UNIVERSITE DE SCIENCES ET TECHNOLOGIES DE LILLE – LILLE 1

ECOLE DOCTORALE BIOLOGIE – SANTE

Doctorat

Discipline: Aspects moléculaires et cellulaires de la biologie

Dries VERDEGEM

Probing the edge of protein (non)-structuration with NMR. A case study of the intrinsically disordered proteins human Tau and HCV NS5A.

Thèse dirigée par Dr. Guy LIPPENS

Soutenue le 18 décembre 2009

Jury:

Dr. Jean-Claude MICHALSKI	Président
Prof. Dr. José MARTINS	Rapporteur
Dr. François PENIN	Rapporteur
Dr. Nico VAN NULAND	Examinateur
Prof. Dr. Bruno KIEFFER	Examinateur
Dr. Guy LIPPENS	Directeur de thèse

UNIVERSITE DE SCIENCES ET TECHNOLOGIES DE LILLE – LILLE 1

ECOLE DOCTORALE BIOLOGIE – SANTE

Doctorat

Discipline: Aspects moléculaires et cellulaires de la biologie

Dries VERDEGEM

Probing the edge of protein (non)-structuration with NMR. A case study of the intrinsically disordered proteins human Tau and HCV NS5A.

Thèse dirigée par Dr. Guy LIPPENS

Soutenue le 18 décembre 2009

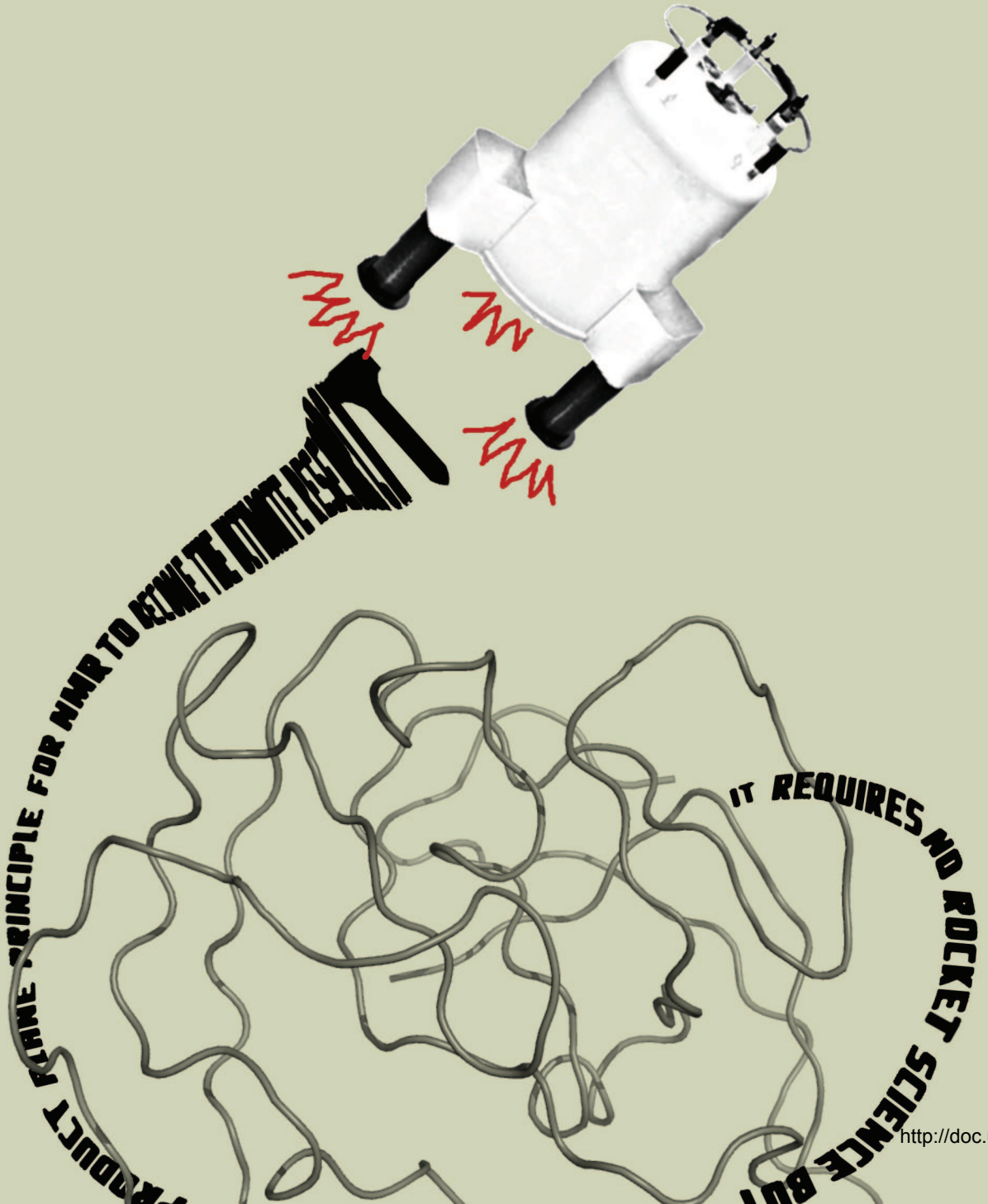
Jury:

Dr. Jean-Claude MICHALSKI	Président
Prof. Dr. José MARTINS	Rapporteur
Dr. François PENIN	Rapporteur
Dr. Nico VAN NULAND	Examinateur
Prof. Dr. Bruno KIEFFER	Examinateur
Dr. Guy LIPPENS	Directeur de thèse

Probing the edge of protein (non)-structuration with NMR.

A case study of the intrinsically disordered proteins human Tau and HCV NS5A.

Dries VERDEGEM



Abstract

Many proteins and protein regions have been shown to lack rigid 3D structure under physiological conditions *in vitro*. Instead, they exist as dynamic ensembles of interconverting structures and were therefore also described as intrinsically unstructured/disordered. Despite this lack of apparent consistent structure, these proteins were found to have diverse and important functions *in vivo*. In fact, their disordered aspect is even a crucial part of the functional state of intrinsically unstructured proteins (IUPs).

Structural biology on these IUPs should be performed in an attempt to elucidate details on their function. X-ray crystallography, usually a powerful method for structure elucidations, is in this case of limited use, since disordered proteins would not be expected to crystallize. Even in the situation where crystals do form (for proteins with both ordered and disordered regions), regions of disorder generally vary in location from one molecule to the next and therefore fail to scatter X-rays coherently. The lack of coherent scattering leads to missing electron density.

The technique of choice for studying IUPs is Nuclear Magnetic Resonance (NMR). Several NMR experiments can be performed that may all contribute to insights in different aspects of the structural and dynamical behaviour of this protein class. One of the great advantages of NMR over other techniques is that it provides information with atomic resolution. However, in order to deduce this information from the recorded NMR spectra, the different NMR signals have to be identified (i.e. linked to individual nuclei) first. This happens during a so-called assignment process of the protein, which is these days done using 3D triple resonance spectra. By matching peaks between complementary spectra, that contain information on either the individual (i) residues or the neighbouring (i-1) residues, sequential connectivities can be made and ultimately all resonances can be assigned.

Unstructured proteins have the unfortunate characteristic of causing very crowded NMR spectra, and this overlap genuinely complicates the assignment process. Therefore, one of the first issues that got our attention was the search for and implementation of an efficient computer algorithm to facilitate this task. This resulted in a graphical semi-automatic assignment tool that uses the concept of products and sums of spectral planes to make the sequential connectivities and to generate residue specific NMR spectra. The principle of product and sum planes also proved useful for a second developed application that provides NMR assignments for structured proteins based on chemical shift predictions.

Once their NMR spectra are assigned, individual IUPs can be further studied. Two unstructured proteins have been examined during my thesis (although one in greater detail than the other). Full-length Tau is a 441 residue-long IUP that regulates the polymerisation of tubulin into microtubules in human neurons. The exact mechanism of this (reversible) polymerisation process is unknown to date. Moreover, the Tau protein

is also implicated in the progression of Alzheimer's disease (AD), as it is observed to aggregate into intracellular filamentous inclusions. Also the details of this latter process are largely unknown. During my PhD work, I have completely assigned the NMR backbone and C_β resonances of two Tau fragments (F3 and F5) and more than 65% of those resonances of full-length Tau P301L. The P301L mutation in Tau has been shown to greatly expedite the development of dementia. The numerous NMR assignments of Tau could eventually lead to more insight in the structural behaviour of the protein in those different situations.

Studied in more detail was the Hepatitis C virus (HCV) non-structural protein 5A (NS5A). Although this IUP is known to be involved in HCV replication and particle assembly, its exact function(s) remains to be elucidated. Upon invasion in a human host, NS5A is anchored to the endoplasmic reticulum membrane via an amphipathic N-terminal α-helix. Its cytoplasmic part consists of three domains: NS5A-D1, -D2 and -D3. We have assessed the structural properties of both the second and third domain and have shown that they are mainly unstructured. However, NS5A-D3 seems to contain some residual α-helical structure towards either ends of the sequence, which could be indicative of regions prone to interaction with other cellular partners. CsA, the well-known cyclophilin inhibitor has anti-HCV properties. Since mutations in HCV NS5A protein have been associated with CsA resistance, cyclophilins might play a role in HCV replication. We have therefore also examined the interaction between both CypA and CypB and domain 2 and 3 of the HCV NS5A protein. Our findings demonstrate that such interactions indeed exist and thus that replication defects of NS5A mutants might result from an altered interaction between the cyclophilin and NS5A. The observation that many prolines in the sequence of NS5A are substrate to the PPIase activity of the cyclophilins raises questions about the importance of the cis/trans peptidyl-prolyl isomerisation in this matter.

Résumé / French Abstract

De nombreuses protéines ou domaines de protéines manquent d'une structure 3D rigide sous conditions physiologiques *in vitro*. Ils existent plutôt comme des ensembles dynamiques de structures interconvertibles et ont en conséquence été aussi decrites comme intrinsèquement non structurés/désordonnés. Malgré ce manque apparent de structure, il a été démontré que ces protéines ont des fonctions diverses et importantes *in vivo*. En fait, leur aspect désordonné constitue une caractéristique cruciale pour l'état fonctionnel des protéines intrinsèquement non structurées (IUPs).

La biologie structurale de ces IUPs doit être effectuée pour essayer de comprendre les détails de leur fonction. La cristallographie par rayons X, méthode puissante habituellement utilisée pour l'élucidation structurale, est dans ce cas peu pertinente, car on ne s'attend pas à ce que des protéines désordonnées cristallisent. Même dans la situation où des cristaux se forment (dans le cas des protéines qui ont aussi bien des parties structurées que non structurées), les domaines désordonnés généralement montrent une hétérogénéité dans leurs positions, et en conséquence ne parviennent pas à diffracter les rayons X de façon cohérente. Ce manque de dispersion cohérente engendre une absence de densité électronique.

La technique de choix pour étudier des IUPs est la Résonance Magnétique Nucléaire (RMN). Plusieurs expériences de RMN sont possibles qui peuvent toutes contribuer à comprendre les différents aspects du comportement structural et dynamique de cette catégorie de protéines. Un des grands avantage de la RMN en comparaison avec d'autres techniques est qu' elle fournit de l'information à une résolution atomique. Cependant, pour déduire cette information des spectres RMN enregistrés, les différents signaux RMN doivent d'abord être identifiés (c'est à dire reliés à des noyaux individuels). Ce processus d'attribution se réalise actuellement en utilisant des spectres 3D triple résonance. En faisant la concordance des pics de spectres complémentaires, qui certains contiennent de l'information sur le résidu individuel (i), d'autres sur le résidu voisin (i-1), des connectivités séquentieles se font et par la suite toutes les résonances peuvent être attribuées.

Des protéines non structurées présentent malheureusement des caractéristiques qui provoquent des spectres RMN très encombrés, et ces superpositions compliquent vraiment l'attribution. Pour cette raison, une des première chose qui a attiré notre attention était la recherche et également l'implémentation d'un algorithme efficace pour simplifier cette tâche. Il en a résulté un outil d'attribution graphique, semi-automatique qui utilise le concept des produits et sommes des plans spectraux pour faire les connectivités séquentielles et pour générer des spectres résidu-spécifique. Le principe des plans produit et plans somme a aussi été utile pour une deuxième application développée qui fournit des attributions RMN pour des protéines structurées basée sur des prédictions de déplacement chimique.

Une fois que leurs spectres RMN sont attribués, des IUPs individuelles

peuvent être étudiées plus en détail. Deux protéines non structurées ont été examinées pendant ma thèse (bien que l'une de façon plus détaillée). Tau entier est une IUP de 441 residus qui règle la polymérisation de la tubuline en microtubule dans les neurones. Le mécanisme exact de ce processus de polymérisation (réversible) est inconnu à ce jour. De plus, la protéine Tau est aussi impliquée dans la progression de la maladie d'Alzheimer (AD), car on a observé son agrégation en filaments intracellulaires. Les details de ce dernier processus sont aussi pour le plupart inconnus. Pendant ma thèse, j'ai entièrement attribué les résonances RMN du squelette et des C_β de deux fragments de Tau (F3 et F5) et plus de 65% des résonances de Tau entier P301L. Cette mutation P301L accélère fortement la dévelopement de la démence. Les nombreuses attributions RMN de Tau pourraient sur le long terme mener à davantage de compréhension sur le comportement structural de cette protéine dans ces situations différentes.

La protéine non structurale 5A (NS5A) du virus de l'Hépatite C (VHC) a été étudiée plus en détail. Bien que cette IUP soit impliquée dans la réplication de VHC et dans l'assemblage de particules, sa/ses fonction(s) exacte(s) ne sont pas encore éclaircie(s). Après invasion dans l'hôte humain, NS5A est ancrée à la membrane du réticulum endoplasmique via une hélice α amphipathique N-terminale. Sa partie cytoplasmique consiste en trois domaines: NS5A-D1, -D2 and -D3. On a évalué les propriétés structurales du deuxième et troisième domaine et on a montré qu'ils sont principalement non structurés. Néanmoins, NS5A-D3 paraît contenir de la structure hélice α résiduelle vers les deux extrémités de la séquence, ce qui pourrait indiquer des régions prédisposées à interagir avec d'autres partenaires cellulaires. CsA, l'inhibiteur de la cyclophiline bien connu, a des propriété anti-VHC. Des mutations de la protéine NS5A VHC ont été associées avec de la résistance CsA, ce qui semble indiquer que des cyclophilines pourraient jouer un rôle dans la réplication de VHC. Par conséquent, on a également examiné l'interaction entre CypA et CypB d'une part, et les domaines 2 et 3 de NS5A VHC d'autre part. Nos résultats démontrent que telles interactions existent en effet et donc que les défauts de réplication des mutants de NS5A pourraient résulter d'une interaction modifiée entre la cyclophiline et NS5A. L'observation que plusieurs prolines de la séquence de NS5A sont substrats de l'activité PPIase des cyclophilines soulève des questions sur l'importance de la cis/trans peptidyl-prolyl isomérisation dans ce processus.

Preface

Structure Determines Functionality is a statement that can often be heard from biochemists, molecular biologists, or the like, when they talk about one of the basic building blocks of life on Earth: proteins. This principle, generally referred to as the structure-function paradigm, states that the amino acid sequence of a protein determines its 3D structure and that the function requires the prior formation of this 3D structure. This point of view grow from the early idea that proteins can be seen as rigid or semi-rigid "blocks", whose specificity and catalytic power are determined by the unique fit of a correct substrate onto the preformed and sturdy surface of the enzyme's active site [134]. This alleged truth dating from the early, infancy days of biochemistry has eventually been enlarged slightly by concepts such as induced fit [220], thereby encompassing some notions of flexibility. Although the paradigm often is correct, a parallel development of microscopic world physics showed it is based on a major simplification.

The correct relationship would be: the structure and functionality of a molecule are both determined by its electron density in a very integrated way. From the moment a protein gradually leaves the ribosome upon its formation, it searches for a thermodynamically stable conformation under the influence of electron density zones that are being drawn to or repelled from other zones of the same protein, surrounding water molecules or possibly interacting chaperone molecules. The residual structure that might thus arise, changes the global protein electron density face with it, allowing new interactions that on their turn evoke additional structuration events, and so on. This general folding mechanism, in the literature referred to as the Zipping and Assembly Hypothesis (ZA), is only recently (after more than fifty years of protein folding research) getting acceptance as the true recipe followed by proteins to obtain their native structure [96, and references therein]. The ZA mechanism provides a plausible answer to Levinthal's paradox[1] [234] because the protein never bothers to search vast stretches of conformational space. But more important in this discourse, it shows us to what degree structure is as much a consequence of the electron density as a cause of it. Ultimately, in a completely folded protein the functional electron density in and around the active site is held together by the rest of the protein structure.

It has taken so much time for scientists to uncover the basics of the protein folding problem because of the transient nature of it. The short living transition states and partially folded intermediates characteristic of a folding process are difficult to access experimentally. The results that experiments did yield allowed statistical surveys that elucidated some crude folding rules leading to simple models. However, it were computer simulations of purely

[1] How does a protein chain succeeds in obtaining its native folded state, which is one of zillions of possible conformations, in so little time?

physics-based methods that gave (and will continue to give in the future) the ability to understand how proteins fold in all-atom detail. Different manifestations of the electron density have been translated to concepts such as hydrophobicity, hydrogen bonds, van der Waals interactions, ... These concepts are used in semi empirical atomic physical force fields that are these days capable of correctly folding small, water-soluble proteins [293]. Whether, slightly enlarging our field of interest, these force fields are also good enough to (*a*) predict conformational changes, such as induced fit, important for computational drug discovery; (*b*) understand protein mechanisms of action, motions, folding processes, enzymatic catalysis, and other situations that require more than just the static native structure; and (*c*) understand how proteins respond to solvents, pH, salts, denaturants, and other factors is a completely different question, to which the answer is probably in the negative. Results of a higher level might be expected from actively using the ultimate molecular descriptor for which the electron density is taken.

It is all described by the quantum theory. In 1905 Albert Einstein discovered that light behaves as a stream of particle-like objects, he called photons [115]. This discovery, arguably the single most important of the 20^{th} century, exposed physicist for one of the first times to the harsh reality of the Alice in Wonderland universe we really live in. However, in order to explain the outcome of certain other experiments with light, the picture of light being an electromagnetic wave remained indispensable. The idea of the dual wave-particle nature of light slowly gained acceptance in the decades to follow. The quantum theory that emerged from this struggle to reconcile light and matter, showed that not only do waves behave as particles (light), but particles (electrons, atoms, molecules, ...) behave like waves as well. That's why in the 1920's Erwin Schrödinger devised an abstract mathematical wave, christened the wave function, to describe the behaviour of any microscopic particle [337]. The wave function is said to contain all possible information on the particle, and, when talking about electrons in atoms or molecules, there is a special link to reality formed by the square of the wave function; the electron density.

Applying this kind of physics to study structure and behaviour of molecules is subject of the field called quantum chemistry. I am a great fan of the mathematical beauty with which quantum chemistry allows to solve chemical and (as extreme offshoot) biochemical issues. The anti-Machiavellian edge of someone like Paul Dirac, who applied the filter of mathematical beauty, and not experimental verifiability throughout almost everything he worked on, are at the least admirable. Unfortunately, the computational cost of this methodology is so high, that we will probably have to wait for advent of commercially available quantum computers [371] before we shall be able to reap the fruits of this highly powerful structure- and function-elucidating method, especially in biological sciences. Another example illustrating the beauty of formulas is the biology related work of mathematical prodigy Erik Demaine. The man gave a public lecture recently (Gembloux, March 5th 2009) where he argued that the earlier discussed co-translational folding (protein-leaving-the-ribosome) issue can be addressed by applying the rules of linkage folding mathematics and demonstrates with it which chain folds are producible and which not (also described in [93]), thereby greatly improving our knowledge on protein folding space.

Anyhow, to come back to our story of structure and function, there is a second and more fundamental reason why the opening statement doesn't

hold an unconditional truth. The discovery of proteins that are wholly disordered or contain lengthy disordered segments, yet are functional, has wreaked havoc on the lock-and-key world view that demands highly organized proteins. Such disorganized proteins, especially prevalent in eukaryotes, have hitherto proven to be abundant, diverse, vital, dynamic and tightly controlled inside the cell. But on their functional and dynamical properties, and their functional repertoire, very little is known.

As for the search for a general theory of disorderedness, the description of the rules that define the conformational behaviour of the protein chain, this domain has been mostly stuck in the bioinformatics phase since many years. Although this has been useful for an initial shaping of the study of the intrinsically disordered proteins, bringing coherence to proteins that were previously viewed as outliers, by for example linking disorderedness to amino acid sequences, a thorough understanding will only become possible by going beyond bioinformatics towards simulation methods, just as the protein folding problem has needed to surpass the simple statistical surveys. And, seen the complexity of the behaviour of these proteins, it will be the quantum chemical approaches, or rather a synergy of these approaches with experimental ones, that will become of ever greater importance as the field develops in the future.

Until those heydays of scientific methodology arrive, our efforts should be concentrated on (*a*) the elaboration of new experimental methods and with them, (*b*) the concrete study of individual natively unfolded proteins in order to accumulate the experimental data, which can then be used for interpretation in computational studies. Only a divided concentration on both of these axes can lead to a full appreciation of this unique protein category. Vladimir Uversky aptly headed one of his reviews in the domain as "Natively unfolded proteins: A point where biology waits for physics" [397]. Written in 2002, the statement has lost only a small amount of its strength over the recent years. The demanding studies of unfolded proteins inside cells and in a crowded milieu *in vitro* might lead to a cure of many of the involved diseases and thus, more generally, to important insights on the functional and dynamical properties of these proteins.

This manuscript reports of two conterminous projects —each handling with one the described demands (*a* and *b*)— that have been tackled during my thesis. First, a powerful NMR assignment method, particularly well suited for the crowded spectra of unstructured proteins, was developed. This method is mainly based on the principle of products (and sums) of spectral planes. If we can believe the message conveyed by the rocket of the front cover picture of this book (which was taken from somewhere far out in the uncharted backwaters of our galaxy, where technology did not quite evolve as ours), this principle could prove an important trump for NMR in the future. Secondly, I took a closer look at two specific intrinsically disordered proteins, human Tau and the NS5A protein of the Hepatitis C virus. The text is divided in four chapters. The first chapter gives some general information on intrinsically disordered proteins, their structural state, their biological importance. The second chapter deals with the NMR of this protein class. In the third chapter the work on the mentioned individual proteins is described. Finally, the fourth chapter discusses a special NMR assignment method for proteins with higher degrees of stable structure.

This work has been influenced both directly and indirectly by many people. Of course, none of this would have happened without my thesis

advisor Guy Lippens. I thank him not only for his guidance and instruction, but also for his patience. His insightful criticism and suggestions on how to improve my work have impacted me far more than any document could possibly reveal. He learned me to see the beauty of having an own research problem and of solving it by proper means, even if this implies muddling through as best as we can. The liberty I was given was very refreshing, and the several excursions to congresses, summer schools and collaborators I was allowed to do, have enlarged my scientific view a great deal. Finally, I thank him for his constructive criticism during my thesis writing. This project would have been much less thorough without his help, and I am thankful for the high standards he has set.

The other (current and past) members of the laboratory (Laziza Amniai, Fanny Bonachera, Sébastien Conilleau, Anthony Daccache, Xavier Hanoulle, Dragos Horvath, Isabelle Huvent, Isabelle Landrieu, Arnaud Leroy, Gérard Montagne, Benjamin Parent, Nathalie Sibille, Alain Sillen, Caroline Smet and Jean-Michel Wieruszeski) have been very helpful as well. Thank you all for the nice conversations. You will be remembered!

I thank Gaëlle Vanstaevel, Michelle Autexier, Nadège Vereecke, Yvon Hu, Lucie Mylondo and Olivier Durreau for their administrative support. They have been of great help during several administrative procedures concerning my recruitment at the university and the many missions.

Although I have worked with Klaas Dijkstra (Groningen University) for only four days, he has had an enormous impact my entire PhD work. During these four days he has learned me to master his NMRpython library and gave me a copy of it. As will become clear in the text, this software package directly and indirectly influenced all of my occupations during the two final thesis years. Besides for these crucial lessons, I also thank him and his colleagues for the friendly welcome.

I also owe much gratitude to Tim Stevens and Wayne Boucher of the CCPN project (Cambridge University). I have met them on several occasions and despite their busy schedule and countless collaborations they were always interested in my research project and the implementation of our tools in their software suite.

Thank you Marie-Christine Slomianny for introducing me to Scilab.

But most importantly: I am so grateful for the friends I have made during the three years I lived and worked in Lille, a city which I've gone to love. In particular I thank Maud S.-A. and Gaëlle E. for their unceasing encouragement. I aspire to be as good a friend to others as they have been to me, and I sincerely hope we can continue to see each other in the future as much as possible. I thank my family and friends (at home) for their love and encouragement. Special thanks to my mother for 26 years of never desisting support.

Enjoy the read...

Lille, October 2009 Dries Verdegem

Contents

Chapter 1

Introduction: Intrinsically Unstructured Proteins

1.1. IUPs, a General Picture - Lessons from Bioinformatics Analysis

It could seem inappropriate to start this discussion with the information that more or less concludes our current knowledge in the field of intrinsically unfolded proteins. The lessons from bioinformatics form a rather unimpressive list, but provide the most general picture of this protein family. Prediction of a protein's predisposition to be intrinsically disordered has been a necessary prerequisite for the understanding of principles and mechanisms of (partial) protein folding and protein function. Future, more profound insights will be build upon the general picture known today, hence its early appearance as introduction to this text. It must also be noted that IUP in the title of this section stands for Intrinsically Unstructured Protein. Since the special term 'natively denatured' was introduced in 1994 to describe the behaviour of the Tau protein [341], many terms are found back in the literature to name this protein category, depending on the author writing about it. Besides intrinsically unstructured [443] and natively denatured, other names are intrinsically disordered [108], natively unfolded [426] and natively disordered [84] proteins. To my opinion (that is not shared by some authors), all of these terms cover the cargo equally well and, especially since none of them have really suffered from inconsistent usage in the biological literature, they will all appear throughout this text.

1.1.1. Discovery of IUPs

More than thirty years ago, at a time when only about twenty protein crystal structures had been determined, some protein segments were discovered to yield no discernable electron density and yet to be essential for function [32, 33]. Missing electron density in protein structures can arise from failure to solve the phase problem, from crystal defects [189], or even from unintentional proteolytic removal during protein purification. However, a common reason for missing electron density is that the unobserved atom, side chain, residue, or region fails to scatter X-rays coherently due to variation in position from one protein to the next, i.e., the unobserved atoms are disordered. Since then, partly due to the developing technique of NMR, that is more certain in its characterisation of disorder than X-ray diffraction, many more protein containing functional, yet disordered regions have been discovered and studied. The discovery rate for such proteins has been increasing continually and has become especially rapid during the last decade [110]. The discovery and characterisation of these proteins is becoming one of the fastest growing areas of protein science.

1.1.2. What Characterises the AA-Sequence of Natively Unfolded Proteins?

Very quick, it became clear that the absence of regular structure in natively unfolded proteins was implemented in their amino-acid sequences. Some of the sequence peculiarities that were recognised over the years include the presence of numerous uncompensated charged groups, i.e. a large net charge at neutral pH, arising from the extreme pI values in such proteins [180, 153, 426], and a low content of hydrophobic amino-acid residues [180, 153]. Thus, they lack all the necessary information (high mean hydrophobicity and relatively low net charge) for the protein or protein part to fold to a compact conformation under physiological conditions. The results of a survey in this matter are presented in fig. 1.1. It shows that natively unfolded proteins are specifically localised within a unique region of the charge-hydrophobicity phase space. The solid line in this figure represents the border between intrinsically unstructured and native proteins. This observation allows the estimation of the 'boundary' main hydrophobicity value $\langle H \rangle_b$ below which a polypeptide chain with a given mean net charge $\langle R \rangle$ will be most probably unfolded:

$$\langle H \rangle_b = \frac{\langle R \rangle + 1.151}{2.785} \tag{1.1}$$

The validity of predictions based on this formula has been successfully shown for several proteins [94]. This means that the degree of compaction of a given polypeptide chain is determined by the balance in the competition between the charge repulsion driving unfolding and the hydrophobic interactions driving folding.

In terms of amino acid types, it was later shown that the majority of the intrinsically disordered proteins are substantially depleted in I, L, V, W, F, Y, C, and N and enriched in E, K, R, G, Q, S, P and A [108], which accounts for the observed behaviour. More recent observations [300] have revealed that differences in amino acid compositions exist between short and long unstructured regions (with thirty amino acids being the threshold commonly applied). Short disordered regions are more depleted in I,V and L, while long disordered regions are more depleted in G and N. In addition, short disordered regions are more enriched in G and D, while long disordered regions are more enriched in K, E and P.

The situation was complicated by yet another discovery. Upon comparison of amino acid compositions of experimentally characterised regions of protein disorder with regions of order, but with high B-factors[1] (that are typically found in catalytic or binding sites), it was found that both have very similar enrichings and depletions of the typical amino acids mentioned above [328, 315]. This similarity was most striking between high-B-factor ordered and short disordered protein segments. Despite the fact that B-factors tend to be influenced by experimental conditions and crystal packing [174, 224], they do show correlation with the amino acid sequence, indicating an implemented drive towards flexibility. These found similarities are logical in a sense in that both kind of protein segments could be associated with large thermal vibrations of individual atoms and with high intramolecular flexibility, but render a common understanding of intrinsically unstructured proteins more difficult.

[1] A measure of of residue flexibility of folded proteins as obtained from X-ray crystallography

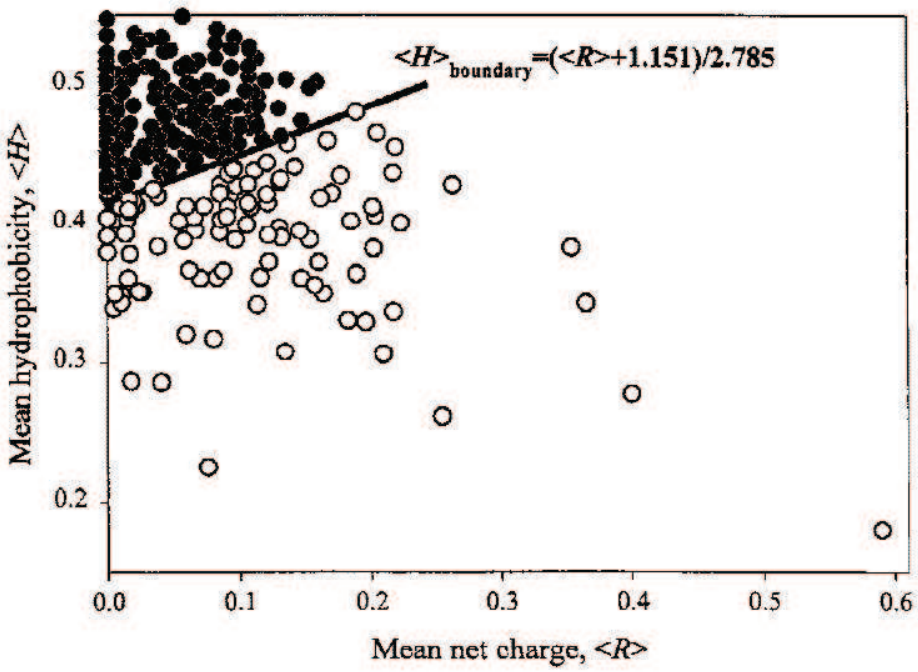


Figure 1.1. The unique property of natively unfolded protein sequences is a combination of low overall hydrophobicity and large net charge. Comparison of the mean hydrophobicity and the mean net charge (at pH 7.0) for a set of 275 folded (black circles) and 102 natively unfolded proteins (open circles). The solid line represents the border between intrinsically unstructured and native proteins calculated using equation 1.1. Figure taken from [397].

Another way of describing the established compositional bias of natively unfolded proteins is by the concept of low-complexity. Indeed, it is expected to find that IUPs have relatively low complex sequences if they are made up preferentially of certain amino acid types. Although this is not a general rule (some limited amino acid sets occur in a specific sequence pattern that results in folding into coiled-coil, super-helical structures [259]), local low-complexity can be used to define and detect unfolded protein sequences. Several statistical local compositional complexity measures have been developed for this purpose [331, 440, 332], and for the sake of clarity one of them is described briefly here. Complexity is a function of the compositional state of a sequence segment or window, as can be represented by a complexity state vector. For example, the numbers (3,3,2,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0), representing in decreasing order counts for the various amino acids, describe one of the 77 possible complexity states of a 12-residue peptide window. Many possible sequences and amino acid compositions, with different residue types corresponding to the 20 numbers, share this complexity state. A local compositional complexity K of a window of length L can thus be defined as:

$$K = \frac{1}{L} log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right) \tag{1.2}$$

where the n_i are the 20 numbers in the complexity state vector. The logarithm is taken to base N to place K in approximately the range 0 to 1. K measures the information needed per position, given the window's composition, to specify a particular residue order. Note that if all 12 residues

in the window were of the same type (which would correspond to the complexity state vector (12,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)), K is 0, while 12 different residue types (1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0) give a K of 0.556. An example of the complexity profile calculated this way of residues 400-500 of the human RING3 protein is given in fig. 1.2.

Figure 1.2. The complexity profile of residues 400-500 of the human RING3 protein calculated using equation 1.2 with $L = 12$. Normalised K-values are given on the vertical axis, while the horizontal axis represents the position in the sequence. Unfortunately, an ordinate axis is not provided by the original authors. So, although it is clear that the maximum value corresponds to one, no relative value can be assigned to valley positions. Nevertheless, in the middle of the sequence stretch, a low-complexity (probably unstructured) region is clearly visible. Figure taken from [440].

Proteins or protein domains being called "proline rich" or "glycine rich" are indeed well-known examples of disordered polypeptide chains exhibiting low complexity and hence enrichment of other types of disorder promoting residues. More than anything else, all these observations combine to strongly indicate that intrinsic protein disorder is generally encoded by the amino acid sequence in very complex and subtle ways.

A final observation pushing this conclusion in the same direction is the typically faster rate of evolution [44], and the distinctive amino acid substitution patterns during evolution [314] of such disordered proteins and protein regions. However, it is unclear whether the high frequency of amino acid substitutions observed for some IUPs is correlated with functional variation. In addition, it is not know whether IUPs in the same functional families will adopt similar dynamic structures. In the absence of this basic information, it is impossible to predict the dynamic structure and molecular function of IUPs based on sequence data. The evolution of intrinsically disordered protein structure seems to depend on selection for other properties. Natively disordered proteins with extended disorder do not form globular structures and therefore do not have a requirement to maintain long range interactions such as those required by globular proteins. This creates a potential for these sequences to accumulate more variation and generate more sequence divergence than globular proteins. On the other hand, some proteins that were indisputably classified as IUP exhibit rather low evolutionary variation (e.g. Tau [313]). Moreover, mutations that do occur are often associated with protein dysfunction and corresponding diseases.

1.1.3. A Series of IUP Predictors

In an attempt to understand the relationship between sequence and disorder more thoroughly, a number of computer programs were created for the prediction of unstructured regions from amino acid sequences. Since the

first predictor for disorder was published [429], more than 50 them have been developed. A non-exhaustive list is: PONDR [326, 236, 328, 301, 289, 300], SEG [439], Disopred and Disopred2 [206, 423, 424, 45], Globplot [240], DisEMBL [239], NORSp [246], FoldIndex [309], Charge/hydropathy method [400], HCA (Hydrophobic Cluster Analysis) [60], PreLink [75], IUPred [100, 101], RONN [449], DRIPPRED [260], FoldUnfold [150, 148, 149], DISpro [68], DisPSSMP and DisPSSMP2 [373, 372], Spritz [415], PrDOS [196], etc. Since CASP5, disorder prediction has also been included in the Critical Assessment of Structure (CASP) Meetings [222], which has played a very positive role in the number and quality of IDP predictors that was developed since. Some of the predictors rely on global physicochemical parameters, such as the charge-hydrophobicity property described above. They are the so called binary disorder predictors that give their verdict (ordered/disordered) for the entire protein. Most of them however are based on artificial neuronal networks (ANNs), support vector machines (SVMs) or logistic regression and look for (*a*) biased amino acid compositions, (*b*) sequence complexity, (c) parameters such as hydropathy, net charge, etc., (*d*) a flexibility index, (*e*) just complete sequence stretch resemblances, and so on. The reference datasets are sometimes previously existing ones, others are build up entirely by the authors. These more sophisticated predictors evaluate intrinsic disorder on a per-residue basis. An exhaustive summary of computational methods for the recognition of unstructured regions is beyond the scope of this text, but some review articles have recently addressed this subject in detail [132, 39, 102, 179]. Moreover, links to many of these predictors can be found in the Disordered Protein Database (http://www.DisProt.org) [411, 358]. Most of the published predictors are similar in the prediction of long disordered regions, but they do differ significantly in the local details of the output. As in other areas of bioinformatics, the available predictors are complementary and several methods based on different concepts should be used to achieve quality predictions. In the light of this observation, very recently so called metapredictors, that combine the outputs of several individual predictors, were developed. Tools in this genre, as there are metaPdDOS [50], MeDor [238] and MetaDisorder [335], indeed seem to give improved prediction accuracies.

What is interesting from the biologist's point of view is that predictions of protein disorder can be used to guide laboratory experiments (as has been done in a countless number of papers and also in chapter 3 of this text), which are in turn leading to the discovery and characterisation of increasing number of unstructured proteins.

1.1.4. *In Silico* Unfoldome Analysis

Whereas classic biochemical methods, of discovering a detectable activity and isolating it by purifying the protein, mainly lead to the acquaintance of ever more structured proteins, disorder prediction programs as described above, showed that the idea of the hegemonic domination of structured proteins in cells is a misconception. Since we now have access to a vast libraries of gene sequences, studies can be based on these data as well. In contrast to protein structure databases such as the PDB [27], sequence databases such as Swiss-Prot [14] and PIR [18], but mainly the genome databases (e.g. NCBI [261]) should provide a much better way of estimating the commonness of intrinsic disorder. The PDB is strongly biased against intrinsic disorder,

because of the many X-ray resolved structures and sequence databases have their own biases as well. Analysis of theoretical translations of sequence data for complete genomes has indicated that intrinsically disordered proteins are indeed highly prevalent [327], and that the proportion of proteins that contain such segments increases with the increasing complexity of an organism [109, 424]. Because of this high prevalence, the ensemble of unfolded proteins in an organism has even been given a special name: the unfoldome [78] and the IDP-ome [146].

