picking process consist in computing the integrals of the detected peaks. These integrals were readily computed by normalizing and rounding to the nearest integer. We did not concern ourselves with multiplicity in 1D traces. This choice was taken after considering that there is significant redundancy between the structural information given by peak multiplicities and the one given by cross-peaks in 2D experiments, and that computing multiplicities comports an additional and quite challenging step.

5.3. Automatic assignment

NMR assignment can be formulated as a combinatorial optimization problem, whose goal is to find the function f from nuclei in the suspect molecule to observed chemical shifts in the spectrum that better fits the observed data, as measured by an appropriate score function. The size of the associated search space scales as $\{m\ n\}n!$, where m is the number of nuclei, n is the number of chemical shifts, and $\{m\ n\}$ is a Stirling number of the second kind. This unfavorable scaling prevents the problem from being solved by full-search. We instead resort to a method that combines symmetry-based constraints with a branch-and-bound search [**74**] (BB), thus achieving an efficient and thorough exploration of the search space.

5.3.1. Condensed Symmetry Structure. Homotopic and enantiotopic nuclei are magnetically equivalent, meaning that they resonate at the same frequencies, thus generating a single peak in a NMR experiment. Assignment of homotopic/enantiotopic nuclei to different chemical shifts is then prohibited. This restriction introduced by topicity can be accounted for by computing what we call the Condensed Symmetry Structure (CSS) of the molecule [**25**], that is, the quotient graph of the molecule under its group of symmetry (see Figure **5-2**), and then performing the assignment on the CSS rather than on the full chemical structure.
The CSS is readily computable using an algorithm described elsewhere [**37**] and is useful for NMR spectra interpretation because, unlike what happens with the ordinary structure formula, each vertex in the CSS corresponds exactly to one chemical shift in a fully resolved NMR spectrum and vice-versa. Furthermore, the size of each vertex in the CSS, i.e. the number of nuclei in the corresponding orbit of its group of symmetry, is equal to the integral observed for the corresponding fully resolved ^{1}H NMR peak, and all the connectivity information allowing for the prediction of cross-peaks in multidimensional experiments is preserved, encoded as n-length paths between ^{n}J-coupled nuclei classes. On top of keeping enough information to allow for the prediction of most relevant NMR properties, the CSS neglects all assignments that identify chemically equivalent nuclei with different chemical shifts in the spectrum, reducing significantly the size of the search space of the assignment problem. Indeed, the reduction achieved in the size of the search space tends to the asymptotic value n^{m-k} where k is the number of vertices in the CSS. In practice, this amounts to an improvement of several orders of magnitude in the number of calculations necessary for determining the optimum assignment [**25**].

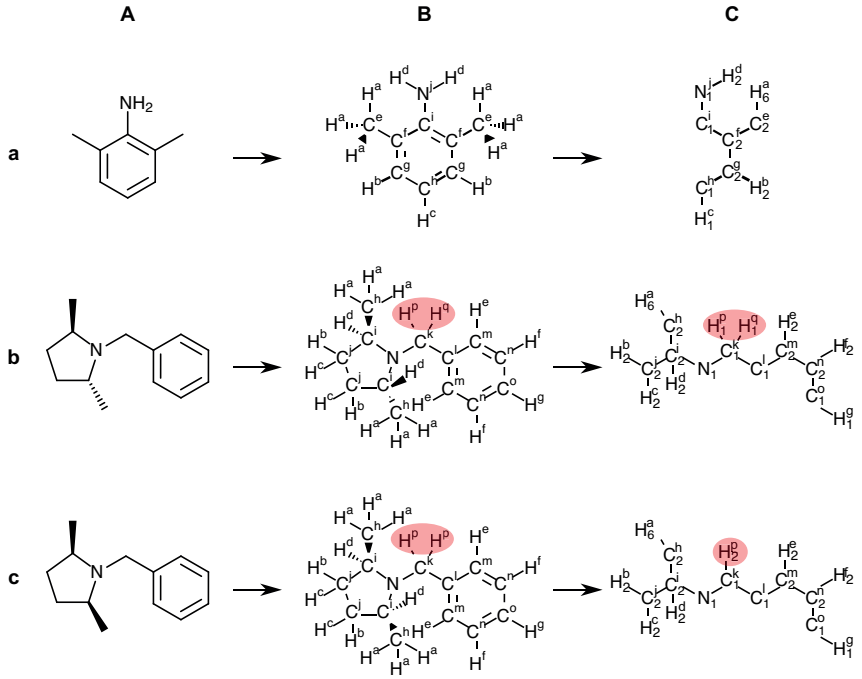


FIGURE **5-2**. Examples of some molecules (**A**) along with their corresponding Condensed Symmetry Structure (**C**). Computation of the CSS involves identification of all classes of magnetically equivalent nuclei in the molecule (**B**), which are determined by symmetry using the algorithm described in [**37**]. Superindices in the CSS are used to identify each family of symmetric nuclei with a unique label, subindices correspond to the number of equivalent nuclei in the family. Note that molecule **b** has a *trans* configuration that causes the H^p and H^q hydrogens (circled in red) to be diastereotopic and thus magnetically non-equivalent, which is reflected in the CSS. Molecule **c**, on the other hand, corresponds to the *cis* isomer, where these hydrogens are enantiotopic and thus magnetically equivalent (H^p), so that they are condensed in a single vertex in the CSS. This difference is in fact reflected in the corresponding ^{1}H NMR spectra, where one molecule presents a singulet (**c**) whereas the other presents two doublets (AB system) (**b**) [**28**]. [**34**]

5.3.2. Assignment scoring. Assignments were scored relative to three properties: ^{1}H NMR peak integrals, ^{1}H chemical shifts and COSY and HMBC correlations. In general, data may be put in a vector **u** comprising the values observed for each property on each (picked) ^{1}H chemical shift coordinate. Now, consider a candidate assignment *f*. Predicted property values can be similarly embedded in a vector **v**(*f*) comprising the corresponding predicted values, aligned with **u** according to *f*. Note that, as the notation points, this vector depends on the assignment being considered. For instance, suppose that we observe three peaks with integrals **u** = (5,3,2) in the ^{1}H NMR spectrum of ethylbenzene. According to prediction 5 peaks should be observed, with integrals 3, 2, 2, 2, and 1. However, more signals are predicted than observed, due to overlap. A possible (incorrect) assignment could link the methyl and para hydrogen to the peak integrating to 5, the ortho and meta hydrogens to the peak integrating to 3, and the methylene hydrogens to the peak integrating to 2. In this case **v**(f) = (3,4,3).
The quality of an assignment then is naturally measured by the similarity between these two vectors, that is:

$$m(f) = s(\mathbf{u}, \mathbf{v}(f)),$$

where $m(f)$ is the score of assignment f and $s()$ is a similarity function. The vectors **u** and **v**(f) can be partitioned into a family of data/prediction vectors, each one corresponding to a property, allowing different similarity measures to be tailored according to each property. In this way m(f) can be written as:

$$\begin{aligned} m(f) &= m_{Integration}(f) + m_{correlation}(f) + m_{Shift}(f) \\ &\quad + s_{correlation}(\mathbf{u}_{correlation}, \mathbf{v}_{correlation}(f)) + s_{Shift}(\mathbf{u}_{Shift}, \mathbf{v}_{Shift}(f)) \end{aligned}$$

Whenever knowing the particular assignment f under consideration is not important, we simplify **v**(f) to v, the dependence on f being understood implicitly. In the same way, the subindices of u and v may be omitted, since they are understood to be the same as the subindices of the function *s,* i.e*:* $s_{Integration}(\mathbf{u}_{Integration}, \mathbf{v}_{Integration}(f)) = s_{Integration}(\mathbf{u}, \mathbf{v})$. Predicted peak integrals v_i are given by the sum of the sizes of all CSS vertices assigned to the corresponding i-th peak. We expect each rounded peak integral to match the predicted value exactly. Similarity between the experimental and predicted ^{1}H integrals vectors is then given by the fraction of correct matches, that is:

$$s_{Integration}(u, v) = \frac{1}{n} \sum_{i=1}^{n} \delta(u_i, v_i)$$

where u_i is the i-th component of u, v_i is the i-th component of v, n is the number of peaks, and

$$\begin{aligned} \delta(u_i, v_i) &= 1 \; if \; u_i = v_i \\ \delta(u_i, v_i) &= 0 \; otherwise \end{aligned}$$

Note that $s_{Integration}$ takes values between 0 and 1, with 1 corresponding to a perfect match between experiment and prediction.

Spin correlations. Components of **u** are the normalized scores of picked peaks in 2D spectra (see Section 5.2). In a similar manner, components of **v** are 2D peak predictions, weighted according to our degree of confidence in that the peak should be observed in the corresponding 2D experiment. This is done so that our similarity function may reflect our degree of belief in the picked and predicted peaks involved. In the present work we used cosine similarity to compare the vectors of observed and predicted correlations:

$$s_{correlation}(u, v) = (u \bullet v)/(|u||v|)$$

,

Alternatively, rather than the full pattern of cross-peaks one may consider the column-wise cross-peak counts. In such case **u** and **v** are vectors of integers and we use the following similarity measure:

$$s_{correlationCounts}(u, v) = \sum_i exp(u_i - v_i)$$

the prediction of correlations relies on predicted coupling constants, on geometric or bond distance, or on a combination of both. Details on how cross-peak prediction was implemented in this work are found in section 3.

Chemical shift. Chemical shift values are probably the oldest and most utilized property in NMR structure elucidation. Here it is important to balance the reliability of data, the amount of structure information it provides, and the cost required for its acquisition. As a rule, 13C chemical shifts are less affected by experimental conditions than proton ones, but the much larger time investment required to acquire carbon spectra discourages their use. Proton spectra are much more popular for this reason, which encourages the design of automatic assignment methodologies that handle the higher variability of 1H chemical shifts.

When chemical shifts are taken into account, vector **v** consists of predicted chemical shift values v_i, that may be obtained through any state-of-the-art chemical shift predictor e.g. [**87**] for protons and [**17**, **101**, **105**, **124**] for ^{13}C. Now, as noted above, though a typical predictor yields a single chemical shift value, one should acknowledge that there is a significant degree of uncertainty related to the quality of the prediction and to variability in the experimental conditions when working with 1H-NMR data [**59**] and choose a similarity function that accounts for it. This can be done by associating the 'exact chemical shift prediction' with a 'diffuse prediction', which would be represented by a normal distribution around the predicted value, such as the one shown in Figure **5-3**A. Let us note this distribution by p_{vi}, where v_i is the i-th component of **v**, that is, the exact value predicted for the i-th chemical shift. We then define the similarity between predicted and observed ^{1}H chemical shifts as

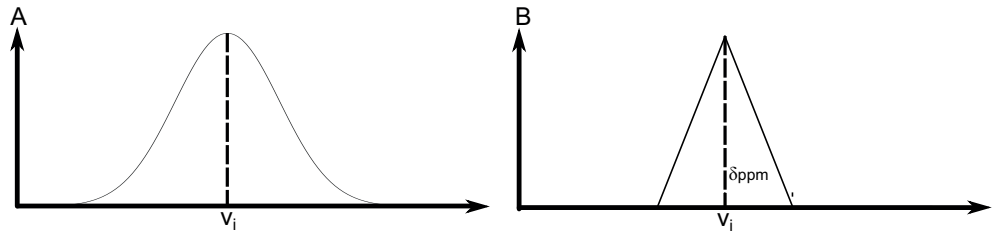


FIGURE **5-3**. **A**) A normal distribution around a predicted ^{1}H chemical shift value v_i represents a 'diffuse' prediction that accounts for the uncertainty in the result produced by the NMR prediction software and different experimental conditions [**59**]. The height of the curve is related to the expectation of observing the peak at the corresponding chemical shift coordinate. **B**) A gross approximation to **A**. In spoken terms, this corresponds to expecting the peak to be at v_i, and with increasing reservations expecting the peak to appear within a plus or minus δ ppm window. Increasing δ will relax the restriction imposed on the chemical shifts and lessen its importance. [**34**]

$$s_{Shift}(u, v) = \frac{1}{n} \sum_i p_{vi}(u_i)$$

where n is the number of peaks. The precise parameters of the p_{vi} distribution are unknown to us, however, so it needs to be approximated. The key parameter here is the width of the distribution, as it allows to tune the importance given to ^{1}H chemical shift in the process.

