Automatic assignment of methyl resonances using experimental NMR data and graph theory



Iva Pritišanac
Magdalen college
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Physical and Theoretical Chemistry
Michaelmas 2016

Automatic assignment of methyl resonances using experimental NMR data and graph theory

Iva Pritišanac, Magdalen College

This dissertation is submitted for the degree of Doctor of Philosophy in Physical and Theoretical Chemistry

Michaelmas term 2016

Abstract

Selective isotope labeling of methyl groups allows for atomic-resolution insight into the structures and dynamics of high molecular mass proteins by nuclear magnetic resonance (NMR) spectroscopy. Widespread application of the methodology has been limited due to the challenges and costs associated with assigning ^{1}H-^{13}C resonances to particular amino acids within the protein. Here, I present a novel structure-based automatic assignment strategy, Methyl Assignment by Graph MAtching (MAGMA), which relates experimentally measured methyl-methyl connectivity (NOESY/DREAM) to inter-methyl distances extracted from a high-resolution protein structure. MAGMA features exact algorithms for graph-subgraph isomorphism and maximal common edge subgraph that can sample all theoretically possible methyl resonance assignments. MAGMA was applied to a benchmark of eight proteins for which high-resolution structures, NOESY or DREAM data, and methyl NMR assignments were available. On this benchmark, MAGMA provides a minimum of 31% and a maximum of 89% methyl residue assignments, with complete accuracy. On larger proteins in the benchmark MAGMA outperforms alternative automated methyl resonance assignment programs in both accuracy and coverage. I assessed the influence of different input structures on the accuracy of MAGMA assignments, and concluded that joining the assignments results from multiple structures guarantees accuracy in cases where the structural form in solution is unknown. Finally, I demonstrate the utility of MAGMA in two novel applications on molecular chaperones. In a ligand binding study involving the N-domain of human HSP90α, MAGMA confidently assigned 38% of the input methyl residues when the results were joined over two protein conformations. MAGMA's assignments were then combined with inter-molecular (ligand-methyl) NOEs and HADDOCK to generate models of the protein-ligand complex. The generated models, while featuring the correct ligand binding site, were of lower quality (0.5 - 2.5 Å higher ligand RMSD) when compared to models

generated with previously published methyl assignments. Incorporating inter-molecular NOE restraints in the calculation enabled MAGMA to discriminate between certain ligand binding modes, and led to an increase in the number of confident assignments, demonstrating the potential for using such restraints in the future. Finally, a generic data collection protocol for MAGMA was established in a *de novo* assignment application on the 400 kDa archaeal small heat shock protein Hsp16.5. The protocol involves measurement of inter-methyl NOEs, determination of intra-residue NOEs between Leu-δ_1/δ_2 and Val-γ_1/γ_2 groups, and discrimination between Leu and Val resonances by preparation of an exclusively Val labelled protein sample. A high sparsity of inter-methyl NOE data obtained for this system led to predominantly ambiguous methyl resonance assignments, which can be used for guiding further experimental efforts.

Table of contents

Abstract 2

Declaration 6

Acknowledgements 7

Chapter 1: NMR spectroscopy of biomolecules of high molecular mass 9

1.1 Brief history of biomolecular NMR spectroscopy 10

1.2 Relaxation in NMR 15

1.3 Transverse relaxation optimized spectroscopy (TROSY) 26

1.4 Nuclear Overhauser effect (NOE) 32

1.5 Methyl-TROSY: Extending the size limit of NMR spectroscopy 40

1.6 Automatic NMR resonance assignment 46

1.7 Thesis content. 63

1.8 References 64

Chapter 2: Methyl assignment by graph matching (MAGMA): Theoretical bases, algorithm design, and testing 71

2.1 Introduction 73

2.1.1 Definitions 73

2.1.2 Structure and data graphs 78

2.1.3 Graph comparisons and similarity measures 79

2.1.4 The maximal common edge subgraph (MCES) algorithm by McGregor 88

2.1.5 NP-completeness of subgraph-isomorphism and MCS problems 92

2.2 Results and discussion 94

2.2.1 Theoretical bases of MAGMA 94

2.2.2 MAGMA protocol - a combination of exact graph matching algorithms and vertex priority sorting heuristics 98

2.2.3 Proof-of-concept 108

2.3 Conclusions and outlook 112

2.4 Methods 114

2.5 References 118

Chapter 3: Testing MAGMA on experimental datasets 126

3.1 Introduction 127

3.1.1 Methyl side-chain assignment strategies 127

3.1.2 Structure-based methyl resonance assignment by FLAMEnGO2.0

and MAP-XSII 128

3.1.3 Methyl Assignment by Graph MAtching: a novel structure-based methyl assignment approach 129

3.1.4 Similarity between the crystal structure and structural restraints derived from NMR observables 130

3.2 Results and discussion 137

3.2.1 Restraints for MAGMA and MAGMA approach 132

3.2.2 Proteins in MAGMA's benchmark: experimental NMR data and data graph characteristics 136

3.2.3 Assignment coverage, accuracy, and ambiguity 140

3.2.4 Performance of alternative automatic methyl resonance assignment algorithms on MAGMA's benchmark 144

3.2.5 Alternative protein structures and MAGMA performance 148

3.3 Conclusions and outlook 153

3.4 Methods 157

3.5 References 162

Chapter 4: Applications of MAGMA to molecular chaperones 173

4.1 Introduction 174

4.2 Results and discussion 180

4.2.1 MAGMA generates correct assignments for N-domain of Hsp90 180

4.2.2 Choice of the N-domain structure influences assignment results 183

4.2.3 MAGMA's methyl assignments can be used in protein-ligand docking 187

4.2.4 Inter-molecular NOEs between methyls and a ligand increase the number of assigned methyl resonances 196

4.2.5 Automatic assignment of the ^{1}H-^{13}C HMQC spectrum of the 400 kDa small heat shock protein 16.5 (Hsp16.5) 199

4.3 Conclusions and Outlook 206

4.4 Methods 209

4.4.1 HSP90α 209

4.4.2 Hsp16.5 211

4.5 References 213

Concluding remarks 224

Declaration

This dissertation is a summary of research performed at the Department of Physical and Theoretical Chemistry, University of Oxford between October 2013 and September 2016. The work described here is wholly my own and not the outcome of a collaboration except where this is specifically stated in the text. It has not, neither as a whole, nor in part been submitted for a degree, diploma or any other academic qualification at any other university or academic institution.

Iva Pritišanac

Oxford, October 2016

Acknowledgements

I owe my gratitude to my supervisor Prof. Andrew Baldwin for his continuous guidance and support during my DPhil. I am especially grateful for our scientific discussions, for giving me the freedom to follow my research path, and for his contagious scientific enthusiasm. It was a privilege to be interacting with Prof. Carol Robinson and Prof. Justin Benesch during my time in Oxford, and I here thank them for their invaluable advice, for providing their insight to my work, and their generous assistance whenever needed. I consider myself fortunate to have been able to work alongside Reid Alderson whose scientific curiosity, expertise, and hard work have been both motivating and vital for the success of our joined project. I am particularly grateful to Reid for proof-reading the entirety of this thesis, for his unconditional support whenever needed, and for being my best friend throughout. I am indebted to Matteo Degiacomi for his inspiring and enthusiastic teaching, "(num)pythonic advices", "single liners", and for all of our graph-related and unrelated conversations. I thank Marta Carneiro, Eiso Ab, and Gregg Siegal from ZoBio for a stimulating collaboration. I would like to particularly thank Carla Schmidt, Heidi Gastall, Elodie Limer, Jun Yan, and Michael Barber, who have been both stimulating colleagues and dear friends, and whose presence in the group is greatly missed. I am grateful to David Dias, Georg Hochberg, Gogulan Karunanithy, Henrik Müller, Idlir Liko, Olga Tkachenko, Shane Chandler, and Timothy Nott for being exceptional and motivating colleagues, and with whom I shared some great times both in and outside the office. It has been a pleasure to share computational seminars at Magdalen with Erik Marklund, Timothy Allison, Michael Marty, and Matteo Degiacomi who continuously sparked my interest in computational structural biology. I extend my thanks to all my colleagues in the Baldwin, Benesch, and Robinson groups who have in different ways motivated me to do my best during my time in Oxford. Finally, my DPhil years would not

have been the same without a wealth of support from my family and I thank my sisters Ana and Ena and my parents Irena, Damir, and Nada.

Chapter 1: NMR spectroscopy of biomolecules of high molecular mass

Abstract

From relatively modest beginnings in the 1950s, the number of NMR spectroscopic-based investigations of biomolecules has been rapidly increasing, in part due to continuous methodological advancements that now enable atomic-resolution studies of the structures and dynamics of biomolecules up to 1 MDa in mass. For many years, however, NMR was limited to systems that were less than 40-50 kDa, due to unfavourable transverse relaxation resulting in rapid signal loss in large systems. Here, I provide a brief historical overview of methodological advances in the biomolecular NMR field that led to the development of transverse relaxation optimized spectroscopy (TROSY) and methyl-TROSY methods. Redfield's semi-classical NMR relaxation theory is used to consider chemical shift anisotropy (CSA) and the spin dipolar interaction, and demonstrate how destructive interference of these relaxation mechanisms leads to the TROSY enhancement that is now routinely exploited in heteronuclear multi-dimensional NMR experiments. I extend the discussion to the interference of dipolar interactions within $^{13}CH_3$ spin systems in the macromolecular limit, and illustrate why the HMQC sequence is well-suited for the TROSY enhancement. In addition, for methyl-labelled high molecular mass systems, the nuclear Overhauser effect (NOE) provides valuable information about the proximity of methyl groups and when embedded between two HMQC sequences, allows for high sensitivity and resolution measurements. Finally, a variety of protocols for the selective labelling of methyl groups were important developments for the applications of methyl-TROSY studies to large systems. Together, methodological advances in both pulse sequences and isotope labeling schemes have led to impressive applications elucidating the functional dynamics in high molecular mass systems such as the 670 kDa 20S core particle (CP) proteasome and 204 kDa SecA molecular chaperone. Automating assignment of NMR resonances in a reliable manner

is necessary for NMR spectroscopy to be more readily and more efficiently applied to biophysical chemistry, drug design, structural biology, structural genomics, and interactomics. Here, an overview of numerous algorithmic strategies employed for automatic assignment of backbone and side chain resonances of proteins is given, with a focus on structure-based strategies employed for assignment of methyl resonance in high molecular mass proteins.

1.1 Brief history of biomolecular NMR spectroscopy

The application of Nuclear Magnetic Resonance (NMR) spectroscopy to investigate proteins first occurred in the late 1950s, with a start roughly placed around the first report of an NMR spectrum of a protein[1]. The ^{1}H continuous-wave (CW) NMR spectrum of bovine pancreatic ribonuclease (13.7 kDa) was obtained at a magnetic field strength of only 40 MHz and featured considerable signal overlap. From such an "unbroadened spectrum of the nonexchangeable protons of ribonuclease"[1], the authors could tentatively assign only the aromatic and aliphatic protons, yet demonstrated that their signals approximately matched the expected signals given the proton composition of the protein.

From these modest beginnings, rapid advances in NMR methodology have produced methods such as multi-dimensional heteronuclear NMR techniques[2-4], Transverse Relaxation Optimized SpectroscopY (TROSY)[5-6], protocols for NMR structure calculations[7-10], methyl-TROSY protocols for studying the structures and dynamics of high molecular mass proteins[11-15], and in-cell NMR protein studies[16-19]. In what follows, a brief overview of a small fraction of the progress in biomolecular NMR spectroscopy will be given, highlighting only a few important contributions. It should be noted that the relevant contributions are too many to consolidate here, and almost certainly some of the crucial works have been omitted.

In the 1950s a theoretical proposal of the nuclear Overhauser effect (NOE) was given by Albert Overhauser, who considered polarization of atomic nuclei in metals upon saturation of

the spin electron resonance of the conducting electrons[20]. The NOE was later experimentally verified by Carver and Slichter[21] and the first simple theoretical description was given by Solomon[22]. Following these discoveries, Kaiser[23] and Anderson and Freeman[24] investigated the NOE effect originating from nuclear spin-spin interactions in proton spectra of organic molecules. Another important theoretical work in the 1950s was that of Martin Karplus who considered scalar or *J*-couplings caused by through-bond hyperfine interactions between electrons and nuclear spins, and empirically related them to nuclear geometric configurations (i.e. the Karplus equation)[25-26]. Together, roughly 30 years after these seminal works, *J*-couplings and NOEs provided the basis for the first NMR-derived biomolecular structures[27]. Other significant developments in the 1950s included works from Carr, Purcell, Meiboom, and Gill that provided the basis for modern NMR measurements of chemical exchange effects in biomolecules[22, 28].

In the beginning of the 1960s, biomolecular NMR suffered from low signal-to-noise (S/N) ratios, a consequence of low magnetic field strengths typically employed at the time (recall that $\mathrm{S/N} \propto B_0^{3/2}$, where B_0 stands for the static magnetic field strength[57]). Furthermore, the presence of many ^{1}H nuclear spins in protein NMR samples resonating at closely spaced frequencies, combined with the limited resolution that could be achieved using one-dimensional continuous wave (CW) spectrometry, resulted in spectral crowding and overlapping signals. The technological developments in this decade included the transition from CW to pulsed excitation NMR, the Fourier transformation of free induction decays (FIDs) from pulsed excitation, the ability to signal average using the combination of multiple FIDs, the introduction of superconducting magnets, and the deployment of isotope labeling strategies to minimize spectral complexity, all of which contributed to enhanced sensitivity[29]. In particular, groundbreaking work from Richard Ernst and Wes Anderson[30] demonstrated the utility of pulsed Fourier transform NMR, which enabled simultaneous excitation of all

spin systems whose signals could be then separated by subsequent Fourier transformation, and significantly sped up data acquisition. Largely due to pulsed excitation and Fourier transformation, NMR rapidly advanced in the 1970s and 1980s. Two-dimensional (2D) NMR spectroscopy was proposed in 1971 by Jean Jeener[31] and was experimentally realized in 1976 by Ernst and coworkers[32]. Many homonuclear 2D NMR experiments followed, including, for example, Correlation SpectroscopY (COSY), NOE SpectroscopY (NOESY), and EXchange SpectroscopY (EXSY)[33]. Such experiments, which spread otherwise overlapping NMR signals into a second dimension, facilitated NMR studies of larger molecules in a site-specific manner; however, heteronuclear applications were challenged by lower sensitivity for ^{13}C and ^{15}N nuclei, due to their lower gyromagnetic ratios (γ) when compared to that of ^{1}H (Table 1.1). To circumvent signal loss, Ray Freeman and Gareth Morris proposed a method called Insensitive Nuclei Enhanced by Polarization Transfer (INEPT)[34], which resulted in the enhancement of heteronuclear NMR signals by the ratio γ_I/γ_S, where *I* corresponds to the more sensitive nucleus (typically ^{1}H) and *S* to the insensitive heteronucleus.

Table 1.1 | NMR-active nuclei that are commonly utilized in biomolecular NMR spectroscopy. For each nucleus, the gyromagnetic ratio (γ), spin quantum number (*I*), and natural abundance are listed. Note that isotopic enrichment of lowly populated isotopes can be readily achieved through bio- or semi-synthetic means.

Nucleus	**γ (rad s^{-1} gauss $^{-1}$)**	***I***	**Natural abundance (%)**
^{1}H	26,753	1/2	99.980
^{2}H	4,106	1	0.016
^{13}C	6,728	1/2	1.108
^{15}N	-2,712	1/2	0.370
^{19}F	25,179	1/2	100.000
^{31}P	10,841	1/2	100.000

To reach their full potential, heteronuclear NMR experiments had to wait for developments in biotechnology and biosynthetic isotope labeling schemes that enabled enrichment of proteins with ^{13}C and ^{15}N isotopes. On the other hand, advances in 2D homonuclear NMR enabled the assignment of most proton resonances in proteins up to 10 kDa in mass, which when

combined with 2D NOESY and COSY experiments enabled generation of some of the first protein structures obtained on the basis of experimental NMR data. Demonstrating a promising role for NMR in structural biology, early structures solved by NMR were in good agreement with crystal structures of homologous proteins (Figure 1.1)[35-36]. These and numerous related efforts in the field provided the basis for biomolecular NMR structure calculation[37], which revolutionized structural biology by adding another method, next to X-ray crystallography, capable of determining structures at atomic resolution. The field of structural biology has been undergoing another revolution due to the rapid improvement of cryo-electron microscopy single-particle analysis[125]. Recent developments in the resolution of the technique are a direct result of advances in both hardware (improved direct electron detectors) and software (improved image processing)[125]. Therefore, close to atomic or atomic resolution structural models for many specimens can now be obtained using three different structural biology techniques.

The field of NMR structural calculations emphasizes the significance of the NOE restraint, its rich information content, and its direct relation to protein structure, which is a topic that will be discussed further in this thesis (Chapters 2 and 3).

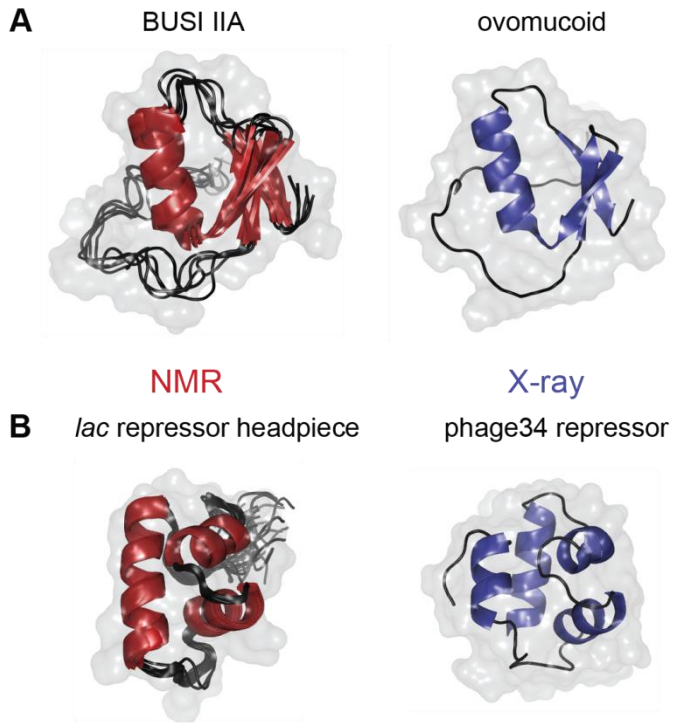


Figure 1.1 | A comparison of some of the first protein structures solved by X-ray crystallography and NMR spectroscopy. (**A**) *Left*, ensemble of five structures that were determined for the protein BUSI IIA (PDB: 1BUS; see text). *Right*, crystal structure of the third domain from ovomucoid, a homologous protein (PDB: 1OVO). (**B**) *Left*, ensemble of 20 NMR structures for the *lac* repressor headpiece (PDB: 1LQC). Note the original structure was not deposited in the PDB. *Right*, crystal structure of the phage34 repressor, a homologous protein. Note the general agreement in tertiary structures.

During the late 1980s and early 1990s, the recombinant production of uniformly ^{13}C and ^{15}N labelled proteins and the development of novel 3D and 4D heteronuclear NMR experiments provided access to heteronuclear backbone and side-chain chemical shifts[2,38]. The large chemical shift ranges of ^{13}C and ^{15}N nuclei alleviated the issues of overlapping resonances in homonuclear spectra, and led to the development of now routinely used backbone resonance assignment strategies[39]. Enabled by these heteronuclear experiments, measurements of transverse, longitudinal, and cross relaxation of ^{13}C and ^{15}N nuclei were realized, which provided insight into protein backbone and side-chain dynamics across the pico- to nanosecond timescales[40-45]. Roughly around the same time, theoretical and experimental framework for the measurement of residual dipolar couplings (RDCs) was established, adding a valuable complementary restraint to the NMR structure calculation method[46-48].

In the 2000s, relaxation dispersion experiments were introduced allowing measurements of molecular dynamics on the micro-to millisecond timescales[49-50]. On the other hand, RDCs can sample intermediate nano- to microsecond timescales[51]. Taken together, these methods allow NMR spectroscopy to span and pico- to millisecond timescales, establishing it as the

primary experimental method for characterizing biomolecular dynamics at the level of individual atoms[50, 52].

However, despite these continuing breakthroughs, biomolecular NMR studies up until late 1990s were restricted to molecules of less than ~30 kDa, due, in part, to enhanced transverse relaxation rates (R_2) in large systems that cause rapid signal decay. The introduction of transverse relaxation-optimized spectroscopy (TROSY)[53] resulted in a ~40-fold suppression of R_2 relaxation rates of amide protons, resulting in significant improvements in resolution. TROSY was subsequently applied to a number of conventional 3D and 4D experiments (e.g. TROSY-HNCA and TROSY-HNCO) and led to up to ~50-fold gains in signal-to-noise, depending on the correlation time of the protein[54-55].

While TROSY facilitated detailed NMR structural investigations of proteins up to ~50 kDa in mass, the need for perdeuteration, introduced to maximize the TROSY enhancement, reduced the number of protons that can give valuable distances restraints for NMR structure determinations. Here, the introduction of side-chain methyl labelling proved valuable as long-range methyl-methyl NOEs can span residues that are separated by more than 5 amino-acids in the primary sequence[56], thus providing important restraints for global protein fold determination. Next to their usefulness in NMR structure calculation, methyl groups have been utilized as the only NMR probes for high resolution studies of large systems[13], as discussed in more detail below.

1.2 Relaxation in NMR

To understand the origin of both TROSY and NOE effects, mechanisms of NMR relaxation need to be considered. This advanced topic of NMR theory falls far beyond the scope of this thesis, and here only a brief introduction is given with focus placed on aspects relevant for understanding chemical shift anisotropy (CSA), spin dipole-dipole interactions, and the basis

of the relaxation interference effects. A comprehensive coverage of these and other topics can be found elsewhere[33, 57-58].

The density operator

Any quantum mechanical system can be described by a wavefunction Ψ. The time-dependent Schrödinger equation considers the evolution of the quantum mechanical system over a time period *t*

$$\frac{\partial}{\partial t}\Psi(t) = -i\hat{\mathrm{H}}(t)\Psi(t) \tag{1.1}$$

where $\Psi(t)$ is the state function of the system and $\hat{H}$(t) is a Hamiltonian (the total energy) operator that acts on the system over time *t*. For convenience of using angular frequency units, the reduced Planck's constant ($\hbar = \frac{h}{2\pi}$) is left out of the equation. Using the equation and the wavefunction of the nuclear spin system, effects of different Hamiltonians on the system can be evaluated, such as a free-precession Hamiltonian ($\hat{H}=\omega_0\hat{I}_z$) or a radiofrequency pulse applied along the x-axis of the rotating frame ($\hat{H}=\omega_1\hat{I}_x$). In what follows, it will be demonstrated that the same observables can be evaluated more easily by using the density matrix formalism.

The wavefunction $\Psi(t)$ can be expanded in terms of a basis set, in Dirac notation:

$$|\Psi(t)> = \sum_{i=1}^{n} c_i(t)|i> \tag{1.2}$$

where *i*, $\{|i>, i = 1,2, \ldots n\}$ is the vector of all allowed basis functions and $c_i(t)$ are the time-dependent coefficients. Importantly, for a nuclear spin, relevant in the context of NMR, the phase space consists of only two states α and β, i.e. i = {α (m=1/2, spin up), β (m=−1/2, spin down)}. The whole phase space can thus be constructed by a linear combination of the two

states; i.e. $|\Psi> = c_1 |\alpha> + c_2 |\beta>$, with the regular normalization condition for the coefficients c_1 and c_2.

Now, a general definition for the density operator $\hat{\rho}(t)$ can be given (using the *bra* and *ket* notation) as a product:

$$\hat{\rho}(t) = |\Psi(t) >< \Psi(t)| \tag{1.3}$$

For an ensemble of spins, the probability p^k that a spin in the ensemble is in one of the possible states $|\Psi^k(t) >$ of the system needs to be taken into account in the definition. The density operator $\hat{\rho}(t)$ for this "mixed state" is an ensemble average, defined as:

$$\hat{\rho}(t) = \sum_k p^k |\Psi^k(t) >< \Psi^k(t)|, \qquad such\ that \sum_k p^k = 1 \tag{1.4}$$

Combining the equations 1.2 and 1.4:

$$\hat{\rho}(t) = \sum_i \sum_j \overline{c_\iota(t) c_j^*(t)}\ |i >< j| \tag{1.5}$$

where the horizontal bar indicates the ensemble average. Considering the elements of a density matrix ρ_{rs} that corresponds to the density operator $\hat{\rho}(t)$ for an ensemble with state probabilities p^k defined as above:

$$\rho_{rs} =< r|\hat{\rho}(t)|s >= \sum_k p^k\, c_r^k(t) c_s^{k*}(t) = \overline{c_r(t) c_s^*(t)} \tag{1.6}$$

it is notable that the elements in the density matrix correspond to the products of the coefficients of the wavefunction *Ψ(t)* (see 1.2). The diagonal density matrix elements correspond to the probability of finding the spin system in the eigenstate $|r >$ (i.e. the populations), which is evident when the density matrix is considered in the eigenbasis of the Hamiltonian *Ĥ:*

$$\rho_{rr} =< r|\hat{\rho}(t)|r >= \overline{|c_r(t)|^2} \tag{1.7}$$

The off-diagonal elements for this basis correspond to:

$$\rho_{rs} =< r|\hat{\rho}(t)|s >= \overline{c_r(t)\, c_s^*(t)} \tag{1.8}$$

which are so called *coherent superpositions* or *coherences*.

The goal of NMR spectroscopy measurements is to obtain certain observable quantities (e.g. spin angular momentum). It is therefore useful to relate those observables to the density matrix of the system under investigation. Considering an application of an operator $\hat{A}$ to a system in a particular state $|\Psi^k(t) >$, the quantity of the observable will be given by the expectation value of the operator $\hat{A}$ denoted $\langle \hat{\mathrm{A}} \rangle$:

$$\langle \hat{\mathrm{A}} \rangle = \sum_k p^k(t) \left\langle \Psi^k(t) | \hat{\mathrm{A}} | \Psi^k(t) \right\rangle \tag{1.9}$$

Using the density matrix ρ_{rs} in equation 1.9, it can be shown that the expectation value of an observable operator $\hat{A}$ corresponds to the trace of its product with the density operator:

$$\langle \hat{\mathrm{A}} \rangle = Tr(\hat{\mathrm{A}}\hat{\rho}(t)) \tag{1.10}$$

This relation has important practical implications as the density operator alone gives an easy handle on the observables removing the need to deal with the wavefunction $\Psi(t)$ (1.2) of the system explicitly. The density matrix formalism allows evaluation of time evolution of the spin system of interest through the time evolution of the density operator $\hat{\rho}(t)$. This can be derived from the time-dependent Schrödinger equation (1.1) and the definition for the density operator (1.3), yielding:

$$\frac{\partial}{\partial t}\hat{\rho}(t) = -i[\hat{\mathrm{H}}(t), \hat{\rho}(t)] \tag{1.11}$$

Equation 1.11 is the famous Liouville-von-Neumann equation and plays a central role in the calculations of the dynamics of quantum mechanical systems, as will be discussed below. Note that, given the general definition for the complete density operator $\hat{\rho}(t)$ it is defined for an entire quantum mechanical system, including all electrons and nuclei. For applications in NMR, it is sufficient to consider the reduced density operator $\hat{\sigma}(t)$, that depends only on the nuclear spins, which can be substituted for $\hat{\rho}(t)$ in the Liouville-von-Neumann equation[33].

Now that the density operator and its time evolution have been introduced, relaxation can be referred to in terms of the state and evolution of the density operator. Starting from the thermal equilibrium state in which no coherences have yet been generated, all off-diagonal elements of the density operator are zero. Given the diagonal elements of the density operator, the populations of the energy levels of the spin system correspond to the Boltzmann distribution. The relaxation of the system is thus a process by which a density operator defined for a particular state of the system returns to equilibrium[57].

Redfield semi-classical relaxation theory

Relaxation in NMR is traditionally described at four levels, which are listed below in the order of increasing rigor:

1) Phenomenological Bloch equations
2) Second order perturbation theory (transition probabilities)
3) Semi-classical relaxation theory (Redfield)
4) Quantum mechanical relaxation theory

Relaxation (super)operator of 1) and 2) cannot be used to treat systems with spin coupling or transverse relaxation, respectively[33]. Semi-classical relaxation theory (3), such as that by Redfield[59], is usually regarded as the most practical approach, as it offers a more fundamental treatment of the density operator that allows for a quantum-mechanical evolution of the spin-

system whilst the interaction of the spins and their surroundings (lattice) is treated with fluctuating random processes.[33] Redfield's semi-classical relaxation theory is adopted herein.

Master equation for relaxation

Time-dependent fluctuations in the lattice that cause relaxation of the spin system can be considered in the Hamiltonian operator. The Hamiltonian can be split into two components; a time-independent component $\hat{H}_0$ and a time-dependent component $\hat{H}_1(t)$ that describes the interaction of the spin system with the lattice.

$$\hat{\mathrm{H}} = \hat{H}_0 + \hat{H}_1(t) \tag{1.12}$$

Spin interactions that contribute to $\hat{H}_0$, also known as the main static Hamiltonian, include the Zeeman interaction, scalar coupling, and the isotropic chemical shift. The two principal components of $\hat{H}_1(t)$ are the anisotropic chemical shift (chemical shift anisotropy, CSA) and dipolar coupling. To account for $\hat{H}_1(t)$, the Liouville von Neumann equation (1.11) is expanded to:

$$\frac{\partial}{\partial t}\hat{\sigma}(t) = -i[\hat{\mathrm{H}}_0 + \hat{\mathrm{H}}_1(t), \hat{\sigma}(t)] \tag{1.13}$$

where $\hat{\sigma}(t)$ stands for the reduced density operator (see above) and the interaction Hamiltonian $\hat{H}_1(t)$ is defined as:

$$\hat{\mathrm{H}}_1(t) = \sum_q F_q(t)A_q = \sum_q F_q^{\dagger}(t)A_q^{\dagger} \tag{1.14}$$

where A_q, the spin term, stands for the operators that act on the spin system alone and $F_q(t)$, the spatial term, for random processes that represent the dynamics of the lattice that are treated classically. The index q distinguishes both the components of a spherical tensor operator and different types of interactions, and † stands for the complex conjugate of the matrix transpose (or just Hermitian conjugate matrix). By moving to the interaction frame, where the effect of the static Hamiltonian $\hat{H}_0$ is effectively removed, the treatment can be focused on $\hat{H}_1(t)$ and equation (1.13) takes the form:

$$\frac{\partial}{\partial t}\tilde{\sigma}(t) = -i[\tilde{H}_1(t), \tilde{\sigma}(t)] \quad (1.15)$$

where a tilde denotes the interaction frame. The derivation proceeds with integration followed by Taylor expansion and truncation at the second term and then differentiation again as detailed in Cavanagh *et al.*[57] The stochastic Hamiltonian, $H_1(t)$ is then expressed in terms of spatial and spin functions as above. In the derivation, the spatial term defines the correlation function $G_{qq'}(t+\tau)$:

$$G_{qq'}(\tau) = \overline{F_q(t)F_{q'}^*\ (t+\tau)} \quad (1.16)$$

that reflects the resemblance of the spatial term at time t ($F_q(t)$) to the term after a time delay τ ($F_{q'}^*\ (t+\tau)$), where * stands for the complex conjugate. The correlation function in its simplest form can be expressed as a mono-exponential decaying function with a characteristic time $\tau_2 = \tau_c$. Importantly, the correlation function of the second order terms ($3\ cos(\theta)^2 - 1$) is considered, rather than the autocorrelation function of the first order term ($cos(\theta)$), which is characterized by a correlation time τ_1.

Finally, when restricted to the secular contributions[33], Redfield's semi-classical master equation for relaxation is given by:

$$\frac{\partial}{\partial t}\tilde{\sigma}(t) = -\frac{1}{2}\sum_q \sum_p J_p^q(w_q)\left[A_p^{q\dagger}, \left[A_p^q, \tilde{\sigma}(t) - \sigma_0\right]\right] \quad (1.17)$$

The equation describes the time evolution of the density matrix in the interaction frame, where $J_{qq}(w_q)$ is the component of the spectral density function J_{qq} at frequency w_q, q stands for the qth component (rank) of the spherical tensor operator of A_q, p iterates over the spin operators that have the same q but different eigenfrequencies, and σ_0 is the equilibrium density operator. Note that the density operator evolving to its equilibrium value σ_0, rather than to the state of complete disorder as would be inferred from equation (1.13), has been added *ad hoc* (1.17) to comply with experimental observations.

Upon transformation to the laboratory frame, equation (1.17) is invariant[33]:

$$\frac{\partial}{\partial t}\sigma(t) = -i\left[\widehat{H}_0, \sigma(t)\right] - \hat{\hat{\Gamma}}\{\sigma(t) - \sigma_0\} \tag{1.18}$$

where $\hat{\hat{\Gamma}}$ denotes the relaxation superoperator:

$$\hat{\hat{\Gamma}}\{\sigma\} = \frac{1}{2}\sum_q \sum_p J_q\left(w_p^q\right)\left[A_p^{q\dagger}, \left[A_p^q, \sigma\right]\right] \tag{1.19}$$

The spectral density functions at different transition frequencies $J_q(w_q)$ act as coefficients of the double commutators in the sum[33], which in turn represent different operators of the interaction Hamiltonian $\widehat{H}_1(t)$. This notion gives the equation a practical usefulness as it can be used to derive relaxation rate constants of different spin interactions. For instance, given two operators I_m and I_n, the relaxation rate $\hat{\hat{\Gamma}}_{mn}$ is given by:

$$\hat{\hat{\Gamma}}_{mn} = \frac{1}{2}\sum_q \sum_p Tr\left(I_m\left[A_p^q, \left[A_p^{q\dagger}, I_n\right]\right]\right) / \left(I_m^\dagger I_m\right) J_q\left(w_p^q\right) \tag{1.20}$$

Note that the equation is correct only for secular product operator terms. The equation will be used below to give relaxation rate constants of specific product operator terms governing the dipolar interaction and CSA effect. More details and a derivation of the master equation can be found in textbooks from Goldman[60], Ernst *et al.*[33], and Cavanagh *et al.*[57].

Relaxation mechanisms in diamagnetic biological macromolecules

The dominant relaxation mechanisms that govern interactions of spin-1/2 nuclei in diamagnetic proteins are chemical shift anisotropy (CSA) and the dipole-dipole interaction[57].

Dipole-dipole interaction

The dipole-dipole interaction is the most important relaxation mechanism for spin-1/2 nuclei[58]. The origin of the interaction is the magnetic dipolar field that is associated with every non-zero spin angular momentum nucleus and is proportional to the magnetic moment

of the nucleus. Due to molecular tumbling in solution, the field fluctuates and causes relaxation of nearby spins[57].

The Hamiltonian of the dipolar interaction $\hat{H}_{DD}$ for the two-spin system of spins *I* and *S* is given by:

$$\hat{\mathrm{H}}_{DD} = b_{IS}\left(I \cdot S - 3(I\,\widehat{r_{IS}})(S\,\widehat{r_{IS}})\right) \tag{1.21}$$

where $b_{IS} = \mu_0\gamma_I\gamma_S\hbar^2/4\pi r_{IS}^3$, μ_0 is the magnetic constant, $\gamma$is the gyromagnetic ratio of a nucleus (Table 1.1), ħ is the reduced Planck's constant, r_{IS} is a vector connecting nuclei *I* and *S*, and $\widehat{r_{IS}} = r_{IS}/|r_{IS}|$ is a normalized internuclear vector in this direction. The dipolar Hamiltonian can be expressed in terms of irreducible tensor operators if the normalized internuclear vector $\widehat{r_{IS}} = r_{IS}/|r_{IS}|$ is expressed in polar coordinates θ_{IS} and ϕ_{IS}

$$\hat{\mathrm{H}}_{DD} = \sum_{q=-2}^{2} F_q^{IS} A_q^{IS} \tag{1.22}$$

where the functions contained in the spatial term F_q^{IS} describe the orientation, and the operators contained in the spin term A_q^{IS} describe the interactions (Table 1.2).

Inserting the terms into the master equation (1.17) gives:

$$\frac{\partial}{\partial t}\tilde{\sigma}(t) = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{(\gamma_I\gamma_S\hbar)^2}{r^6} \sum_q \sum_p J_p^q(w_q)\left[A_p^{q\dagger}, \left[A_p^q, \tilde{\sigma}(t) - \sigma_0\right]\right] \tag{1.23}$$

This equation (1.23) shows how the efficiency of dipolar relaxation is proportional to both the inverse sixth power of the distance between the nuclei and their gyromagnetic ratios (Table 1.1). This is why the dipolar interaction acts over a short range and is the strongest for proton nuclei, which have the largest gyromagnetic ratio. These findings have direct consequences on the outcome of NOESY experiments (see below).

Table 1.2 | Spatial (F_q^{IS}) and spin (A_q^{IS}) terms of dipolar Hamiltonian for two spins

$A_0 = b_{IS}\{I_zS_z - \frac{1}{4}(I_+S_- + I_-S_+)\}$	$F_0 = 1 - 3cos^2\theta_{IS}$
$A_1 = -\frac{3}{2}b_{IS}(I_zS_+ + I_+S_z)$	$F_1 = sin\theta_{IS}\cos\theta_{IS}\exp(-i\,\phi_{IS})$
$A_{-1} = -\frac{3}{2}b_{IS}(I_zS_- + I_-S_z)$	$F_{-1} = sin\theta_{IS}\cos\theta_{IS}\exp(+i\,\phi_{IS})$
$A_2 = -\frac{3}{4}b_{IS}I_+S_+$	$F_2 = sin^2\theta_{IS}\exp(-2i\,\phi_{IS})$
$A_{-2} = -\frac{3}{4}b_{IS}I_-S_-$	$F_{-2} = sin^2\theta_{IS}\exp(+2i\,\phi_{IS})$

Chemical shift anisotropy (CSA)

The external magnetic field induces motion of electrons surrounding atomic nuclei, which in turn gives rise to secondary nuclear magnetic fields. The total magnetic field experienced by any given nucleus is a sum of the contributions from the static magnetic field B_0 and secondary fields (i.e. chemical shielding). The anisotropy of local fields, combined with molecular motion, results in modulation of the different components of the fields in the laboratory frame, thereby causing relaxation. Shielding is described by the second rank tensor σ_{kk}, and the relation between the chemical shift tensor and shielding is given as:

$$\delta = (1 - \sigma_{kk}) \tag{1.24}$$

A transformation to the principal coordinate system of the chemical shift tensor (denoted δ^P) gives a diagonal matrix:

$$\delta^P = \begin{bmatrix} \delta_{xx} & 0 & 0 \\ 0 & \delta_{yy} & 0 \\ 0 & 0 & \delta_{zz} \end{bmatrix} \tag{1.25}$$

and the Hamiltonian for chemical shift anisotropy in the principle coordinate system ($\hat{\mathrm{H}}_{CSA}^P$) is given by:

$$\hat{\mathrm{H}}_{CSA}^P = \gamma\begin{bmatrix} I_xI_yI_z \end{bmatrix}\begin{bmatrix} \delta_{xx} & 0 & 0 \\ 0 & \delta_{yy} & 0 \\ 0 & 0 & \delta_{zz} \end{bmatrix}\begin{bmatrix} B_x \\ B_y \\ B_z \end{bmatrix} \tag{1.26}$$

The principal components of δ define the average (isotropic) chemical shift:

$$\delta_{iso} = (\delta_{xx} + \delta_{yy} + \delta_{zz})/3 \quad (1.27)$$

The chemical shift anisotropy ($\Delta\delta$) and asymmetry of the tensor (η) are given by:

$$\Delta\delta = \delta_{zz} - (\delta_{xx} + \delta_{yy})/2 \quad (1.26)$$

$$\eta = \frac{\delta_{yy} - \delta_{xx}}{\delta_{zz}} \quad (1.27)$$

The maximum size of the chemical shift anisotropy of a given nucleus is approximately given by the range of the chemical shift for the nucleus, which is why this mechanism contributes significantly to relaxation of nuclei with a large chemical shift range, for instance ^{15}N and ^{13}C in carbonyls, but has a small effect on relaxation of ^{1}H and ^{13}C in methyl groups. In many cases the shielding tensor, and hence the chemical shift shows axial symmetry (i.e. $\delta_{xx} = \delta_{yy} \neq \delta_{zz}$), in which case the full CSA Hamiltonian in the laboratory frame[61] is given by:

$$\hat{\mathrm{H}}_{CSA} = \Delta\sigma\,\gamma B_0 \left(\frac{1}{3}I_z(3cos^2\theta - 1) + \frac{1}{2}I_+(sin\theta\cos\theta\ \exp(-i\,\phi)) + \frac{1}{2}I_-(sin\theta\cos\theta\ \exp(+i\,\phi))\right) \quad (1.28)$$

The equation 1.28 shows the angular dependence of the CSA (θ, ϕ). When the spin and spatial terms of the $\hat{\mathrm{H}}_{CSA}$ Hamiltonian are substituted in the master equation (1.17), the square dependence of the CSA relaxation mechanism on the magnetic field strength (B_0) and the chemical shift anisotropy ($\Delta\sigma$) becomes evident:

$$\frac{\partial}{\partial t}\tilde{\sigma}(t) = \frac{4}{9}(\Delta\sigma\gamma B_0)^2 \sum_q \sum_p J_p^q(w_q)\left[A_p^{q\dagger}, \left[A_p^q, \tilde{\sigma}(t) - \sigma_0\right]\right] \quad (1.29)$$

This is why CSA competes with the dominant dipolar-dipolar relaxation mechanism at high magnetic field strengths.

Interference of CSA and dipolar relaxation mechanisms

The coupled movement of nuclear spins and electrons due to molecular rotation results in correlation of the fluctuations in the dipole-dipole interactions and the CSA, giving rise to cross-correlation of these different relaxation mechanisms.

1.3 Transverse relaxation optimized spectroscopy (TROSY)

The origin of the TROSY effect is interference between dipole-dipole and CSA relaxation mechanisms. The interference influences the relaxation rates of particular coherences and, if one selects for coherences with rate contributions from the destructively interfering components, can have great practical use[6, 53, 62]. For instance, in a scalar coupled *AXn* spin system, the relaxation rates of single quantum coherences can differ, giving rise to differential linewidths of the individual components of a multiplet. To demonstrate how two components of the scalar coupled doublet can have different relaxation rates and therefore different spectral linewidths, I will examine cross-relaxation in the two spin system of scalar coupled spins *I* and *S* that interact through dipole-dipole mechanism and where spin *I* has a significant CSA[57, 60]. A relevant example is that of an isolated ^{1}H-^{15}N coupled spin system for which the usefulness of the TROSY effect has been established by Pervushin *et al.*[6, 53]. The interference of the dipole-dipole interaction and CSA effects the evolution of the transverse components of the density operator (I_+ and $2I_+S_z$ magnetization) as follows:

$$\frac{d\langle I_+\rangle(t)}{dt} = -i\pi J_{IS}\langle 2I_+S_z\rangle(t) - \left(\rho_{2I^+}^{DD} + \rho_{2I^+}^{CSA}\right)\langle I_+\rangle(t) - \eta_{xy}\langle 2I_+S_z\rangle(t) \qquad (1.30)$$

$$\frac{d\langle 2I_+S_z\rangle(t)}{dt} = -i\pi J_{IS}\langle I_+\rangle(t) - \left(\rho_{2I^+Sz}^{DD} + \rho_{2I^+}^{CSA}\right)\langle 2I_+S_z\rangle(t) - \eta_{xy}\langle I_+\rangle(t) \qquad (1.31)$$

where η_{xy} is the interference rate constant (see below), J_{IS} is the scalar coupling of spins *I* and *S*, and $\rho_{2I^+}^{DD}$, $\rho_{2I^+Sz}^{DD}$ and $\rho_{2I^+}^{CSA}$ are individual relaxation rates of the density operator

components I_+ and I_+S_z for the dipolar interaction (ρ^{DD}) and CSA (ρ^{CSA}), respectively, which are obtained by application of equation 1.23. The interference rate constant η_{xy} is given by:

$$\eta_{xy} = \frac{\sqrt{3}}{6} c_I d_{IS} P_2(cos\theta)\{4J(0) + 3J(\omega_I)\} \tag{1.32}$$

where $d_{IS} = -(\mu_0\gamma_I\gamma_S\hbar)/(4\pi r_{IS}^3)$, $c_I = \gamma_I(\delta_{\text{II}} - \delta_\perp)B_0/\sqrt{3}$, $J(\omega)$ is the spectral density function at a given frequency, θ is the angle between the principle axis of the CSA tensor and the internuclear vector of the dipolar coupling, and assuming an axially symmetric chemical shift tensor $\delta_{zz} = \delta_{\text{II}}$, and $\delta_{xx} = \delta_{yy} = \delta_\perp$.

Given the equations 1.30 and 1.31 and the operator basis relations ($I_+ = (I_+S_\alpha + I_+S_\beta)$, $I_+S_z = (I_+S_\alpha - I_+S_\beta)$), with an assumption that the spin I is on resonance, the interference effect on the scalar coupled doublet of the spin system[57] can be examined by considering the evolution of I^+S^α and I^+S^β :

$$\frac{d}{dt}\begin{bmatrix}\langle I^+S^\alpha\rangle(t)\\ \langle I^+S^\beta\rangle(t)\end{bmatrix} = -\begin{bmatrix} i\pi J_{IS} + \bar{R}_2 + \eta_{xy} & (\rho_{2I+}^{DD} - \rho_{2I+Sz}^{DD})/2 \\ (\rho_{2I+}^{DD} - \rho_{2I+Sz}^{DD})/2 & -i\pi J_{IS} + \bar{R}_2 - \eta_{xy}\end{bmatrix}\begin{bmatrix}\langle I^+S^\alpha\rangle(t)\\ \langle I^+S^\beta\rangle(t)\end{bmatrix} \tag{1.33}$$

where $\bar{R}_2 = (\rho_{2I^+}^{DD} + \rho_{2I^+Sz}^{DD})/2 + \rho_{2I+}^{CSA}$. If $(2\pi J_{IS}^2) \gg (\rho_{2I+}^{DD} - \rho_{2I+Sz}^{DD})^2$ the off-diagonal terms can be ignored[57] and the remaining diagonal terms reveal how the DD-CSA interference term contributes to $(\bar{R}_2 + \eta_{xy})$ or subtracts from $(\bar{R}_2 - \eta_{xy})$ the relaxation rates of the two components of the doublet that resonate on +/- J_{IS}/2 Hz. This notion provided the basis for the TROSY technique at high B_0 fields[6, 53].

Methyl-TROSY

A similar approach to that of Pervushin *et al.* has been demonstrated for methyl groups by Tugarinov *et al.*[63-64]. The former TROSY approach has focused on backbone amides and side chain aromatics and made use of the interference between dipole-dipole and CSA relaxation

mechanisms (1.3), while the latter approach considers side-chain methyl groups and focuses on cross-correlation of the dipole-dipole relaxation interactions in the $^{13}C^{1}H_3$ spin system[15, 63-65]. The spin system ($^{13}C^{1}H_3$) is considered as attached to a slow-tumbling, isotopically rotating macromolecule, and rotating rapidly around the axis of the methyl group. The chemical shift anisotropy contribution of the ^{13}C nucleus in this spin system is small and is therefore ignored[63].

For a $^{13}C^{1}H_3$ spin system, an analysis of the relaxation effects is far more complex than for *AX* spin systems due to many more transitions that need to be considered whilst analyzing the flow of coherences through the pulse sequence, e.g. eight and ten single-quantum transitions for ^{13}C and ^{1}H spins, respectively[63]. Since the $^{13}C^{1}H_3$ spin system is not weakly coupled, the tensor product basis and the basis of eigenfunctions are not the same[63], which is why the product operator formalism cannot be used and the pulse sequence needs to be examined at the level of each off-diagonal density matrix element.

In their detailed account on this subject, Ollerenshaw *et al*. described their choice of the $^{13}C^{1}H_3$ density matrix in the basis of eigenfunctions, derived the operators for evolution of chemical shifts, scalar coupling, pulses and equilibrium magnetization, and evaluated how these transform the density matrix through the pulse sequence (Figure 1.2). The density matrix could be split for each of three independent spin manifolds, with two different total spin angular momenta (one of $I = 3/2$ and two with $I = 1/2$), in which the ^{1}H part of the state space is divided[66]. The density matrix elements of one manifold do not evolve into elements of the other by applications of pulses[66] and, in the macromolecular limit with the assumption of ^{1}H in a rapidly rotating methyl group, are also not coupled through intra-methyl dipole-dipole interactions[63]. Importantly, in the chosen basis set and in the macromolecular limit, if the contributions to relaxation from external protons are ignored, the relaxation rate of each of the ^{1}H and ^{13}C single quantum transition corresponds to a single exponential. The ^{1}H, ^{13}C

single quantum, and heteronuclear double and zero quantum transitions can be separated into rapidly and slowly decaying transitions, where the rapidly decaying ones result from respectively the ^{1}H-^{1}H, ^{1}H-^{13}C, and a combination of ^{1}H-^{1}H and ^{1}H-^{13}C dipolar interactions (see Table 1.2)[63].

Table 1.2 | Relaxation rates of single quantum ($R_{2,H/C}^{f/s}$) heteronuclear zero and double quantum transitions ($R_{2,CH}^{f/s}$), with *f* and *s* indicating rapidly and slowly decaying density matrix elements, respectively. The expressions are adapted from the work of Ollerenshaw *et al.*[63]. R_{1ext} and R_{2ext} are relaxation rate constants that approximate the contributions from the interactions of ^{1}H^{13}C spin system with external spins (not taking into the account cross-correlation effects).[63] d=$(\frac{\mu_0}{4\pi})^2 \frac{S_{axis}^2 \gamma_H^2 \gamma_C^2 \hbar^2 \tau_c}{r_{HC}^6}$, S_{axis}^2 is an order parameter for the rotation of the methyl threefold axis, γ_i is the gyromagnetic ratio of spin *i*, r_{ij} is the distance between spins *i* and *j*, $\hbar$ is the reduced Planck's constant ($\hbar=\frac{h}{2\pi}$), and τ_c is an assumed isotropic rotational correlation time.

Rate	Equation
$R^{f}_{2,C}$	$\frac{1}{5}d + \frac{3}{2}R_{1,ext}$
$R^{s}_{2,C}$	$\frac{1}{45}d + \frac{3}{2}R_{1,ext}$
$R^{f}_{2,H}$	$\frac{9}{20}(\frac{\mu_0}{4\pi})^2 \frac{S_{axis}^2 \gamma_H^4 \hbar^2 \tau_c}{r_{HH}^6} + \frac{1}{45}d + R_{2,ext}$
$R^{s}_{2,H}$	$\frac{1}{45}d + R_{2,ext}$
$R^{f}_{2,CH}$	$\frac{9}{20}(\frac{\mu_0}{4\pi})^2 \frac{S_{axis}^2 \gamma_H^4 \hbar^2 \tau_c}{r_{HH}^6} + \frac{4}{45}d + R_{1,ext} + R_{2,ext}$
$R^{s}_{2,CH}$	$R_{1,ext} + R_{2,ext}$
$R_{1,ext}$	$\frac{1}{10}(\frac{\mu_0}{4\pi})^2 \sum_i \frac{\gamma_H^4 \hbar^2 \tau_c}{r_{HH,i}^6}$
$R_{2,ext}$	$\frac{8}{15}(\frac{\mu_0}{4\pi})^2 \sum_i \frac{\gamma_H^2 \gamma_D^2 \hbar^2 \tau_c}{r_{HD,i}^6} + \frac{1}{4}(\frac{\mu_0}{4\pi})^2 \sum_i \frac{\gamma_H^4 \hbar^2 \tau_c}{r_{HH,i}^6}$

It should be noted that the contributions to the relaxation from other spins (outside of the isolated ^{13}C^{1}H$_3$ spin system) are significant in fully protonated macromolecules, resulting in a much faster relaxation of the off-diagonal elements of the density matrix, and hence to decrease in a TROSY enhancement. Therefore, the effectiveness of the methyl-TROSY is

achieved on uniformly deuterated protein samples for which relaxation contributions of external ^{1}H and ^{2}H spins are minimized (Table 1.2).

^{1}H-^{13}C TROSY pulse sequence

In their search for a TROSY ^{1}H-^{13}C correlation experiment Tugarinov *et al.* showed that the standard heteronuclear multiple quantum coherence (HMQC) sequence[67] is optimized for TROSY "as is"[64] and when applied to ^{1}H-^{13}C labelled, uniformly deuterated macromolecules provides significantly higher sensitivity than its HSQC counterpart[15,63-64]. A particular advantage of the HMQC is that the sequence contains a single $\pi/2$ ^{1}H pulse in the sequence (Figure 1.2), opposed to multiple $\pi/2$ ^{1}H pulses of the HSQC pulse sequence, which mix rapidly and slowly decaying coherences at several instances and prevent any prospective TROSY effect[63].

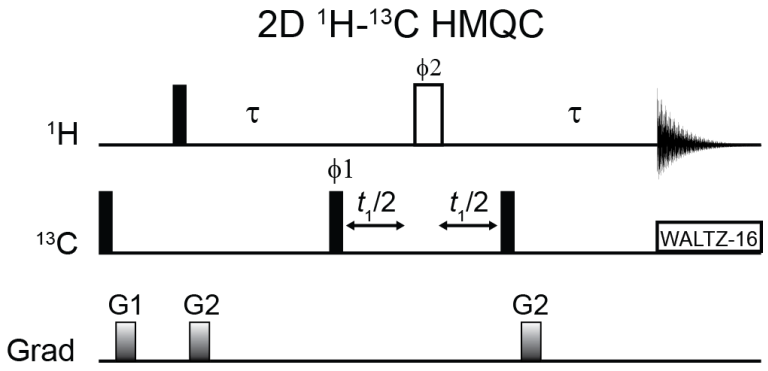


Figure 1.2 | HMQC[67] pulse sequence based on Tugarinov *et al.*[64]. All narrow ($\pi/2$) and wide (π) rectangular pulses are applied along the *x*-axis unless indicated otherwise. The phase cycle is employed as follows: ϕ_1 =(*x*,-*x*, *x*,-*x*), ϕ_2 =2(*x*), 2(*y*), 2(-*x*), 2(-*y*), rec. = *x*,-*x*,-*x*,*x*. Indicated pulse field gradients (G_1 and G_2) are applied along the *z*-axis. The time delay τ is set to $1/(2J_{HC})$, where J_{HC} ~139 Hz.

In the HMQC, after rapidly and slowly decaying ^{1}H coherences have been created by the first $\pi/2$ ^{1}H pulse, the two coherences result in separate slowly and rapidly decaying antiphase ^{1}H single quantum coherence, $^{1}H^{13}C$ multiple quantum, and ^{1}H single quantum coherences

(Table 1.2). To monitor the flow of the magnetization through the sequence and find a form of the observable, it is sufficient to use the product operator formalism as given in Figure 1.3.

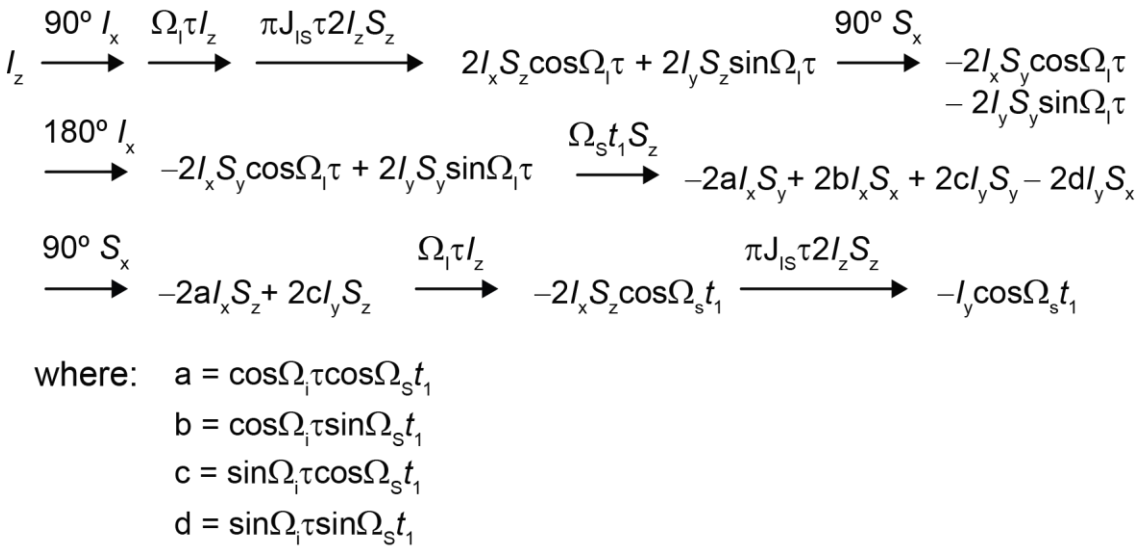


Figure 1.3 | Product operator analysis of the basic HMQC pulse sequence as given in Figure 1.2. The sequence correlates *I* (^{1}H) and *S* (^{13}C) nuclei through their *J* couplings. The representation ignores the effects of relaxation and assumes an initial state with *z*-magnetization of spin *I* (in this case ^{1}H). Note that τ is set to $1/2J_{HC}$ leading to cancelation of terms containing any variation of cos τ, which were omitted for clarity.

In a more rigorous consideration, the pulse sequence is a series of density matrix transformations by the operators that can be summarized as[63]:

$$\boldsymbol{\rho_0} \rightarrow (\frac{\pi}{2}H_x) \rightarrow \boldsymbol{\rho_1} \rightarrow (\frac{\pi}{2}2C_zH_z) \rightarrow \boldsymbol{\rho_2} \rightarrow (\frac{\pi}{2}C_x)(\omega_c t_1 C_z)(\pi H_x) \rightarrow \boldsymbol{\rho_3} \rightarrow (\frac{\pi}{2}C_x) \rightarrow \boldsymbol{\rho_4}$$

$$\rightarrow (\frac{\pi}{2}2C_zH_z) \rightarrow \boldsymbol{\rho_5}$$

A detailed reference to particular points in the transformations of the density matrix, the generated coherences, and their relaxation rates is given in work from Ollerenshaw *et al.*[63]. The sequence ultimately leads to the signal (S_{HMQC}):

$$\boldsymbol{S_{HMQC}} = 6\cos(\omega_C t_1)\{\exp\left(-[4\tau + t_2]R^f_{2,H} - t_1 R^f_{2,CH}\right)$$

$$+\exp(-[4\tau + t_2]\boldsymbol{R^s_{2,H}} - t_1\boldsymbol{R^s_{2,CH}})\}.$$

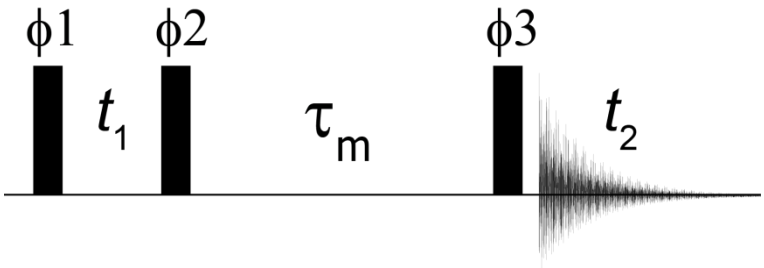
detected in a t_2 period, which is composed of two terms, one of which ($\exp(-[4\tau + t_2]R^S_{2,H} - t_1 R^S_{2,CH})$) is defined solely by slow relaxation rate constants[63].

1.4 Nuclear Overhauser effect (NOE)

NOE Spectroscopy (NOESY) remains one of the oldest and arguably most important multidimensional NMR approaches. The NOESY experiment has been widely used in biomolecular NMR spectroscopy ever since its utility was first demonstrated for assessing the proximity of nuclear spins and in protein NMR structure calculation (1.1). The experiment detects the spatial proximity of nuclear spins by measuring through space coherence transfer that originates from dipole-dipole cross-relaxation (see below). Distance restraints, derived from the measured NOEs, represent the essential restraints in classical NMR structure calculations and are also, as introduced in the following chapters, the basic restraints in the Methyl Assignment by Graph MAtching (MAGMA) strategy.

NOESY pulse sequence

The pulse sequence for a 2D NOESY experiment is given below, where 90° pulses are indicated by the black rectangles and are considered to have a phase *x* in what follows.



The first 90° pulse brings the magnetization onto the transverse plane (for phase *x*, the magnetization is rotated to *–y*). Reducing the consideration to a two-spin system, with spins of type 1 and 2, it can be shown that during time t_1, magnetization evolves to $-\cos\Omega_1 t_1 I_{1y} + \sin\Omega_1 t_1 I_{1x}$. Upon application of the second 90° pulse, the first term ($-\cos\Omega_1 t_1 I_{1y}$) is rotated

to the *z*-axis $-\cos\Omega_1 t_1 I_{1z}$; since the pulse is of phase *x*, it will not affect the second term $\sin\Omega_1 t_1 I_{1x}$.