A first study performed in 2000 [109] on a limited amount of genomes demonstrated that eukaryotes exhibit the greatest amount of disordered as measured by segment lengths (L) $\geq$ 50 as compared to archaea or bacteria. That is, the eukaryotes yielded 24-41% of chains with predicted disorder of L $\geq$ 50 compared to 24% for the highest bacterium. A similar study has also included viral protein information [84]. A comparison between the PDB-derived dataset containing proteins with missing coordinates (PDB_S25) and the Swiss-Prot databases, provided a convenient means to study disorder across the different kingdoms. The proteins were divided in two classes: (*a*) proteins with no or only short disorder and (*b*) proteins with substantial regions of disorder. The percentage of proteins in the set for these mostly ordered proteins were 19% for eukaryotes, 40% for bacteria, 48% for archaea, and 16% for viruses. The number of proteins in the set for the ones likely to contain substantial amounts of disorder were 32% for eukaryotes, 6.5% for bacteria, 2.3% for archaea, and 48% for viruses. These data suggest that both eukaryotes and viruses are most likely to have proteins with large regions of disorder.

A more profound investigation was done by Ward and co-workers [424]. Their use of Disopred2 indicated, with a rather small estimated error, that an average of 2.0% of archaean, 4.2% of eubacteria and 33.0% of eukaryotic proteins contain long regions of disorder. Moreover, the authors have also carried out an analysis of the functions associated with predicted disorder by using the SGD (*Saccharomyces* genome database [111]) that links GO terms (gene ontology annotations[2] [76]) to a large set of several thousands of proteins of budding yeast *Saccharomyces cerevisiae*. The results from this analysis show that many of the functions associated with disordered regions are involved in processes such as the organisation and biogenesis of the cytoskeleton, but mainly involve molecular signalling and regulation [114, 107]. Many of the protein functions associated with disorder (see Fig. 1.3), such as DNA- and cytoskeleton-binding, have in fact previously and since been indicated by numerous experiments. Binding to DNA to facilitate processes such as transcription, transposition, packaging, repair and replication is found to be a particularly important function of IUPs.

There have been some explanations for the lower occurrence of disorder in prokaryotes. For example, the absence of cell compartments reduces the ability of prokaryotic cells to physically protect unfolded structures from degradation. Indeed, the majority of putative disorder-containing proteins are located in cellular components that provide some protection from proteolysis such as the cell cortex and nucleus, as was also indicated by the GO term analysis. Part of the reason for the existence of IUPs might also be

[2] Annotations used to provide descriptions to proteins in a more or less general way. The terms "cell growth and maintenance", "cell proliferation" and "cell cycle" are three examples of GO's belonging to a hierarchical structure as a more specific description is given in going from the first to the third annotation.

assigned to the fact they allow for more complexity, by integrating many binding partners, post translational modification possibilities, ..., which is useful in the more complex prokaryotes.

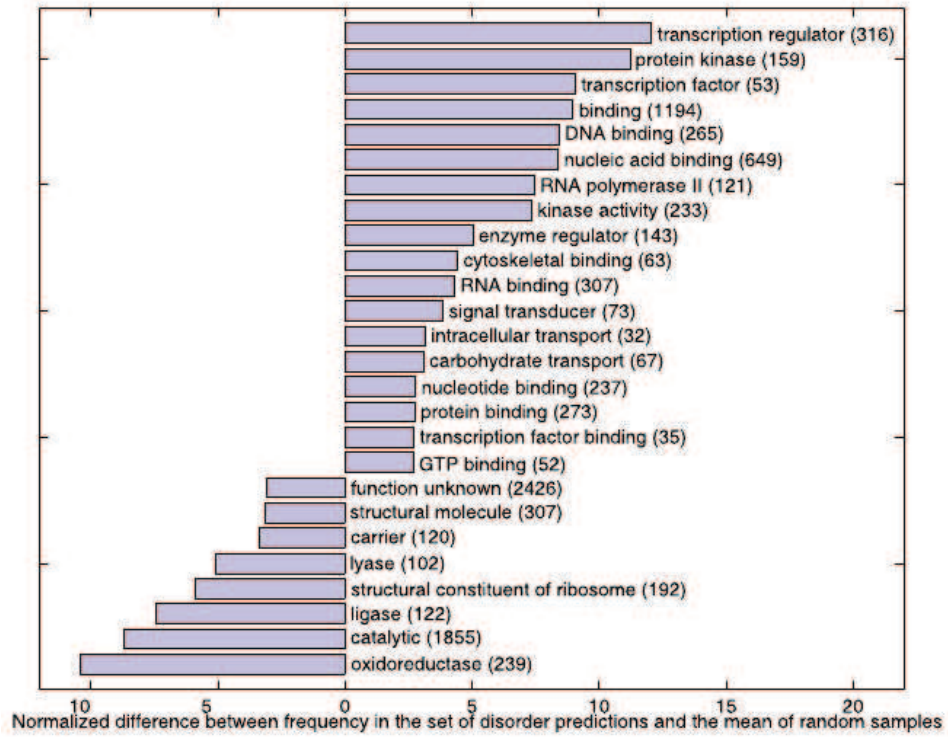


Figure 1.3. GO terms from the molecular function ontology that are significantly over- or under-represented in the set of proteins predicted to contain long regions of disorder. Each term is followed by the number of proteins in the yeast proteome that have been assigned this annotation. The terms are ordered by the normalised differences between the terms frequency of occurrence in the random samples and the set of disordered predictions. Figure taken from [424].

Bioinformatics surveys of this kind have inspired large-scale experimental investigations of the intrinsically unstructured mammalian proteomes. It was reasoned that it is not only important to understand the theoretical upper limit of the number of all IUPs encoded by genomes, but also which IUPs are actually expressed under certain physiological conditions and how cell vary their expressed IUP repertoire in response to changing conditions and external stimuli. Because it is not currently possible to predict protein expression patterns on the basis of genome sequence information alone, experimental methods are required for large-scale detection of expressed IUPs. Recently, Galea et al. [146] were able to identify a total of 1320 of thermostable proteins of the mouse proteome using state of the art proteomics techniques, 900 of which were predicted to be significantly disordered by several bioinformatics tools. They estimate that this amount constitutes ∼38-75% of the mouse unfoldome. The same analysis of GO terms (see above) was also performed in this study. The results were very similar to the ones of the *in silico* analysis.

The bioinformatics analyses are capable of providing even more interesting information. For example, it was recently demonstrated that predicted disorder is the most important feature of proteins that are harmful when the their corresponding genes are overexpressed, probably due the many

promiscuous molecular interactions they make when their concentration is increased [407]. This is an important finding for understanding the effects of increased gene expression in disease. IDPs are indeed involved in the pathogenesis of a wide range of human diseases, including cancer, malaria, AIDS, and amyloid diseases (see also fig. 1.4) [424, 193, 445]. Besides an altered gene expression, previous findings also suggest that diseases may result from misidentification and missignalling (which indicates that protein conformational diseases can involve more than only protein misfolding).

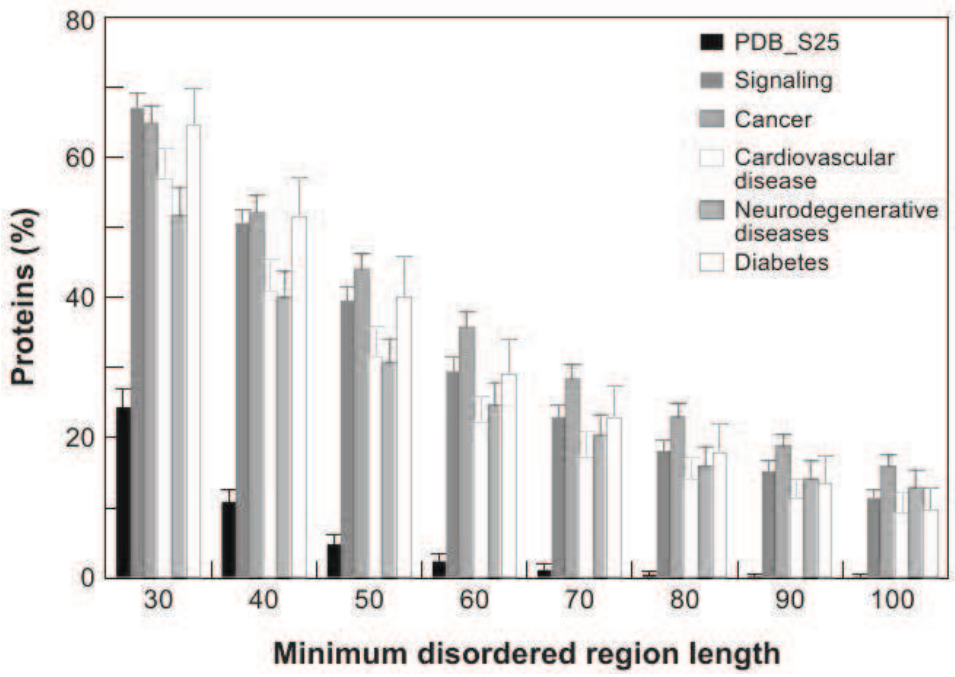


Figure 1.4. Abundance of intrinsic disorder in disease-associated proteins. Percentages of disease-associated proteins with >30 to >100 consecutive residues predicted to be disordered. The error bars represent 95% confidence intervals and were calculated using 1000 bootstrap resampling. Corresponding data for signalling and ordered proteins are shown for comparison. Analysed sets of disease-related proteins included 1786, 487, 689, and 285 proteins for cancer, CVD, neurodegenerative disease, and diabetes, respectively. Figure taken from [402].

Furthermore, Gsponer and colleagues [169] made a few remarkable observations by integrating multiple large-scale datasets that describe control mechanisms during transcription, translation, and post-translational modification with structural information on proteins, obtained from disorder predictors. It was found that the mRNA half-lives of the transcripts that encode highly unstructured proteins were lower than transcripts that encode more structured proteins because of higher decay rates. Consequently, the rate of protein synthesis was found to be significantly lower and protein half-life was shorter for highly unstructured than for more structured proteins. As also a significantly greater fraction of the unstructured proteins contains PEST motifs (regions rich in proline, glutamic acid, serine, and threonine) [108, 392], it appears that the availability of many IUPs is regulated via proteolytic degradation and a reduced translational rate.

Recent computational studies using phosphorylation site-prediction methods have suggested that unstructured regions are enriched for sites that can be posttranslationally modified [194]. In addition, Gsponer et al. [169]

also showed that 85% of the kinases for which more than 50% of their substrates are highly unstructured are either regulated in a cell cycle-dependent manner or activated upon exposure to particular stimuli or stress. These results suggest that several mechanisms contribute to the fine-tuning of IUP function and possibly their availability under different conditions. All observations were done both in unicellular and multicellular eukaryotic species.

Interestingly, of the 900 proteomics-identified proteins predicted to be either completely or partly unstructured in the study of Galea et al. [147], only 53 had (at the time of publication; 2008) previously been experimentally characterised as being either partially or wholly disordered, illustrating the limitations of our current knowledge of disordered proteins that are expressed in living cells. In trying to answer the question of how IUPs succeed in doing their complex and diverse tasks and how their faulty functioning results in disease, it becomes clear quite quickly that structuration events play a crucial role. In the next section, a thorough discussion of the structure of IUPs is thus given.

1.2. The Structural Feature of the Intrinsically Disordered State

The Ramachandran diagram [317], which plots ϕ versus ψ backbone conformational angles for each residue in a protein, has been with us nearly as long as macromolecular crystal structures. This same coordinate system is used to show either empirical scatter plots of the conformations observed in the database of known 3D structures, or else contours of calculated energies or steric criteria as a function of ϕ and ψ for a dipeptide. Especially in recent years, ϕ,ψ plots for individual proteins have also become central for structure validation, because ϕ,ψ values are not optimised in the refinement process and, therefore, provide a sensitive indicator of local problem areas [280]. Ramachandran plots showing favoured and allowed regions of ϕ,ψ space derived from highly resoluted structured proteins are given in fig. 1.5. Different typical secondary structure zones are presented in fig. 1.6.

Although the structural aspect of IDPs is not known in the detail, it can be said for certain that they contain much higher degrees of flexibility than their folded counterparts. This results in the fact that the ϕ and ψ angles in disordered proteins interchange rapidly between many different positions in the Ramachandran plot. As every protein molecule in an ensemble of molecules does this independent of the others, it makes no sense to concentrate on just one molecule. The unstructured state of a proteins can only be adequately described by the entire ensemble. IDPs are believed to cover a few families of structuration: random coils, premolten globules and molten globules.

1.2.1. Protein Random Coils

Over time, a few definitions have been given for the random coil state. Since observed experimental parameters cannot be directly related to a unique protein structure, using these values in a constraint-like manner to obtain a structural ensemble, is difficult. What can be done is, starting from a model for random coils, generating a structural ensemble and comparing back-calculated parameters with measured ones. Therefore, definitions have often (but not always) gone hand in hand with the description of a model.

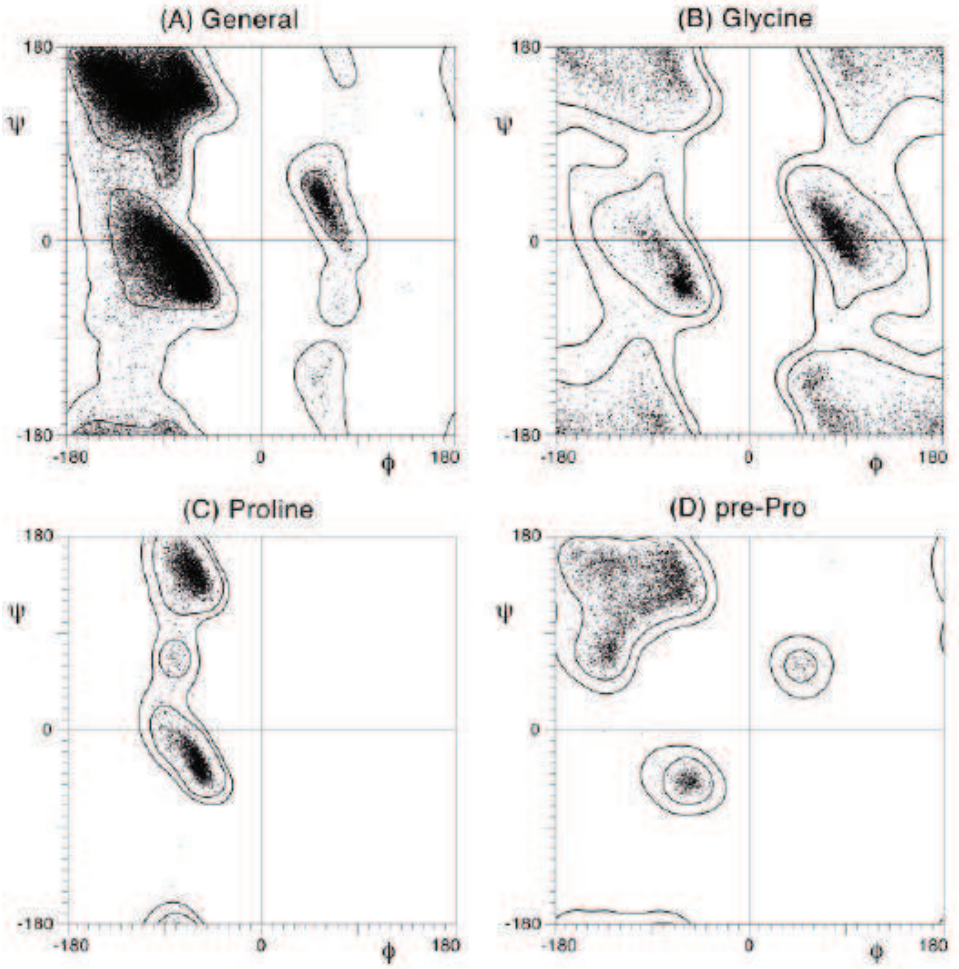


Figure 1.5. ϕ,ψ angle distributions for 97,368 residues with backbone B-factor <30 from a 500-structure high-resolution X-ray database, along with validation contours for favoured (comprising 98% of all data points) and allowed (including 99.95% of the data) regions. (a): The general case of 81,234 non-Gly, non-Pro, non-prePro residues. (b): The 7705 Gly residues, shown with twofold symmetrised contours. (c): The 4415 Pro residues with contours. (d): The 4014 pre-Pro residues (excluding those that are Gly or Pro) with contours. Figure taken from [252].

Until now, none of the given definitions seems to be completely satisfying. The initial random coil model was the freely joined, or random flight, chain [223]. In this model, monomers are connected by bonds of fixed length and uncorrelated (random directions). Although this model was capable of predicting e.g. the observed average radius of gyration of random coil polymers, it is clear that proteins do not really correspond to this model because of its oversimplicity.

According to Tanford [377], a polymer molecule is randomly coiled when internal rotation can take place at about every single bond of the molecule with the same freedom with which it would take place in a molecule of low molecular weight containing the same kind of bonds. Consequently, there are no strongly preferred conformations, the energy landscape is essentially featureless, and a Boltzmann-weighted ensemble of such polymers would populate this landscape uniformly. Protein random coils differ from true random coils in important respects, such as a non-random population of internal bond angles along the chain. That is, because of local, mainly

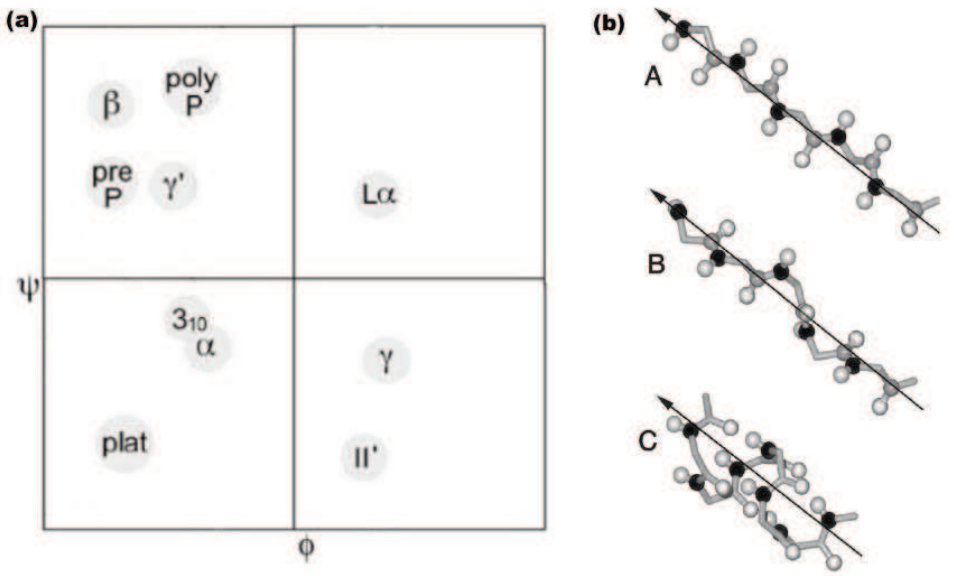


Figure 1.6. (a) contains the labels that indicate the approximate location of regions in the Ramachandran plot. The most favourable, low-energy regions are α+3_{10}, β + polyPro, and L_α plus some areas bridging α and β. More recently characterised regions correspond to the γ-turn, the mirror-image γ' conformation, II' turn, and below-α plateau regions. Residues preceding a proline tend to populate an additional region in ϕ,ψ space. Part (b) of the figure shows schematic diagrams of the backbone conformations of polypeptide segments that preferentially populate dihedral angles in the (A) β, (B) polyproline II and (C) α_R minima on the ϕ,ψ energy surface. (a) is taken from [252], (b) from [276].

sterical interactions, individual residues in a polypeptide chain will never adopt all ϕ,ψ combinations of the Ramachandran plot with equal probability, which leads to structures with highly anisotropic shape [1, 2]. However, to what amount random coil proteins contain residual structure is a question to which today still no satisfying answer has been given.

The idea that random coil proteins could contain important levels of residual secondary structure gained importance after some initial NMR experiments [97, 247, 355, 17, 354]. In particular, the non-zero RDC signals[3] obtained from presumed random-coil protein samples raised doubt about the existence of featureless random coils and about whether or not (residual) secondary structure elements might form an inevitable part of the protein random coil state [354, 1, 2]. On the other hand, as those samples still had the typical coil dimensions, a sort of a reconciliation problem arose [267].

The reconciliation problem was seemingly (but not really, see further) solved by the work of Fitzkee and Rose [137]. They calculated two related measures: the radius of gyration, R_G, and the end-to-end distance, $\langle L^2 \rangle$, (with both properties having typical values for random coil from the number of residues [377, 139]) for large generated ensembles of a number of hypothetical, artificially generated proteins that contained mainly (∼92%) short segments of rigid structure (α-helices and β-strands) linked by much

[3] Residual Dipolar Couplings (RDCs) are detected by NMR (see section 2.7.10). These signals probe the orientation of bond vectors, generally backbone amide NH, relative to an alignment tensor fixed in the molecular frame. RDCs generally vanish when molecules freely tumble, but remain when proteins are confined in weakly aligning media, at least if it contains enough structure to prevent random movements from averaging out the signals to zero.

smaller fragments (∼8%) of which the backbone torsion angles were allowed to vary freely. They obtained predicted R_G and $\langle L^2 \rangle$ values very similar to what would be expected for these values if the proteins were pure random coil. Also calculated Kratky plots (see section 2.3) greatly resembled those of the random coil variants. They thus showed that the random-coil statistics are not a unique signature for featureless polymers and that it could indeed be possible for presumed random coil proteins to have a significant amount of secondary structure.

Some time before these considerations, Dobson and Schwalbe and colleagues had introduced their so-called coil model [133, 365, 366, 338, 181] to interpret some observed experimental NMR parameters (mainly NOEs and J-couplings) in unstructured proteins. The coil model derives amino acid specific local conformations from the distribution of amino acid typical torsion angles in the Protein Data Bank (PDB), thus independent of neighbouring residues, in conjunction with steric exclusion. Deviations from the predictions of such models can then be interpreted as increased local order or order resulting from long-range contacts within the unstructured ensemble. In some cases, NMR parameters were successfully back-calculated using this model, while in others, observed deviations were indicative of some amount of secondary structure [338]. When the model was updated to include nearest-neighbour influenced torsion angles [302], observed J-couplings could be reproduced more precisely, thereby eliminating the requirement to invoke the existence of nascent secondary structure elements in certain intrinsically disordered regions. Introducing a degree of cooperation in the model (e.g. if a certain residue has α-typical ϕ,ψ angles, the neighbouring residue should have more than random chance to be sampling α-typical ϕ,ψ angles as well) improved the prediction of NOE signals in other cases [133]. These results indicate that in unstructured, in these cases denatured (random coil) proteins, residual secondary structure is mostly absent with small local exceptions. However, α and polyproline II (PPII) conformations have similar ϕ values (see figure 1.6) (that determine the measured J-couplings), and nearly as good agreement with experiment can be obtained by using conformational preferences based on the entire Protein Data Bank, which is dominated by α conformers, than by using ones based on unstructured regions in a coil library [203], which is dominated by PPII and β conformations. Hence, these measurements do not yield a stringent test for unstructured protein behaviour.

Later, two novel coil models, but based on the same principle, were independently applied to show that unstructured proteins do not necessarily contain this rich structural diversity [202, 28]. They both focused on the observed RDC values that had previously been interpreted in term of the presence of secondary structure (see above). Jha et al. [202] has constructed a coil library containing only residues (from X-ray structures) within stretches of certain length that lie outside of helices, sheets and turns. Unstructured conformations are built by assigning ϕ,ψ backbone angles based on a statistical potential derived from that coil library and to which an additional term has been added to account for the identity and conformation of neighbouring residues. To remove steric overlap a simple excluded volume energy function was also used during structure generation with Monte Carlo simulations. The flexible-Meccano's Monte Carlo algorithm used in [28] for generating the backbone of the unfolded-state conformations uses a subset of the database of amino-acid-specific ϕ- and ψ-torsion angles obtained by

exclusion of all residues in α-helices and β-sheets. The database has special cases for residues preceding a proline. In this case protein unstructured-state conformations are built by adding residues with a randomly selected pair of ϕ- and ψ-angles from the torsional subset database. If this introduces clashes, the angle pair is rejected and another one is randomly selected, thereby also implicitly introducing the influence of the preceding neighbour. Both studies succeeded reasonably well in simulating the observed RDC patterns of unstructured denatured proteins using their coil models, which means secondary structure must be mainly absent. The fact that RDCs of such unstructured proteins are found to be almost always of a single sign more likely indicates the preponderance of extended (β and PPII) conformations that align along the molecular axis.

Just before these studies, Wright, Dyson and co-workers proposed a similar view in that observed RDCs originate from transient alignment of short segments composed of extended conformations [276]. That is, they applied the rotational isomeric state theory of Flory [139] for the unstructured protein state, in which the chain is treated as a polymer of jointed statistical segments that are randomly oriented with respect to each other. Such statistical segments comprise several amino acid residues (determined by small-angle X-ray scattering and NMR relaxation measurements to be five to seven amino acids), with a propensity towards extended backbone (β or poly-proline II) conformations and that those segments are highly anisotropic in shape with their own alignment tensor. It was proposed that the observed RDCs of unstructured proteins thus arise form transient alignment of local regions of the chain, i.e. from alignment of the statistical segments, i.e. they will be a population weighted average over all low-energy backbone conformations in the ensemble, which in this case are a lot of β and polyproline II typical dihedral angles. The agreement with experiment is slightly less than in work mentioned above, but the results are largely the same. Observed RDCs in unstructured proteins can be explained without having to invoke the presence of much secondary structuration as was done before, and local steric restrictions on rotations about the dihedral angles results in local stiffness leading to extended conformations.

It could be noticed that, since a lot of this work is based on the study of unfolded proteins, not completely the same holds for intrinsically unstructured proteins. In fact, a total lack of intraresidue interactions would be unexpected in the unfolded state because certain (e.g., hydrophobic) side chains have high affinity for each other in the folded state [175]. Thus, in addition to what is expected from the preferential distribution of phi and psi angles, some secondary structure within unfolded proteins could be expected due to residual hydrophobic interactions [175, 352].

However, conclusions do not necessarily have to suffer from this issue. If the proposed models have learned us anything, it is that the ϕ,ψ-behaviour of residues in an unstructured protein in not only determined by the amino-acid type of that residue, but also by that of its neighbours. The number of adjacent residues that determine the conformational behaviour is given by the so-called persistence length, beyond which the remainder of the chain can be considered to exert a negligible effect. This concept of persistent length inspired Lippens and co-workers to define random coil proteins as proteins of which derived small peptides (for example 13-17 amino acids in length) give rise to exactly the same HSQC peak positions, at least for the central residues (were two amino acids on either side are considered as border) as

in the full length protein [364]. Such behaviour is indicative for the fact the amino acids sample the same conformational space in both contexts. However, this definition again views unstructured proteins as featureless random coils and thus only applies to genuine unstructured protein if they are free of any residual secondary structure. A recent study [162] showed, based on scalar coupling constant measurements and MD simulations, that the natural Ala_3 sequence in the protein hen egg white lysozyme (HEWL) adopts different conformations in a HEWL-9mer with the AAA pattern in the centre compared to a HEWL-19mer, again with AAA in the centre. The conformational distribution of the central three alanine residues in the 9mer is similar as for the small peptides Ala_3-Ala_7 (i.e. 80-90% PP_{II}). However, major differences are found for the 19mer, which significantly (30-40%) samples α_R helical structures. These results demonstrate explicitly that α-helix residual secondary structure can, in some cases, only be obtained in polypeptide stretches of lengths greater than proposed in the latter definition.

The consensus view would be that residues in intrinsically disordered proteins rapidly sample ϕ,ψ angles according to their own amino acid type and that of their neighbours in a more or less randomised manner. In general there is probably a preferences for more extended conformations. Polyproline II (P_{II}) helical conformations have indeed also been shown by both theory [217, 297, 281, 104, 275, 408, 151, 12] and experiment [386, 441, 349, 330, 131] to be a preferred conformation in unfolded peptide ensembles. Some regions (that correspond to the ones with decreased residual dipolar couplings) have for example increased flexibility of the polypeptide chain backbone. It is observed [276] that those regions are often rich in Gly and Ala residues, which thus function as flexible molecular hinges in the polypeptide chain. Other regions exhibit greater than average RDC values, which correspond to less flexible regions. The most flexibility-restricting residue has been identified as proline. Sign inversion of the RDCs on the other hand, could indicate residual α-helices in the chain. All these intrinsic conformational propensities of intrinsically unstructured proteins have probably direct relevance to their biological function (see further). The local concentration of residues in the PII region of the conformational space likely lowers the entropy of the unfolded protein chain, thereby facilitating folding under appropriate conditions [350]. More difficult to include in any random coil model are the long-range interactions that are more and more observed in unstructured proteins.

1.2.2. Between Random Coil and Complete ordered: the (Pre)molten Globule

Today it is known, from several experiments of protein unfolding, that the folded and random coil state do not completely cover all structural possibilities. Indeed, upon several denaturing conditions, globular proteins are known to exist in at least four different conformations: ordered, molten globule, premolten globule and unfolded [404, 310, 405]. The structural properties of the molten globule are well known, and have been systematised in a number of reviews (e.g. [310]). It has been established that the protein molecule in this intermediate state has no (or has only a trace of) rigid cooperatively melted tertiary structure, that is, it is denatured. Small-angle X-ray scattering showed that the protein molecule in this intermediate state has a globular structure typical of native globular proteins [118, 212, 213,

343]. 2D NMR coupled with hydrogen-deuterium exchange showed that the protein molecule in the molten globule state is characterised not only by the native-like secondary structure content, but also by the native-like folding pattern [22, 53, 199, 70, 444, 119, 38]. A considerable increase in the accessibility of a protein molecule to proteases was noted as a specific property of the molten globule state [273]. Finally, it was established that the averaged value for the increase in the hydrodynamic radius in the molten globule state compared with the native state is no more than 15%, which corresponds to volume increase of ∼50%.

A protein molecule in the premolten globule state is denatured; it has no rigid tertiary structure. It is characterised by a considerable secondary structure, although much less pronounced than that of the native or the molten globule protein (protein in the premolten globule state has ∼50% native secondary structure, whereas in the molten globule state the corresponding value is close to 100%). The protein molecule in the premolten globule state is considerably less compact than in the molten globule or native states, but it is still more compact than the random coil (its hydrodynamic volume in the molten globule, the premolten globule, and the unfolded states, in comparison to that of the native state, increases 1.5, ∼3, and ∼12 times, respectively). It has also been established that in the premolten globule state the protein molecule has no globular structure [401]. The last observation indicates that the premolten globule probably represents a "squeezed" and partially ordered form of a coil. Finally, it has been shown that the premolten globule is separated from the molten globule state by an all-or-none transition, which represents an intramolecular analog of the first-order phase transition [404, 310, 405]. This means that the molten globule and premolten globule represent divers thermodynamic (phase) states.

These two different novel thermodynamic states (molten globule and premolten globule) have been proposed not to exclusively characterise globular proteins in denaturing conditions. Some proteins have been shown to possess the same properties in physiological conditions. The molten globule and premolten globule state can in this case be considered as intrinsically disordered states, hence the family of IDPs has to be extended to comprise the two. However, concerning (pre)molten globule-like intrinsically disorder proteins, much less is known, and there has been some debate about their prevalence and importance. The intrinsic molten globule state has however been proposed to a number of biological implications. The insertion of proteins into membranes [57] and the transfer of retinal from its bloodstream carrier to its cell-surface receptor [55, 56] have both been suggested to depend on the molten globular state. These and other examples [342, 171, 63, 460, 455, 312, 108] support the existence of the molten globule form *in vivo*.

Experimental evidence for the existence of the native premolten globule state is less abundant, although studies using a multitude of experimental techniques have deduced the premolten globular state of their protein under study (e.g. [249]). Using multiple techniques might in this case be crucial. In particular, the seemingly simple CD-based method[4] Uversky proposes in [397] to distinguish between intrinsically random coil and premolten globule proteins (presented in fig. 1.7) is perhaps not completely up to the mark. At least one protein in the list of [397] of presumed premolten globules, the

[4] Circular Dichroism (CD) is briefly discussed in section 2.2

urea-denatured Δ131Δ domain of *Staphylococcal* nuclease, has recently been shown to fit experimental NMR RDC values surprisingly well when modelled using the previously mentioned flexible-meccano random coil model (see figure 2.11 on page 70). However this observation does not exclude the very existence of native premolten globular protein. It is probable that proteins rarely behave as true random coils, especially in non-denaturing media. Even in their most highly unfolded states, proteins show a propensity to form local elements of secondary structure or hydrophobic clusters. Hence, there is probably a subtle transition between protein random coils, protein random coils exhibiting some short- or even long-range interaction and possibly premolten globule proteins. The residual intramolecular interactions that typify the premolten globule state may enable a more efficient start of the folding process induced by a partner [389, 145, 30, 229]. Thus, the functional relevance of premolten globules may reside in a more pronounced propensity to undergo induced folding compared to random coil-like structures.

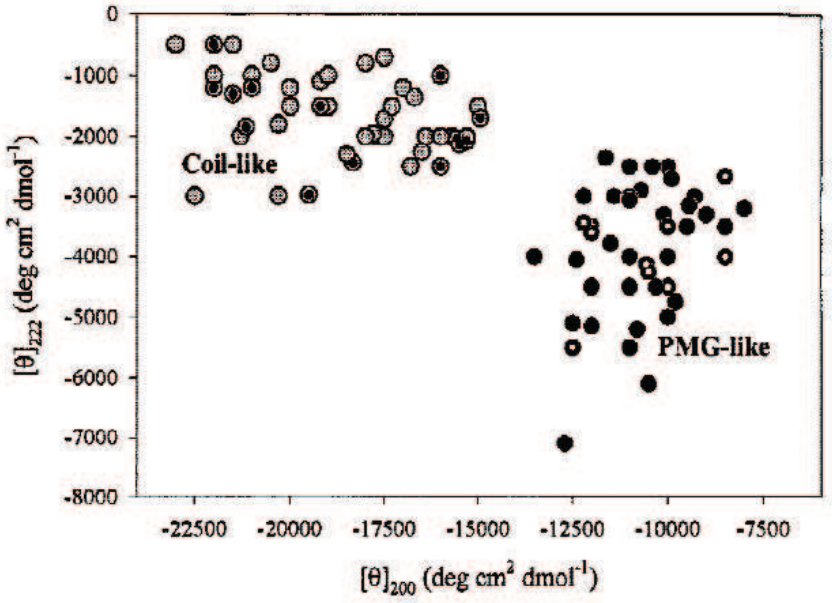


Figure 1.7. Analysis of far-UV CD spectra in terms of double wavelength plot, $[\theta]_{222}$ versus $[\theta]_{200}$, allows (according to [397]) the natively unfolded proteins division on coil-like (grey circles) and premolten globule-like subclasses (black circles). The black and white dots refer to proteins for which the hydrodynamic parameters were measured at time of publication. Figure taken from [397].

This observed continuum of structure in proteins has lead researchers to propose the "protein trinity" [108] or "protein quartet" [397] as alternative for the traditional protein structure-function paradigm (see figure 1.8).

1.2.3. Function-Related Structural Organisation in IUPs

Besides previous considerations about the intrinsic structure (or lack of it) of IUPs in free solution, more interesting from a biological point of view is how they behave when performing their function. However, in the following it will become clear that both aspects are closely related. Deviations from pure random coil behaviour often have a biological importance. In contrast

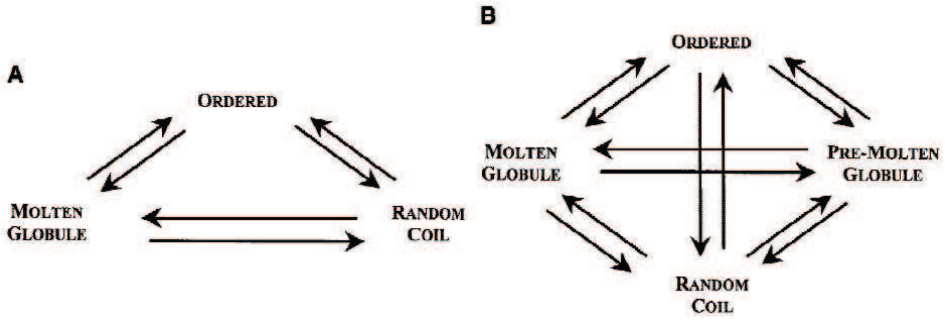


Figure 1.8. The Protein Trinity (A) and the Protein Quartet (B) model of protein functioning. In accordance with these models, function arises from three/four specific conformations of the polypeptide chain (ordered forms, molten globules, (premolten globules), and random coils) and transitions between any of the states. Figure taken from [397].

with the functions elucidated by bioinformatics analyses (see section 1.1.4), the following considerations are based on experimental approaches.

Functions of Intrinsic Disorder in Proteins

For all proteins containing disordered regions of 30 consecutive residues or longer, Dunker et al. argues those regions correspond to not less than twenty-eight separate functions [107]. By considering unifying mechanistic details of these various modes of action, the many different functions of IUPs actually segregate, according to Tompa and co-workers [389], into only six general categories. Although novel IUPs are identified regularly, this classification scheme (see fig. 1.9) appears to accommodate most examples known today [411, 358]. It is not my intention to give an extensive list of examples, so one or a few corresponding proteins will be mentioned for each class.

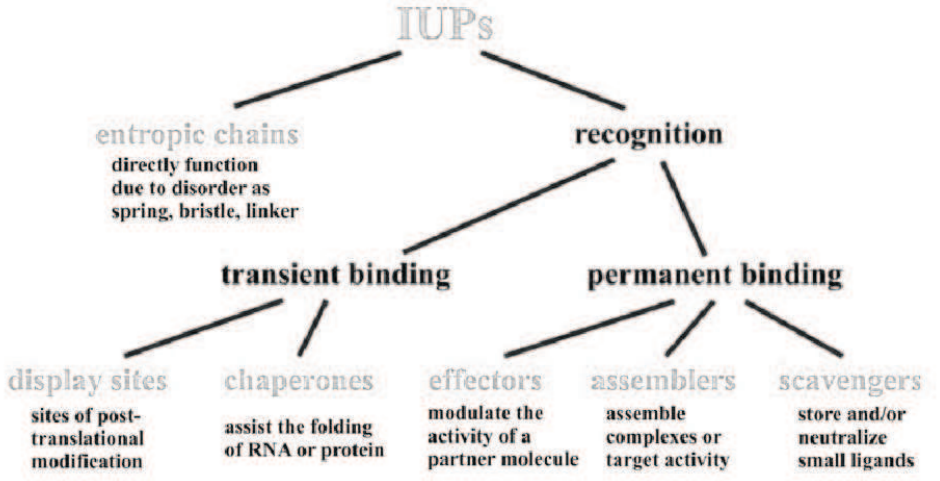


Figure 1.9. Functional classification scheme of IUPs. The function of IUPs stems either directly from their capacity to fluctuate freely about a large configurational space (entropic chain functions) or ability to transiently or permanently bind partner molecule(s). For each functional class, a short definition of function is given. More extended description are found in the text. Diagram from [390].