5.4. Solution search strategy

Efficient exploration of the search space is key to the success of the automatic assignment method. We use a branch-and-bound (BB) algorithm for this purpose (see Figure **5-4**):
Branch generation. On each level of the BB search tree an unassigned ^{1}H chemical shift is selected. Branches are generated by considering each possible combination of unassigned CSS vertices to it. In this way, branches of the search tree consist of sets of complete assignments that are identical up to the last assigned shift. Combinations that add extra vertices to a branch that has already exceeded the observed integral are not generated for the sake of efficiency.
Bounding and pruning. In a strict BB algorithm branches should be rejected only when their scores are bounded below those of alternative branches, which ensures a thorough exploration of the search space and guarantees that the absolute optima are found. However, in practical implementations it is often convenient or even necessary to establish a more relaxed criterion for pruning low-scoring branches before they fall out-of-bounds,

in order to gain additional efficiency. Such criterion should be chosen so that there is reasonable confidence that the pruned branch would not include an optimal (e.g. maximum score) solution, else the solution removed may very well be the correct one.
We associate each branch with an upper bound equal to the final assignment score that would be obtained if all yet unassigned shifts provided perfect matches, and reject branches whose bound falls below a chosen threshold. This strategy is inspired by the way experts perform structure verification by NMR assignment: they assume the structure to be correct until enough mismatches persuade them such is not the case. Equivalently, on the root of the BB tree we assume each and every assignment to be correct, and give them all the top score; then, for each descendant, we update this optimistic expectation according to a new prediction-experiment match. As we go down the BB tree and the mismatches grow larger the branch's bound gets lower, eventually it will fall too low and the branch will be discarded along with all its solutions.

5.5. Experimental section

To examine the performance of the proposed strategy, a data set consisting of experimental spectra of 74 molecules, in the 4 - 20 heavy atoms range, was used to validate the assignment methodology proposed. The data set comprises ^{1}H NMR, COSY, HSQC and HMBC spectra for each molecule. Spectra were recorded on a Bruker 400 MHz spectrometer equipped with a triple-axis gradient indirect probe and using standard Bruker pulse sequences. For proton spectra, 32 transients were acquired, added and stored into a 32k complex point vector, using a 16 kHz excitation pulse centered at 6.5 ppm and a sweep width of 14 ppm. For COSY spectra, 128 complex points were recorded in the indirect dimension, each consisting of a single transient stored into a vector of 2k complex point. The center and width of the observation window were adjusted to the necessity. High concentrations and deuterated solvents were used so that no routine is required for solvent peaks exclusion prior to peak picking.
Peak-picking was carried out by using the methodology described above. The peak-picking algorithm was run twice with different selection thresholds (2.25 and 2.75 standard deviations) to generate the initial peak set. Peaks were then clustered by the single-linkage method using euclidean metric. Clusters with a maximal intracluster distance of 24 Hz were acknowledge as meaningful signals. Afterwards, self-consistent peak selection was performed using the parameters $\varepsilon = 0.025$ ppm and $\varepsilon^{'} = 0.05$ ppm for alignment in proton and carbon dimensions. HSQC peaks were not used directly for scoring assignments, but as a mean to validate ^{1}J cross-peaks in HMBC spectra. Only peaks with final scores greater or equal to 0.6 were accepted.
Observed peak integrals were obtained by simple rounding to the nearest integer, and scores were obtained by simple comparison with the subindex of the CSS vertices (Figure **5-2**).

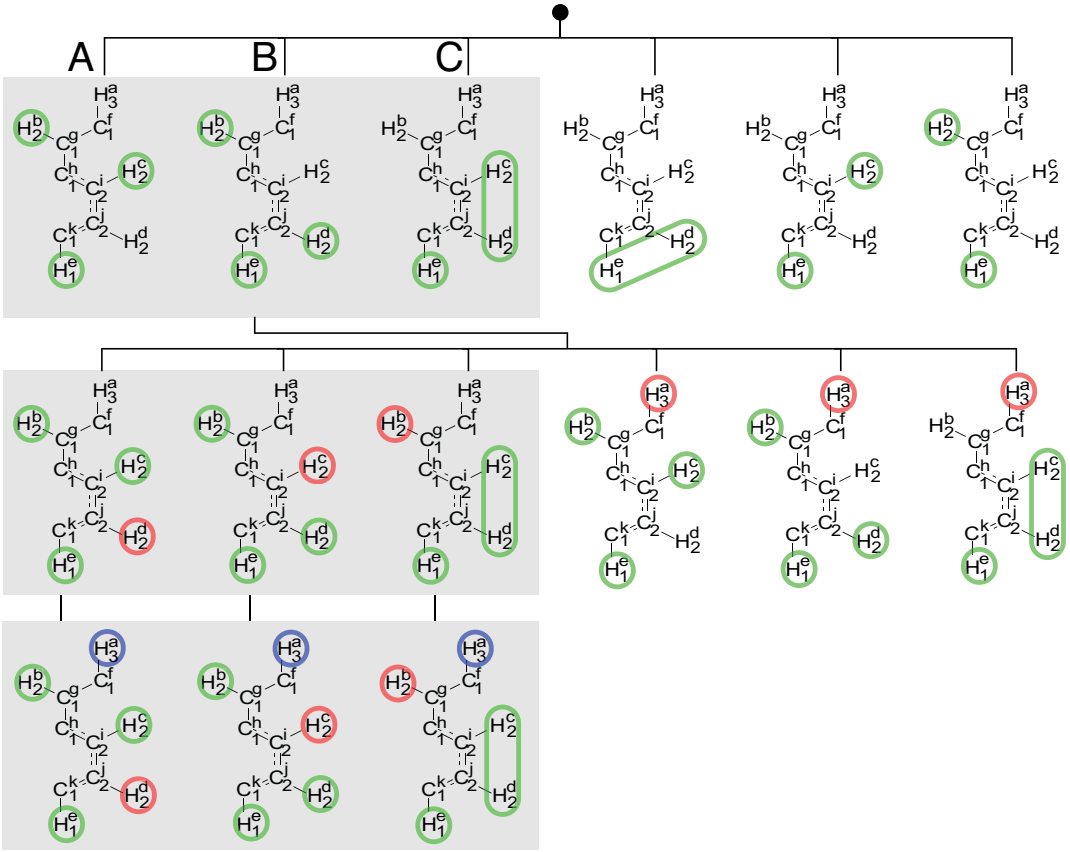


FIGURE **5-4**. Illustration of the assignment algorithm. Three ^{1}H peaks observed in the spectrum of ethylbenzene are to be assigned based on integration data. The branch rejection threshold is set to 0.9. On the first level, combinations of CSS vertices (circled in green) are assigned to a peak integrating to 5. Each combination generates a new branch, but only a few branches are depicted. Candidates in the light gray squares match with observed integrals, and are accepted as they can reach the maximum score $S_{integration} = 1 > 0.9$ provided the other two peak integrals fit as well. The other branches depicted on this level are rejected considering that at best they will get a final score $S_{integration} = (0+1+1)/3 = 2/3 < 0.9$, having already missed one peak integral. Note that each rejected branch implicitly contains several different candidate solutions (not depicted here), comprising all possible ways in which encircled vertices can be assigned to the remaining two peaks. All these solutions are simultaneously rejected without further testing, that is the basis of the efficiency of the BB method. On the second level branches are generated by assigning combinations of vertices, circled in red, to a new peak, this time with integral of 2. Once again, only a few branches are depicted and branches outside the blue square are rejected considering their best possible final score would be $2/3 < 0.9$. On the last level the remaining vertices are assigned to the last peak, which integrates to 3. All possibilities match with integral values and thus generate viable leaves, each one determining a suitable assignment of the spectrum with the maximum possible score of 1. **[34]**

Cross-peaks were predicted based primarily on bond distances. For COSY, HSQC and HMBC spectra, a coupled pair of atoms (a, b) is assumed to produce cross-peaks that were given a score of $v_i = 1$ (maximum) in any of the following cases:

- there is a path of length lesser or equal to 3 (COSY)
- there is a path of length 1 (HSQC)
- there is path of length up to 4 (HMBC)

For four-bond ^{1}H-^{1}H correlations, the scores of COSY cross-peaks were set to $v_{\mathrm{i}} = 0.75$ provided that the predicted coupling constants [**101**] fulfill the condition ^{4}J > 1 Hz. This gives more importance to cross-peaks that are indeed observed rather than to peaks that are expected to be absent.

^{1}H chemical shifts were predicted using the on-line tool Spinus [**27**]. The probability distribution around the predicted value depicted in Figure **5-3**A was approximated by the curve of Figure **5-3**B, with $\delta = 0.5$ ppm. This value was chosen to account for ^{1}H chemical shift fluctuations induced by variation in experimental conditions and for prediction accuracies of standard NMR predictors [**59**, **101**].

Automatic assignment was performed using the BB algorithm described in Section 2.3. We chose to use a very demanding branch rejection threshold of 0.9; that is, we only allow solutions with a 90% agreement with the prediction. This value is chosen based on experience: correct assignments have been found to fall above this threshold in real cases tested. Furthermore, this result agrees with our expectations, as correlation and integration data usually present very good agreement with theoretical predictions. However, if the method does not find any solution with the given threshold, then, it is reduced iteratively in steps of 0.05 until at least one solution is found or the rejection threshold is lower than 0.75. In such case, we say that the algorithm fails.

All the functions described in this work were written in Java and wrapped using a JavaScript interface to make it accessible from a JavaScript interpreter (Rhino) [**1**]. This allowed us to build a web-service to automatically assign the spectra of the test molecules and display the results into a web browser, in an interactive and dynamic fashion [**3**]. Visual inspection of the results is crucial, since it enables determining the causes of erroneous assignments.

5.6. Results

Including ^{1}H chemical shifts. Some results from this test are presented in Figure **5-5**; the full set of results can be visualized using the service provided in the supplementary materials. As it stands, this implementation permits to assign the complete set of 74 molecules in 520 ms. The algorithm failed only for 5% of the assignment problems, which is a very good rate. The correct assignment was found to be the highest score in 89% of the cases and within the best four scores in 95% of the cases. In most cases where the correct assignment did not have the highest score, both the correct assignment and those of higher ranking were identical regarding the predicted integration and spin correlations, and the difference between chemical shifts they assign differently lies within the width

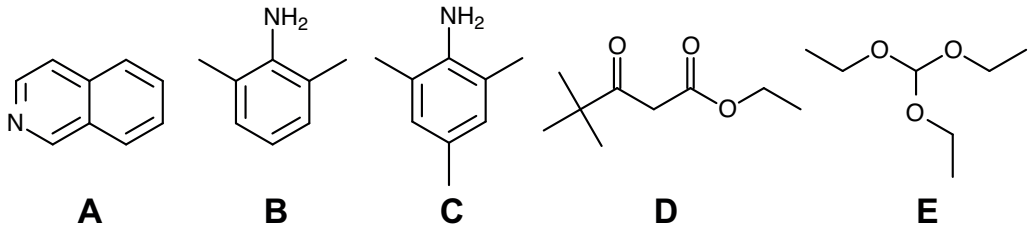


FIGURE **5-5**. Molecules that posed challenges to the automatic assigner in a test with 74 molecules. A) The correct assignment was found but it was not the highest ranking solution (4th rank for the molecule pictured). This happened to 11% of the molecules tested. B-C) No solution was found due to incorrect integral for the methyl groups. D) No solution was found because one of the CH_2 integrates to 1 E) No solution was found because of incorrect integral due to interferences with unwanted peaks. All other molecules, other than B - E, could be assigned and correcting the integrals, using a more complex routine would lead to the correct assignment in all cases. [**34**]

of the chosen range of incertitude (δ in Figure **5-3**B, in this case 0.5 ppm). One could then say that in this cases the results of the automatic assignment were correct within the expected experimental accuracy. The molecule of Figure **5-5**A was an exception to this rule, as the correct and the highest-scored assignments differ by permutation of two peaks in the aromatic region separated by over 1 ppm.

The four cases where the correct assignment could not be found at all, on the other hand, where due to integration mismatches and are shown in Figure **5-5**B-D. The cause of these unsuccessful runs was a failure to read the correct integration during peak-picking. For instance, the ^{1}H spectra of the molecules B and C present a peak exceeding the expected integration by 1 in the aliphatic region. In a similar manner, the ^{1}H spectrum of molecule D gives a methylene peak with an integral of 1 instead of 2, and the integration of molecule E is distorted due to interference with unwanted peaks arising from ethanol impurities.

The problem here is that, given the small number of signals in these spectra, a single integration error has a large impact in the overall score, causing the rejection of all solutions due to their inability to reproduce the observed integrals. These kind of disproportion in peak integrals is not unusual, e.g. it can be caused by inadequate recycling times or by impurities or signals from minor compounds overlapping with the observed peaks. Our integration routine is unable to work around such imperfections, thus leading to failure to assign the spectrum. A more complex strategy might be worked out in future works.