Over the course of the mixing time (τ_m) spin 1 may exchange magnetization through cross-relaxation with spin 2 if they are sufficiently close in space (see below), resulting in the presence of a $\cos\Omega_1 t_1 I_{2z}$ term at the end of the mixing time. The final 90° pulse rotates both terms to the transverse plane, and they evolve under their respective chemical shifts during the time t_2 to finally give four terms: $cos\Omega_1\ t_2\ cos\Omega_1\ t_1 I_{1y}$, $sin\Omega_1 t_2 I_{1y}\ sin\Omega_1\ t_1 I_{1x}$, $cos\Omega_2 t_2\ cos\Omega_1\ t_1 I_{2y}$, $sin\Omega_2 t_2\ cos\Omega_1\ t_1 I_{2x}$. Note that the I_{1y} and I_{1x} terms are scaled with (1-*w*) and I_{2y} and I_{2x} with *w*, where *w* stands for a fraction of the spins of type 1 that have exchanged their magnetization with the spins of type 2 during the mixing period. Evolution during the t_1 period allows frequency labelling ($cos\Omega_1$) and results in a modulation of the amplitude of the signal measured during t_2. Ultimately, Fourier transformation of the two cosine terms ($cos\Omega_1 t_2$ and $cos\Omega_2 t_2$) in the *F2* dimension gives absorptive line signals centered at the Ω_1 and Ω_2 position. In the presented analysis, no *J*-coupling between the spins 1 and 2 was assumed, so that the two spins are considered close in space but not connected through a bond. This notion is relevant for consideration of the NOE measurements between protons of two different methyl groups.

Measurement of the through space proximity of methyl groups in uniformly deuterated, ^{13}C-methyl labelled, and methyl protonated high molecular mass proteins can be achieved if the NOESY sequence[68] is embedded between the two HMQC components. Such a combined sequence benefits from the methyl-TROSY enhancement (see above) and allows for the collection of high quality 3D/4D methyl-methyl NOESY spectra.

Here, I consider the 4D ^{1}H-^{13}C HMQC-NOESY-HMQC pulse sequence, which has advantageous resolution and sensitivity compared to the alternative 3D/4D NOESY

sequences for methyl groups[69] (Figure 1.4). In the first HMQC component of the sequence, considerations are given to the evolution of ^{1}H magnetization during the t_1 period such that the total acquisition time is minimized (Figure 1.4 A), and the reduced relaxation rates of methyl multiple quantum coherences can be exploited (Figure 1.4 B) (see Tugarinov *et al.*[69]). This HMQC is attached to a standard component of the NOESY sequence ($90°_x - t_m$ - $90°_x$, see above), followed by the second HMQC component, identical to the sequence analyzed above.

A general flow of the magnetization through this pulse sequence can be coarsely described as:

$$^{1}H_i\ (J_{CH}) \rightarrow\ ^{13}C_i\ (t_2, J_{CH}) \rightarrow\ ^{1}H_i\ (NOE) \rightarrow\ ^{1}H_j\ (J_{CH}) \rightarrow\ ^{13}C_j\ (t_3, J_{CH}) \rightarrow\ ^{1}H_j$$

The description shows that the methyl protons between which the NOE is measured ($^{1}H_i$, $^{1}H_j$) are *J* coupled to ^{13}C in their respective $^{1}H^{13}C$ systems (Figure 1.4), which allows for correlating ^{1}H nuclei to their respective ^{13}C nucleus, the chemical shift of which is recorded during the t_2 and t_3 periods. This allows for construction of a 4D spectrum in which each ^{1}H-^{13}C plane features diagonal (auto-peak) and a number of off-diagonal (cross-peaks) signals with both ^{1}H and ^{13}C frequency labels.

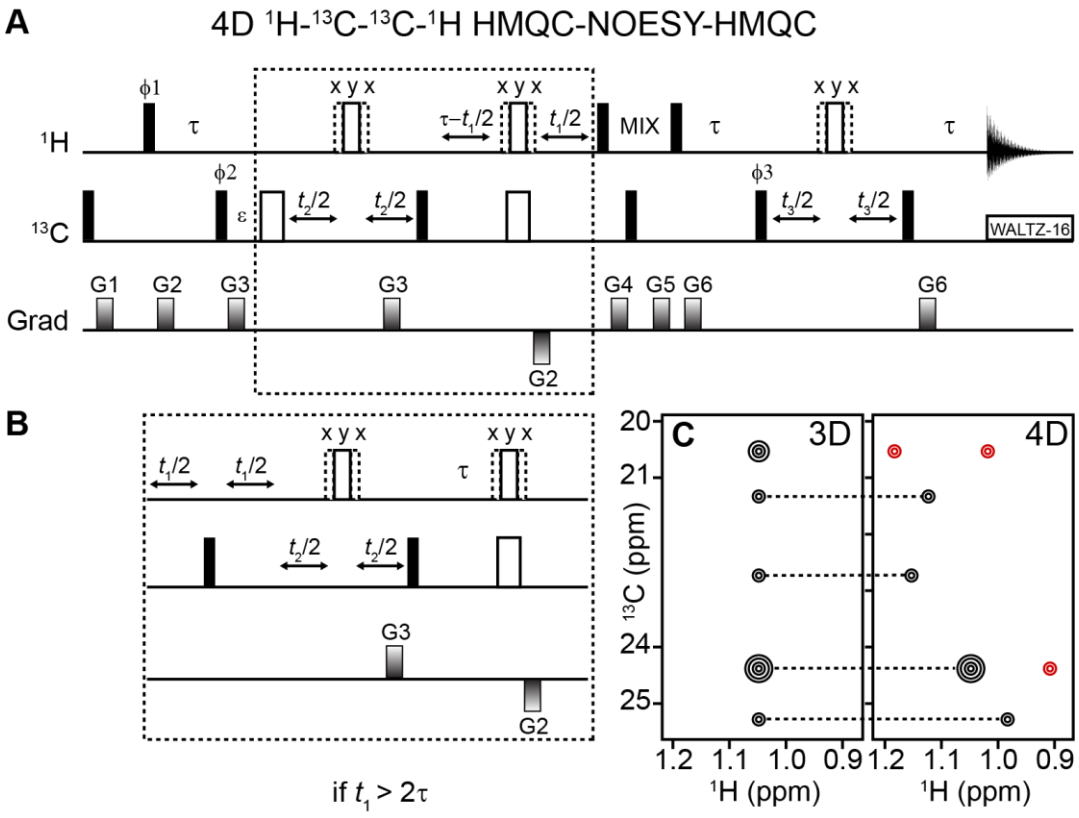


Figure 1.4 | A Methyl-methyl 4D HMQC-NOESY-HMQC pulse sequence based on that from Tugarinov *et al.* wherein more details can be found[69]. All narrow (π/2) and wide (π) rectangular pulses are applied along the *x*-axis unless indicated otherwise. The phase cycle is employed as follows: ϕ_1 =(*x*), ϕ_2 =(*x*,-*x*), ϕ_3 =2(*x*), 2(-*x*), rec. =*x*,-*x*,-*x*, *x*. Pulse field gradients (G_1-G_6) are applied along the *z*-axis. The time delay τ is set to $1/(2J_{HC})$, where J_{HC} ~139 Hz. The element highlighted in the rectangle is employed for $t_1 \leq 2\tau$, and substituted with **B** for $t_1 > 2\tau$ (for details see Tugarinov *et al.*)[69]. **C** Illustration of the advantage of an additional ^{1}H dimension obtained with the 4D (right) compared to the 3D (left) NOE data set. Hypothetical slices of the 3D/4D NOE data sets are shown. The peaks labeled in red resonate at the same/similar ^{13}C frequencies and are overlapped in the 3D spectrum. Improved resolution in the 4D experiment[69]allows their unambiguous assignment, which is also aided by an additional ^{1}H dimension.

Origin of the NOE effect

Since the description of the HMQC sequence was already provided, I will here focus on the events during the NOE mixing time (Figure 1.4, *MIX*) in which magnetization transfer occurs through cross-relaxation between the protons of proximal methyl groups. For ease of explanation, a homonuclear two spin system *I S* is considered, in which spins *I* and *S* are close in space but not *J*-coupled. This system is a considerable simplification of the spin system of two methyl groups, however it is sufficient to cover the basics of the effect. The

energy-level diagram of such a system is given in Figure 1.5. How the *z*-magnetization transfer occurs during the NOE mixing period, can be explained at the most basic level using Solomon's equations (classical treatment)[22].

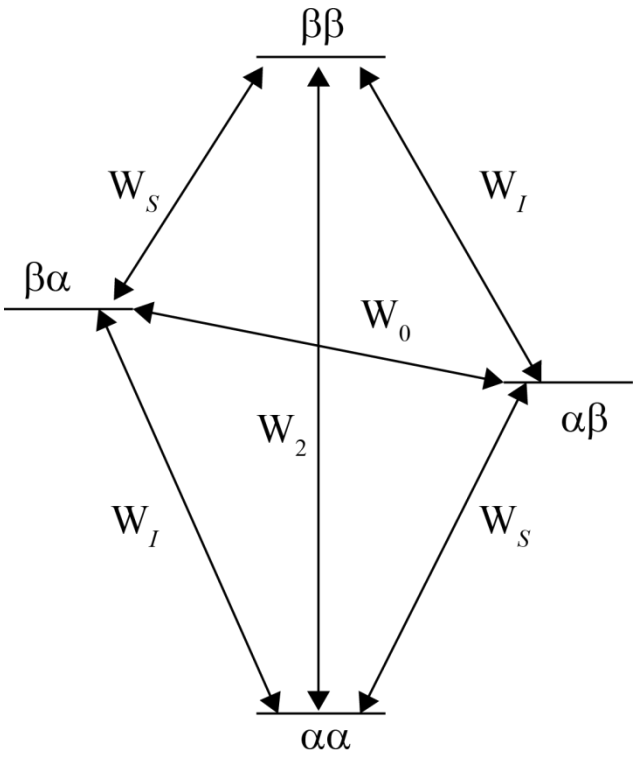


Figure 1.5 | Energy level diagram associated with a spin system *I S* with indicated transitions (cross-relaxation pathways) and their rate constants. The rates are labelled with W_I and W_S for single quantum transitions (Δm=+/-1), which flip the α and β states of spin *I* and *S*, respectively. The zero and double quantum transitions W_0 and W_2 are indicated only with the change in the magnetic quantum number *m* (Δm = 0, Δm = +/-2).

During the mixing period, dipolar interactions between the spins *I* and *S* cause transitions between the energy levels as denoted in the diagram (Figure 1.5). Deviations in the populations from equilibrium are given by differential equations, for example for the spin state |αα>:

$$\frac{d\Delta P_{\alpha\alpha}}{dt} = -(W_I + W_s + W_2)\Delta P_{\alpha\alpha} + W_I\Delta P_{\beta\alpha} + W_s\Delta P_{\alpha\beta} + W_2\Delta P_{\beta\beta} \qquad (1.34)$$

where $i,j = \{\alpha, \beta\}$ $\Delta P_{ij} = P_{ij} + P_{ij}^0$, where P_{ij} is the population of the state |*ij*> and P_{ij}^0 is the equilibrium value of P_{ij}. These differential equations can then be written out for the remaining three states (|αβ>, |βα> and |ββ>). Equation (1.10) defines the expectation value of the operator as the trace of the product of the operator and the density matrix, which for $2\langle\widehat{Iz}\rangle(t) = 2Tr(\hat{\rho}(t)\widehat{Iz})$ equals $P_{\alpha\alpha} + P_{\alpha\beta} - P_{\beta\alpha} - P_{\beta\beta}$, and for $2\langle\widehat{Sz}\rangle(t) = 2Tr(\hat{\rho}(t)\widehat{Sz})$

equals $P_{\alpha\alpha} - P_{\alpha\beta} + P_{\beta\alpha} - P_{\beta\beta}$. Rearrangement of the terms leads to two coupled rate equations:

$$\frac{d\Delta I_z}{dt} = -\rho_I \Delta I_z(t) - \sigma_{IS} \Delta S_z(t) \tag{1.35}$$

$$\frac{d\Delta S_z}{dt} = -\rho_S \Delta S_z(t) - \sigma_{IS} \Delta I_z(t) \tag{1.36}$$

where $\rho_I = W_0 + 2W_I + W_2$, $\rho_S = W_0 + 2W_S + W_2$, and $\sigma_{IS} = W_2 - W_0$, and $\Delta X_z(t) = \langle \widehat{Xz} \rangle(t) - \langle X\widehat{z^0} \rangle$, and $\langle X\widehat{z^0} \rangle$ is the magnitude of the *Xz* operator at equilibrium(X = {*I*,*S*}). These coupled equations show the temporal dependence of the *Iz* / *Sz* magnetization evolution on both the auto-relaxation rate constant ($\rho_{S/I}$) and the cross-relaxation rate σ_{IS}. In addition, they indicate that in order for magnetization transfer to occur between the spins *I* and *S*, the cross-relaxation rate must be nonzero (*i.e.* $W_2 - W_0 \neq 0$).

It should be noted that the rates (*Wx, x=0, 2, I, S*) considered in this simplified explanation cannot be related to any particular stochastic Hamiltonian[57], which for a more rigorous approach to the NOE, a complete quantum-mechanical approach, or a quantum-mechanical approach in combination with the classical lattice treatment has to be used (see 1.2). For quantitative evaluation of the NOE, such treatment is essential, and a detailed account on this topic can be found in work from Vögeli[70]. In that regard, in section 1.2 (*dipole-dipole interaction*) I have briefly introduced the stochastic dipolar Hamiltonian (1.22) and showed how its insertion in the master equation, reveals the $1/r^6$ dependence of the dipole-dipole relaxation efficiency (1.23).

Practical aspects of NOE measurements

As indicated above, the efficiency of the *z*-magnetization transfer depends on the length of the NOE mixing time. Given the theoretical considerations, a mixing time of the order of $1/R_1$ should maximize the NOE cross-peak build-up. In practice, the optimal NOE mixing time for a system is usually determined experimentally, through the measurement of the buildup of

NOE cross-peak intensities over a range of mixing times[69]. For uniformly deuterated methyl-protonated proteins, the optimal NOE mixing time can be on the order of hundreds of milliseconds due to the reduced spin-diffusion effects in deuterated versus protonated systems[128-129]. However, the length of the mixing time in practice does not approach the theoretical limit (i.e. $1/R_1$). Even though cross-relaxation continues to build up over the course of the mixing time, auto-relaxation will ultimately decrease the signal, as the cross-peak intensities are divided by their corresponding diagonal (auto) peak intensities.

Extending the mixing time can lead to an above-mentioned effect known as *spin-diffusion.* Spin-diffusion causes cross-peaks to build-up between otherwise distant spins, but reduces the intensity of the cross-peaks measured between the proximal spins (see below). Due to this effect, in a three-spin system (*i, j, k*) in which spins *i* and *k* are proximal and *i* and *j* are spatially distant, and hence unlikely to have sufficient dipolar coupling for a direct magnetization transfer, an NOE between *i* and *j* can still be measured. The reason for this can be found through an analysis of the relaxation rate matrix for this system[57]. Calculation of the time dependence of the spin *i* magnetization shows that, next to the term that defines magnetization transfer between *i* and *k* ($-\sigma_{ik}\tau_{\mathrm{m}} < I_{kz} > (0)$), an additional second-order term ($-\frac{1}{2}\sigma_{ik}\sigma_{ij}\tau_m^2 \ < I_{jz} > (0)$) also plays a role in the NOE build-up[57]. To eliminate effects from spin diffusion, considerations should be given to shorter mixing times, which unfortunately result in a lower NOESY sensitivity[71]. For methyl-methyl NOE measurements, combining the NOESY with the HMQC sequence is particularly advantageous due to the TROSY enhancement. Using a 4D HMQC-NOESY-HMQC sequence (Figure 1.4) in combination with a non-uniform sampling schedule results in significant gains in both resolution and sensitivity[69].

Relating the measured NOE cross-peak intensities to inter-atomic distances is not a simple task[130]. If the initial magnetization rates can be measured and if one can assume that motion

of the spin system is simply isotropic, the relative magnitude of an NOE is proportional to $\frac{1}{r^6}$, where *r* stands for the distance between the two interacting spins. In practice however, differential internal motions in different parts of a protein, low signal to noise ratios, and the inability to precisely measure the time dependence of all NOE build-ups prevent applications of the simple r^{-6} relation[98].

To estimate the range of distances that can be sampled by methyl-methyl NOEs in highly deuterated proteins, Sounier *et al*[128] measured NOEs in 98% deuterated, Ileδ_1-protonated γS-crystallin and ubiquitin and related them to ^1H–^1H inter-methyl distances of up to 12 Å. Reduced effects of spin diffusion in the highly deuterated samples allowed long NOE mixing times and extraction of the distance restraints even from weak cross-peaks. To obtain better precision in the distance calculation, the authors used a full-relaxation-matrix analysis, determining the cross-relaxation rate directly from the measured diagonal and cross-peak volumes, and assuming fast rotation of the methyl group attached to an isotropically tumbling macromolecule[128]. Kerfah *et al.*[129] subsequently applied this approach to a considerably larger system, the U-[H^2], Ala-[Cβ], Ile-[Cδ1] and $(LeuVal)^{proS}$ -protonated malate synthase G (MSG, ~81 kDa). The authors showed that in the distance range from 2.5 to 8.5 Å, 96% of the NOEs, expected based on the crystal structure, could be observed for this system. In addition, ~50% of the NOEs in the distance range from 9.5 to 10.5 Å could be measured[129]. To improve accuracy in those calculations, considerations could be given to spin-diffusion contributions due to very long mixing times employed[129]. This could perhaps be achieved with an 'exact-NOE' approach, which has already been successfully applied to $^1H^N$-$^1H^N$ NOEs[130].

1.5 Methyl-TROSY: Extending the size limit of NMR spectroscopy

For several reasons, methyl groups are particularly attractive probes for high-resolution biomolecular NMR studies. First, methyl-bearing residues are well represented in the primary sequences of proteins and are found both in the hydrophobic cores[72] and at protein surfaces (e.g. Ala, Met, Thr). Peripheral localization of methyl groups in the side chains of amino acids accounts for their enhanced dynamics, leading to slower relaxation that is favorable for investigations of large systems. Furthermore, in the macromolecular limit, the TROSY enhancement can be obtained by exploitation of destructive interference of dipolar interactions in the $^{13}C^{1}H_3$ spin system (see above). Rapid rotation around the methyl symmetry axis renders the three ^{1}H spins degenerate, leading to enhanced sensitivity, whilst good dispersion of ^{1}H-^{13}C correlation maps allows for optimal resolution.

Methyl labelling strategies

To use methyl groups as NMR probes in studies of large systems, selectively protonated, ^{13}C-labeled methyl groups need to be introduced to an otherwise deuterated, ^{12}C-labeled protein background (Figure 1.6). In the initial methyl labeling strategies ^{1}H, ^{13}C, ^{15}N-labeled isoleucine, leucine, and valine[73], were added to the minimal media that was used for the preparation of [U-^{2}H, ^{15}N]-labeled protein in *E.coli*-based protein production. Alternatively, adding ^{1}H, ^{13}C-labelled pyruvate as the main carbon source to such media was explored, leading to successful protonation of 40-80% of the methyl sites in proteins[74].

Both early labelling strategies had drawbacks, either due to the requirement of expensive isotopically labelled amino-acids[73] or due to the generation of a significant amount of methyl isotopomers (i.e. $^{13}CH_3$, $^{13}CH_2D$, $^{13}CHD_2$)[74]. Ever since these initial efforts, cost-effective and robust methyl labelling protocols for all methyl-bearing residues have been established, with significant improvements in labelling efficiency[75].

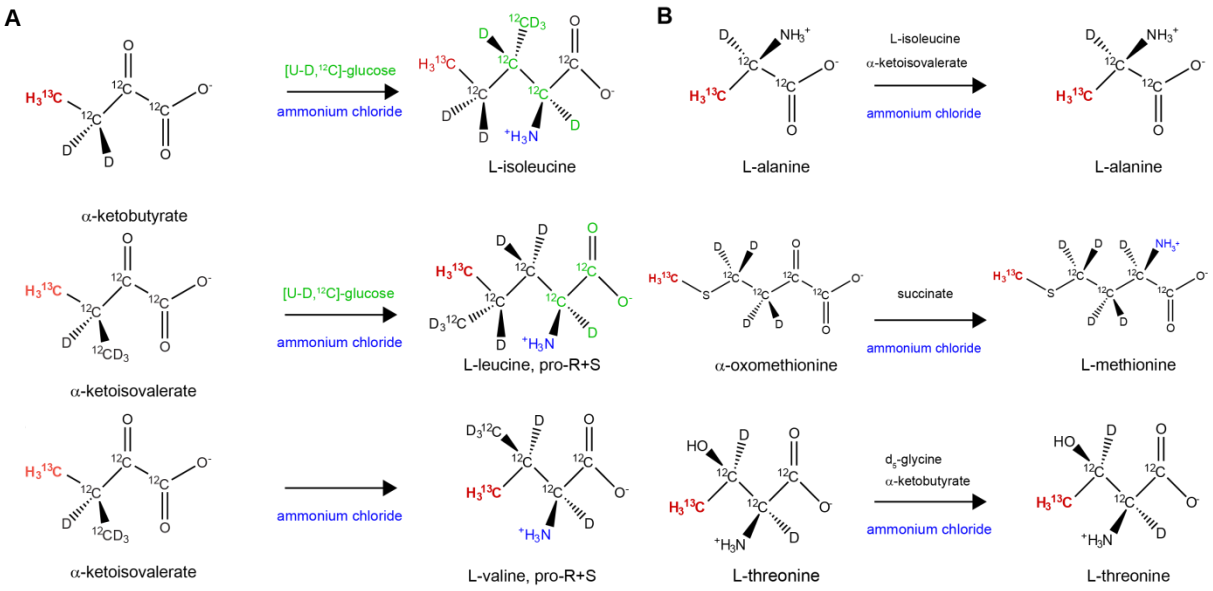


Figure 1.6| Isotopically labelled biosynthetic precursors that contained $^{13}CH_3$ groups (red) used for selective methyl labelling in the *E.coli*-based production of [U-^{2}H,^{13}C]-ILV/A/M/T labelled proteins. The nuclei that are labeled with glucose and ammonium chloride are labeled green and blue, respectively.

Recently, labelling of particular amino acids has been realized by using specific biosynthetic precursors, such as (3, 3-H_2) ^{13}C-2-ketobutyrate (α-ketobutyrate) for Ile-δ_1[56], or (3-H_2) ^{13}C-2-ketoisovalerate (α-ketoisovalerate), a shared metabolic precursor for Leu-δ_1,δ_2 and Val-γ_1,γ_2[76], which removed the problem of isotopomers. Overlap in the chemical shift ranges of Leu-δ_1,δ_2 and Val-γ_1,γ_2 groups motivated the establishment of selective labelling strategies for either of the residue types. Lichtenecker *et al.* showed that by using labeled α-ketoisocaproate, selective Leu labelling can be achieved[77], whereas the combined addition of unlabeled α-ketoisocaproate and labeled α-ketoisovalerate leads to exclusive labeling of Val[78]. To avoid a racemic mixture of pro-*R* and pro-*S* labeled Leu-δ_1,δ_2 and Val-γ_1,γ_2 groups, which results from the use of α-ketoisovalerate, 2-hydroxy-2-[^{13}C]methyl-3-oxo-4-[$^{2}H_3$]-butanoic acid can be used for selective labelling of the pro-*S* groups[79]. In addition, selective labelling of pro-*R* or pro-*S* groups of either Val or Leu residues can be achieved by cell-free synthesis with the SAIL method[80] or by using *E. coli* auxotrophs[81]. Such stereospecific

labelling significantly reduces spectral complexity and enhances signal-to-noise ratios, which is particularly useful for studies of large systems.

Finally, biosynthetic strategies for selective labeling of the Ile-γ_2 group[82], Ala-β[83-84], Met-ε[85-86], and Thr-γ[75, 87] have been established, and recently, a method for simultaneous labelling of all methyl bearing residues (ILVAMT) by combining all biosynthetic precursors in a single protein production medium has been introduced[88]. A schematic depicting the biosynthetic precursors that can be used to label methyl-bearing residues is shown in Figure 1.6.

Methyl-TROSY applications

Since its initial development in the early 2000s, there have been numerous applications of methyl-TROSY. The interested reader is directed to more comprehensive reviews by Rosenzweig *et al.*[13], Ruschak *et al*[12], and Tugarinov *et al.*[15]. Here, I will highlight only a few applications of methyl-TROSY that feature the wide range of dynamical timescales that are accessible to methyl probes (Figure 1.7).

Methyl-TROSY was employed by Sprangers and Kay to investigate the 670 kDa 20S core particle (CP) proteasome complex from *Thermoplasma acidophilum*[89]. The authors produced uniformly ^{2}H, Ile[$\delta1^{13}CH_3$], Leu, Val-[$^{13}CH_3$,$^{12}CD_3$]-labeled α-subunits of the 20S CP (see below). It was known that archaeal 20S CP proteasomes were comprised of concentric, stacked rings that formed a hollow cylinder. The outer (α) and inner (β) rings contain seven subunits, and the N-termini of α-subunits were known to close the entrance to the channel in the yeast and bovine proteasomes, thus functioning as "gates" at the ends of the 20S CP proteasome. Assignment of the Ile, Leu, and Val methyl resonances to the appropriate residues was obtained through a "divide-and-conquer" approach (see Chapter 3), supplemented with inspection of the NOE connectivity pattern with reference to the protein X-ray structure. The methyl assignments allowed for site-specific characterization of dynamics across a range of time-scales. It could be confirmed that the N-termini of the α-

subunits, dynamic on the millisecond timescale, could fold into the antechamber to form the flexible gate of the archaeal proteasome. In addition, a previously unknown region of highly dynamic residues (pico- to nanosecond timescale), located at the interface between the α-subunits, was identified. Furthermore, concerted motions on the millisecond timescale were observed on the inside surface of the antechamber, which links the entrance pore and the catalytic chamber. These dynamics potentially serve functional roles, as they localize to the area where substrates are shuttled for proteolysis[89].

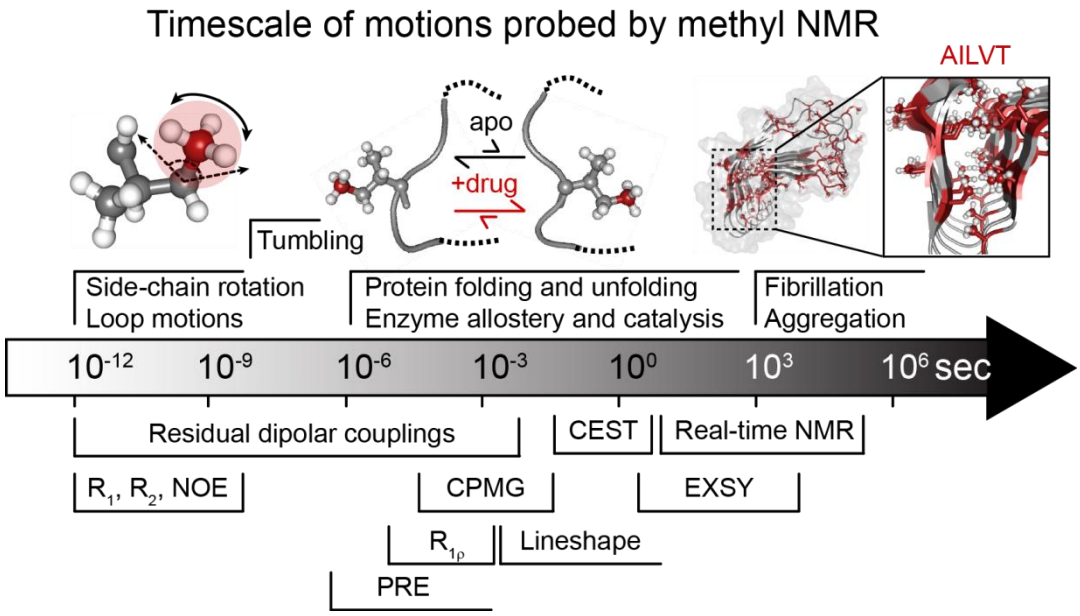


Figure 1.7 | Schematic depicting experimentally accessible times for biomolecular dynamics measurements with methyl-based NMR spectroscopy. The central arrow depicts a timescale spanning picoseconds (10^{-12}) to hours (10^{6}). Different NMR experiments can be used to characterize a particular timescale within this range, as indicated.

In a subsequent study, Religa *et al.*[90] reported that the aforementioned dynamic N-terminal residues of the 20S CP adopt inside or outside facing conformation, and can thereby control substrate access to the antechamber by opening or closing the gate. Paramagnetic relaxation enhancement (PRE) and exchange spectroscopy (EXSY) measurements were used to investigate Met-ε methyl dynamics and characterize interconversion between the multiple exchanging (two "in" and one "out") states of the N-termini of the α-subunits. In addition, by inspecting methyl chemical shift perturbations induced by the presence of the 20S CP

proteasome-activating particle (11S RP), Ruschak and Kay concluded that the 20S CP adopts a set of distinct conformations in solution and that their equilibrium is modulated by the binding of the 11S RP[91]. Moreover, by mapping the measured methyl chemical shift changes onto the structure, the authors identified a continuous region of residues that allosterically link 11S RP binding with CP active sites[91]. This study provides an example of how methyl-TROSY NMR spectroscopy can be used to provide unprecedented structural and dynamical insight into the mechanism of a 670 kDa protein complex.

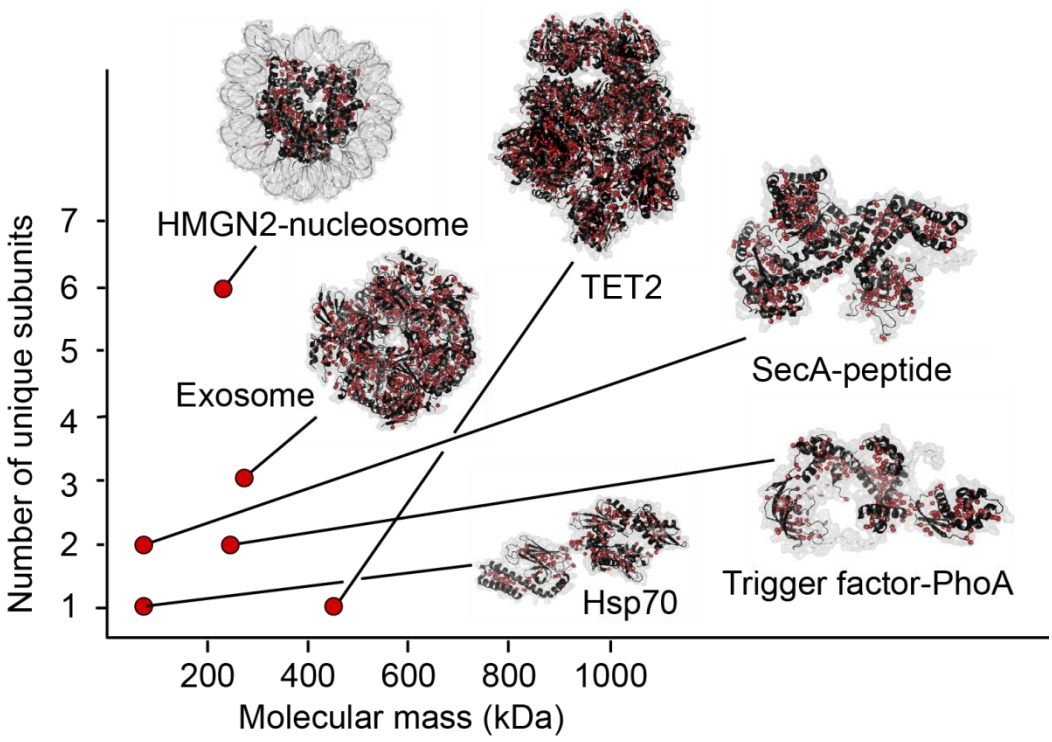


Figure 1.8 | A small subset of proteins that have been studied by methyl-TROSY. Surface and secondary structure representations of several proteins that have been studied by methyl-TROSY are displayed, with links to their molecular masses. Amongst others, the examples include monomeric, single polypeptide chains (Hsp70), homo-dodecamer (TET2), and heteroligomer with six unique subunits (HMGN2-nucleosome).

The 20S proteasome CP served as a model system for the development of additional methyl-based NMR measurements, including measurements on methyl axis orientations using residual dipolar coupling (RDC)[92-93], chemical exchange saturation transfer (CEST)

measurements of transiently populated states, and determination of side-chain rotamer conformations[94].

In another application of methyl-TROSY, Gelis *et al.* investigated SecA, a 204 kDa chaperone (Figure 1.8) involved in transporting newly synthesized proteins across bacterial membranes[95]. It was previously known that the chaperone recognizes specific signal sequences of the substrate; it was unknown how signal sequences interact with the signal recognition particle (SRP) or SecA. The ILVM labelling was employed and the methyl resonances were assigned through "divide-and-conquer" approach, wherein resonances were initially assigned to domains and fragments of contiguous domains of SecA, and subsequently transferred to the spectrum of the 204 kDa complex. The methyl resonance assignments of the full-length protein, in combination with the X-ray structure of apo-SecA, enabled identification of a groove at the interface of two domains of SecA that binds signal peptides of diverse sequences. This finding established that SecA exploits a unique binding site to recognize a variety of signal sequences, information which could not have been determined using X-ray crystallography[95].

These and many other not discussed applications[13,52] demonstrate the usefulness of methyl-TROSY in addressing complex biological questions. The prerequisite for such insights is obtaining methyl resonance assignment, which is often a laborious and time-consuming part of the procedure. Automation of the methyl resonance assignment would therefore ease and widen the applications of this powerful methodology.

1.6 Automatic NMR resonance assignment

NMR spectroscopy as applied to biological macromolecules has played a significant role in biophysical chemistry, drug design, structural biology, structural genomics, and interactomics for decades[96,97]. For biomolecular NMR spectroscopy to be more readily applied and continuously useful in these and other fields, it needs to become comparable to more mature and highly automated structural biology techniques such as X-ray crystallography. Alongside automation, NMR data analyses at all levels need to be standardized, from data acquisition and processing to resonance assignment, structure calculation, and data fitting for the interpretation of NMR relaxation experiments.

Fortunately, the information content of NMR spectra and the interactive process of their analysis are particularly amenable to automation, as most of the key steps performed by an expert can in principle be completed more efficiently, accurately, and rapidly by computers[96-102]. Of all the different manual steps involved in the study of a macromolecule by NMR, resonance assignment, or the attribution of observed NMR signals to atoms in the molecule of interest, can be particularly laborious and time consuming, especially as the sizes of the investigated systems keep increasing[101].

Next to its practical importance, reliable automation of resonance assignments constitutes a particularly rewarding point in the computerization of NMR spectroscopy, as accurate NMR resonance assignments provide insight into proteins' secondary structures, form the basis for the interpretation of macromolecular dynamics, allow the interpretation of intermolecular interaction studies, and are a prerequisite for NMR structure calculations[97, 100-103]. As such, the automation of resonance assignments has been a continuous pursuit in the field and has seen numerous different approaches, as reviewed by Moseley and Montelione[96], Baran *et al.*[99], Guntert[97], and Guerry and Herrmann[101]. Most of the available software require high-quality chemical shift peak lists, which are often manually curated by an experienced

spectroscopist. Community-wide efforts in automation and the standardization of data collection, referencing and spectral processing, peak picking, inter-spectral resonance registering, and peak list editing should remove this requirement for the next generation of automatic NMR resonance assignment methods.

In their review on the topic of automatic NMR resonance assignment in biomolecular NMR spectroscopy, Guerry and Hermann[101] identified a total of 44 research publications documenting novel automatic assignment approaches developed through the late 1990s and the first decade of 2000s. Of those, only five distinct algorithms automate the assignments of both backbone and side-chain resonances exclusively based on the input peak lists[102].

Automatic backbone resonance assignment approaches

Most of the currently available software for the automatic assignment of NMR resonances focuses on resonances from residues in the protein backbone (H^N, N, CO, CA) and side-chain CB resonances. The reason for this prevalence is the ability to utilize data from a set of through-bond NMR spectra specifically recorded to provide connectivities between protein backbone and side-chain atoms of the neighboring residues (e.g. spectra from HNCO, HN(CA)CO, HNCA, HN(CA)CO, HNCACB, HN(CO)CACB experiments). Given that the manual assignment strategies essentially rely on the same information, the through-bond automation approaches focus on mimicking the resonance-linking decisions made by an experienced user, followed by heuristics to determine an optimal mapping of the sequential resonance fragments to the primary sequence of a protein.

Table 1.3 lists generic steps typically encountered in such automatic backbone assignment protocols. Differences between various protocols are usually encountered in the implementation of the listed steps and in the requirements for input restraints. The classification of the assignment protocols is based on the approach used in the final mapping step. Commonly employed algorithmic approaches for this step include: exhaustive search

algorithms, Monte Carlo search with a simulated annealing optimizer, genetic algorithms, best-first search, and a heuristic comparison of experimental and predicted chemical shift values of homologous proteins[99, 101].

Table 1.3 | Generic steps involved in automatic NMR resonance assignment strategies, defined based on the classical strategy outlined by the Wüthrich laboratory[27,96,99].

Step	Description
1) Filtering, registering, and aligning	Filter peaks and relate resonances from different spectra (registering/aligning)
2) Grouping	Group resonances into spin systems
3) Typing	Determine amino acid types of spin systems
4) Linking	Identify and link sequential spin systems into longer segments
5) Mapping	Map these segments onto protein's primary sequence

Side-chain resonance assignment

Only a small fraction of the algorithms focused on automating backbone resonance assignments also take into the account resonances of the side-chain atoms, which is reflected in the lack of generic and robust approaches to side-chain resonance automation[99, 101-102]. The classical NOE-based NMR structure calculation method relies extensively on side-chain assignments, and thus finding approaches for reliable automation of the side-chain assignments remains an important task[102].

Existing approaches either deal with both backbone and side-chain resonances simultaneously, by including the side-chain atoms to the initially defined spin systems (Table 1.3, step 2), or treat the side-chain resonances separately upon completion of the backbone resonance assignments[101]. The main difficulty for the automation of the side-chain resonance assignment is the commonly encountered incompleteness of the experimental peak lists. This is a result of the differences in TOCSY transfer efficiencies along different amino-acid side-chains, and the insensitivity of the experiment, particularly when applied to high molecular mass targets due to fast relaxation. Some of the experimental hurdles could be overcome by combining HCC(CO)NH-TOCSY with multiple mixing times, HCCH-COSY, and HCCH-

TOCSY experiments to provide better quality data as input for automatic assignment algorithms[99].

Of the more recent efforts to provide a generic and robust protocol for both backbone and side-chain resonance assignments, it is worth mentioning the FLYA resonance assignment algorithm[102]. FLYA supports the input of experimental peak lists from a variety of through-bond J-coupling-based experiments for the assignment of protein backbones, through-space ^{13}C- and ^{15}N-resolved NOESY spectra, and side-chain TOCSY and COSY spectra. The algorithm compares the network of expected peaks obtained based on the protein sequence and the available experimental chemical shifts to the network of measured chemical shifts. To match the chemical shifts, FLYA employs approximate optimization, an evolutionary algorithm that operates on a population of assignment solutions (individuals) governed by two scoring functions that take into the account the fitness of the assignment of individual atoms (local) and of the complete assignment solution (global)[102]. FLYA is incorporated in the NMR structure calculation program CYANA.

Structure based resonance assignment

Next to the *J*-coupling based approaches that look to establish a sequential walk along atoms of the protein sequence, another category of automatic resonance assignment strategies has emerged. Structure-based NMR resonance assignment constitutes a class of methods developed as a response to the availability of high-resolution 3D structures for many protein targets of interest. The structure-based strategies reverse the more common order of operation, i.e. from resonance assignments to NMR protein structure, and instead use a protein's structure in conjunction with sparse NMR data to infer upon the assignment. An assumption taken by these approaches is that the average structure of the protein in solution is equivalent to its structure in the crystalline state[98].

A general feature of the structure-based strategies is the reliance on sources of information other than the *J*-couplings[104]. Typically employed restraints are NOEs (most often $^{1}H^{N}$ - $^{1}H^{N}$), residual dipolar couplings (RDCs), and chemical shifts (CS), although occasionally these programs use any combination of these restraints and $^{1}H^{\alpha}$ - $^{1}H^{N}$ NOEs, ^{1}H-^{1}H NOEs, solvent accessibility data (HD-exchange), PREs, and TOCSY data[104]. These NMR data can in different ways be directly compared to the available 3D structure; however, the data are typically sparse and ambiguous. This feature of experimental data prevents a unique assignment solution to be found and demands high sampling rates, which is why structure-based assignment strategies are rarely applied to large proteins (> 200 amino-acids)[104].

To this end, different structure-based assignment approaches have been explored for backbone and side-chain methyl assignments, typically differing in the input restraints they support and algorithmic approach they utilize in sampling. In Table 1.4, a few representative and conceptually interesting approaches are presented that feature similarities to the strategy employed by Methyl Assignment by Graph MAtching (MAGMA), the central topic of this thesis (Chapters 2 and 3). A more complete list of structure-based NMR assignment strategies includes the works from Dobson *et al.*[98], Bartels *et al.*[105], Gronwald *et al.*[106], Bailey-Kellog *et al.*[107], Pristovsek *et al.*[108], Pristovsek and Franzoni[109], Hus *et al.*[110], Erdmann and Rule[111], Pintacuda *et al.*[112], Langmead *et al.*[113], Langmead and Donald[114], Apaydin *et al.*[115], Xiong and Bailey-Kellog[116], Xiong et al.[117], Stratmann et al.[104,124], Venditti *et al.*[118], Chao *et al.*[119-120], Xu *et al.*[121-122], and Jung and Zweckstetter[123].

Table 1.4 | Selected structure-based NMR resonance assignment protocols use different restraints and algorithmic approaches, and share similarities with the structure-based assignment strategy employed by MAGMA (Chapters 2 and 3).

Protocol (reference)/authors	Restraints used	Algorithmic approach
Dobson, Howarth, and Redfield[98]	^{1}H-^{1}H NOE, *J*-coupling data	exhaustive search
Nuclear vector replacement[113,114]	unambiguous $^{1}H^{N}$-$^{1}H^{N}$ NOEs, $^{1}H^{N}$, ^{15}N chemical shifts, two sets of ^{15}N-^{1}H RDCs (from different alignment media), H/D exchange data	constrained bipartite matching
Contact replacement[116,117]	$^{1}H^{N}$-$^{1}H^{N}$ NOEs (with intensities), $^{1}H^{N}$ and ^{15}N chemical shifts, HNHA and TOCSY spectra	subgraph monomorphism (approximate)
PEPMORPH[111]	$^{1}H^{N}$-$^{1}H^{N}$ NOEs (with intensities from which distances are computed), RDCs	matching graph polytopes (approximate)
NOEnet[104,124]	unambiguous $^{1}H^{N}$-$^{1}H^{N}$ NOEs (with optional use of intensities), optional: CS-data and RDC-data	subgraph monomorphism

Amongst the earliest predecessors of the NOESY-based approaches for structure-based assignments is that of Dobson and Redfield[98], which utilized an extensive set of NOEs measured between the system of proton resonances in a hydrophobic rich region (hydrophobic box) of lysozyme. The devised computer-based protocol relied on inter-proton NOEs, a protein crystal structure, and *J*-coupling data used to determine residue-types. It could generate assignment possibilities consistent with all input restraints and a given distance threshold, which was treated as a variable in the optimization. The results of the distance threshold optimization established two characteristic distances for every resonance, a minimal distance necessary for a unique assignment of the resonance, and a second distance at which alternative assignment possibilities become acceptable[98]. The level of confidence in

a particular resonance assignment is thus proportional to the increase in the difference between the two characteristic distances.

From a variety of structure-based approaches focusing on the assignment of ^{1}H and ^{15}N resonances, it is worth mentioning NOEnet[104,124], the nuclear vector replacement (NVR) algorithm[113-114], contact replacement algorithm[116-117], and PEPMORPH[111], as these protocols use graphs to represent experimental NMR data and protein structures, and employ various algorithms to relate the two. These features relate the approaches to the MAGMA strategy that is explored in this thesis (Table 1.4, Chapter 2).

NOEnet

The NOEnet approach[104,124] is of particular interest, as it in principle accounts for 100% assignment accuracy and includes a number of considerations in the algorithmic design and optimization that are similar to those described in this thesis, albeit in the context of a different problem, input restraints, and algorithmic approach. NOEnet is designed for the assignment of ^{1}H-^{15}N HSQC spectra and features an efficient, complete search algorithm that solves for subgraph monomorphism[104, 124] (for definitions see Chapter 2).

The minimal input requires a list of unambiguous $^{1}H^{N}$-$^{1}H^{N}$ NOEs and a 3D protein structure in the PDB format, although CS and RDCs can be included as additional restraints[104]. Next to the theoretical maximum distance for NOE observation (d_{max}^{theo}), which defines the theoretical graph extracted from the protein's 3D structure, the number of allowed NOE outliers (T_{NOE}) and their range (D_d in Å) are additional, necessary parameters. Using the algorithm, the correct solution to the assignment problem can be found in the form of a complete assignment ensemble, provided that the input constraints are unambiguous and the associated parameters are estimated correctly. The protocol can make use of the NOE cross-peak intensities in a semi-quantitative fashion by attributing them to the edges in the NOE-graph, which become labelled according to the NOE intensity class (weak, medium, strong). The

theoretical graph extracted from the protein structure is then also edge-labelled, according to the corresponding distance classes.

Finding all possible subgraph monomorphisms, meaning all possible ways of matching a complete NOE graph of a protein to the theoretical graph extracted from its 3D structure, is an NP-complete problem for which no algorithm of polynomial time complexity is known (Chapter 2). To reduce the complexity of the complete search, NOEnet treats separate subgraphs (fragments) of a disconnected NOE graph separately in an iterative fashion, starting from the largest one. In addition, prior to the graph matching step, the algorithm preprocesses the two graphs to detect higher-order graph structures, *n*-dimensional simplices, and polytopes (for details see Stratmann *et al.*[124]). The preprocessing is flexible in the definition of polytopes and their adjacency relationships (neighbourship), as well as their matching, so that an *n*-simplex of the NOE graph can be matched to any simplex of size $d \geq n$ of the structure graph. The usage of higher order graph structures allows a simultaneous check of the matching feasibility of larger units of experimental data, rather than dealing with individual vertices alone. However, to ensure the completeness of the search, all permutations at the vertex level of any given pair of simplices are tested.

The complete search is done by a backtracking algorithm that orders the NOE graph simplices of a fragment according to the breadth-first-search. The algorithm ensures that all possible matchings of all simplices are tested by backtracking to the root simplex and altering it in an iterative fashion upon each completion of the fragment matching. The detected feasible simplex matchings are reused in further iterations of the algorithm to prune the search space. In addition, the algorithm utilizes a data-structure called an assignment table, a matrix of all possible peak/residue correspondences, which is altered iteratively to remove impossible matching possibilities and thus exploits the result of the incomplete search during the complete search.

The second version of the protocol supports inclusion of additional data (CS and RDCs)[104], next to the NOE network, which can reduce size of the resulting assignment ensemble. The additional data are introduced by the "filter approach", such that the assignments that show significant inconsistencies with the imposed restraints are rejected during the backtracking search. An assignment of a set of peaks (a) is rejected if $RMSD_a > T_R(n)$, where

$$RMSD_a = \frac{\sum_{k=1}^{n} R_{jk}^{exp} - R_{ik}^{theo}}{n}$$

R stands for restraint type (CS, RDC), j for peak, i for residue, and n for the number of assigned peaks in the current state of the backtracking search. T_R is defined by an exponential function that decreases with the number of assigned peaks, such that the tolerance to the discrepancy in theoretical and predicted values of the restraints decreases as more resonances get assigned (for details see Stratmann *et al.*[104]).

A critical assumption for a successful outcome of the NOEnet calculation is that each observed NOE is unambiguous and has a corresponding distance in the extracted protein structure graph. The inherent computational complexity of the graph-subgraph monomorphism problem and data sparsity allowed NOEnet to obtain at best 28% uniquely assigned resonances with 99.2% accuracy using NOE restraints alone on the 28 kDa EIN protein[124]. Using the same NOE data set in addition with ^{15}N-^{1}H CS and simulated RDC restraints in the second version of the program led to an increase in the fraction of uniquely assigned resonances (63%) and the authors demonstrated how the program can be combined with MARS, another automated assignment program, to further increase this fraction (>90%) with optimal (>90%) accuracy[104].

The exploitation of the NOESY data, usage of graph data structures, focus on the complete search algorithm, and generation of an assignment ensemble are the features that make this approach similar to that of MAGMA, as detailed in Chapters 2 and 3.

Structure-based automated assignment of methyl resonances

Methyl resonances are often the only available source of information in NMR studies of high molecular mass proteins. Since $^{13}C^{1}H_3$ groups of methyl-bearing residues are the only isotopically labelled part of the protein and the rest of the protein is uniformly deuterated, automatic assignment strategies cannot rely on through-bond experiments. For this reason, specialized automatic assignment strategies are required that can rely on sparse through-space restraints, as discussed in more detail below.

Approaches that are particularly adept to address the methyl resonance assignment problem in high molecular mass targets have been listed in Table 1.5.

Table 1.5 | Approaches that are specialized in the structure-based assignment of methyl resonances

Program/authors	Input data	Algorithmic approach
Venditti *et al.*[118]	PREs (for multiple nitroxide labelled sites), methyl-methyl NOEs	Metropolis Monte Carlo assignment swapping routine (approximate optimization)
MAP-XS[121-122]	obligatory: peak picked methyl-methyl NOEs, measured [^{1}H, ^{13}C]- chemical shifts, predicted [^{1}H, ^{13}C]- chemical shifts, protein structure containing hydrogen atoms optional: PREs, PCS, RDCs	Metropolis Monte Carlo assignment swapping routine (approximate optimization)
FLAMEnGO[119-120]	obligatory: peak picked methyl-methyl NOEs, measured [^{1}H, ^{13}C]- chemical shifts, predicted [^{1}H, ^{13}C]- chemical shifts, random initial assignment seed, protein structure containing hydrogen atoms optional: peak picked amide-methyl NOEs, PREs, methine-methyl TOCSY, partial assignments (obtained through bond or mutagenesis)	Metropolis Monte Carlo assignment swapping routine with fuzzy logic (approximate optimization)
Xiao *et al.*[126]	methyl-methyl NOEs, measured methyl ^{13}C chemical shifts, 3D HMCMCBCA data	Manual, hierarchical matching of observed and theoretical NOE clusters

FLAMEnGO

Fuzzy Logic Assignment of Methyl Group (FLAMEnGO) is a structure-based method designed for the automatic assignment of methyl resonances of high molecular mass proteins[119-120]. The protocol relies primarily on methyl-methyl NOE restraints in conjunction with a high-resolution protein structure, and supports PREs and methine-methyl TOCSY data as additional restraints[119-120].

At the beginning of the protocol, theoretical methyl-methyl NOEs are simulated based on the protein's structure and a distance cut-off *x*. A random assignment is initially given to the measured ^{1}H and ^{13}C chemical shifts. The optimization step of the protocol uses Monte Carlo sampling to iteratively swap the initial assignment of the resonances. Upon every swap, the total matching function $Match_{tot}$ evaluates the assignment result. The function quantifies the agreement between the observed (experimental) NMR data (e.g. methyl-methyl NOEs) and the simulated ones, in the context of a given assignment *a*. Changes in $Match_{tot}$ upon each new assignment combination *a* are monitored with the aim of finding a maximum for a global scoring function *G(x).*

$$G(x) = max_{a \in A}\{Match_{tot}\}$$

where *x* is a distance cut-off used in the simulation of theoretical methyl-methyl NOEs, *a* is an assignment solution (a member of a set of all possible assignment solutions *A; a*ϵ*A*), and $Match_{tot}$ is a linear combination of the matching functions defined for each of the NMR restraints entering the calculation (methyl-methyl NOEs, CS, PREs). $Match_{tot}$ is defined as:

$$Match_{tot} = \sum_{m \in M, r \in R} Match_{m,r}$$

where *m* is an observed methyl resonance from a set of all methyl resonances *M*, and *r* is a restraint from a set of all input restraints *R* = {NOE, CS, PRE} associated with the methyl resonance *m*. The CS term is weighted with a default value of 0.2.

An obligatory and the most important component of the total matching function is $Match_{NOE}$. The function takes a Gaussian form and is maximized when the best possible agreement between experimental (*E*) and simulated NOEs ($S_{NOE}(x, a)$) is found for one or more assignments $a \epsilon A$ at a distance cut-off *x*.

$$Match_{NOE}\,(S_{NOE}(x,a),E) = \sum_{p1\in E} \begin{cases} 1; if\ p2 \in S|D(p1,p2) \leq 1| \\ max_{p2\in S} e^{-4ln2xD^2(p1,p2)}; otherwise \end{cases}$$

The term *D* (p_1, p_2) denotes the distance between experimental ($p_1 \in E$) and simulated ($p_2 \in S$) NOE cross-peaks. In the version 2.0, the scoring was adjusted such that any particularly well scored match $D(p1, p2) \leq 1$ obtains the highest score, i.e. 1, as indicated by the expression above (for details see Chao *et al.*[119-120]).

It is important to note that the algorithm evaluates each assignment swap, i.e. each Monte Carlo move, fuzzily. This means that each move gets a score in the range 0-100% (complete mismatch-perfect match), rather than a binary decision 'accept' or 'reject'. This gives greater flexibility in the matching and better evaluation of assignment combinations that is necessary to address ambiguities present in the input restraints. FLAMEnGO optimization treats both the distance cut-off *x* and the methyl assignment $a \epsilon A$ as variables, allowing a simultaneous search for the best possible solution for both. The graphical interface of the program enables a range of distance cut-offs to be tested (see Chapter 3) by displaying a graph of global score values at different distance cut-off values. To evaluate the confidence of methyl assignments, a user can choose the optimal cut-off and re-run the calculation multiple times with the same parameters.

FLAMEnGO requires manual peak picking of methyl-methyl NOESY spectra, but does not require assignment of the NOEs nor information about the cross-peak intensities. A set of *Match* functions for the PRE-, CS-, and TOCSY-restraints are defined in FLAMEnGO so that these restraints can be incorporated in the calculation[119-120]. Given that each assignment swap

consists of swapping an assignment of two methyl resonances of the same residue type and prochirality (pro-*R* and pro-*S*), the user is required to provide this information in the input. In the second version of the software, optional methine-methyl TOCSY restraints are introduced, which enables the algorithm to identify two methyl resonances of one same Leu/Val-residue, thus reducing the sampling space and speeding up the search. The requirement for these restraints is that a protein sample with protonated, ^{13}C-labelled, methine positions in Leu/Val-side chains is produced and a 3D TOCSY-HMQC experiment recorded.

MAP-XS

Methyl Assignment Prediction from X-ray Structures (MAP-XS)[121-122], was the first automatic methyl assignment program to be introduced, which shares many similarities with FLAMEnGO. The latter was in fact developed after the former, with the goal of addressing high sparsity and the ambiguity of input restraints. A second version of the MAP-XS protocol has since been developed, which addressed those same issues and expanded the range of supported input restraints[122]. The beginning of the MAP-XS protocol is equivalent to that of FLAMEnGO, in that NOE contacts are simulated on the basis of a high-resolution protein structure, distance cut-off, and experimental ^{1}H and ^{13}C chemical shifts. Obligatory and optional input restraints for the protocol are listed in Table 1.5**.**

The first version of the protocol generated the starting assignment by a best-first search, unlike the random starting assignment used by FLAMEnGO. However, the second version of the protocol starts from an initial, random assignment, followed by a Metropolis Monte Carlo assignment swapping routine that aims at optimizing a global scoring function. As such it is conceptually equivalent to FLAMnGO, albeit with a different form of the scoring function. In addition, the assignment swapping routine of MAP-XS differs from the fuzzy approach of FLAMEnGO, as an increase or a decrease in the global scoring function constitutes acceptance or rejection, respectively[119-120].

The global scoring function of MAP-XS is of the form:

$$S = -w_{cs} \sum_i \left(\frac{\delta_i^{exp} - \delta_i^{calc}}{\sigma_i} \right)^2 + \sum_p S_{NOE} - w_R \sum_N (R^{exp}(N) - R^{calc}(N))^2$$

where w_{cs} and w_R are weighting factors adjusted through optimization and *R* stands for a measured quantity of any of the optional restraints (PRE, PCS or RDC). Each of the optional restraints has an individual scoring term, which is similar to the form shown above. δ and σ are chemical shift and chemical shift prediction error, respectively; *i*, *p*, and *N* are iterators over all measured chemical shifts, NOE cross-peaks, and additional restraints, respectively; S_{NOE} is the NOE scoring term of MAP-XS that is governed by the following equation:

$$S_{NOE} = s\ x\ log \frac{D}{d}\ x\ e^{-0.5\ \sum_i \left(\frac{\delta_i^{NOE} - \delta_i^p}{\sigma_i} \right)}$$

where *s*=1 or 0.4, if the symmetrical NOE cross-peak is found or if it is missing, respectively; *D* is a constant set to 5 Å, *d* is the distance between the two methyl groups that share an NOE in the structure (according to the running methyl assignment), $\delta^{NOE} - \delta^p$ evaluates the difference between the NOE cross-peak position and methyl chemical shifts, and σ is the NOESY chemical shift error. The iterator *i* iterates over all observed NOEs. The score decays with mismatches between experimental NOEs and methyl chemical shifts, and increases with decreasing distances between methyls and symmetry of the NOE cross-peaks. Unlike MAP-XS, FLAMEnGO does not consider the NOE cross-peak symmetry in its scoring function (see above).

The above discussed form of the NOE scoring function is introduced in the updated, second version of the MAP-XS protocol, which better accounts for the ambiguity in NOE restraints. Like FLAMEnGO, MAP-XS does not require NOE cross-peak assignment, but relies on chemical shift values of the cross-peaks. Based on NOE- and [^{1}H, ^{13}C]-HMQC-chemical shifts, the protocol assesses the most probable assignment for NOEs. Due to the degeneracy

in methyl chemical shifts, multiple assignment options for some of the NOE peaks need to be considered. During the execution, the protocol checks all possible ambiguous NOE assignments by individually assessing their compatibility with simulated NOE contacts and the current methyl assignment. Only once the best possible NOE assignment has been found for one of the ambiguous possibilities will the match be evaluated by the scoring function. An advantage of MAP-XS over FLAMEnGO is speed of the protocol. The matching score is evaluated over individual methyl groups so that upon each assignment swap only the appropriate "local" scores need to be evaluated[122].

Both MAP-XSII and FLAMEnGO support multiple, parallel calculations to be performed with the optimized parameter values to allow the assessment of assignment confidence at the level of individual methyl assignments. The assignments that are consistent across the runs are considered reliable[119-120, 122]. Given that neither of the protocols guarantees sampling completeness, it is possible for different calculations to obtain different results, which makes cross-checking necessary. Cross-checking also assesses the convergence of the protocols, as a high degree of inconsistency in the results from different runs could indicate that the calculations have not converged.

While FLAMEnGO helps the user to optimize the cut-off distance for methyl-methyl NOEs, MAP-XS requires a single distance cut-off to be specified by the user. However, the user is encouraged to assess score values and assignment results at different distance cut-offs to optimize this parameter for their dataset[121-122] (see Chapter 3).

Next to the synthetic data[121], MAP-XS was tested on three experimental datasets that represent a range of molecular shapes and masses: the 27 kDa *E.coli* Enzyme I (EIN), the 42 kDa maltose binding protein (MBP), and the 360 kDa α7α7 double ring of the 20S proteasome core particle (α7α7)[122]. After multiple runs and manual inspection of assignment (in)consistencies, the program could assign 92%, 66% and 88% of methyls for EIN, MBP

and α7α7, respectively, albeit with errors in confident assignments in the range from 1.2 to 7.5%[122].

FLAMEnGO2.0 was tested on sparse synthetic data comprising NOE- and PRE-restraints generated from a crystal structure of the maltose binding protein (MBP), and on experimental NMR data collected on the C-terminal domain of cAMP-dependent protein kinase A (PKA-C, ~41 kDa)[120]. For this dataset, methyl-methyl NOEs, PREs, and methine-methyl TOCSY data were used, resulting in ~79% of confidently and accurately assigned methyl-bearing residues, and 16% of confident, but incorrect assignments. The accuracy of the assignments was determined by cross-validation with through-bond experiments[120].

Although careful design of the two protocols aims at accounting for ambiguity in input experimental data, the major issues for these approximate approaches are the potentially limited sampling, which could result from traps in local optima, and the accuracy of the used scoring functions. Local minima can prevent the correct solution to be found, even if the derived global scoring function is optimal, and it is therefore not guaranteed that the protocols will always find the optimal solution. Especially when considering high-molecular mass targets with a large number of methyl groups, the combinatorics of the search becomes intractable and, as such, exhaustive Monte Carlo sampling is required to increase the probability of evaluating the correct solution.

Finally, in the context of automatic methyl assignment, it is worth mentioning a structure-based, PRE-method introduced by Venditti *et al.*[118]. The strategy requires the production of multiple NMR protein samples, where each has a nitroxide spin label introduced to a different surface-exposed cysteine residue. Methyl-proton R_2 relaxation rates ($^1H_{methyl}$-Γ_2) are measured for each cysteine variant, and these values can also be predicted based on a crystal structure. The protocol is Metropolis Monte Carlo-based and minimizes the PRE-energy term, which calculates the discrepancy between the observed and calculated $^1H_{methyl}$-Γ_2[118]. In

addition, a comparison of the predicted and observed methyl-methyl NOE connectivity patterns can be performed to refine the PRE-based methyl assignment. The protocol was tested on the 27 kDa protein (N-terminal EIN1) using five different sets of $^{1}H_{methyl}$-Γ_2 rates, methyl-methyl NOE restraints, and HC(CO)NH-TOCSY data. An assignment for all Ala, Ile, and Met methyl resonances and ~90% of Leu and Val resonances could be obtained.