Entropic Chains The first general functional class of IUPs is that of entropic chains. These disordered regions apparently carry out function without becoming ordered. The class includes first of all flexible linkers/spacers between domains. Flexible linkers allow two domains to move relative to each other, and some also act as spacers that regulate the distance between adjacent domains [198, 207]. The functional, native state of flexible linkers/spacers is likely to be a random coil, or the polypeptide approximation of the random coil (see above) and hence carry out function without undergoing a disorder-to-order transition. A similar lack of a requirement for an ordered state characterises proteins that function as entropic springs (e.g. the pliable giant filamentous protein titin was shown to generate the passive force in muscle [394, 227, 216]) or entropic bristles (e.g. the thermally driven motion of unstructured polypeptide side-arms (part of MAP2) of microtubules are believed to maintain interfilament spacing [283]).

Display Sites In the other five classes, IUPs function via molecular recognition, i.e. they permanently or transiently bind another macromolecule or small ligand(s). Of those transiently binding their partner(s), display sites function as substrates for post-translational modification (PTM). Chemical modification of side chains requires close association between the target protein and the modifying enzyme. If the side chain being modified is within a structured region, steric factors would typically prevent or slow down the association. On the other hand, a side chain within a disordered region facilitates substrate binding because the disordered region can fold onto the modifying enzyme. Several types of chemical modification occur in intrinsically disordered regions [107, and references therein]. These include for example acetylation, fatty acid acylation, glycosylation, methylation, phosphorylation, and ADP-ribosylation. Phosphorylation is by far the most abundant kind of chemical PTM, and often acts to reversibly regulate the activity of the protein as a whole [209, 412].

Cis/trans isomerisation of prolyl peptide bonds (peptide bond preceding a proline, see fig. 1.10) by a peptidyl prolyl isomerase (PPIase) is another kind of PTM, that will prove important when discussing the behaviour of the NS5A protein in chapter 3.

The local environment of proline within a protein can influence the relative free energies of the cis and trans isomeric states, leading to wide variations in different proteins. Most folded proteins require proline to adopt one or the other isomer in the context of native protein folds. Hence, PPIases were originally discovered as helper enzymes for accelerating restructuring of the polypeptide backbone [135]. In contrast, due to their flexibility, completely disordered proteins do not exhibit the structural pressure for one of either isomers and an equilibrium system containing both forms often is the result. Indeed, adjacent amino acids exert local sequence effects on the cis/trans ratio but cannot liberate more than about 8 kJ/mol free energy difference between both isomers [451, 321]. This energetic contribution does not suffice to shift the cis/trans equilibrium completely to a certain side. The natural presence of both isomers rises questions about the utility of an accelerated interconversion between them. Cis and trans prolyl bonds will cause the IUP to sample different conformational spaces, but it is unclear whether this has functional consequences. Alternatively, the mere proline-directed binding capability of the PPIases has been proposed to be of exclusive biological importance [336, 255]. Moreover, catalysis of

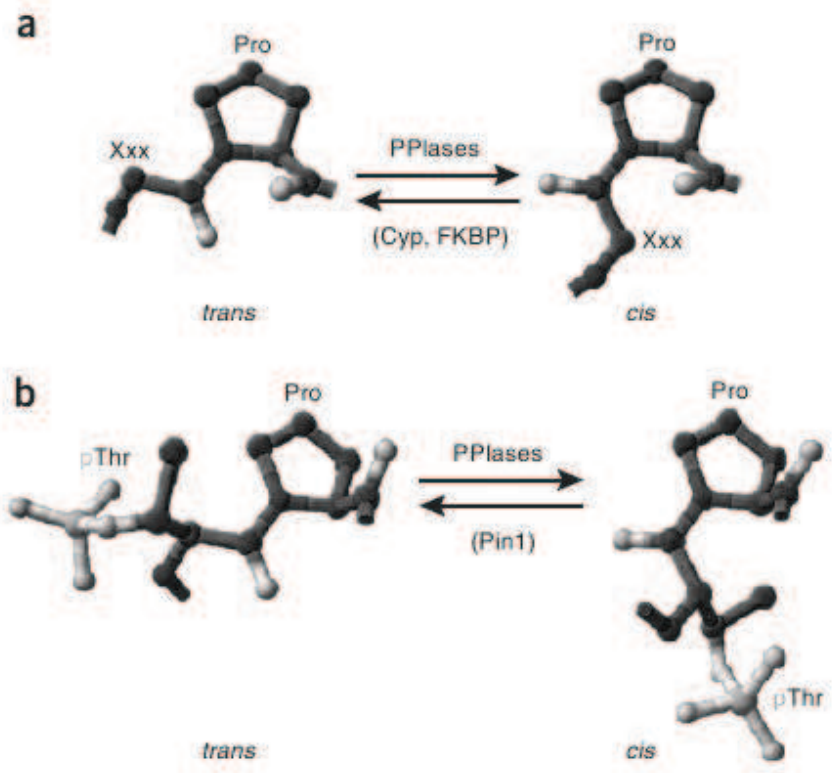


Figure 1.10. Because of the relatively large energy barrier, uncatalysed isomerisation is a rather slow process, but it can be greatly accelerated by PPIases. Based on the substrate specificity, PPIases can be divided into phosphorylation-independent and phosphorylation-dependent enzymes. The former group includes Cyps, FKBPs, PTPA and many parvulins that catalyse isomerisation of Xxx-Pro motifs, where Xxx indicates any amino acid except pSer or pThr (a). Pin1 and Pin1-type enzymes are the only known phosphorylation-dependent PPIases that isomerise on the pSer/Thr-Pro motifs (b). Image taken from [254].

prolyl bond isomerisation was discussed as a side effect attributable to the hydrophobic nature of the substrate binding site of the PPIases [136].

Protease cleavage could be considered as a third kind of post-translational modification. This is not so much the case for the proteasomal destruction that has been observed for non-ubiquinated disordered proteins such as casein [86], Tau [85] and $p21^{Cip1}$ [345]. Here the proteolic activity would be more an effective control allowing rapid turnover. However, the separation of a viral polyprotein into different smaller proteins is a common PTM theme.

Chaperones Another subclass within the category of transiently binding IUPs is chaperones. It was found (by statistical analysis) that RNA chaperones have a much higher incidence of disorder than any other functional class: 40% of their residues fall into long disordered regions (>30 residues), whereas the same number is 15% for protein chaperones [391]. Further, the function of many, or possibly all, of these proteins depends directly on disorder in a way that the disordered segment serves for either recognising, solubilising or loosening the structure of the misfolded ligand. To account for these mechanistic details, an entropy transfer model of disorder in chaperone function has been suggested [391].

Effectors Disordered proteins that function by permanent partner binding belong to either of the three classes of effectors, assemblers and scavengers. Effectors bind and modify the activity of their partner enzyme [389]. Their action is mostly inhibitory, but in some cases they may also activate another protein. The classical effector protein [389], p21Cip1 and its homologue, p27Kip2 have been shown not only to inhibit cyclin-dependent kinases (Cdks), but also to be able to assemble the cyclin-Cdk complex leading to Cdk activation [292].

Assemblers The next class is that of assemblers, which assemble multiprotein complexes and/or target the activity of attached domains [389]. Such proteins/domains have been noted in the assembly of the ribosome, cytoskeleton, transcription preinitiation complex and the chromatin, for example. The unusual complexity of interaction networks supported by such disordered assembly domains have been demonstrated within the partners of CBP, a multidomain transcription coactivator, which forms complexes with a variety of partners [114].

Scavengers The third subclass within this category, scavengers, store and/or neutralise small ligands. The classical examples of this mode of action are casein(s), which prevent calcium phosphate precipitation in the milk by capturing small seeds as they form and salivary proline-rich glycoproteins, which form tight complexes with tannins that can resist harsh conditions encountered in the digestive tract (cf. [389]).

Coupled Folding and Binding

Unstructured proteins executing functions in one of the five classes involving molecular recognition, interact with diverse partners such as DNA, RNA, other proteins or smaller ligands, during several key cellular processes, such as transcription, translation, signal transduction and the cell cycle. Upon the interaction with their partner(s), the energy landscape changes, which in several cases leads to disorder-to-order transitions upon binding. Closer examination reveals that this mechanism can be of crucial importance for regulatory functions, as it enables interactions of high specificity coupled with low affinity. This is because of the so-called isothermal enthalpy-entropy compensation (the free energy arising from the contacts of protein with ligand is reduced by the free energy needed to fold the intrinsic disorder) [443, 107, 389]. The low affinity assures the reversibility of the binding process. That the latter is of utmost importance for regulation, is also underlined by the fact that the action of IUPs is often modulated by phosphorylation.

To approach the issue of structural preorganisation, the actual bound structures have been compared to the inherent structural preferences of IUPs, assessed by secondary structure predictions [145]. It was shown that the prediction accuracy of IUP structures is comparable with that of their ordered partners, which suggests a strong preference of IUPs for the structure they adopt in the bound state. This implies the presence of preformed structural elements, which may limit the conformational search accompanying folding. A special case of such elements is termed primary contact sites (PCSs) [80], i.e. structurally primed, exposed recognition motifs that dock to the partner and lead to the formation of a native-like encounter complex. The presence of such sites has been inferred in MAP2 and CST

and suggested in several other IUPs [80]. These sites are conceptually closely related to anchor sites thus far reported for globular proteins [316] (the anchor site refers to a specific side chain on either of the interacting proteins that buries in a binding groove of the other protein to form a native-like encounter complex), hot spots also implicated in protein-protein interactions [35] (hot spots are residues that contribute much to the change in free energy upon binding) and molecular recognition elements (MoRes) associated with short ordered motifs apparent in disorder patterns [42]. Three basic types of MoREs were proposed: those that form an α-helix upon binding, those that form β-strands (in which the peptide forms a β-sheet with additional β-strands provided by the protein partner), and those that form irregular extended structures (in which the peptide backbone is stabilised by extensive hydrogen bonding with side chains of the protein partner). Respectively the names α-MoRE, β-MoRE and I-MoRE were given for these segments.

Although the underlying concepts are closely related, a good deal of kinetic/thermodynamic work will be needed to sort these things out, since a PCS/anchor site is defined in kinetic terms as recognition element that forms the initial contact with the partner, whereas a hot spot/MoRe is more of a thermodynamic term that signifies the region in the molecular interface that contributes the major part of the free energy of binding. It is to be noted that both may be interpreted in terms of the current "fly-casting" model [353] of IUP recognition, which suggests that IUPs bind weakly and non-specifically to their target and then fold, according to the characteristic energy landscape, as they approach the binding site. This mechanism invokes both the greater capture radius of IUPs and the mechanistic coupling of the recognition process to folding, in which pre-formed, exposed, recognition elements may be effective mechanistic devices. Besides the theoretically demonstrated kinetic advantages of this fly-casting mechanism [353], its experimental validity has also been proposed, based on NMR relaxation measurements [374].

Arguments in the above discussion suggest that an important advantage of disorderedness of proteins could hence be the possibility of faster protein-protein interaction on the one side, the low affinity/high specificity character of the interaction on the other side. However, these are not the only benefits of the increased structural plasticity. An additional prominent feature is that IUPs can adopt different structures upon different stimuli or with different partners, which enables their versatile interaction with various targets. This phenomenon termed binding promiscuity [221] or one-to-many signalling [108], enables an exceptionally plastic behaviour in response to the needs of the cell. An example is the Cdk inhibitor $p21^{Cip1}$, which can interact with CycA-Cdk2, CycE-Cdk2m CycD-Cdk4 complexes [221] and the Rho kinase [376].

A final coupled binding/folding related advantage is that their extended structure enables them to contact their partner(s) over a large binding surface for a protein of the given size, which allows the same interaction potential to be realised by shorter proteins overall, encoded by a more economical genome [170]. In addition, the flexibility is instrumental to the assembly process itself, as certain complexes cannot be assembled from rigid components due to topological constraints.

Although the most common theme of the involved induced folding is a coil-to-helix transition, interactions resulting in β-strand or β-sheet formation are equally observed (e.g. [5]). Furthermore, induced folding does not

necessarily have to be associated with a notion of gain of regular secondary structure [40, 303]. In those cases, binding to the partner would nevertheless lead to a more ordered state, through selection of a conformer of the IDP out of the numerous possible conformational states adopted by the unbound form in solution. The reduced conformational entropy of the IDP would then fall in the range of structural transitions typifying induced folding.

The mentioned disorder-to-order transitions would correspond to the vertical arrows of fig. 1.8.B, at least when considering only the fragment undergoing the structural changes. As for the protein as a whole, all the other transitions in the figure have also been observed. An example of a premolten globule to molten globule transition is given by the DNA-binding domain of the 1,25-dihydroxyvitamin D_3 receptor that makes this transition as a result of specific Zn^{2+} binding. This structuration enables the binding of this domain to, among other thing, DNA [79].

1.2.4. IUPs and Structural Biology

The direct examination of proteins with a well defined structure in free solution often leads to some immediately deducible biological information. The resolved structure of such proteins in most cases hints about the functioning, as for example the active site (in the case of an enzyme) becomes visible its physicochemical properties can straightforwardly be obtained. For intrinsically disordered proteins, this is much less the case. Although the previous discussion has clearly indicated that the IUP function is also implemented in its primary sequence and that function-related structuration events could be deduced from the residual structure contained in the polypeptide chain, obtaining biological information from pure free solution (structure elucidation) experiments is less evident. Therefore, in order to obtain the complete picture of IUP functioning, experiments should, whenever possible, also be performed in presence of the interacting agent(s).

In the case this interacting partner is unknown, free solution investigations of the IUP are the only possibility and provide an initial shaping of the full-characterisation study of the individual protein. There is an important note to this, however. Characterisation of an IUP ensemble in absence of a binding partner is mostly done in highly diluted solutions *in vitro*, which cannot fully represent the situation inside cells. It has been shown that the crowding effect elicited by extreme macromolecular concentrations (up to 400 mg/ml) in living cells may significantly shift their conformational equilibrium towards a folded state [120]. In fact, the intrinsically unstructured inhibitor of the transcription factor sigma28 [83], when expressed in *Escherichia coli*, undergoes significant ordering, as demonstrated by NMR [90]. In other cases studying the effect of such crowding, evidence is mostly against overall folding with only a marginal tendency to form structure [138, 279, 311]. Hence, these considerations do not suggest one should rule out the existence of intrinsic disorder *in vivo*. Observations such as the higher rates of evolution (see above) that characterise unstructured proteins (which is completely independent of experimental conditions) indicate these proteins most likely do not possess the appropriately shaped energy landscape to fold into a 3D structure. Also the fact that some proteins that are highly disordered in the laboratory, have short lifetimes in the cell (for functional reasons), provides a further argument for the existence of disorder

in vivo [443]. However, the detailed (residual) structuration of an IDP might be different *in vitro* compared to *in vivo*.

The fact that many intrinsically disordered proteins have only been studied at this free solution level and wait for further characterisation in presence of their interacting partner(s) is nicely demonstrated by DisProt, a databank for IDPs [411, 358]. To account for the constantly increasing number of experimentally characterised IDPs, DisProt has been built and is being continually updated. The DisProt Release 2.0 (14 February 2005) included 179 IDPs and 290 disordered regions, whereas the current DisProt Release 4.9 (1 May 2009) included 523 IDPs and 1195 disordered regions. This indicates that the number experimentally determined and annotated disordered structures is increasing rapidly, but there still is an enormous gap compared with the actual number of IDPs in nature and compared to the number of experimentally validated IDPs. However, more importantly, of all the IDPs contained in DisProt, which were studied in free solution, the biological functions of only a limited number of them are known.

It could be instructive to exemplify the different levels of structural and functional knowledge with a few situations with which our research group has some affinity. Three situations can be considered. (*a*) Stathmin is an IDP of which the function is known and of which the structural cause of its functionality was later equally elucidated. (*b*) Of the Tau protein at least one function is known. However, the structural details of its working are unknown to date. (*c*) For the NS5A protein of the hepatitis C virus both function and obviously also the underlying structural principles still need to be discovered. Contributions to the two situations that are not fully understood (*b* and *c*) have been made in this thesis (see chapter 3).

Functional and Structural Aspect Known: Stathmin

Stathmin is an intrinsically disordered protein implicated in the regulation of microtubule dynamics. It influences this microtubule dynamics *in vitro* and *in vivo* either by preventing assembly or promoting disassembly of microtubules in a concentration-dependent manner. The protein therefore plays a central role in cell proliferation, cell migration, and mitotic spindle formation [65, 329]. *In vivo*, the activity of stathmin is down-regulated by posttranslational phosphorylation in response to a number of signals on four serine residues, Ser16, Ser25, Ser38, and Ser63 [232, 183, 225, 438]. In mitotic cells, for example, phosphorylation by an unknown kinase-phosphatase system allows creating local stathmin activity gradients, a process essential for regulating microtubule dynamics and spindle formation. Phosphorylation of Ser16 and Ser63 strongly down-regulates the microtubule destabilising activity of stathmin [232, 183, 225, 438, 95, 8]. In contrast, phosphorylation of Ser25 and Ser38 has only a moderate effect on down-regulation but is a prerequisite for allowing phosphorylation of Ser16 and Ser63 *in vivo* [232].

It has been demonstrated, using techniques like video microscopy, that non-phosphorylated Stathmin promotes disassembly by stimulating so-called catastrophes (which refers to the transition of microtubule growth to rapid shortening) [185, 26, 263]. However, the explanation of how Stathmin prevents assembly came from structural biology. Stathmin binds to two head-to-tail aligned α/β-tubulin heterodimers [369, 81, 208, 319, 154, 99]. Upon this binding the IUP N-terminus folds into a β-hairpin, and the C-terminal helical domain becomes stabilised [319, 154, 71, 99] and forms a

~90-residue α-helix interacting with the tubulin dimer. This evokes first of all a curved structure of the tubulin-stathmin complex which does not allow lateral interactions to be established as occurs in growing microtubules and secondly, the tubulin capping capability of the β-hairpin also impedes tubulin polymerisation. Both these issues provide a structural basis for understanding how stathmin family proteins destabilise microtubules (see fig. 1.11).

Finally, phosphorylation is able to cancel microtubule-disassembly function of Stathmin as phospho-Ser16 and phospho-Ser63 disrupt the formation of a tubulin-interacting β-hairpin and a helical segment, respectively [182].

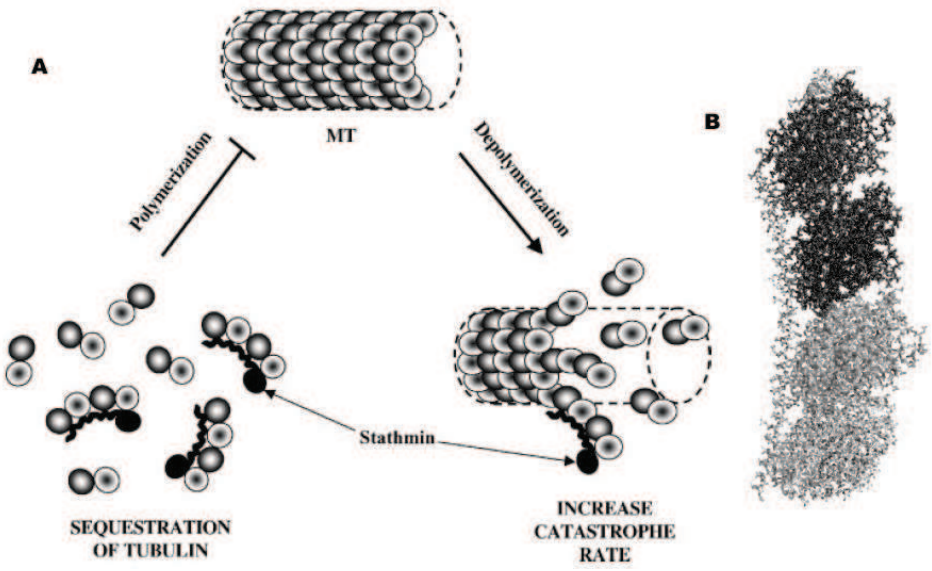


Figure 1.11. (A) Model for the role of stathmin in the regulation of microtubule dynamics. Microtubules (MT) continuously switch between phases of polymerisation and depolymerisation. Stathmin can sequester unpolymerised tubulin by binding two α/β-tubulin heterodimers (represented by light and dark shaded circles), thus diminishing the pool of tubulin heterodimers available for polymerisation. Stathmin can also bind to the end of polymerised microtubules and increase the rate of catastrophe by inducing a conformational change that promotes microtubule depolymerisation. Figure taken from [329]. (B) shows the crystal structure of T2R, the Tubulin:RB3-SLD (Stathmin-like domain) complex. Structural and biochemical data suggest that stathmin and RB3-SLD interact in the same way with tubulin. The curvature of the complex and the capping β-hairpin are clearly visible.

Functional Aspect Known but Structural Aspect Unknown: Tau

The Tau protein is a microtubule-associated protein (MAP) that is abundant in neurons in the central nervous system. The protein interacts with tubulin to stabilise microtubules (MTs) [54] and promote tubulin assembly into microtubules [436]. Furthermore, Tau is involved in the transport of vesicles and organelles along MTs [231] and serves as an anchor for enzymes [367].

Six Tau isoforms exist in brain tissue ranging from 352-441 amino acids in length. The major, full-length Tau protein in the human brain (htau40) is encoded by 11 exons. However, exon 2, 3 and 10 are alternative spliced, allowing five further combinations (besides 2+3+10+, there are 2-3-10-; 2+3-10-; 2+3+10-; 2-3-10+ and 2+3-10+). The exclusion/inclusion of exon 10 determines the number of repeat domains (three vs. four) of the protein.

The isoforms with four repeat domains are better at stabilising microtubules than those with three binding domains.

Indeed, Tau has been shown to interact with microtubules via its microtubule-binding (MTB) domain [54], which consists of the repeat regions (amino acids 244 to 369 in full-length Tau; see also fig. 3.1 on page 92). This binding domain is located in the carboxy-terminus of the protein and is positively-charged (allowing it to bind to the negatively-charged microtubule). However, the binding is also influenced by the proline-rich regulatory region flanking the MTB region upstream [262, 161, 308].

Besides by varying the isoform ratio Tau-3R/4R, binding of Tau to microtubules is dynamically regulated by the degree of phosphorylation (phosphorylation slows down the polymerisation) [242], which is controlled by a host of kinases and probably involves protein (de)structuration events. This affinity regulation is particularly done by phosphorylation at KXGS-motifs in the repeats [31].

Interestingly however, hyperphosphorylation of the Tau protein, can result in the self-assembly of tangles of paired helical filaments (PHF) and straight filaments, which are involved in the pathogenesis of Alzheimer's disease and other Tauopathies [178, 6]. All of the six Tau isoforms are present in an often hyperphosphorylated state in paired helical filaments from Alzheimer's Disease brain. The same regions involved in MT binding are also important for PHF aggregation, i.e. two hexapeptides at the beginning of the second and third repeats ^{275}VQIINK^{280} and ^{306}VQIVYK^{311} were shown to able to initiate the aggregation process [409].

A further interesting case is that of Tau-P301L (proline 301 is converted to a leucine). This mutation of a microtubule binding domain residue, that has been linked to frontotemporal dementia (FTD), a non-Alzheimers dementia [325, 106], has been shown to involve a partial loss of the microtubule assembly/binding properties of Tau [87].

These several functional aspects of Tau seem well understood, however, its mode of action is still enigmatic. Structural biology will still have to explain these issues. NMR being a powerful method in the case of Tau, NMR chemical shift assignments of the protein are an important prerequisite.

Functional and Structural Aspect Unknown: NS5A

Hepatitis C virus (HCV), known to infect humans and chimpanzees and responsible for illnesses such as chronic hepatitis, liver cirrhosis and hepatocellular carcinoma (HCC), has a life cycle that is to this day far from elucidated. HCV is characterised by a high genetic variability and accordingly six genotypes and further subtypes have been identified [362], which are mostly dependent on the geographical origin. Despite this variability, the HCV positive-strand RNA genome encodes for a consistent set of structural and non-structural[5] proteins (derived from a polyprotein precursor of about 3000 amino acids), that each have their function in the virus' replication process. The function(s) of some of these proteins has been elucidated and a structural basis providing further insights in their functioning, has been proposed in some cases [278].

[5] Unlike structural proteins that are part of the viral particle, nonstructural proteins are encoded by the virus but are not part of the viral particle. Their function is often unknown, but some play structural roles within the infected cell during replication or act in virus regulation.

However, the detailed mechanisms regulating translation, replication and packaging of the viral genome are unknown. One of the viral proteins implicated in these mechanisms is NS5A. Current evidence indicates that NS5A might function as a molecular switch between replication and assembly, as the phosphorylation state of this protein affects HCV RNA replication, and there is an inverse correlation between NS5A adaptive mutations facilitating replication and virus production [122, 286, 10]. The NS5A protein has been demonstrated to bind RNA [187], but this process is not understood in detail. NS5A is anchored to the endoplasmic reticulum membrane via an amphipathic N-terminal α-helix. After this helix region, the protein is shown to consist of three domains [382]. The first domain was shown to form dimers and is partly structured, a state that is maintained through Zinc association [383]. Domains II and III of NS5A further extent in the cytosol, but very little is known on their detailed structural and functional behaviour (see fig. 1.12).

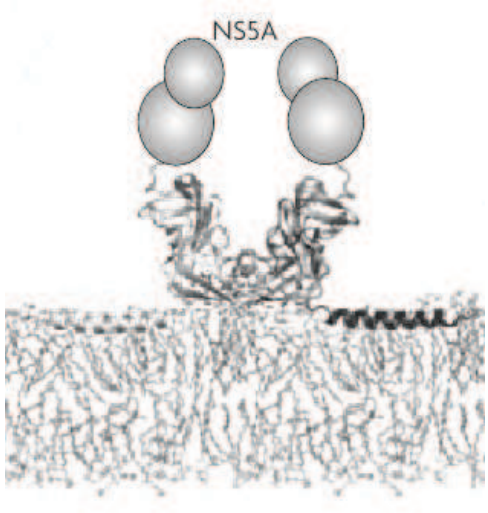


Figure 1.12. The HCV NS5A protein presented as a dimer bound to the endoplasmic reticulum membrane. Immediately after the membrane binding helices are the are the first domains of the two monomers (of which the structure is also shown). Domains II and III are represented as spheres since at the time this image was published (2007) no structural details were known for these domains. Image taken from [278].

In the course of this thesis, a closer look was taken at the structural properties and the possibly functionally relevant interactions with cyclophilins of the domains 2 and 3 of NS5A.

Chapter 2

Biophysics for IUPs and the special role of NMR

IUPs can be characterised by more than twenty biophysical methods, each of which gives slightly different information. Most of the described techniques can be used both for a characterisation of the IDP in free solution as for the IDP in presence of an interacting partner. Perhaps the most important difference to bear in mind when performing experiment on IDPs in free solution and trying to interpret the results, is the difference between a structural state and a thermodynamic state. For folded proteins, the native state is both a structural state and thermodynamic state, but the disordered state of IDPs is only a thermodynamic state. That is, all the molecules in a sample of the native state of a globular protein have nearly the same structure. On the other hand, the disordered state consists, as previously mentioned, of a broad ensemble of molecules, each having a different conformation. Therefore, averaged quantities have different meanings for folded and disordered states. For a native globular protein, an averaged quantity gives information about each molecule in the sample because nearly all the molecules are in the same structural state. For a disordered protein, an averaged quantity contains information about the ensemble, and this information may or may not be applicable to individual molecules in the sample.

A few of the techniques most commonly applied on IUPs are presented hereafter. The emphasis in this chapter will be on the particularly powerful technique of NMR when it comes to IUPs. An NMR assignment tool developed during my thesis will be presented by the end of the chapter, when it is argued that the several discussed (and possibly subsequently applied) NMR methods require the individual nuclei to be identified first.

2.1. X-ray Crystallography

As mentioned at the beginning of the previous chapter, disorder leads to missing electron density in protein structures determined by X-ray crystallography. Two types of disorder have been recognised: static and dynamic [189, 190]. If a dynamic region freezes into a single preferred structure upon cooling, then collecting data at lower temperatures distinguishes dynamic from static disorder in some cases [103]. However, from our point of view, more important than static or dynamic, is whether the missing region assumes one set of the Ramachandran ϕ, ψ angles along the backbone or whether the missing region exists as an ensemble of angles. We call a missing region with one set of ϕ, ψ angles, whether static or dynamic, a "wobbly domain" because such a region assumes different positions as a rigid body, with the transitions between the different positions being slow (static disorder) or fast (dynamic disorder) on the X-ray analysis time scale. From our point of view, a region existing as an ensemble of ϕ, ψ angles, whether static

or dynamic, is intrinsically disordered. The major uncertainty regarding information from X-ray diffraction is that, without additional experiments, it is unclear whether a region of missing electron density is a wobbly domain, is intrinsically disordered, or is the result of technical difficulties. However, neglecting this last possibility, the amino acid compositions of long regions of missing electron density are very similar to the amino acid compositions of disordered ensembles characterised by NMR. Furthermore, predictors based on NMR-characterised disorder for the most part predict disorder for the long regions of missing electron density. Thus, as an explanation of long regions of missing electron density, wobbly, ordered domains are probably the exception rather than the rule [152].

Recently, attempts to derive high resolution structural information of IDPs have been reported, based on crystallisation of such proteins as fusion with GST or in the presence of binding partners or antibodies [453, 46, 250]. However, these structures remain representative only of one particular member of the conformational ensemble of the free protein in solution.

2.2. Circular Dichroism (CD) spectroscopy

Structural information for proteins in solution can be provided by circular dichroism [128, 3]. There are two types of optically active chromophores in proteins: side groups of aromatic amino acid residues, and peptide bonds. Far-UV CD spectra (that observe peptide bond properties) provide estimates of secondary structure and so distinguish ordered and molten globular forms from random coil. A typical mostly unstructured protein has a large negative peak at 200 nm and a value close to zero at 220 nm. Partial folding is reflected by a loss of CD spectral signal at 200 nm and a gain of CD signal at 220 nm. On the other hand, near-UV CD (250-350 nm) reflect the symmetry of the environment of aromatic amino acid residues and show sharp peaks for aromatic groups when the protein is ordered, but these peaks disappear for molten globules and random coils due to motional averaging [98, 291, 226]. Thus, combined use of near and far-UV CD can distinguish whether a protein is ordered, molten globular or random coil. However, this method is only semi-quantitative and lacks residue-specific information and so does not provide clear information for proteins that contain both ordered and disordered regions.

2.3. Small-Angle X-ray Scattering (SAXS)

SAXS is useful in obtaining the very important structural parameter of the degree of globularity, which reflects the presence or absence of a tightly packed core in the protein molecule. This information may be extracted form the SAXS data derived Kratky plot (s versus $s^2I(s)$, where s is the small-angle scattering vector and $I(s)$ is the corresponding scattering intensity), whose shape is sensitive to the conformational state of the scattering protein molecules [157, 129, 343]. It has been shown that a scattering curve in the Kratky coordinates has a characteristic maximum for globular proteins in either their native or molten globule states (i.e., states with globular structure). However, if a protein is completely unfolded or in a premolten globule conformation (i.e., with no globular structure), such a maximum will be absent.

2.4. Fluorescence Resonance Energy Transfer (FRET)

Fluorescence labels are indispensable tools in studies on energy transfer between two chromophores. The essence of the phenomenon is that, at interaction of oscillators at a small distance, the electromagnetic field of the excited oscillator can induce oscillations with the same frequency in the nonexcited oscillator [141, 228]. The transfer of excitation energy between the donor and the acceptor originates only with the fulfilment of several conditions: (1) the absorption (excitation) spectrum of the acceptor overlaps with the emission (luminescence) spectrum of the donor. This is an important prerequisite for resonance; (2) a sufficient spatial proximity of the donor and the acceptor is necessary; they must be at distance not exceeding a few dozens of Angstroms; (3) a sufficiently high quantum yield of the donor is also necessary; (4) spatial orientation of donor and acceptor also plays an important role for the effective energy transfer. The biggest advantage and attractiveness of FRET are that this method can be used as molecular ruler to measure distances between the donor and acceptor. In fact, according to Förster, the efficiency of energy transfer, E, from the excited donor, D, to the nonexcited acceptor, A, located from the D at a distance R_{DA} is determined by an equation [141]:

$$E = \frac{1}{1 + \left(\frac{R_{DA}}{R_o}\right)^6} \tag{2.1}$$

where R_o is the characteristic donor-acceptor distance, so-called Förster distance. Usually, in FRET experiments, one uses intrinsic chromophores (tyrosines or tryptophanes as donors and covalently attached chromophores emitting light in visible region) as acceptors. As it follows from Equation 2.1, the efficiency of energy transfer is proportional to the inverse sixth power of the distance between donor and acceptor. Obviously, structural changes within a protein molecule might be accompanied by the changes of this distance, giving rise to the considerable changes of this parameter.

An elegant approach based on the unique spectroscopic properties of nitrated tyrosine (which has maximal absorbance in the vicinity of 350 nm, does not emit light and renders as an acceptor for Trp electronic energy) has been recently elaborated [324, 399, 380, 379]. For these experiments Tyr residues have to be modified by reaction with tetranitromethane to convert them to a nitro form, Tyr(NO_2). The extent of decrease of Trp fluorescence in the presence of Tyr(NO_2) provides a measure of average distance R_{DA} between these residues. This decrease reflects variations in the distance between these residues as resulting from conformational changes.

2.5. Limited Proteolysis

The role of residual structure in IUP function can be approached by limited proteolysis. This technique is traditionally used to probe the topology of globular proteins and their folding intermediates [140], as proteases generally attack spatially exposed and flexible sites. Under conditions of extremely low protease concentrations, however, IUPs also undergo limited proteolysis, which implies their non-random structural organisation. It has been shown that the location of the preferential cleavage site(s) correlate with their domain organisation [266, 322, 320, 264, 80]. An appealing interpretation of

this observation is that transient short- and/or long-range structural organisation ensures the spatial exposure of certain regions (and the concealing of other regions) in these IUPs. This is of particular relevance for their binding functions as the large-scale binding-coupled folding of IUPs is hardly compatible with a fully disordered structure prior to binding. Rather, it may be anticipated that IUPs exploit some sort of structural preorganisation in effectively recognising their partner and initiating the subsequent induced folding process.

2.6. Electron Paramagnetic Resonance (EPR)

A powerful and sensitive technique to probe protein structure is provided by EPR spectroscopy. This technique is based on the covalent modification of a cysteine side chain to yield a nitroxide side chain that possesses an unpaired electron. Because the mobility of a spin label covalently attached to a protein is influenced by its environment, analysis of its EPR spectra provides insights into the local protein structure [188], and into the equilibria between ordered and disordered conformations within poorly structured proteins [59]. Conformational changes of the spin labelled protein are reflected in the variations of the EPR lineshape [304]. A requirement is obviously that the nitroxide side chain does not impair the formation of a complex with a ligand partner.

2.7. Nuclear Magnetic Resonance (NMR)

Among the most powerful techniques for characterising the unstructured states of proteins is NMR, since it enables to study the proteins with atomic resolution. In the case of a characterisation of the free unstructured states, the normally encountered size problem in NMR can partially be neglected thanks to fast dynamics (on a variety of timescales) that accompany the intrinsic flexibility. I.e., T_2 is longer than for the natively folded protein. Also because of this, pulse sequences with a larger number of delays will still be applicable, where they would not be in a folded protein of comparable molecular weight. However, because of conformational averaging and also because all amino acid residues share a similar chemical environment (being all prevalently exposed to the solvent), the reduced chemical shift dispersion makes spectral overlap (particularly for proton and aliphatic carbon resonances) a more severe problem than for globular proteins. The backbone ^{15}N and ^{13}CO resonances are influenced both by residue type and by the local amino acid sequence and therefore remain well-dispersed, even in fully disordered states [43, 456].

The same is, unhappily, not true for partly folded proteins, or molten globule states. Compared with ordered proteins, relatively few molten globules have been structurally characterised by NMR, indicating the existence of significant experimental difficulties. First, proteins with molten globular regions often aggregate at the concentrations needed for NMR. Second, the molten globules are heterogeneous with structural interconversions on the millisecond (spectral) timescale; this leads to extreme broadening of the side-chain NMR peaks, as is explained schematically in fig. 2.1.

High-resolution NMR studies of disordered proteins have been largely enabled by the advent of high-field spectrometers and multidimensional he-

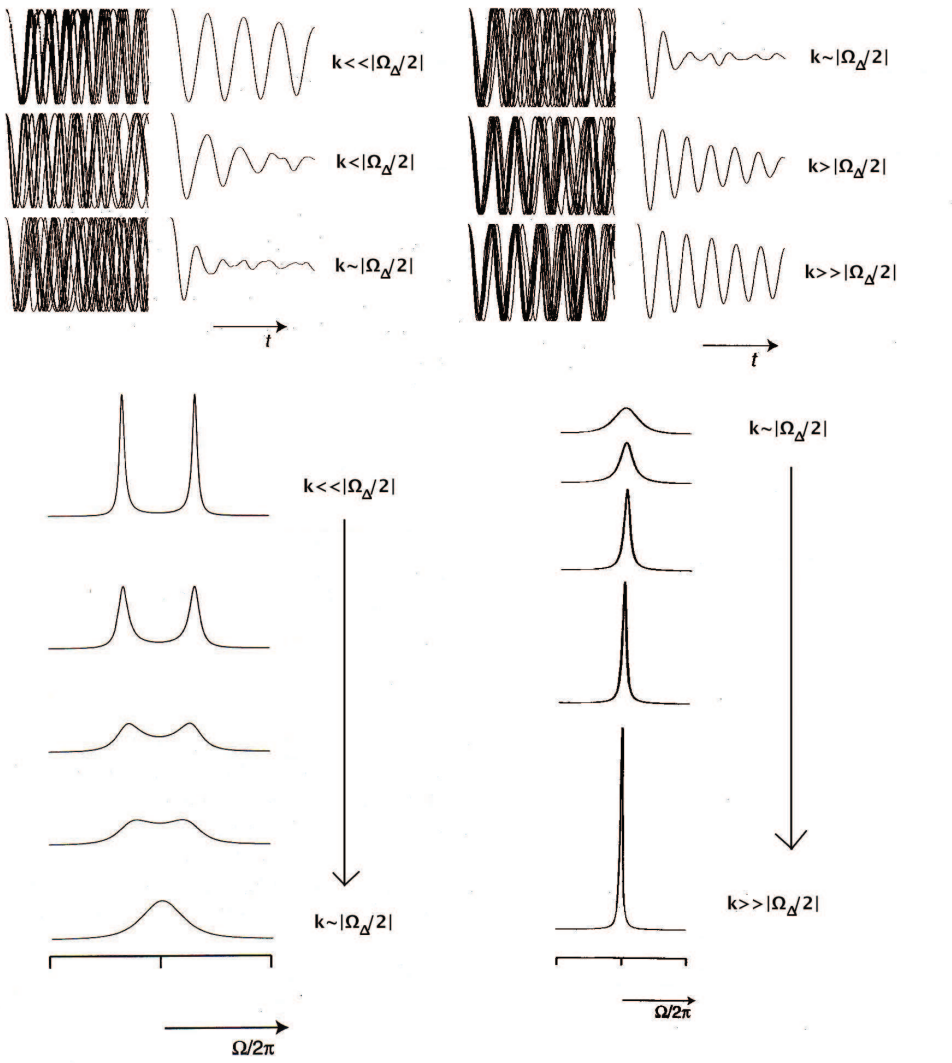


Figure 2.1. The upper part of the figure shows the behaviour of an ensemble averaged FID coming from spins involved in structural interconversions. During the motion, a hypothetical part of a protein leaves structure A (which gives the spin considered a chemical shift frequency of Ω_A^0) and adopts structure B (corresponding to a chemical shift frequency Ω_B^0). The change in chemical shift is $\Omega_\Delta = \Omega_A^0 - \Omega_B^0$ and the interconversion rate is k. Both FIDs (top) and Fourier transformed spectra (bottom) indicate that motional events happening on the spectral timescale ($\tau_{spect} = |2/\Omega_\Delta|$) leads to serious line broadening. Parts of the figure were taken from [235].

teronuclear NMR experiments. High sensitivity is a critical factor in the study of unstructured proteins, since NMR experiments must often be performed at very low concentrations (of the order of 50-300μM) to prevent aggregation.