Without including ^{1}H chemical shifts. An interesting outcome of this test is that the correct solution can be found for most assignments without resorting to ^{1}H chemical shift data, i.e. relying only on spin correlation data and peak integration. As shown in Figure **5-6**, prescinding of ^{1}H chemical shift data causes an increase in the number of viable solutions presented by the algorithm, but has little effect in the rank of the correct solution. Indeed, even without factoring ^{1}H chemical shifts the correct solution was found

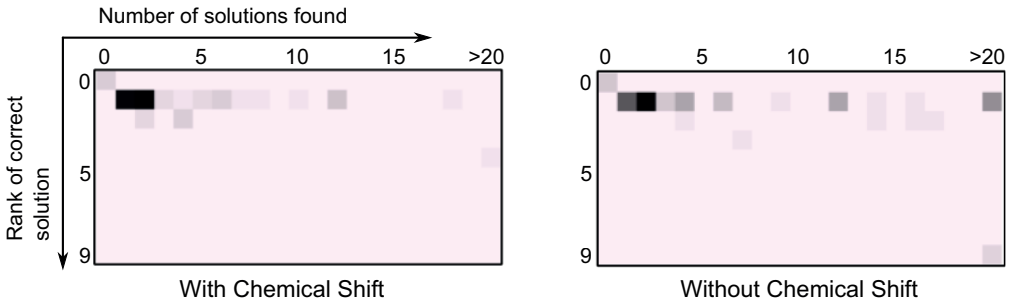


FIGURE **5-6**. Comparison of the results obtained using and not using ^{1}H chemical shift data for the assignment process. Grey squares mark the locations of each assignment problem tested in the Number of solutions vs. Rank of Correct solution plane. The cell on the upper left corresponds to the case where no solution is found. The darkness of the grey fill is proportional to the number of instances comprised. For example, it can be seen that in most cases the correct assignment was found as the highest scoring solution out of 1 or 2 possibilities, regardless of whether ^{1}H chemical shift data was included or not. [**34**]

within the four highest ranks in 89% of the test cases and the assignment failed for the same cases described above and due to identical reasons. Overall, ^{1}H chemical shifts help to settle the score between a narrowed set of candidates, but most importantly allow for faster pruning of the search tree. This amounts to a 50% reduction in assignment time, from 1040 ms to 520 ms for the 74 molecules. This is a significant gain in relative terms, but it becomes irrelevant when compared to the 160 s to complete the full analysis, including file parsing (74 s), peak-picking (55 s), etc.

5.7. Conclusions

The proposed methodology allows for robust peak-picking and accurate assignment of candidate molecules with no other input from users than NMR experiments. Peak integrals and coupling patterns allow for quick filtering of unsuitable solutions, avoiding reliance on fine chemical shift matching criteria. One of the advantages of this approach is that it avoids the need of a large database of correctly assigned spectra that serves as a basis for a powerful and highly accurate ^{1}H chemical shift prediction tool. Furthermore, reliance on accurate chemical shift prediction limits the reach of an auto-assignment tool by the scope of the sampling of chemical space achieved by such database, an issue that we also avoid by tuning down the role of chemical shift matching.

On the other hand, the method is very reliant on the quality of picked peak integrals. Failure to read the correct integrals was indeed responsible for all failed attempts to find the correct assignment in the tests run presented above. A very reliable integration routine thus becomes a key component of an assignment system based in the methodology proposed. We are currently implementing the approach developed by Cobas *et al* [**42**], after having found that it can successfully handle the problematic cases found in our test.

Regarding computation time, peak-picking turned to be the clear-cut bottleneck of the process, taking a hundred times longer than the assignment of picked shifts. Current efforts are directed towards development of more efficient code and exploring modifications to the peak picking strategy that may remove or at least soften this bottleneck.

CHAPTER 6

"*Ask Ernö*": A self-learning tool for assignment and prediction of Nuclear Magnetic Resonance spectra

On chapter 1 and 2, we described briefly the NMR prediction problem. On chapter 3 we suggested a methodology to evaluate the performance of a NMR chemical shift predictor without the necessity of an assigned data set. On chapters 4 and 5 we presented a strategy to automatically score all the possible assignments given a set of NMR spectra and its molecule. Although it was pointed out that such methodology does not relay on the prediction of chemical shifts, we suggested the possibility of its use as a way to improve the auto-assignment results. Clearly both the forward problem of predicting chemical shifts using a database of assigned data and the reverse problem of finding the correct assignment are related. In this chapter we present "Ask *Ernö*", a paired automatic assignment / chemical shift prediction system that can improve its performance while solving new assignment problems.
Follows the accepted version of the paper to the Journal of Chemoinformatics.

6.1. Background

The automation of chemical analysis by Nuclear Magnetic Resonance (NMR) spins around two problems: the *forward* problem of predicting the NMR spectra of a given molecule, and the *inverse* problem of elucidating the molecular structure that generates a given experimental spectrum. The forward problem is solved, in principle, by quantum mechanics: molecular structures determine a unique Hamiltonian from which all measurable NMR parameters can be computed. However, there are several considerations that make this solution impractical in most cases of interest:

- *Ab-initio* calculation of NMR parameters is too slow for contemporary needs. Predicting the full spectrum of small molecules of interest takes from minutes to hours, comparable or even longer than the time required to acquire the experimental data.
- An isolated resonating molecule is actually a very poor model for a real NMR spin-system. For instance, it usually ignores solvent effects, or the existence of multiple conformations of the sample compound under the experimental conditions. Accounting for these factors in *ab-initio* calculations pushes the computation complexity further, probably beyond the capabilities of most research groups. Thus, *ab-initio* NMR prediction turns out to be an ideal rather than a reality.

In practice, the forward problem of NMR prediction is handled by semi-empirical methods supported by previous knowledge of typical chemical shifts, not unlike the way human experts perform the task. Indeed, several commercial packages exist that perform NMR prediction based on models adjusted to large databases of observed chemical shifts [**4**, **5**, **6**, **27**, **28**, **73**]. Building such a database demands the assignment of observed chemical shifts to nuclei, a task that concerns the much more challenging inverse problem. Furthermore, prediction of expected chemical shifts from molecular structure plays a role in the assignment process as well, a necessary step to solve the inverse problem. The two problems are thus strongly related, a fact that poses an important limitation to the automation of NMR analysis. This reflects in existing computational tools for NMR elucidation and assignment: either they are not fully-automatic, requiring preliminary analysis by the user [**50**, **98**], or resort to automatic chemical shift prediction tools [**5**, **43**, **56**, **75**, **81**, **84**, **88**, **112**], which in turn rely on databases of correctly assigned spectra that must be built 'manually' by trained experts. Regardless of the approach, a significant amount of labour is involved that is certainly not devoid of human errors.
We can turn this issue around by noting that the strong relation between the forward and inverse problems means that progress in one direction improves the other [**14**]. Each time a spectrum is successfully assigned, nuclei are attributed a chemical shift that can enrich a database feeding a NMR predictor. In the opposite direction, more accurate and reliable chemical shift predictions provide better restrictions to discard non-viable assignments. The idea behind this article is to exploit this relation in order to create a fully automatic self-learning assignment and prediction system that progressively improves its capabilities as it solves more instances of assignment. Learning from scratch requires being able to automatically assign several spectra without resorting to chemical shift prediction, a task that we showed to be possible in a previous contribution [**34**]. We applied the experience and tools designed in that work to create *Ask Ernö*[1], a fully autonomous and self-improving assignment - prediction program. In this chapter we disclose its fundamentals along with an evaluation of its capabilities.

6.2. Methods

The concept behind *Ask Ernö* is summarized in Figure **6-1**. Automatic assignment of a nucleus in a molecule creates a link between the corresponding substructure and a chemical shift. This data can be transferred to a database for chemical shift predictions. As the database grows, the accuracy of the predictor improves. The improved predictor, in turn, provides better chemical shift constraints that can be used as input for subsequent assignments. *Ask Ernö* is trained by running repeated auto-assignment cycles on a given problem set, appending new assignments found to its prediction database, and using the improved predictions as input for the next assignment cycle.

[1] in reference to Ernö Pretsch and his classic book compiling thoroughly the necessary information for humans to assign spectra [**101**].

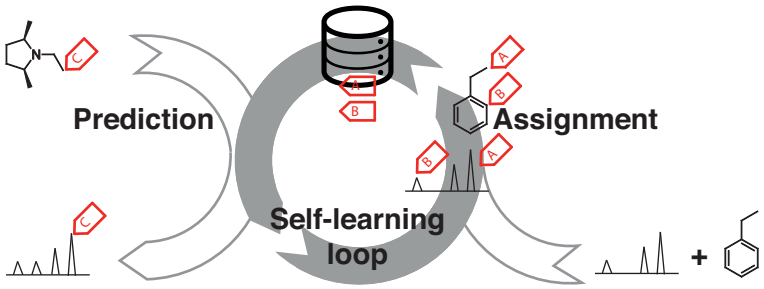


FIGURE **6-1**. The logic behind *Ask Ernö*. The automatic assignment of nuclei to their signals (right) produces entries to a database (mid) for chemical shift prediction (left). Predicted chemical shifts in turn provide further restrictions for assignment. *Ask Ernö* is trained by repeatedly looping on this assignment-prediction cycle.

Ask Ernö was implemented as a proof of concept rather than a full-fleshed assignment and prediction tool. For this reason, *Ask Ernö* was designed with small molecules in mind and tested using ^{1}H-NMR data only.

6.2.1. Chemical shift prediction. The database for chemical shift prediction consists of entries relating a nucleus and its molecular environment to a chemical shift value. Each entry consists of two terms:

- F: a molecular fragment around a proton, comprising the substructure spanned by all atoms up to n bonds from it, with $n \in 1, 2, ...$ We refer to this fragment as the n-sphere around the proton and to n as its radius or size (see Figure **6-2**). These fragments are stored as Hierarchically Ordered Spherical description of Environment (HOSE) codes **[31]** .
- δ: an observed chemical shift value for the proton.

Database registers are generated by automatic assignment of experimental spectra (see **Learning** 6.2.3 for details). Since the same fragment may be observed and assigned in different molecules, multiple entries may exist for the same F.
The methodology used for prediction is the HOSE-based methodology implemented in Modgraph NMRPredict for ^{13}C spectra **[6]**, adapted for ^{1}H spectra as follows: when asked to predict the chemical shift of a proton, the predictor spans the n-sphere of radius n_{max} around it, encodes this fragment and sends a query to the database. If the query is successful, the median $\bar{\delta}$ over all matches is returned as the predicted chemical shift, with an uncertainty εequal to their standard deviation. If the query finds no matching fragments, a new query is sent with the n-sphere of radius n_{max}-1 around the proton and so on, until a successful match is found or the radius of the sphere is below n_{min}. In the latter case no prediction is possible and the predictor returns failed status.

6.2.2. Assignment. We used an automatic assignment method previously described **[34]** that performs fully automatic peak-picking and assignment of chemical shifts based on peak integrals, correlations, and, when available, chemical shifts. The assignment routine

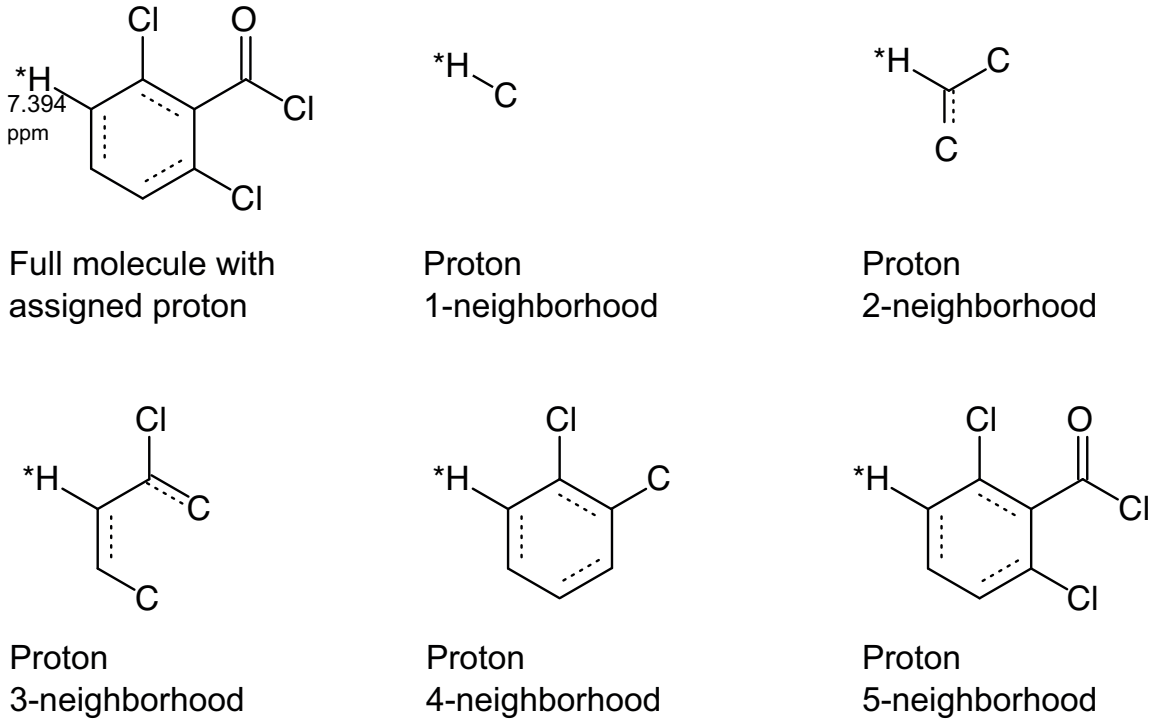


FIGURE **6-2**. n-spheres with radius 1-5 of a proton assigned to a chemical shift of 7.394 ppm. Dotted lines indicate aromatic bonds.

uses a symmetry-constrained branch and bound optimization that achieves a thorough exploration of the whole solution space. The result returned is a ranking of assignments, scored according to how good they fit the observed data. This automatic assigner has been shown to yield good results even if no chemical shift data is provided, which is of great importance for the present development.