More recently, another structure-based approach to methyl resonance assignment was proposed by Xiao *et al*[126]., which at the moment does not have a dedicated software implementation. The approach requires methyl-methyl NOESY and through-bond side-chain NMR experiments to be performed[127]. First, methyl residue types are established through inspection of ^{13}C-methyl chemical shifts and through-bond correlations of methyl resonances with Cα and Cβ chemical shifts. This is followed by a manual inspection of the best match between the theoretical (structure-based) and observed Ile-Ile, Ile-Val, Ile-Leu, Val-Val, Val-Leu, and Leu-Leu NOE clusters. Mapping between the clusters is hierarchical and starts from the largest observed NOE clusters, subsequently progressing towards the smallest clusters, favoring Ile-only or Ile-rich NOE clusters in the ordering. The sequential mapping of clusters from the largest to the smallest ones reduces the number of possibilities for the matching of the smaller clusters. Once the theoretically observed NOE cluster correspondences have been established, i.e. the NOE data have been mapped to regions of the X-ray structure, a qualitative inspection of the NOE connectivities within the clusters is performed to assign individual methyl-resonances from an observed NOE cluster to the methyl-bearing residues in the associated, 'theoretical' NOE cluster. Any ambiguity in the initial NOE cluster mapping is resolved through an inspection of the theoretical and predicted NOE correspondences of different clusters. A qualitative inspection of NOE cross-peak intensities in the later stages of the analysis can be performed to reduce assignment ambiguities.

The approach was applied to the Ile-, Leu-,Val-methyl labelled 42 kDa extracellular signal-regulated kinase 2 (ERK2), assigning 60% of methyls, including 90% of Ile resonances[126]. Even though the approach is conceptually attractive, the lack of a dedicated computational algorithm prevents its benchmarking against a larger set of theoretical and experimental NMR datasets. As such, the protocol relies on human assessment of the data matching ambiguity and an empirical determination of the optimal distance cut-offs for the definitions of different theoretical NOE clusters.

1.7 Thesis content

The different structure-based assignment approaches discussed above demonstrated advantages and disadvantages of different algorithmic approaches and experimental data input. As such, they have paved the way for the development of the MAGMA method, the central topic of this thesis. The aim of this thesis is to introduce a novel, exact, structure-based automatic assignment strategy for methyl resonances in high molecular mass proteins, and to demonstrate its effectiveness on a range of simulated and experimental datasets. In Chapter 2, the theoretical basis and design of MAGMA's algorithm is discussed, and the performance of the devised protocol is presented on synthetic data. In Chapter 3, MAGMA is applied to eight different experimental NMR datasets, featuring proteins that vary in mass, shape, methyl labelling scheme, and sparsity of input restraints. MAGMA's performance on the benchmark is compared to that of the alternative methyl resonance assignment programs, FLAMEnGO and MAP-XS. The introduction of ^{1}H and ^{13}C chemical shift restraints and the effect of different input structures on MAGMA-derived methyl assignments is discussed. In Chapter 4, novel applications of MAGMA to methyl resonance assignments of two molecular chaperones are presented. Finally, the potential of MAGMA's assignment results to guide protein-ligand docking studies is discussed.

1.8 References

1. Saunders, M.; Wishnia, A.; Kirkwood, J. G., The nuclear magnetic resonance spectrum of ribonuclease1. *Journal of the American Chemical Society* **1957,** *79* (12), 3289-3290.
2. Ikura, M.; Kay, L. E.; Bax, A., A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **1990,** *29* (19), 4659-4667.
3. Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M., Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *Journal of the American Chemical Society* **1990,** *112* (12), 4989-4991.
4. Bax, A.; Grzesiek, S., Methodological advances in protein NMR. In *NMR of Proteins*, Springer: 1993; pp 33-52.
5. Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94* (23), 12366-71.
6. Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K., Transverse relaxation-optimized spectroscopy (TROSY) for NMR studies of aromatic spin systems in 13C-labeled proteins. *Journal of the American Chemical Society* **1998,** *120* (25), 6394-6400.
7. Güntert, P.; Mumenthaler, C.; Wüthrich, K., Torsion angle dynamics for NMR structure calculation with the new program Dyana11Edited by P. E. Wright. *Journal of Molecular Biology* **1997,** *273* (1), 283-298.
8. Guntert, P., Automated NMR protein structure calculation. *Prog Nucl Mag Res Sp* **2003,** *43* (3-4), 105-125.
9. Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M., The Xplor-NIH NMR molecular structure determination package. *Journal of magnetic resonance* **2003,** *160* (1), 65-73.
10. Güntert, P., Automated structure determination from NMR spectra. *European Biophysics Journal* **2009,** *38* (2), 129-143.
11. Baldwin, A. J.; Kay, L. E., NMR spectroscopy brings invisible protein states into focus. *Nat Chem Biol* **2009,** *5* (11), 808-14.
12. Ruschak, A. M.; Kay, L. E., Methyl groups as probes of supra-molecular structure, dynamics and function. *Journal of biomolecular NMR* **2010,** *46* (1), 75-87.
13. Rosenzweig, R.; Kay, L. E., Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu Rev Biochem* **2014,** *83*, 291-315.
14. Tugarinov, V.; Hwang, P. M.; Kay, L. E., Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. *Annual review of biochemistry* **2004,** *73* (1), 107-146.
15. Tugarinov, V.; Kay, L. E., Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. *Chembiochem* **2005,** *6* (9), 1567-77.
16. Sakakibara, D.; Sasaki, A.; Ikeya, T.; Hamatsu, J.; Hanashima, T.; Mishima, M.; Yoshimasu, M.; Hayashi, N.; Mikawa, T.; Walchli, M.; Smith, B. O.; Shirakawa, M.; Guntert, P.; Ito, Y., Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature* **2009,** *458* (7234), 102-5.
17. Sakai, T.; Tochio, H.; Tenno, T.; Ito, Y.; Kokubo, T.; Hiroaki, H.; Shirakawa, M., In-cell NMR spectroscopy of proteins inside Xenopus laevis oocytes. *Journal of biomolecular NMR* **2006,** *36* (3), 179-188.

18. Sakakibara, D.; Sasaki, A.; Ikeya, T.; Hamatsu, J.; Hanashima, T.; Mishima, M.; Yoshimasu, M.; Hayashi, N.; Mikawa, T.; Wälchli, M., Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature* **2009,** *458* (7234), 102-105.
19. Selenko, P.; Wagner, G., Looking into live cells with in-cell NMR spectroscopy. *Journal of structural biology* **2007,** *158* (2), 244-253.
20. Overhauser, A. W., Polarization of Nuclei in Metals. *Physical Review* **1953,** *92* (2), 411-415.
21. Carver, T. R.; Slichter, C. P., Polarization of Nuclear Spins in Metals. *Physical Review* **1953,** *92* (1), 212-213.
22. Solomon, I., Relaxation Processes in a System of 2 Spins. *Physical Review* **1955,** *99* (2), 559-565.
23. Kaiser, R., Use of Nuclear Overhauser Effect in Analysis of High-Resolution Nuclear Magnetic Resonance Spectra. *J Chem Phys* **1963,** *39* (10), 2435-&.
24. Anderson, W. A.; Freeman, R., Influence of a Second Radiofrequency Field on High-Resolution Nuclear Magnetic Resonance Spectra. *J Chem Phys* **1962,** *37* (1), 85-&.
25. Karplus, M., Contact electron-spin coupling of nuclear magnetic moments. *The Journal of chemical physics* **1959,** *30* (1), 11-15.
26. Karplus, M., Vicinal Proton Coupling in Nuclear Magnetic Resonance. *Journal of the American Chemical Society* **1963,** *85* (18), 2870-2871.
27. Wuthrich, K., *NMR of proteins and nucleic acids*. Wiley: 1986.
28. Meiboom, S.; Gill, D., Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Review of Scientific Instruments* **1958,** *29* (8), 688-691.
29. Ernst, R. R., Nuclear magnetic resonance Fourier transform spectroscopy (Nobel lecture). *Angewandte Chemie International Edition in English* **1992,** *31* (7), 805-823.
30. Ernst, R.; Anderson, W., Application of Fourier transform spectroscopy to magnetic resonance. *Review of Scientific Instruments* **1966,** *37* (1), 93-102.
31. Jeener, J., Ampere International Summer School, Basko Polje, Yugoslavia. *Unpublished* **1971**.
32. Aue, W.; Bartholdi, E.; Ernst, R. R., Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *The Journal of Chemical Physics* **1976,** *64* (5), 2229-2246.
33. Ernst, R. R.; Bodenhausen, G.; Wokaun, A., *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Clarendon Press: 1990.
34. Morris, G. A., Sensitivity enhancement in nitrogen-15 NMR: polarization transfer using the INEPT pulse sequence. *Journal of the American Chemical Society* **1980,** *102* (1), 428-429.
35. Williamson, M. P.; Havel, T. F.; Wüthrich, K., Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *Journal of molecular biology* **1985,** *182* (2), 295-315.
36. Kaptein, R.; Zuiderweg, E.; Scheek, R.; Boelens, R.; Van Gunsteren, W., A protein structure from nuclear magnetic resonance data: lac repressor headpiece. *Journal of molecular biology* **1985,** *182* (1), 179-182.
37. Güntert, P., Structure calculation of biological macromolecules from NMR data. *Quarterly reviews of biophysics* **1998,** *31* (02), 145-237.
38. Kay, L. E.; Clore, G. M.; Bax, A.; Gronenborn, A. M., Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1b in solution. *Science* **1990,** *249*, 411-414.
39. Sattler, M.; Schleucher, J.; Griesinger, C., Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution. *Prog Nucl Mag Res Sp* **1999,** *34*, 93-158.

40. Kay, L. E.; Torchia, D. A.; Bax, A., Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **1989,** *28* (23), 8972-8979.
41. Sklenář, V.; Torchia, D.; Bax, A., Measurement of carbon-13 longitudinal relaxation using 1H detection. *Journal of Magnetic Resonance (1969)* **1987,** *73* (2), 375-379.
42. Skelton, N.; Palmer, A.; Akke, M.; Kordel, J.; Rance, M.; Chazin, W., Practical aspects of two-dimensional proton-detected 15 N spin relaxation measurements. *Journal of Magnetic Resonance, Series B* **1993,** *102* (3), 253-264.
43. Nirmala, N.; Wagner, G., Measurement of 13C relaxation times in proteins by two-dimensional heteronuclear 1H-13C correlation spectroscopy. *Journal of the American Chemical Society* **1988,** *110* (22), 7557-7558.
44. Nicholson, L. K.; Kay, L. E.; Baldisseri, D. M.; Arango, J.; Young, P. E.; Bax, A.; Torchia, D. A., Dynamics of methyl groups in proteins as studied by proton-detected carbon-13 NMR spectroscopy. Application to the leucine residues of staphylococcal nuclease. *Biochemistry* **1992,** *31* (23), 5253-5263.
45. Jarymowycz, V. A.; Stone, M. J., Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences. *Chemical Reviews* **2006,** *106* (5), 1624-1671.
46. Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H., Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc Natl Acad Sci U S A* **1995,** *92* (20), 9279-83.
47. Tjandra, N.; Bax, A., Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **1997,** *278* (5340), 1111-4.
48. Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A., Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol* **1997,** *4* (9), 732-8.
49. Palmer 3rd, A.; Kroenke, C. D.; Loria, J. P., Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods in enzymology* **2000,** *339*, 204-238.
50. Palmer, A. G., Chemical exchange in biomacromolecules: past, present, and future. *Journal of magnetic resonance* **2014,** *241*, 3-17.
51. Tolman, J. R.; Ruan, K., NMR Residual Dipolar Couplings as Probes of Biomolecular Dynamics. *Chemical Reviews* **2006,** *106* (5), 1720-1736.
52. Kay, L. E., New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy. *J Mol Biol* **2016,** *428* (2 Pt A), 323-31.
53. Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K., Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences* **1997,** *94* (23), 12366-12371.
54. Salzmann, M.; Pervushin, K.; Wider, G.; Senn, H.; Wüthrich, K., TROSY in triple-resonance experiments: New perspectives for sequential NMR assignment of large proteins. *Proceedings of the National Academy of Sciences* **1998,** *95* (23), 13585-13590.
55. Salzmann, M.; Wider, G.; Pervushin, K.; Wüthrich, K., Improved sensitivity and coherence selection for [15N, 1H]-TROSY elements in triple resonance experiments. *Journal of biomolecular NMR* **1999,** *15* (2), 181-184.
56. Gardner, K. H.; Kay, L. E., Production and incorporation of 15N, 13C, 2H (1H-δ1 methyl) isoleucine into proteins for multidimensional NMR studies. *Journal of the American Chemical Society* **1997,** *119* (32), 7599-7600.
57. Cavanagh, J.; Fairbrother, W. J.; Palmer, A. G.; Skelton, N. J.; Rance, M., *Protein NMR Spectroscopy: Principles and Practice*. Elsevier Science: 2010.

58. Levitt, M. H., *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley: 2013.
59. Redfield, A. G., On the Theory of Relaxation Processes. *IBM Journal of Research and Development* **1957,** *1* (1), 19-31.
60. Goldman, M., *Quantum description of high-resolution NMR in liquids*. Clarendon Press: 1988.
61. Fischer, M.; Majudmar, A.; Zuiderweg, E., Protein NMR relaxation: theory, applications and outlook. *Prog Nucl Mag Res Sp* **1998,** *33* (3-4), 207-272.
62. Karunanithy, G.; Cnossen, A.; Muller, H.; Peeks, M. D.; Rees, N. H.; Claridge, T. D. W.; Anderson, H. L.; Baldwin, A. J., Harnessing NMR relaxation interference effects to characterise supramolecular assemblies. *Chemical Communications* **2016,** *52* (47), 7450-7453.
63. Ollerenshaw, J. E.; Tugarinov, V.; Kay, L. E., Methyl TROSY: explanation and experimental verification. *Magnetic Resonance in Chemistry* **2003,** *41* (10), 843-852.
64. Tugarinov, V.; Hwang, P. M.; Ollerenshaw, J. E.; Kay, L. E., Cross-correlated relaxation enhanced 1H[bond]13C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *J Am Chem Soc* **2003,** *125* (34), 10420-8.
65. Kay, L. E., NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance* **2005,** *173* (2), 193-207.
66. Kay, L. E.; Torchia, D., The effects of dipolar cross correlation on 13 C methyl-carbon T 1, T 2, and NOE measurements in macromolecules. *Journal of Magnetic Resonance (1969)* **1991,** *95* (3), 536-547.
67. Bax, A.; Griffey, R. H.; Hawkins, B. L., Correlation of Proton and N-15 Chemical-Shifts by Multiple Quantum Nmr. *Journal of Magnetic Resonance* **1983,** *55* (2), 301-315.
68. Jeener, J.; Meier, B. H.; Bachmann, P.; Ernst, R. R., Investigation of exchange processes by two-dimensional NMR spectroscopy. *The Journal of Chemical Physics* **1979,** *71* (11), 4546-4553.
69. Tugarinov, V.; Kay, L. E.; Ibraghimov, I.; Orekhov, V. Y., High-resolution four-dimensional 1H-13C NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. *J Am Chem Soc* **2005,** *127* (8), 2767-75.
70. Vögeli, B., The nuclear Overhauser effect from a quantitative perspective. *Prog Nucl Mag Res Sp* **2014,** *78*, 1-46.
71. Wagner, G.; Hyberts, S. G.; Havel, T. F., NMR structure determination in solution: a critique and comparison with X-ray crystallography. *Annu Rev Biophys Biomol Struct* **1992,** *21*, 167-98.
72. Janin, J.; Miller, S.; Chothia, C., Surface, subunit interfaces and interior of oligomeric proteins. *Journal of molecular biology* **1988,** *204* (1), 155-164.
73. Metzler, W. J.; Wittekind, M.; Goldfarb, V.; Mueller, L.; Farmer, B. T., Incorporation of 1H/13C/15N-{Ile, Leu, Val} into a Perdeuterated, 15N-Labeled Protein: Potential in Structure Determination of Large Proteins by NMR. *Journal of the American Chemical Society* **1996,** *118* (28), 6800-6801.
74. Rosen, M. K.; Gardner, K. H.; Willis, R. C.; Parris, W. E.; Pawson, T.; Kay, L. E., Selective Methyl Group Protonation of Perdeuterated Proteins. *Journal of Molecular Biology* **1996,** *263* (5), 627-636.
75. Velyvis, A.; Ruschak, A. M.; Kay, L. E., An economical method for production of 2 H, 13 CH 3-threonine for solution NMR studies of large protein complexes: application to the 670 kDa proteasome. *PLoS One* **2012,** *7* (9), e43725.
76. Goto, N. K.; Gardner, K. H.; Mueller, G. A.; Willis, R. C.; Kay, L. E., A robust and cost-effective method for the production of Val, Leu, Ile (δ1) methyl-protonated 15N-, 13C-, 2H-labeled proteins. *Journal of biomolecular NMR* **1999,** *13* (4), 369-374.

77. Lichtenecker, R. J.; Coudevylle, N.; Konrat, R.; Schmid, W., Selective Isotope Labelling of Leucine Residues by Using α-Ketoacid Precursor Compounds. *ChemBioChem* **2013,** *14* (7), 818-821.
78. Lichtenecker, R. J.; Weinhäupl, K.; Reuther, L.; Schörghuber, J.; Schmid, W.; Konrat, R., Independent valine and leucine isotope labeling in Escherichia coli protein overexpression systems. *Journal of biomolecular NMR* **2013,** *57* (3), 205-209.
79. Gans, P.; Hamelin, O.; Sounier, R.; Ayala, I.; Durá, M. A.; Amero, C. D.; Noirclerc-Savoye, M.; Franzetti, B.; Plevin, M. J.; Boisbouvier, J., Stereospecific Isotopic Labeling of Methyl Groups for NMR Spectroscopic Studies of High-Molecular-Weight Proteins. *Angewandte Chemie International Edition* **2010,** *49* (11), 1958-1962.
80. Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Ono, A. M.; GÃžntert, P., Optimal isotope labelling for NMR protein structure determinations. *Nature* **2006,** *440* (7080), 52-57.
81. Monneau, Y. R.; Ishida, Y.; Rossi, P.; Saio, T.; Tzeng, S.-R.; Inouye, M.; Kalodimos, C. G., Exploiting E. coli auxotrophs for leucine, valine, and threonine specific methyl labeling of large proteins for NMR applications. *Journal of biomolecular NMR* **2016**, 1-10.
82. Ruschak, A. M.; Velyvis, A.; Kay, L. E., A simple strategy for 13C, 1H labeling at the Ile-γ2 methyl position in highly deuterated proteins. *Journal of biomolecular NMR* **2010,** *48* (3), 129-135.
83. Ayala, I.; Sounier, R.; Usé, N.; Gans, P.; Boisbouvier, J., An efficient protocol for the complete incorporation of methyl-protonated alanine in perdeuterated protein. *Journal of biomolecular NMR* **2009,** *43* (2), 111-119.
84. Isaacson, R. L.; Simpson, P. J.; Liu, M.; Cota, E.; Zhang, X.; Freemont, P.; Matthews, S., A new labeling method for methyl transverse relaxation-optimized spectroscopy NMR spectra of alanine residues. *Journal of the American Chemical Society* **2007,** *129* (50), 15428-15429.
85. Serber, Z.; Straub, W.; Corsini, L.; Nomura, A. M.; Shimba, N.; Craik, C. S.; Ortiz de Montellano, P.; Dötsch, V., Methyl groups as probes for proteins and complexes in in-cell NMR experiments. *Journal of the American Chemical Society* **2004,** *126* (22), 7119-7125.
86. Fischer, M.; Kloiber, K.; Häusler, J.; Ledolter, K.; Konrat, R.; Schmid, W., Synthesis of a 13C-Methyl-Group-Labeled Methionine Precursor as a Useful Tool for Simplifying Protein Structural Analysis by NMR Spectroscopy. *Chembiochem* **2007,** *8* (6), 610-612.
87. Sinha, K.; Jen-Jacobson, L.; Rule, G. S., Specific labeling of threonine methyl groups for NMR studies of protein–nucleic acid complexes. *Biochemistry* **2011,** *50* (47), 10189-10191.
88. Proudfoot, A.; Frank, A. O.; Ruggiu, F.; Mamo, M.; Lingel, A., Facilitating unambiguous NMR assignments and enabling higher probe density through selective labeling of all methyl containing amino acids. *Journal of Biomolecular NMR* **2016,** *65* (1), 15-27.
89. Sprangers, R.; Kay, L. E., Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **2007,** *445* (7128), 618-22.
90. Religa, T. L.; Sprangers, R.; Kay, L. E., Dynamic Regulation of Archaeal Proteasome Gate Opening As Studied by TROSY NMR. *Science* **2010,** *328* (5974), 98-102.
91. Ruschak, A. M.; Kay, L. E., Proteasome allostery as a population shift between interchanging conformers. *Proceedings of the National Academy of Sciences* **2012,** *109* (50), E3454–E3462.
92. Sprangers, R.; Kay, L. E., Probing Supramolecular Structure from Measurement of Methyl 1H−13C Residual Dipolar Couplings. *Journal of the American Chemical Society* **2007,** *129* (42), 12668-12669.

93. Baldwin, A. J.; Hansen, D. F.; Vallurupalli, P.; Kay, L. E., Measurement of Methyl Axis Orientations in Invisible, Excited States of Proteins by Relaxation Dispersion NMR Spectroscopy. *Journal of the American Chemical Society* **2009,** *131* (33), 11939-11948.
94. Hansen, D. F.; Kay, L. E., Determining Valine Side-Chain Rotamer Conformations in Proteins from Methyl 13C Chemical Shifts: Application to the 360 kDa Half-Proteasome. *Journal of the American Chemical Society* **2011,** *133* (21), 8272-8281.
95. Gelis, I.; Bonvin, A. M. J. J.; Keramisanou, D.; Koukaki, M.; Gouridis, G.; Karamanou, S.; Economou, A.; Kalodimos, C. G., Structural Basis for Signal-Sequence Recognition by the Translocase Motor SecA as Determined by NMR. *Cell 131* (4), 756-769.
96. Moseley, H. N. B. & Montelione, G. T. Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struc Biol* **1999**, *9*, 635-642, doi:Doi 10.1016/S0959-440x(99)00019-6
97. Güntert, P. Automated structure determination from NMR spectra. *Eur Biophys J* **2009**, *38*, 129-143, doi:10.1007/s00249-008-0367-z
98. Dobson, C.M.; Howarth M.A.; Redfield C. Nuclear Overhauser Effects and the assignment of the proton NMR spectra of proteins, *FEBS Lett.* **1984**, *176*, 307-312.
99. Baran, M.C.; Huang, Y.J.; Moseley, H.N.B; Montelione GT. Automated analysis of protein NMR assignments and structures. **2004**. *Chem Rev* 104(8):3541–3556, PMID: 15303826
100. Güntert, P. Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*. **2003**. *43*(3), 105-125.
101. Guerry, P. & Herrmann, T. Advances in automated NMR protein structure determination. *Quarterly reviews of biophysics*, **2011**, *44*(03), 257-309.
102. Schmidt, E. & Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *Journal of the American Chemical Society.* **2012.** *134*(30), 12817-12829.
103. Schmidt, E. & Güntert, P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *Journal of biomolecular NMR*, **2013**, *57*(2), 193-204.
104. Stratmann, D.; Guittet, E.; van Heijenoort, C. Robust structure-based resonance assignment for functional protein studies by NMR. *Journal of biomolecular NMR.* **2010**, *46*, 157-173, doi:10.1007/s10858-009-9390-3
105. Bartels, C.; Billeter, M.; Güntert, P.; Wüthrich, K. Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *Journal of biomolecular NMR*, **1996**, *7*, 207–213
106. Gronwald, W.; Willard, L.; Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD. CAMRA: chemical shift based computer aided protein NMR assignments. . *Journal of biomolecular NMR*, **1998,** *12*, 395–405
107. Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol*. **2000**, *7*, 537–558
108. Pristovsek P, Rüterjans H, Jerala R Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program st2nmr. *J Comput Chem*. **2002**, *23*, 335–340.
109. Pristovsek P, Franzoni L Stereospecific assignments of protein NMR resonances based on the tertiary structure and 2D/3D NOE data. J Comput Chem **2006**, *27*, 791–797.
110. Hus JC, Prompers JJ, Brüschweiler R. Assignment strategy for proteins with known structure. *J Magn Reson*. **2002**, *157*, 119–123.
111. Erdmann MA, Rule GS. Rapid protein structure detection and assignment using residual dipolar couplings. Technical Report CMU-CS-02-195, **2002**, School of Computer Science, Carnegie Mellon University.

112. Pintacuda G, Keniry MA, Huber T, Park AY, Dixon NE, Otting G. Fast structure-based assignment of 15N HSQC spectra of selectively 15N-labeled paramagnetic proteins. *J Am Chem Soc*. **2004**, *126*, 2963–2970
113. Langmead CJ, Yan A, Lilien R, Wang L, Donald BR. A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *J Comput Biol* .BR. **2004**, *11*, 277–298
114. Langmead CJ, Donald BR. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J Biomol NMR*. **2004**, *29*,111–138.
115. Apaydin M, Conitzer V, Donald B. Structure-based protein NMR assignments using native structural ensembles. *J Biomol NMR*, **2008**, 40:263–276.
116. Xiong F, Bailey-Kellogg C. A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. *Bioinformatics and Bioengineering, 2007 BIBE 2007 Proceedings of the 7th IEEE International Conference*. **2007**, pp 403–410.
117. Xiong F, Pandurangan G, Bailey-Kellogg C. Contact replacement for NMR resonance assignment. *Bioinformatics*, **2008**, 24:i205–i213.
118. Venditti, V., Fawzi, N.L. and Clore, G.M. Automated sequence-and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl–methyl nuclear overhauser enhancement spectroscopy. *Journal of biomolecular NMR.* **2011**, *51*(3), pp.319-328.
119. Chao, F. A., Shi, L., Masterson, L. R., Veglia, G. FLAMEnGO: a fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *Journal of magnetic resonance,* **2012**, *214*, 103-110.
120. Chao, F. A., Kim, J., Xia, Y., Milligan, M., Rowe, N., Veglia, G. FLAMEnGO 2.0: an enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J Magn Reson* **2014** , *245*, 17-23
121. Xu, Y., Liu, M., Simpson, P.J., Isaacson, R., Cota, E., Marchant, J., Yang, D., Zhang, X.., Freemont, P., Matthews, S. Automated assignment in selectively methyl-labeled proteins. Journal of the American Chemical Society. **2009**, 131, 9480-9481.
122. Xu, Y., Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J Biomol NMR* **2013**, 179-187
123. Jung, Y. S., & Zweckstetter, M. Backbone assignment of proteins with known structure using residual dipolar couplings. *Journal of biomolecular NMR*, **2004**, *30*(1), 25-35.
124. Stratmann D, van Heijenoort C, Guittet E (2009) NOEnet – Use of NOE networks for NMR resonance assignment of proteins with known 3D structure. Bioinformatics 25(4):474–481.
125. Xiao-Chen, B., McMullan, G., Scheres, SH. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*. **2015,** *40.1*, 49-57.
126. Xiao et al. Biochemistry 2015 (sequential strategy)
127. Tugarinov, V. & Kay, L. E. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *Journal of the American Chemical Society.* **2003**, *125*, 13868-13878, doi:10.1021/ja030345s.
128. Sounier R, Blanchard L, Wu Z, Boisbouvier J. High-accuracy distance measurement between remote methyls in specifically protonated proteins. *Journal of the American Chemical Society*. **2007**, *129*, 472–473
129. Kerfah, R., Plevin, M. J., Pessey, O., Hamelin, O., Gans, P., & Boisbouvier, J. Scrambling free combinatorial labeling of alanine-β, isoleucine-δ1, leucine-proS and valine-proS methyl groups for the detection of long range NOEs. *Journal of biomolecular NMR*, **2015**, *61*(1), 73-82.
130. Orts, J., Vogeli, B., Riek, R. Relaxation Matrix Analysis of Spin Diffusion for the NMR Structure Calculation with eNOEs. *J Chem Theory Comput.* **2012**, *8*, 3483-3492.

Chapter 2: Methyl assignment by graph matching (MAGMA): Theoretical bases, algorithm design, and testing

Abstract

A graph is a mathematical structure defined with a set of vertices and edges that can be used to represent objects. In the field of structural pattern recognition, graph matching is routinely applied to map similar structures or substructures of one graph to another graph, in order to find correspondences between the vertices and edges, subject to more or less stringent matching constraints. In this thesis, focus is placed on simple, vertex-labelled graphs defined on the basis of high-resolution proteins structures (structure graph) and experimental NMR data generated on exclusively methyl-labelled proteins (data graph). The aim is to relate the two graphs in order to obtain an assignment of data graph vertices, and hence assignments of methyl resonances to methyl bearing residues in the protein structure. However, the structure and data graphs are considerably different in that the data graph is expected to be sparser (lower number of edges) than the structure graph, although there may be edges in the data graph that are absent in the structure graph. Given these features, I defined a weaker form of subgraph isomorphism (monomorphism) and maximal common edge subgraph (MCES) as appropriate matching forms for a comparison of these graphs. I devised a protocol featuring two exact matching algorithms for monomorphism (*VF2*) and MCES (McGregor) combined with heuristics to optimize the search space by giving considerations to the vertex matching priorities and the order of the data graph traversal. Here, I demonstrate the design of the protocol and its utility on synthetic data graphs generated on the basis of the crystal structure of Cyclin-dependent kinase 2, a 34 kDa protein. I show that increase in data sparsity results in increased assignment ambiguity, however does not compromise accuracy. Finally, I show that high diversity in vertex labels and density of the synthetic data graphs results in higher

confidence in vertex assignments thus providing useful guidelines for experimental considerations.

In this chapter, the theoretical bases of the Methyl Assignment by Graph MAtching (MAGMA) algorithm are presented, along with detailed descriptions of the algorithmic design and testing. Considerations are given to definitions from graph theory and exact graph matching algorithms that were originally described in the field of structural pattern recognition. I start with the basic definitions that are relevant for understanding the concepts and algorithms implemented in MAGMA, and continue presenting the final protocol in which individual segments are discussed in considerable detail.

2.1 Introduction

To define the methyl-TROSY assignment problem in the domain of graph theory, definition of a graph, a subgraph, and useful relations between them are first provided. I then proceed to explain how graphs can be defined from experimental NMR data and protein structures. More details about specific graphs defined from experimental NMR datasets can be found in Chapter 3, where MAGMA's benchmark of variously sized proteins with high-resolution structures and known methyl-TROSY assignments is presented. Here, focus is placed on the theoretical bases and algorithm design.

2.1.1 Definitions

Graph

A graph $G = (V, E)$ is a mathematical structure defined by two finite sets. The set V contains elements called vertices (or nodes) and the set E contains elements called edges (or arcs).

V is a non-empty set of vertices

$E \subseteq V \times V$ is a set of edges

Each edge is defined by a set of one or more vertices, which are called endpoints[1]. Sets V and E are disjoint, i.e. they do not share common elements:

$$\mathrm{V} \cap \mathrm{E} \equiv \emptyset \tag{2.1}$$

Some authors extend the definition of graph to include an explicit definition of an incidence function, Ψ, and define a graph as an ordered triple $G = (V, E, \Psi)$, where Ψ is a function that associates a set of vertices with each edge of G[2]. Vertices in the set are not necessarily distinct. Finally, in order to include attributes on vertices and/or edges, the definition of a graph is further extended:

$G = (V, E, \alpha, \beta)$, where

V is the set of vertices

$E \subseteq V \times V$ is the set of edges

$\alpha : V \rightarrow L$ is a function assigning attributes to vertices

$\beta : E \rightarrow L$ is a function assigning attributes to edges

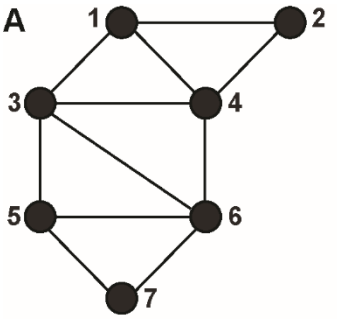
Next, I define some adjacency and incidence relations that are commonly encountered in the literature[2]. Let G be a graph, $G = (V, E)$, and e be an edge $e = (u, v)$; $u, v \in V$. The following statements apply:

1. u is incident with the edge e;
2. v is incident with the edge e;
3. e is incident with the vertex u;
4. e is incident with the vertex v;
5. u and v are adjacent;

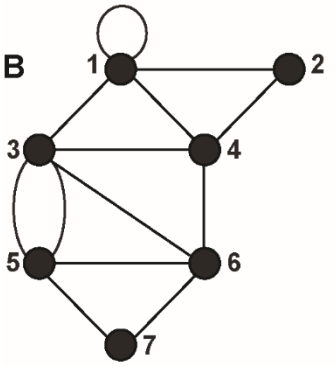
In summary, adjacent vertices are two vertices incident with a common edge, and adjacent edges are two edges incident with a common vertex. Adjacency relations of vertices of a graph G are usually presented in an adjacency matrix A_G. The adjacency matrix is defined as the symmetric matrix with identically indexed ordering of V in rows and columns, such that

$$A_G[u, v] = \begin{Bmatrix} 1 \text{ if } u \text{ and } v \text{ are adjacent} \\ 0 \text{ otherwise} \end{Bmatrix} \quad (2.2)$$

where $u \in V,\ v \in V$[1,3]. The adjacency matrix is a convenient representation of a graph for computational analysis. Two examples of such matrices are given in Figure 2.1.



	1	2	3	4	5	6	7
1	0	1	1	1	0	0	0
2	1	0	0	1	0	0	0
3	1	0	0	1	1	1	0
4	1	1	1	0	0	1	0
5	0	0	1	0	0	1	1
6	0	0	1	1	1	0	1
7	0	0	0	0	1	1	0



	1	2	3	4	5	6	7
1	2	1	1	1	0	0	0
2	1	0	0	1	0	0	0
3	1	0	0	1	2	1	0
4	1	1	1	0	0	1	0
5	0	0	2	0	0	1	1
6	0	0	1	1	1	0	1
7	0	0	0	0	1	1	0

Figure 2.1| Diagrams of a simple (**A**) and a non-simple graph (**B**). Unlike the graph depicted in **B**, the graph in **A** has no loops (edges starting and ending in the same vertex) or double edges. Vertices are numbered to aid the interpretation of the adjacency matrices (right). A loop and a multi-edge in **B** require the adjacency matrix elements to be > 1**.**

In Figure 2.1, two graph diagrams are given (A, B). One should note the difference between a graph and a diagram of graph. The sole purpose of a diagram is to depict the incidence relations. Once a graph has been defined it may be drawn in many different ways, e.g. the positions of the vertices and shape and length of edges may be presented differently. Nonetheless, a graph is mathematically indifferent with respect to the differences in diagrams. Note that in what follows, some diagrams of graphs might occasionally be referred to as graphs.

Simple undirected and labelled graph

All graphs considered in this thesis are simple undirected and labelled graphs. Simple graphs contain neither self-loops (an edge that has a single endpoint) nor multi-edges (a collection of two or more edges with the same endpoints)[1]. Hence, every edge $(u, v) \in E$ of a simple

graph $G = (V, E)$ is defined by a pair of distinct vertices. G is undirected if its edges have no orientation such that:

$$\forall\,(u,v) \in E \;\Rightarrow (v,u) \in E \tag{2.3}$$

A simple graph can be connected or disconnected. For a connected simple graph, there exists a path from any vertex to any other vertex; if not, it is disconnected. Graphical representations of some connected and disconnected simple (sub)graphs are given in Figure 2.2.

Finally, in the application presented in this thesis, I consider vertex-labelled graphs, and hence from this point on, any reference to a graph assumes a simple, undirected graph that may be connected or disconnected. A complete definition is $G = (V, E, \alpha)$, where:

V is the finite set of vertices;

$E \subseteq V \; x \; V$ is the finite set of edges;

$\alpha: V \;\rightarrow L$ is a function that assigns labels to vertices.

For this specific application, the elements of the set L are strings corresponding to the single-atom identifiers of amino acids (e.g. I, L, V, A, etc.).

Subgraph

A subgraph of a graph $G = (V, E, \alpha)$ is a graph $H = (V_H, E_H, \alpha)$ that has all of its vertices and edges in G, i.e. V_H is a subset of V and E_H is a subset of E

$$H \subseteq G \;\Leftrightarrow\; V_H \subseteq V \;\wedge\; E_H \subseteq E \tag{2.4}$$

Different subgraphs can be obtained from a graph by vertex and/or edge deletions. The definition of a subgraph is stricter if there is a requirement for the subgraph to be *vertex-induced* or *edge-induced*. A *vertex-induced* subgraph, also simply called an *induced* subgraph, is defined by any subset of vertices of the original graph and all edges that have both of their endpoints in this subset. An *edge-induced* subgraph is defined by any subset of edges of the original graph and all vertices that are the endpoints of those edges.

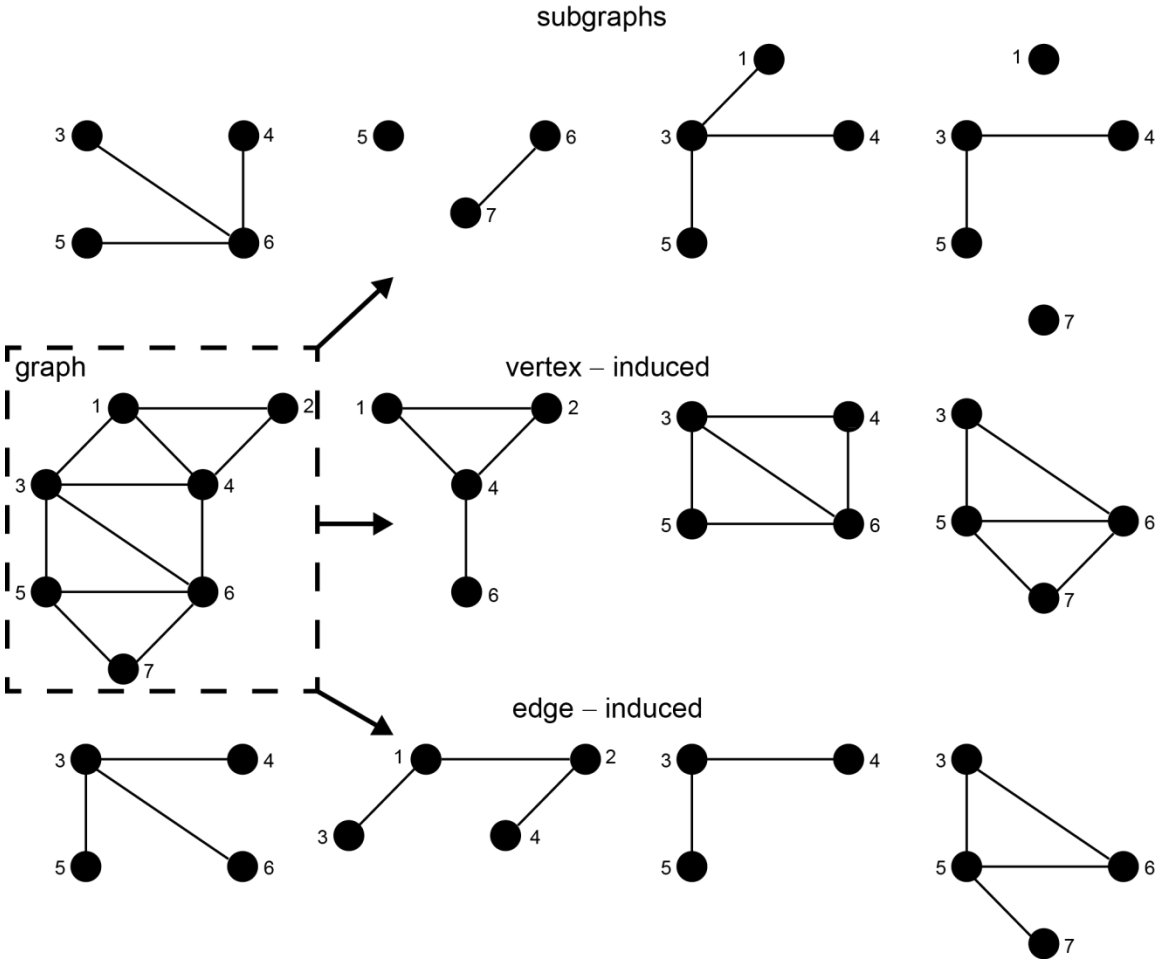


Figure 2.2 | A graphical representation of a simple graph (shown in the dashed rectangle) and its subgraphs. The upper row features subgraphs of the graph in the rectangle, which are not necessarily *induced*. Note that the graph is connected, but some of the subgraphs are disconnected. The subgraphs in the middle and the bottom row are *vertex*- and *edge-induced*, respectively. According to the definitions in the text, the first and the third subgraph in the upper row are also *edge-induced*.

In this application, I will be referring to subgraphs considering the simplest definition (4), unless when explicitly indicated otherwise.

Graph density

For a simple graph $G = (V, E, \alpha)$, graph density is defined as:

$$\eta = \frac{2x\ |E|}{|V|x(|V| - 1)} \tag{2.5}$$

Where $|E|$ is the total number of edges and $|V|$ is the total number of vertices. The graph density relation compares the number of edges of graph to a theoretical maximum defined as:

$$\frac{|V|x(|V|-1)}{2} \quad (2.6)$$

If for any given graph $\eta \ll 1$, it is considered sparse, whereas if $\eta \sim 1$, it is considered dense. A simple graph with the maximum number of edges possible is called *complete* and in this graph every vertex is connected to every other vertex ($\eta \equiv 1$).

2.1.2 Structure and data graphs

In this thesis, I consider two seemingly separate categories of graphs that are named *structure* and *data* graphs. However, graphs in both categories have the same mathematical properties as defined above, i.e. they are simple, undirected, vertex-labelled, and connected or disconnected. The difference between the two is in the origin of the graphs.

The set of vertices of V_s of any given structure graph G_s = (V_s, *Es*, α), is defined by the atomic positions (Cartesian coordinates) of methyl carbon atoms extracted from a protein structure (i.e. a PDB file). Methyl carbon-methyl carbon distances that are smaller or equal to a defined threshold (in Å) define the set of edges (Methods). For a data graph, G_d = (V_d, E_d, α), the set of vertices is defined by pairs of (^{1}H, ^{13}C)-resonant frequencies (i.e. methyl peaks) extracted from a 2D ^{1}H-^{13}C TROSY-HMQC/HSQC spectrum of a protein. Edges are defined by inter-methyl NOEs measured at the ^{1}H-^{13}C frequency positions, which have been extracted from a processed 3D or 4D NOESY spectrum (Methods). The labels in the *structure* graph are assigned to vertices on the basis of the methyl amino-acid type, which is known *a priori* and listed in the structure file. The labels for vertices in the *data* graph are defined from values of amino-acid specific ^{1}H and ^{13}C resonant frequencies (e.g. for ^{13}C chemical shifts of Ile-δ_1) or upon inspection of 2D ^{1}H-^{13}C TROSY- HMQC/HSQC spectra collected from protein samples prepared with differentially labelled methyl groups (Methods, Chapter 3), such that a specific amino-acid type label is given to every ^{1}H-^{13}C peak (data graph vertex). I also consider *synthetic data* graphs (2.2.5), which are different from data graphs in that they are defined from a structure graph. *Synthetic data* graphs were designed to

study and mimic the features of data graphs, and as such have been subjected to certain modifications (e.g. deletion or addition of edges).

2.1.3 Graph comparisons and similarity measures

Comparing two or more objects, or an object and a model to which the object is related to, has found important applications in different fields. In the field of structural pattern recognition (SPR), objects are first decomposed into graphs, also known as *graph-based pattern representation*, and then compared using "*graph matching*"[4-6]. Graph-matching approaches have been explored in the SPR field ever since the early 1970's[4,5], and a number of algorithms have been developed. These algorithms have since found applications in many other fields such as bioinformatics, analysis of complex networks, machine learning, data mining, and social sciences, just to name a few. Recently, graph-matching approaches have been receiving more attention as computationally costly graph-based algorithms are becoming more compatible with the computational power of the latest generation of computers[5].

Graph-based SPR methods can be broadly categorised to graph matching and methods related to graph comparison (Figure 2.3). Here, the focus is placed on exact graph matching problems, as these are directly relevant to MAGMA's assignment strategy. Some of the related approaches listed in Figure 2.3 have since recently received significant attention both in and outside the SPR field[4]. For a comprehensive overview of exact and inexact matching, as well as related problems, an interested reader is directed to excellent reviews by Conte *et al.*[5] and Foggia *et al.*[4].

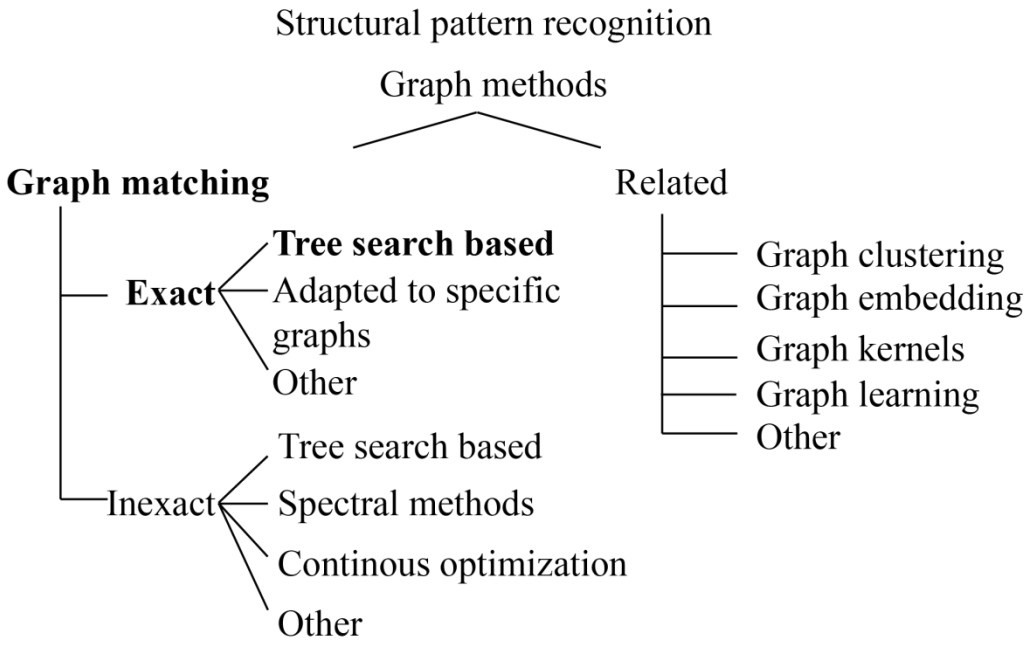


Figure 2.3 | Schematic of graph methods explored in the field of structural pattern recognition. Topics of special interest for this work are given in bold.

Exact graph matching

Graph matching can be defined as a process of mapping the similar (sub)structures of one graph to those of the other graph by finding a correspondence between the vertices and edges of the two graphs, subject to the constraints defined by the specific matching problem[5]. The constraints can be more or less stringent, as discussed in more detail below. Methods employed in graph matching can be broadly categorised to exact and inexact (Figure 2.4). Exact methods have stricter requirements for correspondence between the two graphs to be matched, or between their substructures, whereas inexact methods allow matching of the graphs that are, to an extent, structurally different.

Exact graph matching is generally defined by its strict requirement on edge preservation in the matching process. The "strictest" matching form is graph isomorphism, whereas subgraph isomorphism, monomorphism, and homomorphism have weaker requirements (see below).

Exact graph matching problems can be solved using exact (optimal) matching algorithms, or by using approximate (suboptimal) ones.

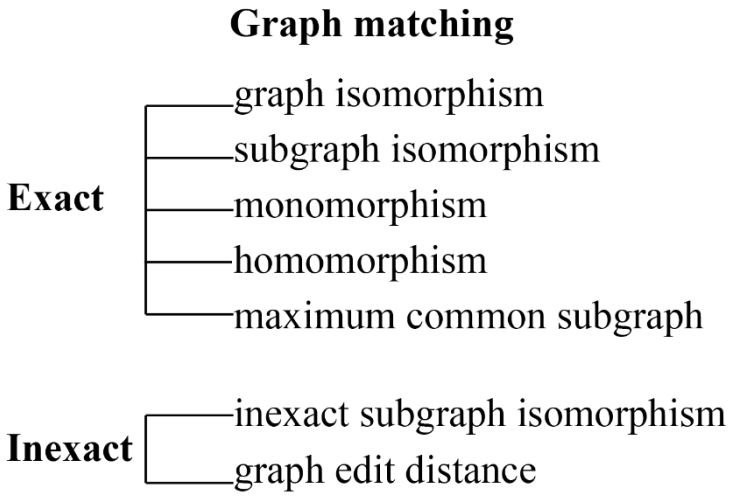


Figure 2.4 | Graph matching can be broadly divided to exact and inexact. Different graph matching problems defined in each category are listed.

Isomorphism, subgraph isomorphism, monomorphism, homomorphism, and maximum common subgraph of two graphs

Edge-preserving mapping requires that any two vertices connected by an edge in one graph, $G_1 = (V_1, E_1)$, must also be connected by an edge in another graph $G_2 = (V_2, E_2)$, if the vertices of the first graph are mapped to the vertices of the second one. For graph isomorphism, denoted $G_1 \cong G_2$, this condition holds in both directions of mapping, such that the mapping function f is a bijection (Appendix, A2.2). This is formally expressed as:

$$\forall\, v_2 \in V_2, \exists\, v_1 = f^{-1}(v_2) \in V_1 : v_2 = f(v_1) \qquad (2.7)$$

where f is a function that maps elements of the vertex set of G_1 to the vertex set of G_2, $f : V(G_1) \rightarrow V(G_2)$[1], and f^{-1} is the inverse of f and also a bijection (A2.2). In simple terms, any two isomorphic graphs have the same number of edges and vertices and the same adjacency structure.

Subgraph isomorphism is a weaker form of graph matching than graph isomorphism. It is important to make a distinction between a more constrained type of subgraph isomorphism,

generally referred to as vertex-induced subgraph isomorphism, and monomorphism[5]. Given two graphs G_1 and G_2, vertex-induced subgraph isomorphism requires that isomorphism holds true between one of the graphs and a vertex-induced subgraph (see 2.1.1) of the other. Formally:

$$\begin{cases} V_2' \subset V_2 = \{v_2 \in V_2 : \exists\, v_1 \in V_1 : v_2 = f(v_1)\} \\ \qquad E_2' \subset E_2 = E_2 \cap (V_2' \times V_2') \end{cases} \tag{2.8}$$

where f is an isomorphism between G_1 and G_2'.

A weaker type of subgraph isomorphism, also referred to as monomorphism, or edge-induced subgraph isomorphism by some authors, only requires that a unique (injective) mapping exists for each vertex and each edge of the first graph to the second graph; however, monomorphism allows the second graph to have extra (remaining) vertices and edges[5]. Formally:

$$\begin{cases} \forall\, u, v \in V_1, (u, v) \in E_1 \Rightarrow (f(u), f(v)) \in E_2 \\ \qquad \forall\ u \neq v \in V_1, \qquad f(u) \neq f(v) \end{cases} \tag{2.9}$$

where f is an injection that maps elements of the vertex set of G_1 to the vertex set of G_2, $f : V(G_1) \rightarrow V(G_2)$. In simpler terms, monomorphism between G_1 and G_2 exists if there exists a subgraph H (not necessarily vertex-induced) of G_2 ($H \subset G_2$), such that $H \cong G_1$. More than one subgraph of G_2 can meet this condition, and therefore monomorphism is not necessarily unique. Monomorphism is of special interest for MAGMA's assignments strategy, as discussed below.

An even weaker form of matching is homomorphism, which removes the requirement for the unique mapping of vertices of one graph to vertices of the other graph, thus allowing the mapping to be many-to-one[4,5]. This is formally expressed as:

$$\forall\, u, v \in V_1, (u, v) \in E_1 \Rightarrow \big(f(u), f(v)\big) \in E_2 \vee f(u) = f(v) \tag{2.10}$$

where f is a function that maps elements of the vertex set of G_1 to the vertex set of G_2, $f: V(G_1) \rightarrow V(G_2)$. The function f is neither an injection nor surjection (A2.2).

Search for a maximum common subgraph (MCS) of two graphs has recently been receiving more attention. Given two graphs G_1 and G_2, a *MCS* (G_1, G_2), is a graph $H = (V, E)$ that is isomorphic to a subgraph of G_1, $G_1' = (V_1', E_1')$ and G_2, ($G_2' = (V_2', E_2')$ and is maximized in the number of vertices or edges, depending on the definition[7,8]. It is often of interest to also retrieve the corresponding mapping between the two subgraphs, which can be expressed in formal notation:

$$\begin{cases} G_1' = \{V_1' \subset V_1, E_1' \subset E_1\}, G_2' = \{V_2' \subset V_2, E_2' \subset E_2\} \\ H \cong G_1' \wedge H \cong G_2' \\ f_1: V(H) \rightarrow V(G_1) \wedge f_2: V(H) \rightarrow V(G_2) \end{cases} \quad (2.11)$$

where f_1 and f_2 are functions that map elements of the vertex set of H to the vertex sets of G_1 and G_2, respectively. The functions f_1 and f_2 are isomorphic mappings[8]. Similar to subgraph isomorphism and monomorphism, the MCS is not necessarily unique. Additionally, the MCS problem has both a more and less "strict" definition depending on whether isomorphism needs to be found between the vertex-induced subgraphs of G_1 and G_2 (strict):

$$\begin{cases} V_1' \subset V_1, E_1' \subset E_1 = E_1 \cap (V_1' \times V_1') \\ V_2' \subset V_2, E_2' \subset E_2 = E_2 \cap (V_2' \times V_2') \end{cases} \quad (2.12)$$

or if it is sufficient that it is found between any subgraphs of G_1 and G_2 (weaker). The definition of the MCS given by McGregor focuses the search of the algorithm on the common subgraph of two graphs that contains the maximum number of edges, termed the MCES, which is also called the *edge-induced* MCS by some authors[7] (Figure 2.2). According to that definition, the common edge subgraph is a maximum common edge subgraph (MCES) if there exists no other common subgraph with more edges[9]. This definition of the MCES is adopted in MAGMA.

Of particular importance for MAGMA's assignment strategies are exact methods to solve subgraph isomorphism and maximum common subgraph, or more precisely for monomorphism, and maximal common edge subgraph.

Two algorithms, *VF2* for subgraph isomorphism and monomorphism and McGregor for maximal common edge subgraph, are implemented in the core of MAGMA's protocol (2.2.2). Both fall into the category of exact, tree search based graph matching algorithms and are discussed below in more detail.

The *VF2* subgraph isomorphism algorithm

Subgraph isomorphism in the context of MAGMA requires searching for all monomorphisms and can be solved using the exact *VF2* algorithm[10,11]. The *VF2* algorithm is a flexible, state-of-art algorithm that is able to address different exact graph matching problems – graph isomorphism, vertex-induced subgraph isomorphism, and monomorphism[12,13]- which contributes to its wide applications.

The algorithm explores the search space based on a depth-first strategy with backtracking and can be described using a state-space representation (SSR)[14]. The SSR requires a definition of the initial state (s_0 in the pseudocode below) and the goal states. Starting from an initial state, each intermediate state, denoted with a generic s in the pseudocode below, represents a partial mapping $M(s)$ between vertices of the two compared graphs (Figure 2.5). The goal states are complete mappings that are consistent with the constraints of the problem, e.g. vertex-induced subgraph isomorphism, monomorphism. When all goal states are found, the algorithm terminates.

The protocol operates using a set of rules defined in the *feasibility function*, *F*. Every candidate pair of vertices that could extend the current state *s*, is checked by *F* for a consistency with the matching constraints. In addition, the rules defined in *F* prune the search

space by preventing further exploration of any partial matching solutions that could not lead to a complete solution, as described in more detail below[10,11].

The search starts with the initial state s_0 and an empty mapping set $M(s_0) = \emptyset$. To understand the operation of the algorithm, it is necessary to define two simple, undirected graphs, a pattern $G_1 = (V_1, E_1)$ and a target $G_2 = (V_2, E_2)$, and several sets:

1. $P(s)$, the pairs of candidate vertices to be added to the current state s.
2. $M_1(s)$ and $M_2(s)$, the vertices of G_1 and G_2, respectively, already included in the partial mapping.
3. $T_1(s)$ and $T_2(s)$, the "direct neighbors", i.e. the vertices adjacent to the vertices in $M_1(s)$ and $M_2(s)$, respectively.
4. $N_{rem1}(s)$ and $N_{rem2}(s)$, the remaining vertices of G_1 and G_2, not contained in either (2) or (3). The set is generally defined as: $N_{rem}(s) = N - M(s) - T(s)$.

The sets are indicated for an illustrated example in the figure below (Figure 2.5).

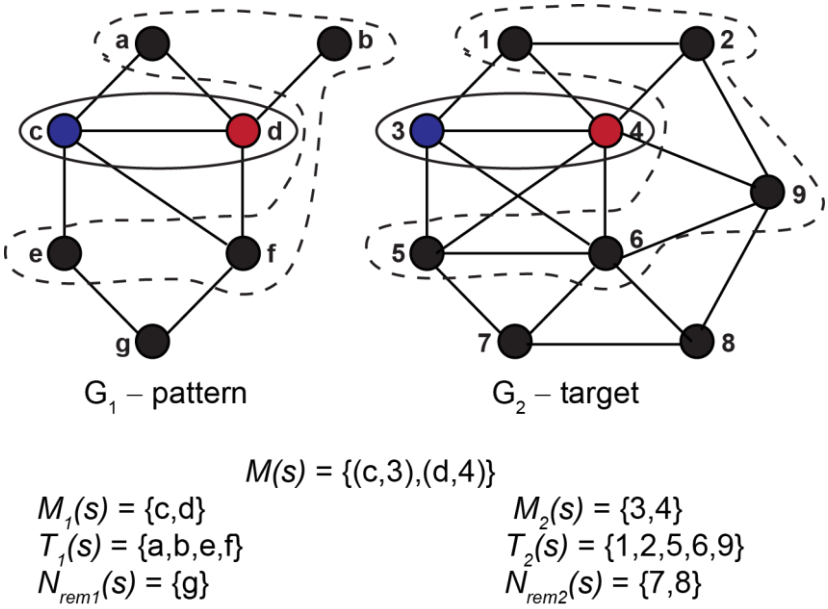


Figure 2.5 | An illustration of a state *s* in the *VF2* subgraph isomorphism search with data structures and their content indicated. See text for details.

THE VF2 ALGORITHM

```
PROCEDURE Match (s)
    INPUT: an intermediate state s; the initial state s0 for which M(s0) = ∅
    OUTPUT: the mappings between the two graphs

    IF M(s) contains all vertices of the pattern graph G1 THEN
      RETURN M(s)
    ELSE
      Compute the set P(s) of the vertex pairs candidates for inclusion in M(s)
        FOREACH (u,v) ∈ P(s)
          IF F(s,u,v) THEN
            Compute the state s' by adding (u,v) to M(s)
            CALL Match(s')
          END IF
        END FOREACH
        Restore data structures
    END IF
END PROCEDURE Match
```

The pseudocode of the algorithm is shown above and based on that from Cordella *et al.*[11]. For each intermediate state *s*, the algorithm computes *P*(*s*), attempting to first assemble it from the sets of direct neighbors, i.e. $T_1(s)$ and $T_2(s)$, or, if these are empty, from $N_{rem1}(s)$ and $N_{rem2}(s)$:

$$P_1(s) = \begin{cases} T_1(s), if\ |T_1(s)| \neq 0 \wedge |T_2(s)| \neq 0 \\ N_{rem1}(s), \quad otherwise \end{cases} \tag{2.13}$$

$$P_2(s) = \begin{cases} T_2(s), if\ |T_1(s)| \neq 0 \wedge |T_2(s)| \neq 0 \\ N_{rem2}(s), \quad otherwise \end{cases} \tag{2.14}$$

For each pair in *P*(*s*), $p = (u, v)$, $u \in P_1(s)$ and $v \in P_2(s)$), the feasibility function *F*(*s, u, v*) is applied, and if it is passed, the successor state *s'* is computed by adding the pair *p* to *M*(*s*), formally denoted $s' = s \cup p$. The protocol then applies recursively to *s'*.