A variety of NMR-based approaches have so far been used to characterise fully or partially unstructured proteins [112, 113, 269, 116]. Several of these approaches will be explored here. Due to the greater maturity of the field of protein folding/unfolding, much of the work reviewed here was performed on non-native states of globular proteins. Because of their great similarity, both protein classes can in principle be studied using the same techniques.

NMR and the Characterisation of the Ensemble of Structures

In the absence of interacting partners, IUPs form a broad, heterogeneous ensemble of structures that undergo conformational fluctuations on multiple time-scales. As mentioned, in a first stage, one would like to characterise the range of conformations consistent with the experimental data and consequently detect the presence of possibly biologically relevant residual structure. All NMR parameters are in fact a population-weighted average over all the structures in the conformational ensemble. Conformational preferences can therefore be identified by comparison of experimental NMR parameters to those expected for a random coil state, in which the polypeptide backbone dihedral angles adopt a Boltzmann distribution over the ϕ,ψ energy surface [133, 366, 202, 28]. In addition, it is implicitly assumed that the dominant conformers in unstructured polypeptides will have backbone dihedral angles that lie within the broad α and β minima on the ϕ,ψ conformational surface.

In some cases, when one suspects the presence of relevant long-range interaction, experimental NMR data can be applied in simulations that try to reconstruct the ensemble of the intrinsically disordered proteins. Most useful data in such simulations are residual dipolar couplings (RDC) and paramagnetic relaxation enhanced (PRE) distance restraints. More detailed information on a structure ensemble (finding a distribution of conformations that accurately reflects the experimental data), can be expected from the use of molecular dynamics simulations (MD) or simulated annealing (SA) that proceed with restraints that are either time averaged over a single molecule or averaged at an instant in time over an ensemble of non-interacting molecules. Such NMR parameter restrained molecular dynamics simulations have already been performed with success [89, 253] and might provide more useful information (complementary to the information obtained from experiments with IUPs in the presence of their actual interacting partner(s)) in the future, showing that long-range contacts (partially collapsed states) in certain IUPs prevent more hydrophobic residues from aggregating, or that local non-random coil behaviour facilitates the interaction between binding partners, and so on.

NMR of Interacting IUPs

If the protein is unstructured and known to have a binding partner, one can obtain interesting information using NMR on the mixture of both partners, provided the interacting agents are differencially labelled to allow to distinguish between them. First of all one could try to detect structuration events such as induced folding. If the resonances do not completely disappear upon interaction, due to the immobilization in a high molecular weight complex which results in an increased relaxation, one can use several of the NMR techniques mentioned hereafter to probe for such structuration events. Furthermore, other information can be deduced. Shift differences or changed peak intensities for residues before and after addition of the interacting partner indicate the interaction of those particular residues. Chemical shift changes can be induced by the introduction of kinetics in the fast exchange regime (see fig. 2.1) or by a changed electronic environment. If a signal intensity at an initial position decreases concomitant with the appearance of a new peak exchange is in the slow exchange limit. Decreasing peak intensity on the other hand can be indicative for either induced NMR-unfavorable kinetics or increased relaxation due to the large size of the complex. Chemical

shift changes can be followed in titration experiments, that are capable of providing dissociation constants of the interaction between partners forming a complex.

As a sidenote, it could be mentioned that if the binding partner of an IDP is unknown, there is an alternative way of assessing the folding propensity of an IDP. Secondary structure stabilisers, such as trifluoroethanol (TFE), are widely used as a probe to identify protein regions with propensity to fold [249, 210, 186]. Trimethylamine N-oxide (TMAO) can also be used to this endeavour [425]. TMAO and other osmolytes may fold unstructured proteins due to the osmophobic effect, a solvophobic thermodynamic force, arising from the unfavourable interaction between the osmolyte and the peptide backbone [20, 19, 21, 36]. Although both solutes, TMAO and TFE, act on the peptide backbone, the molecular mechanisms underlying their effects are different. It has long been known that TFE increases the propensity of amino acids to form an α-helix, presumably by strengthening peptide hydrogen bonds in TFE/H_2O mixtures and through favourable interactions of hydrophobic amino acid side chains with TFE [257, 61]. Peptide hydrogen bonds in helices are believed to be stabilised indirectly by weakening the hydrogen bonding of water molecules to the peptide backbone in the coil form [61]. As a result of weakening the hydrophobic interactions within the protein interior, TFE might promote helical structure in most peptides and proteins, even though this helical structure is non-native [47, 123, 258]. In contrast, TMAO increases the driving forces for protein folding due to its solvophobic effect on the backbone, forcing thermodynamically unstable proteins to fold without altering the rules for folding to a native-like conformation [19]. Furthermore, in opposition to TFE solution, the propensities of hydrophobic groups to interact with solvent are essentially the same in water as they are in TMAO solution [403, 416]. Thus, due to the weakening of hydrophobic interactions, the dominant effect of TFE on proteins is protein denaturation accompanied by the preferential formation of α-helices as a result of the strengthening of peptide hydrogen bonds. Contrarily to that, TMAO promotes folding of unfolded proteins by providing an additional force for folding that has no preference for any particular secondary structure [36]. Based on the molecular origin of TMAO-driven protein folding, if biologically relevant structure can be induced in any intrinsically unstructured protein without its target molecule, it is more likely to be induced by solutes (such as TMAO) that have been selected by nature for their ability to fold and stabilise proteins than by alcohols [21].

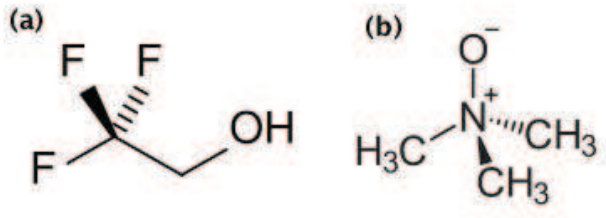


Figure 2.2. The Structures of (a) 2,2,2-trifluoroethanol (TFE) and (b) Trimethylamine N-oxide (TMAO)

It should be pointed out, however, that even if TFE is known to stabilise α-helices more than β-strands, some proteins, as for instance the GCN4

acidic activation domain [406], form little or no α-helix in TFE concentrations as high as 30% and fold as β-sheets at higher TFE concentrations. This observation points the ability of TFE to show propensities. Nevertheless, the general reliability of information derived from studies that make use of TFE is still a matter of debate. Indeed, no systematic study attempting to validate or to refute this approach has been undertaken so far. Definitive conclusions await studies focused on structural comparisons between TFE-induced structures and structures arising from induced folding. However, it could be said that if they are carefully combined with (residual) secondary structure predictors, TFE experiments can be used for the characterisation of the α-helices that might arise in the IDP upon binding.

Obviously, the use TFE or TMAO are not unique to NMR. They have been successfully applied in combination with many other biophysical methods, but are only introduced here because of the emphasis on NMR in this chapter. These combined techniques have also been applied in the work explained in chapter 3.

2.7.1. Chemical Shift Investigation

Chemical shifts report essentially on the local physicochemical environment of the nucleus of interest [164, 375, 299, 368]. Secondary structures such as helix and β-sheet can be readily identified in folded proteins from Hα, Hβ, Cα, Cβ and C' chemical shifts. These values are standardised by subtraction of the appropriate random coil shift, and the secondary structure at a given position in the amino acid sequence is frequently assessed by calculating the chemical shift index (CSI) which combines the data from these nuclei [435, 433]. Interestingly, ^{15}N chemical shifts are not commonly used to predict secondary structure of proteins, although [77] has proposed an elegant way to extract the angle information from these chemical shifts to restrain ϕ,ψ torsion angles in the structure calculation of native proteins using a database approach.

For unstructured proteins, the average chemical shift measured from a broad conformational equilibrium can be interpreted in terms of local conformational propensity of the ensemble. This is because the deviations of experimental chemical shifts from their expected random-coil values are diagnostic for the presence of secondary structure regardless of stability [434], as long as the interconversion rate is fast and the deviation form the random coil chemical shift value is greater than the spectral resolution. Once a random coil shift has been subtracted from the measured value, the so-called secondary chemical shift clearly identifies the presence of transient structure in flexible chains [431, 421]. For example, in the case of C_α and C' spins, successive positive secondary shifts can be interpreted in terms of populations of α-helical segments, while stretches of negative secondary shifts are indicative of nascent β-structure. However, for these unstructured proteins, the difference between the observed and random coil shifts is generally much smaller, and thus a meaningful assessment of the presence and extent of residual secondary structure in such proteins requires accurate reference values obtained under similar experimental conditions.

A first requirement when using chemical shifts to assess secondary structure is that a unique chemical shift referencing is used. Indeed, many entries in the BMRB are incorrectly referenced, including some that claim to follow the proper DSS referencing protocol [454]. However, if incor-

rect referencing were to be applied for disordered proteins, small mistakes could be made in the prediction of the conformational propensities. Besides 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS), that has been imposed as the standard for biomolecular NMR referencing because of its good solubility, pH insensitivity, temperature insensitivity, distant resonance position, inertness and line width [434], 3-(Trimethylsilyl) propionate, sodium salt (TSP) is also of common use. However, the latter is known to be more pH sensitive. Of course, when interpreting ^{13}C NMR data, the chosen reference has to be taken into account. If ^{13}C referencing is done directly via the C-atoms of DSS or TSP, the difference between both standards is 0.12-0.15 ppm. When ^{13}C referencing is done indirectly via the proton signal of the reference molecule, using the ratio published in [432], differences between are of the order of 0.0-0.015 ppm. To perform a ^{13}C chemical shift re-referencing, a method has been implemented based only upon relative $^{13}C_\alpha$ and $^{13}C_\beta$ chemical shifts [265]. The authors have assumed that, for a given amino acid, the relative difference between the $\Delta\delta C_\alpha$ value observed and expected for fully formed α- or β-structure should be similar to the relative difference between the $\Delta\delta C_\beta$ value observed and expected for fully formed α- or β-structure. Since the dependencies of $\Delta\delta C_\alpha$ and $\Delta\delta C_\beta$ values on α- and β-structure are inversely correlated, the tool [265] adjusts the chemical shift referencing offset until those differences are minimised. A similar procedure, by Wang et al. [420], called linear analysis of chemical shifts (LACS) relies on relative $^{13}C_\alpha$ and $^{13}C_\beta$ chemical shifts. A third method, by Wishart and co-workers [422], is not appropriate for disordered proteins since it assumes they exist as random coils; thus the presence of structural propensities would cause improper calculation of referencing offsets.

A more fundamental problem is formed by the choice of the random coil chemical shifts. A meaningful assessment of the presence and extent of residual secondary structure in unfolded proteins requires accurate random coil reference values obtained under similar experimental conditions, since chemical shifts, mainly ^{1}HN, ^{15}N and ^{13}C', are affected by conditions such as temperature and pH. Over time, a number of different sets of random coil shifts have been determined. Some are based on statistical approaches, deriving data from protein databases [164, 430, 434, 454], but these inevitably display some bias. Therefore, it has become common-use to obtain random coil shifts from short unstructured peptides [323, 51, 204, 158, 43, 384, 431, 274, 306, 340, 339]. This variety of random coil models makes it hard to select an appropriate set. Factors like solvent/cosolvent influence, temperature, end effects, chemical shift referencing and nearest neighbour effect must be taken into account. The following points can be taken in consideration: the tables published by Wüthrich and coworkers [51, 323], Thanabal et al. [384] and Glushka and co-workers [158] do not adhere to the standard DSS chemical shift referencing procedure and because of it, give differences in both ^{13}C and ^{15}N shifts of more than 1.5 ppm compared to other sets. They should thus be avoided. Some sets have incorporated nearest neighbour effects [43, 431, 340, 339], which makes them more precise. The identity of the neighbouring amino acid has a significant effect on ^{15}N [305] and amide ^{1}H chemical shifts, while the effect of neighbouring amino acids (mainly the following residue) is smaller on $^1H_\alpha$ and ^{13}C chemical shifts. Finally, some sample properties differ between the different published shift tables. The random coil chemical shift sets of Schwarzinger and colleagues [340, 339] have for example been collected in conditions usually

applied for protein denaturation (8 M urea and pH 2.3), which makes them highly suited for chemical shift investigation in denatured proteins. Other sets were obtained in less extreme conditions, but at different temperatures: 5°C [274], 25°C [431] and 35°C [51, 323, 43]. To conclude, experimental chemical shifts from intrinsically disordered proteins can best be compared to the (most complete) random coil chemical shift set of Wishart et al. [431] and chemical shifts of denatured proteins with the sets Schwarzinger et al. [340, 339]. Several of the main random coil set similarities and differences are summarised in Table 2.1. Finally, fig. 2.3 demonstrates the importance of a correct random coil chemical shift choice.

A final issue in the quantification of the extent to which secondary structure is populated in disordered states is formed by the fact that secondary chemical shifts from different nuclei and residues are not equally sensitive to secondary structure [421]. In order to overcome this problem, a program (SSP) to provide a quantitative measure of the fractional secondary structure propensity, based on differential weighting of the secondary shifts of nuclei in different residues depending on their individual sensitivity to α or β structure [265], has been developed. SSP scores provide the fraction of sampled secondary structure at each residue and can therefore be used in structure calculations of ensembles of conformers. However, by default, they base their calculations on random coils chemical shifts obtained from Zang et al. [454], which is statistically based. Such random coil shifts cannot and do not represent a sufficiently broad sampling of conformational space and therefore should not be regarded as true random coil chemical shifts. The last thing about random coil chemical shifts has obviously not been told yet ...

Random Coil Shift List	S.B./P.B.	Nuclei	Conditions	Referencing
Groß et al., 1988 [164]	S.B.	H (amide & side chain)	n/a	DSS
Wishart et al., 1991 [430]	S.B.	H_N, H_α, C_α, C' & backbone N	n/a	DSS
Wishart and Sykes, 1994 [434]	S.B.	C_α, C'	n/a	DSS
Zhang et al., 2003 [454]	S.B.	backbone H, C & N	n/a	DSS (with error cor)
Richarz and Wüthrich, 1978 [323]	P.B.	C_α, C'	35 °C, pH 7	TSP
Bundi and Wüthrich, 1979 [51]	P.B.	H (amide & side chain)	35 °C, pH 7	TMS, dioxane
Jimenez et al., 1986 [204]	P.B.	N (amide & side chain)	0 and 24 °C, multiple pH, multiple urea concentrations	/
Glushka et al., 1989 [158]	P.B.	backbone N	solvent H_2O, 68 °C, pH 4.6, neighbouring residue effect	NH_3
Braun et al., 1994 [43]	P.B.	backbone N	H_2O, 35 °C, pH 2 and pH 5 neighbouring residue effect	NH_3
Thanabal et al., 1994 [384]	P.B.	all protonated C's	H_2O, acetonitrile & TFE solutions, 298 K, pH 2-3.5	dioxane & TMS
Wishart et al., 1995 [431]	P.B.	H (amide & side chain), amide N, C_α, C_β & C'	1 M urea, 25 °C, pH 5, neighbouring residue effect	NH_3 & DSS
Merutka et al., 1995 [274]	P.B.	H (amide & side chain)	H_2O & TFE solutions, 278-328 K, pH 5	TSP & DSS
Plaxco et al., 1997 [306]	P.B.	H (amide & side chain)	2, 4, 6, 8 M GuHCl, 20 °C, pH 5	DSS
Schwarzinger et al., 2000/2001 [340, 339]	P.B.	amide H, H_α, $H_\beta1$, $H_\beta2$, amide N, C_α, C_β & C'	8 M urea, 293 K, pH 2.3 neighbouring residue effect	DSS

Table 2.1. A summary of the available random coil lists. S.B. and P.B. stands for statitics based and peptide based respectively. DSS (with error cor) refers to the fact that the authors used chemical shift predictions to identify and correct mis-assignments, typographical errors and chemical referencing errors. TMS was not mentioned in the text earlier and stands for tetramethylsilane. GuHCl is short for the commonly used denaturant guanidine hydrochloride. A "/" symbol is given if there are unclarities about the corresponding data.

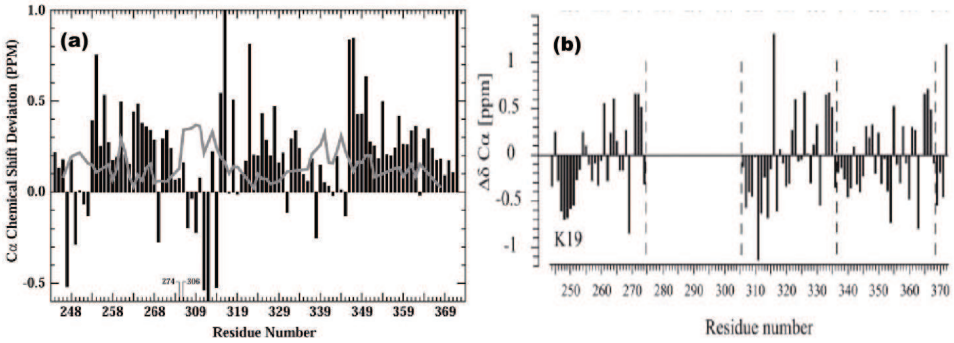


Figure 2.3. A comparison of two independent chemical shift investigations of K19, the repeat domain of Tau in which repeat domain two (out of four) is deleted. (a) comes from a study by Eliezer et al. [117] and represents the secondary C_α chemical shifts of the polypeptide with the spectra recorded at 10°C; protein concentration 1-3 mM, 100 mM NaCl and 10 mM Na_2HPO_4 at pH 6.7. (b) is taken from the study of Mukrasch et al. [285] and corresponds to secondary C_α chemical shifts of K19 with the following sample conditions spectrum recording: 5°C, protein concentration 1 mM, 50 mM phosphate buffer, 1 mM DTT at pH 6.8. Despite similar general conditions, (b) predicts considerably more β-structure propensity than (a). The cause is probably the different set of random coil shifts used in both studies. Respectively the lists by Wishart et al. [431] and Schwarzinger et al. [340] were used in (a) and (b). Since the list of [340] was designed for chemical shift investigation of chemically denatured proteins, (a) approaches probably more the truth than (b).

2.7.2. Pulsed Field Gradient Methods to Measure Translational Diffusion

An easily applied NMR method is the measurement of the translational diffusion coefficient, D, using pulsed field gradients (PFG) [393, 370, 7, 296, 428, 144, 91]. This approach relies on the fact that two protein molecules undergoing translational diffusion at different speeds will differentially sense a field gradient applied to (partially) destroy acquired magnetisation. An array of gradient strength can be used to calculate the translational diffusion coefficient, D [7]. The most common experimental approach to measuring such diffusion coefficients involves the use of a spin echo (SE) [91] (see fig. 2.4), and the resulting PFGSE experiments have been widely applied. The PFGSE signal intensity S_i can be related to the signal intensity obtained in the absence of the gradients, S_0, as follows:

$$S_i = S_0 \, e^{-\alpha^2 D(\Delta - \delta/3)(G/G_{max})^2} \tag{2.2}$$

where $\alpha = G_{max}\gamma\delta$. G_{max} is the maximum field strength of the gradient, γ is the gyromagnetic ratio of the affected nucleus, δ is the duration of the pulsed gradient and Δ is the time between PGF pulses. Eq. 2.2 is the Stejskal-Tanner (ST) equation, and a plot of $ln(S_i/S_0)/\alpha^2\Delta$ against $(G/G_{max})^2$ yields a straight line of gradient $-D$ [370, 378].

The knowledge of D can be used to calculate the hydrodynamic radius. The hydrodynamic radius can then be compared to the expected value based on molecular weight to determine whether the protein is compact and globular or extended and flexible.

It was recently proposed by Baldwin et al. [16] that PFGSE signals do not solely depend on translational diffusion, but equally on rotational diffusion, which leads to a signal intensity (S_i)/gradient dependency slightly

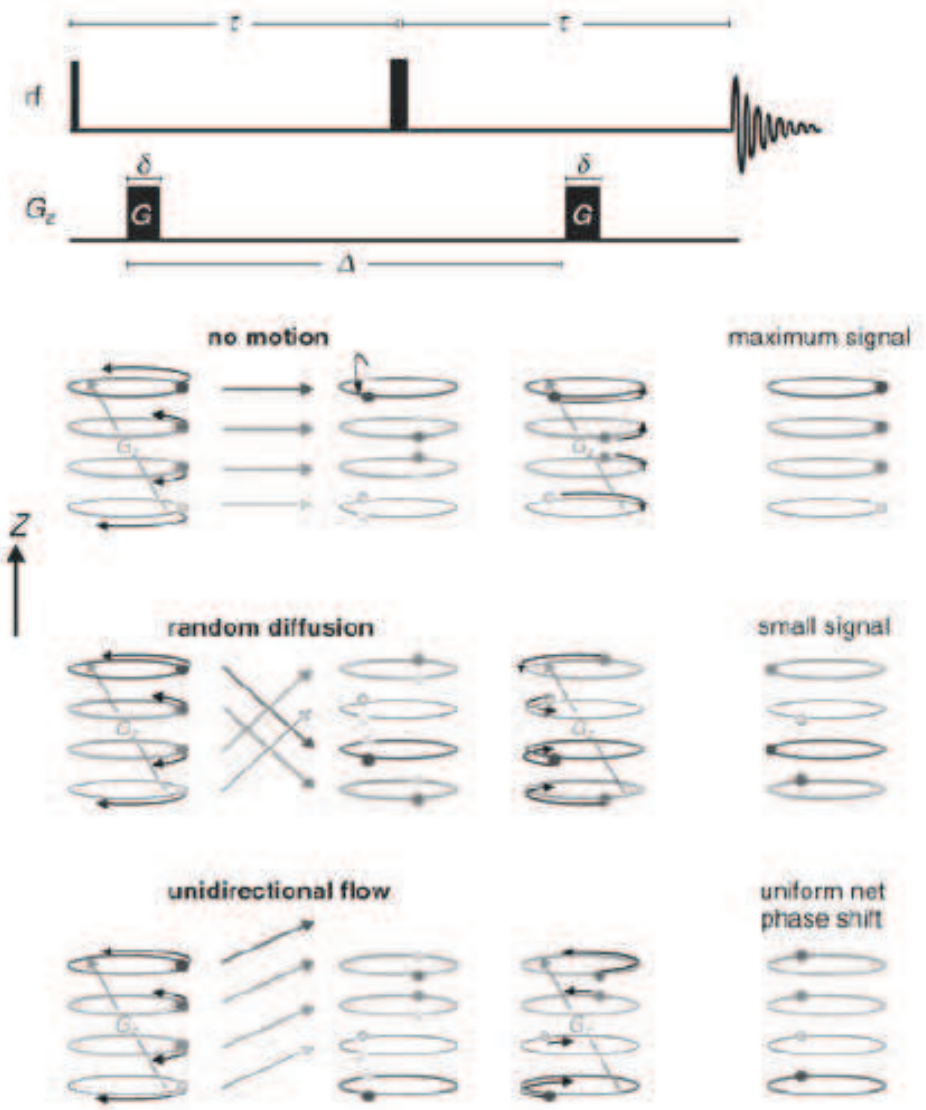


Figure 2.4. Schematic representation of the Stejskal and Tanner pulse sequence and its effect on the spins due to random diffusion (second row) and unidirectional flow (third row). As only coherent magnetisation is observable, random diffusion leads to a loss in signal intensity. In each delay, τ, a gradient pulse of duration δ and magnitude G is inserted. The separation between the gradient pulses is denoted by Δ. Only the precession due to gradients is considered in the rotating reference frame rotating at ω_0. Image taken from [91].

more complicated than given by eq. 2.2. Rotational motion has been shown to increase the apparent diffusion of a nuclear spin. However, since this effect only become important for bodies (proteins) having dimensions of $> 10^{-6}m$, much larger than typical protein dimensions (even disordered ones), the ST equation can be applied with success in most of the cases. Deviations are only observed for large systems, for example, protein amyloid fibrils [15].

Empirical relationships have been established between the hydrodynamic radius determined using PFG translational diffusion measurements, and the number of residues in the polypeptide chain for natively folded proteins and highly denatured states [428]. This study provided evidence for significant coupling between local and global features of the conformational ensembles

adopted by disordered polypeptides. As expected, the hydrodynamic radius of the polypeptide was dependent on the level of persistent secondary structure or the presence of hydrophobic clusters. Many diffusion experiments have been conducted on intrinsically disordered and unfolded proteins ever since (e.g. [218, 279, 82]).

2.7.3. Nuclear Overhauser Effect Spectroscopy (NOESY)

As in folded proteins, the use of nuclear Overhauser effect spectroscopy (NOESY) provides valuable information on secondary structure formation and, to a lower extent, possible long-range interactions in intrinsically disordered proteins. Since most IUPs have been found to interact with binding partners through the formation of an α-helix, one will mostly find back the α-helix typical NOEs in the zones the IUP uses for interacting with other agents. Other regions exhibit less consistent NOEs. In these disordered states, the NOE is difficult to interpret quantitatively because of the ubiquitous conformational averaging. To worsen the situation even more, this conformational averaging is likely to introduce a bias toward those conformations with the shortest contact distance, even though their populations within the ensemble may be quite small. This behaviour is obtained because of the r^{-6} relation between distance and NOE intensity. Indeed, suppose two proton spins in a molecule are at 2Å during 5% of the time and at 20Å the rest of the time, this leads to an effective distance between the spins of 4.3Å($0.05 \times 1/2^6 + 0.95 \times 1/20^6 = 1/4.3^6$), which can still be measured but does not accurately describe the situation. Despite this feature, the $d_{\alpha N}$(i, i + 1), d_{NN}(i, i + 1), and $d_{\beta N}$(i, i + 1) NOEs between sequential amino acid residues can provide information on the local polypeptide backbone conformational preferences, i.e., on the relative population of dihedral angles in the α and β regions of ϕ,ψ space). Such NOEs constitute a valuable supplement to chemical shifts in the analysis of backbone conformational preferences. But, it is important to note that the $d_{\alpha N}$(i, i + 1) and d_{NN}(i, i + 1) sequential NOEs provide information only on local ϕ and ψ dihedral angle preferences and do not by themselves indicate the presence of folded conformations. Definitive identification of folded elements of secondary structure requires additional information from medium-range NOEs [e.g., $d_{\alpha N}$(i, i + 2), $d_{\alpha N}$(i, i + 3), $d_{\alpha\beta}$(i , i + 3) NOE connectivities]. So, if only $d_{\alpha N}$(i, i + 1) and d_{NN}(i, i + 1) sequential NOEs are observed (together with the observation of e.g. the small size of the secondary $^{13}C_\alpha$ shifts), this indicates clearly that conformational averaging occurs over both the α-helical and β regions of ϕ,ψ space. An example of this behaviour can be found in [119]. Some α-helix-characteristic NOEs for are represented in fig. 2.5.

Long-range NOEs, indicative of transient tertiary structure, are very difficult to observe in IDPs. This is likely due to the fact that either the population of the transiently structured forms is too low, or the ensemble containing them is too heterogeneous. Might they still be observed, they provide a definite indication for a close proximity, at least in some structures of the ensemble. For IDPs, long range distance information can be probed better by the use of covalently attached paramagnetic nitroxide spin labels (see below).

The most common way for obtaining NOE connectivities in unfolded proteins is with ^{15}N-edited 3D NOESY-HSQC spectra. However, in some cases, the severe overlap of the ^{1}H resonances can reduce the utility of these

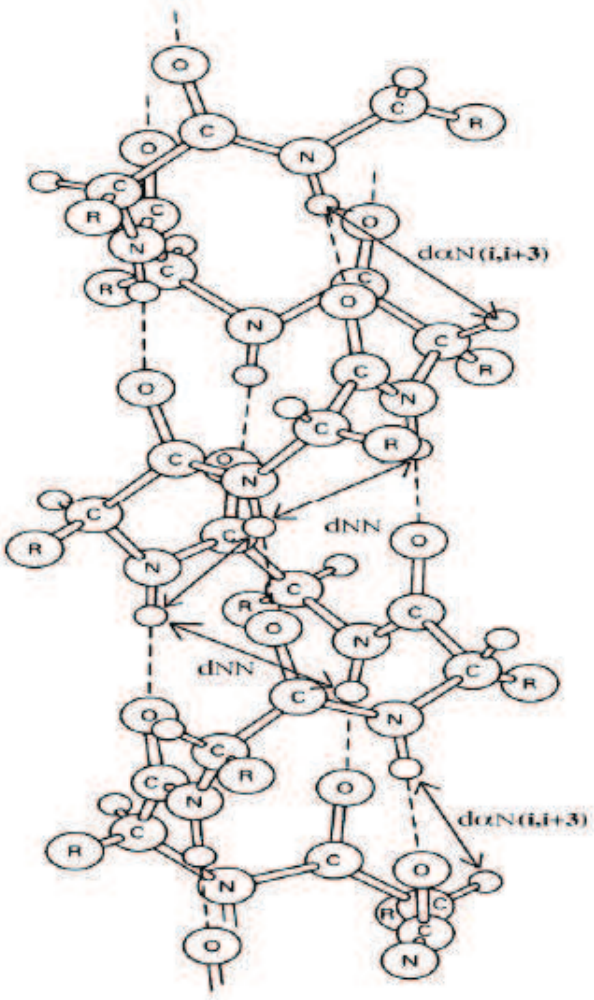


Figure 2.5. Some α-helix typical NOEs. Especially the $d_{\alpha N}(i, i + 3)$ NOE connectivities are very characteristic for this secondary structure element. Figure taken from slide show of irretraceable author.

experiments and seriously limit the number of NOE peaks that can be assigned. Improved resolution can be achieved for NOEs between NH groups by labelling both protons involved with their attached 15N frequencies, using the 3D/4D 15N-HSQC-NOESY-HSQC experiment [456, 143, 195]. Although some long-range interactions can be observed in favourable cases [277], this experiment primarily provides information about backbone conformational preferences. Detailed characterisation of structured conformers (for example obtained in TFE) requires identification and assignment of NOE connectivities involving side-chain protons, which are of course poorly dispersed in disordered states. To overcome this problem, Kay and co-workers have developed an elegant series of triple-resonance based NOESY experiments that exploit the dispersion of the ^{15}N and ^{13}CO resonances to resolve ambiguities in the aliphatic ^{1}H and ^{13}C chemical shifts [456, 458]. These experiments transfer magnetisation from aliphatic protons to the ^{15}N or ^{13}C' nuclei before or after the NOE mixing period, so that the NOEs involving aliphatic protons are observed at well-resolved nuclei with minimal resonance overlap.

2.7.4. Hydrogen/Deuterium Exchange (HDX)

Monitoring the exchange rate between main chain amides and solvent hydrogens as a method to study the structure of proteins has seen increased usage over the past 40 years. Hydrogen-deuterium exchange (HDX) rates are dependent on thermodynamics and dynamic behaviour and thus yield information regarding the structural stability of the protein under study. In partially disordered proteins, amide proton exchange rates tend to be very rapid in unstructured regions, but slower for amide protons that are protected from exchange by structuration. HDX can be followed up by recording multiple consecutive HSQC spectra in which signals corresponding to residues that change their amide hydrogen faster for deuterium diminishes faster in intensity than others. Application of HDX in combination with NMR can hence yield data to near single-amide resolution. To this day, concerning IUPs, HDX-NMR has been mainly used to characterise structure and flexibility of unstructured proteins forming amyloid fibrils common to several disease pathogenesises [184, 256].

2.7.5. Relaxation Methods

NMR spin relaxation methods can be used to assess the dynamics and mobility of proteins on different timescales [294, 214, 41]. The degree of protein flexibility and disorder as well as the changes in protein flexibility can be assessed by NMR spin relaxation methods at individual residues within the protein. Furthermore, kinetic processes in the ms time regime may be directly investigated to extract the rates of conformational interconversion, ligand binding, and protein folding processes.

A variety of heteronuclear relaxation experiments now exist to monitor ^{1}H, ^{2}H, ^{13}C and ^{15}N nuclei allowing access to motional information at multiple locations within the protein backbone and side chain. The most commonly employed NMR relaxation experiments monitor the ^{15}N amide resonances of the protein backbone. The advantages of ^{15}N relaxation measurements for the study of proteins are the low cost of the ^{15}N isotope, the absence of complex scalar homonuclear and heteronuclear couplings and the absence of significant remote interactions with neighbouring spins. Indeed, in solution, the relaxation rates of the isolated ^{1}H-^{15}N heteronuclear atoms are dominated by the dipolar interaction with the attached proton and by the chemical shift anisotropy interaction[1]. In addition, the interpretation and methodology for the application of ^{15}N spin relaxation experiments to proteins is well established [363, 125]. A downside is that amide protons often undergo fast exchange with solvent in unstructured proteins, which can reduce the sensibility and limit the range of experimental conditions employed. In particular, the NH-NOE experiment provides a fast and easy way to interpret diagnostic for the presence of intrinsic protein disorder. The sign of the NH-NOE resonance is sensitive to the rotational correlation time and is positive for N-H bond vectors with a long rational correlation time (>1-10 ns) and negative for N-H bond vectors with a short rotational correlation time (<0.1-1 ns).

More generally, pulse sequences have been developed for collecting several types of ^{15}N related relaxation rates [125]. The most commonly measured

[1] This entails the relaxation by fluctuating local magnetic fields caused by molecular electron currents induced by he external magnetic field.

relaxation rates are $R_N(N_z)$ (longitudinal ^{15}N relaxation rate), $R_N(N_{x,y})$ (transverse ^{15}N relaxation rate) and $R_N(H_z \rightarrow N_z)$ (heteronuclear NOE data) and the most widely applied pulse sequences use two dimensional heteronuclear correlation spectra to allow observation of the relaxation effects of the given nucleus indirectly through the attached proton.

Measured NMR ^{15}N relaxation parameters are related to the motions of the ^{1}H-^{15}N vector through their spectral densities at the five frequencies: 0, ω_N, $\omega_{H^N} - \omega_N$, ω_{H^N} and $\omega_{H^N} + \omega_N$. The spectral density function describes the frequency spectrum of rotational and intramolecular motions of the N-H bond vector relative to the external magnetic field and is derived from the Fourier transform of the spherical harmonics describing the rotational motions (the autocorrelation function $G(t)$). It gives the proportions of the energy used for the motions at the corresponding frequencies. A characteristic feature of the spectral density function is that the integral over the whole frequency range is a constant. As a result, the density of the frequency of motions is shifted towards lower frequencies for slowly tumbling molecules, i.e. larger-sized macromolecules. Knowing the spectral density function implies a deeper knowledge on the molecular motion and their timescales of a molecule.

Using the reduced density matrix formalism [127, 197, 233], values of the spectral density function can be derived from the relaxation parameters at three frequencies: $J(0)$, $J(\omega_N)$ and $\langle J(\omega_H)\rangle$. It is observed that high-frequency spectral density terms are approximately equal in magnitude ($J(\omega_H \pm \omega_N) \simeq J(\omega_H)$). Hence, these three terms have been set equal to an average value ($\langle J(\omega_H)\rangle$) close to $J(\omega_H + \omega_N)$ [233] or to three slightly different values ($J(0.87\omega_H)$, $J(0.92\omega_H)$ and $J(0.96\omega_H)$) that can be calculated one out of the other since it is assumed that $dJ(\omega)/d\omega^2$ is linear for high frequencies [127]. The calculation is done using:

$$\begin{pmatrix} R_N(N_z) \\ R_N(N_{x,y}) \\ R_N(H_z \rightarrow N_z) \end{pmatrix} = \begin{pmatrix} 0 & E & 7A \\ \frac{2E}{3} & \frac{E}{2} & \frac{13A}{2} \\ 0 & 0 & 5A \end{pmatrix} \begin{pmatrix} J(0) \\ J(\omega_N) \\ \langle J(\omega_H)\rangle \end{pmatrix} \tag{2.3}$$

where $A = \gamma_{H^N}^2 \gamma_N^2 \hbar^2 / 4 r_{NH^N}^6$, $B = \Delta^2 \omega_N^2 / 3$ and $E = 3A + B$, with Δ the CSA value (difference between the parallel and perpendicular components of the assumed axially symmetric chemical shift tensor), and r_{NH^N} the internuclear ^{15}N-^{1}H bond distance. The right-hand column vector consists of the unknown power spectral density function $J(\omega)$. Hence, the set of three relaxation rates allows an approximate calculation of $J(\omega)$ (in the absence of chemical exchange contributions).

The direct analysis of the spectral densities can provide a picture of the distribution of the frequencies of N-H bond motions along the protein backbone. Since the area under the spectral density function $J(\omega)$ is constant, a decrease of the $J(0)$ value is compensated by an increase of the value of $J(\omega)$ at high frequencies. The more flexible the protein, the lower the values of $J(0)$. Largely unstructured proteins containing almost no residual secondary structure exhibit a flat profile of the spectral density function along the frequency dimension. If a protein contains some partially folded regions, variations of the spectral density values along the sequence are observed. Fig. 2.6 shows spectral density values for a protein domain containing both a structured and flexible region. In order to obtain a crude idea on the timescale of the motions leading to the observed relaxations, experimentally

derived values of $J(0)$, $J(\omega_N)$ and $\langle J(\omega_H)\rangle$ have been compared to theoretical values of these variables corresponding to a unique isotropic motion and calculated (for a certain magnetic field strength) at different values of correlation time (see fig. 2.7) [290].