Since here we only used ^1H-NMR spectra, assignments were performed exclusively on the basis of integration and chemical shift data. The auto-assigner was configured to seek for assignments that perfectly reproduced the observed integrals, and that matched the predicted and observed chemical shifts (when available) with an error no greater than 3 times the prediction's uncertainty at the current iteration. For this purpose, the uncertainty was estimated as the standard deviation of the sample of observed chemical shifts on which the prediction is based (see **Chemical shift prediction** 6.2.1 above), multiplied by a factor that depends on the size of the sample and the number of training iterations (see **Learning** 6.2.3 below) :

$$1 + n^{-I/2}$$

where n is the size of the sample and I is the index of the current iteration. This factor contributes significantly to controlling error propagation, accounting for the fact that the standard deviation is a poor estimator of the uncertainty for small n or I. Furthermore, for predictions based on less than two matches the allowed chemical shift error was set to the maximum of 20 ppm in order to accept any chemical shift, since no reasonable

Possible assignments found					
	Proton a	**Proton b**	**Proton c**	**Proton d**	**Proton e**
1	1.30	2.52	4.16	7.47	8.27
2	2.52	1.30	4.16	7.47	8.27
3	1.30	2.52	4.16	8.27	7.47
4	2.52	1.30	4.16	8.27	7.47

TABLE **6-1**. Results of the automatic assignment of a 5 proton molecule performed based on integrals exclusively. Despite the ambiguity introduced by the existence of 4 possible solutions, assignment of proton c to the peak at 4.16 ppm is repeated in all of them. This nucleus - chemical shift pair is thus deemed correct and selected to be learnt.

estimation of uncertainty exists at this point.

6.2.3. Learning. Data for training *Ask Ernö* consists of a set of molecules along with their corresponding ^1H-NMR spectra. The learning algorithm is based on a self-organizing map [**69**] and consists of a recursive learning cycle which is repeated until nothing new is learnt from the training dataset.
The first learning iteration starts by running the automatic assignment algorithm without taking chemical shifts into account. We refer to it as *iteration 0*. Redundancy (e.g. multiple occurrences of methyl groups) is expected so that several possible assignments may be found for any given molecule; this is particularly true when no correlations (2D) are available. Though the correct solution is unknown, it is often possible to find some nuclei - chemical shift pairs that are present in all computed assignments for a molecule and that can thus be assumed to be correct (see Table **6-1**). These pairs are learnt by creating database entries for the corresponding n-spheres, for $n = n_{min},...,n_{max}$ (see **Chemical shift prediction** 6.2.1 above). Completing this process on all molecules of the training set finishes iteration 0. The system then proceeds with iterations 1, 2, etc., in which newly learnt chemical shifts are used as additional restrictions for the assignment process. Note that new database entries are batch-generated, that is, newly learnt chemical shifts are only available as additional knowledge after a full iteration of the learning cycle. We found in preliminary tests that this approach slows down the learning process but yields better results than the "on-line" approach. Learning continues until two consecutive iterations yield no improvement.

6.3. Experimental

Ask Ernö was implemented in Java (automatic assigner), MySQL (prediction database) and JavaScript (chemical shift predictor, self-learning loop and integration of the system's components). The project is open source and available on GitHub [**2**], along with links to the data used for training and testing. A web service is available at http://visualizer.epfl.ch/tiny/r8cjIcLSVbEorw11WfGO to evaluate the system.

Data used for the evaluation consisted of 2639 molecules along with their experimental ^1H-NMR spectra. Examples of these spectra are included as supplementary materials. The dataset was assembled by random sampling from the Maybridge catalogue (2198 registers selected) and our own library (441 registers selected). Data was split between training set (2341 molecules) and test set (298 molecules). No assignment information was provided with the training set.
Spectra in the test set were manually assigned to determine the reference experimental chemical shift values for the calculation of prediction error. Not all protons in the set were assigned; most remarkably, labile protons were avoided considering that they are known to pose challenges to the components of *Ask Ernö* [**34**] and that we aimed to evaluate the potential and issues of the self-learning loop rather than those of its components. Overall, there were 2007 assigned protons that were used to benchmark *Ask Ernö's* predictions.
10 iterations of training were run, with $n_{max} = 4$ and $n_{min} = 2$. At the end of each iteration, chemical shifts for test molecules were predicted and compared with the observed values.

6.4. Results and discussion

Figure **6-3** shows the evolution of the correlation between predicted and correct chemical shifts as the system learns through 10 iterations. It can be seen that predictions oscillate from one iteration to the other as they converge towards the correct value (diagonal). At the last iteration, most predictions concentrate on the diagonal, though a few large errors persist.
To get a more detailed picture of *Ask Ernö's* performance and learning process we looked at three indicators: prediction error, prediction uncertainty, and the fraction of chemical shifts from the test set that could be predicted.

6.4.1. Prediction error. The overall prediction error is expected to decreases as the system iterates, and final errors to be the lowest possible. Figure**6-4** (top) shows the evolution of the average error across the iterations for $n_{min} = 2, 3, 4$. It is found that larger n_{min} values yield lower errors, but also that it improves less through each iteration (slower learning). Indeed, larger n-spheres give a better representation of the magnetic environment of the proton, producing more accurate predictions that can hardly be improved. For smaller fragments the distribution of observed chemical shifts is wider, so there is more room for improvement. Thus, as the system iterates and the fragment database grows, the average chemical shift of matching fragments moves closer to the true average of the full distribution, improving the average prediction error. For $n_{min} = 2$ this error decreased by 17% across 10 cycles, for a final value of 0.265 ppm.
Since the average error can be dominated by a few predictions with large errors, the cumulative error distributions were plotted (Figure **6-5**). It can be seen that larger n_{min} values yield more accurate predictions (< 0.2 ppm) and fewer predictions with high error (>1 ppm). Also, the number of accurate predictions grows faster with larger n_{min} (12% at $n_{min} = 4$ vs. 4% at $n_{min} = 2$), while the number of less accurate predictions reduces

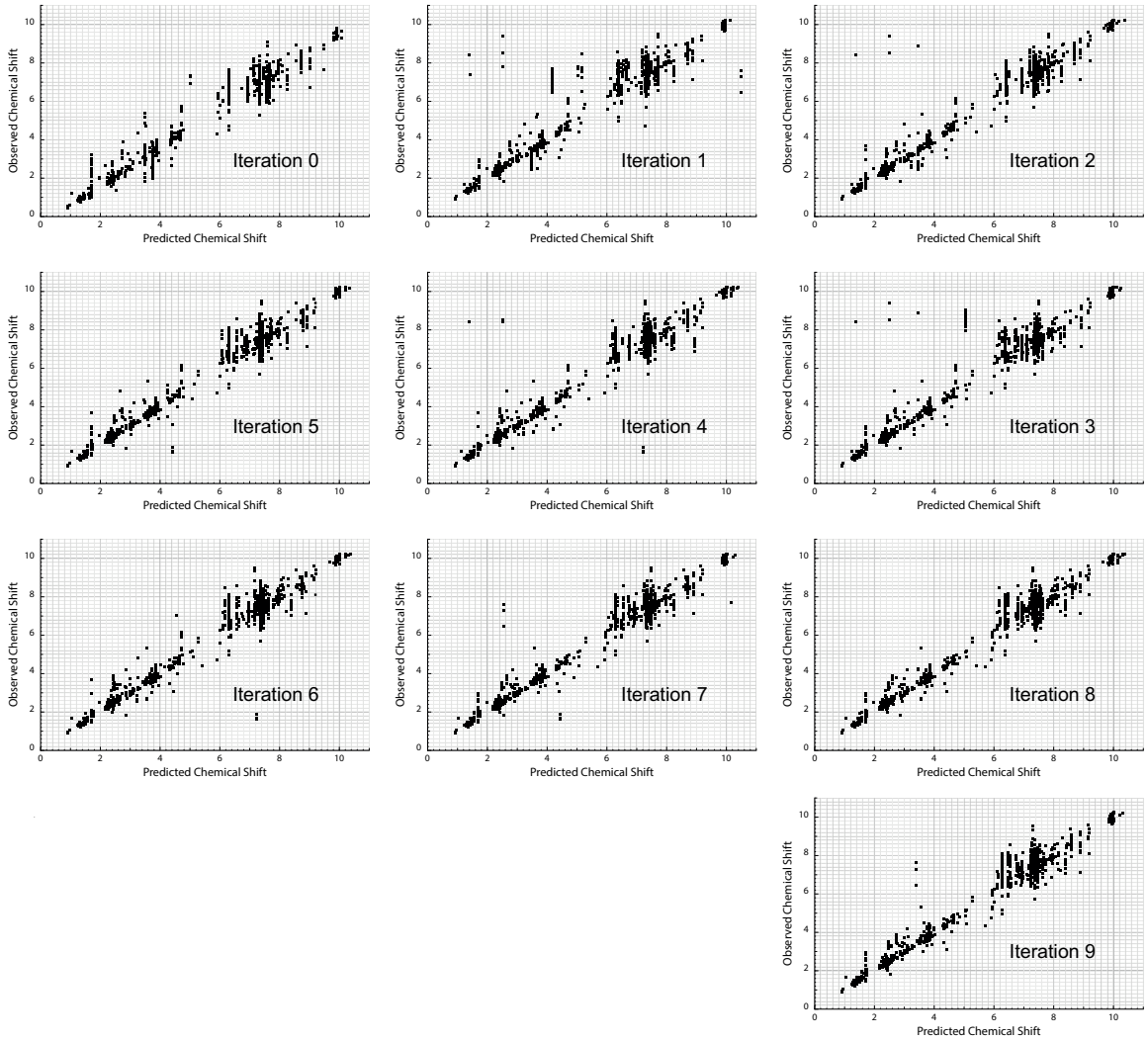


FIGURE **6-3**. Correlation between observed and predicted chemical shift values for the test molecules at each iteration of the training loop.

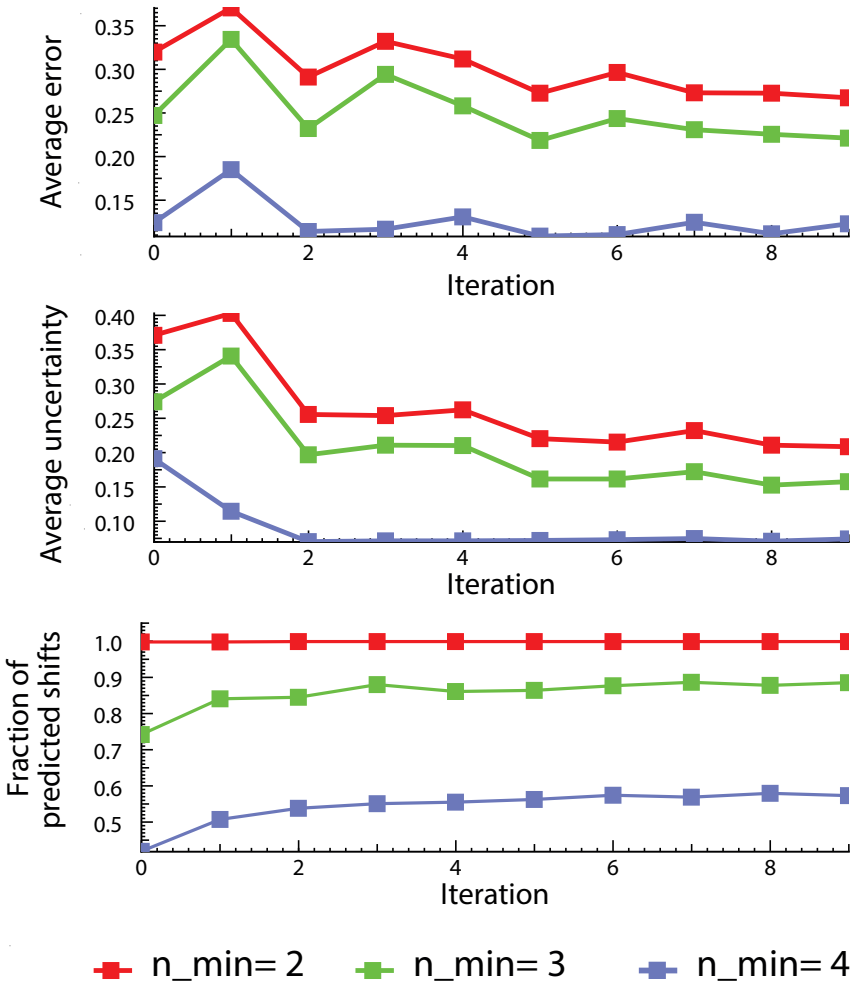


FIGURE **6-4**. Evolution during the training loop of prediction error (top), prediction uncertainty(middle) and fraction of predicted chemical shifts (bottom) for the test molecules.

more slowly (<1% for n_{min}=4 vs. 4% at $n_{min} = 2$). This is consistent with the observed behavior of the average error and again is explained by the naturally higher accuracy of predictions achieved with larger n-spheres. In the end, with $n_{min} = 2$, over 60% of the tested chemical shifts were predicted with less than 0.2 ppm error, and only 5% of them were found with error exceeding 1 ppm.