The feasibility function is composed of a set of rules which, for simple, undirected graphs can be summarized as[13]:

1. If vertices adjacent to *u* and *v* are in the current partial mapping *M*(*s*), each branch from *u* has a correspondence in the branches of *v* and *vice versa*, formally expressed as:

$$R_1(s,u,v) \Leftrightarrow \forall u' \in adj(u) \cap M_1(s)\ \exists!\ v' \in adj(v) \cap M_2(s) : (u',v') \in M(s) \quad (2.15)$$
$$\wedge\ \forall v' \in adj(v) \cap M_2(s)\ \exists!\ u' \in adj(v) \cap M_1(s) : (u',v') \in M(s)$$

2. The number of vertices adjacent to *u* and *v*, and not in the current partial mapping *M*(*s*), allows for the matching to result in a subgraph isomorphism

$$R_2(s,u,v) \Leftrightarrow |adj(u) \cap T_1(s)| \leq |adj(v) \cap T_2(s)| \quad (2.16)$$

3. The total number of vertices remaining in the graphs allows for the matching to result in a subgraph isomorphism

$$R_3(s,u,v) \Leftrightarrow |adj(u) \cap N_{rem1}(s)| \leq |adj(v) \cap N_{rem2}(s)| \quad (2.17)$$

If the search for graph isomorphism is required, the ≤ operator can instead be replaced by =. R_1 is also referred to as the *core* rule, whereas R_2 and R_3 are *look-ahead* rules. The *core* rule makes sure that the generated states *s* are consistent with the problem at hand (e.g. graph isomorphism, graph-subgraph isomorphism), whereas the *look-ahead* rules ensure that any future inconsistencies are detected early and the search space is pruned. Moreover, the *look-ahead* rules can be extended to *k-look ahead* rules, checking for consistency over *k* steps ahead rather than only one[11]. If vertices or edges of the graph contain attributes, i.e. labels, the semantic checks are included in the feasibility rules, so that only the pairs of vertices with the same label are considered. For more details on the feasibility rules, computation of the set of candidate pairs, data structures used, and implementation issues, the interested reader is directed to original works[10,11,13,15].

The heuristics implemented in the *VF2* algorithm focus on sets of vertices adjacent to those already included in the partial mapping and are fast to compute. *VF2* has been shown to result in significant improvement in performance over Ullmann's algorithm and several other graph-subgraph isomorphism algorithms[12,16]. Finally, the new version of *VF2* has reduced

spatial complexity, of the order $O(N)$, where N is the maximum number of vertices of the two graphs. This compares favorably when considering the original memory requirements of the order $O(N^2)$ and $O(N^3)$ for Ullmann's algorithm[17], making the new version of *VF2* particularly suitable for applications to large graphs[11].

2.1.4 The maximal common edge subgraph (MCES) algorithm by McGregor

The maximal common edge subgraph (MCES), as defined by McGregor[9] and as considered in this thesis, is a common subgraph of two graphs that contains the largest possible number of edges, also referred to as the edge-induced MCS by some authors[7]. The algorithm was developed for the purpose of classifying chemical reactions by comparing the structures of molecules before and after chemical reactions[9].

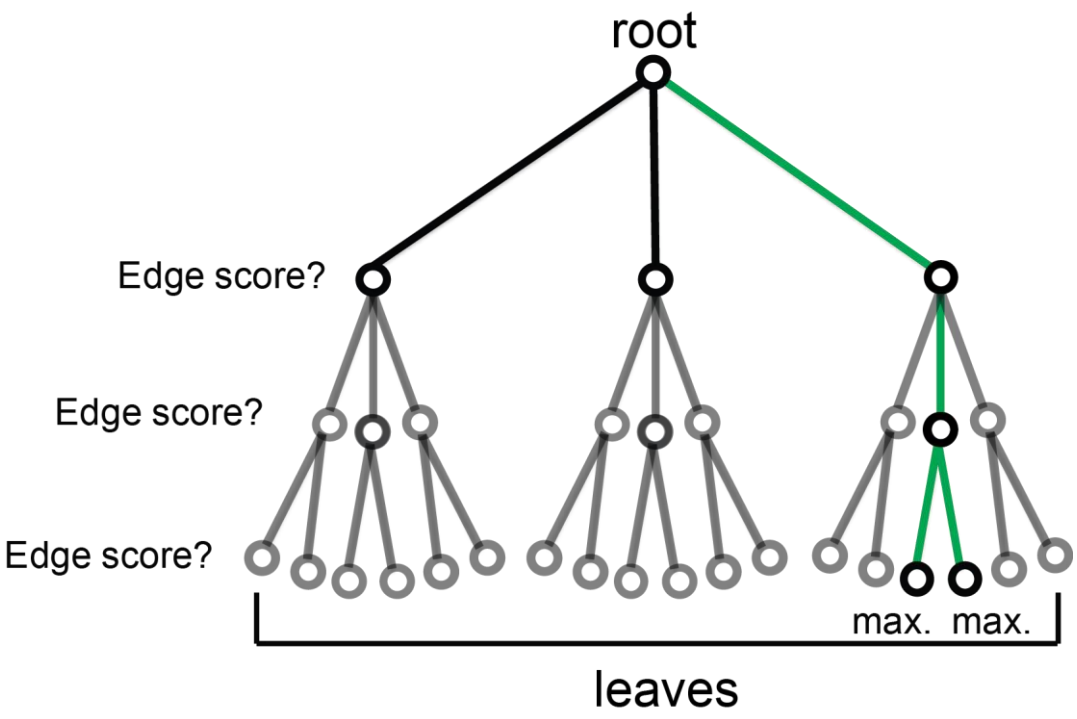


Figure 2.6 | An illustration of a tree search performed by McGregor's MCES algorithm. Each vertex in the search tree corresponds to a tentative assignment decision, i.e. mapping of a vertex of a data graph to a vertex of a structure graph. Edge correspondences are assessed after every tentative assignment decision. The score corresponds to the number of edges that could still be explained given the current (partial) assignment solution. An ideal outcome of the search is illustrated, where initially a maximum size MCES was found (one of the green branches). Upon backtracking from one of the leaves, and changing the last assignment decision, another MCES was found (another green branch). The rest of the branches were less fruitful and were pruned early (grey).

The algorithm is based on a depth-first tree search with backtracking[18] (Figure 2.6), and includes an effective pruning technique that reduces the search space[9]. Similar to the *VF2* algorithm, introduced earlier, the McGregor algorithm can be described using the SSR[14]. The initial state s_0 corresponds to the empty mapping, which is extended to the generic state *s* through progression of the search. Each state *s* is a common subgraph of the two graphs that is consistent with the matching constraints. The goal states are complete mappings that correspond to the maximal common edge subgraphs of the two compared graphs. The algorithm terminates upon backtracking to the root of the search tree with no other vertex mappings that could lead to the MCES left untried.

For the purpose of algorithmic description, I will define two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, such that $|V_1| \leq |V_2|$[9]. Such definition of the graphs entails that every vertex of the graph G_1 will be included in the final mapping. I will also define the following data-structures:

1. $M(s)$, the set of vertices mapped in the current state *s;* the set can be described as $M(s) = \{(u_1, v_1), (u_2, v_2), (u_3, v_3) \ldots\}$, where $u \in V_1(s)$ and $v \in V_2(s)$.
2. *edgesleft*, the number of edges of G_1 that could still be explained, i.e. mapped to the edges of G_2, given the current state *s*; $edgesleft \leq |E_1|$.
3. *bestedgesleft*, the number of edges explained by the current best mapping, i.e. the size of the current MCES, $bestedgesleft \leq |E_1|$.
4. *tried-mappings*, the list of all tried vertex correspondences defined for every $u \in V_1(s)$, given the current state *s*.
5. *MEDGES*, matrix of edge correspondences defined by E_1 and E_2

$$MEDGES = E_1 \times E_2 \tag{2.18}$$

Each position in the matrix indicates whether any given edge $e_1 \in E_1$ is allowed to correspond to any given edge $e_2 \in E_2$. For simple, undirected and unlabelled graphs, all

matrix elements are initially set to 1 to indicate that all possible edge correspondences are allowed[9].

$$MEDGES[e_1, e_2] = 1, \qquad \forall\, e_1 \in E_1, e_2 \in E_2 \tag{2.19}$$

THE MCS ALGORITHM

```
PROCEDURE McGregor_MCS
    INPUT: the initial state s0 for which M(s0) := ∅, edgesleft := |E1|,
           bestedgesleft = 0, MEDGES[e1][e2] := 1, e1 ∈ E1, e2 ∈ E2;
    OUTPUT: the mappings between the two graphs
    REPEAT
      IF (IsFeasiblePair(s,u,v)); u ∈ V1, v ∈ V2 THEN
        s' = AddPair(s,u,v)
        UpdateState(s')
        IF edgesleft(s') > bestedgesleft THEN
          IF LeafOfSearchTree THEN
            s' = MCS
            SaveMCS(s')
            bestedgesleft = edgesleft
          ELSE
            CopyState(s')
            INCREMENT u := u+1
            DeleteTried(u)
          END IF
      ELSE
        Backtrack(s')
        DECREMENT u := u-1
        Restore data structures
      END IF
    UNTIL u =0
END PROCEDURE McGregor_MCS
```

The symbol ":=" in the above pseudocode is interpreted as "set to". The search starts with the initial state s_0, and an empty mapping set $M(s_0) = \emptyset$. Data structures are prepared as indicated in the "INPUT" above. Note that the algorithm operates on a matrix that keeps track of the edge correspondences for every computed state (*MEDGES*). This data structure allows for explicit representation of edges of the two graphs and is used to direct the search toward the edge-induced maximum common subgraphs, as described in more detail below.

The first intermediate state *s* is computed from s_0 by matching the first vertex of the graph G_1 to one of its options from G_2. For each tentative vertex correspondence (*u*,*v*) - $u \in V_1, v \in V_2$that might be added to the mapping *M*, the algorithm checks the feasibility of the matching, i.e. the state *s'* is computed only if the vertex $v \in V_2$ is an option for mapping to the vertex $u \in V_1$, and if it has not already been tentatively mapped to this vertex given the current state *s*. If the pair passes the check, the state *s'* is computed ($s' = s \cup (u, v)$). The data structures are then updated to take into account the new tentative correspondence of the vertex *u* (*UpdateState*(*s'*)).

The data structures are updated as follows:

1. *MEDGES* is updated such that any edge incident with *u* is allowed to correspond only to the edges incident with *v*. This is achieved by setting to 0 any alternative correspondences for edges incident with *u*
2. *edgesleft* is computed from *MEDGES* by counting the number of edges of G_1 that could still be assigned given the state *s'*.
3. The vertex $v \in V_2$ is queued to the *tried-mappings* associated with the vertex $u \in V_1$.

$$tried - mappings(u, s') = tried - mappings(u, s) \cup v \in V_2 \quad (2.20)$$

By keeping track of the *tried-mappings* for each vertex of V_1 for each computed state *s*, the algorithm prevents repetition of any previously computed states (*IsFeasiblePair(s,u,v)*). To decide if the new state *s'* is worth pursuing, the *pruning rule* is enforced and the state *s'* is pursued only if it corresponds to a common subgraph for which the maximum theoretical number of edges that could still be mapped is larger than that of the MCS found so far:

$$edgesleft(s') > bestedgesleft \quad (2.21)$$

If this condition is passed, the algorithm will check if it reached the leaf of the search tree, i.e. if for every vertex of G_1 there is a tentative correspondence in G_2.

$$\forall\, u \in V_1\ \exists!\ v \in V_2 \quad (2.22)$$

If this is the case, the computed mapping $M(s)$ correspond to the MCS and it sets the lower boundary that needs to be exceeded by any new MCS, as well as any partial solution leading to it (*bestedgesleft* := *edgesleft*).

If, however, there are vertices of G_1 remaining to be mapped, the algorithm will copy the current state (*CopyState(s')*) prior to proceeding to a yet unassigned vertex of G_1 ($u := u+1$). The copy of the current state includes all data structures associated with it (e.g. *MEDGES, edgesleft, tried-mappings, etc.*) so that, when the backtracking takes place, these structures can be restored and the algorithm can operate on them while testing alternative vertex correspondences. Importantly, every time a vertex of G_1 is reached through $u := u+1$, its memory is wiped, i.e. it is allowed to try out any allowed correspondences in G_2. This allows the algorithm to test, given a state *s*, all possibilities of subsequent vertex matching, provided that these possibilities meet the pruning condition. For any further algorithmic details, the reader is appointed to the original work by McGregor[9].

Note that at the first pass, an initial state will progress to completion, as *edgesleft* is set to be the maximum possible and is reset only after the first MCS has been found. It is therefore critical that the first computed solution be the MCS with the highest possible edge score, which will reinforce the effect of the condition *edgesleft(s') > bestedgesleft,* and allow for a substantial reduction in the search space, as discussed in more detail in 2.2.1. The McGregor algorithm, similar to the *VF2* algorithm, has spatial complexity of the order $O(N)$[19] making it suitable for applications on large graphs.

2.1.5 NP-completeness of subgraph-isomorphism and MCS problems

From a computational complexity point of view, subgraph isomorphism and MCS fall in the category of Nondeterministic Polynomial-time (NP)-complete decision problems[7,20-22]. For this class of decision problems, no algorithms of polynomial time complexity are known, however a known (or guessed) solution to the problem can be verified in polynomial time[20].

In the worst case, the time complexity for any exact algorithm applied to this category is exponential. In case of the exact graph matching algorithms for subgraph isomorphism (*VF2*) and MCS (*McGregor*) presented above, the worst case time complexity is of the order *O(N! x N)* and $O(\frac{(N_2+1)!}{(N_2-N_1+1)!})$, respectively[11,21]. For the *VF2* algorithm, *N* stands for the maximum number of vertices of the two graphs, whereas for McGregor's MCS algorithm N_1 and N_2 stand for the number of vertices of graphs G_1 and G_2, respectively. The factorial time complexity may render applications of these algorithms to large graphs (hundreds or thousands of vertices) prohibitive. In practice however, the actual computation time is often acceptable as, in many cases, attributes can be assigned to the vertices and/or edges, which, especially when combined with pruning and search space-ordering heuristics in the implementation, leads to a significant reduction of the search space, as is the case for MAGMA (2.2.3, Chapter 3).

Graph matching and methyl resonance assignment

Here, I present the Methyl Assignment by Graph MAtching protocol, devised from the two exact matching algorithms, *VF2* for monomorphism and McGregor for MCES combined with heuristics to optimize the vertex matching priorities and the order of the data graph traversal, allowing successful applications to large graphs. I discuss details of the implementation and considerations taken in the design, and show that the protocol is robust to graph sparsity and errors in the input data. Furthermore, I show that different vertex labelling and edge sparsity levels influence the number of subgraph isomorphism/MCES solutions found and, consequently, the confidence of vertex assignments. The improvements made herein to the MCES algorithm and the flexibility of the MAGMA protocol can be useful outside of the applications presented in this thesis, and could be particularly interesting for applications on simple, undirected, labelled graphs with around 100 vertices.

2.2 Results and discussion

2.2.1 Theoretical bases of MAGMA

I previously introduced two graph categories relevant to MAGMA, structure and data graphs (2.1.2). Here I will examine in some detail general features of these graphs, as these have important consequences for my choice of graph matching algorithms and their subsequent implementation.

Features of the structure and data graphs

Both the structure graph $G_s = (V_s, E_s, \alpha_s)$ and the data graph $G_d = (V_d, E_d, \alpha_d)$ are defined by the same protein molecule and their features will be determined by the shape and size of the protein, the total number and locations of (isotopically labelled) methyl bearing residues, the diversity of the methyl labelling scheme that determines the variety of vertex labels, and, for the structure graph, the distance threshold used to define the set of edges. In general, the more globular the protein and the more uniformly distributed the methyl groups, the higher the chance for the graph to be connected. On the other hand, the more elongated the protein, the more likely for the graphs to be disconnected with smaller connected components. The connectivity of a graph is also directly influenced by the diversity of the methyl labelling scheme. The more diverse the labelling, the larger in size and denser the graphs, leading to a higher chance for non-adjacent vertices to be linked through a path.

For the structure graph, the density (and hence the connectivity) will directly depend on the distance threshold used, which is set to reflect the expectation for observable NOEs. The longer the threshold, the denser the graph (def. 2.5):

$$\eta_{structure} \propto d \tag{2.23}$$

where $\eta_{structure}$ stands for the density of the structure graph and d for the distance threshold used. Assuming that a correct estimate of the distance threshold is found (see Chapter 3), it can be expected that in general the data graph will be less dense than the structure graph, due

to the $\frac{1}{r^6}$ dependence of the NOE intensity on the inter-atomic distance *r*, and the difficulties inherent to the NOE measurements (see Chapter 1). Therefore:

$$\eta_{data} \leq \eta_{structure} \tag{2.24}$$

Erroneous edges in the data graph can arise from noise and ambiguity in the NOE signals, differences between the available structure of the protein and the conformations sampled in solution, and too short distance threshold used to define the structure graph.

Furthermore, even if a perfect match between the protein construct used for structure determination and that used for NMR measurements is assumed, the data graph may still lack some vertices present in the structure graph ($|V_d| \leq |V_s|$) due to the potential absence of NMR signals caused by, for instance, chemical exchange in the intermediate exchange regime. In contrast, chemical exchange in the slow exchange regime may manifest in the duplication of some methyl peaks, i.e. the vertices of the data graph ($|V_d| \geq |V_s|$). Therefore, in order to relate the data graph to the structure graph by graph matching, and obtain mapping of vertices that solves the methyl assignment problem, the definition of the appropriate graph matching form needs to take into account the considered features and differences of the two graphs.

The relevant exact graph matching forms

Graph isomorphism, with its strict requirement for the structural equality of the two graphs that are compared (def. 2.7), is too "rigid" of an exact graph matching form to be of practical usefulness in MAGMA. Subgraph isomorphism and maximum common subgraph, on the other hand, can sustain some forms of noise, e.g. missing or extra vertices and edges, and are as such more applicable. In the best case, i.e. the available protein structure is representative of that in solution and the distance threshold was determined accurately, a sparse data graph G_d will be isomorphic to a subgraph G_s' of the structure graph ($G_s' \subset G_s$),

$$G_d \cong G_s' \tag{2.25}$$

This will allow the entire set of the data graph vertices and edges to be mapped to the structure graph. Importantly, the mapping function should be an injection (A2.2), and the matching constraint should allow some edges and/or vertices of the structure graph be left out of the mapping (i.e. G_s' need not be a vertex-induced subgraph of G_s). This "weaker form" of subgraph isomorphism, monomorphism, fits the requirements noted above and is used in MAGMA (2.1.3, def. 2.4, 2.9). Critically, not only one, but all monomorphisms need to be found to provide a set of G_d vertex assignments that all explain all experimental inter-methyl NOEs.

Any erroneous edges in the data graph (see above) will preclude monomorphism, as the subset of erroneous edges will have no match in the structure graph. A graph-matching form that allows some edges of the data graph to be left out of the mapping is the maximum common subgraph (2.1.3, def. 2.11). Here, the definition introduced by McGregor[9] is of particular interest, as it defines a *maximal common edge* subgraph H, as a common subgraph isomorphic to a subgraph of G_d ($G_d' \subset G_d$) and to a subgraph of G_s' ($G_s' \subset G_s$) that has the maximum possible number of edges:

$$\begin{aligned} H = MCES(G_d, G_s) \Leftrightarrow H \cong G_d' \wedge H \cong G_s' \\ \wedge \nexists H' \cong G_d' \wedge H \cong G_s' : |E(H')| > |E(H)| \end{aligned} \tag{2.26}$$

Such definition allows the maximum number of experimental inter-methyl NOEs ($E_d' \subset E_d$) to be included in the mapping, without the requirement for G_d' and G_s' to be vertex-induced subgraphs of G_d and G_s, respectively, and is the definition adopted in MAGMA. As noted earlier, the MCES is not necessarily unique, and for the application here it is important to find all MCESs to retrieve all possible vertex mappings that are equally viable ^{1}H-^{13}C resonance assignment options.

Homomorphism (def. 2.10), the matching form that would allow more than one vertex of G_d to be mapped to exactly one vertex in G_s has not yet been considered for MAGMA. However, this form might prove important in the future for experimental datasets where peak doubling occurs in the ^{1}H-^{13}C HMQC/HSQC spectrum. Alternatively, any peak doubling could be assessed experimentally, using 2D EXchange SpectroscopY (EXSY) or ZZ-exchange spectroscopy to identify methyl peaks that are in slow-exchange and join them into a single G_d vertex.

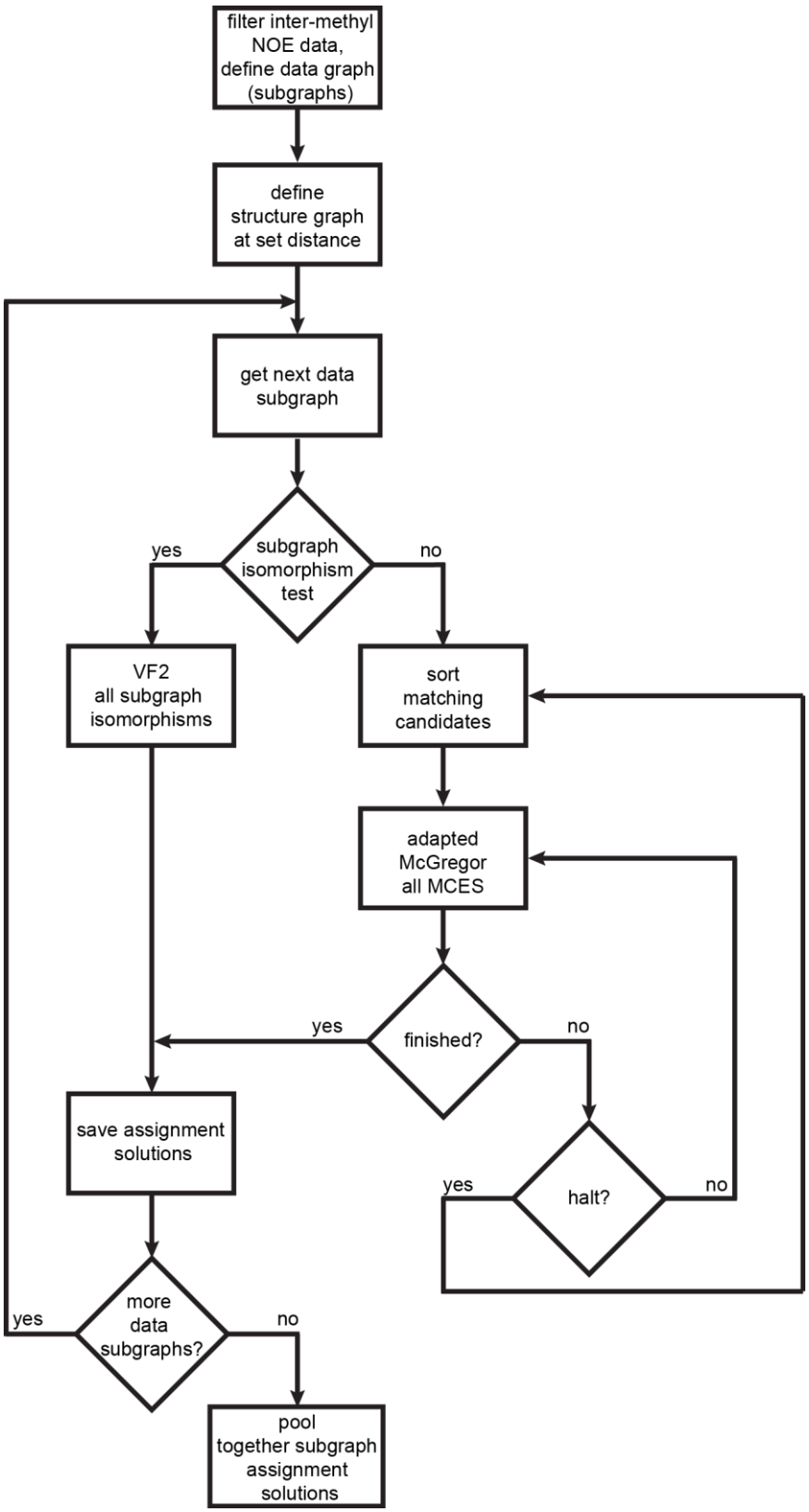
Choice of graph matching algorithms

Exact graph matching problems (2.1.3) can be solved using exact or inexact graph matching algorithms. Exact algorithms, if run to completion, will always find a correct (optimal) solution, whereas inexact algorithms might not find a correct solution and can only guarantee finding an approximate solution. In the design of MAGMA, I opted for two exact matching algorithms – the *VF2* algorithm[10,11,15] for monomorphism, and McGregor's algorithm[9] for MCES.

This choice was motivated by the flexibility of the *VF2* algorithm that allows it to address monomorphism (a weaker form of subgraph isomorphism)[12,13], its favourable space complexity ($O(N)$), making it ideal for applications on large graphs[11], and recent demonstrations of its usefulness in comparisons of graphs defined from protein structures deposited in the PDB[12,13]. McGregor's algorithm, with its focus on maximizing the edge compatibility in matching,[7,9,21] enables exact searches for assignment solutions that explain most of the inter-methyl NOEs. The favourable space complexity $O(N)$ and performance on sparse, attributed graphs[21], further support this choice.

2.2.2 MAGMA protocol - a combination of exact graph matching algorithms and vertex priority sorting heuristics

MAGMA protocol is illustrated below.



1. Defining methyl-methyl edges for the structure graph

Unless indicated differently, pseudo-atoms are created by averaging the coordinates of the two methyl carbon atoms of L (δ_1 and δ_2) and V (γ_1 and γ_2). The matrix of all Euclidean distances between I-δ_1, L-pseudo-δ, V-pseudo-γ, and when applicable, A-β, T-γ, and M-ε, is computed from protein coordinates (provided in a PDB structure file). If more than a single distance is measured for any of the methyl-methyl pairs, the matrix is reduced to the set of shortest distances between any two carbon atoms. The distance cut-off is applied and only the distances smaller or equal to the defined distance threshold are kept (Methods).

This final set of distances defines a set of non-weighted edges for the structure graph (E_s), for which the methyl carbon atoms define a set of vertices (V_s). Each vertex is assigned a label according to its amino acid type (α). The structure graph is then defined as $G_s = (V_s, E_s, \alpha)$. In calculations in which discrimination between *L* and *V* labels is omitted (Figure 2.8C), the same label (*LV)* is assigned to methyls of both L and V residues.

2. Analysis of experimental NMR data

The list of ^{1}H-^{13}C frequencies from an HMQC/HSQC spectrum should be provided in *Sparky* format. The ^{1}H-^{13}C peak names should be given as strings composed of single letter amino acid identifiers followed by a number (e.g. 'I1'). Processed 3D- or 4D-HMQC NOESY spectra must be pre-analyzed using appropriate visualization software (e.g. *Sparky*[23]) and resulting NOESY peak lists need to be provided in *Sparky* format. The identifiers of the auto- and cross-peaks in the NOESY spectrum list should match those from the HMQC/HSQC list.

Alternatively, a two column text file input is supported, in which two ^{1}H-^{13}C peak identifiers are listed as appropriately formatted strings separated by a tab (e.g. 'I1' 'V2'). Each row in the file indicates that the two methyls (e.g. I1, V2) give rise to an NOE. If synthetic data are

given, the two column NOE text file is sufficient, and the ^{1}H-^{13}C HMQC/HSQC list can be an empty file.

If the NOESY peak list is provided with included signal intensities (or data heights), MAGMA will impose the following rules to obtain confident methyl-methyl NOE restraints:

1) An NOE is reciprocated, i.e. there exists a mutual cross-peak in the plane of each ^{1}H-^{13}C auto-peak

2) Cross peaks have similar normalized intensities

3) Cross peaks have signal-to-noise ratios above a given threshold

Normalised cross-peak intensities I_n are calculated between any given methyls *a* and *b* as:

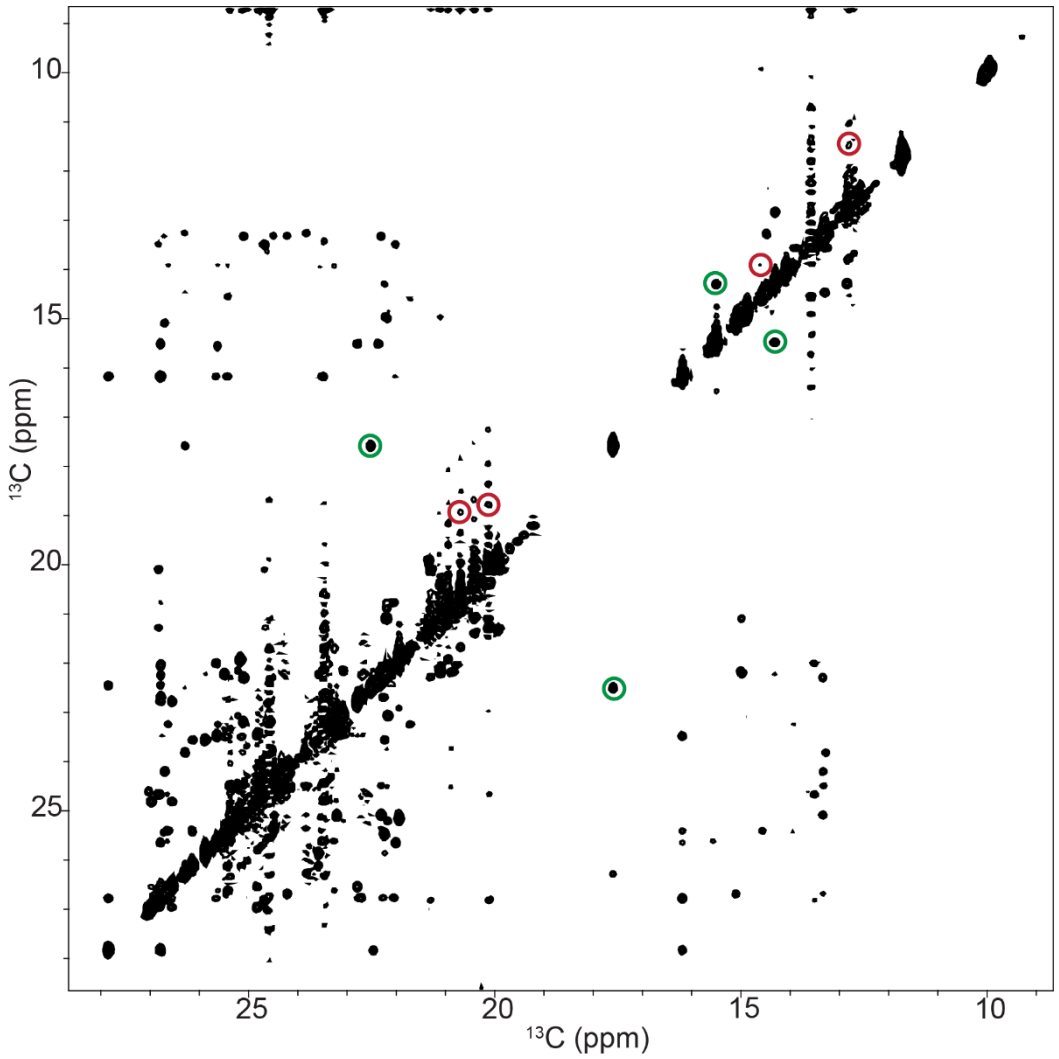
$$I_n^{a,b} = \frac{I_{\text{cross}}^{\text{a,b}}}{\sqrt{I_{\text{diagonal}}^{\text{a}} \, x \, \text{I}_{\text{diagonal}}^{\text{b}}}} \tag{2.27}$$

Signal-to-noise ratios are determined as:

$$\frac{I_{\text{cross /diagonal}}^{\text{a,b}}}{noise\ level}, \tag{2.28}$$

where intensities $I_{\text{cross /diagonal}}^{a,b}$ and *noise level* can be estimates provided by the software for the analysis of NMR spectra (e.g. *Sparky*). The expected similarity between normalized cross-peak intensities $I_n^{a,b}$ and $I_n^{b,a}$ (2.27, rule 2)), is defined by the user and given in MAGMA input file (A2.4, *cross_peak_similarity_percent*). If the similarity percent (%) is set for instance to 80, the program will accept an NOE between two methyls *a* and *b* only if $\frac{|I_n^{a,b}|}{|I_n^{b,a}|} \times 100 \vee \frac{|I_n^{b,a}|}{|I_n^{a,b}|} \times 100 \geq 80$. The similarity rule filters experimental information to

reduce the likelihood of erroneous NOEs entering the calculation, however comes at the expense of total number of utilized experimental restraints. In benchmarking of MAGMA, accuracy of input restraints was prioritized over their total number and hence the rule was applied (Chapter 3). Note however that in the future, user could choose to omit the rule by setting *cross_peak_similarity_pecent* to 0 in the input file (A2.4). An example of the choice or filtering of cross-peaks according to the MAGMA rules is illustrated below on a $^{13}C/^{13}C$ projection of a 3D NOESY data set.



Two examples of reciprocated cross peaks, which also have sufficiently similar normalized intensities, are marked with green circles. In dark red circles are examples of unreciprocated NOEs that are always filtered by MAGMA.

If the two column NOE input file is provided, only the first rule (NOE reciprocity) applies. Finally, if it is known which two methyl groups (pro-*R* and pro-*S*) belong to the same L or V residue, they should be given the same ^{1}H-^{13}C peak identifier, which will allow MAGMA to join their methyl-methyl NOEs to a single vertex in the data graph. As for the structure graph, if the amino acid type of ^{1}H-^{13}C resonances for L and V is unknown, MAGMA supports giving these an *LV* label.

3. Defining the data graph

In MAGMA, appropriately formatted identifiers of ^{1}H-^{13}C frequencies from ^{1}H-^{13}C HMQC/HSQC spectrum are used to define the set of vertices (V_d), and filtered methyl-methyl NOEs are used to define the set of edges (E_d). The vertices in V_d are labelled according to the amino acid types indicated in the input files (α), which finalizes the definition of data graph, $G_d = (V_d, E_d, \alpha)$.

4. Data graph connectivity analysis and vertex removal

MAGMA analyses the connectivity of every data graph to check if it is disconnected (2.1.1). Disconnected data graphs are separated into their connected components (data subgraphs) and the cardinality (the total number of vertices) of each is determined. The subgraphs are then ordered according to their cardinality (from the highest to the lowest) and are henceforth treated separately. Any trivial subgraphs, i.e. vertices of the data graph that are not incident with any edges, are removed from further analysis. Note that the ^{1}H-^{13}C resonances of such vertices present will not be assigned by MAGMA.

5. Subgraph isomorphism (monomorphism) testing

The subgraph isomorphism, or more precisely, monomorphism (def. 2.9) test is performed iteratively on each of the connected data subgraphs, according to the aforementioned order. If a data subgraph passes the test, it is matched to a structure graph using the *VF2* algorithm[11]. The calculation will return all vertex mappings that result in unique subgraph isomorphism. The data subgraphs that do not pass the test are collected and queued for matching using McGregor's MCES algorithm.

6. Heuristics for ordering the MCES search

Before the exact MCES search is executed on each of the remaining connected data subgraphs, each subgraph passes through an iterative protocol that determines:

A) The optimal vertex matching priorities

B) The optimal order for the matching of data (sub)graph vertices

If a data graph is connected, the entire graph passes through the same protocol. From here on, I will refer to every data subgraph as a data graph. Determining A) and B) is critical for obtaining high initial edge correspondence scores in the MCES search. Establishing optimal orders can significantly reduce the search space, as the large-size MCES found early in the search and McGregor's algorithm heuristics prevent the completion of any less optimal partial solution[9] (2.1.4). Good sorting heuristics will also allow for faster convergence of a complete MCES search.

A Sorting the vertex matching priorities

The vertex matching priority is a set of vertices of the structure graph associated with each data graph vertex. All structure graph vertices in the set are assignment options for the data graph vertex. The order in the set determines the order in which alternative assignment

options will be tested for matching. My implementation of McGregor's MCES algorithm supports associating an ordered set of assignment options to each data vertex. The Munkres algorithm[24,25], a variant of the Hungarian method for the assignment problem, and the Jaccard similarity coefficient are used to determine the order.

To demonstrate how MAGMA determines the vertex matching priorities, two vertices are considered, one from a data graph and another from a structure graph with the same label as the first one.

$$u \in V_d\,, v \in V_s : \alpha(u) = \alpha(v) \tag{2.29}$$

MAGMA first computes the neighborhood of *u* and *v*, denoted *N*(*u*) and *N*(*v*), respectively. The neighborhood of a vertex is defined as a set of all vertices adjacent to it. From each neighborhood, MAGMA extracts the set of vertex labels and assigns them to the sets *Nα*(*u*), and *Nα*(*v*), for *u* and *v*, respectively. It then computes the intersection of the sets and determines its size:

$$|N\alpha(u) \cap N\alpha(v)| \tag{2.30}$$

The larger in size the intersection, the more similar the neighborhoods of *u* and *v* will be.

For example, if a vertex $u \in V_d$, *α*(*u*) = *I* is adjacent to three vertices, with labels *I, V, L*, then *Nα*(*u*) = {*I,V,L*}. For a vertex $v \in V_s$, *α*(*v*) = *I* that is directly adjacent to five structure graph vertices with labels *I, I, V, V,* and *L*, then *N*(*v*) = {*I, I, V, V, L*}. The intersection of the two neighborhoods is $N\alpha(u) \cap N\alpha(v)$ = {*I, L, V*}, and its size equals 3.

The neighborhood intersections are computed for every label-compatible pair of vertices from the two graphs. The sizes of the neighborhood intersections then define the elements of

a matrix, whose rows correspond to vertices of a data graph and whose columns correspond to the vertices of a structure graph.

An example of one such matrix of neighborhood intersections is given below. The example features a plausible outcome of the computation of the neighborhood intersections for a data graph and a structure graph with four vertices.

	1'-Ile	**2'-Ile**	**3'-Ile**	**4'-Ile**
1-Ile	3	4	1	3
2-Ile	1	1	2	2
3-Ile	2	5	4	4
4-Ile	4	5	1	1

Given the matrix, an outcome of the Munkres algorithm is in orange. The result is an assignment solution assigning each of the data graph vertices (rows) to one of the structure graph vertices (columns), such that the total sum of neighborhood intersections is maximized (max = 3 + 5 + 2 + 4 = 14). Note that this mapping is not necessarily the only solution to the matching problem. For instance, the same maximum (14) can be obtained by an alternative vertex assignment, as shown below:

	1'-Ile	**2'-Ile**	**3'-Ile**	**4'-Ile**
1-Ile	3	4	1	3
2-Ile	1	1	2	2
3-Ile	2	5	4	4
4-Ile	4	5	1	1

Nonetheless, using the Munkres algorithm as a heuristic for sorting the sets proved highly beneficial, and is a good starting point for the MCES search. Every vertex of a data graph will have its first assignment option set to the one determined by the outcome of the Munkres algorithm. Since this algorithm returns only one assignment possibility, the order for the rest is determined by decreasing Jaccard coefficients, defined as:

$$\frac{|N\alpha(u) \cap N\alpha(v)|}{|N\alpha(u) \cup N\alpha(v)|} \tag{2.31}$$

where $u \in V_d$, $v \in V_s$ and $\alpha(u) = \alpha(v)$. The Jaccard coefficient can be interpreted as a measure of the probability that an element given in at least one of the sets is found in both, and has been used before as a measure of the overlap of sets[26].

A combination of the Munkres algorithm and Jaccard similarity coefficients gives priority orders that take into account global structural similarities of the two graphs and local similarities of the individual vertices. It should be noted that once a complete MCES search starts, i.e. the halting condition is false (see schematic), the determined vertex matching priorities may change. Namely, during the exact MCES search, every time a tentative assignment decision is made for a vertex, the matching priorities of all vertices adjacent to it are updated, i.e. reordered (see below).

B Sorting the order for matching of data graph vertices

McGregor's MCES algorithm constructs a search tree, which has a root representing an assignment for the first vertex of the data graph to one of its option in the structure graph (Figure 2.6). From the root, the order in which the remaining vertices of the data graph are matched to the structure graph can significantly impact the size of the first found MCES, i.e. the edge correspondence score (Figure 2.6). In order to determine the most optimal order for the matching of data graph vertices, a single iteration of the MCES search is performed with the assignment of every data vertex considered for a root. The size of the first MCES found is then used to determine which data graph vertex will be matched first. In order to perform sequential matching of adjacent data graph vertices, the remaining order is determined by a breadth-first search from the first vertex.

After determining the optimal matching order, MAGMA performs several iterations of the MCES search prior to the execution of a complete search. The goal of this short search is to

collect a number of high scoring MCES mapping solutions, which are then used to update the initially determined vertex matching orders. The resulting mappings are inserted to the front of the set of assignment options for every data graph vertex, and the initially determined matching order follows. The final result obtained by a complete run of the algorithm is exact and thus not affected by the described heuristics. However, the considerations to the ordering of the search are critical determinants of the speed of convergence and completion of the search.

7. Complete MCES search

Once the optimal order of matching priorities for all data graph vertices and their matching order have been established, a complete MCES search is executed using the implementation of McGregor's MCES algorithm (2.1.4, Methods). McGregor's MCES algorithm used in steps 6 and 7 of the protocol was implemented as described in the Methods.

8. MAGMA scoring

At the end of the protocol, MAGMA joins all assignment solutions that comprise subgraph isomorphisms and/or MCESs and thereby associates each methyl resonance (vertex in the data graph) to a set of atoms (vertices in the structure graph) to which it was uniquely assigned over a set of all solutions. For each peak, the set of all possible structure graph (PDB) assignments were then found. If multiple assignment options were found, the score given to each peak was defined as the reciprocal of the total number of possible assignments for that peak. Therefore, for a confident assignment, where there is only one possibility, the score is one.

MAGMA usage

MAGMA can be used as a standalone command-line tool on Windows and Linux/Unix operating systems. MAGMA is capable of reading extracted xyz coordinates or PDB files containing single or multiple models, and reading in files containing inter-methyl NOE connectivity information in a two column (for synthetic and simple test cases) or *Sparky* format[23]. An example MAGMA input text file is shown in Figure A2.4, for clarity. MAGMA outputs all methyl-to-residue mappings that result in subgraph isomorphism or MCESs to a two column text file, where methyl identifiers are given in the first column and their residue pairs in the second.

2.2.3 Proof-of-concept

To illustrate how MAGMA's graph matching strategy works and to demonstrate that the approach is well-suited for automatic methyl resonance assignment, two proof-of-concept examples are presented.

A simplified problem

Two simple graphs were defined, each with eight labelled vertices representing eight methyl-labelled residues (Figure 2.7). One of the graphs was sparser than the other to serve as a mimic of a data graph. In addition, the data graph mimic contained an erroneous edge not present in the structure graph mimic (Figure 2.7 B).

Due to the presence of an erroneous edge, the matching was solved using McGregor's MCES algorithm in MAGMA. To generate all solutions to this relatively simple problem, the pruning condition of the algorithm was omitted (see 2.1.4, Figure 2.7 C). Out of 288 possible vertex mappings (assignment solutions) only two define the MCES (Figure 2.7 C and E). The assignment solution is represented in Figure 2.7 E, where a black square shows every confident (i.e. single option) methyl assignment, and multiple squares of lighter colours (i.e. lower confidence) illustrate ambiguous assignments.

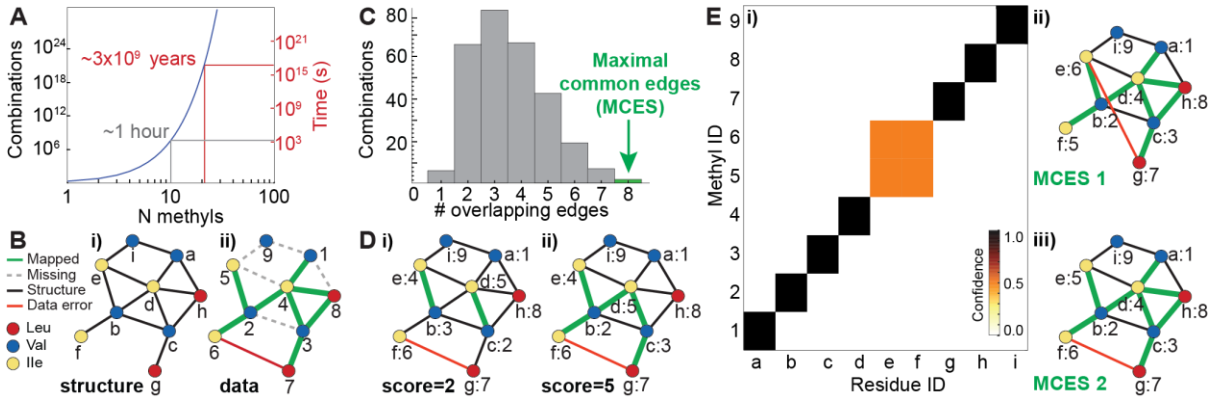


Figure 2.7 | Illustration of MAGMA's graph matching strategy on two small graphs mimicking a data and structure graph. **A)** If the methyl assignment problem was addressed by brute force, the factorial number of combinations (blue line) would render the approach unfeasible. Even if each assignment combination takes 10^{-3} seconds to evaluate, the assignment of 10 residues would take ~1 hr. However, even a small increase in the number of methyls leads to prohibitive combinatorics. **B**) Graph diagrams of illustrative structure (i) and data graphs (ii) with 9 vertices (methyls) and 13 and 9 edges, respectively. The data graph has fewer edges than the structure graph (dotted lines) and contains a deliberate error (red line) - i.e. an edge not present in the structure graph. Common edges between the two graphs are indicated on the data graph in green. (**C**) There are 288 vertex mapping combinations between the two small graphs that are a complete set of possible assignments. These can be classified by the number of common edges, and the two combinations with the highest number of overlapping edges (8 edges; MCES) are indicated in green. **(D)** Two examples of suboptimal assignments; edge score = 2 (i) and edge score = 5 (ii). **(E)** The assignment result obtained by MAGMA, indicating the correspondence between the vertices of the structure graph (*x*-axis) and the vertices of the data graph (*y*-axis). Two maximum common edge subgraphs (ii, iii) explain eight data graph edges (inter-methyl NOEs). Joining their assignments gives the final result, where seven vertices (methyl resonances) are confidently assigned, and two are ambiguous (orange). The correct assignment follows the $x = y$ line.

Cyclin-dependent kinase 2

To demonstrate the effectiveness of MAGMA on larger graphs with features closer to the experimental reality, I performed a series of proof-of-principle calculations using synthetic inter-methyl NOE restraints. The structure of 34 kDa human Cyclin-dependent kinase 2 (CDK2, PDB ID: 2C6O)[27] was selected, and the maximum distance threshold r for the observation of an inter-methyl NOE was set to 8 Å. The same distance was used to define the edges of different structure graphs, in which their vertices were defined by several different methyl labelling schemes, including I-δ, L-δ, V-γ and A-β labelling (Figure 2.8). Increasingly sparse data graphs were obtained by randomly removing edges from the structure graphs, and

the edge removal was weighted to enhance the probability of retaining shorter distances (Methods).

Increasing the sparsity of the data graphs reduced the number of vertices (i.e. methyls) that could be confidently assigned (Figure 2.8 C). This significant increase in assignment ambiguity is expected, as increased removal of edges can result in disconnections in the data graphs. In such scenarios, the small connected data subgraphs can then be mapped to structure graph in multiple ways, resulting in the ambiguity of vertex assignments. Notably, having a larger number of graph labels, i.e. methyl groups with different residue labels, increased the success of the method. Synthetic data with I-δ, L-δ, V-γ, and A-β methyls (vertices) provided a higher proportion of confident assignments than those with vertices labelled in only I-δ, L-δ and V-γ. In addition, being able to distinguish L from V labels, an experimentally achievable goal[28], was advantageous (Figure 2.8 C). All vertex (methyl) assignments solutions generated by MAGMA for the two data graphs of different sparsity (30% and 60% edges removed) are shown in Figure 2.8 B. Importantly, for each labeling scheme and data graph sparsity tested, the correct vertex assignment was always sampled within the set of ambiguous options and all confident (single option) assignments were correct (Figure 2.8 B, C).

Finally, I compared the performance of MAGMA on the synthetic datasets to two other freely available automated methyl assignment approaches, FLAMEnGO2.0[29] and MAP-XSII[30] (Supplementary, Figure A2.3). On these data sets, MAGMA outperformed both approaches in assignment accuracy and provided 50% and 16.6% more correct, confident assignments than FLAMEnGO2.0 and MAP-XSII, respectively. FLAMEnGO2.0 and MAP-XSII were respectively incorrect in 49% and 41% of their confident methyl assignments.

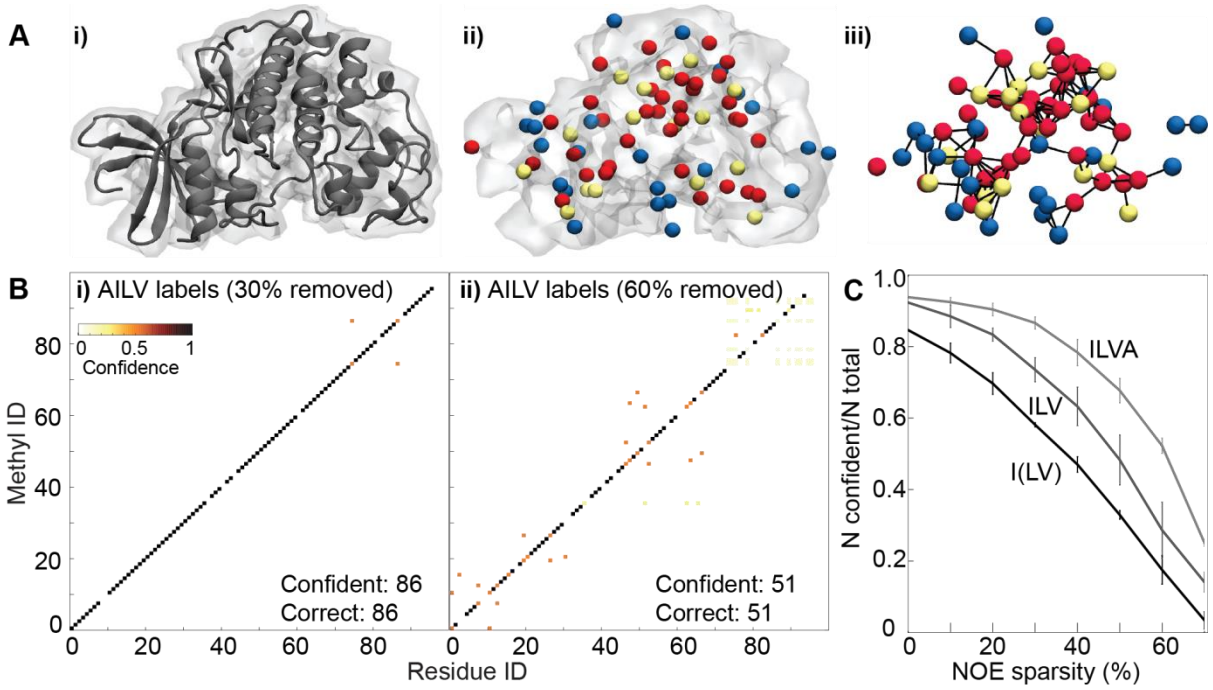


Figure 2.8 | Performance of MAGMA on synthetic data graphs (**A**) i) Secondary structure of CDK2.[27] ii) Locations of carbon atoms of I-δ1 (yellow), L-δ_1/δ_2 (red) and V-γ_1/γ_2 (blue) are shown as spheres. iii) Atoms from ii) within an 8 Å distance are connected by edges (black) giving a representation of a structure graph. (**B**) AILV-labelled synthetic data graphs were generated from the structure graph by the removal of 30% and 60% of inter-methyl edges for i) and ii), respectively. The assignment results obtained by MAGMA are shown. The correct assignment follows the $x = y$ line. Confident assignments are single option methyl assignments in black squares. (**C**) Results of a comparison of synthetic data graphs featuring different labelling schemes and sparsity levels. The (LV) label denotes labelling in which the L and V labels are merged. Even on highly sparse data graphs, a large fraction of confident (single option) assignments can be obtained. Increasing the diversity of labels increases the success of MAGMA (*ILVA*, light grey).

2.3 Conclusions and outlook

Alternatives to the exact algorithms implemented in MAGMA exist. For graph-subgraph isomorphism, a number of exact algorithms based on tree search are known. The most popular ones include the early Ullmann's algorithm and its recent improvement[17,31], and two versions of the *VF2* algorithm by Cordella *et al*[10,11]. Most recently *VF2 Plus*, an improvement to *VF2*, has been presented with an application featuring protein graphs defined on structures extracted from the PDB[12,13]. This version of the algorithm may be a useful addition to MAGMA in the future. A review of a number of exact algorithms for graph isomorphism and graph-subgraph isomorphism and their performance comparison can be found in works by Foggia *et al.*[32] and De Santo *et al.*[16]. Notably, next to the well-known tree-based algorithm, alternative algorithmic approaches to graph-subgraph isomorphism using constrained programming have been suggested[33-36]. These approaches propose matching constraints that could be useful additions to any graph-subgraph matching algorithm.

The MCS is the most computationally demanding form of exact matching. The most commonly employed algorithmic approaches include depth-first tree search with backtracking[9,18] or a reduction of the problem to finding a maximal/maximum clique after defining a *modular product graph* (for a definition see Levi[37] or Willett *et al.*[38]) of the two input graphs[37,39-41 42] (Appendix 2.1, Figure A2.2). Different MCS algorithms have most recently been reviewed by McChreesh *et al.*[36]

The exact MCS algorithms are limited to relatively small graphs[5], which is why they have so far mostly found applications in assessing the similarity of chemical structures of small molecules. For instance McGregor[9], Cone *et al.*[43], and Durand *et al.*[39] developed their algorithms for applications on small molecules. In the field of cheminformatics, the MC(E)S is investigated as a metric of molecular similarity. For instance, in a publication by Raymond

et al.[44], a comparison of molecular similarity derived from the MCES and fingerprint-based measures, typically employed in chemical databases,[45,46] was made. In work from Gardiner *et al.*[47], six algorithms for maximal clique detection [40,48] [41,42,49,50] were compared for their efficiency in finding the MCS on graphs of small molecules and proteins, albeit restricted to a representation of the secondary structure elements. Applications of the MCS for structural comparisons of protein graphs have so far mostly featured approximate approaches, or those restricted to a coarse representation and assessment of global fold similarities[51], probably due to their large graph sizes. The topics of usefulness of the MC(E)S in chemistry and biology have recently been reviewed[38,51].

The performance of three exact MCS algorithms - McGregor[9], Durand-Pasari[39], and Balas-Yu[41] - on a large database featuring diverse, labelled graphs was assessed by Conte *et al.*[21]. The study concluded that, depending on the structural features of the graphs, different algorithms show advantages. Importantly, McGregor's algorithm was the best of the three in its performance on sparse graphs and graphs with low vertex degrees (i.e. number of adjacent vertices). However, on dense, large graphs Balas-Yu is the fastest algorithm, due to its efficient heuristics[21]. The addition of this algorithm to MAGMA might be a useful future extension.

Finally, I should note that the exact graph matching problems addressed by MAGMA are NP-complete and their worst-case time complexity is exponential (more precisely factorial), which challenges their usefulness on very large and dense graphs. Therefore in the future, some of the numerous approximate graph matching approaches[4] may be worth exploring.

2.4 Methods

MAGMA implementation

MAGMA is written in the Python programming language and supported by the *igraph* high performance graph library *python-igraph0.7*[52], *SciPy* library[53], and Python packages *networkx-1.9*[54], and *munkres1.0.8*. MAGMA uses the *igraph* implementation of the *VF2* algorithm[10,11] and the *munkres1.0.8* implementation of the Munkres algorithm[24]. Analyses and plotting of graphs in MAGMA are supported by *networkx-1.9*[54] and the *Mayavi* application[55]. The MAGMA program together with documentation for usage will be available for download at http://magma.chem.ox.ac.uk/.

McGregor MCES algorithm implementation

The MCES algorithm based on the work from McGregor[9] was implemented in MAGMA in Python. To make the algorithm run with an acceptable time on large graphs, and to minimize its space complexity, I considered data structures used in the computation. The core data structures of the algorithm are:

- The matrix of edge correspondences (*MEDGES*), with rows and columns indexed by enumerated data and structure graph edges, and dimensions determined by $|E_d|$ and $|E_s|$, respectively.
- Vectors, containing vertices of the data and structure graphs, with dimensions that were determined by $|V_d|$, and $|V_s|$, respectively.
- Two hash tables, one containing the current vertex mapping *M*; and one containing the vectors of *matching_priority* and *tried_vertices* associated with every data graph vertex (see 2.2.2). The size of these structures varied with the requirement of values stored in them, but never exceeded the number of the structure graph vertices ($|V_s|$).

The *MEDGES* and, where applicable, vectors, are defined either as regular *NumPy* arrays or boolean (“masked”) *NumPy* arrays to allow advanced and fast indexing options[56]. Copies of

all data structures of a given state *s* in the search are stored in the memory associated with the last matched data vertex, just before testing the options for a subsequent data vertex. This is necessary so that the data structures can be restored upon backtracking (see 2.1.4). If the data graph is disconnected, the data structures are prepared iteratively for every connected data subgraph.

The following modifications to the original McGregor's algorithm were made:

1. The operator > in the pruning conditions *edgesleft* > *bestedgesleft* was replaced by ≥ to account for the collection of all MCES
2. Resorting of vertex matching priorities (*matching-priority*) followed each tentative assignment decision, as proposed by McGregor[9], and was, where appropriate, supplemented with the originally determined matching priorities (MAGMA's heuristics; 2.2.2)

To illustrate modification 2, two graphs are defined - G_1 and G_2; $G_1 = (V_1, E_1, \alpha)$ and $G_2 = (V_2, E_2, \alpha)$, such that $|V_1| \leq |V_2|$. When searching for the MCES (G_1, G_2), every time a vertex $u \in V_1$ is tentatively assigned to a vertex $v \in V_2$, the matching priority of every subsequent vertex $k \in V_1$, which has not yet been tentatively mapped given the current search state *s*, is updated according to the following rule:

$$\forall\, k \in V_1 : k > u \wedge k \in V_1 \cap adj(u)$$
$$matching_priority\,(k) := matching_priority(k) \cap \{l : l \in adj(v),\ l \in V_2 \tag{2.32}$$

The update in the matching priorities ensures that given the current partial mapping, *M(s)* is extended by mapping of the adjacent vertices of G_1 to the adjacent vertices of G_2. Any vertices contained in the original matching priorities that were determined by MAGMA's heuristics (2.2.2) and are not in the intersection defined by def. 2.32 and not in the current partial mapping *M(s)* will follow the updated vertices (i.e. these are appended to the end of

the queue). Note that the change in the pruning condition in 1 makes the algorithm less restrictive. The increased sampling requirement can result in longer execution times.

MAGMA input parameters

The input parameters that need to be specified for a MAGMA calculation are:

- The distance threshold (Å) to define a structure graph
- The number of iterations of *step 6* of the protocol (Appendix 2.4, *priority_iterations*)
- The number of MCESs to collect for *matching_priority* resorting (Appendix 2.4, *n_priority_mces*)

The input parameters are read by the program from the input text file. The input text file is prepared on the basis of a template (Appendix 2.4). Setting *priority_iterations* between 10 and 50 is recommended, where higher values are more suitable for large data graphs. Setting *n_priority_mces* between 3 and 20 is recommended.

CDK2 – definition of the structure and data graphs

To provide proof-of-principle results with MAGMA, data were simulated from the crystal structure of the protein CDK2 (PDB ID: 2C6O)[27] for a variety of methyl labeling schemes and sparsity levels (Fig 3). A structure graph was determined at a threshold of 8 Å. Data graphs were then generated from this structure graph by removing a percentage of edges (Figures 2.8, A2.3). Edges were removed by random sampling without replacement from the structure graph, and the removal of individual edges was weighted by the probability $\frac{r_i^6}{\Sigma r_i^6}$, where r_i is the distance of any given edge i. Longer distances are therefore removed with a higher probability, such that the final distributions of inter-methyl restraints versus length resembles experimental data. The assignment result from MAGMA will depend on exactly which inter-methyl restraints (structure graph edges) have been removed, the sparsity of the

generated data graph, and the methyl labeling scheme. MAGMA calculations were run on multiple inter-methyl restraint lists generated by changing the seed of the random number generator to calculate the mean and the standard deviation of confidently assigned methyls.

2.5 References

1 Gross, J. L., Yellen, J. *Graph Theory and its Applications*. Second edn, 779 (CRC Press Taylor&Francis Group, 2006).

2 Bondy, J. A., Murty U.S.R. *Graph theory with applications*. 9 (Elsevier Science Ltd/North-Holland, 1976).

3 Bollobas, B. *Graph theory: an introductory course*. Vol. 63 (Springer Science & Business Media, 2012).

4 Foggia, P., Percannella, G. & Vento, M. Graph matching and learning in pattern recognition in the last 10 years. *International Journal of Pattern Recognition and Artificial Intelligence* **28**, 1450001, doi:doi:10.1142/S0218001414500013 (2014).

5 Conte, D., Foggia, P., Sansone, C. & Vento, M. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence* **18**, 265-298, doi:doi:10.1142/S0218001404003228 (2004).

6 Bunke, H. Graph Matching: Theoretical Foundations, Algorithms, and Applications. *International Conference on Vision Interface*, 82-84 (2000).

7 Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C., Vento, M. in *Structural, Syntactic, and Statistical Pattern Recognition* (ed Adnan Amin Terry Caelli, Robert P. W. Duin,Dick de Ridder,Mohamed Kamel) Ch. 10, pp 123-132 (Springer Berlin Heidelberg, 2002).

8 Koch, I. Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science* **250**, 1-30, doi:http://dx.doi.org/10.1016/S0304-3975(00)00286-3 (2001).

9 McGregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software Pract Exper* **12**, 23-34, doi:DOI 10.1002/spe.4380120103 (1982).

10 Cordella, L., Foggia, P., Sansone, C., Vento, M. An improved algorithm for matching large graphs. *3rd IAPR-TC15 workshop on graph based representations in pattern recognition*, 149-159 (2001).

11 Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *Ieee T Pattern Anal* **26**, 1367-1372, doi:Doi 10.1109/Tpami.2004.75 (2004).