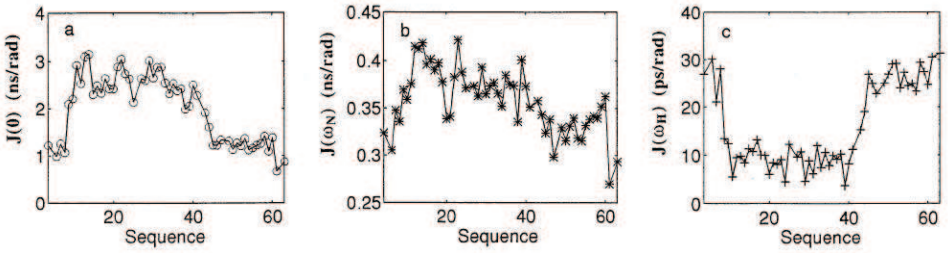


Figure 2.6. Values of the spectral density functions versus the protein sequence at the frequencies of (a) 0, (b) ω_N and (c) ω_H for the DNA binding domain (residues 1-65) of the yeast transcriptional activator GAL4. The results demonstrate well the partially folded nature of this protein domain. Residues 9-40 of this protein segment form a compact, globular metal-binding domain while the first nine residues and the residues C-terminal to Ser41 (42-65) are disordered. This transition between ordered and disordered is clearly visible in the figure. Picture taken from [233].

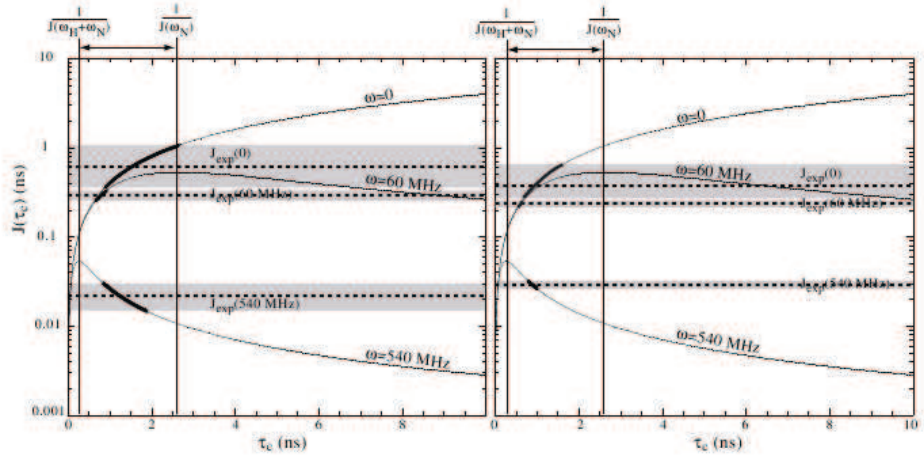


Figure 2.7. Results of a relaxation study on the isolated, partially unstructured D2 domain of annexin I [290]. Theoretical values of the spectral densities $J(0)$, $J(\omega_N)$ and $\langle J(\omega_H)\rangle$ at 600 MHz as a function of the correlation time τ_c assuming a unique isotropic motion (thin continuous curves) compared with the experimental spectral densities obtained for the N-terminal segment 5-51 (left panel) and the C-terminal 52-68 (right panel) of the D2 domain at 600 MHz. The dispersion of experimental spectral density values is represented by horizontal grey bars and the mean values as horizontal dotted lines. The segments of the theoretical curves that match the experimental values are thickened. This figure indicates that the observed relaxation rates for this protein are mainly due to motions between 180 ps and 2.5 ns. It should be noticed that the ranges of τ_c do not strictly coincide for all spectral densities, indicating that the dynamics of the domain cannot be described by a unique isotropic motion. Picture taken from [290].

However, extraction of more pictorial information about motions, such as correlation times and amplitudes, requires the use of models and thus the formulation of hypotheses on motions. The Lipari-Szabo model-free approach [243, 73] states that the power spectral density function for an

isotropic molecule can be described as:

$$J(\omega) = \frac{2}{5}\left\{\frac{S^2\tau_m}{1+\omega^2\tau_m^2} + \frac{(1-S^2)\tau}{1+\omega^2\tau^2}\right\} \tag{2.4}$$

in which τ_m is the isotropic rotational correlation time of the molecule, $\tau = \tau_m\tau_e/(\tau_m+\tau_e)$ where τ_e describes internal ps motional processes and S^2 is the square of the generalised order parameter describing the amplitude of the internal motions. The previously obtained spectral density values can be fitted to eq. 2.4, allowing the direct extraction of overall rotational and internal correlation times and order parameters for each residue. It is unclear how valid the assumption of isotropic rotation is for intrinsically disordered proteins, as anisotropic structures also populate the conformational ensemble. Regardless of this limitation, the model free analysis of some intrinsically disordered proteins has ultimately provided a useful qualitative picture of the heterogeneity in rotational diffusion observed in these proteins (e.g. [4, 48, 83, 49]). As a first order approximation for unstructured proteins, τ_e is ignored and a global τ_m is never optimised. The obtained τ_m can then be treated as an "effective local" correlation time for each residue. Changes in S^2 have been used to estimate changes in conformational entropy due to changes in ns-ps bond vector motions during the protein folding that is coupled to binding in intrinsically disordered proteins [447, 448, 83, 442, 41].

Although the applications and theory for characterisation of protein dynamics from ^{13}C nuclear spin relaxation are not as well established or as widely applied as ^{15}N spectroscopic methods, it is particularly well suited for investigation of conformational dynamics of unstructured proteins. Proteins and peptides can be selectively enriched at specific sites ^{13}C via peptide synthesis or selective labelling strategies to avoid the complications encountered from ^{13}C-^{13}C scalar and dipolar couplings in uniformly labelled samples. The great advantage is that unlike amide protons, carbon attached protons are stable, preventing fast exchange with solvent in unstructured proteins that can reduce the sensitivity of ^{15}N spectroscopy and limits the range of experimental conditions employed.

To conclude, relaxation measurements have been a useful tool in several studies of motion in disordered proteins and show that the amount of secondary structure in the protein is inversely related to the level of internal dynamics [88, 233, 219, 395, 163, 287, 388].

2.7.6. Interconversion Rate Measures - Exchange spectroscopy

Exchange spectroscopy can be used for assessing two-state structural changes in the slow regime ($k \ll |\Omega_\Delta/2|$). As pointed out previously, intrinsically disordered proteins for example often undergo induced structuration (such as α-helix formations) or peptidyl-prolyl cis/trans isomerisation that correspond to this kind of exchange. Exchange spectroscopy is based on the very simple principle that, during a certain mixing interval (τ_m), magnetisation can be transported from a system in one conformation to the system in the other conformation if the protein molecule in question makes the structural change between the two conformations during that interval. Several pulse schemes have been developed to record these exchanges peaks of which the proton-detected heteronuclear correlation experiments are the most useful. The ones most adapted for the study of disordered proteins have again taken care of minimising the water signal [126]. The exchange

can pass through two-spin heteronuclear order (I_zS_z) or through the heteronuclear longitudinal magnetisation (S_z) [126]. Longitudinal nitrogen magnetisation exchange is most common and will be discussed in the following. This longitudinal magnetisation transfer is depicted in fig. 2.9. In heteronuclear coherence spectra, the exchange process generates, besides the 'auto' peaks (I_{aa} and I_{bb}; corresponding to individual conformations a and b not undergoing chemical exchange), the so-called exchange peaks (I_{ab} and I_{ba}) of which the intensity depends on the exchange rate k between the two conformations of the protein and on the length of the mixing interval.

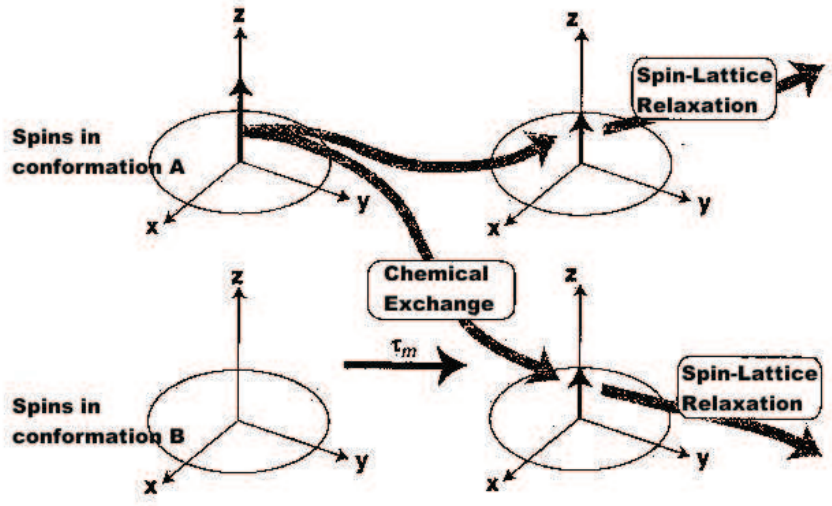


Figure 2.8. Physical processes during the mixing interval of a 2D exchange experiment. Picture taken from [235].

Of course, after being exchanged, the magnetisation is again subject to (longitudinal) relaxation, which will also be measured. This leads to a typical buildup-decay course of exchange peak intensities with growing mixing times. Longitudinal relaxation and chemical exchange rates can be extracted by fitting theoretical exchange/decay curves to the measured data. The theoretical curves can easily be calculated from the Bloch equations for a system undergoing chemical exchange between two sites [173, 172, 268, 191] and are given by:

$$I_{aa}(T) = I_a(0)(-(\lambda_2 - c_{11})e^{-\lambda_1 T} + (\lambda_1 - c_{11})e^{-\lambda_2 T})/(\lambda_1 - \lambda_2) \qquad (2.5)$$

$$I_{bb}(T) = I_b(0)(-(\lambda_2 - c_{22})e^{-\lambda_1 T} + (\lambda_1 - c_{22})e^{-\lambda_2 T})/(\lambda_1 - \lambda_2) \qquad (2.6)$$

$$I_{ab}(T) = I_a(0)(c_{21}e^{-\lambda_1 T} - c_{21}e^{-\lambda_2 T})/(\lambda_1 - \lambda_2) \qquad (2.7)$$

$$I_{ba}(T) = I_b(0)(c_{12}e^{-\lambda_1 T} - c_{12}e^{-\lambda_2 T})/(\lambda_1 - \lambda_2) \qquad (2.8)$$

where $\lambda_{1,2} = 1/2\{(c_{11} + c_{22}) \pm [(c_{11} - c_{22})^2 + 4k_{ab}k_{ba}]^{1/2}\}$, $c_{11} = R_a + k_{ab}$, $c_{12} = -k_{ba}$, $c_{21} = -k_{ab}$, $c_{22} = R_b + k_{ba}$, R_a and R_b are the longitudinal relaxation rates of magnetisation in sites a and b, $I_a(0)$ and $I_b(0)$ denote the amount of longitudinal nitrogen magnetisation in sites a and b, and k_{ab} and k_{ba} are the exchange rates for magnetisation converting from site a to b and b to a, respectively. A least-square fitting procedure can be employed to extract the longitudinal decay and chemical exchange rates by fitting the

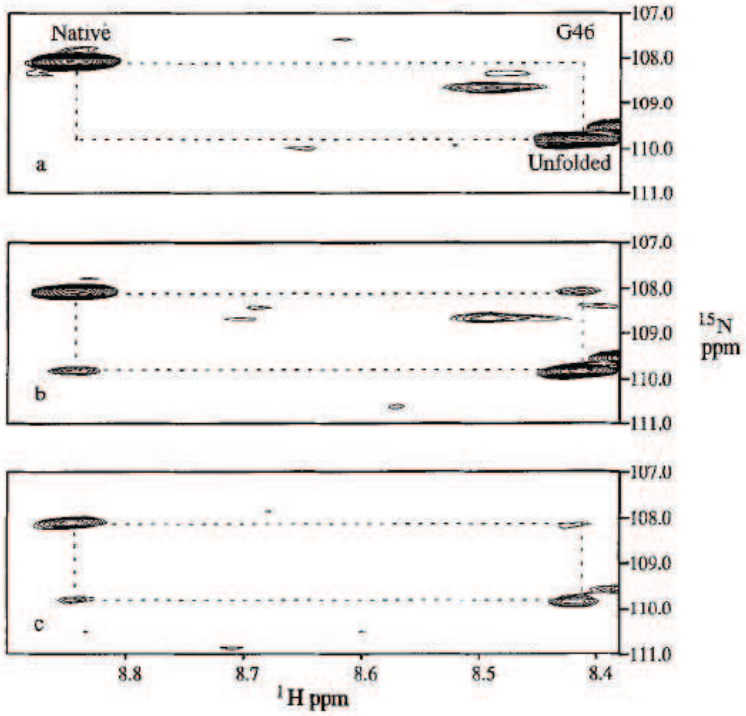


Figure 2.9. Selected regions of spectra containing peaks of Gly46 from the N-terminal SH3 domain of drk that exchanges between a folded and unfolded form. The spectra correspond to the following delays T: (a) 0.011 s; (b) 0.155 s; and (c) 0.843 s. The image is taken from [126].

expressions given above to the measured intensities of auto and exchange peaks. The volumes of the auto peaks recorded in an experiment with zero mixing time provides a measure of the equilibrium constants at each site of the molecule:

$$K_{eq} = \frac{k_{ba}}{k_{ab}} = \frac{p_a}{p_b} = \frac{I_a(0)}{I_b(0)} \tag{2.9}$$

p_a and p_b in this equation are the populations in either conformation. The chemical exchange rate is the sum of the individual rate constants ($k_{ex} = k_{ab} + k_{ba}$). If one is solely interested in determining k_{ex} and, moreover, the longitudinal relaxation rates are known to be identical in both conformations ($R_a = R_b$), the fitting procedure can be greatly simplified by concentrating on the ratio $I_{ab}(T)/I_{aa}(T)$. Indeed, dividing eq. 2.7 by eq. 2.5 and multiplying both numerator and denominator with $exp(\lambda_1)$ gives:

$$\frac{I_{ab}(T)}{I_{aa}(T)} = \frac{c_{21} - c_{21}e^{(\lambda_1-\lambda_2)T}}{-(\lambda_2 - c_{11}) + (\lambda_1 - c_{11})e^{(\lambda_1-\lambda_2)T}} \tag{2.10}$$

Then substituting the above values for $c_{11,21}$ and $\lambda_{1,2}$ results in:

$$\frac{I_{ab}(T)}{I_{aa}(T)} = \frac{-k_{ab} + k_{ab}e^{(k_{ab}+k_{ba})T}}{k_{ab} + k_{ba}e^{(k_{ab}+k_{ba})T}} \tag{2.11}$$

which, considering eq. 2.9 and the definition for k_{ex} gives:

$$\frac{I_{ab}(T)}{I_{aa}(T)} = \frac{-p_b + p_b e^{k_{ex}T}}{p_b + p_a e^{k_{ex}T}} \tag{2.12}$$

Using a similar reasoning one can also obtain:

$$\frac{I_{ba}(T)}{I_{bb}(T)} = \frac{-p_a + p_a e^{k_{ex}T}}{p_a + p_b e^{k_{ex}T}} \tag{2.13}$$

These equations will be used in chapter 3 to determine PPIase activities of cyclophilins on disordered proteins.

2.7.7. NMR Titrations

If an intrinsically unstructured protein interacts through the formation of a bi-molecular interaction complex, a straighforward follow-up of the gradual ^{1}H,^{15}N-HSQC chemical shift shifts, measured in a titration experiment as follows:

$$\delta\Delta = |\delta\Delta(^{1}HN)| + 0.2|\delta\Delta(^{15}N)| \tag{2.14}$$

allows for a calculation of the dissociation constant (K_D) of the interaction complex. The weighting factor of 0.2 for the nitrogen chemical shift changes is necessary to obtain a more or less equal importance of changes in both x- and y-dimensions.

Indeed, for formation/dissociation of a bi-molecular interaction complex ($A + B \rightleftharpoons AB$), the dissociation constant is given by:

$$K_D = \frac{[A][B]}{[AB]} = \frac{([A]_0 - [AB])([B]_0 - [AB])}{[AB]} \tag{2.15}$$

where $[A]_0$ and $[B]_0$ are the initial concentrations of A and B. Solving eq. 2.15 for $[AB]$ gives us:

$$[AB] = \frac{1}{2}\left([A]_0 + [B]_0 + K_D - \sqrt{([A]_0 + [B]_0 + K_D)^2 - 4[A]_0[B]_0}\right) \tag{2.16}$$

The chemical shift observed for partner A residues in the HSQC spectra are rationalised by the following equation:

$$\Delta = \Delta_{free}\frac{[A]}{[A] + [AB]} + \Delta_{bonded}\frac{[AB]}{[A] + [AB]} \tag{2.17}$$

Since $\delta\Delta = \Delta - \Delta_{free}$ and $\delta\Delta_{max} = \Delta_{bonded} - \Delta_{free}$, eq. 2.16 can be written as:

$$\delta\Delta = \frac{\delta\Delta_{max}}{2}\left(1 + X + \frac{K_D}{[A]_0} - \sqrt{\left(1 + X + \frac{K_D}{[A]_0}\right)^2 - 4X}\right) \tag{2.18}$$

where $X = [A]_0/[B]_0$. Fitting this equation to the observed chemical shift changes (eq. 2.14) provides one with a value for K_D. The smaller the found back values K_D, the stronger the two binding partners interact. Since the dissociation constant is also defined as $K_D = k_{off}/k_{on}$, with k_{off} and k_{on} the dissociation and association rate constants respectively, smaller K_Ds also be interpreted as smaller k_{off}s (expressed in s^{-1}) and longer residence times.

2.7.8. Backbone $^3J_{HN,H\alpha}$ Coupling Constants

ϕ and ψ conformations in solution can be probed experimentally by NMR via spin-spin coupling constants, which in the case of ^3J-coupling constants can be related to specific torsion angles by Karplus relationships [211] (see fig. 2.10). When, as is the case for unstructured proteins, multiple conformations are adopted and there is a rapid interconversion between them, the experimental coupling constant values will be averaged over the contributing conformers. This averaging greatly complicates the interpretation of coupling constants in terms of specific torsion angles.

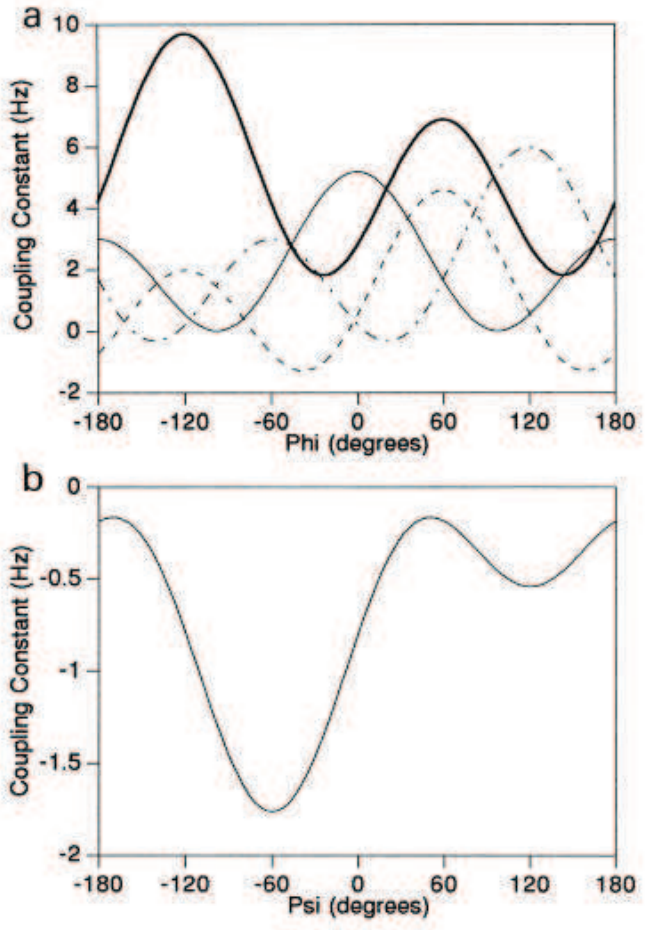


Figure 2.10. Graphs showing the relationship, predicted by the Karplus equation, between (a) the ϕ torsion angle and $^3J_{HN,H\alpha}$ (bold line), $^3J_{HN,CO}$ (thin line), $^3J_{HN,C\beta}$ (dot-dash line) and $^3J_{C_{i-1},H\alpha_i}$ (broken line) coupling constants. (b) the ψ torsion angle and the $^3J_{N_i,H\alpha_{i-1}}$ coupling constant. The graphs in (a) were calculated using the Karplus equation ($^3J = Acos^2\theta + Bcos\theta + C$) [211] with A=6.4, B=-1.4, C=1.9 and $\theta = \phi - 60°$ for $^3J_{HN,H\alpha}$ [298], A=4.0, B=1.1, C=0.07 and $\theta = \phi$ for $^3J_{HN,CO}$ [417], A=4.7, B=-1.5, C=-0.2 and $\theta = \phi + 60°$ for $^3J_{HN,C\beta}$ [58] and A=4.5, B=-1.3, C=-1.2 and $\theta = \phi + 120°$ for $^3J_{C_{i-1},H\alpha_i}$ [58]. The $^3J_{N_i,H\alpha_{i-1}}$ coupling constants (b) were calculated using the relationship $^3J = -0.88cos^2(\psi + 60°) - 0.61cos(\psi + 60°) - 0.27$ [417]. The image is taken from [365].

Strategies have been proposed to describe the backbone conformations sampled by unfolded states of proteins based on experimentally observed $^3J_{HN,H\alpha}$ coupling constants. By extracting the distributions of backbone dihedral angles from a database of high-resolution protein crystal structures, coupling constants for "statistical" coil structures of a given sequence can

be predicted. Originally this lead to per-amino acid type characteristic $^3J_{HN,H\alpha}$ coupling constant values [344, 365, 133]. This can be rationalised in terms of the steric properties and hydrogen bonding characteristics of the amino acid side-chain concerned. However, it was later recognised that second-order effects also need to be accounted for. I.e., subtle variations in coupling constants were observed caused by the nature of the residue in position (i-1). The $^3J_{HN,H\alpha}$ values are increased when this neighbouring residue is one with a β-branched or aromatic residue [302]. Based on this framework, deviations of the observed NMR parameters for a protein in the unfolded state from the predicted coil values indicate the presence of residual structure.

If $^3J_{HN,H\alpha}$ values are smaller than for random coil, this indicates a preference for dihedral angles in the α-region of (ϕ,ψ) space. Of course, these tendencies should be in agreement with the chemical shift data. If $^3J_{HN,H\alpha}$ coupling constants are significantly larger than the random coil values, a propensity for these residues to populate backbone dihedral angles in the β-region of (ϕ,ψ) space is suggested (again if this is in agreement with the chemical shift data). It should be noted, however, that deviations of $^3J_{HN,H\alpha}$ from random coil values are usually very small and it is difficult to discern subtle variations in backbone conformational propensities from coupling constants alone. Despite this, $^3J_{HN,H\alpha}$ coupling values have been used in a few thorough structural characterisation studies of intrinsically unstructured proteins [69, 13, 62, 29]. The coupling constant values obtained for regions undergoing major structuration upon binding are obviously more easily interpreted, as the deviations are much more pronounced.

Although a variety of methods have been introduced for obtaining the $^3J_{HN,H\alpha}$ coupling constants, only the HNHA experiment [414] has found general support in protein science because of many advantages. Basically, it does not rely on the measurement of multiplet splittings, which is problematic when the line widths become larger than the J coupling, but instead relies on the measurement of the diagonal-peak to cross-peak intensity ratio in a 3D ^{15}N-separated quantitative J-correlation spectrum. In this experiment, an HMQC signal is split in two signals (a negative H^N-H_α cross peak and a positive H^N diagonal peak) appearing at δH^N and δH_α in the third (indirect) proton dimension. Their intensity ratio relates as:

$$\frac{S_{cross}}{S_{diag}} = -tan^2(2\pi J_{HH}\zeta) \tag{2.19}$$

where ζ is half of the duration of the de- and rephasing delays during which the homonuclear H^N-H_α J-coupling is active.

2.7.9. Paramagnetic Relaxation Enhancement (PRE)

While chemical shifts, and short and medium-range NOEs provide valuable insights into the secondary structural properties of the polypeptide backbone in unfolded and partly folded proteins, it generally proved difficult to observe long-range NOEs for disordered or partly ordered states. To circumvent this problem, several groups have used site-specific nitroxide spin labelling to probe transient long-range interaction in disordered proteins [155, 156, 452, 381, 237, 89, 253]. Because this PRE method relies on stronger interactions (the gyromagnetic ratio of the electron spin is 660 times larger than that of the proton and enters quadratically in all formulae

describing relaxation), it allows to measure ensemble-averaged, transient contacts over distances as large as 20 Å.

A nitroxide spin-label is attached to a region of a protein, and in its oxidised (paramagnetic) state, it has a free electron that enhances relaxation of heteronuclear coherences observed in a ^{1}H-^{15}N correlation (HSQC) experiment. No effects are observed when the label is in a reduced (diamagnetic) state. The nitroxide spin-label is most often accomplished by site-directed mutagenesis, substituting a single amino acid residue with cysteine to provide a site for coupling of an iodoacetamide derivative of the spin label. Care must be taken to select sites for labelling that are unlikely to perturb the residual structure of the unstructured protein. Relaxation enhancement scales as r^{-6} with label proximity, allowing a crude (or more precise in combination with simulations) interpretation of long-range distances and ensemble tendencies from ratios of HSQC cross-peak intensities acquired for both oxidised and reduced label states. Alternatively, by inserting spin labels at multiple sites, a sufficient number of long-range distance constraints can be obtained to allow determination of the global topology [155, 156].

However, due to the discrete position of spin-labels in PRE experiments, more distance restraints are available at spin-labelling sites relative to unlabelled sequence locations. This can potentially be wrongly interpreted as a structural compaction surrounding the site of labelling more than in the other regions. When using the PRE distances as restraints in MD simulations it was shown that this wrong interpretation effect is minimised and that the zones of closest proximity do not necessarily coincide with the labelling zones [89, 253]. The structure ensembles obtained with such MD simulation can afterwards be validated (to a minor extent) by comparing calculated gyration radii (R_g) with experimentally observed Stokes radii (R_s), where the relationship between the two values can be found in [241][2].

2.7.10. Residual Dipolar Couplings (RDCs)

Also the use of residual dipolar couplings (RDCs) has been introduced for measurements on unstructured proteins. Dipolar couplings contain information on the orientation of internuclear vectors and have become and important adjunct to traditional structural constraints in refinement of NMR structures of globular proteins [307, 23]. RDCs are measured by weakly aligning a macromolecule in slightly anisotropic nematic liquid crystalline media, such as detergent bicelles or filamentous phages, or in anisotropically compressed gels (e.g. strained polyacrylamide gels) that interfere with the isotropic tumbling of the macromolecule in solution [387, 177, 334, 396]. The small degree of alignment resulting from the anisotropic environment leads to incomplete averaging of the dipolar coupling between magnetic nuclei close in space. The magnitude of the RDC is dependent on the orientation of an internuclear vector relative to the alignment tensor of the protein as a whole. In other words, bond vectors such as ^{15}N-^{1}H or ^{13}C-^{1}H can be oriented relative to a global alignment tensor fixed in the molecular frame, regardless of their location in the molecule [387]. Alternatively, the information can be interpreted in terms of angular relations between pairs of bond vectors that are independent of the intervening distance. The magnitude of

[2] The hydrodynamic radius (R_H) or Stokes radius (R_s) is the radius of a hard sphere that diffuses at the same rate as the molecule. R_s is smaller then R_g.

the dipolar coupling D_{ij} between two spins i and j is given by equation 2.20 [121]:

$$\begin{aligned} D_{ij} &= -\frac{\gamma_i\gamma_j\hbar\mu_0}{4\pi^2 r_{ij}^3}\left\langle\frac{(3cos^2\theta(t)-1)}{2}\right\rangle \\ &= D_{max}\langle P(cos\theta(t))\rangle \end{aligned} \tag{2.20}$$

with $D_{max} = -\gamma_i\gamma_j\hbar\mu_0/4\pi^2 r_{ij}^3$, θ the angle of the internuclear vector relative to the static field and r_{ij} the internuclear distance, which is assumed constant in the case of covalently bound nuclei, and in all cases represents a vibrationally averaged distance. The angular parentheses define an average over all conformations exchanging on timescales faster than the millisecond. Dipolar couplings between covalently bound spins can be intrinsically very strong (around 11 kHz for an amide ^{15}N-^{1}H spin pair), nevertheless if all possible orientations θ are sampled with equal probability, as is the case free solution, the value of the measured coupling averages very efficiently to zero. However, as said, in the case of an anisotropic environment, the averaging is incomplete and eq. 2.20 becomes:

$$D = D_{max}A_{zz}[P(cos\vartheta) + \eta/2sin^2\vartheta cos2\varphi] \tag{2.21}$$

where A_{zz} is the longitudinal component of the alignment tensor, which describes the net alignment of the protein relative to the magnetic field in terms of a second rank order matrix. η is the rhombicity defined as $\eta = (A_{xx} - A_{yy})/A_{zz}$, and ϑ and φ are the angles relating the orientation of the internuclear vector to the alignment tensor. Measured RDCs can then be interpreted in terms of different orientations of the internuclear vectors relative to the molecular frame.

In the case where the IDP is examined in free solution, the alignment of all conformations of the molecule contributing to the time and ensemble average can be expected to vary significantly as a function of the shape and size of the individual conformation. In this case the RDC must be described in terms of the sum over the different time averages of all N molecules in the ensemble:

$$D = D_{max}\frac{1}{N}\sum_{k=1}^{N}\frac{1}{t_{max}}\int_{t=0}^{t_{max}} P(cos\theta_k(t))dt \tag{2.22}$$

where t_{max} is the maximal time, during which experimental averaging occurs, i.e. on the order of tens to hundreds of milliseconds. Assuming that each copy of the protein samples the conformational space of the ensemble [269], this can be further simplified to:

$$D = D_{max}\frac{1}{t_{max}}\int_{t=0}^{t_{max}} P(cos\theta(t))dt \tag{2.23}$$

Given the intrinsic flexibility and high degree of disorder of the intrinsically unstructured proteins, it is at first sight surprising that RDCs can be observed at all. Intuitively, one would expect the magnitude of the RDCs to be reduced to zero by averaging of $(3cos^2\theta - 1)$ over all possible orientations of the internuclear vector in the conformational ensemble. And yet, all

intrinsically unstructured or unfolded proteins studied with this NMR technique exhibit non-zero RDCs. To a first approximation, when for example considering $^1D_{NH}$ coupling, the measured values are found to have negative sign, with maximal values measured in the centre of the protein, tapering off via a so-called bell-shaped distribution the zero at the extremities. The key to understanding this time and ensemble average came mainly from Annila and co-workers [251, 142, 288], who used polymer models to describe the unfolded protein as a series of connected segments of equal length experiencing restricted random walk. Integration of eq. 2.23 over available orientations of each segment formalises the idea that in the presence of an obstacle, orientational sampling is more restricted in the centre of the chain than at the termini, leading to non-vanishing RDCs, even when the torsion angles along the polymer chain can adopt random conformations. Segments in the centre have more neighbours, and are therefore less flexible than those at the ends. This rationalises the experimentally observed bell-shaped distribution.

However, in the detail, intrinsically unstructured proteins show deviations from the bell-shaped distributions, that must originate from structural elements in the protein chain. The fact the gel medium could induce changes in structure or dynamics [1], is denounced by the fact that ^{15}N-^{1}H HSQC spectra recorded both in presence and absence of the alignment medium are reported to be identical [276]. The cross-peaks are neither shifted nor broadened. Moreover, there were no significant changes in ^{15}N R_1 or R_2 relaxation rates. However, this conclusion could be rather rash, since the very low relative population of the aligned state ($10^{-4} - 10^{-3}$) together with fast chemical exchange between different structurations will lead to negligible chemical shift and line broadening effects.

The observed RDCs can be explained without the need to invoke the presence of secondary structure elements or native-like structure (in the case of denatured protein) and were shown to arise only from the intrinsic properties of the unstructured polypeptide chain. As explained in section 1.2.1, protein random coils are no statistical random coils. Real polypeptides do not behave as ideal random coils in which the backbone dihedral angles of each amino acid residue are independent of its neighbours. Especially the introduction of amino acid-specific conformational behaviour proved an important measure in explaining the observed RDC values. Jha et al [202] and Bernadó et al. [28] have constructed conformational ensembles by sampling amino-acid-specific (ϕ/ψ) propensities (taken from a coil library), provided that these did not result in steric clashes in the chain, and showed that the averaged predicted RDCs of such ensembles closely matched the experimentally observed values. This approach, called Flexible-Meccano or FM, thus explains the experiment without the need to invoke residual secondary or tertiary structuration. The recalculated RDCs compared to the experimental ones of one such study are shown in fig. 2.11 The RDCs are of a single sign because of the prevalence of extended conformations and the global alignment of the entire chain. However, when opposite sign RDCs appear for a certain unstructured protein, this does indicate the propensity for α-helices or local hydrophobic collapse [450]. This can, for the case of helix formation be seen in fig. 2.12. Such nascent α-helical structuration could thus indicate the presence of molecular recognition elements in the protein chain. Possibly equally important, the direction in which the disordered strands adjacent to such helical structuration are projected, which depends on the length of the helix [200], could be detected by the effective tilt of the helix relative to

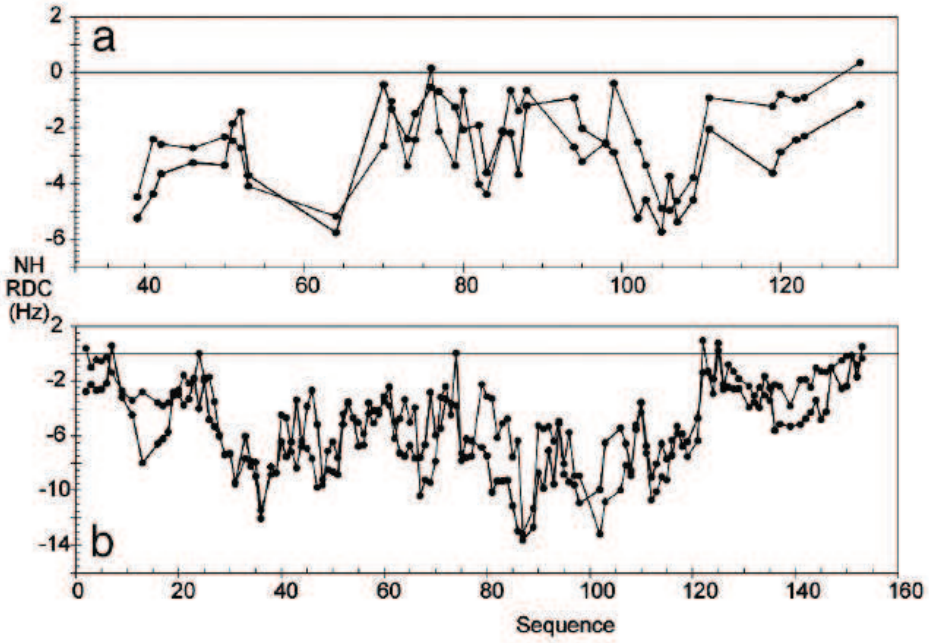


Figure 2.11. Prediction of RDCs from urea-unfolded proteins. (a) Comparison of RDCs calculated by using a combined side-chain volume exclusion and amino acid-specific backbone dihedral angle (ϕ/ψ) propensities method with experimental RDCs measured from 8 M urea-denatured Δ131Δ form of Staphylococcal nuclease [354]. (b) Comparison of RDCs calculated by using the same method with experimental RDCs measured from 8 M urea-denatured apo-myoglobin [276]. In both cases, simulated data were scaled to reproduce the total range of the experimentally measured couplings (grey). The image is taken from [28].

the alignment axis. Such information can be important since there probably exist optimal directions that lead to the most functional interactions of the molecular recognition element.

Although Flexible-Meccano (as it is called by Blackledge and co-workers and by others slightly less poetically referred to as the generate-and-test approach [419]), using the amino-acid-specific statistical coil description, provides a straightforward method for calculating RDC profiles that would be expected if the protein behaved as a random coil, devoid of any specific or persistent local or long-range structure, established deviations from experiment are more challenging to explain. Clearly, the only current option is to generate additional molecular ensembles in a trial-and-error kind of way, in the hope this might result in the prediction of the observed RDCs. But since deviations can originate from very different structuration types (from very local to long-range) and a priori nothing indicates structuration of either type, this method is not practical indeed. Moreover, given a particular alignment tensor, an infinite number of backbone conformations can agree with the single ^{15}N-^{1}H RDC most oftenly used [418]. To avoid this scenario, additional (compared to the traditional $^{1}D_{NH}$) RDCs can be measured. A generated molecular ensemble then has to match a set of RDC measurements including for example $^{1}H^{N}$-$^{1}H^{N}$, $^{1}H^{N}$-$^{1}H^{\alpha}$, $^{1}D_{C\alpha H\alpha}$ and $^{1}D_{C\alpha C'}$ RDCs [270, 271, 200]. Alternatively, a method has been proposed to search systematically for an ensemble of structures directly from sparse experimental RDC restraints [419].

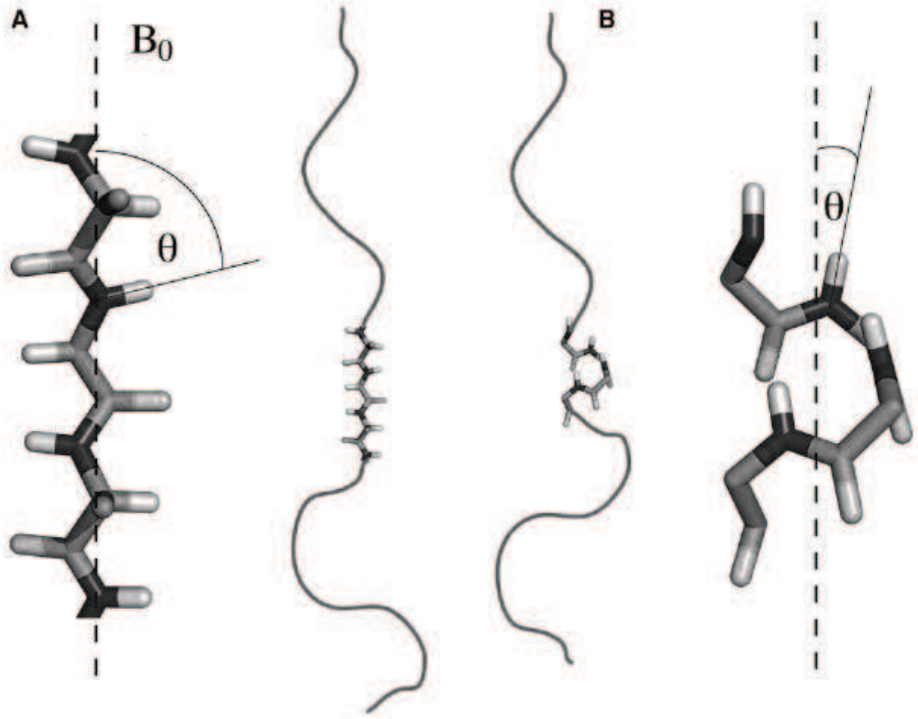


Figure 2.12. Figurative representation of effective angular averaging properties of ^{15}N-^{1}H vectors in an unfolded protein dissolved in weakly aligning medium with the director along the magnetic field. Dipolar couplings measured for ^{15}N-^{1}H vectors in more extended conformations ($\theta \simeq 90°$), more commonly found in unfolded proteins, will have negative values (A), whereas those in helical or turn conformations align more or less parallel with the direction of the chain ($\theta \simeq 0°$) and will have larger positive values (B). The image is taken from [201].