6.4.2. Prediction uncertainty. In *Ask Ernö*, the uncertainty of a prediction is associated with the standard error of the distribution of chemical shifts of matching fragments (see **Methods6.2, Chemical shift prediction**6.2.1). While the prediction error

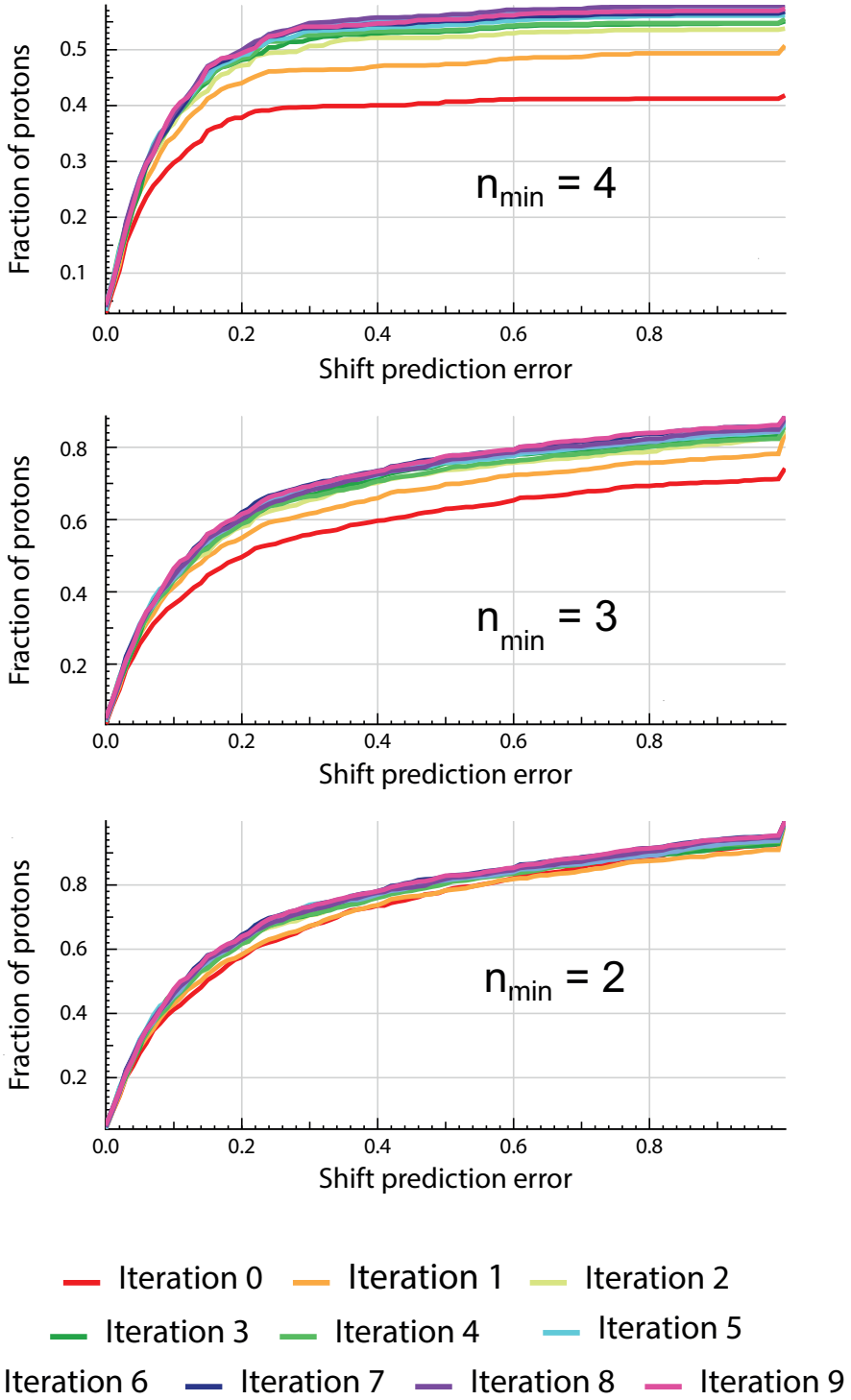


FIGURE **6-5**. Evolution of the cumulative error distributions during training. The fraction of predictions is given relative to the total number of protons in the testing set for which chemical shift can be predicted (2007 protons in total). To generate these curves, the set of chemical shift prediction errors was split into 100 bins of 0.01 ppm, plus a last bin containing predictions with an error equal or greater than 1 ppm. This last bin being larger explains the sudden increase observed at the end of the curves.

validates the results against an external reference (the correct chemical shifts), the uncertainty provides an internal validation. It is expected that as the system learns it gives predictions with lower uncertainty.
Figure **6-4** (mid) presents the evolution of this statistic through the training. It can be seen that the uncertainty quickly decreases, reaching a limit value. Both the rate and limit value are related to n_{min}: the smaller n_{min} the faster the uncertainty decreases and the lower it reaches (0.23 ppm for $n_{min} = 2$ and <0.1 ppm for $n_{min} = 4$).
Note that this limit is nothing but the standard deviation of the distribution of chemical shifts on the population of all possible n-spheres for the corresponding n_{min}. This allows for an interesting interpretation of the limit uncertainty as the theoretical best that *Ask Ernö* can achieve. Noting how the final average error in Figure 4 (top) is above the limit uncertainty in Figure **6-4** (mid), we conjecture that *Ask Ernö's* accuracy can still be improved by around 13% through further training with more data.

6.4.3. Amount of predicted chemical shifts. For a chemical shift to be predicted, it is necessary that a matching substructure is found in the database. The fraction of chemical shifts from the test set that can be predicted then constitutes a third descriptor of learning. Figure **6-4** (bottom) shows that though larger n-spheres provide better predictions, they only cover around half of the test problems (54% for $n_{min} = 4$ at the end of learning). Including predictions with $n_{min} = 3$ and $n_{min} = 2$ allows for a major leap in coverage, up to 85% and 99%. It is clear that no significant improvement can be gained by considering 1-spheres.
It is worth noting that the fraction of predictions with larger n-spheres increases by 13% during training. This is pivotal to *Ask Ernö's* performance: as its database grows, larger n-sphere matches becomes possible, which translates into better predictions.

6.4.4. Sources of error. *Ask Ernö* is particularly prone to errors when working with structures underrepresented in the training set. For instance, consider a prediction based on a small fragment that is present in numerous molecules of the training set. Since this small fragment is unable to properly account for all relevant interactions, it is associated with a broad range of chemical shifts and the uncertainty of the prediction is very high. Although such fragments are only used until a bigger match is found, no better match will ever be found for underrepresented fragments. In other words, *Ask Ernö* can't learn to correctly predict spin systems that are not properly represented in the training set. The situation just described is reflected in the large lines of vertically aligned points, observed in Figure **6-6**. Most mistakes are located along these vertical series of points, proving that this was the main source of error in the test.
Other recurring mistakes can also be related to underrepresented structures. For instance, the biggest errors for predictions based on 4-spheres (see Figure **6-6**, bottom) arise when the query returns a single matching fragment. In these cases the maximum uncertainty (20 ppm) given by the assigner to predictions based in less than 3 fragments allows for the propagation of an error that in principle should be rectified by new observations, but remains due to lack of the necessary data.

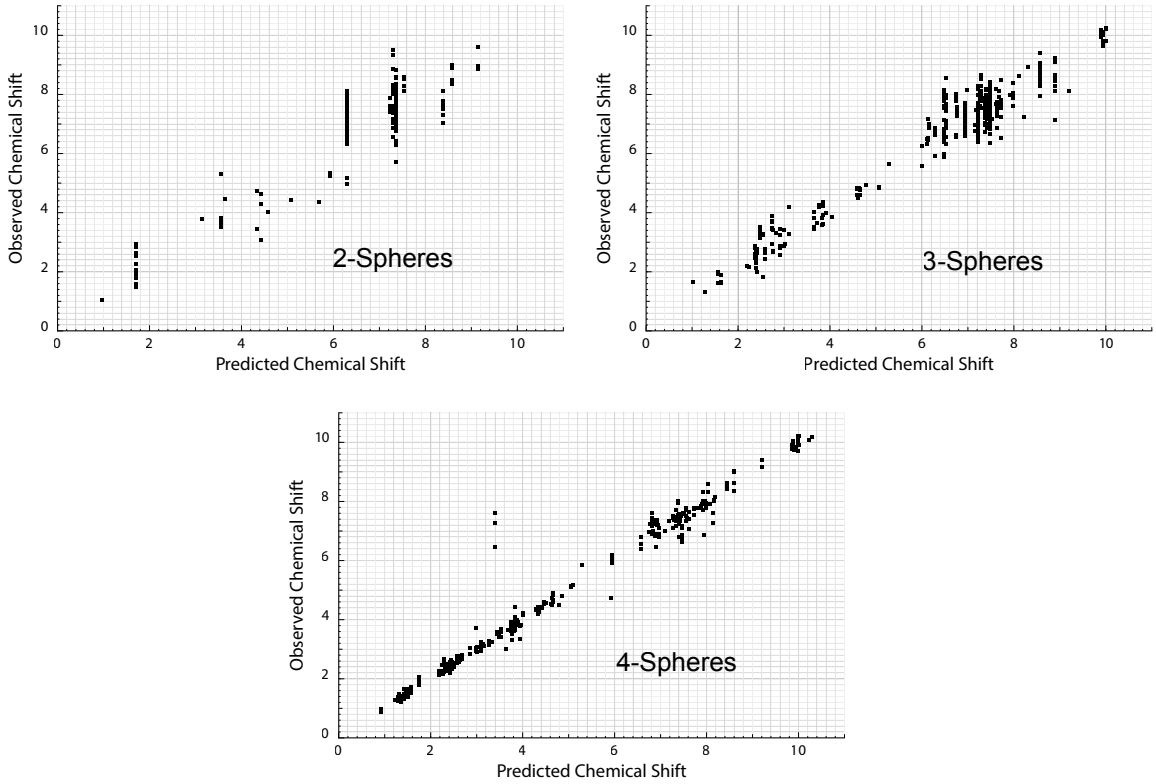


FIGURE **6-6**. Correlation between observed and predicted chemical shift values after learning for different sphere radius (iteration 9).

Detailed examples are given in the supplementary materials.

6.5. Conclusions

The reduction in error and uncertainty and the increase in the amount of predictions proves that *Ask Ernö* is indeed improving its prediction capabilities as it iterates on the assignment-prediction cycle. After 10 iterations using a set of 2341 assignment problems, *Ask Ernö* was able to predict the chemical shifts of protons in a set of 298 molecules with an average error of no more than 0.265 ppm. At least 60% of the chemical shifts were predicted with an error of less than 0.2 ppm. These are very promising results, especially for such a basic implementation of the concept.

It must be emphasized that *Ask Ernö* developed this capability fully autonomously: at no point it was fed with the fruits of the labour of human experts. The learning process of *Ask Ernö* is akin to that of a newcomer to the realm of NMR analysis, who is told the basic rules of assignment and through experience and induction develops his own NMR tables.