12 Carletti, V., Foggia, P. & Vento, M. in *International Conference on Image Analysis and Processing*. 409-417 (Springer).

13 Carletti, V., Foggia, P. & Vento, M. in *Graph-Based Representations in Pattern Recognition: 10th IAPR-TC-15 International Workshop, GbRPR 2015, Beijing, China, May 13-15, 2015. Proceedings* (eds Cheng-Lin Liu, Bin Luo, G. Walter Kropatsch, & Jian Cheng) 168-177 (Springer International Publishing, 2015).

14 Nilsson, N. J. *Principles of artificial intelligence*. (Morgan Kaufmann, 2014).

15 Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. in *Image Analysis and Processing, 1999. Proceedings. International Conference on.* 1172-1177 (IEEE).

16 De Santo, M., Foggia, P., Sansone, C. & Vento, M. A large database of graphs and its use for benchmarking graph isomorphism algorithms. *Pattern Recognition Letters* **24**, 1067-1079, doi:http://dx.doi.org/10.1016/S0167-8655(02)00253-2 (2003).

17 Ullmann, J. R. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)* **23**, 31-42 (1976).

18 Golomb, S. W. & Baumert, L. D. Backtrack Programming. *J Acm* **12**, 516-&, doi:Doi 10.1145/321296.321300 (1965).

19 Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C. & Vento, M. in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR).* 123-132 (Springer).

20 M.R. Garey, D. S. J. (W. H. Freeman & Co. New York, NY, USA, 1970-2002).

21 Conte, D., Foggia, P., Vento, M. Challenging Complexity of Maximum Common Subgraph Detection Algorithms: A Performance Analysis of Three Algorithms on a Wide Database of Graphs. *Journal of Graph Algorithms and Applications* **11**, 99-143 (2007).
22 Johnson, D. S. The Np-Completeness Column - an Ongoing Guide. *J Algorithm* **6**, 291-305, doi:Doi 10.1016/0196-6774(85)90046-X (1985).
23 Goddard, T. D. K., D. G. SPARKY 3. *University of California, San Francisco*.
24 Munkres, J. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics* **5**, 32-38, doi:doi:10.1137/0105003 (1957).
25 Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2**, 83-97, doi:10.1002/nav.3800020109 (1955).
26 Levandowsky, M. & Winter, D. Distance between Sets. *Nature* **234**, 34-35 (1971).
27 Richardson, C. M., Williamson, D.S., Parratt, M.J., Borgognoni, J., Cansfield, A.D., Dokurno, P., Francis, G.L., Howes, R., Moore, J.D., Murray, J.B., Robertson, A., Surgenor, A.E., Torrance, C.J. Triazolo[1,5-a]pyrimidines as novel CDK2 inhibitors: Protein structure-guided design and SAR. *Bioorganic & Medicinal Chemistry Letters* **16**, 1353-1357, doi:http://dx.doi.org/10.1016/j.bmcl.2005.11.048 (2006).
28 Lichtenecker, R. J., Coudevylle, N., Konrat, R. & Schmid, W. Selective isotope labelling of leucine residues by using alpha-ketoacid precursor compounds. *Chembiochem : a European journal of chemical biology* **14**, 818-821, doi:10.1002/cbic.201200737 (2013).
29 Chao, F. A., Kim, J., Xia, Y., Milligan, M., Rowe, N., Veglia, G. FLAMEnGO 2.0: an enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J Magn Reson* **245**, 17-23, doi:10.1016/j.jmr.2014.04.012 (2014).
30 Xu, Y., Liu, M., Simpson, P.J., Isaacson, R., Cota, E., Marchant, J., Yang, D., Zhang, X., Freemont, P., Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J Biomol NMR* **55**, 179-187, doi:10.1007/s10858-012-9700-z (2013).
31 Ullmann, J. R. Bit-vector algorithms for binary constraint satisfaction and subgraph isomorphism. *J. Exp. Algorithmics* **15**, 1.1-1.64, doi:10.1145/1671970.1921702 (2011).
32 Foggia, P., Sansone, C. & Vento, M. in *Proceedings of the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition.* 188-199.
33 Zampelli, S., Deville, Y. & Solnon, C. Solving subgraph isomorphism problems with constraint programming. *Constraints* **15**, 327-353 (2010).
34 Larrosa, J. & Valiente, G. Constraint satisfaction algorithms for graph pattern matching. *Mathematical structures in computer science* **12**, 403-422 (2002).
35 Dahm, N., Bunke, H., Caelli, T. & Gao, Y. in *Pattern Recognition (ICPR), 2012 21st International Conference on.* 1164-1167 (IEEE).
36 McCreesh, C., Ndiaye, S. N., Prosser, P. & Solnon, C. in *International Conference on Principles and Practice of Constraint Programming.* 350-368 (Springer).
37 Levi, G. A note on the derivation of maximal common subgraphs or two directed or undirected graphs. *CALCOLO* **9**, 341-352 (1973).
38 Willett, P. Maximum Common Subgraph Isomorphism Algorithms: A Review. *MATCH Communications in Mathematical and in Computer Chemistry* (2016).
39 Durand, P. J., Pasari, R., Baker, J. W. & Tsai, C. C. An efficient algorithm for similarity analysis of molecules. *Internet J Chem* **2**, art. no.-17 (1999).
40 Bron, C. & Kerbosch, J. Finding All Cliques of an Undirected Graph [H]. *Commun Acm* **16**, 575-577, doi:Doi 10.1145/362342.362367 (1973).

41 Balas, E. & Yu, C. S. Finding a Maximum Clique in an Arbitrary Graph. *Siam J Comput* **15**, 1054-1068, doi:Doi 10.1137/0215075 (1986).
42 Carraghan, R. & Pardalos, P. M. An Exact Algorithm for the Maximum Clique Problem. *Oper Res Lett* **9**, 375-382, doi:Doi 10.1016/0167-6377(90)90057-C (1990).
43 Cone, M. M., Venkataraghavan, R. & Mclafferty, F. W. Molecular-Structure Comparison Program for Identification of Maximal Common Substructures. *J Am Chem Soc* **99**, 7668-7671, doi:DOI 10.1021/ja00465a041 (1977).
44 Raymond, J. W., Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of computer-aided molecular design* **16**, 59-71 (2002).
45 Raymond, J. W., Gardiner, E. J. & Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *Journal of chemical information and computer sciences* **42**, 305-316 (2002).
46 Raymond, J. W., Gardiner, E. J. & Willett, P. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *The Computer Journal* **45**, 631-644 (2002).
47 Gardiner, E. J., Artymiuk, P. J. & Willett, P. Clique-Detection Algorithms for Matching Three-Dimensional Molecular Structures. *Journal of Molecular Graphics and Modelling* **15**, 245-253, doi:http://dx.doi.org/10.1016/S1093-3263(97)00089-2 (1997).
48 Babel, L. Finding maximum cliques in arbitrary and in special graphs. *Computing* **46**, 321-341 (1991).
49 Gendreau, M., Picard, J.-C. & Zubieta, L. in *Advances in Optimization and Control: Proceedings of the Conference "Optimization Days 86" Held at Montreal, Canada, April 30 – May 2, 1986* (eds H. A. Eiselt & G. Pederzoli) 79-91 (Springer Berlin Heidelberg, 1988).
50 Shindo, M. & Tomita, E. A Simple Algorithm for Finding a Maximum Clique and Its Worst-Case Time Complexity. *Systems and Computers in Japan* **21**, 1-13, doi:10.1002/scj.4690210301 (1990).
51 Ehrlich, H. C. & Rarey, M. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 68-79 (2011).
52 Csardi, G., Nepusz, T. The igraph software package for complex network research. *Inter Journal Complex Systems, 1695* (2006.).
53 Jones, E., Oliphant, T. & Peterson, P. SciPy: Open source scientific tools for Python. *http://www. scipy. org/* (2001).
54 Hagberg, A., Schult DA.,Swart, PJ. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, 11-15 (2008).
55 Ramachandran, P. & Varoquaux, G. Mayavi: 3D Visualization of Scientific Data. *Comput Sci Eng* **13**, 40-50, doi:Doi 10.1109/Mcse.2011.35 (2011).
56 Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* **13**, 22-30 (2011).
57 Han, B., Liu, Y. F., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR* **50**, 43-57, doi:10.1007/s10858-011-9478-4 (2011).
58 Sahakyan, A. B., Vranken, W. F., Cavalli, A. & Vendruscolo, M. Structure-based prediction of methyl chemical shifts in proteins. *Journal of biomolecular NMR* **50**, 331-346, doi:10.1007/s10858-011-9524-2 (2011).

Appendix

A2.1 List of mathematical symbols

≅	Isomorphic/structurally identical
≡	Identical to
∈	Is an element of
⇒	If … then
⇔	If and only if
→	From … to
⊆	Subset of
∩	Intersect
∨	Or
∧	And
:	Such that
∀	For every
∃	There exists
∃!	There exists exactly one
∅	Empty set

A2.2 Definitions

Injection (also called an injective function): a function that maps every element of the domain to at most one unique element of the codomain.

Surjection (also called a surjective function): a function that maps at least one element of the domain to every element of the codomain.

A function is a **bijection** if it is both an injection and a surjection.

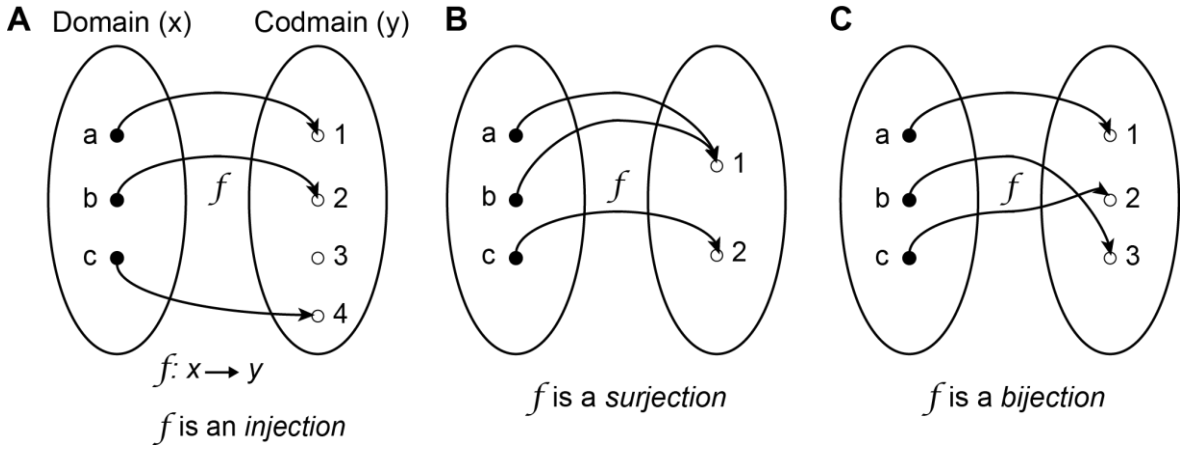


Figure A2.1 | An illustration of an injection (**A**), surjection (**B**) and bijection (**C**).

Clique

A clique C of a simple graph $G = (V, E, \alpha)$, is a subgraph of G ($C \subseteq G$) such that all of its vertices are adjacent, i.e. a complete subgraph of G. A clique is *maximal* if it cannot be further extended by the addition of a vertex that is adjacent to the vertices in the clique. A clique is *maximum,* if there are no other cliques of G that have more vertices. Note that a maximal clique can be a subset of a larger one. All cliques of a simple graph are illustrated on its graphical representation below (Figure A2.2).

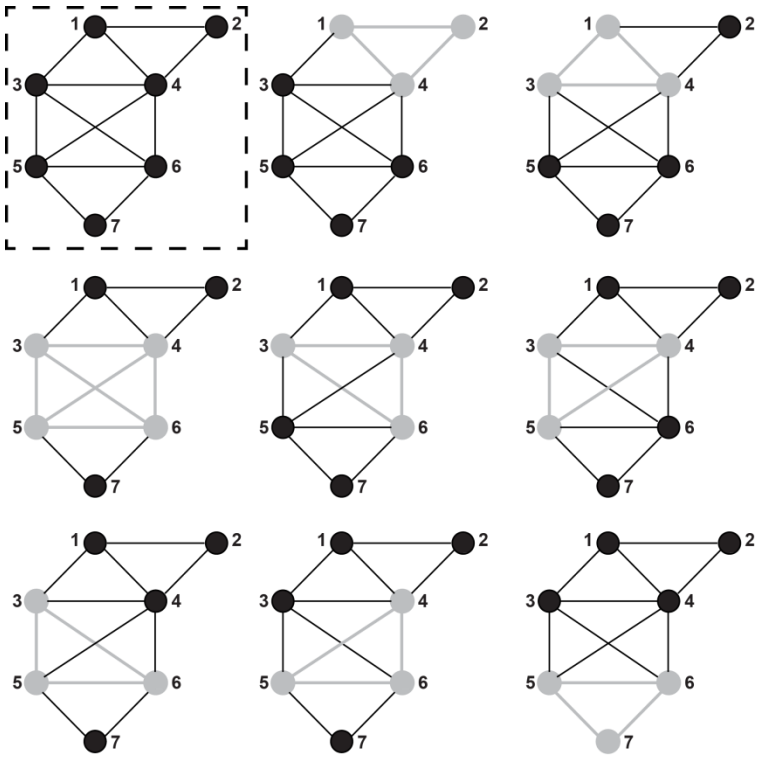


Figure A2.2 | An illustration of a simple graph (in the rectangle) and all of its cliques. The *maximum* clique is the left-most in the middle row. The two cliques in the top row and the right most clique in the bottom row are *maximal.*

A2.3 Comparison of three algorithms for automatic methyl resonance assignment on synthetic dataset

In addition to a protein structure and a list of inter-methyl NOEs, FLAMEnGO2.0 and MAP-XSII require additional inputs: a list of experimental ^{1}H-^{13}C chemical shifts and a list of predicted ^{1}H-^{13}C chemical shifts[29,30]. For a performance comparison to MAGMA, one set of chemical shifts was predicted by ShiftX2[57] ("experimental"), and another set was calculated by CH3Shift[58] (predicted). Both sets of chemical shifts were calculated on the basis of the same protein structure (PDB ID: 2C6O). The list of synthetic inter-methyl NOEs was provided by removing 50% of edges from the structure graph (at 8 Å) with I-δ, L-δ, and V-γ methyl-labelling. The NOE lists for the two algorithms were prepared from the remaining inter-methyl distances and the "experimental" methyl chemical shifts (see above). The number of correct/incorrect assignments was 36/0 for MAGMA. FLAMEnGO2.0 and MAP-XSII provided 18/17 and 30/21 correct/incorrect assignments, respectively (Figure A2.3).

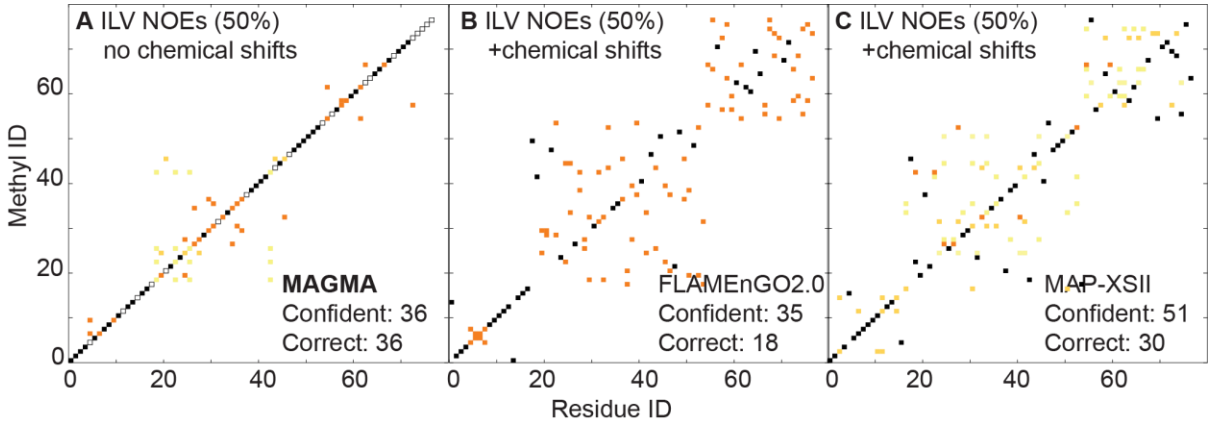


Figure A2.3 | Comparison of the performance of MAGMA, FLAMEnGO2.0,[29] and MAP-XSII[30] on a synthetic ILV-labelled dataset with 50% of the structure graph edges removed. In white squares are the methyls (vertices) that have no edges incident with them, and hence cannot be assigned by MAGMA (see 2.2.2, step 4).

These results demonstrate that MAGMA can provide accurate and reliable methyl assignments.

A2.4 An example of a MAGMA input text file

```
#       PDB structure part
pdb_file        my_pdb_file.pdb
chains  A
residues        ILE VAL LEU
distance_threshold      10
#       NMR part
HMQC_list       hmqc.list
format_peak_list        sparky
labelled_residues       ILE VAL LEU
read_NOEs_from_file     off
NOE_list        methyl_noe.list
format_NOE_list         sparky
dimensions_NOEs         3
read_extracted_NOEs     off
extracted_NOE_list      two_column_noe_file.txt
check_NOE_reciprocity           on
check_NOE_intensity     on
cross_peak_similarity_percent           80
#       System dependent requirements
run_subgraph_isomorphism        on
priority_iterations     20
n_priority_mces         5
#       Results/outputs
output_file_McGregor            mces_assignment_results.txt
output_file_vf2         si_assignment_results.txt
merge_LV_label          off
merge_NOEs      on
merge_stereo_pairs      on
min_size        1
mode    connected_subgraphs
run_at  short
mces_mode       all
maximum_run_time        100
```

Chapter 3: Testing MAGMA on experimental datasets

Abstract

Due to their favourable spectroscopic properties and abundance in proteins, methyl groups have become the NMR probes of choice for studying the structure and dynamics of high molecular mass biomolecular assemblies. The widespread application of methyl-based NMR techniques has been limited due to the challenges and costs associated with the assignment of methyl resonances. Methyl Assignment by Graph MAtching (MAGMA) is a novel method for the automatic assignment of methyl resonances based on exact graph matching algorithms that compare inter-methyl connectivity patterns, which are extracted from either solution-state (NOESY) or solid-state (DREAM) NMR spectra, to inter-methyl connections within a known protein structure. Given the input restraints and a distance threshold for the calculation, MAGMA generates a set of all theoretically possible assignments for each methyl resonance. Here, to my knowledge, I assembled the largest benchmark for the testing of automatic methyl resonance assignment protocols, composed of experimental NMR datasets and high-resolution structures of eight proteins, for which cross-validated methyl assignments were available. The benchmark features a range of molecular masses, three-dimensional shapes, methyl labelling schemes, and sparsity of inter-methyl distance restraints. On this benchmark, MAGMA provides 100% accuracy in confident assignments and assigns a minimum of 31% and a maximum of 89% of methyl-bearing residues. A comparison to two alternative automatic methyl resonance assignment protocols, FLAMEnGO2.0 and MAP-XSII, shows MAGMA's advantages in accuracy and better assignment coverage on the three largest proteins in the benchmark. Considerations are given to the optimality of the distance threshold for the calculation and differences in input protein structures, and their effects on the assignment results. MAGMA promises to facilitate wider applications of methyl-based NMR experiments in both academia and industry.

3.1 Introduction

3.1.1 Methyl side-chain assignment strategies

NMR spectroscopy is the only experimental method capable of simultaneously probing both the structure and dynamics of biomolecules at atomic resolution[1-8]. For decades, solution-state NMR spectroscopy has been successfully applied to small proteins up to ~40 kDa in molecular mass[2]. However, 70-80% of eukaryotic proteins assemble into larger oligomers that exceed this molecular mass threshold and are thus challenging to study using 'traditional' NMR methods[2,9]. This is due to the fast decay of coherences caused by rapid spin relaxation and limited resolution produced by overlapping resonances in crowded NMR spectra of large systems[10]. These limitations have been removed through the combination of advanced nuclear spin labeling strategies[11-21] and optimised pulse sequences[22] (Chapter 1). Of particular note, NMR studies using ^{13}C labelled methyl groups ($^{13}CH_3$) in the side-chains of otherwise deuterated proteins[11-21] now allow for detailed structural and dynamical studies of protein complexes up to 1 MDa in molecular mass using methyl-TROSY protocols[23-25, 26].

A major bottleneck for methyl-TROSY, however, is to relate the observed ^{1}H-^{13}C coherences in a heteronuclear multiple quantum coherence (HMQC) spectrum (Figure 3.1 B) to specific methyl groups within the molecule, a challenge known as spectral (or resonance) assignment. For smaller proteins, backbone and aliphatic ^{13}C resonances are assigned using approaches that correlate the backbone and side-chain peak positions[27,28]. However, the linewidths of non-methyl side-chain and backbone nuclei become increasingly broad in high molecular mass molecules, for which these correlation methods have limited applicability. Therefore, NMR signals from methyl groups in high molecular mass molecules are usually assigned either by transferring the assignments obtained from smaller protein fragments to spectra of the intact protein[25,29], an approach known as 'divide and conquer'[25], or by monitoring spectral

changes after targeted site-directed mutagenesis (SDM)[30] (Figure 3.1). Both approaches are laborious, time consuming, and expensive. Moreover, they require that the isolated domains or introduced mutations do not lead to major structural rearrangements, such that assignments of the spectra acquired from divide and conquer or SDM can faithfully be transferred to that of the intact protein. These obstacles discourage research groups from routinely applying methyl-TROSY methods.

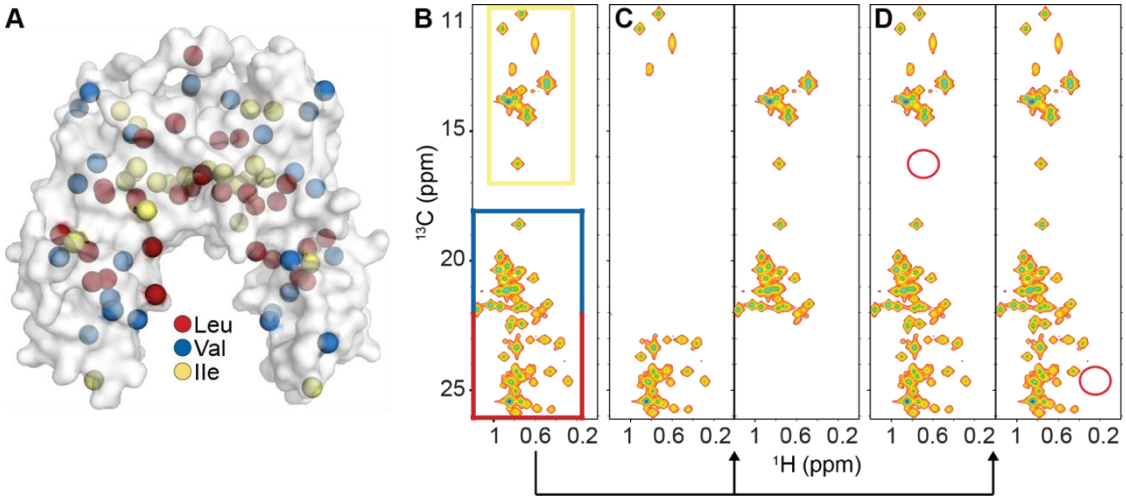


Figure 3.1 | Illustration of the two approaches used to obtain methyl resonance assignment of high molecular mass proteins on the dimer of aspartate transcarbamoylase (ATCase R_2). **A)** Surface representation of the structure of ATCase R_2 with methyl groups in spheres coloured according to their amino acid type (PDB ID: 1D09)[40]. **B)** ^{1}H-^{13}C HMQC spectrum of ATCase R_2 with the region of Ile and Val/Leu resonances in yellow and red/blue rectangles, respectively. Each peak, i.e. ^{1}H-^{13}C correlation, corresponds to a labelled methyl group. **C)** Illustration of the possible outcome of the "divide-and-conquer" assignment strategy. If the full-length protein construct could be split between e.g. two smaller domains for which backbone resonances assignment can be obtained, the spectra of the side-chain methyl groups for the domains (left and right) would be a subset of the ^{1}H-^{13}C HMQC spectrum. The resonances could be assigned through approaches that correlate the backbone and side-chain peak positions[27,28]. **D)** Illustration of the possible outcome of the SDM assignment strategy. If ^{1}H-^{13}C HMQC spectra of the protein construct with a point mutation in an Ile (left) and a Leu/Val (right) position would be recorded, ideally, the disappearance of one signal in the otherwise unchanged spectrum would allow for assignment of the mutated residue. Both of the approaches are less straight-forward in practice than in the provided illustrations.

3.1.2 Structure-based methyl resonance assignment by FLAMEnGO2.0 and MAP-XSII

As an alternative methyl-TROSY assignment method, a high-resolution structure of the protein under study can be combined with NMR distance restraints, such as methyl-methyl

NOEs and/or paramagnetic relaxation enhancements (PREs), to automate the assignment of methyl resonances in large systems[31,32-34]. Automating the assignment of protein backbone resonances in biomolecular NMR has been explored for decades, and these efforts have resulted in numerous automated backbone assignment protocols[35-39].

In contrast, for assignment of methyl resonances recorded on exclusively methyl-protonated, uniformly deuterated proteins using methyl-TROSY method (Chapter 1), only two dedicated automatic assignment approaches exist – FLAMEnGO2.0 and MAP-XSII[34,60]. Both approaches link experimentally determined NOEs, as well as other NMR restraints, to a reference structure of the protein. The challenge for these structure-based approaches is the high-dimensionality of the assignment search space, which these methods address by using approximate computing (Metropolis Monte Carlo sampling, Chapter 2). In addition, the optimization includes a scoring function with a number of terms that require optimal weights that are determined through approximate optimization. As such, these approaches cannot guarantee finding the true (global) optimum of the matching (assignment) problem and can therefore result in erroneous assignments, which can prevent users from drawing confident biochemical conclusions from their results.

3.1.3 Methyl Assignment by Graph MAtching: a novel structure-based methyl assignment approach

Methyl Assignment by Graph MAtching (MAGMA) is a novel structure-based automatic assignment approach based on graph theory. This method combines protein structures and information on methyl-methyl connectivity, obtained from 3D/4D NOESY or DREAM spectra, with exact graph-matching algorithms (Chapter 2) to determine a set of all possible methyl assignments that maximize the experimental restraints. The approach is motivated by the applications of graph theory in chemistry[41-44], structural[45,46] and computational biology[47-

[49], and numerous other fields[50]. In Chapter 2, I defined the resonance assignment problem from inter-methyl distance restraints and a molecular structure as a problem of determining subgraph isomorphism (more precisely, monomorphism) or maximal common edge subgraph (MCES). I also discussed how the exact graph matching algorithms for subgraph isomorphism[51] or MCES[41] allow MAGMA to accurately identify confident (single option) and ambiguous (multiple options) assignments that are consistent with a set of synthetic methyl-methyl distance restraints and available protein structure (Chapter 2).

3.1.4 Similarity between the crystal structure and structural restraints derived from NMR observables

Structure-based assignment (SBA) strategies assume that a structure of a protein in the crystalline state and structural restraints derived from NMR observables have a high degree of compatibility. MAGMA, for instance, relates inter-methyl distances extracted from the crystal structure of the protein to the methyl-methyl NOEs. The availability of a good quality crystal structure and its similarity to the distance restraints extracted from the NMR observables are therefore the most important determinants for the success of the method.

Similarity in structural restraints derived from NMR measurements with those derived from X-ray diffraction is an old problem,[52] and has thus far mostly been explored in the context of comparing structures obtained by NMR spectroscopy and X-ray crystallography. A comparison of the two techniques, and thus the structural restraints derived from them, is not straightforward. In the process that leads to extraction of the restraints, differences exist at every stage, from the raw data and interpretation of the peaks to processing of the signals, i.e. Fourier transform (FT) of the diffraction pattern for X-ray versus FT of the time-dependent signal for NMR[52]. The NOE measurements are an averaged property affected by spin-diffusion, technical limitations, and low signal-to-noise ratios[53]. As such, these measurements

do not give an insight into “an averaged structure”, but rather provide information about the average of the distances sampled by the different states between which a molecule fluctuates[53-55]. Finally, the averaged distance restraints derived from NMR measurements ultimately generate an ensemble structure that coexists within the boundaries of mobility that affect the NMR data, whereas X-ray crystallography typically results in a single structure[52].

Finally, a protein can sample different conformations, some of which can be a result of ligand binding. Sometimes structures of the same protein with different ligands exist whereas NMR studies have only been performed in the ligand-free state or only in the *apo* state. Furthermore, “errors” can exist in X-ray structures due to, for instance, the influence of crystal packing interactions on the surface residues. In these scenarios, the differences in conformations between solution and crystal structure are almost certainly significant and could affect the compatibility of the NOESY-derived inter-methyl connectivity pattern with the corresponding pattern extracted from the crystal structure.

Here, I show the utility of the graph matching approach on experimental NMR data. When applied to a benchmark of eight experimental datasets with known methyl-TROSY assignments, MAGMA assigned up to 89% of methyl resonances using solely NOE restraints, with no errors. Moreover, on higher molecular mass targets (> 40 kDa) MAGMA performed particularly favorably when compared to alternative automated methyl assignment programs and identified all possibilities for assignments in the sites where only ambiguous assignments could be provided. The calculations converged and completed for all tested datasets apart from the largest protein (α7α7), which only converged (Appendix A3.1). In cases where both convergence and completeness were achieved, all theoretically possible resonance assignment options were sampled.

Furthermore, I explore the effect of conformational changes and other structural differences on performance of MAGMA on four proteins from the benchmark. The different structures

for some of the proteins represent different ligand bound states (ATCase R_2, MBP, MSG), NMR structures versus their crystal counterparts (EIN1, MSG), or even structures obtained using different sets of NMR restraints (MSG). Differences in sets of confident methyl assignments and in their accuracy may arise when using different input structures, however, joining their assignment results ensures assignment accuracy (3.2.5). These observations highlight the importance for further exploration of input structure(s) for use in MAGMA.

3.2 Results and discussion

3.2.1. Restraints for MAGMA and MAGMA approach

Graphs of NMR data and methyl-methyl connections in a protein structure

When performing a methyl-TROSY-based study, a protein sample is prepared with uniform deuteration and $^{13}CH_3$ labelled methyl groups, typically isoleucine δ_1, valine γ, and leucine δ; however, alanine β, threonine γ and methionine ε methyl labeling methods are also possible (Chapter 1). The next step is to obtain a 2D ^{1}H-^{13}C HMQC spectrum, which acts as a fingerprint of the molecule. In the simplest case, one resonance is expected per methyl group (Figure 3.2 A), including one for each of the pro-*R* and pro-*S* valine and leucine methyl groups in the absence of stereospecific labelling. It is then possible to record a NOESY spectrum that reveals which methyl groups are close together in space (Figure 3.2 B). In addition, a short mixing-time NOESY experiment can be performed on a uniformly deuterated leucine-[$^{13}C\delta_1H_3$, $^{13}C\delta_2H_3$], valine-[$^{13}C\gamma_1H_3$, $^{13}C\gamma_2H_3$] labelled sample to reveal which two methyl groups belong to the same leucine or valine residue (Figure 3.2 C).

The data from 2D ^{1}H-^{13} HSQC/HMQC and 3D/4D NOESY/DREAM spectra can be readily used to define a graph, where each ^{1}H-^{13}C correlation from the 2D spectrum is a vertex and each cross-peak from a plane in the NOESY/DREAM spectrum connecting such two 2D correlations is an edge (Figure 3.3, Methods). When a high-resolution structure of the protein

under study is available, another graph can be defined in which the edges of the graph comprise all connections between carbon atoms of methyl groups within a certain distance threshold (Figure 3.3 D).

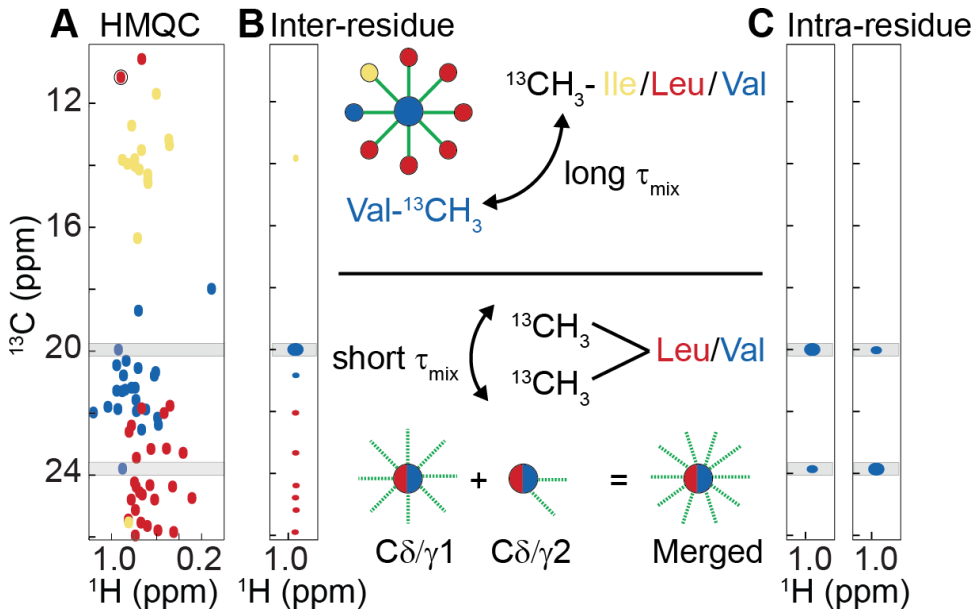


Figure 3.2 | Schematic of NMR restraints used by MAGMA. **A**) Resonance positions in the ^{1}H-^{13}C HMQC spectrum indicated with circles and coloured according to their amino acid type. The assignment of the amino acid types is enabled by preparation of differentially labelled protein samples and inspection of the ^{1}H, ^{13}C resonant frequencies that have characteristic ranges for, for instance, Met and Ile. **B)** An illustration of a strip extracted from a plane of a 3D or 4D methyl-methyl NOESY spectrum showing correlations to the indicated Val peak (**A**, shaded top). These correlations are obtained from spectra recorded with a long NOESY mixing time (for this system 250 ms at 25°C) and constitute inter-methyl restraints used by MAGMA to build the data graph. **C**) An illustration of a strip that could be extracted from a plane of a 3D or 4D methyl-methyl NOESY spectrum if recorded on a double labelled {Leu-[$^{13}C\delta_1H_3$, $^{13}C\delta_2H_3$], Val-[$^{13}C\gamma_1H_3$, $^{13}C\gamma_2H_3$}, U-[^{2}H] protein sample with a short-mixing time (usually between 30-80 ms). The short-mixing time spectrum can provide intra-residue correlations between two Leu/Val resonances that belong to the same residue (as illustrated here for two Val resonances in **A**; shaded). The short-mixing time correlations are used by MAGMA to join the two resonances into a single data graph vertex, attributing all the inter-methyl correlations to the vertex. In general, both long and short mixing times have to be optimized for a particular system.

The resonance assignment process then requires the two graphs to be compared. As introduced in depth in Chapter 2, the problem of comparing graphs to determine their similarity has been known as a graph-matching problem[56] in the field of structural pattern recognition. The objective of the graph-matching is to find a mapping between the vertices of

one graph to the vertices of the other graph, subject to the restrictions relevant to the matching form in question (Chapter 2). In the context of the methyl-methyl NOESY/DREAM data and the protein structure, the objective is to find such mapping (from vertices of the data graph to vertices of the structure graph) that results in the highest overlap of edges. In addition, the search for all such mappings is required to provide a complete set of assignment solutions that all explain the maximum number of methyl-methyl distance restraints. The problem cannot be addressed by a brute force approach, i.e. by evaluating the total number of explained methyl-methyl distance restraints for all possible ^{1}H-^{13}C resonance-to-methyl-bearing-residue assignment combinations due to its exponential (factorial) time requirement (Chapter 2). Therefore, a more efficient sampling strategy is necessary in order to solve the assignment problem exactly.

Exact graph-matching algorithms sample all possible methyl assignment solutions

MAGMA utilizes two exact graph matching algorithms[41,51] to solve for the subgraph isomorphism and MCES problems (Chapter 2). In this application, both algorithms operate on simple, undirected labelled graphs. The labels on the vertices of the graphs reflect different methyl residue types, and these labels account for a significant reduction in the search space for both algorithms. In the case where a residue has two methyl-labelled groups, frequently the case with leucine $\delta_{1/2}$ and valine $\gamma_{1/2}$, it is beneficial to join the two methyl groups into a single pseudo-atom. In general, the more diverse the vertex labels, and the smaller the graph, the faster the search. In addition, more diversity in labels accounts for less ambiguity in the final set of subgraph isomorphism or MCES solutions (Chapter 2).

Different parts of the overall MAGMA workflow are described in detail in Chapter 2. As input, MAGMA takes a set of side-chain atoms from a protein structure and Euclidean distances between them to create an input structure graph (Figure 3.3 D). To define a data

graph, a set of ^{1}H-^{13}C correlations from 2D HSQC/HMQC and a set of cross-peaks from methyl-methyl 3D/4D NOESY/DREAM spectra are required (Methods, Figure 3.3 C).

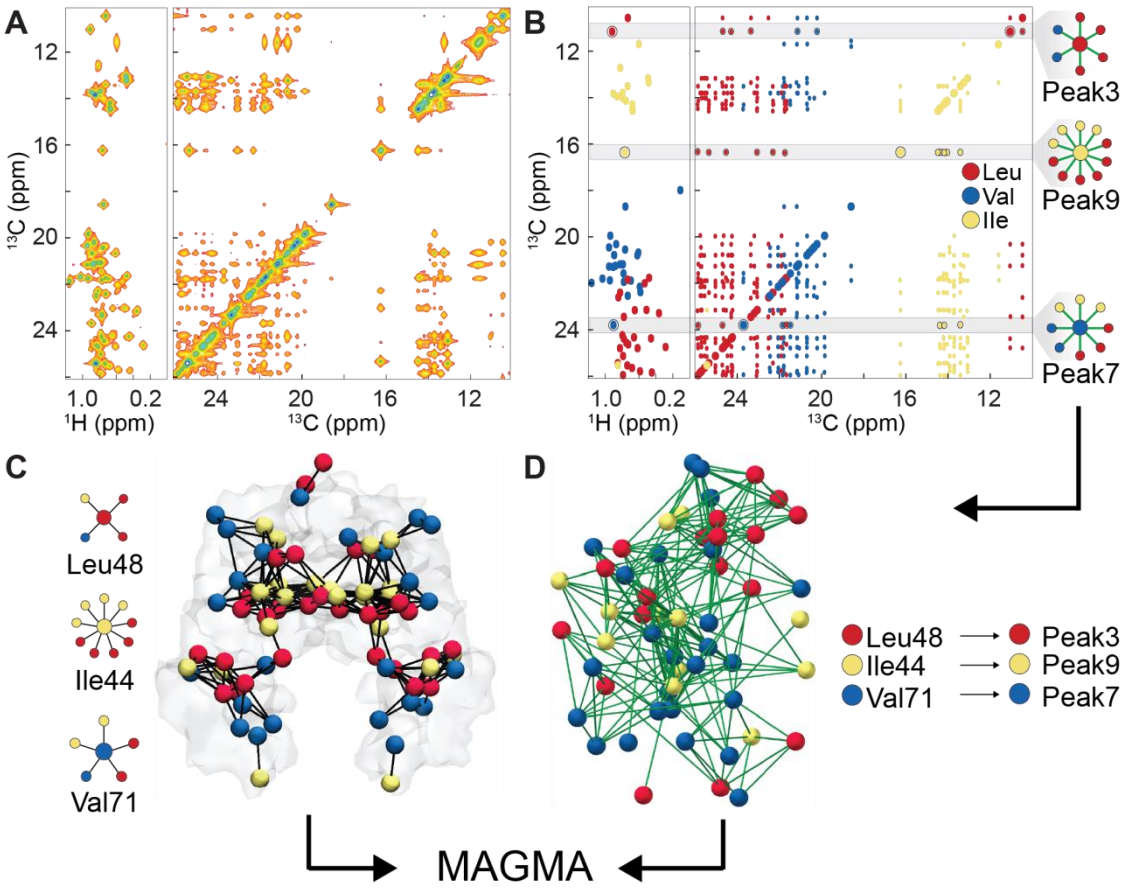


Figure 3.3 | Illustration of the automated methyl resonance assignment approach by MAGMA. **A) A** 2D ^{1}H-^{13}C HMQC spectrum and a 2D ^{13}C-^{13}C projection of a CCH-HMQC–NOESY spectrum from the dimer of aspartate transcarbamoylase (ATCase R_2), recorded from a sample with [$^{13}CH_3$]-labelled methyl groups of Ile, Leu, and Val residues in an otherwise deuterated background. In the case of Leu and Val, only one methyl group is labelled for any one residue. The diagonal peaks in the NOESY spectrum correspond to the methyl peaks from the ^{1}H-^{13}C HMQC spectrum. Off-diagonal peaks reveal which methyl residues undergo significant cross relaxation during the mixing time, where shorter range distances are expected to give higher intensity correlations. **B)** Interpretation of a methyl-TROSY spectrum and 3D ^{13}C-HMQC–NOESY spectrum by MAGMA. Methyl resonances from different residues types are shown in different colours. Extraction of strips from the 3D or 4D NOESY spectrum (**B**, shaded) for each auto peak allows a network to be constructed showing which peaks have connections (i.e. NOEs) to others. **C)** The information from each resonance can be combined into a graph. **D)** Similarly, methyl-methyl connections from a protein structure model can be derived, within a MAGMA determined distance cut-off (see text), and used to define a graph. **E)** MAGMA applies a graph matching protocol to produce assignments.

3.2.2 Proteins in MAGMA's benchmark: experimental NMR data and data graph characteristics

Testing MAGMA on experimental NOE data

MAGMA was tested against a benchmark of experimental data obtained from either 3D or 4D solution-state NOESY[25,60-63] or proton-detected 4D solid-state DREAM[64] NMR spectra of methyl-labelled proteins[63,65-69] (Table 1), whose assignments were previously determined and cross-validated (Figures 3.4 and 3.5, Table 1)[25,32,61-64].

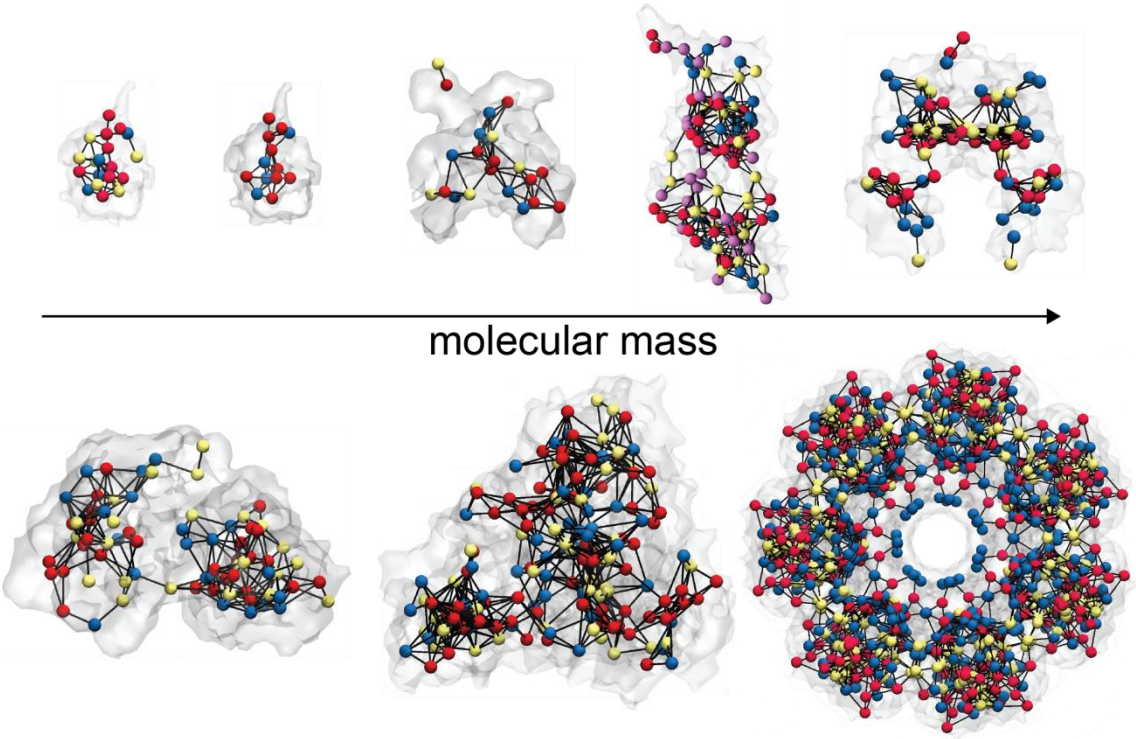


Figure 3.4 | The seven proteins comprising MAGMA's benchmark with eight NMR datasets are shown in surface representation shown approximately to scale, with Ala, Ile, Leu, and Val groups shown as spheres in purple, red, yellow, and blue, respectively. The methyl groups are linked through space within the determined distance cut-off (see text). The proteins are ordered according to increasing molecular mass (left to right). In the top row: ubiquitin (solution-state), ubiquitin (solid-state), MsrB, EIN1, MBP; in the bottom row: MBP, MSG, α7α7 double ring. The range of molecular masses within this dataset spans 8 kDa (ubiquitin) to 360 kDa (α7α7 double ring).

Ubiquitin (Ubq)[32,64,65], methionine-R-sulfoxide reductase B protein (MsrB)[63], the N-terminal domain of *E.coli* Enzyme I (EIN1)[31], a dimer of regulatory chains of aspartate transcarbamoylase from *E. coli* (ATCase R_2)[62,66], maltose binding protein (MSB)[63], malate synthase G (MSG)[61,69], and the double-α_7-ring of the 20S proteasome core particle ($\alpha_7\alpha_7$)[25,68] were selected (Figure 3.4). This benchmark is, to my knowledge, the largest assembled for testing automatic methyl assignment programs.

The eight experimental data sets are diverse in molecular mass, shape, amino acid labeling scheme, and data sparsity (Figure 3.4; Table 1). Using the known NMR assignments, the distribution of inter-methyl distance restraints observed in the data was compared with the distribution of the total possible inter-methyl distances within the structures (Figure 3.6 A and C). In general, the majority of short-range inter-methyl distances are observed (< 8 Å), while relatively few restraints are observed at distances greater than 10 Å.

Determination of the distance threshold

Importantly, an optimal distance threshold needs to be set for every protein for the method to be successful. The distance threshold defines the edges of the structure graph. If the distance threshold is set too short, the inter-methyl NOEs that are critical determinants of accurate assignment are prevented from their accurate mapping in the structure graph, resulting in errors. On the other hand, if the distance threshold is set too long, the time for the calculation is significantly increased and greater heterogeneity is seen in the final result. In this benchmark, the majority of NOEs fall within ~10 Å (Figure 3.6). To determine the optimal distance threshold for MAGMA's benchmark, short (incomplete) MAGMA calculations were run at a range of distances, monitoring the increase in the fraction of explained NOE restraints with increasing distance thresholds (Methods, Figure 3.5 A). From these calculations and by reference to the cross-validated methyl assignments, the optimal

threshold could be determined as the shorter of either the distance at which all NOE restraints are explained, or 10 Å.

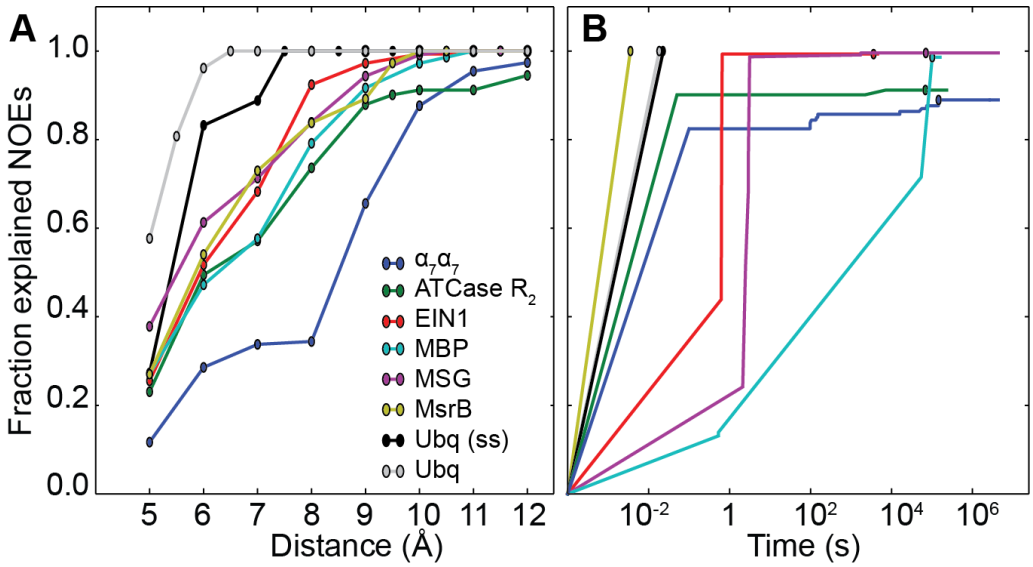


Figure 3.5 | Determining the optimal distance threshold and the time required for convergence and completion of MAGMA calculations. **A)** Short MAGMA calculations are performed at different distance thresholds to determine the number of explained inter-methyl NOEs given the early-found assignment solutions (see Chapter 2, Methods). The distance chosen by the program is the shorter of either the distance at which all inter-methyl distance restraints are explained, or 10 Å. MAGMA can be used to determine the optimal distance threshold according to the above empirical rule, or the user can impose their own threshold in the input file. **B)** The total time taken for the calculation (Table 1). The circle indicates the point at which no further change was found in the final assignment outputs (convergence point). The maximum time for convergence was ~2 days for $\alpha_7\alpha_7$. For this dataset, the completion of the calculation has not observed over a period exceeding 3 months.

On the presented benchmark, the calculation was run with a 10 Å cutoff, with the exception of the smallest protein ubiquitin, for which 6.5 Å and 7.5 Å for solution- and solid-state data sets were respectively used. The small size of the protein resulted in subgraph isomorphism at these distance thresholds, thus explaining all input NOE restraints. For these two datasets, the confidence in assignments decreased at higher distance thresholds. Changing the distance threshold significantly above or below the optimal threshold for each protein can affect the assignment outcome, which is why the use of this empirical rule that led to error free assignments on the presented benchmark is recommended.

On this benchmark, increasing the distance threshold above 10 Å generally led to an increase in ambiguity, resulting in the loss of confident assignments. Decreasing the threshold, however, affected the accuracy of the assignments in several cases. To conclude, this empirical rule (the smaller of 10 Å or the distance at which all NOEs are satisfied) used on MAGMA's benchmark provides 100% accuracy in assignments without introducing unnecessary additional heterogeneity (ambiguity) in the assignment. In the current implementation, the user can either set their own distance threshold, or use MAGMA to determine the optimal threshold according to the described protocol (Methods).

The accuracy of the available protein structure and determination of the correct distance threshold are important factors for the success of MAGMA. From a computational complexity point of view, both subgraph isomorphism and MCES are NP-complete problems (Chapter 2) that in the worst case require exponential time to be solved exactly[56]. The larger the number of edges in the compared graphs, the longer the time needed for comparison. While too large a threshold will increase the ambiguity in the final result and lead to extremely long calculations, too short a threshold might exclude the mapping of restraints that are crucial determinants of the correct assignment, potentially leading to errors. Therefore, the most appropriate threshold takes into account both the experimental reality, and the computational complexity of the graph matching problem.

In practice, it cannot be known *a priori* which methyl-methyl connections will be present or absent in an NMR dataset of a given protein. For instance, larger and slower tumbling molecules will lead to increased cross relaxation rates and enhanced NOEs. However, reduced signal-to-noise, spin diffusion, chemical exchange, and, in some cases, different local mobility in different parts of the molecule could all reduce the expected NOE intensity, thereby ultimately preventing the otherwise expected NOEs from being observed[70]. These

features pose challenges to any method reliant on NOE restraints to generate automatic assignment.

MAGMA interprets experimental inter-methyl restraints in the context of an available protein structure (3.2.5). From this standpoint, if the available protein structure is representative of the protein in solution and if the distance threshold for the observation of the restraints was correctly determined, subgraph isomorphism (more precisely, monomorphism) can be used. Solving for all subgraph isomorphisms will result in a set of assignments that can explain all inter-methyl restraints. Such a solution is the best-case scenario that will maximize the restraints, but remains limited to data with high accuracy, a correct structure, and a correctly determined distance threshold. If the available protein structure is not the most accurate representative of the protein in solution, if noise or ambiguity is present in data, or if the distance threshold is set below the span of long-range NOEs, then the data graph can be compared to the structure graph by solving for a maximum common subgraph. Finding all maximal common edge subgraphs will result in a set of assignment solutions that all maximize the input restraints, but do not explain all of them. Due to the problem of potential errors in the data, it is in most cases not advisable to force all NOEs to be assigned.

3.2.3 Assignment coverage, accuracy, and ambiguity

The number of confident assignments generated with MAGMA on the benchmark varied with the details of the specific experimental dataset but coarsely scaled with the number of restraints in the dataset (Figure 3.6 B and D; Table 1). On average, 3.6 inter-methyl restraints were needed for each confident assignment (Figure 3.7). Two features of the data graphs, average vertex degrees and graph densities (Table 1, Appendix Table A1), justify the choice of the McGregor's MCES algorithm (Chapter 2). The algorithm was shown to outperform

alternatives on graphs with vertex degree of 3 (Table 1)[71], sparse graphs (graph density η = 0.05), and graphs with increasing density (η = 0.1), if these are diversely attributed[71].

Most notably, one of the larger proteins under study, the α7α7 proteasome double-ring resulted in 89% confidently assigned methyls. This exceptional performance could be attributed to the quality of the NMR data, which yielded high ratio of $\frac{|\text{data graph}|}{|\text{structure graph}|}$ (80%) and high local graph connectivity (with graph vertex degrees of ~ 6 on average). The NOE data result in a single graph with only three methyls without any methyl-methyl NOEs, for which the assignments could consequently not be obtained (Figure 3.6 D iv)). The number of generated assignments is likely also related to the shape of the protein under consideration, in addition to the quality of experimental data. Another interesting case is EIN1. This protein was a benchmark for the automatic assignment strategy based primarily on PRE data, with the NOE data being used only for cross validation[31]. Here, I show that 39% of methyl assignments can be obtained using NOE data alone. In general, MAGMA will provide more confident assignments when there are more inter-methyl restraints (Figure 3.7), and when the reference structure is in the closest possible agreement with the distance restraints extracted from the NMR measurements.

In all cases with ambiguous assignments, apart from one (MBP), the experimentally determined answer was sampled as one of the possibilities. For MBP, extending the distance threshold by 0.5 Å (to 10.5 Å) was sufficient for the correct assignment to be sampled within the ambiguous options. Whilst the confident assignments matched the cross-validated ones, using the MBP structure 1ez9 at 10 Å, a few methyl resonances did not sample the cross-validated methyl assignment within the set of ambiguous options. Extension of the distance threshold by only +0.5 Å was sufficient for the ambiguous sets to include the cross-validated assignment. Interestingly, for this dataset, the problem was not circumvented by combining the assignments from multiple structures generated at the 10 Å thresholds. This suggests that

the inter-methyl distances “at the edge” of the distance threshold can be critical in determining the correct assignment outcome.

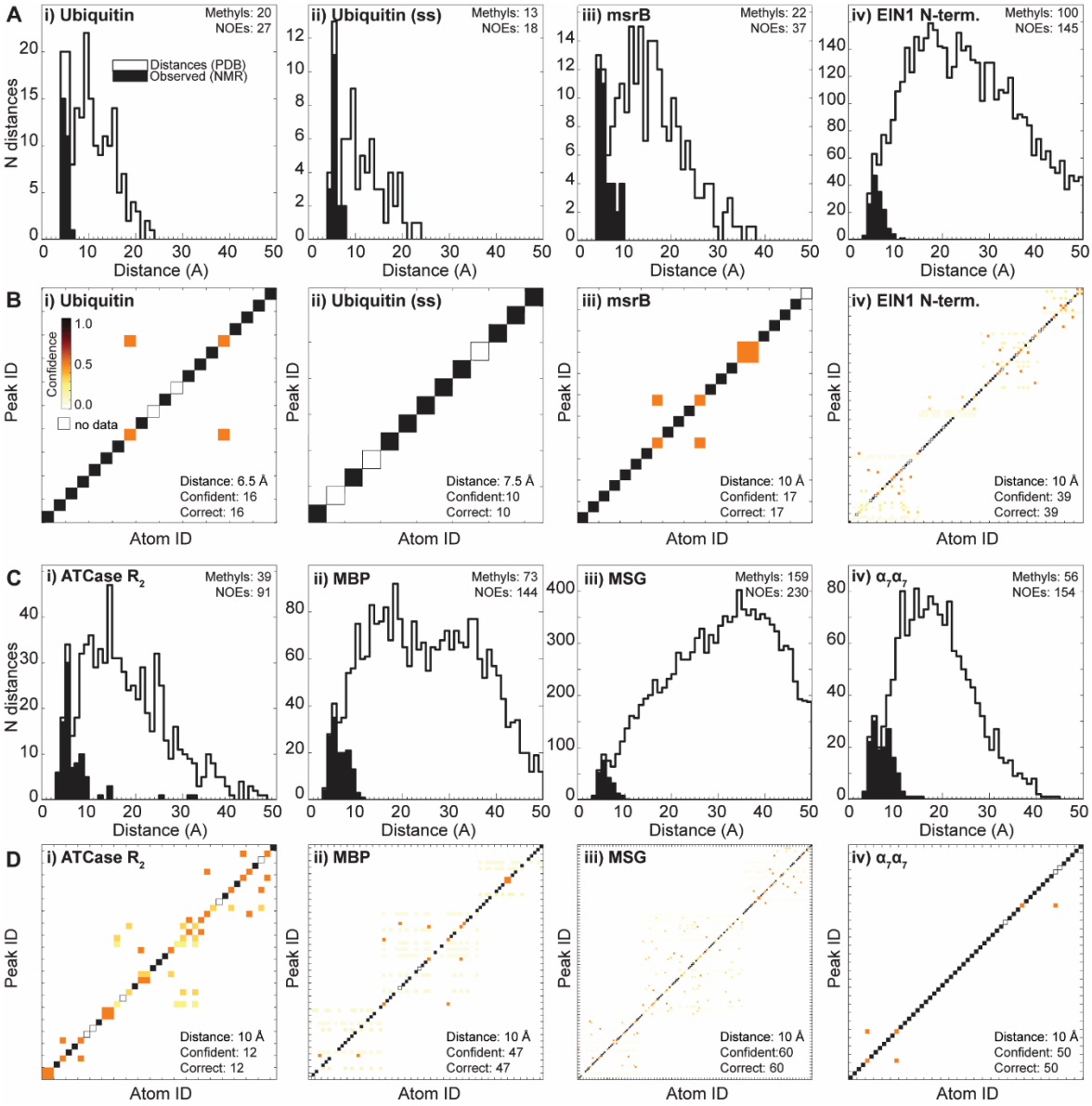


Figure 3.6 | Performance of MAGMA on experimental data (benchmark shown in Figure 3.4, Table 1). **(A, C)** Histograms of inter-methyl carbon distances calculated from the structures of the indicated proteins (white, Figure 3.4), and compared to those experimentally observed (black). **(B, D)** Representation of the assignment results from MAGMA. Methyl resonances are given on the *y*-axis (Peak ID), and methyl carbon atoms on the *x*-axis (Atom ID). The assignments are indexed such that the cross-validated methyl assignments follow the diagonal, i.e. the correct assignment for Peak ID at position 1 is Atom ID position 1. The color bar indicates the confidence in assignments as determined by MAGMA (see text). The methyl resonances that have no inter-methyl restraints are given in white boxes and these cannot be assigned with MAGMA. The total number of confidently assigned methyl groups, where a confident assignment is given to methyl with only one residue assignment option (black, confident), and accuracy (correct) are indicated.

Table 1 | The inter-methyl data defining MAGMA's benchmark and the performance of MAGMA. For each protein, the molecular mass, PDB structure, labelling scheme, and type of NMR experiment is specified. Restraints in solution- and solid-state NMR experiments were respectively obtained from 3D/4D methyl-methyl NOESY spectra and a 4D methyl-methyl DREAM spectrum. The number of restraints and connected data subgraphs are indicated. **(i)** Assignable residues are those with inter-methyl restraints. **(ii)** The average number of inter-methyl restraints per assignable methyl group. **(iii)** The distance threshold derived by MAGMA for each case (Methods). **(iv)** Data density, here defined as number of observed NOEs divided by the number of total number of possible distances, calculated at the optimized distance threshold, iii. Note the difference between this definition of density and that of graph density η (Table A3.1). **(v,vi)** The number of confident (unambiguous) assignments returned by MAGMA, together with the overall fraction of confidently assigned methyls. **(vii)** The time taken to converge on the final result. The maximum overall time was ~2 days (for α7α7). **(viii)** The time for the calculation to run to completion. This was less than 3 days in all cases other than MSG (30 days) and α7α7, which had not concluded in > 50 days (Figure 3.5 B). For these larger proteins, the final result, sampling all ambiguity (Figure 3.6), was obtained after 1 day (MSG) and 2 days (α7α7).