2.7.11. Isotopically Discriminated (IDIS) NMR spectroscopy

The study of interactions in mixture samples may be extremely interesting, for NMR spectroscopy it requires differentially labelled proteins. This has long been done by adding non-labelled interaction partners to a ^{15}N- or ^{15}N- and ^{13}C-labelled protein. The subsequent changes in the fingerprint ^{1}H-^{15}N correlation spectra can then be interpreted to deduce information about the complex, interactions and binding sites [356, 459, 11, 37, 105, 351]. However, such an approach suffers from the disadvantage that, to see what happens with unlabelled partners in a protein mixture, the mixture needs to be recreated again, with a different combination of labelled-unlabelled components. This necessity to prepare multiple samples makes it impossible to run all the experiments required to study a particular system under identical conditions. Moreover, the correlated changes, happening to several polypeptide components as further ligands are added, cannot be detected, again making it difficult to study cooperative, competitive, and allosteric binding events.

To circumvent this problem, IDIS NMR (isotopically discriminated NMR) was developed [160]. The principle is simple. Two polypeptide components, where one is isotopically labelled with ^{15}N and the other with ^{15}N and ^{13}C, are mixed in a sample. The usual ^{1}H-^{15}N correlation spectra are supplied with an isotope filter selecting or discriminating against ^{12}C' versus ^{13}C'

atoms connected to ^{15}N. As a result, two normal-looking ^{1}H-^{15}N correlation subspectra are obtained in a single experiment for the two differently labelled components, one for ^{1}H-^{15}N(^{12}C') and another for ^{1}H-^{15}N(^{13}C'), thus allowing each polypeptide to be monitored separately and independently. Of course, unlabelled thirds can also be added to the mixture.

2.7.12. Prerequisite: Resonance Assignments

Almost all of the discussed NMR techniques (all those that give per-residue information) have one hugely important conditio sine qua non in common; the backbone resonances must be assigned. As outlined earlier, chemical shift dispersions are poor for flexible unstructured proteins. Exceptions are the backbone ^{15}N and ^{13}CO chemical shifts that are influenced both by residue type and by the local amino acid sequence and therefore remain well-dispersed. Thus, the classical three-D triple resonance experiments to establish sequential connectivities can be used for resonance assignments, but the emphasis should be on the well-resolved ^{15}N and ^{13}CO resonances in uniformly ^{15}N,^{13}C-labelled proteins. Pulse sequences that are appropriate for this purpose are summarised in table 2.2.

Experiment	Connectivity	Refs.
^{1}H-^{15}N HSQC	$^{15}N_i$-$^{1}H_i$	[34, 295, 457]
HNCACB	$(^{15}$N-^{1}H$)_i$-$(^{13}C_\alpha$-$^{13}C_\beta)_i$,$(^{13}C_\alpha$-$^{13}C_\beta)_{i-1}$	[437, 282]
CBCA(CO)NH	$(^{15}$N-^{1}H$)_i$-$(^{13}C_\alpha$-$^{13}C_\beta)_{i-1}$	[166, 282]
HNCO	$(^{15}$N-^{1}H$)_i$-$^{13}CO_{i-1}$	[215, 168, 282]
(HCA)CO(CA)NH	$(^{15}$N-^{1}H$)_i$-$^{13}CO_i$,$^{13}CO_{i-1}$	[248]
HNCA	$(^{15}$N-^{1}H$)_i$-$^{13}C_{\alpha,i}$,$^{13}C_{\alpha,i-1}$	[215, 168, 124]
HN(CO)CA	$(^{15}$N-^{1}H$)_i$-$^{13}C_{\alpha,i-1}$	[25, 168]
TOCSY-HSQC	$(^{15}$N-^{1}H$)_i$-$H_{\alpha,i}$-$(H_\beta$,H_γ,$\dots)_i$	[457]
C(CO)NH-TOCSY	$(^{15}$N-^{1}H$)_i$-$H_{\alpha,i-1}$-$(H_\beta$,H_γ,$\dots)_{i-1}$	[165]
HN(CA)NNH	$(^{15}$N-^{1}H$)_i$-$^{15}N_i$,$^{15}N_{i-1}$,$^{15}N_{i+1}$	[427]
(H)N(CO-TOCSY)NH	$(^{15}$N-^{1}H$)_i$-$^{15}N_i$,$^{15}N_{i-1}$,$^{15}N_{i+1}$	[245]
(H)CA(CO-TOCSY)NH	$(^{15}$N-^{1}H$)_i$-$^{13}C_{\alpha,i}$,$^{13}C_{\alpha,i-1}$,$^{13}C_{\alpha,i-2}$	[245]
(H)CBCA(CO-TOCSY)NH	$(^{15}$N-^{1}H$)_i$-$(^{13}C_\alpha$-$^{13}C_\beta)_i$,$(^{13}C_\alpha$-$^{13}C_\beta)_{i-1}$,$(^{13}C_\alpha$-$^{13}C_\beta)_{i-2}$	[245]

Table 2.2. The HNCACB, CBCA(CO)NH and (HCA)CO(CA)NH experiment can be replaced by the CBCANH [167], HN(CO)CACB [446] and HN(CA)CO [74] experiments respectively that provide the same connectivities.

Besides the HNCO and HN(CA)CO experiment (well dispersed ^{13}CO), the HN(CA)NNH and (H)N(CO-TOCSY)NH experiment are of particular interest in the study of unstructured proteins since they allow to make sequential connectivities based on the more dispersed ^{15}N chemical shifts.

The last three entries of table 2.2 refer to a set of high resolution constant-time[3] triple resonance experiments that transfer magnetisation sequentially along the amino acid sequence using carbonyl ^{13}C homonuclear isotropic mixing and have been developed specifically for assignment of unfolded proteins [245]. In flexible ^{13}C,^{15}N labelled polypeptide chains in H_2O solution, the backbone carbonyl carbons have long transverse relaxation times when compared to the amide ^{15}N and $^{13}C_\alpha$ spins, since they have no directly bound protons and relax almost exclusively due to chemical shift anisotropy (CSA). Carbonyl carbon homonuclear isotropic mixing through the sequential three-bond $^3J_{C'C'}$ is therefore an interesting method to transfer magnetisation along the polypeptide backbone. Using such a scheme, backbone carbonyl magnetisation of a given residue is transferred with equal efficiency to both sequentially adjoining residues, which enables to establish connectivities as far as the i+2 residue. This can be useful in the case of numerous proline residues that can then be bridged. The magnetisation transfer of several of the experiments in table 2.2 is given in fig. 2.13.

The carbonyl-carbon homonuclear isotropic mixing based 3D triple resonance spectra have the overall advantage that more (C_α,C_β) information is contained in the third dimension which allows to make more reliable sequential connectivities. On the other hand, this crowdedness is more or less only compatible with manual assignment since automated assignment approaches would quickly be unable to see the wood for the trees. Obviously, manual assignments are tedious and even impossible for larger proteins. Moreover, as is shown further, even though C_α and C_β chemical shifts are poorly dispersed in flexible proteins, these resonances are sharp compared to rigid folded proteins because of the significantly longer T2 and thus often still selective enough for sequential assignments. As a conclusion it could be said that C_α and C_β chemical shift information from the HNCACB and CBCA(CO)NH spectrum, combined with the more dispersed N and CO signals form the optimal set of chemical shifts for an sequential assignment on intrinsically unstructured proteins.

As for the choice of performing a manual or rather an automatic assignment, fig. 2.14 is interesting to consider. Manual assignments can be long and tedious and even practically impossible in the case of bigger proteins as the set of possible matching (neighbouring) residues becomes to large to cope with. Manual assignments obviously require a lot of time. Completely automated assignments can be useful for smaller folded proteins that exhibit a minimal amount of signal overlap. However, if the amount of signal overlap gets bigger, these methods run into trouble and require a subsequent manual validation, which again increases the assignment time. Semi-automatic procedures, which almost always imply a graphical presentation of window slides on screen and thus allow for an interactive assignment, have proven to

[3] The substitution of a $t_1/2$—180°(^{1}H)—$t_1/2$ sequence by $t_1/2$—180°(^{1}H)—$t_e/2$—180°(X)—$(t_e - t_1)/2$, in which X is a low γ isotope of which the chemical shift is encoded in single-quantum state and with the t_e period constant (constant time) during the entire experiment, is very useful to prevent signal dispersion during the experiment (caused by the scalar coupling to other NMR active spins) by collapsing the multiplet structure [24, 318, 333, 413].

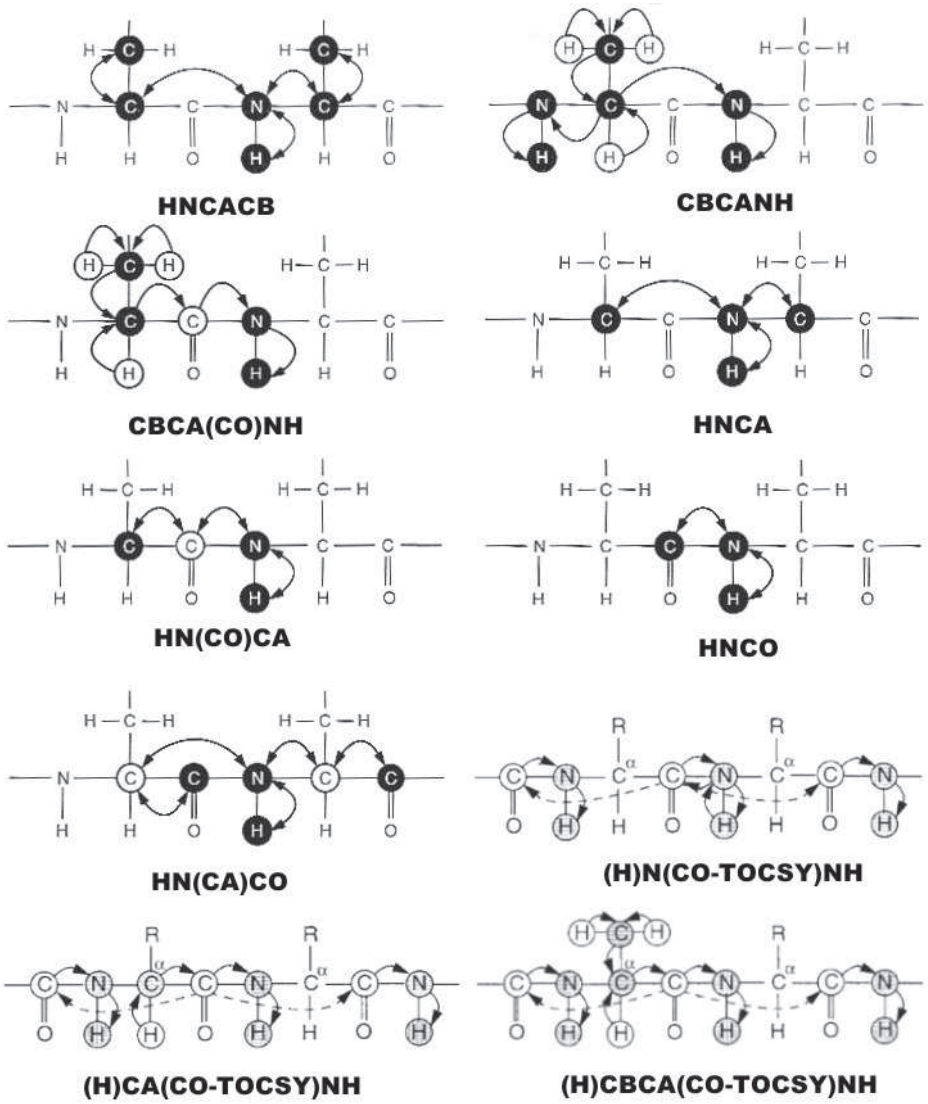


Figure 2.13. The magnetisation transfer of several triple resonance spectra useful for the sequential assignment of disordered protein NMR spectra. Magnetisation transfer diagrams of the (H)N(CO-TOCSY)NH, (H)CA(CO-TOCSY)NH and (H)CBCA(CO-TOCSY)NH experiment are taken from [245], the other from [67].

be the fastest assignment strategies, but have hitherto still required a preceding peak picking of all spectra. Peak pickings are awkward in the case of unstructured proteins because of the abundant signal overlap. Avoiding a peak picking would lead to additional considerable gain of assignment time (as indicated by the cross in fig. 2.14).

Not only do the mentioned advantages (the use of N and CO signals in addition to C_α and C_β, working with a semi-automated assignment strategy, abandoning peak pickings) lead to faster assignment in the case of unstructured proteins, they often are indispensable to obtain complete backbone assignments. During my thesis, I have developed a tool that combines all three advantages. It is described in the article hereafter. The reader will notice that it was presented as an assignment method for both structured and unstructured proteins, in order to address the largest possible NMR-community. Indeed, any tool that successfully tackles large unstruc-

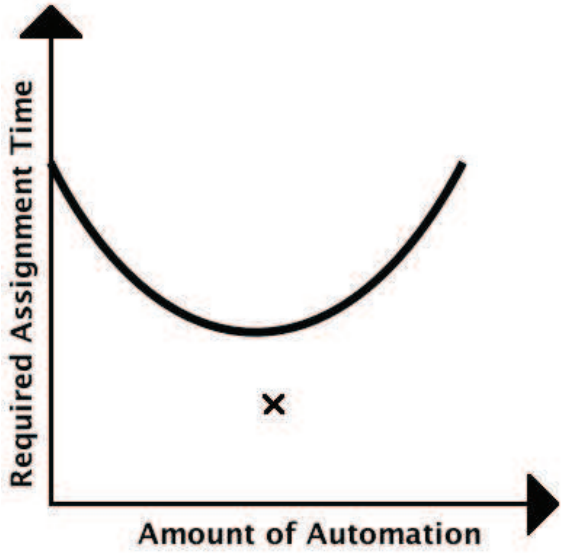


Figure 2.14. Diagram demonstrating the average assignment effort (expressed as required man time) per residue of a big folded protein or an unstructured protein in function of the amount of automation of the assignment tool. The minimum in the graph corresponds to semi-automated tools. Moreover, such semi-automatic tools are the only to allow a resonance assignment without a preceding peak picking phase, which can further reduce execution time (cross).

tured proteins, should be readily applicable on the easier case of structured proteins.

J Biomol NMR (2008) 42:11–21
DOI 10.1007/s10858-008-9262-2

ARTICLE

Graphical interpretation of Boolean operators for protein NMR assignments

Dries Verdegem · Klaas Dijkstra · Xavier Hanoulle · Guy Lippens

Received: 31 March 2008 / Accepted: 9 June 2008 / Published online: 2 September 2008

Abstract We have developed a graphics based algorithm for semi-automated protein NMR assignments. Using the basic sequential triple resonance assignment strategy, the method is inspired by the Boolean operators as it applies "AND"-, "OR"- and "NOT"-like operations on planes pulled out of the classical three-dimensional spectra to obtain its functionality. The method's strength lies in the continuous graphical presentation of the spectra, allowing both a semi-automatic peaklist construction and sequential assignment. We demonstrate here its general use for the case of a folded protein with a well-dispersed spectrum, but equally for a natively unfolded protein where spectral resolution is minimal.

Keywords Computer-aided sequential assignment · Graphical semi-automatic protein assignment method · Boolean operators in NMR · Assignment of structured proteins · Assignment of unfolded proteins

Introduction

The first step in protein structure determination by NMR consists in the sequence specific assignment of the backbone and side chain resonances. A large number of programs have been developed over the last years to assist or automate this assignment process (Andrec and Levy 2002; Atreya et al. 2000, 2002; Bailey-Kellogg et al. 2000, 2005; Bartels et al. 1996, 1997; Bernstein et al. 1993; Buchler et al. 1997; Choy et al. 1997; Coggins and Zhou 2003; Croft et al. 1997; Eads and Kuntz 1989; Eccles et al. 1991; Eghbalnia et al. 2005; Friedrichs et al. 1994; Goddard and Kneller 1989; Görler et al. 1999; Grishaev and Llinás 2004; Gronwald et al. 1998, 2002; Güntert et al. 2000; Hare and Prestegard 1994; Helgstrand et al. 2000; Herrmann et al. 2002a, b; Hitchens et al. 2003; Hyberts and Wagner 2003; Johnson and Blevins 1994; Jung and Zweckstetter 2004; Kjaer et al. 1994; Kleywegt et al. 1991; Kobayashi et al. 2007; Kraulis 1989, 1994; Langmead and Donald 2004; Langmead et al. 2004; Leutner et al. 1998; Li and Sanctuary 1996, 1997a, b; Lin et al. 2003, 2006, 2005; Lukin et al. 1997; Malliavin et al. 1998; Malmodin et al. 2003; Masse and Keller 2005; Masse et al. 2006; Meadows et al. 1994; Morelle et al. 1995; Morris et al. 2004; Moseley and Montelione 1999; Moseley et al. 2001; Mumenthaler and Braun 1995; Mumenthaler et al. 1997; Neidig et al. 1995; Oezguen et al. 2002; Olson and Markley 1994; Orekhov et al. 2001; Oschkinat et al. 1991; Oschkinat and Croft 1994; Ou et al. 2001; Pons and Delsuc 1999; Pristovšek et al. 2002; Slupsky et al. 2003; Szyperski et al. 1998, 2002; Tian et al. 2001; van de Ven 1990; Vitek et al. 2005, 2006; Wan et al. 2003; Wan and Lin 2006; Wang et al. 2005; Wehrens et al. 1991, 1993a, b; Wu et al. 2006; Xu and Sanctuary 1993; Xu et al. 1994, 2002, 2006; Zimmerman et al. 1994, 1997; Zimmerman and Montelione 1995). One of the most common assignment strategies, on which indeed most of the mentioned methods are based, consists of a peak list construction and the subsequent matching of the C_α, C_β and CO chemical shifts between successive residues (Ikura et al. 1990; Kay et al. 1990; Montelione and Wagner 1990). Although successful for small to medium sized proteins, many programs using

D. Verdegem · X. Hanoulle · G. Lippens (✉)
Unité de Glycobiologie Structurale et Fonctionelle, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France
e-mail: guy.lippens@univ-lille1.fr

K. Dijkstra
Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747AG Groningen, The Netherlands

this strategy run into trouble when (i) overlap of the amide resonances increases due to the size or the unstructured nature of the protein, or (ii) spectral incompleteness due to intermediate line broadening or other phenomena. Both operations of peak list construction and frequency matching will suffer under those conditions, leading the operator back to the physical spectra, where one will manually try to complete the data.

We present here an assignment strategy that simultaneously allows both peak list construction and frequency matching in a semi-automatic manner, while remaining close to the initial spectra. It is based on a graphical interpretation of the Boolean AND operator, i.e. a point-by-point multiplication of 2D spectra. The main advantage is that the operator can walk graphically through the protein sequence, while maintaining a quality evaluation of the experimental data that lead to a decision on a sequential assignment. Although point-by-point operations (addition, subtraction, multiplication or division) are very commonly performed on FID's, they can also be done on frequency domain spectra and have even been introduced yet in the field of NMR spectra assignments (Masse et al. 2006). However, in our method, these operations play a more prominent role. Demonstrating the principles first on the well-folded Cyclophilin B protein, we extend its application towards a fragment of the natively unfolded Tau protein (Tau F3, amino acids 208–324), where extreme spectral overlap leads to strong degeneracies in the resonance frequencies.

Theory and methods

The assignment principle

Starting from the root $^{1}H,^{15}N$ HSQC spectrum and clicking on a certain appearing peak, our program readily extracts the corresponding $^{1}H,^{13}C$ planes from the CBCA(CO)NH and HNCO spectra (Fig. 1). On the basis of these two $^{1}H,^{13}C$ spectra, the operator defines with the mouse the carbon frequencies corresponding to the (i − 1) residue. These are automatically stored in a peak list (without

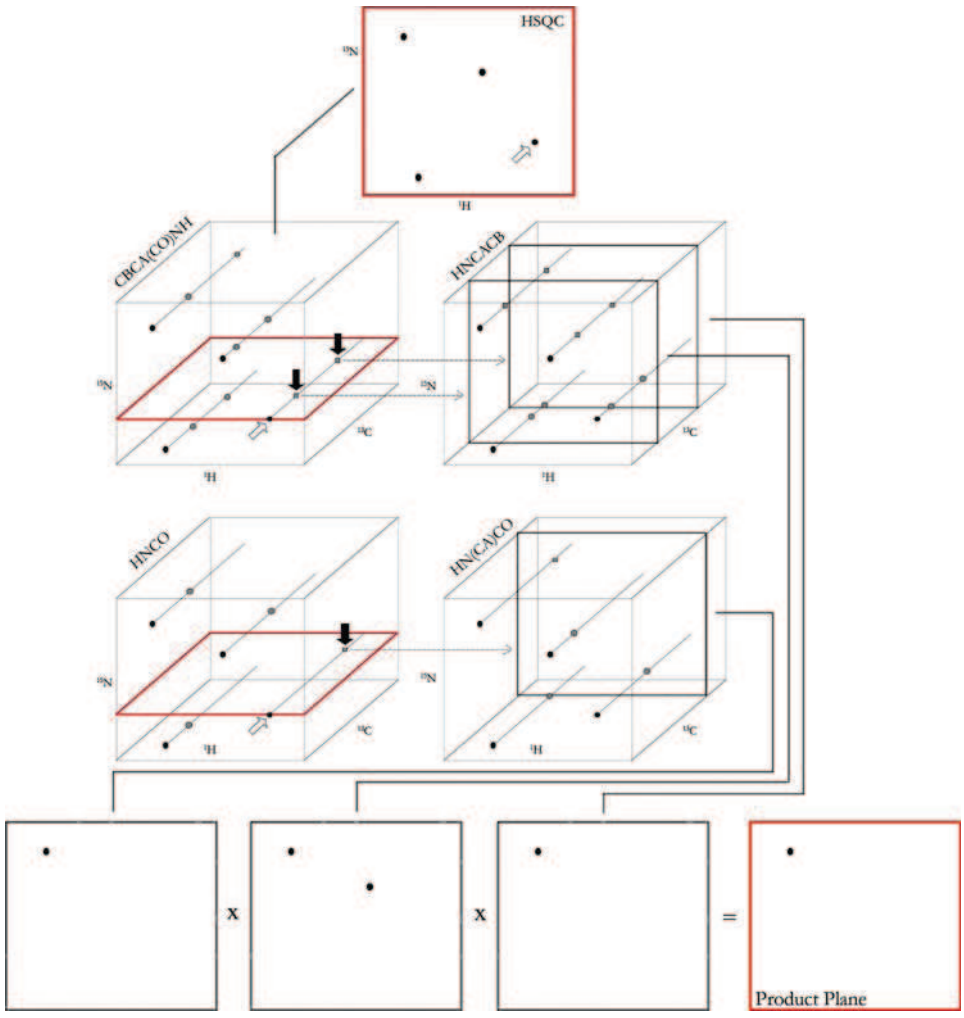


Fig. 1 The product plane approach applied to a protein subset of four consecutive residues. The planes presented on screen during execution in order to be able to click the necessary peaks are drawn in red. Clicking on the rightmost amide peak in a first step (hollow arrow) and the (i − 1) ^{13}C signals in a second step (black arrows) results in the selection of three planes whose point-by-point multiplication leaves only one major peak indicating that the leftmost amide peak is the (i − 1) residue. All 3D-spectra are joined with the HSQC spectra in front to indicate the root of each spin system. For simplicity, the smaller (i − 1) peaks occurring in the HNCACB and HN(CA)CO spectra have been left out of this scheme. It should be noted however, that these can lead to a small product plane signal in the residue (i) position

assignment at this moment), and the corresponding $^{1}H,^{15}N$ planes are extracted from the HNCACB and HN(CA)CO spectra. Rather than displaying the three corresponding planes together on screen and determine by eye the coordinates where there is simultaneous intensity, we impose this criterion by a point-by-point multiplication of the planes. This corresponds to a graphical interpretation of the Boolean AND operator, that requires simultaneous intensity in the spectra to obtain a resulting spectrum with a detectable intensity (Fig. 2). Applied to the three $^{1}H,^{15}N$ spectra extracted at the carbon frequencies of the $(i - 1)$ residue, the point-by-point multiplication therefore defines a novel $^{1}H,^{15}N$ HSQC spectrum that contains intensity at the position of the $(i - 1)$ residue.

Once the $(i - 1)$ residue position has been found, it can be used as the starting point of another run of the algorithm. The repeated execution of the routine allows for an assigning walk through the spectrum towards the N-terminus of the protein.

In order to obtain product planes with constant highest peak intensities, the final product plane is initially normalized by dividing it by its maximum value (or minimum value if an odd number of negative peaks was involved in the multiplication) and is afterwards multiplied by a constant factor (e.g. 1e10) to finish with a spectrum with "natural" intensities (i.e. with peaks of more or less the same magnitude as the ones in real spectra).

Boolean operators in NMR

The assignment method is based on a graphical interpretation of the Boolean AND operator, that can be implemented as the point-by-point multiplication of spectral matrices (Fig. 2). Likewise, the OR operator would correspond with point-by-point summation. The NOT operation applied to a spectrum does not result in a new spectrum as such, but rather a 0/1 filled matrix of the same size as the original spectrum. Whether a certain element of this matrix is zero or one is determined by a chosen threshold (see light grey plane in the 2D-case of Fig. 2). If the intensity at a certain point in the original spectrum exceeds this threshold, the corresponding value in the NOT-matrix is set to zero. In the other case, the NOT-matrix value is set to one. This results in a "spectrum" that display holes at the places where the original spectrum contained peaks. Both OR and NOT operations on spectral planes will prove useful further on when trying to assign proteins with unfavorable amino acid sequences.

Input spectra

In its most basic form, the described algorithm uses the HNCACB, CBCA(CO)NH, HN(CA)CO, HNCO and of course HSQC spectra as input. When these five spectra are applied as depicted in Fig. 1, an assigning "walk" towards the N-terminus of the protein is made. It is however interesting to note that a simple exchange of sequential and intra-residue spectra in the algorithm results in the opposite functionality that allows a "walk" in the opposite direction, towards the C-terminus.

Here, the assignment of Cyclophilin B and Tau F3 spectra will be discussed.

The NMR measurements of both protein samples were performed on a Bruker Avance 600 MHz equipped with a cryogenic triple resonance probe head by using standard Bruker pulse programs. The CypB (185aa, 20.4 kDa) sample contained 600μM CypB in an aqueous buffer with 50 mM Na_2HPO_4/NaH_2PO_4, 60 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 6.35 at 293 K. The Tau F3 (124aa, 13.3 kDa) sample contained 250 μM of protein in a 25 mM Tris-D11, 25 mM NaCl, 2.5 mM EDTA, 2.5 mM DTT aqueous buffer (pH 6.8, 293 K). The acquisition

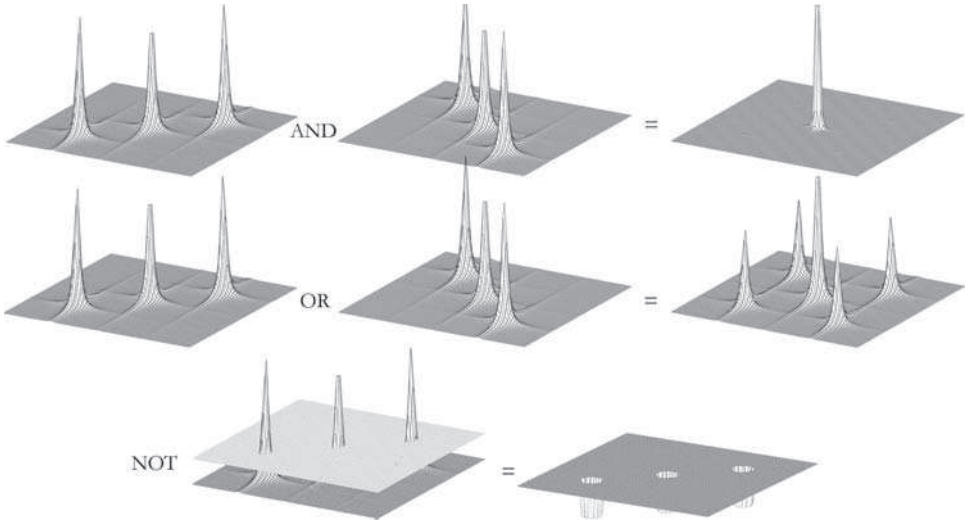


Fig. 2 The Boolean operators applied to 2D spectra presented as topographic maps

parameters for the CypB spectra were: 2048 (^{1}H) and 256 (^{15}N) complex points and 32 scans per increment for the HSQC (exp time: 2 h 41 min), 1024 (^{1}H), 68 (^{15}N) and 128 (^{13}C) complex points and 16 scans per increment for the HNCACB (exp time: 1 day 21 h 23 min), 1024 (^{1}H), 104 (^{15}N) and 142 (^{13}C) complex points and 8 scans per increment for the CBCA(CO)NH (exp time: 1 day 15 h 20 min) and 1024 (^{1}H), 104 (^{15}N) and 128 (^{13}C) complex points and 8 scans per increment for the HN(CA)CO and HNCO (exp times: 1 day 10 h 58 min and 1 day 10 h 25 min). For the Tau F3 sample, the acquisition parameters were: 2048 (^{1}H) and 256 (^{15}N) complex points and 64 scans per increment for the HSQC (exp time: 5 h 25 min), 2048 (^{1}H), 96 (^{15}N) and 232 (^{13}C) complex points and 8 scans per increment for the HNCACB (exp time: 2 days 13 h 36 min), 2048 (^{1}H), 96 (^{15}N) and 132 (^{13}C) complex points and 8 scans per increment for the CBCA(CO)NH (exp time: 1 day 11 h 42 min), 2048 (^{1}H), 92 (^{15}N) and 96 (^{13}C) complex points and 8 scans per increment for the HN(CA)CO and HNCO (exp times: 1 d 10 h 58 min and 1 d 10 h 25 min) and 2048 (^{1}H), 86 (^{15}N) and 96 (^{13}C) complex points and 8 scans per increment for the HNN (exp time: 23 h 26 min).

It is important to notice that the different spectra required for an assignment of this kind should all be recorded under the same sample conditions. Any technique based on the point-by-point multiplication of spectrum slices originating from different spectra is obviously quite sensitive to small differences in chemical shift across those spectra, which might arise if nonidentical parameters are used.

Results and discussion

Graphical walk through the triple resonance spectra

To demonstrate the procedure on a real-life example, we start from the cross peak at 7.55, 121.72 ppm in the CypB HSQC spectrum, that we previously assigned to Lys 149 (Hanoulle et al. 2007). Extracting the ^{1}H,^{15}N planes from the HN(CA)CO, HNCACB (C_α and C_β) at the (i − 1) ^{13}CO, $^{13}C_\alpha$ and $^{13}C_\beta$ carbon frequencies as defined by the HNCO and HN(CO)CACB lines at the Lys 149 position yields the three planes shown in Fig. 3. The product plane (4) (in green) comes into being as a result of their point-by-point multiplication. This plane superposed on the Cyclophilin B HSQC indisputably points out the position of the root signal of Arg 148.

When lowering the threshold, we do see other amide resonances of lower intensity, indicating that due to the limited resolution in the carbon dimension residues can have some residual intensity that matches the three required frequencies. When a given (^{1}H,^{15}N) correlation peak represents two or more residues, the operator is faced with the same problem as the number based algorithms. However, as in other semi-automated assignment programs, our method, inherent to its principle, constantly shows the relevant spectrum slices on screen. The obvious advantage is that the raw data with all the information about subtle frequency differences and/or peak forms are still available. Two real situations where only working with raw data helps to exclude ambiguity in the assignment of the CypB protein are considered here.

Figure 4 shows the plane pulled out from the cyclophilin B CBCA(CO)NH spectrum after clicking the Val 12 residue signal. This plane can then, according to the product plane (AND) algorithm, be used to select the C_α and C_β (i − 1) signals. The Val 12 (^{1}H,^{15}N) correlation peak appears in a more crowded region of the HSQC. Four ^{13}C peaks can be distinguished in the ^{1}H,^{13}C plane, but visual inspection readily allows to pair the peaks at 61.5 and 69.0 ppm. The two other signals at 41.0 and 54.0 ppm have a proton frequency that differs by 0.006 ppm from the previous pair, and would therefore probably be assigned to the same peak by automated assignment routines that commonly apply a proton uncertainty of 0.05 ppm.

A second example illustrating the advantage of having ready access to the raw data is found when trying to assign the amide peak that corresponds to Gly 31. The superposition of the HSQC and the product plane obtained after clicking Leu 32 is shown in Fig. 5. We are faced here with the extreme, but possible situation in which the authentic (i − 1) signal is not the most intense one in the product plane. To establish and overcome this problem however, a simple feedback strategy, that exploits once again the usefulness of being able to graphically present the slight chemical shift differences, is sufficient.

This feedback functionality graphically compares the set of peaks involved, as in Fig. 6 for the Leu 32 case, and reveals clearly that the Gly 138 C_α chemical shift is shifted slightly downfield compared to the Leu 32 C_α (i − 1) shift.

At any point, the spectroscopist can decide not to include one of the three plane subject to the multiplication (bottom Fig. 1). If for example a certain residue has a weak C_β peak in the HNCACB spectrum, one can exclude the corresponding C_β-plane from the (i − 1) product plane calculation (and thus treat the residue as if it were a glycine). Although this practice will in theory lead to a less selective product plane, it can in some cases avoid the situation where the product plane exhibits a too low signal/noise to be useful.

Using our graphical walk, that is in this case only interrupted when one encounters a proline residue as these do not appear in a HSQC spectrum, we were able to repeat the full assignment of CypB in a minimal time (less than a

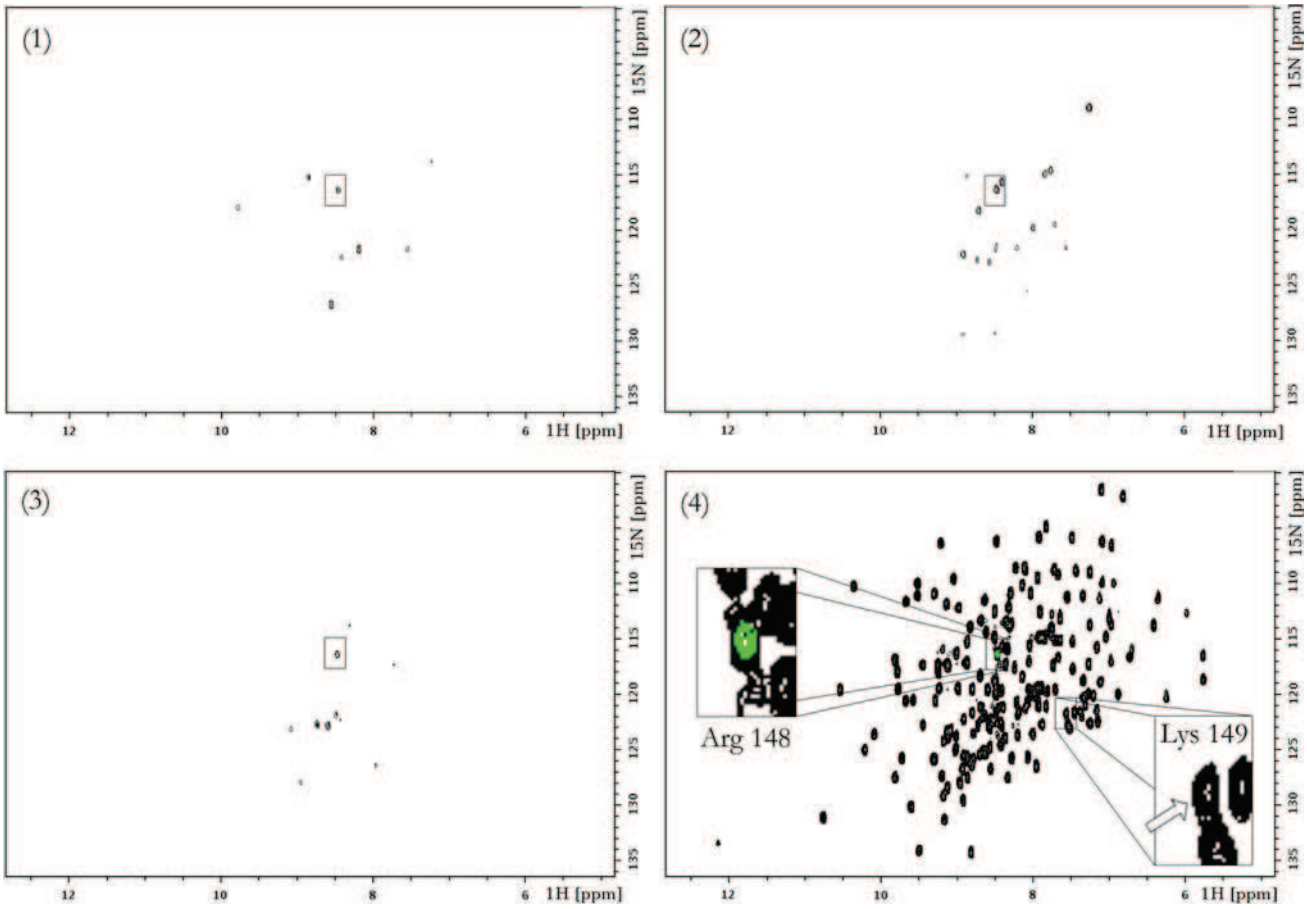


Fig. 3 When clicking the Lys 149 ^{1}H, ^{15}N signal in the CypB HSQC, automatic extraction of the corresponding ^{1}H,^{13}C planes of the HNCO and HN(CO)CACB allow the manual definition of the (i − 1) CO, C_{α} and C_{β} frequencies. Extracting the ^{1}H,^{15}N planes from the HN(CA)CO, HNCACB (C_{α} and C_{β}) at these (i − 1) ^{13}CO, $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ carbon frequencies yields the three planes (**1**), (**2**) and (**3**). The resulting product plane (in green) superposed on the original HSQC-spectrum is presented in (**4**). In it, the Arg 148 position can be clearly identified. All clicking required to obtain the first three planes can be done fairly precisely because of an available zoom function. This also enables one to backup precise backbone ^{15}N, ^{13}C, ^{1}H, and side-chain $^{13}C_{\beta}$ assignments to an output file

day), and comparison with our previous assignment based on peak lists showed perfect agreement.

Extending to natively unstructured proteins

Natively unfolded proteins represent a different challenge to assignment programs. With a reduced amide proton chemical shift range (often inferior to 1 ppm), overlap becomes very severe, leading to many branching points for the automatic matching algorithms. Therefore, manual intervention of the operator becomes even more important than in the case of folded proteins such as CypB, as it allows to alleviate possible ambiguities on the basis of subtle peak position or shape differences.