As expected, larger n-spheres provide better but fewer predictions. Furthermore, it was found that most errors occurred for underrepresented molecules when forcing highly uncertain predictions based on smaller fragments. For these reasons, it is expected that

with more data the database could grow to a point where any query would match a large n-sphere. Thus, though the system currently tops at an average error of 0.265 ppm,the limit of $<$0.1 ppm error could be reached with enough data. Further improvements to this limit would require taking into account other experimental parameters such as solvent, concentration and temperature of acquisition, as major source of experimental errors.
Based on the results presented here, we expect to develop *Ask Ernö* into a state-of-the-art tool for automatic NMR analysis in the near future. Current efforts are focused in reforming the estimator of uncertainty in order to enhance the system's capability to rectify its mistakes as it iterates. Correlation data from 2D experiments should also lead to significant improvement, when available.

CHAPTER 7

Conclusions and future work

7.1. Conclusion

Although automatic assignment and structure elucidation have long been foreseen, they remain challenging problems not fully addressed yet. This work shows the importance of considering both the forward and reverse problem as an iteration loop, not unlike the way human experts propose a structure and then validate this candidate by assigning each atom until a coherent solution is found that explains the observations. In turn the candidate is proposed on the basis of prior knowledge of similar molecules or similar molecular fragments. Recognizing this is certainly part of the novelty of the work presented here.
Assignment of atoms to signals is still the gold standard for structure confirmation by NMR spectroscopy. All around the world, large investments in trained labour and research time are devoted to this task. This is already a big enough reason to make of NMR assignment a top target for automatization; but on top of it, its relationship to the problems of chemical shift prediction and automatic elucidation turn it into a cornerstone for the automatization of NMR analysis. Yet, after decades of efforts, the ideal of an autonomous computer algorithm capable of assigning NMR spectra without human intervention has escaped the reach of the discipline of chemoinformatics.
We have shown that it is possible to fully automatize the NMR assignment process. In this case, "fully" makes reference not only to the assignment process itself, but to all other tasks involved in the generation of an assigned spectrum; most remarkably, peak picking of meaningful signals and prediction of expected chemical shifts. We also have shown, arguably for the first time, that it is possible to carry all these tasks in a completely unsupervised manner: in principle, no 'manual' analysis is necessary to reproduce all the results presented in this work -nothing needs to be provided by the user other than the spectra and structures to be analyzed. These are really good news, taking into account that scientists are not very actively publishing their spectra assignments on facebook or tweeter, and that high quality datasets are treasures zealously protected by the pharmaceutical industry. Opposing this trend, all our tools were implemented and released as open source contributions by the *cheminfo* project https://github.com/cheminfo. We also have useful demos online, currently used for teaching purposes http://www.cheminfo.org/.

7.1.1. Real world applications and perspectives. It would be difficult to close this chapter, and those 4 years of work, without pointing out some aspects about this research and its outcome. Parts of all of them are just opinions:

- Although we developed an expert system for NMR analysis, the system is able to gain experience and improve its performance with time. This feature has been made thanks to the fact that NMR assignment and NMR prediction are correlated tasks in both ways. Improvements in the capacity to predict NMR chemical shifts, translates into improvements of the automatic assignment performance. Less ambiguities in the assignment of molecules translates to better chemical shift prediction through the enlargement of the assignments database. Nevertheless, there is a bottleneck in the loop: Despite of our aim to make the automatic assignment a "robust to noise" algorithm, the noise at the input is still impacting negatively the performance of the method. Such input noise arises from the peak picking algorithm. Those are bad news because at the end, our work seems to have the same weakness that the other state of the art methods. Nevertheless, this means that improvements on that field, will translate also in improvements in the automatic assignment and the chemical shift prediction. We also can start to think in including this task directly in the loop. I mean, we can make the peak picking not the entry step to the assignment process, but an initial guess that can be modified by the assignment-prediction loop itself.
- Although one of our main aims in this work was to develop methods that do not depend on databases of assigned NMR data, we recognize the great value of such information for the NMR automatic analysis. By now, our only complain about this, is the lack of publicly available data bases of high quality assigned NMR data. Therefore, one of our current challenges is to build a free, open-source and open-access repository of NMR data where they can be automatically analyzed and validated, just by the submission of raw data. *Ask Ernö* is a great step towards this end and a first prototype of such a repository may be found at wiley.cheminfo.org and was developed with the agreement of Wiley to be incorporated in the publication process of *Magnetic Resonance in Chemistry*.
- A related problem to the creation of such repository is to ensure the perennity of the data: How to guarantee that the information can be accessed and interpreted far away in the future independently of the changes in the current technologies. This problem is in part solved by the use of open-source projects, shared through GitHub and open formats duly described, such as JSON, JCAMP, etc. That ensures you only need to fire up your modern browsers to start working with the data.
- Since the Java-Applets started to die, because of their many security issues, we decided to migrate most of our applications from java to javascript. Although this milestone has not been fully reached yet, in the near future we will be able to make 'science in the browser': Prediction, simulation, peak-picking and assignment, indeed part of this future is already at hand . Everything working for you directly in the browser. No installation needed, no fees. For now, we have part of the tools working in the browser, and part on a server.

- In spite of the speed, We choose JavaScript for our applications because of it great versatility and because it is natively supported by all the modern browsers, allowing to write applications that provide interaction between the HTML web pages and the Cloud, and more recently also in the server side thanks to NodeJS. The community of JavaScript developers is one of the highest and most active communities nowadays, and it is driving javascript to become more than the simple HTML control layer. On the other hand, JavaScript is far away of being Python for *science programing*. For example JavaScript still has speed problems and there is not direct access to the hardware resources, which makes the use of GPU a very complex and tricky task.
- A missing aspect yet to this project is the visualization of the information. Although it has not been the scope of this work, parallel to its development, Norman Pellet, Luc Patiny, Michaël Zasso, and many other contributors developed the visualizer https://github.com/NPellet/visualizer **[3]** responding to a real necessity to be able to visualize the results during the development of tools. Another open-source project resulted with many original features that currently allows to create web applications readily to tackle chemical problems). The visualizer has been used extensively during this work, not only as a way to understand, and debug our processes, but also to share our results with others.
- All the research we did in this thesis, was implemented and released as open-source contributions, all available at https://github.com/cheminfo. We also released and maintain the site http://www.nmrdb.org/ where we aim to share useful tools for teachers and research fellows (See Figure **7-1**). The former site receives more than 1900 visits per day from different addresses, while the latter is ranked first in google for several keywords, such as "nmr prediction" and "nmr simulation" for example. This is a different, yet valid, way of measuring the impact of your research, an alternative to the scores in the paper citations scientists are proud of.

- Figure **7-2** shows a snapshot of another web tool, this time for 1D-2D NMR automatic assignment. In this site you can drag and drop or draw the molecule in the JSME editor and drag-and-drop your 1D and 2D NMR spectra. The web tool will perform the peak-picking directly on the browser, and can then be edited manually if necessary. Then, the tool may be asked to get all the possible assignments by clicking the button "Auto-assign". It will send the molecule and the peak-picking information to the server. The result of the process is a table that lists all the feasible solutions sorted according to the score described in chapter 5. This system allows to validate the correct assignment , and will then store all the related information and assignment to a database.

- In spite of recent developments in machine learning sciences, expert systems can achieve better results in most of the fields, whenever you can understand the true nature of the problem. In chemoinformatics, for example, there is, usually, a

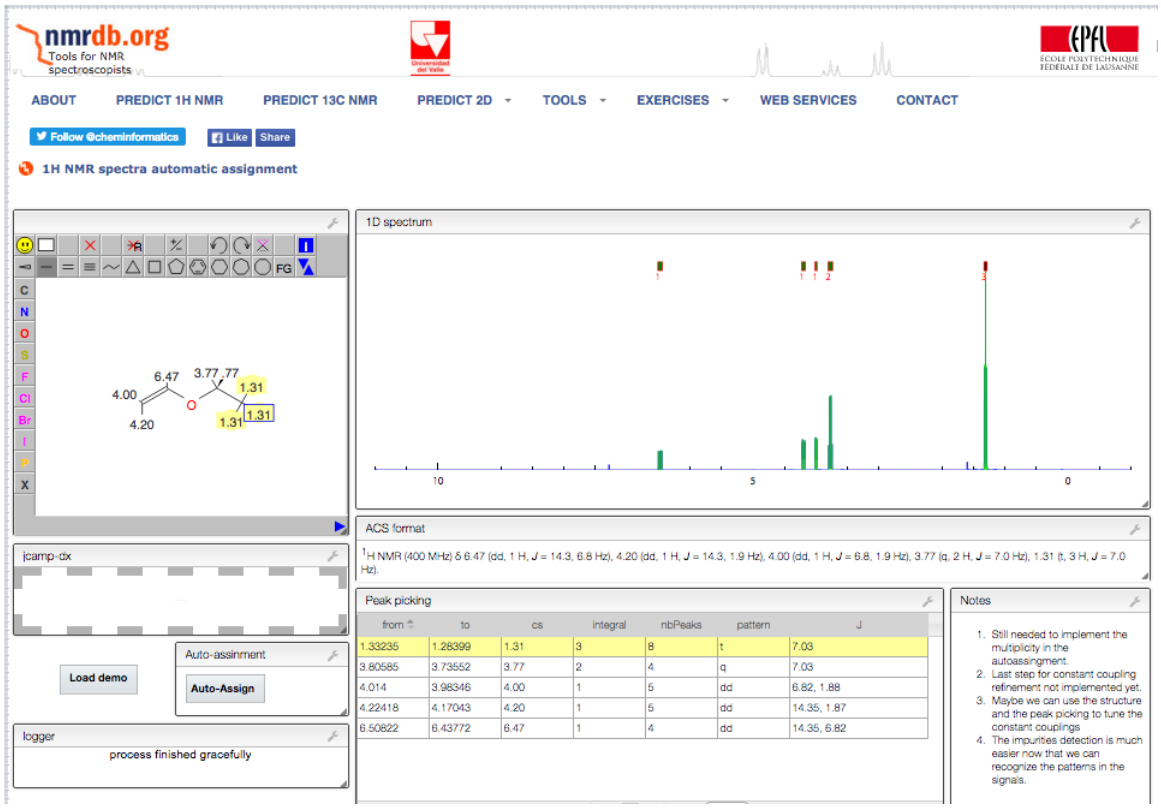


FIGURE **7-1**. ^1H-NMR automatic assignment web tool at http://www.nmrdb.org/assigner_1h/index.shtml?v=v2.23.0

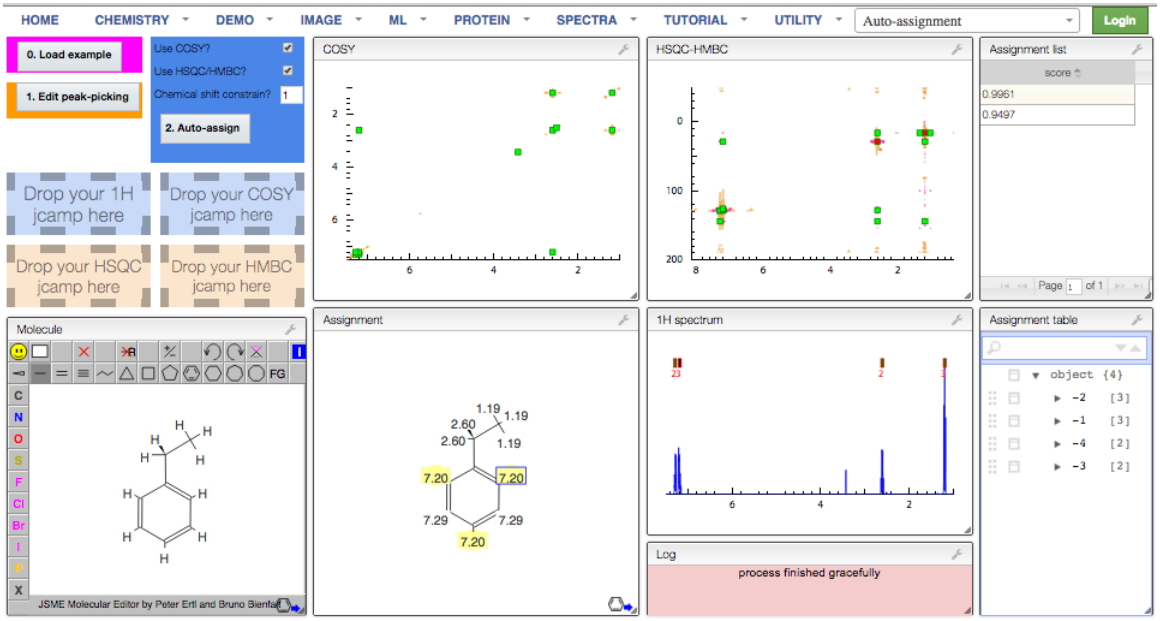


FIGURE **7-2**. 1D and 2D automatic assignment at http://www.cheminfo.org/Spectra/NMR/Tools/Auto-assignment/index.html

good set of fundamentals (physical and chemical laws) that explains the observed phenomenon. This situation contrasts with the problems associated e.g. with image or natural language, where the nature of the problem is not supported by formal models but by our belief and interpretations of it. This has been the reason why this work is not a compendium of several machine learning algorithms applied to NMR data.