	Mw (kDa)	**PDB identifier**	**Labelling scheme**	**Methyl groups**	**Labelled residues**	**Data type**	**Inter-methyl restraints**	**Connected data subgraphs**	**Assignable resonances (i)**	**Mean vertex degree (ii)**	**Distance threshold (Å) (iii)**	**Data density(iv)**	**Confident assignments (v)**	**Assigned fraction (vi)**	**Time to converge (s) (vii)**	**Total time (s) (viii)**
Ubiquitin	8.6	1ubq	I,L,V	33	20	NOESY	25	3	18	2.78	6.5	0.56	16	0.8	$2x10^{-2}$	$2x10^{-2}$
Ubiquitin(ss)	8.6	1ubq	L,V	26	13	DREAM	18	1	10	3.6	7.5	0.78	10	0.77	$2x10^{-2}$	$2x10^{-2}$
MsrB	16.6	3e0o	I,L,V	39	22	NOESY	37	2	21	3.52	10	0.62	17	0.77	$4x10^{-3}$	$4x10^{-3}$
EIN1	27	1eza	I,L,V,A	146	100	NOESY	145	6	84	3.45	10	0.37	39	0.39	$4x10^{3}$	$5x10^{3}$
ATCase R_2	30	1d09	I,L,V	66	39	NOESY	91	1	34	5.35	10	0.58	12	0.31	$7x10^{4}$	$2x10^{5}$
MBP	43.4	1ez9	I,L,V	123	73	NOESY	144	4	70	4.11	10	0.56	47	0.64	$1x10^{5}$	$2x10^{5}$
MSG	81.4	1y8b	I,L,V	274	159	NOESY	230	12	141	3.26	10	0.47	60	0.38	$7x10^{4}$	$4x10^{6}$
α7α7	360	1yau	I,L,V	96	56	NOESY	154	1	54	5.81	10	0.8	50	0.89	$2x10^{5}$	-

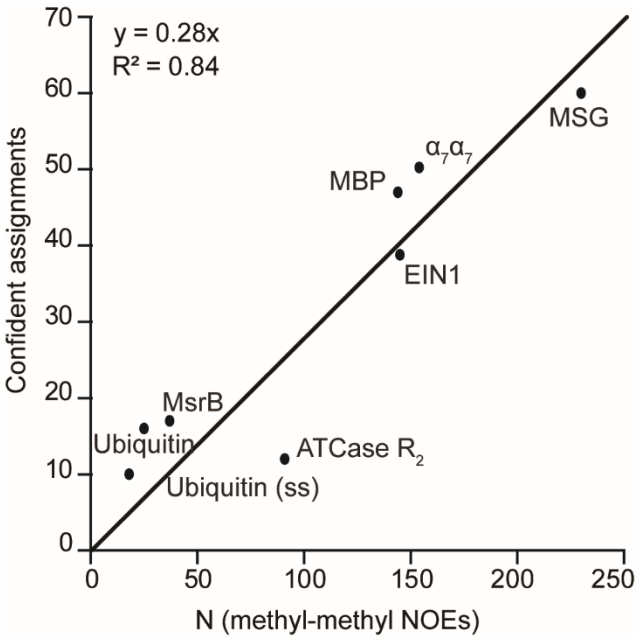


Figure 3.7 | Number of confident assignments for MAGMA's benchmark versus the total number of inter-methyl distance restraints used in the calculation. The performance depends on the data sparsity, the labelling scheme, and the features of the structure, which contains an associated distribution of methyl groups for each protein. On average over the benchmark, 3.6 restraints were required for each confident assignment. A best fit line is shown and the slope of the line and R^2 correlation coefficient are indicated.

3.2.4 Performance of alternative automatic methyl resonance assignment algorithms on MAGMA's benchmark

I compared the performance of MAGMA to that of the two alternative automatic methyl assignment programs (FLAMEnGO2.0 and MAP-XSII)[34,60] on my benchmark (Figure 3.8). The two protocols were run at distance thresholds determined as optimal according to their protocols, and including chemical shift prediction data (Methods, Figure A3.2). Otherwise, the same input data were given to all three programs.

On smaller proteins (<20 kDa), all three algorithms performed comparably, with FLAMEnGO2.0 being the best and yielding 3 and 5 more assignments than MAGMA for ubiquitin (ss-data) and MsrB, respectively. On the intermediate size proteins (20-40 kDa), MAP-XSII is giving the most assignments, respectively covering 12 and 6 more accurately assigned methyl residues than MAGMA for EIN1 and ATCase R_2. However, the algorithm also makes 10 and 5 errors in assignment for EIN1 and ATCase R_2, respectively. On the two larger proteins in the benchmark (MBP and MSG), MAGMA outperforms the alternatives in both accuracy and coverage yielding 34 more assignments than MAP-XSII for MBP, and 57

more for MSG. For EIN1 and MSG, FLAMEnGO2.0 could not provide any confident assignments. Interestingly, on the largest dataset (α7α7), all three algorithms performed comparably. MAP-XSII gave a slightly higher number of confident assignments than MAGMA, albeit with one mistake. FLAMEnGO2.0 had the lowest accuracy, with 7 assignment errors. Comparable performance of the three algorithms on the three datasets with the highest data density (ubiquitin(ss), MsrB, α7α7) suggests that this parameter may be the most important factor for success of the methods.

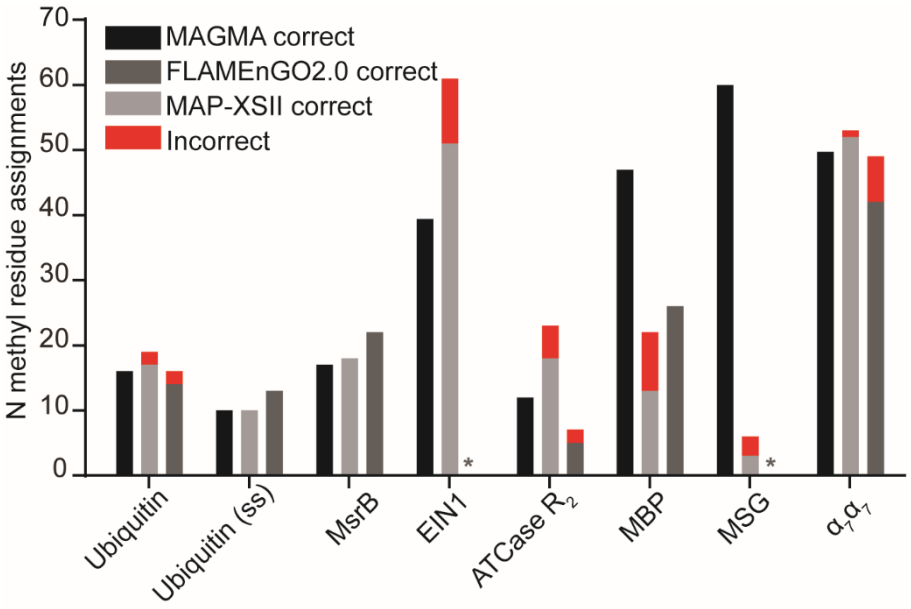


Figure 3.8 | Performance comparison of MAGMA to alternative automated methyl assignment protocols. For each protein in the benchmark, the total number of assigned methyl residues is shown for MAGMA (black), FLAMEnGO2.0 (dark grey), and MAP-XSII (light grey). The confident assignments that were incorrect are indicated (red). FLAMEnGO2.0 returned no confident assignments for EIN1 and MSG (*). MAGMA obtains more confident assignments on targets of high molecular mass (MBP, MSG) and ensures their reliability (accuracy).

FLAMEnGO2.0 and MAP-XSII treat the methyl assignment problem as an optimization problem and use approximate computing (Metropolis Monte Carlo) to minimize a scoring function that is composed of scores for methyl-methyl NOEs, ^{13}C and ^{1}H chemical shifts, PREs and, in the case of FLAMEnGO2.0, optional TOCSY and amide-methyl NOE restraints[32-34,72]. By contrast, MAGMA treats the methyl-TROSY assignment problem as a

graph matching problem, where the objective is to find an exact solution that maximizes methyl-methyl NOE restraints. Unless combinatorically prohibitive, MAGMA's core algorithms will produce all possible assignment solutions that explain the experimental data in the context of the available structure. In doing so, MAGMA provides a complete description of assignments that are ambiguous. Importantly, on the presented benchmark MAGMA showed tolerance to both erroneously assigned NOEs and structural noise (Figure 3.6). In the future, it could be beneficial to include as much NMR data in the calculation as possible; chemical shifts and PREs are two commonly used restraints that could supplement NOEs[31,34,60]. One might expect to find residues whose proton chemical shifts have been substantially shifted by local ring currents. In the benchmark presented here, this does not appear to be the case as, in my hands, beyond distinguishing between methyl residue types, no improvement was seen when chemical shifts were considered (Figure 3.9), for both ^{1}H and ^{13}C nuclei (Figure 3.9 A), or either ^{1}H (Figure 3.9 B) or ^{13}C nuclei (Figure 3.9 C).

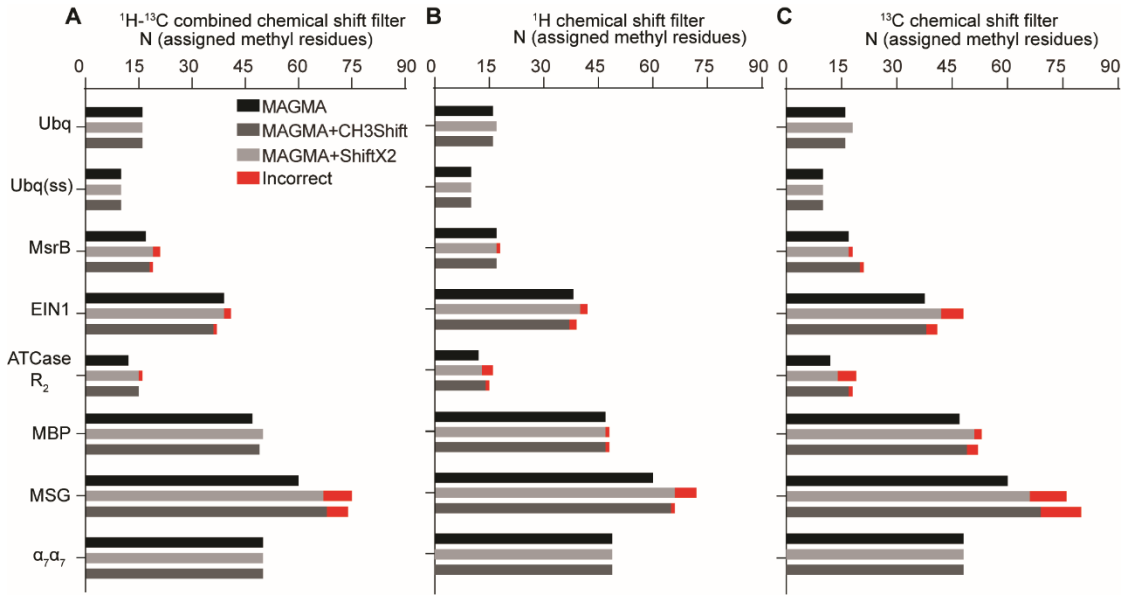


Figure 3.9 | Usefulness of methyl chemical shift prediction in the refinement of MAGMA's assignments. Assignment results obtained by MAGMA were rescored based on ^{1}H-^{13}C chemical shifts predicted from a protein structure using ShiftX2 and CH3Shift protocols. Scoring was applied conservatively (see text). Briefly, if one member of the ambiguous assignment set was the only one to have a chemical shift within the statistical prediction accuracy for the predictor, it was taken as a new confident assignment. This approach generally increased the number of confident assignments, however, at a cost of accuracy (red). The MAGMA solutions for $\alpha_7\alpha_7$ and solid-state NMR ubiquitin could not be improved as there were no ambiguities in the assignments of methyls (*). In all cases other than ubiquitin, the new confident assignments introduced some errors. This was the case if either ^{1}H (**a**), or ^{13}C (**b**), or the combined chemical shifts(**c)** were used. The narrow range of methyl chemical shifts makes them challenging to predict accurately.

Discrimination between different methyl residue types is accessible through producing protein samples with different labelling schemes and recording multiple HMQC spectra. Merging the NOEs arising from the two methyl groups of Leu and Val residues is highly recommended (Chapter 2) and can be accomplished either by recording intra-residue NOEs (via short mixing time NOESY experiments with double methyl-labelled Leu and Val) or, if using divide and conquer, through a TOCSY experiment on a uniformly ^{13}C-labelled sample. In addition, PRE data can be highly valuable for providing assignments as PREs report on both local proximity and long range distances[31]. At present, this restraint is not included in MAGMA; however it would be important to evaluate its usefulness in the future.

Finally, it should be noted that MAGMA's definition of the assignment problem from NOE restraints as graph-subgraph monomorphism and its focus on the graph data structures has similarities to the NOEnet structure-based assignment approach, previously introduced for assignment of the ^{1}H-^{15}N HSQC spectra[78]. Although conceptually similar, NOEnet focuses on solving only subgraph monomorphism and differs from MAGMA in a number of features including the algorithmic approach, input restraints requirements, parametrization, and distance threshold optimization (Chapter 1).

3.2.5 Alternative protein structures and MAGMA performance

MAGMA is a structure-based assignment strategy that extracts the inter-methyl connectivity from a structure of a protein and compares these to the inter-methyl NOE connectivity derived from NMR spectroscopy measurements. The assumption is that the connectivity patterns have a high degree of similarity and can be used to deduce the most optimal mapping between the data graph and the structure graph. MAGMA is thus reliant on the availability of a high-resolution protein structure, and is sensitive to potential differences between the NMR-derived distance restraints and inter-methyl distances measured within the protein structure, as well as to potential differences in protein conformations in solution-, solid-, or crystalline phase. As NMR-derived distance restraints coexist with the boundaries of protein dynamics, and are affected by technical limitations, spin-diffusion, and low signal-to-noise ratios, comparing distance restraints to distances extracted from a crystal structure might influence the accuracy of the assignment.

It is important for any structure-based assignment method to determine the degree of sensitivity to differences in the restraints and to explore the possibilities for circumventing errors in the assignment that may occur. Since MAGMA is applied to proteins for which a crystal structure is available, the only way to assess the potential differences in assignments due to differences in the input structures is to obtain crystal structures in different

conformations, or, if these are not available, to obtain models featuring different conformations.

Here, I explore the degree of sensitivity of MAGMA to the differences in crystal structures used as input in the calculation. The structure benchmark assembled for this purpose is an extension of MAGMA's benchmark presented above, and features four proteins – EIN1, ATCase R_2, MBP, and MSG – for which multiple structural forms were available in the PDB (Figure 3.10 A). ATCase R_2, MBP, and MSG show different conformations when bound to different ligands. For EIN1 and MSG, slightly different structures have been derived from NMR restraints compared to X-ray crystallography (Figure 3.10). In addition, for MSG, two different NMR structures are available[61,73]. Both structures were calculated using the same set of NOE restraints, but one structure included RDC and SAXS restraints (PDB ID: 2qjx)[73].

Different input structures can yield different numbers of confident assignments, different accuracies of these assignments, and different fractions of explained NOEs. The extent of the differences is case-dependent and varies across the benchmark, with the most pronounced differences in the ensemble of ATCase R_2 and negligible differences for that of EIN1 (Figure 3.10 B-D, A3.3). The spread in the number of confident assignments and the assignment accuracy is comparable (Figure 3.10 B, C), however the differences in the fraction of explained inter-methyl NOE restraints across the ensembles are almost negligible (Figure 3.10 D). This means that the structures resulting in different methyl assignment results of varying accuracies can all explain the same (high) fraction of the input NOE restraints (Figure A3.3). Hence, the fraction of explained NOEs does not constitute a reliable metric of the degree of compatibility between the inter-methyl NOE restraints and the different protein structures.

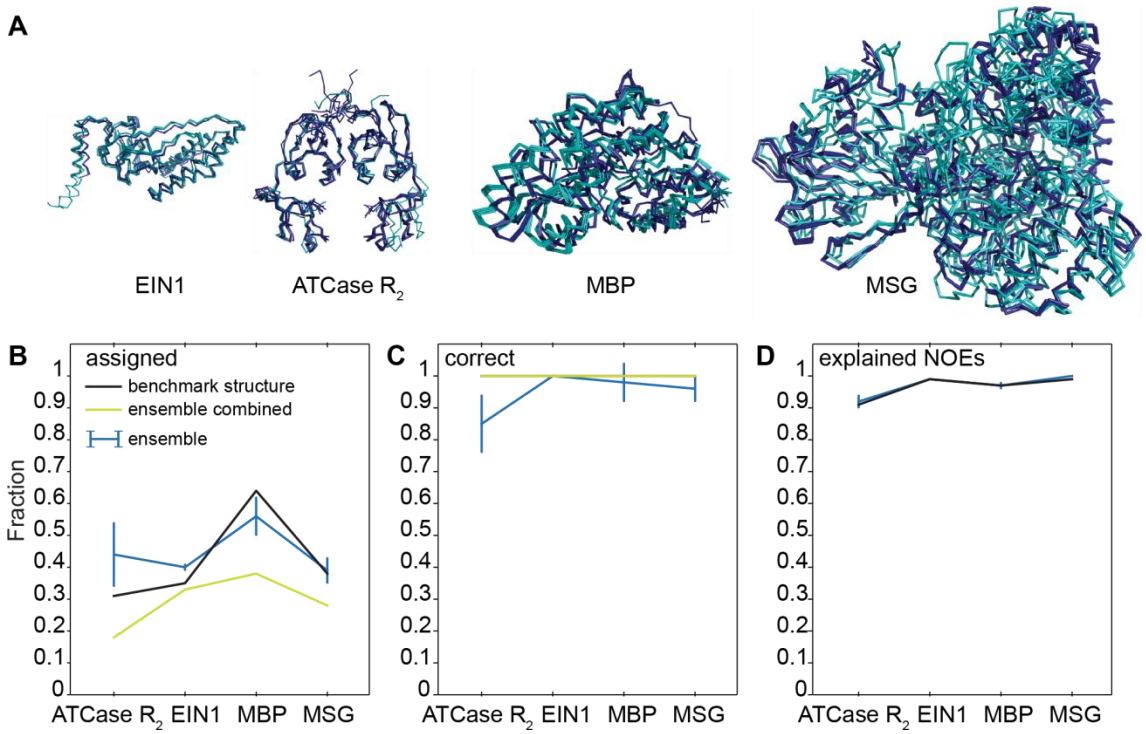


Figure 3.10 | Performance of MAGMA on ensembles of protein structures. Four proteins in MAGMA's benchmark sample more than one conformation for which crystal structures and/or NMR structures are available. These proteins were part of MAGMA's benchmark and are shown in cartoon representation in **A**; from left to right: EIN1 (cyan – NMR structures [PDB ID: 1eza, 2eza], dark blue – crystal structures [2mp0, 3eza, 3eze, 1zym]), ATCase R_2 (cyan – R state [1d09], dark blue – T state [1tug, 1za1, 3d7s, 4fyx, 4fyy]), MBP (cyan – open conformation [1ez9, 1omp, 1lls, 1jw5, 1jw4, 1fqb, 1dmb], dark blue – closed conformation [1anf, 1fqc, 3mbp, 3rum, 4mbp]), MSG (cyan – NMR structures [2jqx, 1y8b], dark blue – crystal structures [1d8c, 1p7t]). ATCase R_2, MBP, and MSG are allosterically modulated and can sample different conformations depending on their interactions with ligands. In **B,** the difference in the number of confident assignments obtained for different structures is indicated with the mean and standard deviation of the ensemble (light blue). The result reported for the structure used in MAGMA's benchmark is given in black. The combined assignment (yellow) is obtained by pooling together assignments for each methyl obtained for all the different structures in the ensemble. **C)** Same as in **B**, but for accuracy of confident assignments. "Combined" and "benchmark" results both achieve 100% accuracy, resulting in the overlap of the black and yellow line. **D)** The fraction of explained NOEs for the ensemble (in blue, mean and standard deviation are displayed) and for a structure reported in MAGMA's benchmark (black); not applicable for combined ensemble results.

Perhaps the best example of the inability of the fraction of explained NOE restraints to discriminate between structures are the assignment results for two protein molecules found in the asymmetric unit of the MSG structure, PDB ID 1p7t[74]. At a first glance, the two structures show negligible differences with all-atom RMS = 0.39 Å, upon removal of the

outliers. Furthermore, each of the structures explains the same maximum number of the input inter-methyl NOE restraints. Nonetheless, one of them results in more confident assignments albeit with lower accuracy: 90% compared to 100% accuracy for the other molecule in the asymmetric unit (Figure A3.3). Further inspection of the two molecules reveals differences in their N-termini where one of them has an extra Ile (Ile 5), which is missing in the other due to lack of electron density for that region of the N-terminus. This difference does not only immediately prevent accurate assignment of the ^{1}H-^{13}C resonance that corresponds to Ile5, but also results in differences in the structure graphs defined from the two molecules, as one of the graphs will have an extra vertex with all of its corresponding edges. Looking at the positions of the methyl carbon coordinates in the best all-atom alignment of the molecules further elucidates differences between the two structures, some of which are significant (A3.4, circled). These carbon coordinate differences are the major cause of differential structure graphs and render one of the structures less compatible with the NMR data graph than the other (Figure A3.3 D). This example illustrates how subtle differences in the primary and tertiary structure of a protein can result in different methyl assignment results obtained by MAGMA and can cause incorrect assignments in some cases (Figure A3.3). It is therefore recommended to always inspect such structural differences prior to MAGMA calculations to avoid treating different protein structures as equivalent.

Interestingly, using two different NMR-derived structures (2jqx[73], 1y8b[61]) for *apo*-MSG resulted in different assignment results (Figure A3.3). For 2jqx, the structure refined using RDCs and SAXS data[73,75], four incorrect methyl assignments were obtained and 11 ambiguously assigned resonances did not sample their cross-validated assignments in the ambiguous set. Importantly, both MSG structures used the same NOE restraints in the structure calculation. This may suggest that 2jqx is less compatible with the inter-methyl NOEs than the original, unrefined structure (PDB: 1y8b)[61]. Before drawing any such

conclusions, it should be noted that in both cases an average NMR structure rather than an ensemble was considered, and that MAGMA's result depends on the distance threshold used in the calculation (here fixed to 10 Å). In the future, it would be important to evaluate MAGMA's performance on the ensembles of the two structures to evaluate their compatibility with the NOE restraints.

If multiple structural forms of a protein are available, the "safest" option is to join the assignments, generated with different structures, for every methyl resonance. This prevents subtle structural differences from compromising assignment accuracy (Figure 3.10, A3.3 "combined"). This approach, however, comes at the cost of confidently assigned methyls (Figure 3.10 B). On average, when joining the assignments, ~16% less confident assignments were obtained when compared to the mean of the ensembles, and ~13% less compared to the structures used in MAGMA's benchmark.

The multiple structure benchmark presented here illustrates the danger of using a single structure as input in MAGMA's calculation, as selecting a "wrong" structure, i.e. the structure that is incompatible with the NMR data, can affect the accuracy of the methyl assignments. Therefore, in the present absence of a more elaborate solution, currently recommended approach is to merge the assignment results from a range of different input structures.

3.3 Conclusions and outlook

The costly and laborious process of assigning resonances has limited the widespread applicability of methyl-TROSY. Typically, methyl-TROSY assignments for large proteins are obtained via site-directed mutagenesis at numerous methyl-bearing residues or through a 'divide-and-conquer' approach, which relies on preparing isolated domains that are small enough to be assigned using standard NMR experiments. Automating this iterative process would enable rapid acquisition of methyl-TROSY assignments and prove beneficial to the NMR community. Two such automatic assignment methods (FLAMEnGO2.0 and MAP-XSII) have been developed; however, these methods are based on approximate computation and cannot guarantee finding the optimal solution.

Here, performance of MAGMA, a novel structure-based methyl-TROSY assignment method based on exact algorithms for graph-matching (Chapter 2), was presented on a benchmark of eight experimental datasets. The benchmark includes proteins with a wide range of molecular masses (8 to 360 kDa), labelling schemes (LV, ILV, or ILVA), and density of inter-methyl restraints (37%-80%), for which high-resolution structures, 2D ^{1}H-^{13} HSQC/HMQC spectra, 3D/4D methyl-methyl NOESY/DREAM spectra, and known methyl-TROSY assignments were available and used for cross-validation. MAGMA requires a distance threshold to be set for a calculation, which should reflect the maximum distance for observation of the inter-methyl NOEs, in the context of the available protein structure. I tested different distance thresholds on the benchmark, noting that too short of distances occasionally yield misassignments and too long of distances led to an increase in assignment ambiguity. On the benchmark, the optimum distance was found to be 10 Å or, for small proteins (ubiquitin), the shortest distance at which all inter-methyl NOEs are explained.

On the benchmark, MAGMA yields a minimum of 31% and a maximum of 89% of the possible methyl-TROSY assignments (i.e. methyl resonances with NOEs), and does not make

an assignment error. In comparison to the alternative methods, MAGMA shows a significant improvement in both the accuracy and assignment coverage on the larger proteins in the benchmark (MBP, MSG).

Accuracy of MAGMA's assignments depends on the protein structure used in the calculation and its relation to the protein conformation sampled during NMR measurements. When multiple structural forms were considered, errors in the methyl assignments were found for some of them. Importantly, not only large conformational differences, but also subtler inconsistencies in the positions of methyl carbons could lead to assignment differences. Therefore, where possible, joining MAGMA's assignments from calculations performed on multiple structures is recommended to achieve optimal accuracy. In the future, more consideration must be given to the treatment of the structure used in MAGMA. For instance, redefining a structure graph such that it contains only the edges that are common to a number of different protein structures would be a straightforward first approach. A more elaborate approach would be starting from a static structure and sampling the range of possible methyl-carbon coordinate positions with a molecular dynamics (MD) simulation. For this approach to be applied, an appropriate way to treat the MD-derived ensemble of methyl carbon positions, and of the fluctuations in their spacing, would have to be devised first. The approach would hopefully lead to a more appropriate structure treatment in MAGMA.

To assess the generality of the conclusions in this chapter, MAGMA needs to be applied to more experimental datasets and a method for the verification of MAGMA's assignments is required. An independent verification method could include site-directed mutagenesis of a subset of the confidently assigned methyl bearing residues.

It would also be of interest to explore the potential of stereo-specific methyl labelling in MAGMA. Such stereospecific labelling would remove the need for an additional protein sample and NOESY experiment, which are currently required to join the two Leu/Val

methyls to a single structure graph vertex. In addition, by reducing the proton density in the system, it could enable more long-range methyl-methyl NOEs to be recorded, consequently leading to a stronger connectivity in the data graph. Finally, measuring the inter-methyl connectivity over a range of NOE mixing times, using some of the existing protocols[53-55], could provide more accurate estimate of the optimal distance threshold for a MAGMA calculation.

The combination of subgraph monomorphism (*VF2*) and MCES (McGregor) algorithm allows MAGMA to be used on sparse, ambiguous, and error-containing experimental data. Nonetheless, a user is advised to remove from the input NOE list any cross-peaks that originate from the degenerate ^{1}H-^{13}C chemical shift positions, to minimize the probability of introducing errors in the input restraints. To identify if any degenerate ^{1}H-^{13}C chemical shifts are present, the user could collect a series of [^{1}H, ^{13}C]-HMQC spectra under different pH, salt, or temperature conditions.

Furthermore, quantitative use of the inter-methyl distances and the cross-peak intensities would be a valuable addition to the method. This information could be used at the end of the protocol to reduce assignment ambiguities and to verify assignment results. To reduce the ambiguity, every assignment list could be checked for compatibility of its NOE cross-peak intensities and their assigned distances. In the most conservative manner, if any MCES or SI gives an assignment that attributes an NOE cross-peak of high intensity to a long inter-methyl distance (> 8 Å), it should be excluded from a pool of viable solutions. In this way, the total number of possible solutions would be reduced, consequently reducing the assignment ambiguity. The same principle could also allow verification of at least some of the confident assignments, as a degree of consistency is expected between the high NOE cross-peak intensities and the distances attributed to them by the assignment.

In the future, the NOE cross-peak intensities could potentially be exploited semi-quantitatively directly in the calculation. For instance, the inter-methyl NOEs could be distributed into intensity classes with approximately assigned distance ranges (e.g. strong, medium, weak NOEs), similarly to approaches typically employed in the classical NMR structure calculation. The inter-methyl distances would then be compared to the NOE cross-peak intensities during the assignment search. This would introduce an approximation to the search, but would also prune the search space and reduce the ambiguity in the final result. To test such a protocol, a larger set of inter-methyl NOE datasets with the cross-peak intensities is required to derive the empirical intensity/distance classes. In all cases, prior to using any distance information, a better treatment of the structure must be found to avoid incorrect classification of the inter-methyl distances to their respective intensity classes.

Finally, next to the NP-completeness of the problem and the employment of the exact graph-matching algorithms, the speed of MAGMA calculation is at the moment limited by the implementation of McGregor's MCES algorithm in an interpreted, dynamically typed, high-level programing language (Python). In the future, a significant speed up will be realized when the algorithm is rewritten into a high-level programming language that can be compiled, such as C or Fortran.

3.4 Methods

Defining methyl-methyl distances within a structure graph

Unless indicated differently, pseudo-atoms were created by averaging the coordinates of the two methyl carbon atoms for both Leu $\delta_{1,2}$ and Val $\gamma_{1,2}$ groups to create a pseudo-atom. The matrix of all Euclidean distances between Ile-δ_1, Leu-pseudo δ, Val-pseudo γ, and when applicable, alanine β, threonine γ, and methionine ε, was computed from protein coordinates (provided in a PDB file format). If more than a single distance was measured for any of the methyl-methyl pairs, the matrix was reduced to the set of shortest distances between any two carbon atoms. Next, the distance cut-off was applied, and only the distances smaller or equal to the defined distance threshold were kept (see Results, Figure 3.5A). This final set of distances defined a set of non-weighted edges for the structure graph, in which the methyl carbon atoms defined a set of vertices with labels assigned according to their amino acid type.

Analysis of experimental NOESY spectra

Processed 3D- and 4D HMQC-NOESY spectra were analyzed using the software *Sparky*[76] for the R2 dimer of ATCase and the N-terminal domain of EIN1, respectively. For other proteins in the benchmark, the spectral analyses were performed by others who provided their NOE peak lists. The NOESY lists were filtered according to the MAGMA's rules (Chapter 2, 2.2.2):

1) An NOE is reciprocated, i.e. there exists a mutual cross-peak in the plane of each ^{1}H-^{13}C auto-peak

2) Cross peaks have similar normalized intensities and signal-to-noise ratios

The required similarity of the normalized cross-peak intensities (I_n) was set to ≥ 90%, and the minimum of the signal-to-noise ratio was ≥ 2. The I_n and the signal-to-noise ratios were

calculated from peak data heights and the estimated noise level determined by the Sparky software[76] (Chapter 2, 2.2.2). In the obtained NOE lists for which data heights (or calculated peak intensities) were missing, only the first (reciprocity) rule was applied.

Defining the data graph

Within MAGMA, ^{1}H-^{13}C frequencies from methyl-TROSY HMQC spectra and the filtered inter-methyl restraints from 3D/4D methyl-methyl NOESY/DREAM spectra are converted to a data graph (Chapter 2). The vertices of the graph are labelled according to their amino acid type. If the assignments for the two methyl groups (pro-*R* and pro-*S*) within a given Leu or Val residue are known, then a single Leu or Val residue is used to join the methyl-methyl NOEs from both methyl peaks, thereby creating a single vertex in a data graph. For all the proteins in the benchmark, the residue types were assumed to be known *a priori*, a requirement that can be accomplished through inspection of the chemical shifts (e.g. discriminating Ile resonances from Leu resonances), or by preparing specifically labeled samples (e.g. to discriminate Leu from Val resonances). Notably, when the resonances of Leu and Val types cannot be distinguished, MAGMA supports giving the same label for Leu/Val ^{1}H-^{13}C resonances, albeit with higher resulting ambiguity.

Determining the optimal distance threshold for graph matching

The distance threshold for the calculation defines a set of edges (methyl-methyl distances) of a structure graph. For MAGMA's benchmark, in all cases, an increase in the distance threshold yields structure graphs with a larger number of edges and graph matching results that explain more methyl-methyl NOEs (i.e. more edges of the data graph find their overlap with methyl-methyl edges of the structure graph). However, due to possible errors in the methyl-methyl NOEs and/or structural noise, optimal distance thresholds should account for the majority – not forcefully all – of methyl-methyl NOEs to be explained. In addition, consideration needs to be taken to a large increase in combinatorics and calculation times for

longer distances (>=12 Å). At 10 Å, the majority of the benchmark proteins samples maximal or near maximal number of NOEs and MAGMA could be run to completion or to convergence (Figure S3, Figure S4). For the small protein ubiquitin, shorter distance thresholds can be used as at 6.5 Å (solution state data) and 7.5 Å (solid state data) all inter-methyl restraints are explained in the context of the available structure. On these bases, an empirical rule for determining the optimal distance for the MAGMA calculation can be proposed, suggesting that either the distance at which all inter-methyl restraints are explained or 10 Å distance should be used in MAGMA calculations.

MAGMA's protocol for the benchmark and scoring

MAGMA was run according to its protocol, as described in detail in Chapter 2. The number of matching-priority sorting iterations was set to 20, and the number of MCESs collected in these iterations and used for the priority resorting to 10. The scoring of the assignment solutions at the end of MAGMA's calculation was performed as described in Chapter 2. Briefly, for each methyl resonance (data graph vertex), the set of all possible assignments to the structure graph vertices (methyl carbon atoms) determined by MAGMA was pooled together and the score given to each resonance assignment was defined as the reciprocal of the total number of possible assignments. Confident resonance assignments are those for which only one residue assignment possibility was found (score = 1). Ambiguous assignments have a score < 1.

Performance comparison

We compared the performance of MAGMA to two alternative structure-based automated assignment protocols, MAP-XSII[34] and FLAMEnGO2.0[60]. The decision had to be taken on the number of iterations for the Metropolis Monte Carlo optimizer (for FLAMEnGO2.0), the number of parallel runs to be performed (for both), and the distance threshold at which the programs are run (for both). The original works provide guidelines for choosing these values.

Information about the amino acid types of ^{1}H-^{13}C resonances (i.e. I, L, V, A, etc.) was provided to all three algorithms, and all three were given the same inter-methyl restraints. In addition, both MAP-XSII and FLAMEnGO2.0 require predicted ^{1}H and ^{13}C chemical shifts, which were calculated using CH3Shift[77].

FLAMEnGO2.0

A single complete calculation was run with 100,000 Monte Carlo steps for each distance in the range of thresholds set from 4.5 to 15 Å in 0.5 Å steps. The optimal distance was obtained by determining the distance value at which the FLAMEnGO2.0 score reached a plateau (Figure A3.2 A). At the optimal distance, 100 repeats of the calculation (each with 100,000 MC steps) were performed. Individual methyl assignments were considered confident if the same result was obtained in all 100 calculations.

MAP-XSII

We ran 20 parallel repeats of the calculation, as proposed by the authors[34]. The weighting between CS and NOEs was set to the default value of 0.2[34]. Upon completion of the calculation, the results of the parallel runs were sorted according to the total score using the graphical interface to the MAP-XSII program. The 10 best scored results were chosen and displayed (Figure A3.2 B). The described protocol was performed at each distance in the range ranging from 6 to 11 Å (1 Å steps). The distance that provided the largest number of confident assignments was taken as optimal. When the same number of confident assignments was obtained at different distance thresholds, the shortest of the distances was chosen.

MAGMA's calculations on multiple input structures

For each input structure, a separate MAGMA calculation was performed with the same input parameters, i.e. the numbers of matching priority re-sorting (N=50) and MCES collected (N=3) and the distance threshold (10 Å). The assignment results from each individual

structure for each methyl resonance were combined, removing the repetitions to obtain the final “combined” result.

3.5 References

1 Clore, G. M., Gronenborn, A. M. Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science* **252**, 1390-1399 (1991).

2 Rosenzweig, R. & Kay, L. E. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu Rev Biochem* **83**, 291-315, doi:10.1146/annurev-biochem-060713-035829 (2014).

3 Barrett, P. J., Chen, J., Cho, M.K., Kim, J.H., Lu, Z., Mathew, S., Peng, D., Song, Y., Van Horn, W.D., Zhuang, T., Sönnichsen, F.D., Sanders, C.R. The quiet renaissance of protein nuclear magnetic resonance. *Biochemistry* **52**, 1303-1320, doi:10.1021/bi4000436 (2013).

4 Sekhar, A., Vallurupalli, P., Kay, L. E. Defining a length scale for millisecond-timescale protein conformational exchange. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 11391-11396, doi:10.1073/pnas.1303273110 (2013).

5 Mittermaier, A., Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224-228, doi:10.1126/science.1124964 (2006).

6 Mittermaier, A. K., Kay, L. E. Observing biological dynamics at atomic resolution using NMR. *Trends in biochemical sciences* **34**, 601-611, doi:10.1016/j.tibs.2009.07.004 (2009).

7 Ishima, R., Torchia, D. A. Protein dynamics from NMR. *Nature structural biology* **7**, 740-743, doi:10.1038/78963 (2000).

8 Baldwin, A. J., Kay, L. E. NMR spectroscopy brings invisible protein states into focus. *Nat Chem Biol* **5**, 808-814, doi:10.1038/nchembio.238 (2009).

9 Goodsell, D.S., Olson, A.J. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* **29**, 105-153, doi:10.1146/annurev.biophys.29.1.105 (2000).

10 Ollerenshaw, J. E., Tugarinov, V., Kay, L. E. Methyl TROSY: explanation and experimental verification. *Magn Reson Chem* **41**, 843-852, doi:Doi 10.1002/Mrc.1256 (2003).

11 Goto, N. K., Gardner, K. H., Mueller, G. A., Willis, R. C. & Kay, L. E. A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated 15N-, 13C-, 2H-labeled proteins. *J Biomol NMR* **13**, 369-374 (1999).

12 Tugarinov, V., Kanelis, V., Kay, L. E. Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nature protocols* **1**, 749-754, doi:10.1038/nprot.2006.101 (2006).

13 Tugarinov, V., Kay, L. E. An isotope labeling strategy for methyl TROSY spectroscopy. *J Biomol NMR* **28**, 165-172, doi:10.1023/B:JNMR.0000013824.93994.1f (2004).

14 Ayala, I., Hamelin, O., Amero, C., Pessey, O., Plevin, M.J., Gans, P., Boisbouvier, J. An optimized isotopic labelling strategy of isoleucine-gamma2 methyl groups for solution NMR studies of high molecular weight proteins. *Chemical communications* **48**, 1434-1436, doi:10.1039/c1cc12932e (2012).

15 Ayala, I., Sounier, R., Use, N., Gans, P., Boisbouvier, J. An efficient protocol for the complete incorporation of methyl-protonated alanine in perdeuterated protein. *J Biomol NMR* **43**, 111-119, doi:10.1007/s10858-008-9294-7 (2009).

16 Gans, P., Hamelin, O., Sounier, R., Ayala, I., Durá, M.A., Amero, C.D., Noirclerc-Savoye, M., Franzetti, B., Plevin, M.J., Boisbouvier, J. Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. *Angew Chem Int Ed Engl* **49**, 1958-1962, doi:10.1002/anie.200905660 (2010).

17 Fischer, M., Kloiber, K., Häusler, J., Ledolter, K., Konrat, R., Schmid, W. Synthesis of a 13C-methyl-group-labeled methionine precursor as a useful tool for simplifying protein structural analysis by NMR spectroscopy. *Chembiochem : a European journal of chemical biology* **8**, 610-612, doi:10.1002/cbic.200600551 (2007).

18 Isaacson, R. L., Simpson, P. J., Liu, M., Cota, E., Zhang, X., Freemont, P., Matthews, S . A new labeling method for methyl transverse relaxation-optimized spectroscopy NMR spectra of alanine residues. *Journal of the American Chemical Society* **129**, 15428-15429, doi:10.1021/ja0761784 (2007).

19 Ruschak, A. M., Velyvis, A. & Kay, L. E. A simple strategy for (1)(3)C, (1)H labeling at the Ile-gamma2 methyl position in highly deuterated proteins. *J Biomol NMR* **48**, 129-135, doi:10.1007/s10858-010-9449-1 (2010).

20 Lichtenecker, R., Ludwiczek, M. L., Schmid, W. & Konrat, R. Simplification of protein NOESY spectra using bioorganic precursor synthesis and NMR spectral editing. *Journal of the American Chemical Society* **126**, 5348-5349, doi:10.1021/ja049679n (2004).

21 Lichtenecker, R. J., Coudevylle, N., Konrat, R. & Schmid, W. Selective isotope labelling of leucine residues by using alpha-ketoacid precursor compounds. *Chembiochem : a European journal of chemical biology* **14**, 818-821, doi:10.1002/cbic.201200737 (2013).

22 Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 12366-12371 (1997).

23 Baldwin, A. J., Walsh, P., Hansen, D. F., Hilton, G. R., Benesch, J. L., Sharpe, S., Kay, L. E. Probing Dynamic Conformations of the High-Molecular-Weight alpha B-Crystallin Heat Shock Protein Ensemble by NMR Spectroscopy. *Journal of the American Chemical Society* **134**, 15343-15350, doi:10.1021/ja307874r (2012).

24 Baldwin, A. J., Hansen, D. F., Vallurupalli, P. & Kay, L. E. Measurement of Methyl Axis Orientations in Invisible, Excited States of Proteins by Relaxation Dispersion NMR Spectroscopy. *Journal of the American Chemical Society* **131**, 11939-11948, doi:10.1021/ja903896p (2009).

25 Sprangers, R. & Kay, L. E. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **445**, 618-622, doi:10.1038/nature05512 (2007).

26 Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wuthrich, K. NMR analysis of a 900K GroEL GroES complex. *Nature* **418**, 207-211, doi:10.1038/nature00860 (2002).

27 Tugarinov, V. & Kay, L. E. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *Journal of the American Chemical Society* **125**, 13868-13878, doi:10.1021/ja030345s (2003).

28 Sheppard, D., Guo, C. & Tugarinov, V. 4D 1H-13C NMR spectroscopy for assignments of alanine methyls in large and complex protein structures. *Journal of the American Chemical Society* **131**, 1364-1365, doi:10.1021/ja808202q (2009).

29 Gelis, I., Bonvin, A. M., Keramisanou, D., Koukaki, M., Gouridis, G., Karamanou, S., Economou, A., Kalodimos, C. G. Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **131**, 756-769, doi:10.1016/j.cell.2007.09.039 (2007).

30 Amero, C., Asuncion Dura M., Noirclerc-Savoye M., Perollier, A., Gallet, B., Plenvin, M. J., Vernet, T., Franzetti, B. Boisbouvier, J. A systematic mutagenesis-driven strategy for site-resolved NMR studies of supramolecular assemblies. *J Biomol NMR* **50**, 229-236, doi:10.1007/s10858-011-9513-5 (2011).

31 Venditti, V., Fawzi, N. L. & Clore, G. M. Automated sequence- and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl-methyl nuclear Overhauser enhancement spectroscopy. *J Biomol NMR* **51**, 319-328, doi:10.1007/s10858-011-9559-4 (2011).
32 Chao, F. A., Shi, L., Masterson, L. R. & Veglia, G. FLAMEnGO: a fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *Journal of magnetic resonance* **214**, 103-110, doi:10.1016/j.jmr.2011.10.008 (2012).
33 Xu, Y., Liu, M., Simpson, P.J., Isaacson, R., Cota, E., Marchant, J., Yang, D., Zhang, X., Freemont, P. Automated assignment in selectively methyl-labeled proteins. *Journal of the American Chemical Society* **131**, 9480-9481, doi:10.1021/ja9020233 (2009).
34 Xu, Y. & Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J Biomol NMR,* **2013**, *55*, 179-187, doi:10.1007/s10858-012-9700-z.
35 Moseley, H. N. B. & Montelione, G. T. Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struc Biol.* **1999**, 635-642, doi:Doi 10.1016/S0959-440x(99)00019-6
36 Bahrami, A., Assadi, A. H., Markley, J. L. & Eghbalnia, H. R. Probabilistic Interaction Network of Evidence Algorithm and its Application to Complete Labeling of Peak Lists from Protein NMR Spectroscopy. *Plos Comput Biol* **2009**, *5*
37 Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R., Montelione, G. T. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol* **269**, 592-610, doi:DOI 10.1006/jmbi.1997.1052 (1997).
38 Zuiderweg, E. R. P., Bagai, I., Rossi, P. & Bertelsen, E. B. EZ-ASSIGN, a program for exhaustive NMR chemical shift assignments of large proteins from complete or incomplete triple-resonance data. *Journal of Biomolecular Nmr* **57**, 179-191, doi:10.1007/s10858-013-9778-y (2013).
39 Crippen, G. M., Rousaki, A., Revington, M., Zhang, Y. B. & Zuiderweg, E. R. P. SAGA: rapid automatic mainchain NMR assignment for large proteins. *Journal of Biomolecular Nmr* **46**, 281-298, doi:10.1007/s10858-010-9403-2 (2010).
40 Jin, L., Stec, B., Lipscomb, W. N. & Kantrowitz, E. R. Insights into the mechanisms of catalysis and heterotropic regulation of Escherichia coli aspartate transcarbamoylase based upon a structure of the enzyme complexed with the bisubstrate analogue N-phosphonacetyl-L-aspartate at 2.1 A. *Proteins* **37**, 729-742 (1999).
41 Mcgregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software Pract Exper* **12**, 23-34, doi:DOI 10.1002/spe.4380120103 (1982).
42 Raymond, J. W., Gardiner, E. J. & Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comp Sci* **42**, 305-316, doi:Doi 10.1021/Ci010381f (2002).
43 Raymond, J. W. & Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J Comput Aid Mol Des* **16**, 59-71, doi:Doi 10.1023/A:1016387816342 (2002).
44 Raymond, J. W. & Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aid Mol Des* **16**, 521-533, doi:Doi 10.1023/A:1021271615909 (2002).
45 Di Paola, L. & Giuliani, A. Protein contact network topology: a natural language for allostery. *Curr Opin Struc Biol* **31**, 43-48, doi:10.1016/j.sbi.2015.03.001 (2015).

46 Vuillon, L. & Lesieur, C. From local to global changes in proteins: a network view. *Curr Opin Struc Biol* **31**, 1-8, doi:10.1016/j.sbi.2015.02.015 (2015).
47 Stam, C. J. Modern network science of neurological disorders. *Nat Rev Neurosci* **15**, 683-695, doi:10.1038/nrn3801 (2014).
48 Parikshak, N. N., Gandal, M. J. & Geschwind, D. H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat Rev Genet* **16**, 441-458, doi:10.1038/nrg3934 (2015).
49 Sirur, A., De Sancho, D. & Best, R. B. Markov state models of protein misfolding. *J Chem Phys* **144**, 075101, doi:10.1063/1.4941579 (2016).
50 Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268-276, doi:10.1038/35065725 (2001).
51 Cordella, L. P., Foggia, P., Sansone, C. & Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *Ieee T Pattern Anal* **26**, 1367-1372, doi:Doi 10.1109/Tpami.2004.75 (2004).
52 Wagner, G., Hyberts, S. G. & Havel, T. F. NMR structure determination in solution: a critique and comparison with X-ray crystallography. *Annu Rev Biophys Biomol Struct* **21**, 167-198, doi:10.1146/annurev.bb.21.060192.001123 (1992).
53 Chi, C. N., Vogeli, B., Bibow, S., Strotz, D., Orts, J., Guntert, P., Riek, R. A Structural Ensemble for the Enzyme Cyclophilin Reveals an Orchestrated Mode of Action at Atomic Resolution. *Angew Chem Int Ed Engl* **54**, 11657-11661, doi:10.1002/anie.201503698 (2015).
54 Orts, J., Vogeli, B. & Riek, R. Relaxation Matrix Analysis of Spin Diffusion for the NMR Structure Calculation with eNOEs. *J Chem Theory Comput* **8**, 3483-3492, doi:10.1021/ct3002249 (2012).
55 Vogeli, B., Orts, J., Strotz, D., Guntert, P. & Riek, R. Discrete three-dimensional representation of macromolecular motion from eNOE-based ensemble calculation. *Chimia (Aarau)* **66**, 787-790, doi:10.2533/chimia.2012.787 (2012).
56 Bunke, H. Graph Matching: Theoretical Foundations, Algorithms, and Applications. *International Conference on Vision Interface*, 82-84 (2000).
57 Golomb, S. W. & Baumert, L. D. Backtrack Programming. *J Acm* **12**, 516-&, doi:Doi 10.1145/321296.321300 (1965).
58 Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C. & Vento, M. in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada, August 6–9, 2002 Proceedings* (eds Terry Caelli *et al.*) 123-132 (Springer Berlin Heidelberg, 2002).
59 Bunke, H. Recent developments in graph matching. *Int C Patt Recog*, 117-124 (2000).
60 Chao, F. A., Kim, J., Xia, Y., Milligan, M., Rowe, N., Veglia, G. FLAMEnGO 2.0: an enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *Journal of magnetic resonance* **245**, 17-23, doi:10.1016/j.jmr.2014.04.012 (2014).
61 Tugarinov, V., Choy, W. Y., Orekhov, V. Y. & Kay, L. E. Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 622-627, doi:DOI 10.1073/pnas.0407792102 (2005).
62 Velyvis, A., Schachman, H. K. & Kay, L. E. Assignment of Ile, Leu, and Val methyl correlations in supra-molecular systems: an application to aspartate transcarbamoylase. *Journal of the American Chemical Society* **131**, 16534-16543, doi:10.1021/ja906978r (2009).

63 Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H. W., Aramini, J. M., Ertekin, A., Xiao, R., Acton, T. B., Montelione, G. T., Baker, D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 10873-10878, doi:10.1073/pnas.1203013109 (2012).
64 Huber, M., Hiller, S., Schanda, P., Ernst, M., Bockmann, A., Verel, R., Meier, B. H. A Proton-Detected 4D Solid-State NMR Experiment for Protein Structure Determination. *Chemphyschem* **12**, 915-918, doi:10.1002/cphc.201100062 (2011).
65 Vijay-Kumar, S., Bugg, C.E., Cook, W.J. Structure of ubiquitin refined at 1.8 A resolution. *J Mol Biol* **194**, 531-544 (1987).
66 Stieglitz, K., Stec, B., Baker, D. P. & Kantrowitz, E. R. Monitoring the transition from the T to the R state in E. coli aspartate transcarbamoylase by X-ray crystallography: Crystal structures of the E50A mutant enzyme in four distinct allosteric states. *J Mol Biol* **341**, 853-868, doi:10.1016/j.jmb.2004.06.002 (2004).
67 Garrett, D. S., Seok, Y. J., Liao, D. I., Peterkofsky, A., Gronenborn, A., Clore, G. M. Solution structure of the 30 kDa N-terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. *Biochemistry* **36**, 2517-2530, doi:Doi 10.1021/Bi962924y (1997).
68 Lowe, J., Stock, D., Jap, B., Zwicki, P., Baumeister, W., Huber, R. Crystal-Structure of the 20S Proteasome from the Archaeon T-Acidophilum at 3.4-Angstrom Resolution. *Science* **268**, 533-539, doi:DOI 10.1126/science.7725097 (1995).
69 Howard, B. R., Endrizzi, J. A. & Remington, S. J. Crystal structure of Escherichia coli malate synthase G complexed with magnesium and glyoxylate at 2.0 angstrom resolution: Mechanistic implications. *Biochemistry* **39**, 3156-3168, doi:Doi 10.1021/Bi992519h (2000).
70 Vogeli, B. The nuclear Overhauser effect from a quantitative perspective. *Prog Nucl Mag Res Sp* **78**, 1-46, doi:10.1016/j.pnmrs.2013.11.001 (2014).
71 Conte, D., Foggia, P., Vento, M. Challenging Complexity of Maximum Common Subgraph Detection Algorithms: A Performance Analysis of Three Algorithms on a Wide Database of Graphs. *Journal of Graph Algorithms and Applications* **11**, 99-143 (2007).
72 Chao, F. A., Kim, J., Xia, Y., Milligan, M., Rowe, N., Veglia, G. FLAMEnGO 2.0: An enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *Journal of Magnetic Resonance* **245**, 17-23, doi:10.1016/j.jmr.2014.04.012 (2014).
73 Grishaev, A., Tugarinov, V., Kay, L. E., Trewhella, J. & Bax, A. Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J Biomol NMR* **40**, 95-106, doi:10.1007/s10858-007-9211-5 (2008).
74 Anstrom, D. M., Kallio, K. & Remington, S. J. Structure of the Escherichia coli malate synthase G:pyruvate:acetyl-coenzyme A abortive ternary complex at 1.95 A resolution. *Protein Sci* **12**, 1822-1832, doi:10.1110/ps.03174303 (2003).
75 Grishaev, A., Wu, J., Trewhella, J. & Bax, A. Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J Am Chem Soc* **127**, 16621-16628, doi:10.1021/ja054342m (2005).
76 Goddard, T. D. K., D. G. SPARKY 3. *University of California, San Francisco*.
77 Sahakyan, A. B., Vranken, W. F., Cavalli, A. & Vendruscolo, M. Structure-based prediction of methyl chemical shifts in proteins. *Journal of biomolecular NMR* **50**, 331-346, doi:10.1007/s10858-011-9524-2 (2011).

78 Stratmann, D., Guittet, E., van Heijenoort, C. Robust structure-based resonance assignment for functional protein studies by NMR. *Journal of biomolecular NMR* **46**, 157-173, doi:10.1007/s10858-009-9390-3 (2010).

Appendix

Table A1 | Graph density of data graphs defined from experimental NOE datasets in MAGMA's benchmark. η is graph density (Chapter 2, 2.1.1 Definitions)

Experimental data-set	η
Ubiquitin	0.08
Ubiquitin(ss)	0.20
MsrB	0.09
EIN1	0.02
ATCase R_2	0.08
MBP	0.03
MSG	0.01
α7α7	0.06

A3.1 Convergence and completion of MAGMA calculation

We should note the difference in definitions of convergence and completion of MAGMA's calculation. Here, a calculation is defined converged when no new assignment solutions have been found after a defined period of time. On the benchmark presented here, no new assignment solutions were found after a maximum of ~48 hours of sampling. The time between convergence and completion of the calculation the algorithms employed here spend on backtracking and testing alternative (partial) assignment solutions to ensure that no optimal solution has been omitted. Once all such alternative options have been assessed the search has completed. Convergence does not guarantee that all solutions to the problem have been found, and the results obtained from "solely" converged and not completed calculation are approximate.

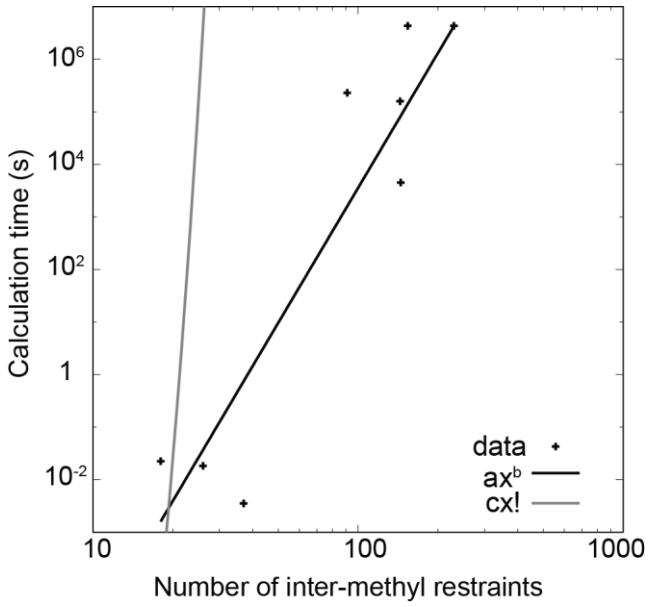


Figure A3.1 | The convergence time on the benchmark scales with the total number of inter-methyl restraints (N). For comparison, the expected time scaling if the time was increasing in a factorial fashion is indicated (grey). Despite this favorable time requirement, the NP-completeness of subgraph isomorphism and MCES problems does not allow prediction of the execution times for any new cases.

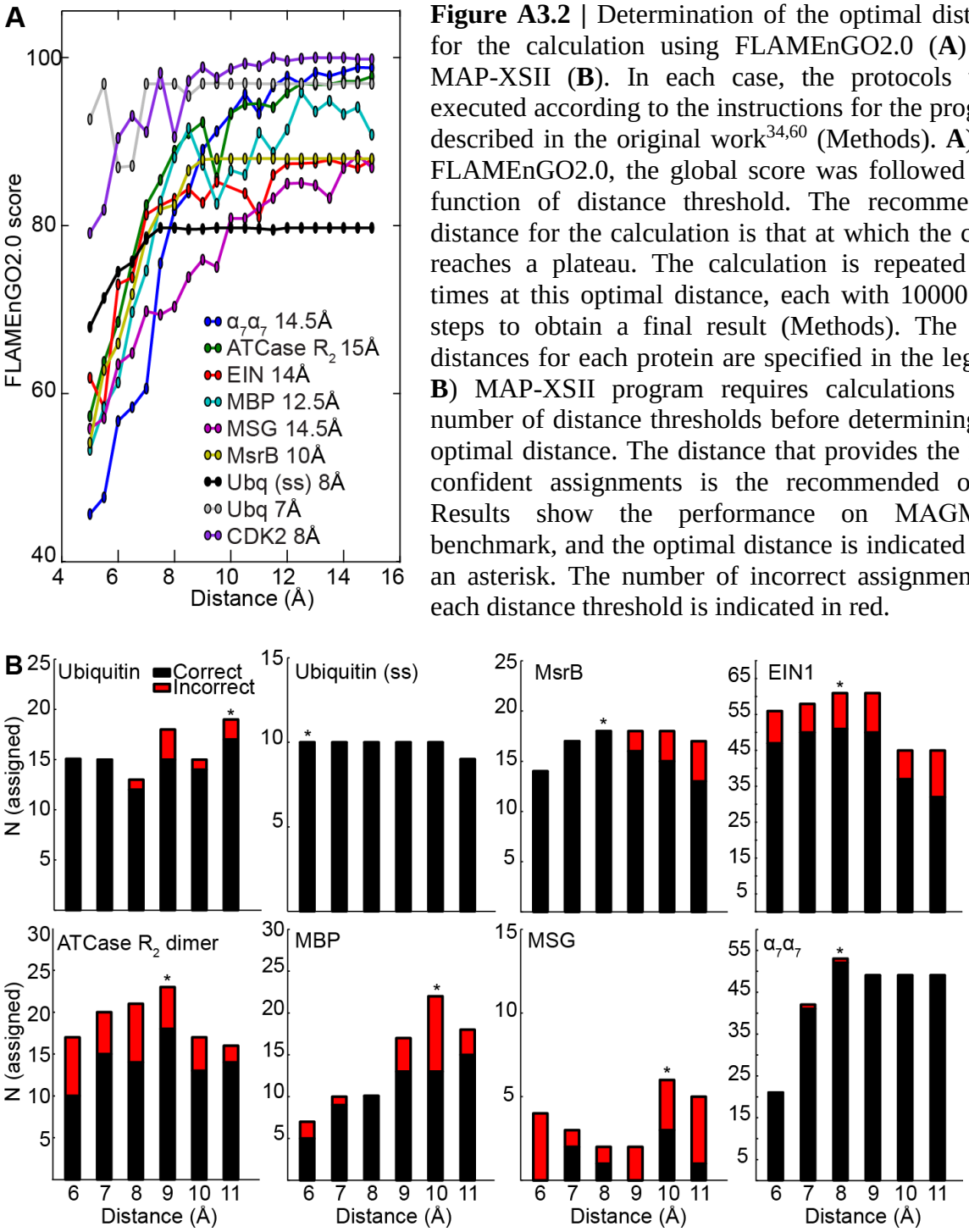


Figure A3.2 | Determination of the optimal distance for the calculation using FLAMEnGO2.0 (**A**) and MAP-XSII (**B**). In each case, the protocols were executed according to the instructions for the program described in the original work[34,60] (Methods). **A**) For FLAMEnGO2.0, the global score was followed as a function of distance threshold. The recommended distance for the calculation is that at which the curve reaches a plateau. The calculation is repeated 100 times at this optimal distance, each with 10000 MC steps to obtain a final result (Methods). The final distances for each protein are specified in the legend. **B**) MAP-XSII program requires calculations at a number of distance thresholds before determining the optimal distance. The distance that provides the most confident assignments is the recommended one[34]. Results show the performance on MAGMA's benchmark, and the optimal distance is indicated with an asterisk. The number of incorrect assignments at each distance threshold is indicated in red.