In this unstructured protein category, the algorithm was powerful enough to determine almost all the proline bordered amino acid stretches of the Tau F3 (amino acids 208–324) protein fragment. We were able to assign all residues except for the S237–S238 pair and the two GGG triplets starting at G271 and G302. The pair and triplets occur in heavily overlapped regions and are moreover preceded by a proline residue preventing the upstream graphical walk.

The performance of the product plane algorithm can however, for natively unfolded proteins, be improved when combined with the information included within the triple resonance HNN spectrum, that can be recorded with a decent sensitivity for these protein due to their narrow line widths. When a certain HSQC root is chosen with the mouse, a corresponding ^{1}H,^{15}N plane can also be pulled out of this latter 3D spectrum, in which one can find the ^{15}N chemical shift of the residues in (i − 1) and (i + 1) position. Drawing this information as horizontal lines on the product plane spectrum will, in case of doubt, reveal the correct neighbor of the clicked root signal.

Figure 7 shows the (i − 1) product plane and the (i − 1) ^{15}N and (i + 1) ^{15}N chemical shift values after executing the algorithm on Val 306. Besides the own (i) signal, the product plane contains three possible (i − 1) signals. The

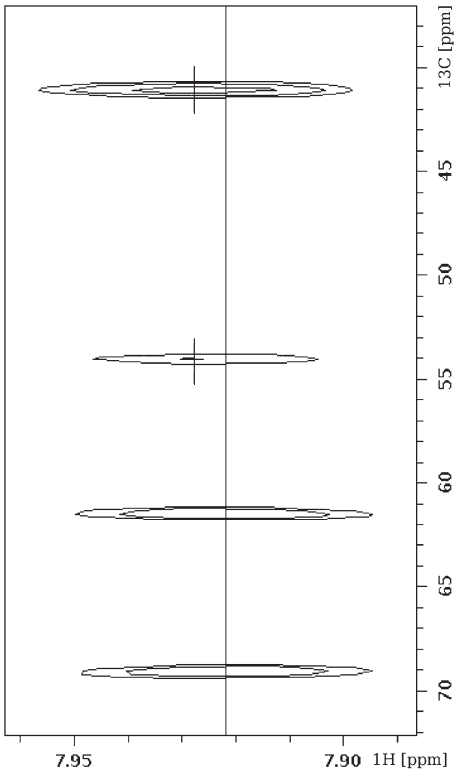


Fig. 4 A fragment of the Val 12 corresponding $^{1}H,^{13}C$ plane extracted from the cyclophilin B CBCA(CO)NH spectrum. The black vertical line indicates the proton chemical shift of the Val 12 residue. This view is projected on the screen as determined by the algorithm

feedback strategy of graphically comparing the C_α and C_β shifts involved shows that there is a perfect match with none of those three. The HNN info on the other hand,

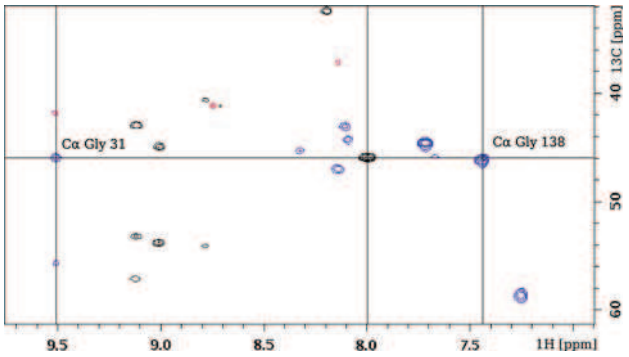
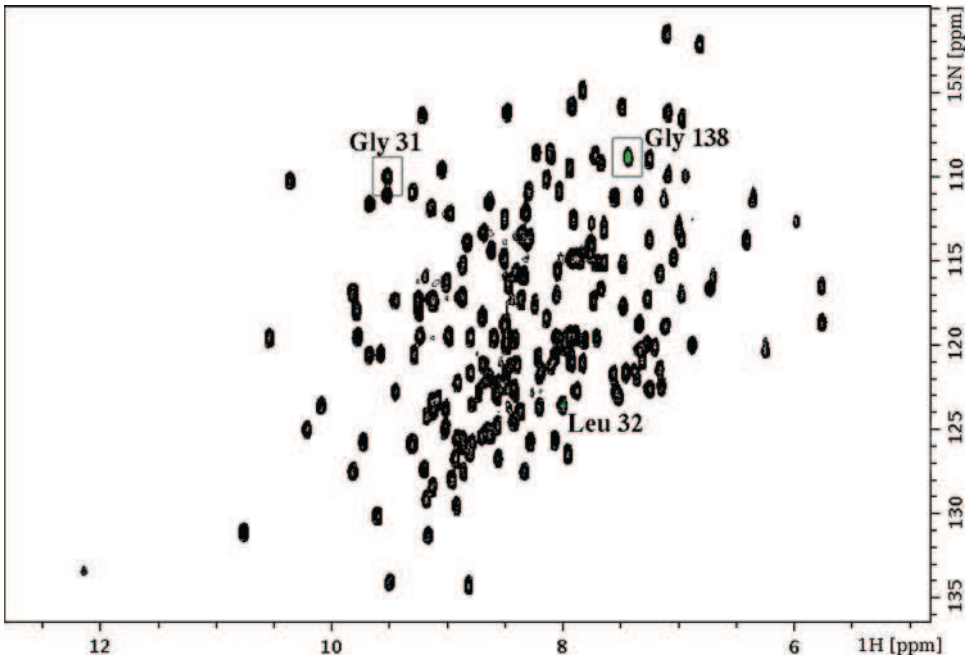


Fig. 6 A superposition of three pulled-out 2D spectra in order to graphically compare the involved signals. The two HNCACB extracted planes are determined by the ^{15}N chemical shifts of Gly 31 and Gly 138 of CypB. The CBCA(CO)NH plane corresponds to the Leu 32 ^{15}N chemical shift. As color convention we use black for the CBCA(CO)NH signals and blue (C_α) and red (C_β) for HNCACB signals. Vertical lines intercept the three comparison partners in their points of highest intensity. The horizontal line crosses the Leu 32 signal at its maximum

allows one to pinpoint the largest product plane signal as the genuine Ser 305 amide resonance. This situation arose because of the almost complete overlap of the Ser 262, Ser 293 and Ser 305 HSQC signals that all have a glycine residue in the (i − 1) position. This caused the individual HNCACB signals to have merged to three new averaged glycine C_α, serine C_α and serine C_β signals at different chemical shifts.

When the HNN was added to list of input spectra we succeeded in completely assigning the Tau F3 spectra, including the eight earlier mentioned difficult cases. For

Fig. 5 The product plane after clicking Leu 32 shown on top of the Cyclophilin B HSQC. The strongest signal is actually that of Gly 138, while Gly 31 has a lower intensity at the given threshold. Also the Leu 32 peak itself shows some intensity due to the presence of minor (i − 1) signals in the HNCACB and HN(CA)CO spectra



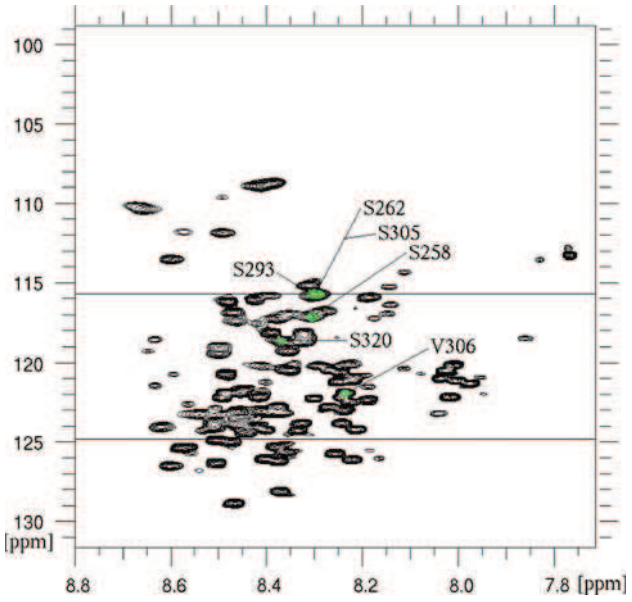


Fig. 7 For natively unstructured proteins such as Tau F3, the product plane functionality has to be reinforced by information from the HNN spectrum to be able to do complete assignments. The (i − 1) and (i + 1) ^{15}N chemical shifts it provides when following the procedure after clicking V306 (drawn on the product plane as horizontal lines), allows to assign the S305 residue

unstructured proteins, the HNN information also makes the assignment a lot faster, since it efficiently prevents the need for a signal position based feedback in the assigning walk.

Generating starting points

The above described procedure leads to the ready assignment of stretches of connected resonances. Based on the residue-type specific carbon chemical shifts, that unambiguously define residues such as Gly, Ala, Ser and Thr, those stretches can in most cases be mapped in a straightforward way onto the protein sequence. For the case of a folded protein such as CypB in our example, this information is ample, and a full assignment can be easily obtained. For the Tau fragment, however, the rapid obtention of suitable starting points helps in the procedure, and avoids problems with repeating stretches in the protein sequence. Generally, the existence of suitable starting points in the assignment procedure gives additional confidence in the method, and leads to a more rapid assignment.

The graphical interpretation of the Boolean operators as defined above can equally be used in a similar way by including the OR operator to allow for some spectral degeneracy. A first manner is to define a given residue type by the requirement that both the C_α and C_β frequencies fall within a certain range of the random coil chemical shift values for this residue type. This requirement can be obtained graphically in two steps (Fig. 8): first, the HNCACB ^{1}H,^{15}N planes with the ^{13}C chemical shift values within the defined range of the random coil values are summed, leading to a C_α- and C_β-defining plane. Formally, this sum procedure is equivalent to the Boolean OR operator. In a second stage, we multiply both resulting sum-planes to obtain a novel ^{1}H,^{15}N plane that contains intensity only for those resonances where the C_α and C_β requirement is fulfilled. This procedure is akin to the MUSIC pulse sequences, where one combines carbon selective pulses and multiple quantum filtering to obtain residue-type specific subspectra (Schubert et al. 1999, 2001a, b). However, the present method does not require novel experiments, as it is a post-processing method based only on the existing HNCACB experiment, and thereby does not suffer from the relaxation losses that inevitably accompany the longer pulse lengths required for selectivity. This is a distinct advantage for larger proteins, but also for unfolded proteins where the selectivity of the post-music procedure can easily be fine-tuned on the basis of the same experiment, without requiring the recording of novel experiments.

Following this procedure starting from the HNCACB spectrum, one obtains (i, i + 1) subspectra, as the given residue (i) will also be seen from the (i + 1) amide resonance because of the (weaker) N(i)-C_α(i − 1) coupling constant. The same principle can equally be applied to the CBCA(CO)NH spectrum, and thereby leads to a subspectrum of only those residues that have the required residue type as their downstream neighbor (i + 1). Applying the third Boolean operator described in Fig. 2, both the HNCACB and CBCA(CO)NH can thus be used to generate

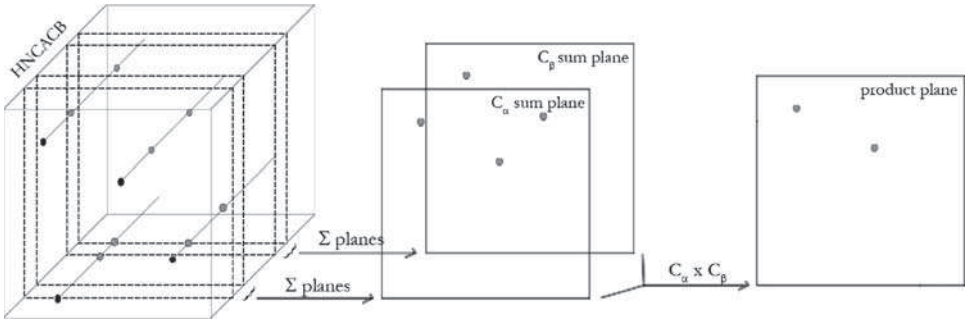


Fig. 8 Boolean operators applied to obtain type-selective ^{1}H,^{15}N spectra. Windows are defined (indicated by the dashed lines) around a residue type characteristic C_α and C_β chemical shift value ($\overline{C_\alpha} \pm x_\alpha$ and $\overline{C_\beta} \pm x_\beta$) and the total of planes enclosed are summed. Subsequently, the resulting sum planes are multiplied to yield the type specific HSQC

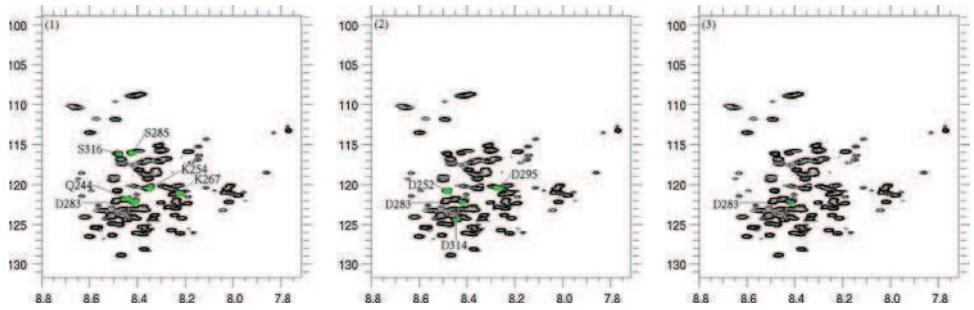


Fig. 9 (**1**) Represents the Leu (i + 1) HSQC, generated from Tau F3's CBCA(CO)NH. For it, windows of 55.1 ± 0.5 ppm (C_α) and 42.4 ± 0.5 ppm (C_β) were chosen. All Leu (i + 1) signals are present in the spectrum. In order to generate the pure Asp (i) HSQC (**2**), the C_α and C_β windows were 54.2 ± 0.5 and 41.1 ± 0.5 ppm, respectively. The cutoff threshold for the NOT operation (Fig. 2) was put at 0.001% of the most intense CBCA(CO)NH derived product plane peak. The Asp type-specific subspectrum contains all four Asp residues in the Tau F3 sequence. Finally, a Boolean AND operation between (**1**) and (**2**) results in a spectrum that only contains intensity at the position of the Asp residue in the unique (L)D283 dipeptide (**3**). The spectrum manipulations that lead to (**3**) are done in a few seconds and thus this procedure provides a very fast generation of starting points. Again, as was the case in the assignment method, all planes are normalized before multiplication and multiplied by a constant factor (1e10) after, as to maintain constant intensity. The scales are values in ppm

pure (i) type specific subspectra. Indeed, (i,i + 1) AND NOT (i + 1) equals (i). We typically do a dipeptide scan over the protein sequence to determine those dipeptides that are unique in the sequence. For the first amino acid of such a dipeptide, the (i + 1) residue-type selective HSQC is calculated, while for the second amino acid, we generate the (i) selective subspectrum. The product plane of those two HSQC's will contain only one major peak, indicating the position of the second residue of the dipeptide.

In a concrete example as the Tau fragment, a simple scan found that 56 residues are in a unique pattern, and this despite the fact that the overall amino acid sequence of Tau is largely unfavorable, with five amino acids making up for over 55% of the sequence. The (L)D283 dipeptide is unique, and the Leu (i + 1) selective spectrum combined with the Asp (i) specific spectrum readily defines it as the peak at 8.41, 122.20 ppm (Fig. 9).

A total of 42 out of 56 residues in a unique pattern, where we note that the pattern XY can be differentiated from XYP because of the proline-directed effect (prolines in (i + 1) induces a −2 ppm chemical shift for the C_α (Wishart et al. 1995)), were immediately assignable. Some residues were unable to be found because of one of three reasons: (i) carbon signals not included in the defined windows, (ii) weak corresponding HNCACB and/or CBCA(CO)NH signals or (iii) both residues of the unique pair are of the same type, which causes the signal to disappear in the (i) type selective subspectrum. All three effects lead to empty product planes at the usual contour threshold and a large amount of meaningless noise peaks at lower thresholds. Reason (ii) is related to the fact that a summation of a number of N subspectra by the OR operator will lead to a decrease in signal/noise of about sqrt(N) (depending on the peak widths) as many planes will contribute to the noise and only a few to the actual signal. However, we found that for unfolded proteins such as Tau, where the defined windows can be kept reasonably small (e.g. 1 ppm) because of the smaller C_α and C_β chemical shift spreads, this signal/noise reduction is disturbing in only a minor number of cases. Thus, the rapid determination of pivotal points, whereby we can even allow for some ambiguous assignments, greatly enhances not only the initial stages of the assignment procedure. As it provides for suitable anchoring points, it facilitates to connect the sequential stretches to the protein sequence. The complete assignment of Tau F3, using our complete package of assignment tools, was done in 1 day time.

Discussion

We have shown here a graphical implementation of the traditional assignment procedure based on connecting complementary triple resonance experiments. The main advantage of the procedure is that the operator remains very close to the experimental spectra at every moment, without relying on peak lists. Whereas the latter allow a rapid computer-assisted assignment in favorable cases, spectral overlap or differential quality of the data in different zones of the spectra can introduce errors that inevitably will lead to problems requiring manual intervention. The product planes as defined in this work represent the Boolean AND operator in its most simple fashion: point-by-point multiplication guarantees that the only remaining intensity comes from planes that both had intensity at the given resonance position. We showed that even for crowded spectra such as obtained for the natively unfolded Tau protein, this graphical procedure can greatly facilitate the assignment process. When complemented in a straightforward way with the HNN experiment, that is particularly favorable for such samples because of their

sharp lines, the assignment becomes as trivial as for a folded protein. An extension to the Boolean OR operator allows to define amino-acid specific subspectra based on the original HNCACB and CBCA(CO)NH spectrum. When compared to the experimental MUSIC pulse sequences, this procedure does not suffer from additional relaxation losses due to the selective and hence longer carbon pulses, does not require novel experiments and can readily build in differential ^{13}C selectivity, but can evidently not reproduce the multiple quantum filtering as was done in the some MUSIC sequences. These residue selective subspectra constitute the input of a very straightforward starting point generation method. We showed that this latter procedure is particularly suitable for unfolded proteins, where the random coil ^{13}C chemical shifts by definition provide an excellent center point for the chemical shift range to be considered. We are currently exploring how the procedure can be combined with quantum-mechanical or semi-empirical chemical shift calculations in order to provide a rapid assignment of the HSQC spectra of proteins with a known 3D structure.

All methods described in this paper were developed using python scripts with the NMR python library functionality (http://linuxnmr02.chem.rug.nl/~dijkstra/NMRpy/). However, we have also implemented them in the CcpNmr software suite (Vranken et al. 2005) as an extension to Analysis for wide distribution. They will become available in the next release (Analysis2.0)

Acknowledgments We thank Dr. I. Landrieu for sample preparation, Dr. J.-M. Wieruszeski for collecting the NMR spectra and Dr. T. Stevens and W. Boucher of the University of Cambridge, Department of Biochemistry for implementing our protein NMR assignment tools in the CcpNmr software suite. The 600 MHz facility used in this study was funded by the Région Nord—Pas de Calais (France), the CNRS and the Institut Pasteur de Lille. Part of this work was funded by a grant of the Agence National de la Recherche (ANR 05 BLAN 0320-0; Tau:Tubulin). D.V. received a predoctoral grant of the French Ministère de la Recherche.

References

Andrec M, Levy RM (2002) Protein sequential resonance assignments by combinatorial enumeration using $^{13}C_\alpha$ chemical shifts and their (i,i − 1) sequential connectivities. J Biomol NMR 23: 263–270

Atreya H, Chary K, Govil G (2000) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. Curr Sci 83:1372–1376

Atreya H, Sahu S, Chary K, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17:125–136

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparce unassigned NMR data. J Comput Biol 7:537–558

Bailey-Kellogg C, Chainraj S, Pandurangan G (2005) A random graph approach to NMR sequential assignment. J Comput Biol 12:569–583

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18: 139–149

Bernstein R, Cieslar C, Ross A, Oschkinat H, Freund J, Holak TA (1993) Computer-assisted assignment of multidimensional NMR spectra of proteins: application to 3D NOESY-HMQC and TOCSY-HMQC spectra. J Biomol NMR 3:245–251

Buchler NE, Zuiderweg ER, Wang H, Goldstein RA (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. J Magn Reson 125:34–42

Choy W, BC S, Zhu G (1997) Using neural network predicted secondary structure information in automatic protein NMR assignment. J Chem Inf Comput Sci 37:1086–1094

Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 26:93–111

Croft D, Kemmink J, Neidig KP, Oschkinat H (1997) Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. J Biomol NMR 10:207–219

Eads C, Kuntz I (1989) Programs for computer-assisted sequential assignment of proteins. J Magn Reson 82:467–482

Eccles C, Güntert P, Billeter M, Wüthrich K (1991) Efficient analysis of protein 2D NMR spectra using the software package EASY. J Biomol NMR 1:111–130

Eghbalnia HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). J Biomol NMR 32:219–233

Friedrichs M, Mueller L, Wittekind M (1994) An automated procedure for the assignment of protein 1HN, 15N, 13C alpha, 1H alpha, 13C beta and 1H beta resonances. J Biomol NMR 4:703–726

Goddard T, Kneller D (1989) Sparky 3. University of California, San Francisco

Görler A, Gronwald W, Neidig KP, Kalbitzer HR (1999) Computer assisted assignment of ^{13}C and ^{15}N edited 3D-NOESY-HSQC spectra using back calculated and experimental spectra. J Magn Reson 137:39–45

Grishaev A, Llinás M (2004) BACUS: a Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. J Biomol NMR 28:1–10

Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR 12:395–405

Gronwald W, Moussa S, Elsner R, Jung A, Ganslmeier B, Trenner J, Kremer W, Neidig KP, Kalbitzer HR (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). J Biomol NMR 23:271–287

Güntert P, Salzmann M, Braun D, Wüthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18: 129–137

Hanoulle X, Melchior A, Sibille N, Parent B, Denys A, Wieruszeski JM, Horvath D, Allain F, Lippens G, Landrieu I (2007) Structural and functional characterisation of the interaction between cyclophilin B and a heparin derived oligosaccharide. J Biol Chem 282:34148–134158

Hare BJ, Prestegard JH (1994) Application of neural networks to automated assignment of NMR spectra in proteins. J Biomol NMR 4:35–46

Helgstrand M, Kraulis P, Allard P, Härd T (2000) Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. J Biomol NMR 18:329–336

Herrmann T, Güntert P, Wüthrich K (2002a) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Biomol NMR 319:209–227

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR 24: 171–189

Hitchens T, Lukin JA, Zhan YP, McCallum SA, Rule GS (2003) MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR 25:1–9

Hyberts SG, Wagner G (2003) IBIS—A tool for automated sequential assignment of proteins spectra from triple resonance experiments. J Biomol NMR 26:335–344

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667

Johnson BA, Blevins RA (1994) NMR view: a computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614

Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Kjaer M, Andersen K, Poulsen F (1994) Automated and semiautomated analysis of homo- and heteronuclear multidimensional nuclear magnetic resonance spectra of proteins: the program PRONTO. Methods Enzymol 239:288–308

Kleywegt GJ, Boelens R, Cox M, Llinás M, Kaptein R (1991) Computer-assisted assignment of 2D ^{1}H NMR spectra of proteins: basic algorithms and application to phoratoxin B. J Biomol NMR 1:23–47

Kobayashi N, Iwahara J, Koshiba S, Tomizawa T, Tochio N, Güntert P, Kigawa T, Yokoyama S (2007) KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR studies. J Biomol NMR 39:31–52

Kraulis P (1989) ANSIG: a computer program for the assignment of 1H NMR spectra by interactive computer graphics. J Magn Reson 84:627–633

Kraulis P (1994) Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space ab initio approach. J Mol Biol 243:696–718

Langmead CJ, Donald BR (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR assignments. J Biomol NMR 29:111–138

Langmead CJ, Yan A, Lilien R, Wang L, Donald BR (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. J Comput Biol 11:277–298

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11:31–43

Li KB, Sanctuary B (1996) Automated extracting of amino acid spin systems in proteins using 3D HCCH-COSY/TOCSY spectroscopy and constrained partioning algorithm. J Chem Inf Comput Sci 36:585–593

Li KB, Sanctuary B (1997a) Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. J Chem Inf Comput Sci 37:359–366

Li KB, Sanctuary B (1997b) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. J Chem Inf Comput Sci 37: 467–477

Lin G, Xiang W, Tegos T, Li Y (2006) Statistical evaluation of NMR backbone resonance assignment. Int J Bioinform Res Appl 2:147–160

Lin G, Xu D, Chen ZZ, Jiang T, Wen J, Xu Y (2003) Computational assignment of protein backbone NMR peaks by efficient bounding and filtering. J Bioinform Comput Biol 1:387–409

Lin HN, Wu KP, Chang JM, Hsu WL (2005) GANA—a genetic algorithm for NMR backbone resonance assignment. Nucleic Acids Res 33:4593–4601

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (13C, 15N)-labeled proteins). J Biomol NMR 9:151–166

Malliavin T, Pons J, Delsuc M (1998) An NMR assignment module implemented in the Gifa NMR processing program. Bioinformatics 14:624–631

Malmodin D, Papavoine CH, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J Biomol NMR 27:69–79

Masse JE, Keller R (2005) AutoLink: automated sequential reso nance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. J Magn Reson 174:133–151

Masse JE, Keller R, Pervushin K (2006) SideLink: automated side-chain assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. J Magn Reson 181:45–67

Meadows RP, Olejniczak ET, Fesik SW (1994) A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. J Biomol NMR 4:79–96

Montelione GT, Wagner G (1990) Conformation-independent sequential NMR connections in isotope-enriched polypeptides by ^{1}H-^{13}C-^{15}N triple resonance experiments. J Magn Reson 87:183–188

Morelle N, Brutscher B, Simorre JP, Marion D (1995) Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments. J Biomol NMR 5:154–160

Morris LC, Valafar H, Prestegard JH (2004) Assignment of protein backbone resonances using connectivity, torsion angles and $^{13}C^{\alpha}$ chemical shifts. J Biomol NMR 29:1–9

Moseley H, Montelione G (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Moseley H, Monleon D, Montelione G (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol 339:91–108

Mumenthaler C, Braun W (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. J Mol Biol 254:465–480

Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY and three-dimensional protein structure determination. J Biomol NMR 10:351–362

Neidig KP, Geyer M, Görler A, Antz C, Saffrich R, Beneicke W, Kalbitzer HR (1995) AURELIA, a program for computer-aided

analysis of multidimensional NMR spectra. J Biomol NMR 6:255–270

Oezguen N, Adamian L, Xu Y, Rajarathnam K, Braun W (2002) Automated assignment and 3D structure calculations using combinations of 2D homonuclear and 3D heteronuclear NOESY spectra. J Biomol NMR 22:249–263

Olson JB, Markley JL (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. J Biomol NMR 4: 385–410

Orekhov VY, Ibraghimov V, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Oschkinat H, Croft D (1994) Automated assignment of multidimensional nuclear magnetic resonance spectra. Methods Enzymol 239:308–318

Oschkinat H, Holak T, Cieslar C (1991) Assignment of protein NMR spectra in the light of homonuclear 3D spectroscopy: an automatable procedure based on 3D TOCSY-TOCSY and 3D TOCSY-NOESY. Biopolymers 31:699–712

Ou HD, Lai HC, Serber Z, Dötsch V (2001) Efficient indentification of amino acid types for fast protein backbone assignments. J Biomol NMR 21:269–273

Pons J, Delsuc M (1999) RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. J Biomol NMR 15:15–26

Pristovšek P, Rüterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program *st2nmr*. J Comput Chem 23:335–340

Schubert M, Smalla M, Schmieder P, Oschkinat H (1999) MUSIC in triple-resonance experiments: amino acid type-selective $^{1}H,^{15}N$ correlations. J Magn Reson 141:34–43

Schubert M, Oschkinat H, Schmieder P (2001a) MUSIC and aromatic residues: amino acid type-selective $^{1}H,^{15}N$ correlations, III. J Magn Reson 153:186–192

Schubert M, Oschkinat H, Schmieder P (2001b) MUSIC, selective pulses, and tuned delays: amino acid type-selective $^{1}H,^{15}N$ correlations, II. J Magn Reson 148:61–72

Slupsky CM, Boyko RF, Booth VK, Sykes BD (2003) Smartnotebook: a semi-automated approach to protein sequential NMR resonance assignments. J Biomol NMR 27:313–321

Szyperski T, Banecki B, Braun D, Glaser RW (1998) Sequential resonance assignment of medium-sized $^{15}N/^{13}C$-labeled proteins with projected 4D triple resonance NMR experiments. J Biomol NMR 11:387–405

Szyperski T, Yeh DC, Sukumaran DK, Moseley HN, Montelione GT (2002) Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 99:8009–8014

Tian F, Valafar H, Prestegard J (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 123: 11791–11796

van de Ven FJ (1990) PROSPECT, a program for automated interpretation of 2D NMR spectra of proteins. J Magn Reson 86:633–644

Vitek O, Bailey-Kellogg C, Craig B, Kuliniewicz P, Vitek J (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. Bioinformatics 21:230–236

Vitek O, Bailey-Kellogg C, Craig B, Vitek J (2006) Interential backbone assignment for sparse data. J Biomol NMR 35: 187–208

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinás M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59:687–696

Wan X, Lin G (2006) A graph-based automated NMR backbone resonance sequential assignment. In: Computational systems bioinformatics 2006 conference proceedings, vol 4, pp 55–66

Wan X, Xu D, Slupsky CM, Lin G (2003) Automated protein NMR resonance assignments. In: Proceedings of the IEEE computer society conference on bioinformatics, vol 2, pp 197–208

Wang J, Wang T, Zuiderweg ER, Crippen GM (2005) CASA: an efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. J Biomol NMR 33:261–279

Wehrens R, Buydens L, Kateman G (1991) Validation and refinement of expert systems—interpretation of NMR-spectra as an application in analytical-chemistry. Chemometr Intell Lab Syst 12:57–67

Wehrens R, Lucasius C, Buydens L, Kateman G (1993a) HIPS, a hybrid self-adapting expert system for nuclear magnetic resonance spectrum interpretation using genetic algorithms. Anal Chim Acta 277:313–324

Wehrens R, Lucasius C, Buydens L, Kateman G (1993b) Sequential assignment of 2D-NMR spectra of proteins using genetic algorithms. J Chem Inf Comput Sci 33:245–251

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigation of nearest-neighbour effects. J Biomol NMR 5:67–81

Wu KP, Chang JM, Chen JB, Chang CF, Wu WJ, Huang TH, Sung TY, Hsu WL (2006) RIBRA—an error-tolerant algorithm for the NMR backbone assignment problem. J Comput Biol 13:229–244

Xu J, Sanctuary B (1993) CPA: constrained partitioning algorithm for initial assignment of protein ^{1}H resonances from MQF-COSY. J Chem Inf Comput Sci 33:490–500

Xu J, Strauss S, Sanctuary B, Trimble L (1994) Use of fuzzy mathematics for complete automated assignment of peptide 1H 2D NMR spectra. J Magn Reson B103:53–58

Xu Y, Xu D, Kim D, Olman V, Razumovskaya J, Jiang T (2002) Automated assignment of backbone NMR peaks using constrained bipartite matching. Comput Sci Eng 4:50–62

Xu Y, Wang X, Yang J, Vaynberg J, Qin J (2006) PASA—a program for automated protein NMR backbone signal assignment by pattern-filtering approach. J Biomol NMR 34:41–56

Zimmerman DE, Montelione GT (1995) Automated analysis of nuclear magnetic resonance assignments for proteins. Curr Opin Struct Biol 5:664–673

Zimmerman D, Kulikowski C, Wang L, Lyons B, Montelione GT (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. J Biomol NMR 4:241–256

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien Cy, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610

The interface of the sequential assignment program (not the post-MUSIC functionality), as written in python/nmrpython is shown in figure 2.15 and the code is given in Appendix A

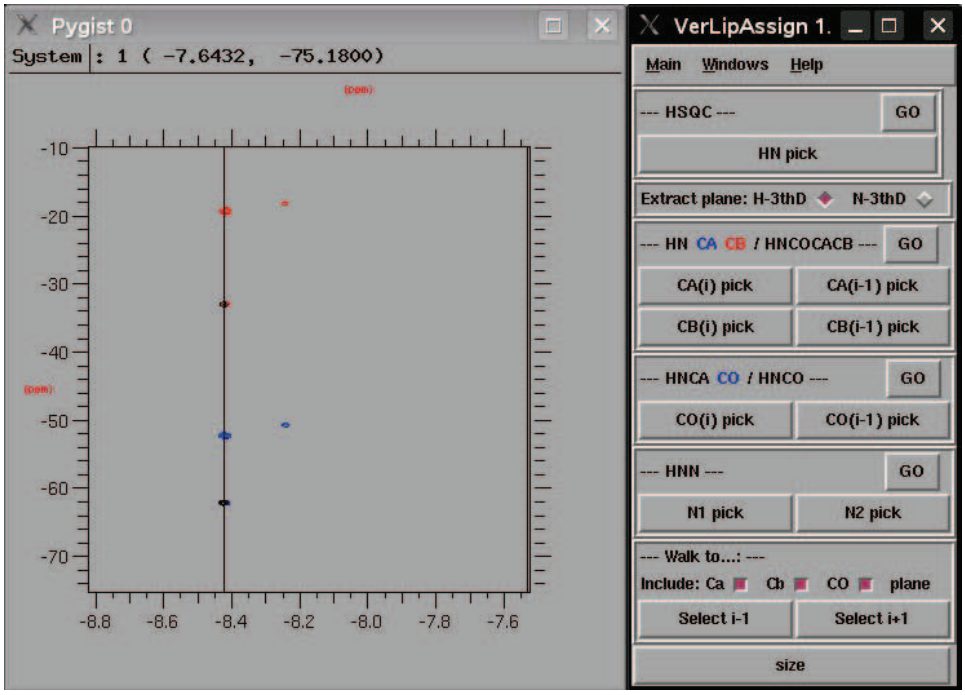


Figure 2.15. The sequential assignment tool implemented in python/NMRpython in its final version.

The plotting window (left panel) and the main (top-level) window (right panel) form the principle parts of the program. More details on the origin of these windows is given in chapter 4. The menu bar of the main window contains three menus: Main, Windows and Help. The main menu contains some items to import spectra, clear memory, ... The three items in the windows menu are "peak signs window", "levels window" and "info window" and open upon clicking, as can be expected, each an additional window. The peak signs window allows to change the expected intensity sign of different resonance signals. Default signs are: C_α(i): +; C_β(i): -; C_α(gly)(i): +; C_α(i-1): +; C_β(i-1): -; C_α(gly)(i-1): +; C'(i): +; C'(i-1): +; N(i+1)/N(i-1): +. The levels window permits to change the contour levels of the different spectra and spectral slices. Finally, the info window allows one to recover the latest clicked values of the different resonances (C_α, C_β, ...) in an easily interpretable format and also to write back this information to a file.

The different functions required for the actual assignment are accessible via the top-level window. The "HSQC GO" button contours the HSQC in the plotting window. Clicking "HN pick" allows to select a peak (proton/nitrogen chemical shift combination) in the HSQC. Both chemical shift values can then be used to indicate the plane to extract from the different triple resonance spectra. The user is given the choice of horizontal or vertical plane extracting (H-3thD vs. N-3thD). This means that e.g. in the case of the HNCACB extracted planes can have the dimensions H-C or N-C. The choice for either option should depend on in which HSQC dimension (H or N) the most overlap occurs for the selected peak. Once an HSQC nitrogen (or hydrogen) chemical shift is selected, the corresponding H-X

(or N-X) plane are extracted from the 3D spectra by clicking the different other "GO" buttons. The example of fig. 2.15 shows the superimposed H-C plane from the HNCACB (blue and red signals on screen, grey signals in image) and HN(CO)CACB (black signals) spectrum of an arbitrary protein after selecting an arbitrary HSQC peak. In these different spectral slices, individual resonances can again be selected using the "pick" buttons. The neighbouring (i-1 or i+1) residue can finally be found by calculating and plotting the product planes, which is done by clicking the "Select" buttons. For these calculations one can leave out any of the involved spectral slices (C_α, C_β, CO) when these contain weak or uninterpretable signals.

As mentioned in the article, we have also wanted to implement the described functionality as a macro in the Ccpnmr suite, developed at Cambridge University, in order to reach the largest possible NMR community. Due to the fact that their software was initially completely unadapted to perform spectral plane summations and multiplications, this implementation process was much longer than expected. Today, the macro still only exists as a test version (see fig. 2.16). Both our laboratory and the people of Ccpnmr are working on this issue. The test version contains most of the functionality, but runs very slow. The macro is straightforward to use if one understands the assignment principle and is used to the Ccpnmr interface.

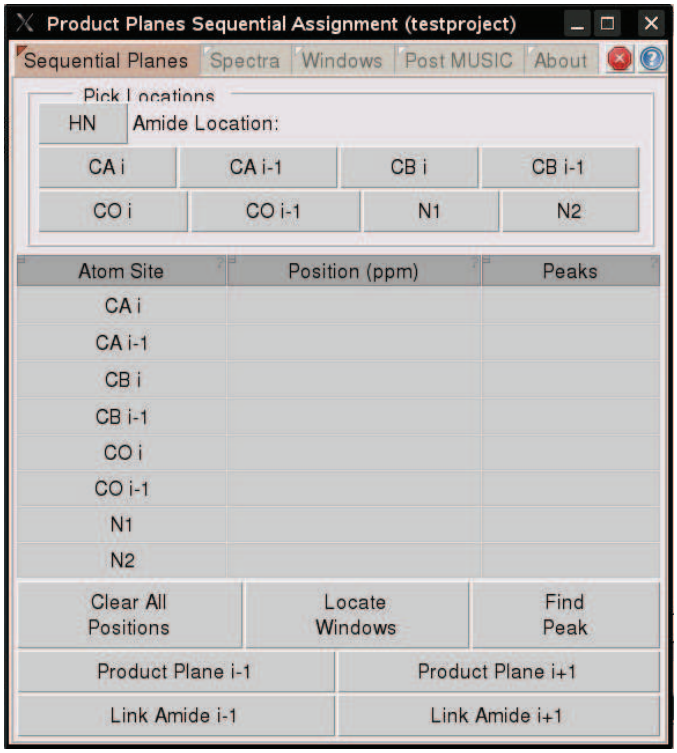


Figure 2.16. The sequential assignment macro as implemented in ccpnmr.

Concludingly, this tool was applied to assign several proteins. Generally, one needs only 1-2 working days for the full-confident assignment of each 100 residues in a polypeptide chain, independent of the fact it concerns a structured or rather an unstructured protein.