7.2. Future work

In the near future we plan to build a public NMR repository which will contain *auto-assigned* spectra. To do this we plan to take advantage of the "learn by solving" nature of the automatic assignment tool developed: we offer services for automatic or computer-assisted validation of structures and NMR chemical shift prediction e.g. to organic chemists preparing or evaluating publications; the paycheck for the use of these services is the contribution of data by the users to enlarge the database supporting the system. It's a win-win situation for all parts involved: for the users (researchers, publishers, reviewers) who get a useful tool freely available, and for the developers (us) who get a growing database to work with. In the meanwhile, the tool itself becomes increasingly accurate and reliable.

It is clear for us, that the inclusion of other kind of information is another way to improve the system. By now we are considering the inclusion of coupling information from ^1H-NMR spectra, but also the more accurate you can obtain through the analysis of J-Resolve NMR spectra.

It has been our dream to work on an elucidator system, and this work already contributes one of the two main building blocks for that project: the module of validation of candidate structures. Parallel to the development of this work, we have undertaken preliminary advances in the creation of the second major component, the structure generation module, which makes us think that we are nearing the point when we finally can propose something in that field.

Bibliography

[1] https://developer.mozilla.org/en-US/docs/Mozilla/Projects/Rhino.

[2] https://github.com/cheminfo/autolearning.

[3] https://github.com/NPellet/visualizer.

[4] http://www2.ccc.uni-erlangen.de/services/spinus/.

[5] http://www.cambridgesoft.com/ensemble_for_chemistry/chemdraw.

[6] http://www.modgraph.co.uk/product_nmr.htm.

[7] http://www.maybridge.com/, 2011.

[8] PERCH-ACA, http://new.perchsolutions.com/. Online, 04 2016.

[9] A Abbas, X Guo, B Y Jing, and X Gao. An automated framework for NMR resonance assignment through simultaneous slice picking and spin system forming. *Journal of Biomolecular NMR*, 59:75–86, 2014.

[10] A Abbas, X B Kong, Z Liu, and X Gao. Automatic Peak Selection by a Benjamini-Hochberg-Based Algorithm. PLoS ONE. *PLos ONE*, 8(1):e53112, 2013.

[11] Raymond J Abraham and Mehdi Mobli. The prediction of 1h nmr chemical shifts in organic compounds. *Spectroscopy Europe*, 16(4):16–22, 2004.

[12] Raymond J Abraham and Mehdi Mobli. A practical approach to 1h nmr calculation and prediction. *Modelling H NMR Spectra of Organic Compounds: Theory, Applications and NMR Prediction Software*, pages 349–368, 2008.

[13] João Aires-de Sousa, Markus C Hemmer, and Johann Gasteiger. Prediction of 1h nmr chemical shifts using neural networks. *Analytical chemistry*, 74(1):80–90, 2002.

[14] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[15] B Alipanahi, X Gao, E Karakoc, L Donaldson, and M Li. PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics*, 25:i268–75, 2009.

[16] Andrej Likar and Tim Vidmar. A peak-search method based on spectrum convolution. *Journal of Physics D: Applied Physics*, 36:1903–1909, 2003.

[17] Andrés M Castillo, Andrés Bernal, Luc Patiny, and Julien Wist. A new method for the comparison of 1H NMR predictors based on tree-similarity of spectra. *Journal of Cheminformatics*, 6, 2014.

[18] C Antz, K P Neidig, and H R Kalbitzer. A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. *Journal of biomolecular NMR*, 5(3):287–96, April 1995.

[19] S. Anzali, G. Barnickel, B. Cezanne, M. Krug, D. Filimonov, and V. Poroikov. Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *Journal of Medicinal Chemistry*, 44(15):2432–7, July 2001.

[20] H. S. Atreya, S. C. Sahu, K. V. R. Chary, and Girjesh Govil. A tracked approach for automated NMR assignments in proteins (TATAPRO). *journal of Biomolecular NMR*, 17:125–136, 2006.

[21] Chris Bailey-Kellogg, Alik Widge, John J Kelley, Marcelo J Berardi, John H Bushweller, and Bruce Randall Donald. The NOESY Jigsaw: Automated Protein Secondary Structure and Main-Chain Assignment from Sparse, Unassigned NMR Data. *Journal of Computational Biology*, 7(3/4):537–558, 2000.

[22] D Balamurugan, Weitao Yanga, and David N Beratan. Exploring chemical space with discrete, gradient, and hybrid optimization methods. *Journal of Chemical Physics*, 129(174105), 2008.

[23] Michael Barfield and William B Smithf. Internal H-C-C Angle Dependence of Vicinal ' H-lH Coupling Constants. *journal of a*, (23):1574–1581, 1992.

[24] Christian Bartels, Peter Güntert, Martin Billeter, and Kurt Wüthrich. GARANT-A General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra. *Journal of Computational Chemistry*, 18(1):139–149, 1996.

[25] A Bernal, A Castillo, F González, L Patiny, and J Wist. Improving the efficiency of branch-and-bound complete-search NMR assignment using the symmetry of molecules and spectra. *Journal of Chemical Physics*, 142:074103, 2015.

[26] Reimond Bernstein, Christian Cieslar, Alfred Ross, Hartmut Oschkinat, Jens Freund, and Tad A. Holak. Computer-assisted assignment of multidimensional NMR spectra of proteins: Application to 3D NOESY-HMQC and TOCSY-HMQC spectra. *Journal of Biomolecular NMR*, 3(114):245–251, March 1993.

[27] Yuri Binev and João Aires-de Sousa. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *Journal of chemical information and computer sciences*, 44(3):940–5, 2004.

[28] Yuri Binev, Maria MB Marques, and João Aires-de Sousa. Prediction of 1h nmr coupling constants with associative neural networks trained for chemical shifts. *Journal of chemical information and modeling*, 47(6):2089–2097, 2007.

[29] Lorant Bodis, Alfred Ross, and Ernö Pretsch. A novel spectra similarity measure. *Chemometrics and intelligent laboratory systems*, 85(1):1–8, 2007.

[30] E. Breitmaier. Structure Elucidation by NMR in Organic Chemistry. A Practical Guide. *Concepts in Magnetic Resonance*, 6:237–238, 1994.

[31] W Bremser. HOSE -a novel substructure code. *Analytica Chimica Actac*, 103:355–365, 1978.

[32] Nicolas E. G. Buchler, Erik R. P. Zuiderweg, Hong Wang, and Richard A. Goldstein. Protein Heteronuclear NMR Assignments Using Mean-Field Simulated Annealing. *Journal of Magnetic Resonance*, 125(125):34–42, 1997.

[33] R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi. Applications of artificial intelligence for chemical inference. XVII. Approach to computer-assisted elucidation of molecular structure. *Journal of the American Chemical Society*, 97(20):5755–5762, 1975.

[34] A M Castillo, A Bernal, L Patiny, and J Wist. Fully automatic assignment of small molecules' NMR spectra without relying on chemical shift predictions. *Magnetic Resonance in Chemistry*, 53:603–611, 2015.

[35] Andrés M Castillo, Luc Patiny, and Julien Wist. Fast and accurate algorithm for the simulation of nmr spectra of large spin systems. *Journal of Magnetic Resonance*, 209(2):123–130, 2011.

[36] Andrés Mauricio Castillo, Lalita Uribe, Luc Patiny, and Julien Wist. Fast and shift-insensitive similarity comparisons of nmr using a tree-representation of spectra. *Chemometrics and Intelligent Laboratory Systems*, 127:1–6, 2013.

[37] Wei Chen, Jing Huang, and Michael K Gilson. Identification of symmetries in molecules and complexes. *Journal of chemical information and computer sciences*, 44(4):1301–13, 2004.

[38] Zhi-Zhong Chen, Tao Jiang, Guohui Lin, Jianjun Wen, Dong Xu, Jinbo Xu, and Ying Xu. Approximation algorithms for NMR spectral peak assignment. *Theoretical Computer Science*, 299:211–229, 2003.

[39] Y Cheng, X Gao, and F Liang. Bayesian peak picking for NMR spectra. *Genomics, Proteomics and Bioinformatics*, 12:39–7, 2014.

[40] Bradley D Christie and Morton E Munk. The role of two-dimensional nuclear magnetic resonance

spectroscopy in computer-enhanced structure elucidation. *Journal of the American Chemical Society*, 113(10):3750–3757, 1991.

[41] Jens Clausen. Branch and Bound Algorithms - Principles and Examples . *Department of Computer Science, University of Copenhagen*, pages 1–30, 1999.

[42] C Cobas, F Seoane, S Domínguez, and S Skyora. A new approach to improving automated analysis of proton NMR spectra through Global Spectral Deconvolution (GSD). *Spectroscopy Europe*, 23(1):25–30, 2011.

[43] C Cobas, F Seoane andE Vaz, M A Bernstein, S Domínguez, M Pérez, and S Sýkora. Automatic assignment of 1H-NMR spectra of small molecules. *Magnetic Resonance in Chemistry*, 51:649–654, 2013.

[44] Brian E Coggins and Pei Zhou. PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR*, 26(2):93–111, June 2003.

[45] Nick Craswell. Mean reciprocal rank. In *Encyclopedia of Database Systems*, pages 1703–1703. Springer, 2009.

[46] W Diettrich, C H Rüdel, and M Neumann. Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra. *Journal of Magnetic Resonance*, 91:1–11, 1991.

[47] C. Eccles, P. Güntert, M. Billeter, and K. Wüthrich. Efficient analysis of protein 2D NMR spectra using the software package EASY. *Journal of Biomolecular NMR*, 1(2):111–30, 1991.

[48] M. E. Elyashberg, K. a. Blinov, S. G. Molodtsov, and E. D. Smurnyi. New computer-assisted methods for the elucidation of molecular structure from 2-D spectra. *Journal of Analytical Chemistry*, 63(1):13–20, January 2011.

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

[50] ME Elyashberg, AJ Williams, and GE Martin. Computer-assisted structure verification and elucidation tools in nmr-based structure elucidation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 53(1):1–104, 2008.

[51] F Fiorito, T Herrmann, F F Damberger, and K Wüthrich. Automated amino acid side-chain NMR assignment of proteins using 13C-and 15N-resolved 3D [1H, 1H]-NOESY. *Journal of Biomolecular NMR*, 42:23–33, 2008.

[52] H. Friebolin. *Basic One-and Two-Dimensional NMR Spectroscopy*. Wiley-VCH, 1-108, 4 edition, 2005.

[53] X Gao. Mathematical approaches to the NMR peak-picking problem. *Journal of applied computational mathematics*, 1:1–12, 2012.

[54] D S Garrett, R Powers, A M Gronenborn, and G M Clore. A Common Sense Approach to Peak Picking in Two-, Three-, and Four-Dimensional Spectra Using Automatic Computer Analysis of Contour Diagrams. *Journal of Magnetic Resonance*, 95:214–220, 1991.

[55] Sergey S Golotvin, Eugene Vodopianov, Brent A Lefebvre, Antony J Williams, and Timothy D Spitzer. Automated structure verification based on 1h nmr prediction. *Magnetic Resonance in Chemistry*, 44(5):524–538, 2006.

[56] Sergey S Golotvin, Eugene Vodopianov, Rostislav Pol, Brent A Lefebvre, Antony J Williams, Randy D Rutkowske, and Timothy D Spitzer. Automated structure verification based on a combination of 1D 1 H NMR and 2D 1 H-13 C HSQC spectra. *Magnetic resonance in chemistry*, 45:803–813, 2007.

[57] Lee Griffiths. Towards the automatic analysis of 1H NMR spectra: Part 2. Accurate integrals and stoichiometry. *Magnetic Resonance in Chemistry*, 39(4):194–202, April 2001.

[58] Lee Griffiths and JD Bright. Towards the automatic analysis of 1H NMR spectra: Part 3. Confirmation of postulated chemical structure. *Magnetic Resonance in Chemistry*, 40:623–634, 2002.

[59] M Grzonka and A N Davies. Empirical Investigation on the Reproducibility of 13C NMR Shift Values. *Journal of Chemical Information and Computer Science*, 38:1096–1101, 1998.

[60] Peter Güntert, Michael Salzmann, Daniel Braun, Kurt Wüthrich, and Ch Zürich. Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *Journal of Biomolecular NMR*, 18:129–137, 2000.

[61] Brian J. Hare and James H. Prestegard. Application of neural networks to automated assignment of NMR spectra of proteins. *Journal of Biomolecular NMR*, 4:35–46, 1994.