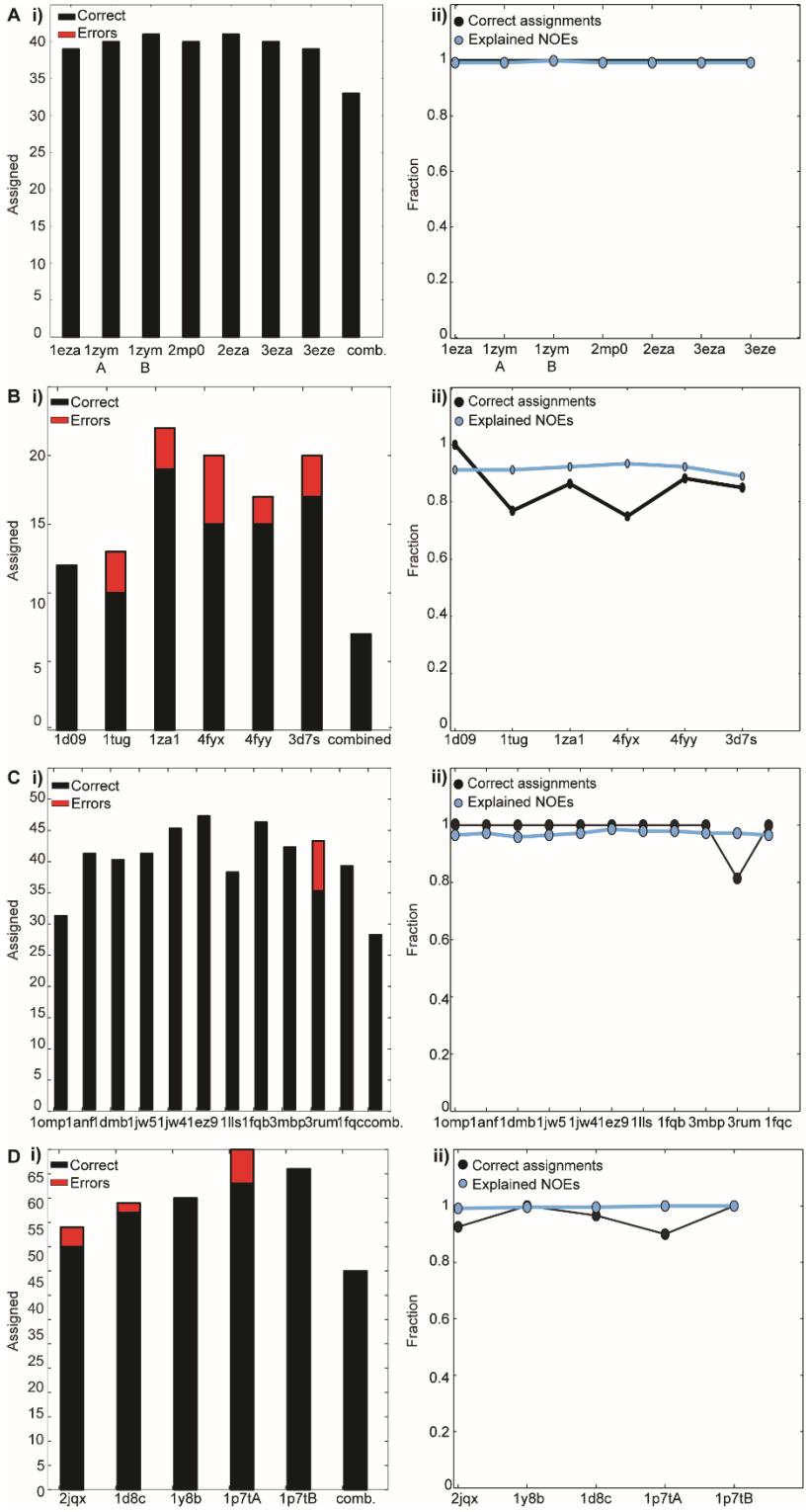


Figure A3.3 | Performance of MAGMA on ensemble of structures. MAGMA was run on multiple different structures for four proteins: EIN1 (**A**), ATCase R_2 (**B**), MBP (**C**), and MSG (**D**). (**i**) PBD ID is given on x-axis for each of the structures used in MAGMA's calculation with the number of confident methyl assignments on the y-axis. The number of correct and incorrect assignments is indicated in black and red, respectively. **ii**) Fraction of explained NOEs and of correct assignments for different input structures. Note that structures with high fraction of explained NOEs do not necessarily result in completely accurate assignments (e.g. ATCase R_2 – 4fyx, MBP – 3rum). For MSG structure 2jqx, MAGMA's calculation converged, but has not completed (calculation has been run for ~3 months).

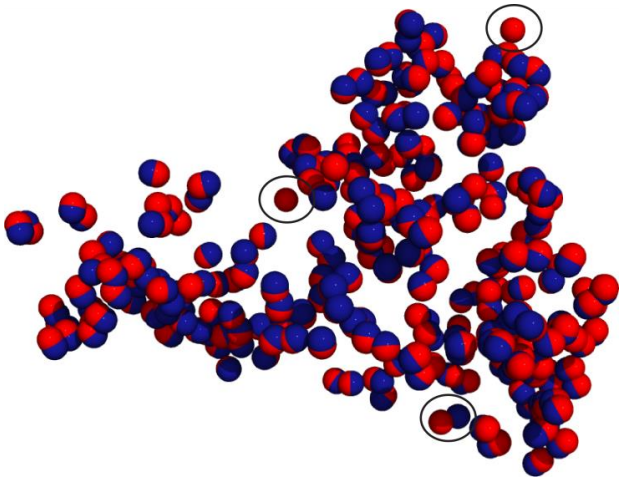


Figure A3.4 | Comparison of methyl carbon positions for the two molecules in the asymmetric unit of the MSG structure (PDB ID: 1p7t). Methyl carbon atoms of the molecules corresponding to chains A and B are given in red and blue, respectively. Most of the methyl carbon positions overlap, however some show significant differences between the two molecules (circled). In addition, chain A has an extra CD1 atom corresponding to Ile5 residue, which is missing in chain B (top right circle).

Chapter 4: Applications of MAGMA to molecular chaperones

Abstract

Heat shock protein 90 kDa (Hsp90) is an essential molecular chaperone in eukaryotes whose clients include a number of oncological targets. The N-domain of human HSP90 contains an important ATP binding site and has been targeted for pharmacological inhibition of ATPase activity. Here, methyl resonance assignments for the N-domain of ILV-labelled HSP90α were obtained using sparse inter-methyl NOEs only. Two different HSP90α conformations were used as input, and, after joining their assignment results, 38% of the ILV residues entering the calculation were confidently assigned. The confident and ambiguous MAGMA assignments were combined with inter-molecular NOE restraints in HADDOCK to obtain complexes of the N-domain with a compound containing an aminotriazine moiety. The generated models featured the correct binding site, but their overall quality was lower when compared to docked complexes that used a complete set of previously published assignments (0.5–2.0 Å higher ligand RMSD). Interestingly, incorporating inter-molecular NOEs in the MAGMA calculation increased the number of confidently assigned methyl resonances, and demonstrates the benefit of including additional restraints in MAGMA. Finally, a generic data collection protocol for MAGMA in a *de novo* assignment application on the 400 kDa archaeal small heat shock protein Hsp16.5 was established. This dataset included an ILVM-labelled sample, a double-LV methyl labelled sample, and a selectively L-labelled sample from which inter-methyl NOEs, intra-LV NOEs, and discrimination between L and V resonances were respectively obtained. A high sparsity of NOE restraints permitted only ambiguous assignment to be obtained for most of the methyl resonances.

In this chapter, I explore the potential of MAGMA to assign methyl resonances of two novel protein targets, which were not part of MAGMA's benchmark. The two proteins are the N-domain of human heat shock protein 90 α (HSP90α) and the 400 kDa oligomer of heat shock protein 16.5 (Hsp16.5). These proteins are molecular chaperones of biological importance that have been investigated together with other researchers. As such, the work presented here is a result of a collaborative effort, with the data contributed by my collaborators (Dr. Eiso Ab (EA), Dr. Marta G Carneiro (MGC)) and colleagues (T Reid Alderson (TRA) and Prof. Andrew J Baldwin (AJB)), as marked in the text with their initials.

4.1 Introduction

Hsp90 – conserved and essential eukaryotic molecular chaperones

The 90 kDa heat shock protein (Hsp90) family constitutes one of the most abundant and most conserved molecular chaperones[1], and is essential in eukaryotes[2,3]. Hsp90 chaperones show specificity for their numerous, diverse clients (over 280 identified)[4,5], which are involved in a range of cellular processes such as signal transduction, protein degradation, intracellular transport, cell cycle progression, transcriptional regulation, etc.[1,3,6,7]. Interestingly, Hsp90 is not required for *de novo* folding of most proteins[1], but it facilitates the maturation of its clients by a mechanism that remains to be elucidated. Hsp90 interacts with more than 20 different co-chaperones[1], which control its ATPase activity and, by acting as adaptors, mediate recruitment of specific clients[6].

Hsp90 is composed of three domains, an organization that is evolutionary conserved from bacteria to humans[1]: an N-terminal nucleotide binding domain (N-domain)[7], a middle domain (M-domain), and a C-terminal domain (C-domain) responsible for dimerization[8] (Figure 4.1). Full-length Hsp90 forms a homodimer with two distinct conformations, as revealed by X-ray crystallography studies on *E.coli* Hsp90 (HtpG)[8] and Hsp90 from yeast[4] (Figure 4.1). The "tense", closed form of the nucleotide-bound yeast Hsp90 in complex with p23/Sba1[4], and

the "relaxed", open form of the apo HtpG[8] represent the end points of the dynamic and conformationally coupled ATPase cycle of Hsp90.

The onset of structural changes that characterize the so-called ATPase-coupled molecular clamp mechanism is mediated by nucleotide binding to the N-domains of the homodimer. Nucleotide binding promotes dimerization of the N-domains within the constitutive dimer, which is held together by interactions involving the C-domains[6]. The "tense", closed state with bound nucleotide is characterized by interactions of both N- and C-domains resulting in a left-handed helical twist along the vertical axis of the homodimer[4]. Completion of the conformationally coupled catalytic cycle involves intramolecular interactions between the N- and M-domains, leading to a conformational change in the M-domain that brings together the "two halves of a split active site"[6]. The hydrolysis of ATP to ADP and inorganic phosphate results in dissociation of the N-domains and return of the homodimer to a "relaxed", open conformation[1,6] (Figure 4.1A).

Hsp90 is an anti-cancer drug target

Mammalian cells have two isoforms of Hsp90 in the cytosol, the stress induced Hsp90α and the constitutively expressed Hsp90β[1]. HSP90α, the human cytosolic isoform, is significantly up-regulated in different forms of cancer[5]. HSP90α can be secreted and its high levels in the plasma are associated with tumour malignancy, while inhibition of HSP90α secretion reduces tumour metastasis[9]. A number of oncological targets are amongst the numerous clients of Hsp90, of which protein kinases and nuclear steroid receptors are the most investigated substrates[6,10]. Pharmacological inhibition of the ATPase activity of Hsp90 has been shown to result in the degradation of client proteins *in vivo*, making Hsp90 an attractive target for anti-cancer therapeutics[6].

Investigations of Hsp90 inhibitors started in late 1990s with geldanamycin and radicicol. Both of these natural compounds were shown to inhibit Hsp90's ATPase activity, and their

binding sites in the N-domain colocalize with the ATP/ADP binding site[11,12]. A modified version of geldanamycin, 17AAG, was the first inhibitor to enter clinical trials, after previously showing antitumor activity on a number of preclinical models *in vitro* and *in vivo*[6,13]. Currently, over 50 different Hsp90 inhibitors are being tested in various stages of clinical trials[13]. The inhibitors compete for the ATP-binding site, thereby obstructing chaperone's dynamic ATPase cycle, and reducing its affinity for clients, which ultimately leads to their proteasome-mediated degradation of the clients[13].

Highly conserved N-domain of Hsp90 is a nucleotide and drug binding site

The binding site for nucleotides and inhibitors in the N-domain of Hsp90 is highly conserved, with 82% of the residues identical from bacteria to humans[7,10,11]. The ATP binding site is non-canonical, featured by a 15 Å deep binding pocket of mixed hydrophobic and polar composition[11]. Crystal structures of the N-domain complex with ADP revealed that the nucleotide forms a single hydrogen bond with the protein, from the adenine base to D79, whilst the rest of the hydrogen bonding is mediated by water molecules coordinated by protein groups within the pocket[7].

For the isolated N-domain of human HSP90α, different conformations have been captured by X-ray crystallography, dependent on which nucleotide or ligand is bound[7,11,14,15]. The most pronounced differences, apart from those in the flexible loops, cluster to the region I104-G114, which alternates between helical and loop conformations in the different complexes (Figures 4.1C, 4.3). A small part of this region contributes to the "lid segment" (A111-G135), important in formation of the active, "closed" conformation of Hsp90 upon ATP binding[5] (Figure 4.1A).

NMR spectroscopy can provide simultaneous, atomic-level insight into the kinetics, molecular interfaces, and conformational changes that accompany biomolecular interactions. As such, various aspects of the dynamic and conformationally rich ATPase cycle of Hsp90,

as well as interactions with Hsp90 and various co-chaperones and clients, are well suited for NMR spectroscopy studies[16]. The early NMR studies of Hsp90 have focused on isolated domains (N-, M-, and C- domains) of yeast and human homologues, due to the high molecular mass of the full-length Hsp90 (~170 kDa for the dimer of the human cytosolic isoform), as reviewed by Didenko *et al.*[16]. An interesting result from such NMR studies on the isolated N-domain revealed differential internal dynamics in the free-state, compared to ATP analogue (AMP-PCP) and ADP-bound states[5].

The NMR studies of the full-length human cytosolic isoforms HSP90α (construct HSP90Δ, with residues 241-268 removed)[17] and Hsp90β[18] were made possible with the methyl-TROSY technique. One of the major difficulties in these studies was the assignment of Ile-δ1 resonances in methyl-TROSY spectra. For HSP90α, partial assignments were achieved by transfer of Ile assignments obtained for the constructs of its constituent domains (N- and M-domains). Those were in turn obtained by relating the backbone resonance assignments of the domains with connectivity patterns in their ^{13}C- and ^{15}N- NOESY spectra with reference to the crystal structure of yeast Hsp82[4,17]. The assignment of 11 Ile-δ_1 resonances from the C-domain was not possible. However, the remaining methyl assignments were used to characterize the dynamics of the different domains of HSP90α in solution, HSP90's interactions with the ATP analogue AMP-PMP, and its client p53[17]. For HSP90β, partial assignments of Ile-δ1 resonances were again achieved by transfer of methyl assignments from the individual N- and M-domains, which were assigned through correlation of the backbone to the side chain resonances using a ^{13}C-resolved CCH-TOCSY experiment[18]. The assignment effort was supplemented with inspection of connectivity patterns from ^{15}N-resolved 3D NOESY experiments[18]. The methyl assignments were used to monitor the interaction of HSP90β with ATP, and the co-chaperone p23, as well as the associated conformational changes. More recently, a combination of methyl-TROSY and SAXS

provided a structural model of Hsp90's interaction with its intrinsically disordered client Tau[19].

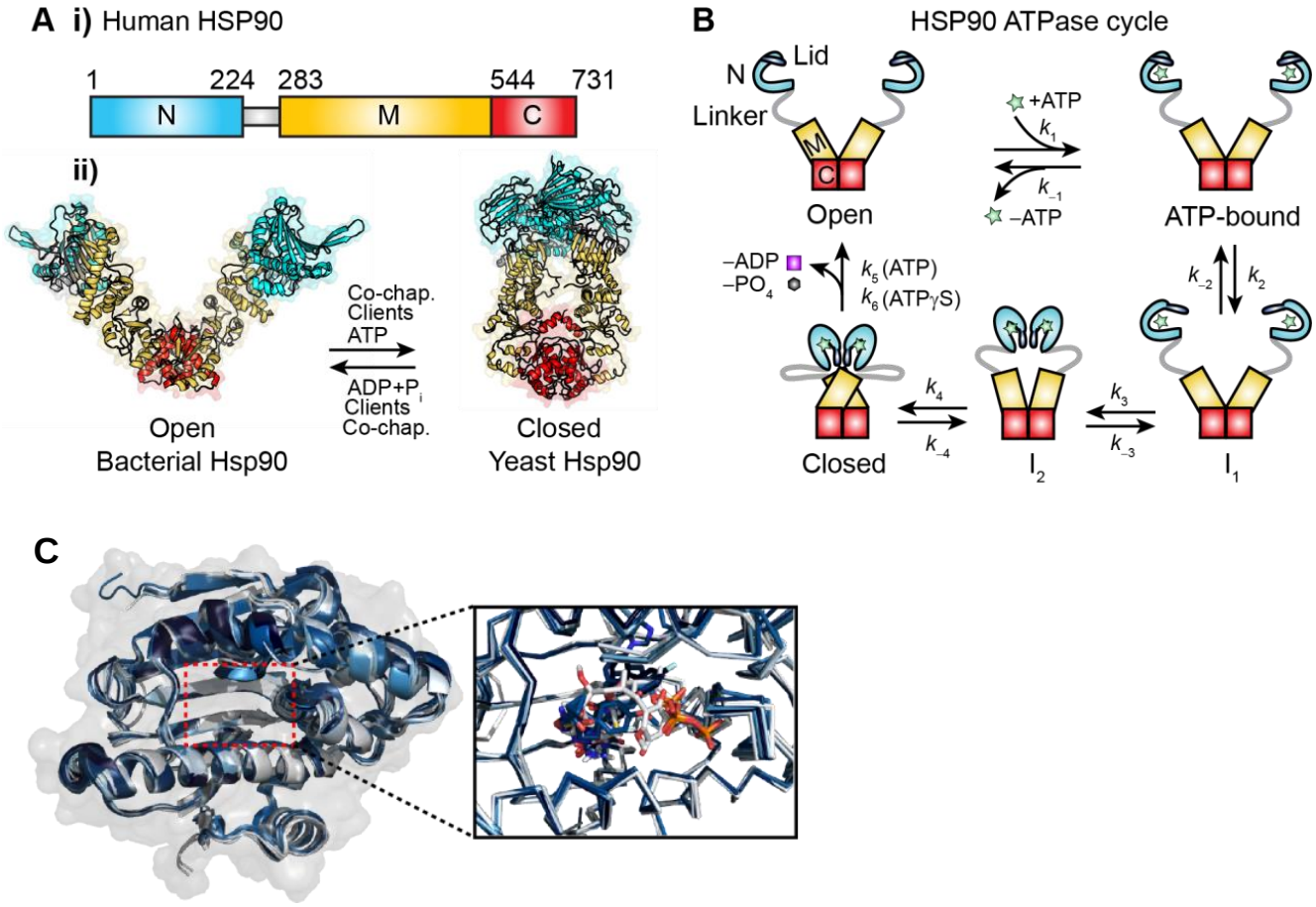


Figure 4.1 | **A** Schematic of the full-length construct of human HSP90α with domain boundaries indicated with amino acid numbers. **B** The "ATPase-coupled molecular clamp" mechanism of the HSP90 homodimer is mediated by nucleotide binding and hydrolysis in the N-domains[1,3,6](see 4.1). **C** Different conformations of the N-domain of HSP90α from different ligand-bound complexes including, amongst others, complexes with geldanamycin[7,11], an aminotriazine compound[14], ADP[7], and ATP[15]. The region of the protein that binds ATP, ADP, and different ligands is highlighted to indicate the overlap in the binding sites.

Hsp16.5 – a 400 kDa molecular chaperone and a target for a methyl-TROSY case study

Small heat shock proteins (sHsps) are a class of molecular chaperones that function in an ATP-independent manner[20,21]. sHsps assemble into large oligomers that vary in size and degree of polydispersity, including monodisperse 24-mers[22] and polydisperse oligomers spanning from 10 to over 40 subunits[23-25]. Insight into the structure and dynamics of

mammalian members of this class of molecular chaperones has proved challenging due to their general tendency to form dynamic, polydisperse oligomers[20,21].

However, high-resolution structural insight into an archaeal sHsp, Hsp16.5, has been gained, featuring a symmetric 24-mer of an external diameter of ~120Å (Figure 4.8)[22]. Hsp16.5 belongs to the extremophile *Methanocaldococcus jannaschii* that lives at 85°C and an optimal pH of 6.0[22], and Hsp16.5 was shown to prevent the aggregation of unfolded protein substrates *in vitro*[26-28], with higher chaperoning activity at elevated temperature[27]. The small monomeric mass of the protein (16.5 kDa), the availability of the crystal structure, and its symmetry together with its stability at elevated temperatures make Hsp16.5 an ideal target for a methyl-TROSY case study with MAGMA.

Here, I investigate both the N-domain of human HSP90α and full-length *M. jannaschii* Hsp16.5 by solution-state NMR spectroscopy with a focus on obtaining methyl assignments based on only sparse methyl-methyl NOE restraints, which were collected in the presence of a ligand with an aminotriazine moiety (HSP90α) or on the *apo* protein (Hsp16.5). For HSP90α, in addition to cross validating MAGMA's assignments, I investigated the potential of the generated unambiguous and ambiguous assignments to determine a structural model for the ligand bound complex by molecular docking. Finally, I explored the influence of different HSP90α conformations and ligand binding modes on MAGMA's assignments, and how MAGMA could be used to assess which conformers explain the experimental NOE restraints and known resonance assignments. The set of restraints supported by MAGMA were extended to include intermolecular NOEs, which show a great potential for future MAGMA applications. Furthermore, it is demonstrated how different NMR restraints, essential for the MAGMA strategy, can be collected and incorporated in a calculation on the 400 kDa molecular chaperone Hsp16.5, for which methyl resonance assignments were not previously known.

4.2 Results and discussion

4.2.1 MAGMA generates correct assignments for N-domain of Hsp90

To assess the accuracy of methyl assignments generated by MAGMA, ^{1}H-^{13}C resonance assignments of a protein of interest need to be available. For the N-terminal (ATPase) domain of HSP90α the backbone resonance assignments were previously obtained (BMRB 5355)[29] and confirmed with TROSY-HNCACB and TROSY-HN(CO)CACB experiments in the study by Shah *et al.*[30]. In the same study, CCH-TOCSY data were used to correlate the Cα and Cβ resonance assignments to the methyl resonances of the same residue, generating assignments for ~80% of the Ile, Leu, and Val side-chains.[30] These assignments were used for cross-validation of MAGMA's results. The restraints for MAGMA comprised inter-methyl NOEs that had been extracted from a NOESY spectrum collected on a fully deuterated, methyl protonated sample of the 27 kDa N-domain construct (see Methods; spectra collected by EA). The spectrum was recorded in the presence of a ligand with an aminotriazine moiety, introduced in an earlier study of the N-domain[30] (Methods). The ^{1}H-^{13}C HSQC spectrum with Ile, Leu and Val resonances that were subject to assignment by MAGMA is given in Figure 4.2 A, and the ^{13}C-^{13}C projection of the NOESY spectrum is shown in Figure 4.2 C.

The amino acid types of the methyl resonances, including discrimination between Leu and Val types, an experimentally achievable goal[31], were assumed to be known. In addition, the information about the two methyl groups of Leu/Val that belong to the same residue was given to MAGMA based on the known assignments[30]. This allowed for merging of the inter-methyl NOEs (edges) to a residue vertex in the data graph (Chapters 2 and 3). The NOE restraints were assigned only to the resonances for which the assignments have been previously known, to allow assessment of accuracy.

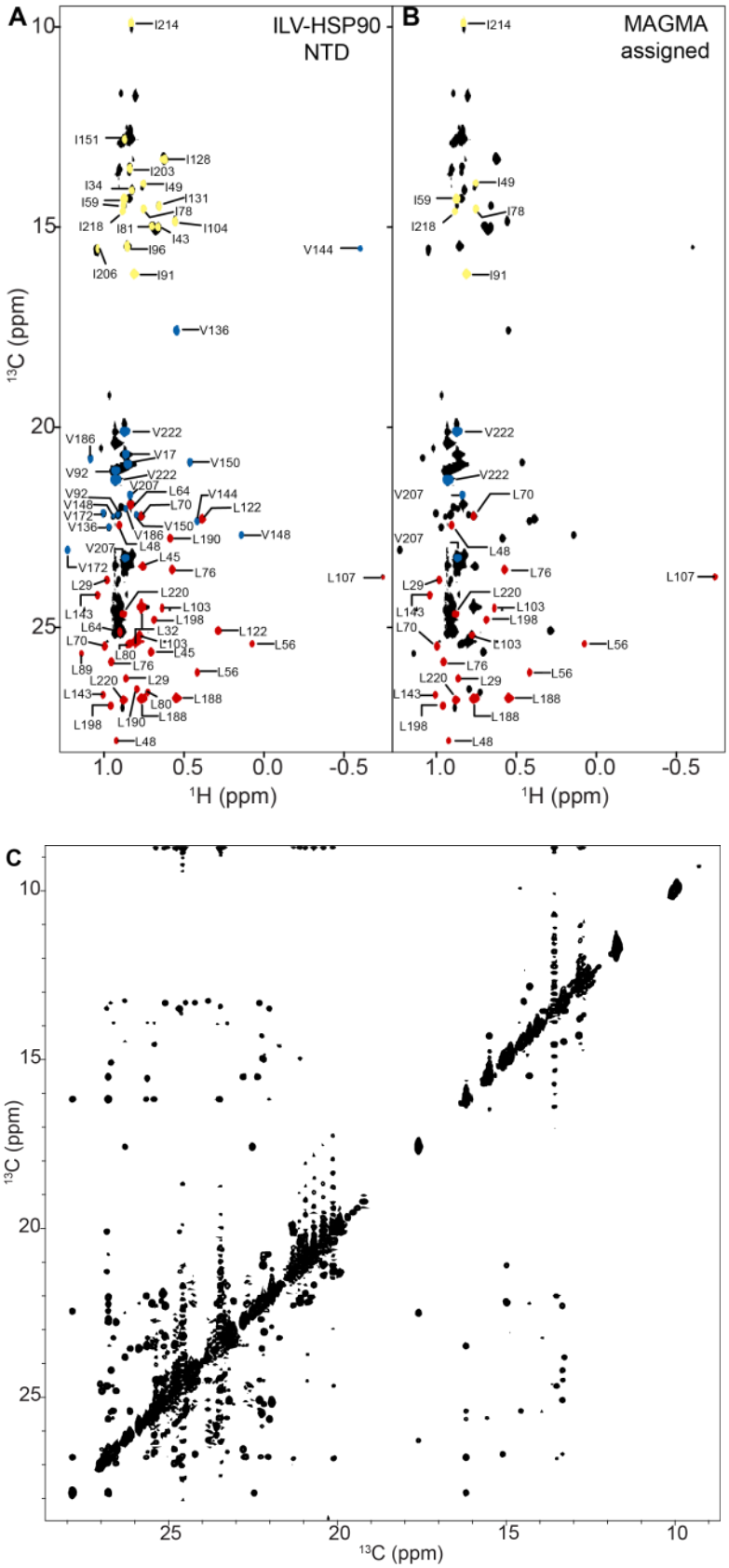
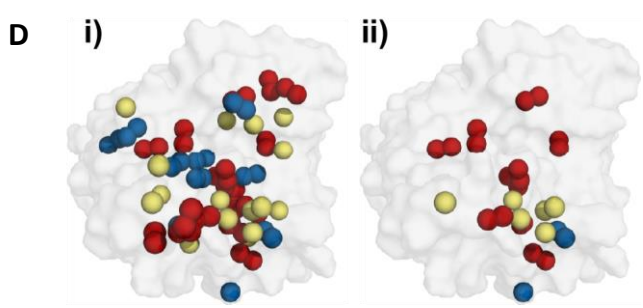


Figure 4.2 | NMR spectra of the N-terminal domain of HSP90α. **A** ^{1}H-^{13}C HSQC spectrum of a deuterated, ILV-methyl protonated sample of the N-domain of Hsp90 with indicated amino acid types: Ile (yellow), Leu (red), Val (blue). Methyl resonances assigned confidently by MAGMA (**B**). Assignment for the resonance indicated in **i)** has been previously determined[30], and these entered MAGMA's calculation to test the accuracy of the automatic assignment (see Methods). **C** ^{13}C-^{13}C projection of the 3D CCH-NOESY spectrum of the N-domain.



D Methyl groups with known resonance assignment displayed on the surface representation of the NTD of Hsp90α (**i)**. Methyl groups assigned unambiguously by MAGMA on the surface representation of the NTD of Hsp90α (**ii**). In both cases carbon atoms of the methyls are shown as spheres colored according to the type as indicated in **A**.

A total of 72 resonances (out of 115 observed), corresponding to 45 methyl-bearing residues with known assignments, were given as input to MAGMA. MAGMA was run according to its usual protocol, with the distance threshold of 10 Å.

Using the available restraints and two different input structures (Figure 4.3), MAGMA provided a unique (confident) assignment for 38% of the methyl bearing residues (Figure 4.2). All of the confident assignments were correct, as determined by the cross-reference to the available assignments. To assess the performance of MAGMA without discriminating between the Leu and Val resonances, MAGMA was run with a joined LV label. As expected, the number of confident assignments was lower and covered only 12% of the methyls entering the calculation (data not shown). Notably, all the confidently assigned resonances were of Ile type, indicating that Leu/Val resonance label merging significantly impacted the confidence of the assignments. This result confirms the already established conclusion that more distinct labels significantly increase the confidence of MAGMA's assignments (Chapter 2, 2.2.5). Moreover, in case of ILV labelling, an average of 4.7 NOEs were required per confident assignment, whereas for I(LV) labelling, up to on average of 23.6 NOEs were required. For all methyl resonances in the HSP90α calculation, apart from Ile128, the known assignment was within the set of ambiguous assignments provided by MAGMA (Appendix Table A1).

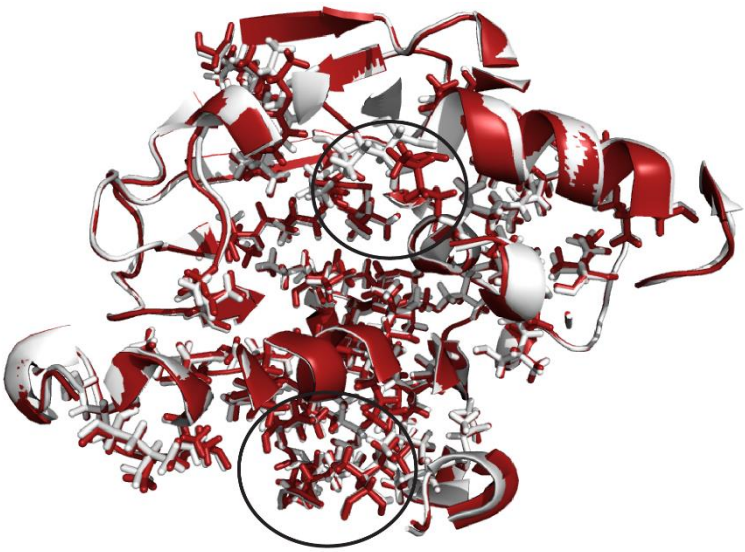


Figure 4.3 | Different conformations of the N-domain of HSP90α used in MAGMA calculation; in white (HSP90_N1), in red (HSP90_N2) in cartoon representation. Side chains of Ile, Leu and Val residues are shown as sticks. Regions with the most significant structural differences and deviations in superimposition of the side-chains are circled (see text for description).

Compared to MAGMA's benchmark, the density of the NOE restraints in these calculations was 0.42, comparable to the proteins EIN1 (0.37) and MSG (0.47) for which similar percentages of confident assignments were obtained, 39% and 38%, respectively. Despite this incomplete coverage of resonances, even the ambiguously assigned methyls can be valuable, as exemplified below with their use in computational docking (4.2.3).

4.2.2 Choice of the N-domain structure influences assignment results

As previously mentioned, the methyl assignments for the N-domain of HSP90α were obtained for two different input structures. The results presented in Section 4.2.1 have been obtained by joining the assignments for each resonance for each of the two structures. The two structures were chosen as representatives of an ensemble of 17 crystal structures of complexes of the N-domain and different ligands (as determined by EA). The structures show two different backbone conformations for the residues I104-A111 (Figure 4.3). In one representative structure, the residues are part of a helix encompassing residues K100-A124 (Figure 4.3 white). In the other, the residues loop out of the helix, effectively breaking it into

two helices: helix (K100-L103)-loop (I104-A111)-helix (K112-A124) (Figure 4.3 red). It was not known *a priori* which of the conformations was predominantly sampled in solution, so MAGMA calculations were performed with individual structures followed by the merging of the assignment results.

The "merging" strategy came at the expense of confident assignments, as one of the structures (HSP90_N1) resulted in a set of 26, whereas the other in a set of 19 confidently assigned methyls (HSP90_N2, Figure 4.4 A, B). Upon merging, 17 confident assignments were in the intersection, HSP90_N1 ∩ HSP90_N2, i.e. common to the two sets (Figure 4.4 C ii)). In the symmetric difference of the confident sets, HSP90_N1 Δ HSP90_N2, six assignments are inconsistent with the known assignment. All of these assignments are elements of the HSP90_N1 set (Figure 4.4 B). This comparison shows that even small structural differences can result in significant differences in the sets of confident assignments obtained by MAGMA. Cross-referencing to the available methyl assignments shows that six methyls swap their assignments when the HSP90_N1 structure is used (Figure 4.4 A, B). On the other hand, using HSP90_N2, the assignments of these methyls are ambiguous, and sample both the HSP90_N1 assignment and the accurate assignment.

To assess the cause of the assignment differences, I assessed differences in the sets of edges of the two structure graphs, as these are the cause for the differences in assignments. Interestingly, the two structure graphs have the exact same size, i.e. the number of edges, and hence same restraint density (Table 1). Nonetheless, the two graphs differ in the identity of 14 edges, which are listed in Appendix (Table A4.3).

One could assume that the data graph vertices of the residues that form the intersection of the confident assignment sets show no differences in edges incident with them. Surprisingly, this is not the case as indicated in the Appendix (Table A4.3, grey). Despite the different edge structures of the two graphs their vertices have been consistently assigned to the same methyl

resonances. In contrast, some of the edge differences have been reflected in differences in the confident assignments (Table A4.3, red). These edge differences contributed to the confident assignment of six methyl bearing residues when using HSP90_N1 as the input structure, which were assigned ambiguously when using HSP90_N2. I further examined the differences in inter-methyl distances that define these edges in the two structure graphs. The details of this analysis are given in the Appendix (A 4.1).

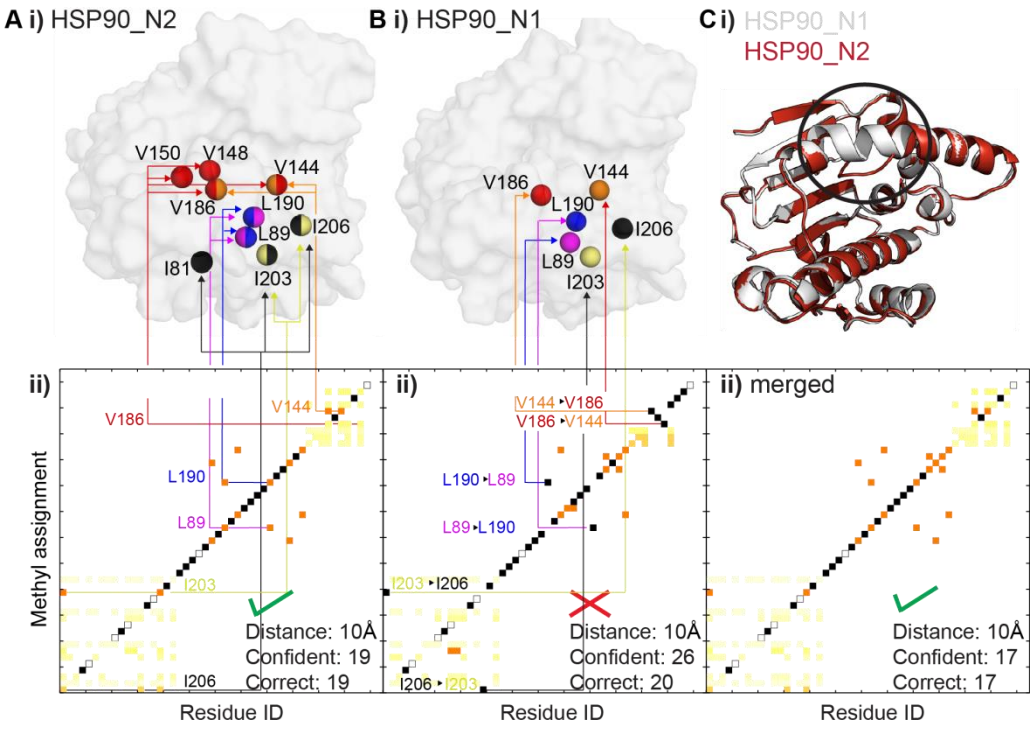


Figure 4.4 | Two different conformations of the NTD of Hsp90α result in different sets of confident assignments (**A, B**). When different structures for the same protein are given as inputs in MAGMA's calculation, differences in the sets of confident assignments can arise (**A, B i)**, **ii)**). In the case of the conformation HSP90_N1 **(B)**, more confident assignments are obtained. The additional assignments are in positions ambiguously assigned in the conformation HSP90_N2 (**A**). The specific differences in assignments are indicated on the assignment plots (*y*-axis = methyl resonance, *x*-axis = residue assignment) (**ii**) and connected to the assigned residue positions on a surface representation of the NTD (**i)**. The differences in the two protein conformations are best seen in the overlay of their cartoon representations (**C i**)) with the only significant change (in helix H5) indicated by an arrow. If the conformation of the protein present during the NMR measurement is not known *a priori*, MAGMA's assignment for the two (or multiple) different structures needs to be combined (**C ii**), see Chapter 3). The total set of confident assignments after combining, is obtained by additional considerations to the mutually exclusive assignment combinations as discussed in the Methods.

Table 1: Comparison of the data graph and the structure graphs of the N-domain of HSP90α. For both conformations, the structure graph has the same number of vertices and edges (HSP90_N1 and HSP90_N2). *Restraint density = $\frac{|data\ graph\ edges|}{|structure\ graph\ edges|}$

Vertices (data graph)	42
Edges (data graph)	71
Vertices (structure graph)	48
Edges (structure graph)	170
Data graph	
Restraint density*	0.42
Average degree	3.38
Median degree	3
N (connected data subgraphs)	5

Next to the two selected structures (EA and MGC), another structure of the N-domain exists in the PDB and contains the same ligand that has been added to the NMR sample during the NOESY measurement (PDB ID: 3b24)[14]. The PDB entry features two protein molecules in the asymmetric unit with slightly different conformations, with the most prominent differences spanning residues I104-G114. The conformation of the loop I104-A111 in one of the structures (chain A) resembles HSP90_N2. In the other structure (chain B), the loop shows a third conformation that is shifted to the residues G108-G114. Interestingly, the two molecules also show differences in the orientation of the ligand in the binding site. These different ligand orientations are discussed in more detail below in the context of ligand-protein docking (4.2.3) and MAGMA's calculation that include intermolecular ligand-protein NOEs (4.2.4).

Since the N-domain of HSP90α shows different conformations in residues 104-111 and 104-114 when bound to different ligands (Figure 4.1 C), and these conformations resemble those found in the four structures of the complex with the ligand of interest, I compared MAGMA's performance on these four different structures (Table 2).

Table 2: Five different structures of the N-domain of HSP90α used as input in MAGMA

Structure ID	N confident methyl residue assignments	Methyl assignment cross-validation
3b24_A	4	✓
3b24_B	22	×
HSP90_N1	26	×
HSP90_N2	19	✓
1YER_1	14	✓

In addition, there exists another structure of the N-domain (1YER_1), which is a model obtained in the previous work on the protein ligand complex[30]. The conformation of this structure in the region I104-G114 resembles that of HSP90_N2, and passes the assignment cross-validation. Surprisingly, 3b24_A, which also resembles the conformations of HSP90_N2 and 1YER_1, yields significantly more ambiguous assignment when compared to the rest of the structures. A significant fraction of the ambiguously assigned resonances have two or three residue assignment possibilities. This nonetheless indicates that not only the most prominent structural differences, but also small changes in the positions of the methyl-bearing side-chains, can affect the assignment outcome. When using MAGMA for assignment purposes, it is at present best to rely on an ensemble-averaged assignment (Chapter 3, Figure 4.4 C).

4.2.3 MAGMA's methyl assignments can be used in protein-ligand docking

Molecular docking is a computational method that generates an ensemble of conformers for a molecular complex on the basis of the known structures of the individual components[32]. Protein-ligand docking is an essential and a well-established part of the drug discovery process, and is capable of predicting the binding modes and affinities *in silico*, in a rapid and economic manner[33]. Any protein-ligand docking protocol will consist of a search algorithm, which should effectively sample the vast degrees of freedom of the protein-ligand complex

and generate different ligand conformations and binding poses (orientations). These docked models are then evaluated using an appropriate scoring function, enabling a ranking of the models and ideally filtering out the true binding modes. In addition, the scoring function should provide a good estimate of the binding affinity for the different binding modes[33].

Intermolecular NOEs, which link proximal chemical groups of a protein and a ligand, are valuable restraints for protein-ligand docking. These restraints can be generated by NMR spectroscopy and supplied directly to the HADDOCK docking protocol[34]. I was interested in the potential of confident and ambiguous methyl assignments, produced by MAGMA, in generating structural models for the complex of N-domain of HSP90α and an aminotriazine ligand. If successful, such a combined MAGMA-HADDOCK protocol could be applied to a range of novel protein-ligand targets.

We focused on the methyl resonances of the N-domain of HSP90α that give rise to the intermolecular NOEs to H1, H2, and H3 groups of the ligand[30] (Figure 4.7). The set of MAGMA's assignments for these resonances defined a set of ambiguous distance restraints for HADDOCK[35] (MGC, Methods). Several docking calculations were performed using MAGMA's or previously published assignments, two different protein structures, and different protein-ligand docking settings in HADDOCK, as summarized in Table 3 (MGC).

Table 3: Summary of the HADDOCK calculations run on the basis of MAGMA's and previously published restraints

HADDOCK run	Description
MAGMA combined	Distance restraints from combined MAGMA assignments for structures HSP90_N1 and HSP90_N2 (Figure 4.4 C)
MAGMA combined + active	Distance restraints from combined MAGMA assignments for structures HSP90_N1 and HSP90_N2 + L103 and L107 defined as active residues
Published-methyl only	Distance restraints from Shah *et al.*[30], excluding those to HN groups + 96,103,107,150,186 as active residues
Published	Distance restraints from Shah *et al.*[30] including those to HN groups + 96,97,98,103,107,150,186 as active residues

In the calculation denoted “MAGMA combined + active”, L103 and L107 were defined as active residues, as these were the only two methyls that contributed the intermolecular NOE restraints and were confidently assigned by MAGMA. Calculations “Published” and “Published–methyl only” respectively used previously published methyl and backbone assignments or only methyl assignments as restraints[30]. These assignments were used in a previously published protocol that was based on sparse intermolecular NOEs, methyl labelling, and HADDOCK, but required an unambiguous assignment of the side chain methyl resonances of the residues at the ligand binding site[30]. In addition, the protocol made use of the amide-ligand NOEs, thereby requiring assignment of the backbone resonances[30].

A comparison of the HADDOCK results for the previous protocol and the new protocol, reliant on MAGMA’s assignments, is summarized in Figure 4.5. Two different structures of the N-domain (Figure 4.3) were used in docking, generating two sets of HADDOCK models (Figure 4.5: A-B for HSP90_N1; C-D for HSP90_N2).

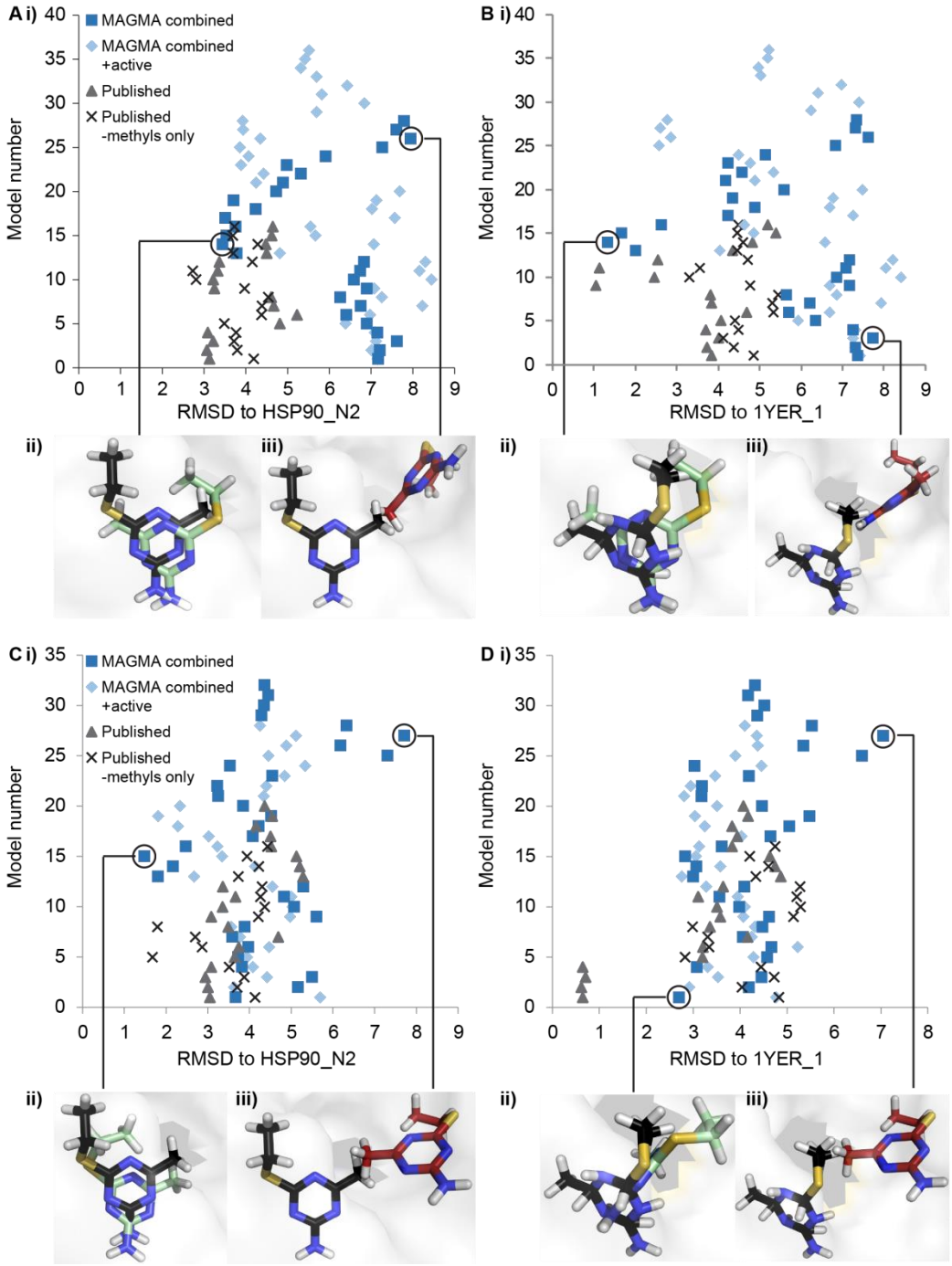


Figure 4.5 | Comparison of HADDOCK protein-ligand docking results using MAGMA's assignments, previously published backbone and methyl side-chain assignments, or only methyl side-chain assignments[30], and two different protein structures (HSP90_N1 and HSP90_N2). In **A** and **B**, the results of docking with HSP90_N1; **C** and **D,** the results of docking with HSP90_N2. **A-D i)** for each HADDOCK model (*y*-axis) the RMSD of the model ligand to the ligand in the structure HSP90_N2 (left), or 1YER_1 (right) is displayed (*x*-axis). **A-D ii)** superposition of the ligand structures from the best MAGMA-HADDOCK model and HSP90_N2 (left) or 1YER_1 (right); **iii)** same as in **ii)** for the worst MAGMA-HADDOCK model.

To assess the quality of the generated models, a reference structure of the N-domain-aminotriazine ligand complex is needed. However, two plausible structures of the complex exist: the crystal structure HSP90_N2 and the HADDOCK model generated using previously published assignments and the old protocol[30] (labelled 1YER_1). Both of the plausible structure models are also supported by a published crystal structure that features two ligand binding modes in two protein molecules in the asymmetric unit (PDB ID: 3b24)[14]. For this reason, in Figures 4.5 and 4.6, the docking results are referenced against both the HSP90_N2 structure (left) and 1YER_1 model (right). The RMSD between the heavy atoms of the ligand in each HADDOCK model and the reference ligand positions (1YER_1, HSP90_N2) were used as a metric of the model's quality.

Table 4: HADDOCK ensembles using the starting structure HSP90_N1. The ensemble with the lowest RMSD on average is in bold.

HADDOCK run	HADDOCK ensemble RMSD±SD (1YER_1)	HADDOCK ensemble RMSD±SD (HSP90_N2)
MAGMA combined	5.57 ± 1.92	5.89 ± 1.53
MAGMA combined + active	5.95 ± 1.65	6.15 ± 1.42
Published-methyl only	4.55 ± 0.57	**3.84 ± 0.52**
Published	**3.66 ± 1.28**	3.95 ± 0.79

Table 5: HADDOCK ensembles using the starting structure HSP90_N2. The ensemble with the lowest RMSD on average is in bold.

HADDOCK run	HADDOCK ensemble RMSD±SD (1YER_1)	HADDOCK ensemble RMSD±SD (HSP90_N2)
MAGMA combined	4.25 ± 1.03	4.32 ± 1.40
MAGMA combined + active	3.74 ± 0.65	4.05 ± 0.99
Published-methyl only	4.33 ± 0.82	**3.61 ± 0.89**
Published	**3.23 ± 1.41**	3.94 ± 0.79

A comparison of the HADDOCK models from "MAGMA combined" and "MAGMA combined+active" shows more spread in "MAGMA combined", reflected in the larger standard deviation (SD) of this ensemble in all comparisons (Tables 4, 5). With HSP90_N1

as a starting protein structure (Table 4), “MAGMA combined” produces a better ensemble (average RMSD = 5.57 and 5.89 Å) compared to “MAGMA combined+active” (average RMSD = 5.95 and 6.15 Å) for both reference structures (1YER_1 and HSP90_N2, respectively). This is reversed when the starting structure is HSP90_N2 (Table 5), with an average RMSD = 3.74 and 4.05 Å for “MAGMA combined+active” versus 4.25 and 4.32 Å for “MAGMA combined”, compared to 1YER_1 and HSP90_N2, respectively. In all cases however, “MAGMA combined” contributes the lowest RMSD model (circled and displayed in Figure 4.5).

When comparing the two different starting structures, on average, HSP90_N2 does better in all tested cases (Table 2 and 3). The most significant differences are seen in the ensemble “MAGMA combined+active”; $\Delta RMSD_{HSP90_N1\text{-}HSP90_N2}$ = 2.21 and 2.1 Å for 1YER_1 and HSP90_N2, respectively.

Finally, if HADDOCK–MAGMA-driven ensembles are compared to those using previously published restraints, the “Published” ensemble that uses unambiguous methyl and backbone assignments always outperforms MAGMA ensembles (Table 4, 5). When compared against 1YER_1 the “Published” ensemble is indeed the best performing. However, when compared against an alternative (HSP90_N2), the “Published-methyl only” ensemble is the best performing ensemble (Table 4, and 5; $\Delta RMSD_{Published\text{-}Published\text{-}methyl\text{-}only}$ = 0.11 and 0.33 Å, respectively). In Figure 4.5, the ligand position in the best and the worst models obtained with MAGMA restraints are compared to the reference. Given the sparsity of the restraints, the generated models are satisfactory, as in both the best and the worst case the ligand is docked to, or close to, the binding pocket.

There is consistency between MAGMA and HADDOCK when using methyl restraints, suggesting that the HSP90_N2 conformation, and perhaps also the HSP90_N2 ligand binding pose, is consistent with the methyl-methyl and methyl-ligand NOEs. For MAGMA, this is

evident in the accuracy of the methyl assignments when using the HSP90_N2 structure (Table 2, 4.2.3); for HADDOCK, this is reflected in the lowest average RMSD values of the "Published-methyl only" ensemble when compared to HSP90_N2 (Tables 4 and 5). Moreover, when running MAGMA with inter-molecular NOEs and the HSP90_N2 complex as an input, the correct methyl and ligand assignments were obtained (4.2.4). On the other hand, using methyl and backbone amide NOE restraints ("Published"), the earlier HADDOCK model 1YER_1 is the closest to the average of the ensemble. In addition, MAGMA generated accurate methyl assignment when using the 1YER_1 protein structure as input (Table 2, 4.2.3), and did not exclude the ligand binding mode from the 1YER_1 complex when using inter-molecular NOEs (4.2.4). Therefore, more investigation will be required in the future to distinguish between the two plausible ligand binding modes.

NMR data collected on the protein-ligand complex suggest a single predominant conformation in solution[30]. To determine which of the reference structures is correct, more work needs to be done. In the earlier work, three backbone amides with intermolecular NOEs to H3 group of the ligand were reported as critical for determining the correct ligand orientation[30]. As the first step toward identifying which structure is correct, MAGMA's calculations that will include the three intermolecular restraints to amide protons need to be performed. These will determine if one of the structures can be excluded on the basis of the inconsistencies with the methyl and/or ligand assignments (see 4.2.4). It will be important to check that the previously reported model (1YER_1)[30] can be confirmed.

Finally, using MAGMA's assignments to guide HADDOCKing is useful only if the correct binding mode can be filtered from the generated ensembles using HADDOCK score or some other relevant metric. In Figure 4.6, the HADDOCK score is plotted versus ligand-RMSD for each model obtained using MAGMA's methyl assignments. When using HSP90_N1 as an input, the HADDOCK score does not help discriminate the models of high quality (low

RMSD) from those of low quality (high RMSD), regardless of the reference used in the comparison (Figure 4.6 A-B). The best scored models feature considerable ligand displacement from the reference binding mode (Figure 4.6 A-B ii)). However, with HSP90_N2 as input, the best scored HADDOCK model is within 2 Å or 3 Å, from the reference (Figure 4.6 C-D). Interestingly, when compared to the ligand in HSP90_N2, the best scoring model features the correct orientation and minimal ligand displacement (Figure 4.6 C ii)). Nonetheless, the models do not show clear clustering of the low RMSD models with the lowest HADDOCK scores.

It will therefore be important to investigate if the HADDOCK score is more discriminative on a benchmark of ligand-protein complexes and if perhaps other metrics can be used in scoring the models. One approach could be to perform MAGMA calculations with both methyl-methyl and intermolecular NOE restraints and HADDOCK models as input (4.2.4). This could determine which models maximise the number of explained methyl-methyl and methyl-ligand NOEs, and remain consistent with methyl and ligand assignments.

Finally, a decision on the treatment of MAGMA's methyl assignments as restraints for HADDOCK needs to be made. In this example, it is not clear whether "MAGMA combined" or "MAGMA combined+active" performs better, since this is dependent on the input structure (see above). Using HSP90_N2, "MAGMA combined+active" performs better on average. To get a clearer conclusion of the best way to treat the restraints and score the models, more benchmarking is essential, featuring the protein-ligand complexes for which a reference structure is unambiguous.

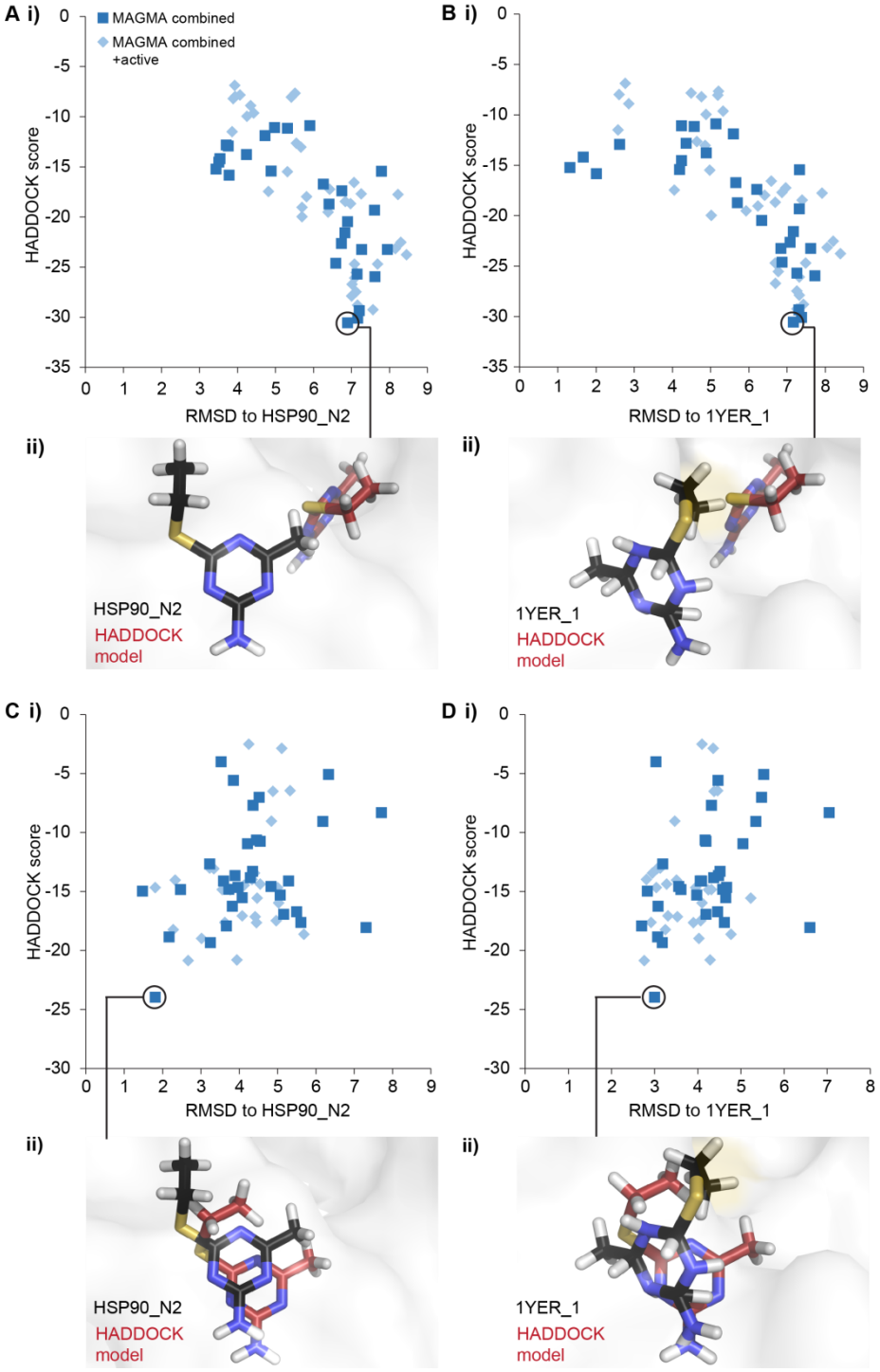


Figure 4.6 | HADDOCK score for the models generated using MAGMA's restraints versus ligand-RMSD of the models. In **A** and **B**, HADDOCK models generated using HSP90_N1 as the HADDOCK input structure and compared against the ligand position in HSP90_N2, and 1YER_1, respectively. In **C** and **D**, same as in **A** and **B** for the models generated using HSP90_N2 as the input structure for HADDOCK. **ii)** The position of the ligand in the best scored HADDOCK model with respect to that in the reference structure (HSP90_N2 or 1YER_1).

4.2.4 Inter-molecular NOEs between methyls and a ligand increase the number of assigned methyl resonances

Crystal structures of the aminotriazine compound in complex with the HSP90α N-domain have been determined previously. One (HSP90_N2) was provided for this study by collaborators of EA, and another two were reported independently (PDB ID: 3b24, chains A and B)[14]. Three binding modes in the available structures differ in the positions of the H1, H2 and H3 groups in the binding pocket, and in the ligand's orientation (Figure 4.7 B). The ligand orientations in 3b24 chain A and HSP90_N2 match, but in 3b24 chain B, the ligand is flipped 180° along the horizontal axis that passes through the nitrogen of the amino group (Figure 4.7 B). In addition, there exists a HADDOCK model of the complex, which was generated on the basis of the intermolecular NOEs between the ligand and the methyl and backbone amide groups of the protein (1YER_1, see above)[30]. Intriguingly, 1YER_1 is in agreement with the ligand orientation captured in 3b24 chain B.

Here, I used a subset of the reported intermolecular NOEs[30] as restraints in MAGMA, including all of those measured between the ligand protons and the methyl groups of the N-domain (Methods). The objective of the calculations was to examine:

1. If including the ligand protons as vertices in the data graph helps by reducing the ambiguity in the assignment of methyl groups close to the ligand
2. If different binding orientations (poses) of the ligand can be evaluated on the basis of their agreement with the measured methyl-methyl and intermolecular NOE restraints

Including the intermolecular NOEs in MAGMA's calculation requires a determination of the optimal distance for extraction of the protein-ligand edges from a structure of the complex. Here, the distance threshold between methyl carbons of the protein and H1, H2, and H3 protons of the ligand was set to 9 Å (Methods). The choice of the threshold was motivated by the fact that the intermolecular NOE data were collected with a large mixing time (~400 ms),

which might have resulted in spin diffusion contributions to the inter-molecular NOEs. A protocol for the determination of the distance threshold for inter-molecular NOEs will have to be determined in the future when a larger benchmark of ligand protein complexes has been assembled.