[62] T Kevin Hitchens, Jonathan A Lukin, Yiping Zhan, Scott A McCallum, and Gordon S Rule. MONTE : An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *Journal of Biomolecular NMR*, 25:1–9, 2003.

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

[64] Bruce Randall Donald Jianyang Zeng, Pei Zhou. chapter A Markov Random Field Framework for Protein Side-Chain Resonance Assignment, pages 550–570. Springer Link, 2010.

[65] Johnny Chang and Robert E Wyatt. Preselecting paths for multiphoton dynamics using artificial intelligence. *Journal of Chemical Physics*, 85(1826), 1986.

[66] B. A. Johnson and R. A. Blevins. NMR View: A computer program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR*, 4(5):603–614, 1994.

[67] Young-Sang Jung and Markus Zweckstetter. Mars - robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR*, 30(1):11–23, September 2004.

[68] Kimito Funatsu, Carlos del Carpio, and Shin-ichi Sasaki. Automated structure elucidation system -CHEMICS. *Fresenius Z Anal Chem*, 324:750–759, 1986.

[69] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[70] R Koradi, M Billeter, M Engeli, P Güntert, and K Wüthrich. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 135(2):288–97, December 1998.

[71] Stefan Kuhn, Bjorn Egert, Steffen Neumann, and Christoph Steinbeck. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, 9(1):400, September 2008.

[72] ACD Labs. http://www.acdlabs.com. Online, 03 2013.

[73] ACD Labs. Acd/hnmr predictor v.9.0,, 2014.

[74] A H Land and A G Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28:497–520, 1960.

[75] Lee Griffiths, Howard H Beeley, and Rob Horton. Towards the automatic analysis of NMR spectra: Part 7. Assignment of 1H by employing both 1H and 1H/13C correlation spectra. *Magnetic resonance in chemistry*, 46:818–827, 2008.

[76] M. Leutner, R. M Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11(1):31–43, 1998.

[77] M. H. Levitt. Spin dynamics: Basics of nuclear magnetic resonance. 2001.

[78] Thomas Lindel, Jochen Junker, and Matthias Köck. Cocon: From nmr correlation data to molecular constitutions. *Molecular modeling annual*, 3(8):364–368, 1997.

[79] Xiaoyu Liu, K Balasubramanian, and ME Munk. Computer-assisted graph-theoretical construction of 13 c nmr signal and intensity patterns. *Journal of Magnetic Resonance (1969)*, 87(3):457–474, 1990.

[80] Jonathan A Lukin, Andrew P Gove, Sarosh N Talukdar, and Chien Ho. Automated probabilistic method for assigning backbone resonances of (13 C , 15 N) -labeled proteins. *Journal of Biomolecular NMR*, 9(416):151–166, 1997.

[81] Hideyuki Masui and Huixiao Hong. Spec2d: a structure elucidation system based on 1h nmr and hh cosy spectra in organic chemistry. *Journal of chemical information and modeling*, 46(2):775–787, 2006.

[82] A. F. Mehlkopf, D. Korbee, and T. A. Tiggelman. Sources of t 1noise in two-dimensional NMR. *Journal of magnetic ...*, 323(58):315–323, 1984.

[83] J Meiler, R Meusinger, and M Will. Fast determination of 13C NMR chemical shifts using artificial neural networks. *Journal of chemical information and computer sciences*, 40(5):1169–76, 2000.

[84] J Meiler and M Will. Automated structure elucidation of organic molecules from 13C NMR spectra using genetic algorithms and neural networks. *Journal of chemical information and modeling*, 41:1535–1546, 2001.

[85] Jens Meiler. PROSHIFT: protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR*, 26(1):25–37, May 2003.

[86] Jens Meiler and Matthias Köck. Novel methods of automated structure elucidation based on 13C NMR spectroscopy. *Magnetic resonance in chemistry : MRC*, 42(12):1042–5, December 2004.

[87] Jens Meiler, Walter Maier, Martin Will, and Reinhard Meusinger. Using neural networks for (13)c NMR chemical shift prediction-comparison with traditional methods. *Journal of Magnetic Resonance (San Diego, Calif.: 1997)*, 157(2):242–52, August 2002.

[88] Sergey G Molodtsov, Mikhail E Elyashberg, Kirill A Blinov, Antony J Williams, Eduard E Martirosian, Gary E Martin, and Brent Lefebvre. Structure elucidation from 2D NMR spectra using the StrucEluc expert system: detection and removal of contradictions in the data. *Journal of Chemical Information and Computer Sciences*, 44(5):1737–1751, 2004.

[89] Nathalie Morelle, Bernhard Brutscher, Jean-pierre Simorre, and Dominique Marion. Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments. *Journal of Biomolecular NMR*, 5:154–160, 1995.

[90] H N B Moseley, N Ríaz, J M Aramini, T Szyperski, and G T Montelione. A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. *Journal of Magnetic Resonance*, 170:263–277, 2004.

[91] K. P. Neidig, H. Bodenmueller, and Hans Robert Kalbitzer. Computer aided evaluation of two-dimensional nmr spectra of proteins. *Biochemical and biophysical research communications: BBRC*, 125(3):1143–1150, 1984.

[92] TJ Norwood. The effects of relaxation on the e. cosy experiment. *Journal of Magnetic Resonance, Series A*, 114(1):92–97, 1995.

[93] Jean-Marc Nuzillard and Massiot Georges. Logic for structure determination. *Tetrahedron*, 47(22):3655–3664, 1991.

[94] Jean-Marc Nuzillard and Georges Massiot. Computer-aided spectral assignment in nuclear magnetic resonance spectroscopy. *Analytica Chimica Acta*, 242:37–41, January 1991.

[95] VladislavYu. Orekhov, Ilghiz Ibraghimov, and Martin Billeter. MUNIN: A new approach to multi-dimensional NMR spectra interpretation. *Journal of Biomolecular NMR*, 20(1):49–60, May 2001.

[96] Luc Patiny, Julien Wist, and Andrés M. Castillo. MyLIMS - My Laboratory Information Management System. [Online]. Available: www.mylims.org.

[97] Harry E. Pence and Antony Williams. ChemSpider: An Online Chemical Information Resource. *journal of chemical education*, 87(11):10–11, 2010.

[98] B Plainchont, V P Emerenciano, and J M Nuzillard. Recent advances in the structure elucidation of small organic molecules by the LSD software. *Magnetic Resonance in Chemistry*, 51:447–453, 2013.

[99] B Plainchont, J M Nuzillard, G V Rodrigues, M J P Ferreira, M T Scotti, and V P Emerenciano. New Improvements in Automatic Structure Elucidation Using the LSD (Logic for Structure Determination) and the SISTEMAT Expert Systems. *Natural Product Communications*, 5:763–770, 2010.

[100] J. A. Pople and D. P. Santry. Molecular orbital theory of nuclear spin coupling constants. *Molecular Physics*, 8(1):1–18., 1964.

[101] E Pretsch, P Bühlmann, and M Badertscher. *Structure Determination of Organic Compounds*. Springer-Verlag, 4 edition, 2009.

[102] Reimond Bernstein, Christian Cieslar, Alfred Ross, Hartmut Oschkinat, Jens Freund, and Tad A Holak. Computer-assited assignment of multidimensional NMR spectra of proteins: Application to 3D NOESY-HMQC and TOCSY-HMQC spectra. *Journal of Biomolecular NMR*, 3:245–251, 1993.

[103] Mestrelab research. http://mestrelab.com. Online, 03 2013.

[104] Richard R Ernst. Nuclear Magnetic Resonance Fourier Transform Spectroscopy. In *Nobel Lectures in Chemistry 1991-1995*, page 308. World Scientific Publishing Co., 1997.

[105] H Satoh, H Koshino, J Uzawa, and T Nakata. H. Satoh, Koshino, J. Uzawa, T. Nakata. Tetrahedron, 2003, 59(25), 4539–4547. *Tetrahedron*, 59(25):4539–4547, 2003.

[106] Renate Bürgin Schaller, Morton E Munk, and Ernö Pretsch. Spectra estimation for computer-aided structure determination. *Journal of chemical information and computer sciences*, 36(2):239–243, 1996.

[107] Ac Schulte, a Gorler, C Antz, Kp Neidig, and Hr Kalbitzer. Use of global symmetries in automated signal class recognition by a bayesian method. *Journal of magnetic resonance (San Diego, Calif. : 1997)*, 129(2):165–72, December 1997.

[108] Sergey Molodtsov, Mikhail E Elyashberg, Kirill A Blinov, Antony J Williams, Eduard E Martirosian, Gary E Martin, and Brent Lefebvre. Structre elucidation from 2D NMR spectra using the StructEluc expert system: detection and removal of contradictions in the data. *Journal of chemical information and computer science*, 44:1737–1751, 2004.

[109] Yegor D Smurnyy, Kirill A Blinov, Tatiana S Churanova, Mikhail E Elyashberg, and Antony J Williams. Toward more reliable 13C and 1H chemical shift prediction: a systematic comparison of neural-network and least-squares regression based approaches. *Journal of Chemical Information and Modeling*, 48(1):128–34, 2008.

[110] Stephen G Spanton and David Whittern. The development of an NMR chemical shift prediction application with the accuracy necessary to grade proton NMR spectra for identity. *Magnetic resonance in chemistry : MRC*, 47(12):1055–61, December 2009.

[111] Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. NMRShiftDB-constructing a free chemical information system with open-source components. *Journal of Chemical Information and Computer Sciences*, 43(6):1733–9.

[112] Christoph Steinbeck and Stefan Kuhn. NMRShiftDB - compound identification and structure elucidation support through a free community-built web database. *Phytochemistry*, 65(19):2711–7, October 2004.

[113] S Tikole, V Jaravine, V Rogov, V Dötsch, and P Güntert. Peak picking NMR spectral data using non-negative matrix factorization. *BMC Bioinformatics*, 15:46, 2014.

[114] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F Schulte, David E Tolmie, R Kent Wenger, Hongyang Yao, and John L Markley. BioMagResBank. *Nucleic acids research*, 36(Database issue):D402–8, January 2008.

[115] Olga Vitek, Chris Bailey-Kellogg, Bruce Craig, Paul Kuliniewicz, and Jan Vitek. Reconsidering

complete search algorithms for protein backbone NMR assignment. *Bioinformatics (Oxford, England)*, 21 Suppl 2:ii230–6, September 2005.

[116] G Vivo Truyols, J R Torres Lapasi, A M van Nederkassel, Y Vander Heyden, and D L Massart. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part I: peak detection. *Journal of Chromatography A*, 1096:146–155, 2005.

[117] Jianyong Wang, Tianzhi Wang, Erik R P Zuiderweg, and Gordon M Crippen. CASA: an efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. *Journal of Biomolecular NMR*, 33(4):261–79, December 2005.

[118] R. Wehrens, C. Lucasius, L. Buydens, and G. Kateman. Sequential Assignment of 2D-NMR Spectra of Proteins Using Genetic Algorithms. *Journal of Chemical Information and Computer Sciences*, 33(2):245–251, 1993.

[119] D S Wishart, M S Watson, R F Boyko, and B D Sykes. Automated 1H and 13C chemical shift prediction using the BioMagResBank. *Journal of Biomolecular NMR*, 10(4):329–36, December 1997.

[120] Xiangqian Hu, David N Beratan, and Weitao Yang. A gradient-directed Monte Carlo method for global optimization in a discrete space: Application to protein sequence design and folding. *Journal of Chemical Physics*, 131(154117), 2009.

[121] Osamu Yamamoto, Kazuo Someno, Nobuhide Wasada, Jiro Hiraishi, Kikuko Hayamizu, Kazutoshi Tanabe, Tadao Tamura, and Masaru Yanagisawa. An Integrated Spectral Data Base System Including ESR and Raman Spectra IR , MS ,. *analytical sciences*, 4(June):233–239, 1988.

[122] Yegor D Smurnyy, Kirill A Blinov, Tatiana S Churanova, Mikhail E Elyashberg, and Antony J Williams. Towards more reliable 13C and 1H chemical shift prediction: a systematic comparison of neural-network and least-squares regression based approaches. *Journal of Chemical Information and Modeling*, 48:128–134, 2008.

[123] Zhi Liu, A Abbas, B Y JIng, and X Gao. WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. *Bioinformatics*, 28(7):914–920, 2012.

[124] L P Zhou, L L Sun, Y Yu, W Lu, and Z L Li. Prediction of carbon-13 NMR chemical shift of alkanes with rooted path vector. *Journal of Molecular Graphics and Modelling*, 25(3):333–339, 2006.

[125] D E Zimmerman, C a Kulikowski, Y Huang, W Feng, M Tashiro, S Shimotakahara, C Chien, R Powers, and G T Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of molecular biology*, 269(4):592–610, June 1997.