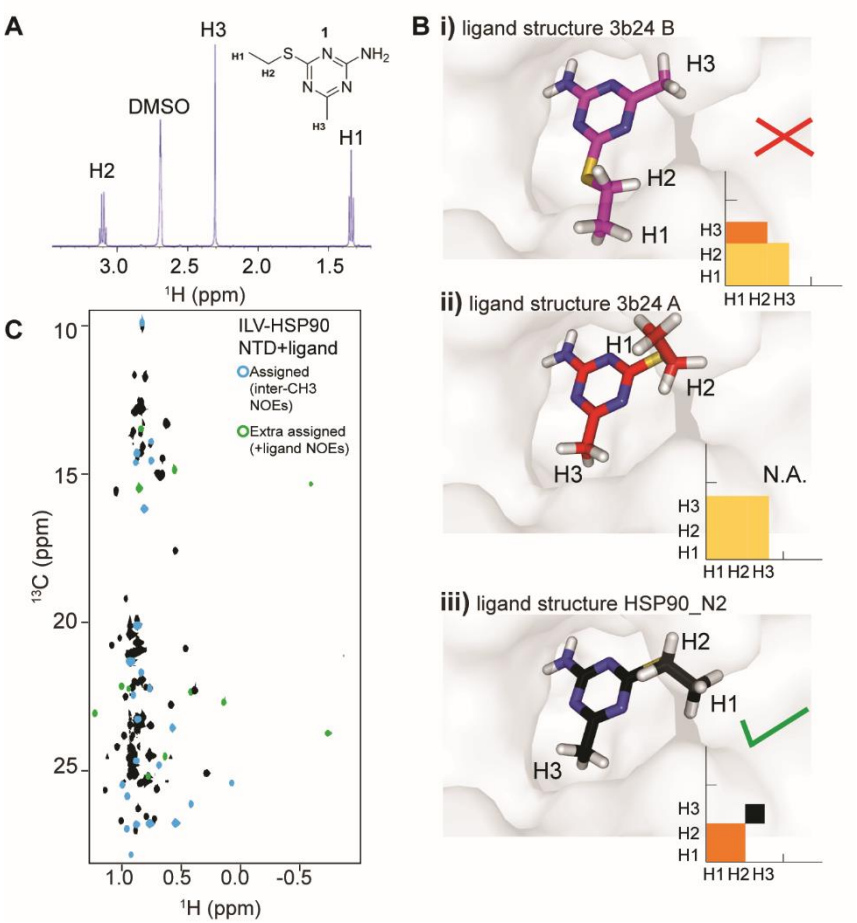


Figure 4.7 | Using intermolecular (protein-ligand) NOEs to generate more methyl assignments and the application of MAGMA to evaluate three different ligand binding modes. Compounds binding to the ATP-binding pocket of the N-domain of Hsp90 are of interest in anti-cancer drug design. **A)** A lead compound with an aminotriazine core motif that binds to ATP-binding site of Hsp90 with its assigned ^{1}H NMR spectrum. **B** Three alternative structures of the compound from **A,** bound to the Hsp90 N-terminus in a surface representation (i,ii,iii). The structures differ in the orientation (i,ii) and positions (i,ii,iii) of the H1, H2, H3 groups of the ligand. MAGMA's assignments for H1, H2, and H3 are shown and related to the structure. Based on the known ^{1}H assignments of the compound, the binding mode of the structure 3b24-B can be excluded. **c)** Confidently assigned methyl resonances are indicated in the ^{1}H-^{13}C HMQC spectrum. The assignments obtained on the basis of the ligand-bound structure and inter-methyl NOEs (blue) and additional assignments obtained in a calculation with intermolecular NOE restraints are shown in green.

Table 6. Result of MAGMA's calculations including intermolecular NOE restraints

Structure ID	N methyl-residue assignments	MAGMA assignment cross-validation	Known assignment[30] cross-validation	Ligand assignment[30] cross-validation
3b24_A	22	✓	✓	N.A./✓
3b24_B	29	✓	×	×
HSP90_N2	24	✓	✓	✓
1YER_1	23	✓	✓	✓

A summary of the MAGMA calculations including intermolecular NOEs is given in Table 6. For all tested complexes, the number of methyl assignments increased compared to calculations based solely on inter-methyl NOEs (see Table 2). The accuracy of the methyl assignments was assessed by cross-reference to MAGMA's assignments (based on inter-methyl NOEs, Figure 4.4 C), and to the known methyl assignment[30]. Since MAGMA's assignment includes all ambiguous possibilities for each resonance, in the cross-validation of MAGMA, the assignment of a methyl to one of the MAGMA's options is considered accurate. The cross-validation is potentially valuable, as for a blind case, MAGMA's assignment would be the only available one. Table 6 shows that MAGMA cross-validation is less restrictive than that against the published assignments and, on its own, would not discriminate against the complex 3b24_B. However, this complex would be excluded if a simple check against ligand proton assignments was performed[30]. These assignments are readily obtained from the analysis of a 1D ^{1}H spectrum of the ligand (Figure 4.7A).

The three complexes that pass all assignment checks (3b24_A, HSP90_N2 and 1YER_1) still feature two different ligand orientations, as previously discussed (4.2.3). In all cases, however, including the intermolecular NOEs increases the number of methyl assignments. For HSP90_N2 and 1YER_1 for instance, next to L107 the confident assignments obtained with the additional restraints included also I96. These are both important residues in the conserved hydrophobic binding pocket of Hsp90[11]. As discussed above, in the future, it will

be important to assess which of the binding modes can explain the majority of inter-methyl and intermolecular NOE restraints.

The results of MAGMA's calculations with intermolecular NOE restraints pave the route for novel applications of MAGMA. Introducing a small set of extra intermolecular restraints led to data graph vertices with unique labels and resulted in significant improvement in methyl assignment. Similar results could be expected from the incorporation of restraints based on paramagnetic relaxation enhancements (PREs). Once a larger benchmark of protein-ligand complexes is assembled, it could be further tested if MAGMA could also be used to rapidly screen multiple structures or models of ligand-protein complexes to reveal which are consistent with the NMR data.

4.2.5 Automatic assignment of the ^{1}H-^{13}C HMQC spectrum of the 400 kDa small heat shock protein 16.5 (Hsp16.5)

In the ^{1}H-^{13}C- HMQC spectrum of full-length Hsp16.5, 59 methyl resonances were observed (Figure 4.8 A-C). On the basis of the amino-acid sequence of the construct, and the employed methyl labelling scheme, 55 cross-peaks were expected: 21 Ile, 28 Val/Leu, and 6 Met (Methods). Since the requirement for a successful application of MAGMA is the determination of the methyl resonance amino acid types, I sought to first attribute an amino acid type to each of the observed resonances. Twenty-one resonances of Ile-δ1 type (as expected on the basis of the protein sequence) were readily detectable due to their characteristic up-field shifted ^{13}Cδ1 chemical shift values[36] (Table 7), and these were verified by the independent preparation of an Ile-only methyl-labelled sample. Six cross-peaks resonating between 2.0 and 1.5 ppm in the ^{1}H dimension were strong candidates for the expected 6 Met signals (Figure 4.8, Table 7), and these Met resonances were confirmed by a constant-time (CT) ^{1}H-^{13}C HSQC spectrum collected on a smaller, core-domain Hsp16.5 construct (data not included). The assignments of Ile and Met resonance types were also

confirmed by a ^{1}H-^{13}C HMQC spectrum collected on a double methyl-labelled {Leu-[$^{13}C\delta_1H_3$, $^{13}C\delta_2H_3$], Val-[$^{13}C\gamma_1H_3$, $^{13}C\gamma_2H_3$}, U-[^{2}H] sample that was used subsequently to determine which two methyl resonances belong to the same Leu/Val residue (see below, Methods).

Due to the narrow spectral range of Leu/Val resonances, the overlap of their methyl resonances required preparation of an additional exclusively Val-[$^{13}C\gamma H_3$], [U-^{2}H] labelled protein sample[31] that allowed discrimination between the resonances of the two amino acid types (performed by TRA, Methods). Figure 4.8 C (i) shows the final outcome of the determination of the methyl resonance amino acid types amounting to: 21 Ile, 6 Met, 18 Leu, and 14 Val resonances. Interestingly, four more resonances than expected on the basis of the sequence of the construct, two of Leu and two of Val types were observed. The extra resonances could be the result of conformational exchange on a slow NMR time-scale or sample degradation, and will be further investigated starting with, for instance, ZZ-exchange spectroscopy[38].

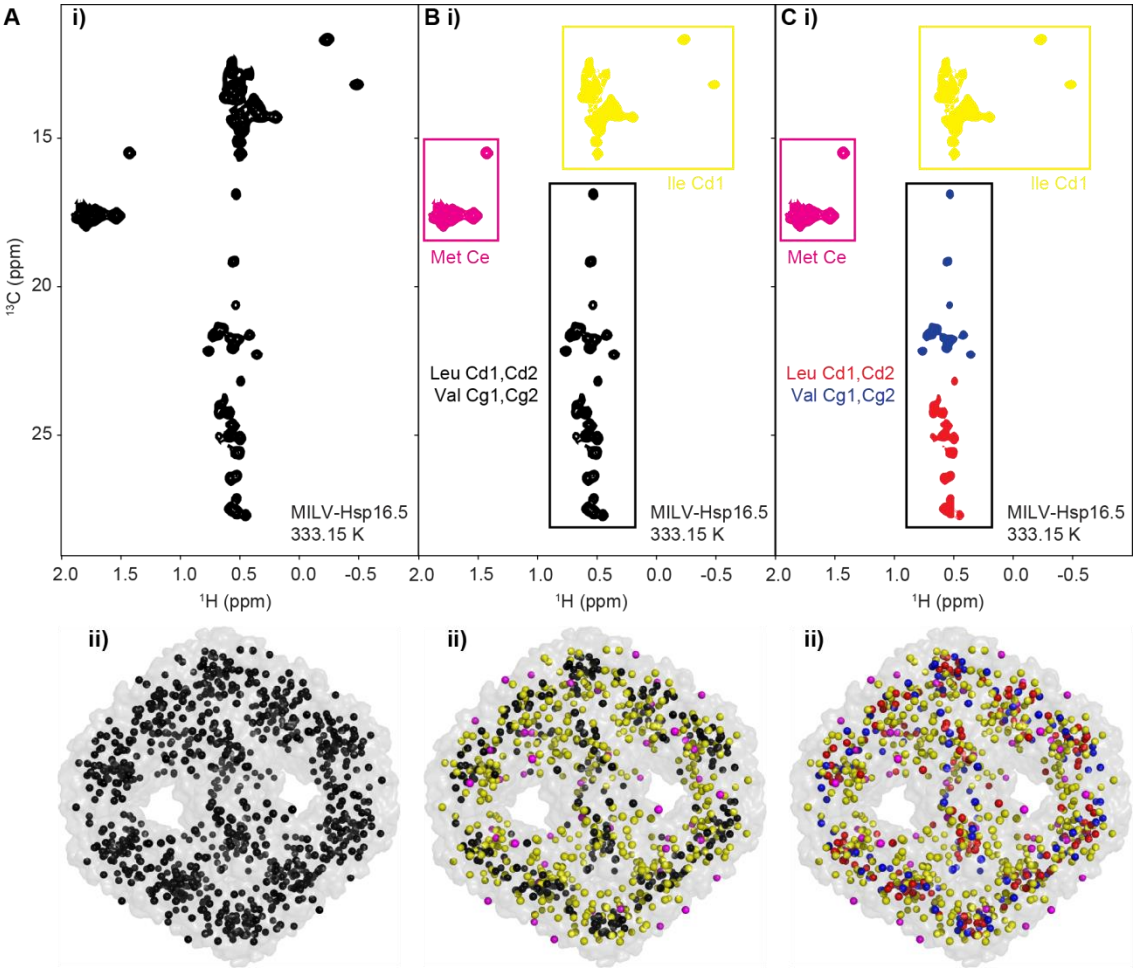


Figure 4.8 | Assignment of resonance types to observed resonances in the ^{1}H-^{13}C HMQC spectrum collected on a methyl protonated {Ile-[δ_1], Met-[ϵ], Leu-[$^{13}CH_3$,$^{12}CD_3$], Val-[$^{13}CH_3$,$^{12}CD_3$]} U-[^{15}N,^{13}C,^{2}H] sample of Hsp16.5 (**A i)**. The spectrum with Ile and Met resonance types indicated, readily determined from the values of ^{13}C- and ^{1}H-resonant frequencies (BMRB) and confirmed by independent experiments (**B i)**. Discrimination between Leu (red) and Val (blue) resonances determined from Val-[$^{13}C\gamma H_3$,$^{12}C\gamma D_3$] U-[^{2}H] labelled sample (TRA) (**C i)**. **ii)** Positions of methyl carbons (Ile, Leu- Cδ_1, Val-Cγ_1, Met-Cε) are indicated as spheres on a surface representation of the 24-mer of Hsp16.5 (PDB ID: 1SHS)[37]. Colours of the spheres match the methyl types as given in **i)**. Only the Cδ1/Cγ1 atom of Leu/Val residues is displayed. The spectra were recorded at 333.15 K, on a 600MHz field strength (see Methods).

Table 7: Values of the methyl resonances according to the Biological Magnetic Resonance Data Bank (BMRB)[36]

Methyl type	Atom name	Average shift (ppm)	Standard deviation (ppm)
LEU	Cδ1	24.7	2.08
LEU	Cδ2	24.11	2.17
LEU	Hδ1	0.75	0.34
LEU	Hδ2	0.72	0.37
VAL	Hγ1	0.82	0.34
VAL	Hγ2	0.8	0.45
VAL	Cγ1	21.55	2.52
VAL	Cγ2	21.34	2.67
MET	Cε	17.27	4.62
MET	Hε	1.73	1.85
ILE	Cδ1	13.51	3.69

In MAGMA's benchmark, methyl-methyl NOE restraints from two methyl groups of the same Leu/Val residue were joined to a single residue label. This significantly reduced the size of data graphs and allowed for faster completion of the MCES/subgraph isomorphism search. In order to apply the same approach to Hsp16.5, a double labelled {Leu-[$^{13}C\delta_1H_3$, $^{13}C\delta_2H_3$], Val-[$^{13}C\gamma_1H_3$, $^{13}C\gamma_2H_3$}, [U-^{2}H] sample of the full-length protein was prepared. A short-mixing time 3D CCH NOESY experiment was collected on this sample, resulting in the measurement of strong intra-residue NOE cross-peaks for 26 out of 32 Leu/Val methyls (results from TRA, Methods, Figure 4.9). The same experiment was previously successfully applied to MSG to confirm and extend the assignments of Leu methyls[39].

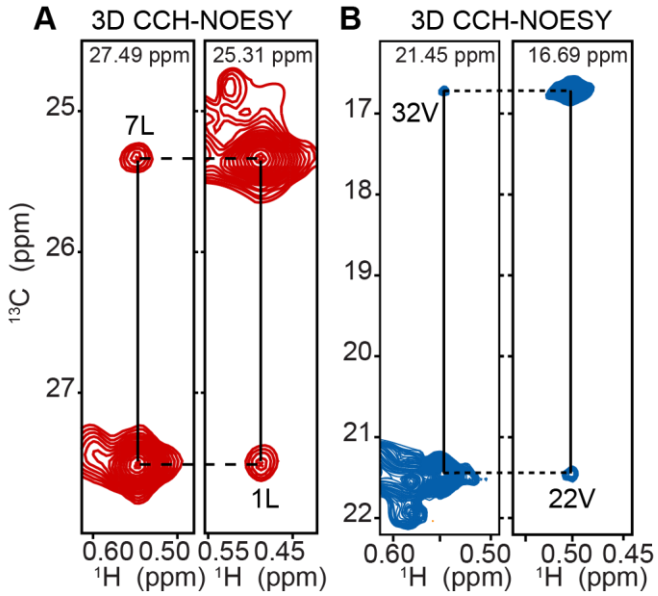


Figure 4.9 | Merging of pro-*R* and pro-*S* methyl group resonances to a single Val/Leu residue label. The merging was enabled by the short mixing time (65 ms) 3D CCH-NOESY experiment performed on a double labelled Leu-[^{13}Cδ_1H$_3$, ^{13}Cδ_2H$_3$], Val-[^{13}Cγ1H$_3$, ^{13}Cγ2H$_3$]. **A)** An example of two Leu methyl resonances (labelled 7L and 1L) from two methyl groups belonging to the same residue. **B)** An example of two Val methyl resonances (labelled 32V and 22V) from two methyl groups belonging to the same residue.

All of the collected restraints were given as input to MAGMA, together with the structure of Hsp16.5[37]. The spherical 24-mer of octahedral symmetry was used as an input for the calculation, to account for the inter-methyl contacts at the interfaces, and provide consistency with the NMR measurement in which the NOEs between monomeric subunits can be measured. The protein construct for which the structure has been obtained shares the primary sequence with the NMR construct; however the structure lacks electron density for the first 32 residues. This eliminated the possibility of assignment for Met1, Met14, Met23 and Met28, Ile29, and Leu10. It also presented a challenge for MAGMA, as the missing N-terminal methyl groups in the structure could have contributed inter-methyl NOEs and hence edges in the data graph. The extra methyl resonances from the N-terminus, as well as the extra resonances in the Leu/Val region (see above), if included in the set of data graph vertices, would present a problem for the requirement of the MCES algorithm of MAGMA, which is N (data graph vertices) ≤ N (structure graph vertices)[40]. To account for this, a set of additional "dummy" vertices, of the appropriate amino acid types, were added to the structure graph to account for the missing residues. The vertices were given no connectivity to any other vertices and hence serve a purpose of "assignment place-holders" for those data graph vertices for which inter-methyl NOE restraints will not be explained by MAGMA's

assignment (note that MAGMA excludes from the calculation any data vertices that have no connections to any other data vertices; Chapter 2).

The characteristics of the data and the structure graph of Hsp16.5 defined by MAGMA are listed in Table 8.

Table 8: Size and other characteristics of the data and structure graphs of Hsp16.5.

Nodes (NOE graph)	31
Edges (NOE graph)	49
Nodes (structure graph)	35
Edges (structure graph)	168
Data graph	
Restraint density	0.29
Average degree	3.16
Median degree	3
N (data subgraphs)	3

The calculation was run at 10 Å, a distance threshold established as optimal on MAGMA's benchmark (Chapter 3). Optimality of the distance threshold for Hsp16.5 was confirmed by running short MAGMA calculations at a range of distances (5-12 Å; Figure A4.1). At 10 Å, ~90% of the measured inter-methyl NOEs could be explained. The assumption taken here is that such a large fraction of the NOEs explained at 10 Å is sufficient to accurately define inter-methyl connectivity patterns and relate them to those in the structure graph. At this distance threshold, the NOE density (Methods) was 0.29, making the Hsp16.5 NOESY data more sparse than any dataset in MAGMA's benchmark, where the most sparse dataset was EIN1 (0.37), and also sparser than Hsp90 (0.42). The high sparsity of the NOE restraints reflected in the high ambiguity of methyl assignments calculated by MAGMA. Confident assignment could be obtained only for two methyl resonances, which were assigned to the methionine residues 43 and 87 (Figure 4.10 A and C). This result suggests that the inter-methyl NOEs on their own are not sufficient to obtain a significant fraction of unambiguous methyl assignments for this dataset (Table A4). Nonetheless, the generated ambiguous

assignments could still prove valuable, as they could be used for targeted site-directed mutagenesis. For instance, residues Ile-54, Ile-144, and Ile-146 are the only assignment possibilities for the methyl resonances labelled I9 and I4. However, the same residues are also plausible options for a number of other Ile type resonances with highly ambiguous assignment. Targeting these residues by site-directed mutagenesis would likely lead to a more significant overall reduction in the assignment ambiguity compare to, for instance, targeting of Ile-79, which is a plausible assignment option for 13 different methyl resonances. The same reasoning would lead to a preferential choice of Leu-70, Leu-129, and/or Leu-133 (Table A4).

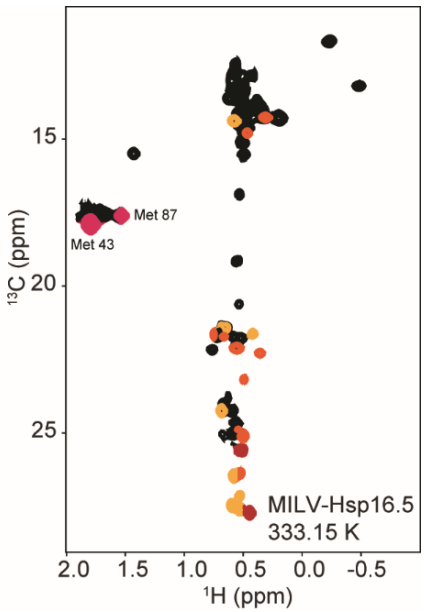


Figure 4.10 | Resonances in the [^{1}H,^{13}C]-HMQC spectrum of Hsp16.5 coloured according to their assignment ambiguity. In magenta, methyl resonances assigned to the residues Met 87, and Met 43 (i.e. confident, single option assignment). Methyl resonances with two, three, and four residue assignment options are shown in dark red, orange, and yellow, respectively. All other resonances (black) have more than four possible residue assignments.

Given the MAGMA result, an alternative approach to the assignment of Hsp16.5 methyl resonances will have to be undertaken in the future. One approach could be the assignment of resonances in HBHA(CO)NH[41], H(CCO)NH[42], CC(CO)NH[42] and HCCH-TOCSY[43] spectra, which were already collected on a truncated Hsp16.5 construct (S31-K141). This truncated construct does not oligomerize and therefore yields high-quality 3D NMR spectra. Even if only a subset of the methyl resonances could be assigned from these spectra, they could already provide a good starting point for a more "restricted" MAGMA calculation. In the

calculation, the known methyl assignments would be “fixed”, which is expected to significantly reduce assignment ambiguity for the rest of the resonances.

4.3 Conclusions and Outlook

HSP90α is a stress-induced cytosolic form of human HSP90 that is up-regulated in different forms of cancer[5]. The N-domain of HSP90α contains an important ATP binding site that has been a target for numerous competitive inhibitors developed in efforts for developing anti-cancer therapeutics[6,13]. NMR spectroscopy has been employed to study the structural and dynamical effects of ligand binding to the N-domain of HSP90α, and methyl resonance assignments have been published[30]. Here, I analyzed sparse methyl-methyl NOE restraints collected on the N-domain and used MAGMA with two different input structures. After combining the assignment results from both structures, MAGMA confidently assigns 38% of the input methyl resonances which is consistent given the data density and comparison to MAGMA’s benchmark, and the accuracy of the assignments was confirmed by cross-validation[29,30]. On average 4.7 NOEs were required per confident assignment, which was slightly higher compared to the average of the benchmark (3.6, Chapter 3). Merging the intra-residue LV NOEs to a single data graph vertex again proved beneficial (see Chapter 2), as separating the two LV methyls led to only 12% of the resonances assigned. If methyl-methyl NOESY data of sufficient quality could be obtained for full-length HSP90, it would be of interest to test MAGMA on such a dataset, as so far only partial assignments for Ile resonances have been obtained[17,18].

A comparison of the assignment results for the two different input structures of the N-domain confirmed the usefulness of the strategy outlined in Chapter 3, whereby merging the assignment results from multiple protein conformations is required to ensure accuracy. Moreover, testing MAGMA on a total of five different crystal structures of the N-domain showed that three of those were consistent with the available methyl assignments.

It was of interest to use the assignment results from MAGMA to obtain structural models of the N-domain bound to a lead compound. The confident and ambiguous methyl assignments from MAGMA were combined with intermolecular (methyl-ligand) NOE restraints, and molecular docking with HADDOCK resulted in a range of bound complexes. Assessment of the quality of the solutions was not straightforward due to two plausible ligand orientations in the bound form, and two different protein conformations (Figures 4.3 and 4.7). In general, the quality of the generated models was lower than models that were previously published with unambiguous methyl and amide resonance assignments[30], with ~2 Å, and ~0.5 Å higher RMSD for the worst and the best case, respectively.

Interestingly in a number of comparisons between MAGMA's assignment cross-validations and the HADDOCK results, one of the crystal structures consistently performed better than the alternatives, which supported its consistency with the solution-state measurements (HSP90_N2, Figures 4.5 and 4.6). Both the large dispersion in ligand-ligand RMSDs versus the HADDOCK scores and a lack of clear clustering indicated potential future difficulties in using this approach on blind cases. To circumvent the problem, MAGMA with intermolecular NOEs could be used in the future to help evaluate which HADDOCK models are consistent with both the methyl-methyl and the methyl-ligand NOEs. To define the optimal docking protocol using MAGMA's assignments in the future, a larger benchmark needs to be assembled that features a range of different ligand-protein complexes for which the ligand binding mode is unambiguous.

It was shown that incorporating ligand-protein NOEs, previously measured between a ligand and the methyl groups of the N-domain[30], can allow for exclusion of certain ligand binding modes. However, MAGMA found consistency with the known ligand and protein methyl assignments for structures that contain two different ligand orientations. Further investigations will thus be required to distinguish between the two possibilities. In all tested

cases, including intermolecular NOEs increased the number of confidently assigned methyl resonances, demonstrating the potential of these, and similar restraints (e.g. PREs), for use in MAGMA. For the future calculations that include ligand-protein NOE restraints, collecting a series of NOESY experiments with different mixing times could enable a more precise determination of the appropriate distance threshold.

Recently, a protocol for the calculation of the ligand binding site structure was introduced that operates on a set of semi-quantitative protein-ligand NOEs and a known protein structure (or a model) without any prior knowledge of resonance assignment[44]. The approach (NMR^2) is computationally demanding owing to a large ambiguity in attributing distance restraints from the assigned resonances of the ligand to a set of ambiguous resonance assignments of the protein. MAGMA could help reduce the ambiguity in the methyl assignments for NMR^2, and thus reduce the number of sub calculations that need to be performed[44].

Hsp16.5

The presented methyl-TROSY study of Hsp16.5 provides a concise summary of experimental efforts that need to be taken to collect an optimal set of restraints for MAGMA calculation. With the complete set of restraints, MAGMA could provide only ambiguous assignment for most of methyl resonances, probably due to the lower restraint density when compared to the cases presented in the benchmark. The methyl resonances will therefore be assigned on a smaller protein construct using through-bond NMR data that were already collected for the purpose. The presence of four extra L/V resonances in the [^{1}H,^{13}C]-HMQC spectrum of full-length Hsp16.5, potentially resulting from a conformational exchange on a slow NMR timescale, requires further experimental investigation.

4.4 Methods

4.4.1 HSP90α

NMR spectroscopy

3D [H]CCH-NOESY data were collected on a methyl protonated {Ile[^{13}Cδ1], Leu[^{13}CδH$_3$,^{12}CδD$_3$], V[^{13}CγH$_3$,^{12}CγD$_3$]} U-[^{15}N,^{13}C,^{2}H] sample of the N-domain of Hsp90α (9-233)[30]. The experiment was collected on 800 MHz Varian spectrometer with a 100 ms NOESY mixing time. The intermolecular ligand-methyl NOE restraints were documented previously (Supplementary Table 1. in publication from Shah *et al.*[30]).

Assignment of NOE restraints, methyl residue-types, and intra-residue methyl resonances of Leu/Val

NOE assignment was performed manually in Sparky[45] and was restricted to the set of methyl resonances for which the assignment was previously obtained[30]. The restraint density is defined as the number of reciprocated inter-methyl NOEs (data graph edges) divided by the number of the Euclidean distances between the methyl carbon atoms, within the distance threshold of MAGMA's calculation (structure graph edges). Information about two intra-residue methyl resonances (pro-*R* and pro-*S*), and of the methyl resonance amino-acid types (Ile, Leu, Val) was extracted from the available methyl resonance assignments[30].

Incorporating intermolecular NOEs in MAGMA

An addition to the regular MAGMA protocol was made to incorporate the intermolecular (protein-ligand) restraints to the data and structure graphs. Proton resonances of the ligand, which contributed intermolecular NOEs to methyl resonances of the protein, were added to the set of data graph vertices and were all given the same label, which was distinct from the labels of other data graph vertices. The intermolecular NOEs were incorporated in the data graph as edges between the specified methyl vertices (^{1}H-^{13}C resonances) and ligand vertices (^{1}H resonances). The proton atoms of the ligand were added to the set of structure graph

vertices and labeled with their atom identifiers. Euclidean distances were measured between the methyl carbons and the specified ligand atoms, and those smaller or equal to the inter-molecular (ligand-methyl) distance threshold (9 Å) were used to define the intermolecular edges. These edges were added to the set of structure graph edges and treated in the calculation in the same manner as the inter-methyl edges.

MAGMA calculations

MAGMA was run according to its usual protocol (see Chapter 2, 2.4) at a distance threshold of 10 Å, with modifications introduced only for the calculations that included the intermolecular NOEs/distances, as described above. In all cases, the number of iterations for re-sorting of the matching priorities was set to 50 (*priority_iterations*). The re-sorting was performed automatically by the algorithm on the basis of the three MCES (*n_priority_mces*) that were generated for each re-sorting iteration. This was followed automatically by a complete MCES search as described in Chapter 2.

Ligand-protein docking

The docking of an aminotriazine ligand and the N-domain of HSP90α was performed on the basis of the sparse inter-molecular NOE restraints, reported previously[30], restricted to the methyl-resonances of the protein and the ligand groups H1, H2, and H3 (Figure 4.7 A). Calculations were run using the HADDOCK2.2 server[46] with a guru interface to the HADDOCK docking program[35], and default parameters for the protein/peptide/ligand docking protocol[47] (performed by MGC). Two different protein conformers were used in docking (Figures 4.3 A, 4.5, 4.6, section 4.2.3).

In the rigid-body stage (it0) 1000 structures were generated, the top scoring 200 progressed to the torsion angle refinement stage (it1), and were finally refined in water. The list of ambiguous distance restraints is given in Appendix in the CNS format. The NOE data were collected with a long mixing time (~400ms), potentially introducing spin diffusion

contributions to the distance restraints. This was taken into account by the defining the upper and the lower distance margin as 2 Å and 1 Å, respectively. The distance restraints from residues assigned ambiguously by MAGMA were defined as ambiguous (i.e. their distance defined as $\sum_i 1/r_i^6$, where i iterates over all residue assignment options), and were given larger lower margins (1.5 Å). In addition, some restraint sets included protein residues defined as active (Table 3). Half of the distance restraints from the active residues are randomly discarded in each docking iteration. Clustering was performed on the water-refined models, on the basis of the RMSD (see below) using a 2 Å cut-off. The 4 best scored models from each cluster were analysed (Tables 4, 5, Figures 4.5, 4.6).

RMSD calculation

The ligand/ligand RMSD was calculated between the heavy atoms of the ligand in each HADDOCK model and the ligand in each of the reference structures (HSP90_N2 or 1YER_1) upon superposition of the protein backbones of all protein-ligand complexes (i.e., all HADDOCK models, 1YER_1, and HSP90_N2.pdb).

4.4.2 Hsp16.5

NMR spectroscopy, assignment of methyl resonance amino acid types, methyl-methyl NOEs, and two methyl resonances of each Leu/Val

The preparation of the protein samples of Hsp16.5 and the NMR spectroscopy measurements were performed by Andrew J Baldwin (AJB) and T. Reid Alderson (TRA) according to the previously established protocols[48].

The NMR experiments were performed on 800 MHz and 600 MHz Varian spectrometers equipped with room-temperature pulsed-field gradient triple resonance probes. 2D ^{1}H-^{13}C HMQC[49] and 4D ^{13}C-HMQC-[^{1}H,^{1}H]-NOESY-^{13}C-HMQC[49,50] spectra were collected on a methyl protonated {Ile-[^{13}Cδ1], Leu-[^{13}CδH$_3$,^{12}CδD$_3$], Val-[^{13}CγH$_3$,^{12}CγD$_3$], Met-[^{13}CϵH$_3$]} U-[^{2}H] labelled Hsp16.5 sample. The 4D ^{13}C-HMQC-[^{1}H,^{1}H]-NOESY-^{13}C-HMQC was

collected following a non-uniform sampling schedule on 800 MHz, at 343.15K (70°C) with the mixing time of 250 ms (AJB). A [^{13}C-*F*1, ^{13}C-*F*2]-edited NOESY experiment [39] with a 65 ms mixing time was recorded on a methyl protonated double labelled {Leu-[^{13}CδH$_3$, ^{13}CδH$_3$], Val-[^{3}CγH$_3$, ^{3}CγH$_3$},U-[2H] sample in D_2O at 333.15K (60°C) (TRA). This experiment allowed for relating the two methyls of Leu/Val residues to a single residue label (single vertex in data graph). In addition, the experiment allowed double checking of methyl amino acid types, confirming the accuracy of Ile and Met resonance type assignments given on the basis of ^{1}H and ^{13}C chemical shifts (Figure 4.8). An additional 2D ^{1}H-^{13}C HMQC[49] spectrum was collected on an exclusively Val-[^{13}CγH$_3$,^{12}CγD$_3$] U-[^{2}H] labelled sample, which was achieved by addition of unlabelled sodium-α-ketoisocaproate to the minima medium in which bacteria were grown, together with sodium-(^{13}CH3)-2-ketoisovalerate, according to the protocol of Lichtenecker *et al*.[31] (TRA). This experiment allowed for discrimination between Leu and Val methyl resonances, completing the set of restraints needed for MAGMA. Assignment on the inter- and intra-methyl NOE restraints was performed manually in Sparky[45] (TRA and myself).

MAGMA calculations

MAGMA was run according to its usual protocol with a distance threshold of 10 Å, and 50 iterations for re-sorting of the matching priorities (*priority_iterations*) on the basis of three MCES generated in each re-sorting iteration (*n_priority_mces*) (see Chapter 2, 2.4). From the deposited PDB file of Hsp16.5 (PDB ID:1SHS)[37], the 24-mer was reconstructed using the *Symexp* command in PyMol[51] within a 3 Å cut-off about the structure of an octamer.

4.5 References

1 Li, J., Buchner, J. Structure, function and regulation of the hsp90 machinery. *Biomed J* **36**, 106-117, doi:10.4103/2319-4170.113230 (2013).

2 Borkovich, K. A., Farrelly, F. W., Finkelstein, D. B., Taulien, J. & Lindquist, S. hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. *Mol Cell Biol* **9**, 3919-3930 (1989).

3 Pearl, L. H., Prodromou, C. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annu Rev Biochem* **75**, 271-294, doi:10.1146/annurev.biochem.75.103004.142738 (2006).

4 Ali, M. M., Roe, S.M., Vaughan, C.K., Meyer, P., Panaretou, B., Piper, P.W., Prodromou, C., Pearl L.H. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature* **440**, 1013-1017, doi:10.1038/nature04716 (2006).

5 Zhang, H., Zhou, C., Chen, W., Xu, Y., Shi, Y., Wen, Y., Zhang, N. A dynamic view of ATP-coupled functioning cycle of Hsp90 N-terminal domain. *Sci Rep* **5**, 9542, doi:10.1038/srep09542 (2015).

6 Pearl, L. H. Review: The HSP90 molecular chaperone-an enigmatic ATPase. *Biopolymers* **105**, 594-607, doi:10.1002/bip.22835 (2016).

7 Prodromou, C., Roe, S. M., O'Brien, R., Ladbury, J. E., Piper, P. W., Pearl, L. H. Identification and structural characterization of the ATP/ADP-binding site in the Hsp90 molecular chaperone. *Cell* **90**, 65-75 (1997).

8 Harris, S. F., Shiau, A. K. & Agard, D. A. The crystal structure of the carboxy-terminal dimerization domain of htpG, the Escherichia coli Hsp90, reveals a potential substrate binding site. *Structure* **12**, 1087-1097, doi:10.1016/j.str.2004.03.020 (2004).

9 Wang, X., Song, X., Zhuo, W., Fu, Y., Shi, H., Liang, Y., Tong, M., Chang, G., Luo, Y. The regulatory mechanism of Hsp90alpha secretion and its function in tumor malignancy. *Proc Natl Acad Sci U S A* **106**, 21288-21293, doi:10.1073/pnas.0908151106 (2009).

10 Trepel, J., Mollapour, M., Giaccone, G. & Neckers, L. Targeting the dynamic HSP90 complex in cancer. *Nat Rev Cancer* **10**, 537-549, doi:10.1038/nrc2887 (2010).

11 Stebbins, C. E., Russo, A. A., Schneider, C., Rosen, N., Hartl, F. U., Pavletich, N. P. Crystal structure of an Hsp90-geldanamycin complex: targeting of a protein chaperone by an antitumor agent. *Cell* **89**, 239-250 (1997).

12 Roe, S. M., Prodromou, C., O'Brien, R., Ladbury, J. E., Piper, P. W., Pearl, L. H. Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J Med Chem* **42**, 260-266, doi:10.1021/jm980403y (1999).

13 Tatokoro, M., Koga, F., Yoshida, S. & Kihara, K. Heat shock protein 90 targeting therapy: state of the art and future perspective. *EXCLI J* **14**, 48-58, doi:10.17179/excli2015-586 (2015).

14 Miura, T., Fukami, T. A., Hasegawa, K., Ono, N., Suda, A., Shindo, H., Yoon, D. O., Kim, S. J., Na, Y. J., Aoki, Y., Shimma, N., Tsukuda, T., Shiratori, Y. Lead generation of heat shock protein 90 inhibitors by a combination of fragment-based approach, virtual screening, and structure-based drug design. *Bioorg Med Chem Lett* **21**, 5778-5783, doi:10.1016/j.bmcl.2011.08.001 (2011).

15 Li, J., Sun, L., Xu, C., Yu, F., Zhou, H., He, J. Crystal Structures of N-terminal Domain of Human Hsp90 With ATP Analogues Reveal The Functional Regulation of Hsp90. *Prog Biochem Biophys* **39**, 995-1002, doi:10.3724/Sp.J.1206.2011.00611 (2012).

16 Didenko, T., Duarte, A. M., Karagoz, G. E. & Rudiger, S. G. Hsp90 structure and function studied by NMR spectroscopy. *Biochim Biophys Acta* **1823**, 636-647, doi:10.1016/j.bbamcr.2011.11.009 (2012).
17 Park, S. J., Kostic, M. & Dyson, H. J. Dynamic Interaction of Hsp90 with Its Client Protein p53. *J Mol Biol* **411**, 158-173, doi:10.1016/j.jmb.2011.05.030 (2011).
18 Karagoz, G. E., Duarte, A. M., Ippel, H., Uetrecht, C., Sinnige, T., van Rosemalen, M., Hausmann, J., Heck, A. J., Boelens, R., Rudiger, S. G. N-terminal domain of human Hsp90 triggers binding to the cochaperone p23. *Proc Natl Acad Sci U S A* **108**, 580-585, doi:10.1073/pnas.1011867108 (2011).
19 Karagoz, G. E., Duarte, A. M., Akoury, E., Ippel, H., Biernat, J., Moran Luengo, T., Radli, M., Didenko, T., Nordhues, B. A., Veprintsev, D. B., Dickey, C. A., Mandelkow, E., Zweckstetter, M., Boelens, R., Madl, T., Rudiger, S. G. Hsp90-Tau complex reveals molecular basis for specificity in chaperone action. *Cell* **156**, 963-974, doi:10.1016/j.cell.2014.01.037 (2014).
20 Haslbeck, M. & Vierling, E. A first line of stress defense: small heat shock proteins and their function in protein homeostasis. *J Mol Biol* **427**, 1537-1548 (2015).
21 Hilton, G. R., Lioe, H., Stengel, F., Baldwin, A. J. & Benesch, J. L. P. in *Molecular Chaperones* (ed Sophie Jackson) 69-98 (Springer Berlin Heidelberg, 2013).
22 Kim, K. K., Kim, R. & Kim, S.-H. Crystal structure of a small heat-shock protein. *Nature* **394**, 595-599 (1998).
23 Aquilina, J. A., Benesch, J. L., Bateman, O. A., Slingsby, C. & Robinson, C. V. Polydispersity of a mammalian chaperone: mass spectrometry reveals the population of oligomers in αB-crystallin. *Proceedings of the National Academy of Sciences* **100**, 10611-10616 (2003).
24 Baldwin, A. J., Hilton, G. R., Lioe, H., Bagneris, C., Benesch, J. L., Kay, L. E. Quaternary dynamics of αB-crystallin as a direct consequence of localised tertiary fluctuations in the C-terminus. *J Mol Biol* **413**, 310-320 (2011).
25 Baldwin, A. J., Lioe, H., Hilton, G. R., Baker, L. A., Rubinstein, J. L., Kay, L. E., Benesch, J. L. The polydispersity of αB-crystallin is rationalized by an interconverting polyhedral architecture. *Structure* **19**, 1855-1863 (2011).
26 Kim, R., Lai, L., Lee, H. H., Cheong, G. W., Kim, K. K., Wu, Z., Yokota, H., Marqusee, S., Kim, S. H. On the mechanism of chaperone activity of the small heat-shock protein of Methanococcus jannaschii. *Proceedings of the National Academy of Sciences* **100**, 8151-8155 (2003).
27 Haslbeck, M., Kastenmüller, A., Buchner, J., Weinkauf, S. & Braun, N. Structural dynamics of archaeal small heat shock proteins. *J Mol Biol* **378**, 362-374 (2008).
28 Shi, J., Koteiche, H. A., McDonald, E. T., Fox, T. L., Stewart, P. L., McHaourab, H. S. Cryoelectron microscopy analysis of small heat shock protein 16.5 (Hsp16. 5) complexes with T4 lysozyme reveals the structural basis of multimode binding. *Journal of Biological Chemistry* **288**, 4819-4830 (2013).
29 Salek, R. M., Williams, M. A., Prodromou, C., Pearl, L. H. & Ladbury, J. E. Backbone resonance assignments of the 25kD N-terminal ATPase domain from the Hsp90 chaperone. *J Biomol NMR* **23**, 327-328 (2002).
30 Shah, D. M., AB, E., Diercks, T., Hass, M. A., van Nuland, N. A., Siegal, G. Rapid Protein-Ligand Costructures from Sparse NOE Data. *J Med Chem* **55**, 10786-10790, doi:10.1021/jm301396d (2012).
31 Lichtenecker, R. J., Weinhaupl, K., Reuther, L., Schorghuber, J., Schmid, W., Konrat, R. Independent valine and leucine isotope labeling in Escherichia coli protein overexpression systems. *J Biomol NMR* **57**, 205-209, doi:10.1007/s10858-013-9786-y (2013).

32 van Dijk, A. D., Boelens, R. & Bonvin, A. M. Data-driven docking for the study of biomolecular complexes. *FEBS J* **272**, 293-312, doi:10.1111/j.1742-4658.2004.04473.x (2005).
33 Sousa, S. F., Ribeiro, A. J., Coimbra, J. T., Neves, R. P., Martins, S. A., Moorthy, N. S., Fernandes, P. A., Ramos, M. J. Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. *Curr Med Chem* **20**, 2296-2314 (2013).
34 de Vries, S. J., van Dijk, M. & Bonvin, A. M. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883-897, doi:10.1038/nprot.2010.32 (2010).
35 Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society* **125**, 1731-1737, doi:10.1021/ja026939x (2003).
36 Biological Magnetic Resonance Bank, B.M.R.D. <http://www.bmrb.wisc.edu/ref_info/statful.htm>
37 Kim, K. K., Kim, R., Kim, S. H. Crystal structure of a small heat-shock protein. *Nature* **394**, 595-599 (1998).
38 Palmer, A. G., 3rd. Chemical exchange in biomacromolecules: past, present, and future. *J Magn Reson* **241**, 3-17, doi:10.1016/j.jmr.2014.01.008 (2014).
39 Tugarinov, V. & Kay, L. E. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *Journal of the American Chemical Society* **125**, 13868-13878, doi:10.1021/ja030345s (2003).
40 Mcgregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software Pract Exper* **12**, 23-34, doi:DOI 10.1002/spe.4380120103 (1982).
41 Grzesiek, S. & Bax, A. Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J Biomol NMR* **3**, 185-204 (1993).
42 Grzesiek, S., Anglister, J. & Bax, A. Correlation of Backbone Amide and Aliphatic Side-Chain Resonances in 13C/15N-Enriched Proteins by Isotropic Mixing of 13C Magnetization. *Journal of Magnetic Resonance, Series B* **101**, 114-119, doi:http://dx.doi.org/10.1006/jmrb.1993.1019 (1993).
43 Bax, A., Clore, G. M. & Gronenborn, A. M. 1H□1H correlation via isotropic mixing of 13C magnetization, a new three-dimensional approach for assigning 1H and 13C spectra of 13C-enriched proteins. *Journal of Magnetic Resonance (1969)* **88**, 425-431, doi:http://dx.doi.org/10.1016/0022-2364(90)90202-K (1990).
44 Orts, J., Walti, M. A., Marsh, M., Vera, L., Gossert, A. D., Guntert, P., Riek, R. NMR-Based Determination of the 3D Structure of the Ligand-Protein Interaction Site without Protein Resonance Assignment. *J Am Chem Soc* **138**, 4393-4400, doi:10.1021/jacs.5b12391 (2016).
45 Goddard, T. D. K., D. G. SPARKY 3. *University of California, San Francisco*.
46 van Zundert, G. C. P., Rodrigues, J. P., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S., van Dijk, M., de Vries, S. J., Bonvin, A. M. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720-725, doi:10.1016/j.jmb.2015.09.014 (2016).
47 Interface, H. G. *http://haddock.science.uu.nl/services/HADDOCK2.2/haddockserver-guru.html*.
48 Kim, K. K., Yokota, H., Santoso, S., Lerner, D., Kim, R., Kim, S. H. Purification, crystallization, and preliminary X-ray crystallographic data analysis of small heat shock protein homolog from Methanococcus jannaschii, a hyperthermophile. *J Struct Biol* **121**, 76-80, doi:10.1006/jsbi.1998.3969 (1998).

49 Tugarinov, V., Hwang, P. M., Ollerenshaw, J. E. & Kay, L. E. Cross-correlated relaxation enhanced H-1-C-13 NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *Journal of the American Chemical Society* **125**, 10420-10428, doi:10.1021/ja030153x (2003).

50 Tugarinov, V., Kay, L. E., Ibraghimov, I. & Orekhov, V. Y. High-resolution four-dimensional H-1-C-13 NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. *Journal of the American Chemical Society* **127**, 2767-2775, doi:10.1021/ja04032o (2005).

51 The PyMOL Molecular Graphics System, V. S., LLC.

Appendix

Table A1: MAGMA's assignment result for the N-domain of HSP90α. In the first column, methyl resonance ID is given by a number and a single letter amino acid identifier. The number corresponds to the known methyl resonance assignment. In the second column assignments with the residue numbers and their amino acid type are listed. For all methyls, apart from 128I, correct residue assignment was within the assignment set determined by MAGMA.

Methyl	Residue												
214I	214I												
78I	78I												
48L	48L												
218I	218I												
56L	56L												
91I	91I												
222V	222V												
188L	188L												
198L	198L												
220L	220L												
207V	207V												
49I	49I												
59I	59I												
29L	29L												
103L	103L												
143L	143L												
107L	107L												
80L	45L	80L											
45L	45L	80L											
89L	190L	89L											
190L	190L	89L											
70L	70L	76L											
76L	70L	76L											
32L	32L	122L											
122L	122L	32L											
144V	144V	186V											
203I	206I	203I											
206I	206I	203I	81I										
186V	144V	150V	148V	186V									
148V	144V	150V	148V	186V									
*128I	206I	203I	33I	131I									
136V	136V	150V	148V	186V	92V								
150V	172V	144V	150V	148V	186V								
172V	144V	172V	150V	148V	186V	92V							
92V	144V	172V	136V	92V	150V	148V	186V						
131I	206I	26I	49I	81I	131I	203I	187I	33I					
104I	206I	151I	49I	81I	218I	43I	203I	187I	104I				
43I	206I	131I	81I	43I	218I	203I	187I	33I	128I				
96I	81I	151I	206I	26I	43I	34I	110I	33I	187I	128I	96I	131I	
81I	206I	187I	151I	81I	43I	26I	34I	33I	203I	104I	110I	131I	
151I	206I	187I	151I	81I	43I	26I	34I	33I	203I	96I	110I	128I	131I
187I	206I	187I	151I	81I	43I	26I	34I	203I	33I	96I	110I	128I	131I

Table A2: Intersect and symmetric difference of the confident assignment sets for the structures HSP90_N1 and HSP90_N2. Assignments in the symmetric difference that are incorrect given known methyl assignments[30] are marked in magenta.

Intersect		Symmetric difference	
214I	214I	206I	203I
78I	78I	80L	80L
48L	48L	186V	144V
218I	218I	89L	190L
56L	56L	144V	186V
91I	91I	203I	206I
222V	222V	190L	89L
188L	188L	45L	45L
198L	198L	148V	148V
220L	220L	150V	150V
207V	207V	70L	70L
49I	49I	76L	76L
59I	59I		
29L	29L		
103L	103L		
143L	143L		
107L	107L		

A 4.1 The edge differences of two different structure graphs of the N-domain of HSP90α

Distances of the edges which differ between the two structures (HSP90_N1, HSP90_N2) show negligible difference (< 0.5 Å), in all cases apart from Leu107-Val186 edge with Δ (edge distance) = 3.57 Å (Table A3). Nonetheless, the small difference was sufficient to push the edges over the 10 Å threshold of the calculation, excluding them from one of the structure graphs.

We extended the threshold for the calculation to 10.5 Å to determine if the assignment differences between the two structures will prevail, even when the edges in question become common to the two structure graphs. The results indeed featured no confident assignment differences and the assignments from both structures were correct. However, only 20% and 8% of the resonances entering the calculation were assigned for the HSP90_N2 and HSP90_N1 structure, respectively.

This result shows how a significant increase in assignment ambiguity can arise from a small increase in the distance threshold used in the calculation (+0.5 Å). Even though this test was performed, it should be noted that the calculation at 10.5 Å resulted in the same fraction of explained NOEs as the calculation at 10 Å (0.985). Since the fraction of explained NOEs did not increase, from a general point of view, there is no justification for the increase in the distance threshold. Moreover, results of the calculations at 10.5 Å did not increase the number of confident assignments, when compared to the combined result of 10 Å calculations (Figure 4.4 C).

Taken together, when two structure forms of the protein are available, and the predominant form in solution is unknown, recommend protocol includes: 1. running the calculation independently for all available structure forms, at an optimal distance threshold; 2. pulling the assignment solutions together. In this example, as well as for several other proteins in

MAGMA's benchmark, such approach maximizes the accuracy of the confident assignments (Chapter 3). It also removes the need for inspection of differences of the protein structure graphs, as, at least in this example, this does not give a straightforward route for prediction of the assignment differences (Table A3).

Table A3: Edge differences of the two structure graphs defined based on the structures HSP90_N1 and HSP90_N2 (Figure 4.3). In grey, edges incident with the residues that were confidently assigned for both structures. HSP90_N2\HSP90_N1 is the relative complement of HSP90_N1 in HSP90_N2, comprising edges that are in HSP90_N1 set, but not in HSP90_N2 set, and *vice versa*. In red, edges incident with the residues that were assigned confidently for HSP90_N1, but ambiguously for HSP90_N2 (Leu89, Val144, Val186, Leu190, Ile203). The distances of the edges in red and their difference in the two structures are displayed.

Edge differences				
HSP90_N2\ HSP90_N1	HSP90_N1\ HSP90_N2			
ILE110-LEU107	ILE110-LEU29			
ILE131-ILE26	ILE110-ILE26			
ILE151-VAL172	ILE110-VAL136			
LEU143-LEU32	ILE151-ILE81			
ILE26-ILE33	LEU107-VAL148			
ILE43-LEU48	LEU107-VAL172			
ILE81-LEU220	ILE151-VAL92	**Distance(Å)**		
ILE218-LEU220	ILE151-VAL148	HSP90_N1	HSP90_N2	Δ distance(Å)
LEU188-LEU198	LEU107-VAL186	11.37	7.80	3.57
LEU56-LEU70	ILE203-VAL222	10.01	9.95	0.06
LEU56-VAL92	LEU103-VAL186	10.22	9.82	0.40
LEU89-VAL207	LEU107-LEU48	9.77	10.05	0.28
LEU190-VAL148	ILE131-LEU48	9.64	10.08	0.43
LEU198-VAL144	ILE78-LEU76	9.89	10.08	0.19

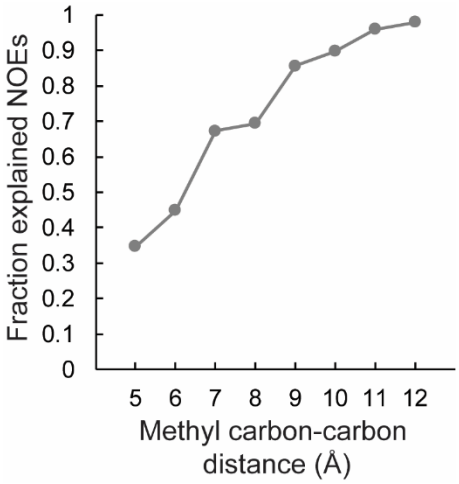


Figure A4.1 | Fraction of explained inter-methyl NOEs over a range of distance thresholds tested for MAGMA calculation on Hsp16.5. The complete MAGMA calculation was run at 10 Å, the threshold found to be optimal on MAGMA's benchmark.

Table A4: MAGMA's assignment result for Hsp16.5. In the first column, a random methyl resonance ID is given by an arbitrary number and a single letter amino acid identifier. In the second column, residue assignments are given by the residue numbers and their amino acid type. Note that every L/V peak ID in the first column corresponds to two "merged" methyl resonances from the [^{1}H,^{13}C]-HMQC spectrum of the protein.

Peak ID	Resid.																		
5M	43M																		
6M	87M																		
1L	77L	129L																	
6L	129L	70L	133L																
9I	54I	144I	146I																
18V	56V	131V	63V																
4I	144I	54I	146I																
11L	86L	60L	111L																
23V	115V	56V	131V																
3L	70L	133L	129L	77L															
7I	37I	47I	146I	57I															
22V	131V	115V	56V	63V															
4L	133L	111L	70L	60L															
2L	111L	133L	77L	70L															
16I	57I	146I	132I	54I	47I														
20V	128V	131V	63V	115V	56V	73V													
10L	60L	86L	133L	77L	129L	111L	70L												
8I	109I	45I	79I	47I	68I	146I	57I	37I											
1I	45I	79I	109I	132I	57I	68I	37I	54I	47I										
6I	68I	45I	109I	79I	146I	57I	54I	47I	132I										
10I	79I	99I	47I	45I	105I	109I	146I	37I	68I	57I									
3I	47I	99I	146I	37I	48I	95I	35I	69I	57I	94I	132I	88I	54I						
2I	37I	68I	47I	132I	99I	109I	140I	146I	54I	144I	105I	79I	48I	45I	57I				
20I	37I	88I	68I	47I	35I	94I	99I	109I	146I	54I	144I	95I	105I	79I	48I	45I	57I		
17I	37I	88I	68I	47I	35I	94I	99I	109I	146I	54I	144I	95I	105I	79I	48I	45I	57I		
13I	37I	88I	68I	47I	35I	94I	99I	109I	146I	54I	144I	95I	105I	79I	48I	45I	57I		
11I	37I	88I	68I	47I	35I	94I	99I	109I	146I	54I	144I	95I	105I	79I	48I	45I	57I		
5I	99I	47I	79I	45I	105I	109I	146I	37I	35I	95I	48I	69I	68I	57I	88I	94I	132I	54I	
12I	105I	146I	94I	48I	35I	95I	47I	99I	79I	109I	69I	45I	37I	68I	57I	88I	132I	54I	
15I	35I	94I	140I	146I	95I	69I	48I	132I	88I	105I	47I	99I	79I	109I	45I	37I	68I	57I	54I
19I	37I	88I	132I	47I	35I	94I	68I	99I	109I	140I	146I	54I	144I	95I	105I	69I	79I	48I	45I	57I

A 4.2 Filtering assignment options after the calculation

The high ambiguity in some of the resonance assignments obtained from MAGMA is in part due to the treatment of a disconnected data graph, i.e. a separate calculation run for each connected data subgraph. As such, different data subgraphs can get assigned to intersecting sets of methyl-residues (vertices) from the structure graphs, thereby artificially increasing assignment ambiguity. To circumvent this problem, and reduce assignment ambiguity, a suggestion was made by MCG to filter the final assignment sets obtained by MAGMA for the N-domain of HSP90α. Filtering was performed such that confidently assigned residues were removed as assignment options from all other methyl resonances. Furthermore, those pairs of residues that were assigned exclusively to a pair of methyls were also removed from the assignment options of other methyl resonances. Consider for instance an assignment result where $M_1 = \{R_1, R_2\}$, $M_2 = \{R_1, R_2\}$, where M_1 and M_2, denote methyls 1 and 2, and R_1 and R_2 denote residues 1 and 2, respectively. For M_1 and M_2 to each have a unique residue assignment, the residues R_1 and R_2 have to be excluded as assignment options for any other methyl resonances. In Figure A4.2 (below) the resulting reduction in the assignment ambiguity of methyls of the N-domain of HSP90α is shown.

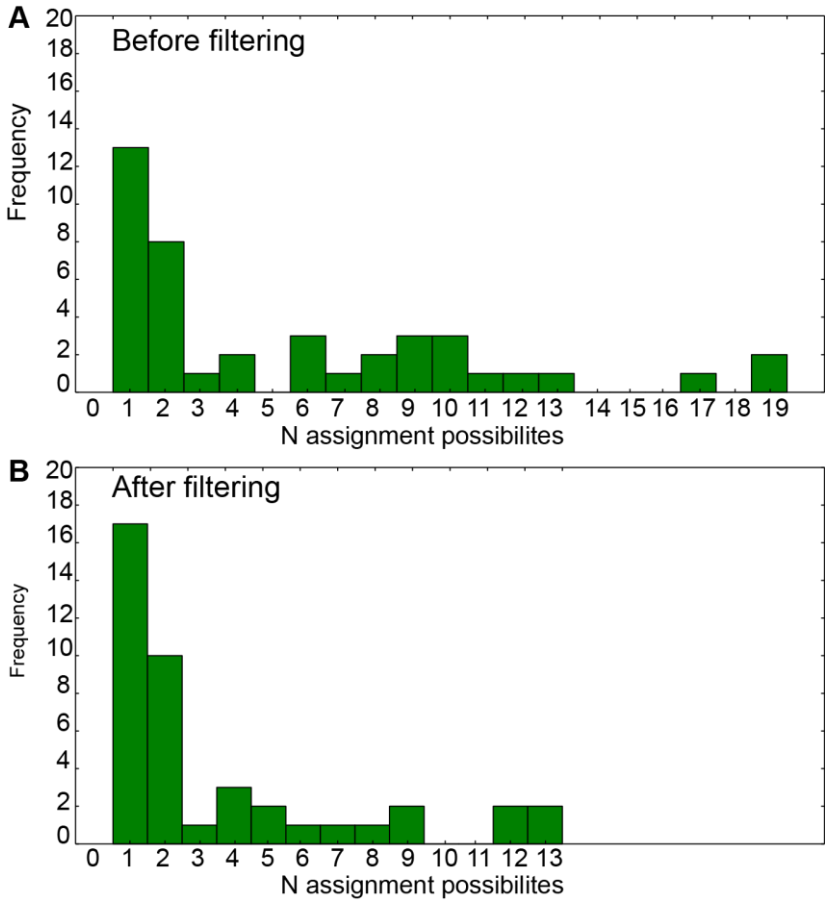


Figure A4.2 | Reducing assignment ambiguity by filtering of the mutually exclusive assignment options for the N-domain of HSP90α. Filtering was performed on the combined assignment results from the input structures HSP90_N1 and HSP90_N2 (Figure 4.2 B, D ii)) Frequency (*y*-axis) of the ambiguous assignments (*x*-axis) before (**A)** and after filtering (**B).**

Concluding remarks

In this thesis a novel, structure-based method for the automatic assignment of methyl resonances of high molecular mass proteins is presented. Defining the methyl resonance assignment as a graph comparison problem allowed the usage of the theoretical framework from the field of structural pattern recognition, wherein graph-subgraph isomorphism and maximum common subgraph have been studied for decades, resulting in a variety of exact and inexact algorithmic solutions.

To calculate a set of all possible methyl resonance assignments, MAGMA relies on the comparison of sparse, NOE-derived connections between isotopically labelled methyl groups and the inter-methyl connectivity extracted from a high-resolution structure of a protein. When dealing with sparse data, as commonly encountered in integrative modelling efforts in structural biology, including as many as possible complementary sources of information typically generates more optimal solutions. The future of MAGMA will include the incorporation of additional restraints, such as PREs, and elaborate on the theoretical framework, particularly regarding the treatment of the input structure. Furthermore, the refinement and expansion of the algorithmic approaches to achieve optimal speed and scalability constitute important future improvements. It is my hope that the presented MAGMA protocol and its further development will extend the number of methyl-TROSY applications on large biomolecular systems.

In principle, the approach outlined in this thesis could be straightforwardly and generically applied to structure-based NMR resonance assignment, including backbone resonances, as considered before by the NOEnet approach. Given the computational complexity of graph comparison problems, and a substantial increase in size of the graphs that would be generated in such applications involving backbone resonances, more consideration might need to be given to approximate graph comparison approaches. Outside of the scope of NMR resonance

assignment automation, the presented approach is more broadly applicable to any problem that requires structural comparison, provided that the objects of such comparisons can be cast to a graph structure and the constraints of the "graph matching" can be suitably defined.

Finally, MAGMA exemplifies the importance of considering theoretical concepts and methodological developments from different areas of science and engineering when encountering related problems. For decades, graph theory-based approaches have found their applications in chemistry, structural biology, systems biology, evolutionary biology, and ecology, where they are used to, for example, cluster chemical structures, elucidate allosteric communication pathways in proteins, analyse the co-expression of genes, study the regulatory mechanisms of cellular signal transduction networks, or understand the evolution of a population. Progress in the theoretical fields of computational complexity and algorithmics, as well as practical improvements in computing speed combined with a greater cross-talk between different scientific disciplines, will pave the way toward the continuous usefulness of graph-based approaches in diverse areas of chemistry and biology in the years to come.