Chapter 3

Application to Human Tau and HCV's NS5A

3.1. Tau

One of the main technical hurdles of this thesis was trying to obtain an as complete as possible NMR assignment of the longest isoform, 441-residue Tau protein (htau40). Indeed, this assignment has been a major challenge in structural biology for several years, and was the initial stimulus for the development of the described assignment tool. The sheer size of this protein and the amount of peak overlap have impeded complete assignments until very recently [284]. Generally, researchers have pleased themselves with assignments of smaller fragments containing certain zones of interest (proline rich, repeats, ...) (e.g. [285, 117]), or incomplete assignments [244, 364]. In their struggle to obtain a full htau40 NMR assignment (more precisely 98-99% of the non-proline residues were assigned), Zweckstetter and co-workers [284] have also concentrated on the assignment of smaller fragments of Tau to, only afterwards, proceed to the full-length protein (see fig. 3.1). To achieve this, they have used a combination of manual and Sparky-aided [159] sequential assignments. It is perhaps interesting to note that the authors of [284] stress considerably on the importance of the premolten globule like structures the protein could populate, where several previous studies have clearly indicated the random coil nature of the protein [72, 341]. The former argue that several long-range interactions prime tau for MT binding and make phosphorylation at certain sites also influence behaviour of remote sequence regions.

Anyway, initially, the assignment of smaller fragments to reconstruct the full assignment has equally been our method of use. For this purpose, fragments F3 and F5 (fig. 3.1) were completely assigned (although their assignment as such can be of biological relevance as well). The spectra engaged were the usual set of HNCACB, HN(CO)CACB, HNCO, HN(CA)CO and HNN. The Tau F3 sample contained 250 μM protein in 0.6 mL solution, 25 mM Tris-D11, 25 mM NaCl, 2.5 mM EDTA, 2.5 mM DTT, 5% D_2O at T = 293 K and pH = 6.8. The Tau F5 sample contained 437 μM protein in 0.4 mL solution, 25 mM Tris-D11, 25 mM NaCl, 2.5 mM EDTA, 5% D_2O, at T = 293 K and pH = 6.80. The lists of assigned chemical shifts are given in appendix B.

However, when it was finally tried to make a full assignment of Tau P301L with the product plane assignment tool, these fragment assignments proved not completely indispensable. The Tau P301L sample used for the triple resonance assignment spectra contained 70 μM protein in 0.6 mL solution, 50 mM Na_2HPO_4/NaH_2PO_4, 25 mM NaCl, 2.5 mM EDTA, 1 mM DTT, 5% D_2O at T = 298 K and pH = 6.7. In five days time, 67% of the backbone/C_β resonances were assigned (see fig. 3.2), and this number was determined more by the absence of some peaks than by spectral overlap. Indeed, several HSQC signals had no corresponding intensity in

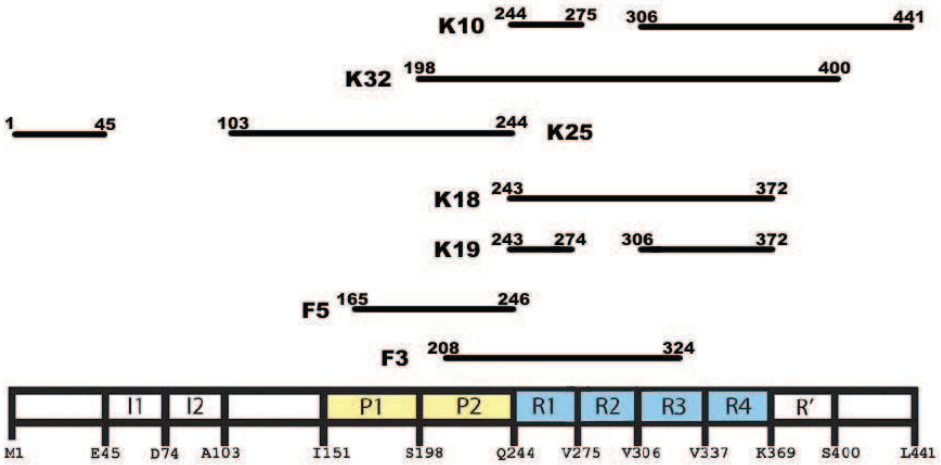


Figure 3.1. Indication of the different domains of Tau. I1 and I2 are the two inserts that are affected by alternative splicing (encoded by exons 2 and 3) and hence are absent in some smaller isoforms. P1 and P2 are the proline rich regions. R1-4 are the pseudo repeat regions. R3 is encoded by exon 10, the third (of eleven) exon subject to alternative splicing. Some of the different fragments that have been used in the study of Tau are indicated above the sequence. The constructs K25, K32 and K10 were used in addition to htau40 in the study of Mukrasch et al. [284] to obtain a full NMR assignment. K18 and K19 are two well-studied constructs as they contain the residue stretches involved in the aggregation of Tau into PHFs, and in the interaction with MTs. F3 and F5 are the fragments completely assigned with the product plane tool and that will be used for future Tau studies in the lab.

the 3D assignment spectra. It is unclear whether this is due to motional issues, pH or even non-perfect isotopic labelling. However, it is striking that unassigned zones of the Tau P301L sequence are often rich in arginine and lysine residues. Respectively the guanidino and amino side chain groups are characterised by high pKa values and therefore are protonated and positively charged except at very high pH values. Due to an inductive effect, this positive charge weakens the backbone H-N bond with greater tendencies for hydrogen exchange as a consequence. The line width broadening caused by this exchange is a possible explanation for the observed weaker intensities.

It would be interesting to repeat this assignment exercise with spectra derived from a different sample, obtained at slightly lower pH and at a few different temperatures, to know the intrinsic possibilities of our assignment tool. Higher temperatures generally increase the protein mobility and hence the spectral resolution. On the other hand, increasing temperature evens out the population of states, thereby lowering the sensitivity of NMR and the intensity of the signals, and also increase the amide hydrogen exchange rates. Temperature and pH should also not be taken too different from their physiological values. It has been demonstrated that major structural changes can occur in IDPs in the range of 3 to 30-50 °C. The structuration effects of elevated temperature may be attributed to increased strength of the hydrophobic interaction at higher temperatures, leading to a stronger hydrophobic driving force for folding [398]. Decreases or increases in pH also induce partial folding of intrinsically disordered proteins due to the minimisation of their large net charge present at neutral pH, thereby decreasing charge/charge intramolecular repulsion and permitting hydrophobic-driven collapse to the partially folded intermediate [398].

```
        10         20         30         40         50         60
MAEPRQEFEV MEDHAGTYGL GDRKDQGGYT MHQDQEGDTD AGLKESPLQT PTEDGSEEPG

        70         80         90        100        110        120
SETSDAKSTP TAEDVTAPLV DEGAPGKQAA AQPHTEIPEG TTAEEAGIGD TPSLEDEAAG

       130        140        150        160        170        180
HVTQARMVSK SKDGTGSDDK KAKGADGKTK IATPRGAAPP GQKGQANATR IPAKTPPAPK

       190        200        210        220        230        240
TPPSSGEPPK SGDRSGYSSP GSPGTPGSRS RTPSLPTPPT REPKKVAVVR TPPKSPSSAK

       250        260        270        280        290        300
SRLQTAPVPM PDLKNVKSKI GSTENLKHQP GGGKVQIINK KLDLSNVQSK CGSKDNIKHV

       310        320        330        340        350        360
LGGGSVQIVY KPVDLSKVTS KCGSLGNIHH KPGGGQVEVK SEKLDFKDRV QSKIGSLDNI

       370        380        390        400        410        420
THVPGGGNKK IETHKLTFRE NAKAKTDHGA EIVYKSPVVS GDTSPRHLSN VSSTGSIDMV

       430        440
DSPQLATLAD EVSASLAKQG L
```

Figure 3.2. The 67% residues of Tau P301L that were assigned using the product plane approach.

The main reason, besides their perseverance, why Mukrasch et al. [284] were able to get to a full assignment of full-length Tau is they have worked at a very high field strength (21.1 Tesla; 900 MHz). The Larmor frequency of a nucleus is determined by $\omega_0 = -\gamma(1-\sigma)B_0$, where γ is the nucleus type specific gyromagnetic ratio, B_0 is the external magnetic field and σ a factor accounting for the electronic environment of a specific nucleus. It follows that the difference in Larmor frequency between two individual spins (with a different σ) increases with the increasing magnetic B_0 field. This results in spectra with higher resolution, i.e., the individual peaks are more separated. Such an increased resolution is crucial for unstructured proteins. In addition, increasing the magnetic field strength also enhances sensitivity. The fundamental relationship describing the influences on the signal-to-noise ratio (N/S) is:

$$S/N \propto n\gamma_e\sqrt{\gamma_d^3 B_0^3 t} \tag{3.1}$$

where n is the number of nuclear spins being observed, γ_e is the gyromagnetic ratio of the spin being excited, γ_d is the gyromagnetic ratio of the spin being detected, B_0 is the magnetic field strength, and t is the experiment acquisition time. It is obvious from this equation that the higher the magnetic field, the better the sensitivity. The highest field available for the described studies of Tau has been 18.8 Tesla, 800 MHz. Moreover, this spectrometer is not equipped with a cryogenically cooled probe (cryoprobe). These latter enable typically a 3-4-fold enhancement of the detection sensitivity in high-resolution NMR compared to the corresponding conventional probes, by lowering the temperature of the coil and the preamplifier, thus essentially reducing the thermal noise in the receiver circuitry.

In this view, the arrival of the 900 MHz spectrometer (with cryoprobe) in our group, will probably allow for further very promising biological investigations. The group has been interested in the study of the phosphorylation of Tau and the influence of this phosphorylation on Tau's binding to and

stabilising of microtubules, and on the aggregation of the Tau protein. Previous incomplete NMR assignments have already allowed the identification and quantification of the phosphorylation of Tau by certain (complexes of) kinases [230, 9]. Also the Alzheimer's-like aggregated paired helical filament form of Tau has already been studied. By following up the HSQC peak intensity decrease (in both solution and magic angle spinning NMR), the motion-sensitive heteronuclear NOE data and H/D exchange data of heparin-induced Tau assembly into PHFs, the regions involved in the Tau aggregation were characterised [360, 361]. Using the same HSQC intensity profile method, the binding domains of Tau, pSer214-Tau (which has been reported to decrease the interaction between Tau and the MTs) and an oxidised state of Tau (containing an intramolecular disulphide bridge between residues Cys291 and Cys322) upon binding to Paclitaxel (Taxol)-stabilised microtubules have been characterised [359]. Equally studied in the past were the binding regions (and binding strengths involved) of Tau with heparin, a polyanion causing Tau aggregation without the need for posttranslational phosphorylation. Interestingly, the use of several earlier described NMR methods (chemical shift deviations, $^3J_{HN,H\alpha}$ couplings, RDCs, and NOE data) showed a structural impact of heparin binding on full-length Tau. It was hence proposed that through increased residual β-sheet propensity within peptides of the R2 and R3 repeat domains and charge neutralisation followed by an overall structural change in the basic regions flanking the microtubule binding repeats, enables aggregation of Tau into PHFs [357].

However, several aspects, such as the precise orientation and structure adopted by Tau (and its several phosphorylated forms) when bound microtubules are still unknown. The influence of single point mutations such as in Tau P301L on the protein's behaviour is equally unknown. Armed with the new spectrometer, experiments are planned that might give further insight in these issues. For example, the interaction of Tau with the T2R complex (see fig. 1.11 on page 40) will be investigated. The availability of more assigned residues will in this matter only increase the amount of possible retrievable information.

3.2. NS5A

As mentioned in the introductory chapter, the possibility of fast NMR assignments has equally allowed a study of the non-structural 5A (NS5A) protein of the Hepatitis C virus. This study is divided in three parts. Part I concentrates on domain 2 of NS5A in the JFH1 virus strain (a variant of the 2a genotype/subtype). We mainly focus on the interaction of this domain with Cyclophilins A and B. Part II briefly discusses the intrinsically disordered nature of the third domain of NS5A in the Con1 strain (belonging to the 1b genotype/subtype). In Part III, we further investigate the residual structure of NS5A-D3 and its interaction with Cyclophilin A. This is done in a comparative manner, confronting the behaviour of NS5A-D3 of the Con1 and JFH1 strain.

3.2.1. Hepatitis C Virus NS5A Protein is a Substrate for the Peptidyl-Prolyl Cis/Trans Isomerase Activity of Cyclophilins A and B

Supplemental Material can be found at:
http://www.jbc.org/cgi/content/full/M809244200/DC1

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 284, NO. 20, pp. 13589–13601, May 15, 2009

Hepatitis C Virus NS5A Protein Is a Substrate for the Peptidyl-prolyl *cis/trans* Isomerase Activity of Cyclophilins A and B*[S]

Received for publication, December 9, 2008, and in revised form, February 11, 2009 Published, JBC Papers in Press, March 18, 2009, DOI 10.1074/jbc.M809244200

Xavier Hanoulle[‡1], Aurélie Badillo[§], Jean-Michel Wieruszeski[‡], Dries Verdegem[‡], Isabelle Landrieu[‡], Ralf Bartenschlager[¶2], François Penin[§], and Guy Lippens[‡3]

From the [‡]Unité de Glycobiologie Structurale et Fonctionnelle, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France, [§]Institut de Biologie et Chimie des Protéines, UMR 5086, CNRS, Université de Lyon, IFR 128, BioSciences Gerland-Lyon Sud, F-69397 Lyon, France, and [¶]Department of Molecular Virology, University of Heidelberg, Im Neuenheimer Feld 345, 69120 Heidelberg, Germany

We report here a biochemical and structural characterization of domain 2 of the nonstructural 5A protein (NS5A) from the JFH1 Hepatitis C virus strain and its interactions with cyclophilins A and B (CypA and CypB). Gel filtration chromatography, circular dichroism spectroscopy, and finally NMR spectroscopy all indicate the natively unfolded nature of this NS5A-D2 domain. Because mutations in this domain have been linked to cyclosporin A resistance, we used NMR spectroscopy to investigate potential interactions between NS5A-D2 and cellular CypA and CypB. We observed a direct molecular interaction between NS5A-D2 and both cyclophilins. The interaction surface on the cyclophilins corresponds to their active site, whereas on NS5A-D2, it proved to be distributed over the many proline residues of the domain. NMR heteronuclear exchange spectroscopy yielded direct evidence that many proline residues in NS5A-D2 form a valid substrate for the enzymatic peptidyl-prolyl *cis*/*trans* isomerase (PPIase) activity of CypA and CypB.

Hepatitis C virus (HCV)[4] is a small, positive strand, RNA-enveloped virus belonging to the Flaviviridae family and the genus *Hepacivirus*. With 120–180 million chronically infected individuals worldwide, hepatitis C virus infection represents a major cause of chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma (1). The HCV viral genome (~9.6 kb) codes for a unique polyprotein of ~3000 amino acids (recently reviewed in Refs. 2–4). Following processing via viral and cellular proteases, this polyprotein gives rise to at least 10 viral proteins, divided into structural (core, E1, and E2 envelope glycoproteins) and nonstructural proteins (p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B). Nonstructural proteins are involved in polyprotein processing and viral replication. The set composed of NS3, NS4A, NS4B, NS5A, and NS5B constitutes the minimal protein component required for viral replication (5).

Cyclophilins are cellular proteins that have been identified first as CsA-binding proteins (6). As FK506-binding proteins (FKBP) and parvulins, cyclophilins are peptidyl-prolyl *cis*/*trans* isomerases (PPIase) that catalyze the *cis/trans* isomerization of the peptide linkage preceding a proline (6, 7). Several subtypes of cyclophilins are present in mammalian cells (8). They share a high sequence homology and a well conserved three-dimensional structure but display significant differences in their primary cellular localization and in abundance (9). CypA, the most abundant of the cyclophilins, is primarily cytoplasmic, whereas CypB is directed to the endoplasmic reticulum lumen or the secretory pathway. CypD, on the other hand, is the mitochondrial cyclophilin. Cyclophilins are involved in numerous physiological processes such as protein folding, immune response, and apoptosis and also in the replication cycle of viruses including vaccinia virus, vesicular stomatitis virus, severe acute respiratory syndrome (SARS)-coronavirus, and human immunodeficiency virus (HIV) (for review see Ref. 10). For HIV, CypA has been shown to interact with the capsid domain of the HIV Gag precursor polyprotein (11). CypA thereby competes with capsid domain/TRIM5 interaction, resulting in a loss of the antiviral protective effect of the cellular restriction factor TRIM5α (12, 13). Moreover, it has been shown that CypA catalyzes the *cis*/*trans* isomerization of Gly^{221}-Pro^{222} in the capsid domain and that it has functional consequences for HIV replication efficiency (14–16). For HCV, Watashi *et al.* (17) have described a molecular and functional interaction between NS5B, the viral RNA-dependent RNA polymerase (RdRp), and cyclophilin B (CypB). CypB may be a key regulator in HCV replication by modulating the affinity of NS5B for RNA. This regulation is abolished in the presence of cyclosporin A (CsA), an inhibitor of cyclophilins (6). These results provided for the first time a molecular mechanism for the early-on observed anti-HCV activity of CsA (18–20). Although this initial report suggests that only CypB would be involved in the HCV replication proc-

* This work was supported by the French Centre National de la Recherche Scientifique (CNRS) and the Universities of Lille and Lyon and by grants from the French National Agency for Research on AIDS and Viral Hepatitis and the European Commission (VIRGIL Network of Excellence on Antiviral Drug Resistance).

The ^{1}H,^{15}N, and ^{13}C backbone resonances for this protein are available in the Biological Magnetic Resonance Data Bank under BMRB accession number 16165.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. 1–6 and Table 1.

[1] Supported by a fellowship from the French National Agency for Research on AIDS and Viral Hepatitis.

[2] Supported by the German Research Council (Contract BA 1505/2-1).

[3] To whom correspondence should be addressed. Tel.: 33-3-20-33-72-41; Fax: 33-3-20-43-65-55; E-mail: guy.lippens@univ-lille1.fr.

[4] The abbreviations used are: HCV, hepatitis C virus; aa, amino acid; CsA, cyclosporin A; Cyp, cyclophilin; EXSY, exchange spectroscopy; HIV, human immunodeficiency virus; HSQC, heteronuclear single quantum correlation; NMR, nuclear magnetic resonance; NOESY, nuclear Overhauser enhancement spectroscopy; NS5A, nonstructural protein 5A; PPIase, peptidyl-prolyl *cis/trans* isomerase; TFE, 2,2,2-trifluoroethanol; IPTG, isopropyl 1-thio-β-D-galactopyranoside; DTT, dithiothreitol; NPSA, network protein sequence analysis; CSI, chemical shift index.

ess (17), a growing number of studies have recently pointed out a role for other cyclophilins (21–25).

In vitro selection of CsA-resistant HCV mutants indicated the importance of two HCV nonstructural proteins, NS5B and NS5A (26), with a preponderant effect for mutations in the C-terminal half of NS5A. NS5A is a large phosphoprotein (49 kDa), indispensable for HCV replication and particle assembly (27–29), but for which the exact function(s) in the HCV replication cycle remain to be elucidated. This nonstructural protein is anchored to the cytoplasmic leaflet of the endoplasmic reticulum membrane via an N-terminal amphipathic α-helix (residues 1–27) (30, 31). Its cytoplasmic sequence can be divided into three domains: D1 (residues 27–213), D2 (residues 250–342), and D3 (residues 356–447), all connected by low complexity sequences (32). D1, a zinc-binding domain, adopts a dimeric claw-shaped structure, which is proposed to interact with RNA (33, 34). NS5A-D2 is essential for HCV replication, whereas NS5A-D3 is a key determinant for virus infectious particle assembly (27, 35). NS5A-D2 and -D3, for which sequence conservation among HCV genotypes is significantly lower than for D1, have been proposed to be natively unfolded domains (28, 32). Molecular and structural characterization of NS5A-D2 from HCV genotype 1a has confirmed the disordered nature of this domain (36, 37).

As it is still not clear which cyclophilins are cofactors for HCV replication, and as mutations in HCV NS5A protein have been associated with CsA resistance, we decided to examine the interaction between both CypA and CypB and domain 2 of the HCV NS5A protein. We first characterized, at the molecular level, NS5A-D2 from the HCV JFH1 infectious strain (genotype 2a) and showed by NMR spectroscopy that this natively unfolded domain indeed interacts with both cyclophilin A and cyclophilin B. Our NMR chemical shift mapping experiments indicated that the interaction occurs at the level of the cyclophilin active site, whereas it lacks a precise localization on NS5A-D2. A peptide derived from the only well conserved amino acid motif in NS5A-D2 did interact with cyclophilin A but only with a 10-fold lower affinity than the full domain. We concluded from this that the many proline residues form multiple anchoring points, especially when they adopt the *cis* conformation. NMR exchange spectroscopy further demonstrated that NS5A-D2 is a substrate for the PPIase activities of both CypA and CypB. Both the NS5A/cyclophilin interaction and the PPIase activity of the cyclophilins on NS5A-D2 were abolished by CsA, underscoring the specificity of the interaction.

EXPERIMENTAL PROCEDURES

Sequence Analysis—Sequence analyses were performed using tools available at the Institut de Biologie et Chimie des Protéines (IBCP) network protein sequence analysis (NPSA) website (38). HCV NS5A sequences were retrieved from the European HCV Database (39). Multiple sequence alignments were performed with ClustalW (40) using default parameters. The repertoire of residues at each amino acid position and their frequencies observed in natural sequence variants were computed by the use of a program developed at the IBCP.[5]

[5] F. Dorkeld, C. Combet, F. Penin, and G. Deléage, unpublished data.

Expression and Purification of Nonlabeled and ^{15}N- and $^{15}N,^{13}C$-Labeled NS5A-D2 (JFH1)—The synthetic sequence coding for domain 2 of the HCV NS5A protein from JFH1 strain (euHCVdb (39); GenBank™ accession number AB047639, genotype 2a) was introduced in the bacterial expression vector pT7.7 with a His_6 tag (41). The resulting recombinant domain 2 of HCV NS5A (NS5A-D2; residues 248–341) has extra M- and -LQHHHHHH extensions at the N and C termini, respectively. The pT7–7-NS5A-D2 plasmid was introduced in *Escherichia coli* BL21(DE3) (Merck-Novagen, Darmstadt, Germany). Cells were grown at 37 °C in Luria-Bertani (LB) medium for nonlabeled protein or in a M9-based semi-rich medium (M9 medium supplemented with $[^{15}N]NH_4Cl$ (1 g/liter), D-$[^{13}C_6]$glucose (2 g/liter) (for ^{13}C labeling only), Isogro $^{13}C,^{15}N$ powder growth medium (1 g/liter, 10%; Sigma-Aldrich). At an A_{600} of ~0.7, the protein production was induced with 0.4 mM isopropyl 1-thio-β-D-galactopyranoside (IPTG), and cells were harvested by centrifugation at 3.5 h post-induction. NS5A-D2 was first purified by Ni^{2+}-affinity chromatography (HisTrap column, 1 ml, GE Healthcare Europe). Selected fractions were pooled, dialyzed against 20 mM Tris-Cl, pH 7.4, 2 mM EDTA, and then submitted to a second purification step by ion exchange chromatography (ResourceQ, 1 ml column, GE Healthcare Europe). Following SDS-PAGE analysis, NS5A-D2-containing fractions were selected and pooled. The protein was concentrated up to 340 μM with a Vivaspin 15 concentrator (cutoff, 5 kDa) (Satorius Stedim Biotech, Aubagne, France) while simultaneously exchanging the buffer against 20 mM NaH_2PO_4/Na_2HPO_4 pH 6.4, 30 mM NaCl, 1 mM DTT (or 1 mM Tris(hydroxypropyl)phosphine), 0.02% NaN_3. After filtration (0.2 μm), NS5A-D2 aliquots were stored at −80 °C with a few Chelex 100 beads (Sigma-Aldrich).

Circular Dichroism (CD)—CD spectra were recorded on a Chirascan dichrographe (Applied Photophysics, Surrey, UK) calibrated with (1*S*)-(+)-10-camphorsulfonic acid. Measurements were carried out at room temperature in a 0.1-cm path length quartz cuvette with protein concentrations ranging from 5 to 15 μM. Spectra were recorded in the 185–260 nm wavelength range at 0.5-nm increments and a 2-s integration time. Spectra were processed, base-line-corrected, and smoothed using Chirascan software. Spectral units were expressed as the molar ellipticity per residue using protein concentrations determined by measuring the UV light absorbance of tyrosine and tryptophan at 280 nm. The α-helix content was estimated using the method of Chen *et al.* (42).

Peptide Synthesis—A synthetic peptide (named PepD2) corresponding to residues 304–323 of NS5A (^{304}GFPRALPAWARPDYNPPLVE^{323}) was obtained from Neosystems (Strasbourg, France). The purity of the peptide was verified by high pressure liquid chromatography and mass spectrometry as greater than 95%.

Expression and Purification of Nonlabeled and $^{15}N,^{13}C$-Labeled Cyclophilin B—Production and purification of recombinant human cyclophilin B in *E. coli* were done as described previously (43). Briefly, the *pET15b-CypB* plasmid was introduced into *E. coli* BL21(DE3) strain, recombinant bacteria were grown in LB medium (or in M9 medium supplemented with $[^{15}N]NH_4Cl$ and $[^{13}C]$glucose for labeled samples), and produc-

tion was induced with 0.4 mM IPTG. Cyclophilin B was purified by ion exchange (SP Sepharose Fast Flow) and then by gel filtration (Superose 12 Prep Grade) chromatography. The purified and concentrated Cyclophilin B was stored at −80 °C.

Expression and Purification of Nonlabeled and $^{15}N,^{13}C$-Labeled Cyclophilin A—Sequence coding for human CypA was amplified from the plasmid pKK233–2-CypA, kindly provided by Prof. Allain (UMR8576, CNRS-University of Sciences and Technologies of Lille, France), using the forward primer 3′-cttcatatggtcaaccccaccgtg-5′ and the reverse primer 5′-caaggatcctattcgagttgtcc-3′ and was then inserted in the pET15b plasmid (Merck-Novagen) between the NdeI and BamHI restriction sites. The *pET15b-CypA* plasmid, coding for a recombinant CypA with an N-terminal His tag, was introduced in *E. coli* BL21 (DE3). Cells were grown in M9 medium supplemented with [^{15}N]NH_4Cl or [^{15}N]NH_4Cl and [^{13}C]glucose. When the culture reach an A_{600} = ~0.8, protein production was induced with 0.4 mM IPTG; cells were harvested 3 h after induction at 37 °C. Recombinant CypA was purified by Ni^{2+}-affinity chromatography (HiTrap Chelating HP, GE Healthcare Europe). Finally, the protein was dialyzed against 50 mM NaH_2PO_4/Na_2HPO_4, pH 6.3, 20 mM NaCl, 2 mM EDTA, 1 mM DTT, concentrated, filtered (0.2 μm), and then stored at 4 °C.

NMR Data Collection and Assignments—Spectra were acquired on either a Bruker Avance 600 MHz equipped with a cryogenic triple resonance probe head or a Bruker Avance 800 MHz with a standard triple resonance probe (Bruker, Karlsruhe, Germany). The proton chemical shifts were referenced using the methyl signal of TMSP (sodium 3-trimethylsill-[2,2′,3,3′-d4]propionate) at 0 ppm. Spectra were processed with the Bruker TopSpin software package 1.3 and analyzed using the product plane approach developed in our laboratory (44).

Assignments of NS5A-D2 backbone resonances were achieved using two-dimensional ^{1}H,^{15}N HSQC and three-dimensional ^{1}H,^{15}N,^{13}C HNCO, HNCACO, HNCACB, HNCOCACB, and HNCANNH spectra (45) acquired at 600 MHz on a 340 μM ^{15}N,^{13}C-labeled NS5A-D2 sample at 298 K (Biological Magnetic Resonance Data Bank (BMRB) accession number 16165).

Assignments of the CypB spectrum were taken from our previous study (43). Assignments of CypA resonances were taken from the literature (46) and confirmed with a HNCACB spectrum acquired at 600 MHz on a 340 μM [^{15}N,^{13}C]CypA sample in 50 mM NaH_2PO_4/Na_2PO_4 pH 6.3, 40 mM NaCl, 2 mM EDTA, 1 mM DTT at 25 °C.

Interaction between NS5A-D2 and Cyclophilins—To study the interaction between NS5A-D2 and CypA or CypB, differentially labeled proteins (^{15}N for NS5A-D2 and ^{15}N,^{13}C for CypA or CypB) were mixed at different molar ratios. The (^{1}H,^{15}N) plane of the HNCO spectrum thereby selects only for the ^{15}N,^{13}C-labeled protein component, whereas the HNCO spectrum with modified phases to select for the non-^{13}C-labeled ^{15}N nuclei (which we will further call the HN(noCO) spectrum (47)) was used for selection of the only-^{15}N-labeled protein. The combined chemical shift perturbations following NS5A-D2 addition were calculated as shown in Equation 1, whereby $\delta\Delta(^1HN)$ and $\delta\Delta(^{15}N)$ are the chemical shift perturbations in the ^{1}H and ^{15}N dimensions, respectively.

$$\delta\Delta = |\delta\Delta(^1HN)| + 0.2 \cdot |\delta\Delta(^{15}N)| \quad \text{(Eq. 1)}$$

Cyclophilin PPIase Activity toward NS5A-D2—PPIase activity of CypA and CypB on NS5A-D2 were assessed using EXSY spectra, whereby the exchange was monitored on the proton resonance (in homonuclear ^{1}H,^{1}H spectra (48)) or on the ^{15}N nucleus (in heteronuclear ^{1}H,^{15}N z-exchange spectra (49)). The ratio between the *cis* and *trans* populations for a given residue (p_c and p_t, respectively) was measured on the basis of a ^{1}H,^{15}N HSQC spectrum in the absence of any cyclophilin assuming that an exchange peak for this residue was observed.

^{1}H,^{1}H EXSY spectra were acquired as ^{1}H,^{1}H planes from a three-dimensional ^{15}N-edited NOESY-HSQC with different mixing times (50, 100, 200, and 400 ms) on a sample of 320 μM [^{15}N,^{13}C]NS5A-D2 and 40 μM [^{15}N]CypB or -CypA in 20 mM NaH_2PO_4/Na_2HPO_4, pH 6.4, 30 mM NaCl, 0.02% NaN_3, 1 mM DTT.

^{15}N z-exchange spectra were recorded on an 800-MHz spectrometer with 0.88, 25, 50, 100, 200, 300, and 400-ms mixing times. PPIase activities were analyzed on a sample of 220 μM [^{15}N]NS5A-D2 and 23 μM CypB or CypA in 20 mM NaH_2PO_4/Na_2HPO_4 pH 6.3, 30 mM NaCl, 0.02% NaN_3, 1 mM DTT. Exchange rates were derived from a simplified version of the analytical form given in Ref. 38 by taking into account only the maximal intensity of the *trans-cis* exchange peak (I_{tc}) and the *trans* diagonal peak (I_{tt}). This procedure minimized any problems with exchange broadening of the *cis* diagonal peak due to the interaction with the Cyp and with significant proton overlap hindering the reliable integration of the weak off-diagonal peaks. The exchange rate (k_{exch}), as a function of mixing time (M_T), was determined by using a least-squares fitting procedure between the experimental data and the theoretical Equation 2 adapted from Ref. 15.

$$\frac{I_{tc}}{I_{tt}} = \frac{-p_c + p_c \times e^{(k_{exch} \times M_T)}}{p_c + p_t \times e^{(k_{exch} \times M_T)}} \quad \text{(Eq. 2)}$$

To confirm that the exchange peaks were due to the PPIase activity of cyclophilins, CsA was added into the sample, and a ^{1}H,^{15}N z-exchange spectra was recorded with a 100-ms mixing time. PyMOL software was used for the molecular graphics (DeLano Scientific).

RESULTS

Sequence Analysis—We performed sequence analysis and structure predictions to assess the degree of conservation of the NS5A-D2 domains across the different strains and to identify potential essential amino acids (aa) and motifs. The aa repertoire deduced from the analysis of 21 HCV isolates of genotype 2a revealed that aa are strictly conserved in 70% of the sequence positions (denoted by *asterisks* in Fig. 1*A*). The apparent variability is limited at most positions because the observed residues exhibit similar physicochemical properties, as indicated both by the similarity pattern (Fig. 1*A*, *colons* and *dots*) as well as the hydropathic pattern, where the letters *o*, *i*, and *n* denote hydrophobic, hydrophilic, and neutral residues, respectively (see leg-

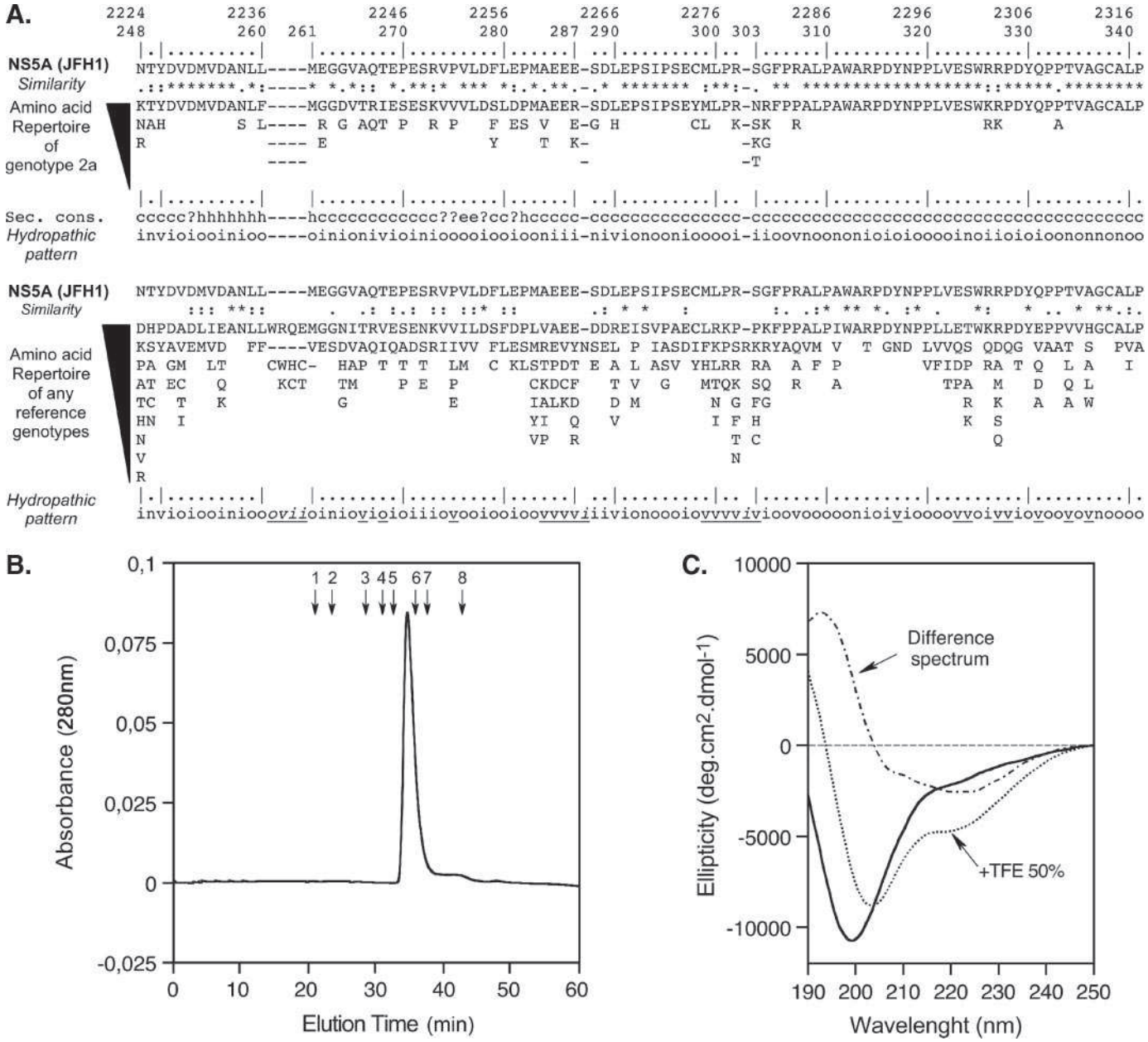


FIGURE 1. **Sequence analysis and biochemical characterization of NS5A-D2 domain 2 from HCV.** *A*, amino acid repertoires. The NS5A-D2 aa 248–341 sequence from the HCV JFH1 strain of genotype 2a (GenBank™ accession number AB047639), which was used in this study, is indicated. Amino acids are numbered with respect to NS5A and the HCV JFH1 polyprotein (*top row*). The *hyphens* indicate the aa deletions compared with the sequence alignment of any genotypes (see below). The aa repertoire deduced from the ClustalW multiple alignments of 21 NS5A sequences of genotype 2a is shown at the *top*. Amino acids observed at a given position less than twice were not included. At the *bottom* is the aa repertoire of the 27 representative NS5A sequences from confirmed HCV genotypes and subtypes (listed with accession numbers in Table 1 in Ref. 71; see the European HCV Database for details). The degree of aa and physicochemical conservation at each position can be inferred from the extent of variability (with the observed aa listed in decreasing order of frequency from top to bottom) together with the similarity index according to ClustalW convention (*asterisk*, invariant; *colon*, highly similar; *dot*, similar (40)) and the consensus hydropathic pattern deduced from the consensus aa repertoire: *o*, hydrophobic position (Phe, Ile, Trp, Tyr, Leu, Val, Met, Pro, Cys); *n*, neutral position (Gly, Ala, Thr, Ser); *i*, hydrophilic position (Lys, Gln, Asn, His, Glu, Asp, Arg); *v*, variable position (*i.e.* when both hydrophobic and hydrophilic residues are observed at a given position). A position that is *underlined* in the hydropathic pattern for any reference genotypes (*bottom*) indicates the change of position status when compared with the hydropathic pattern for genotype 2a (*middle*). Secondary structure predictions of NS5A-D2 are indicated as helical (*h*), extended (*e*), undetermined (coil (*c*)), or ambiguous (?). *Sec. Cons.*, consensus of protein secondary structures predictions for NS5A-D2 from JFH1 strain deduced from a large set of prediction methods available at the NPSA website, including DSC, HNNC, MLRC, PHD, Predator, SOPM, and SIMPA96 available at the NPSA website (see Ref. 38 and references therein). *B*, gel filtration analysis of NS5A-D2 was performed on a Superdex S200 column equilibrated in 50 mM sodium phosphate, pH 7.4, 1 mM Tris(2-carboxyethyl) phosphine hydrochloride with a flow rate of 0.5 ml/min. Elution volumes of globular protein standards are indicated by *black arrows* with the following corresponding molecular masses: *1*, thyroglobulin (669,000 Da); *2*, ferritin (44,0000 Da); *3*, aldolase (158,000 Da); *4*, conalbumin (75,000 Da); *5*, ovalbumin (43,000 Da); *6*, chymotrypsin (25,000 Da); *7*, ribonuclease (13,700 Da); *8*, vitamin B_{12} (1,355 Da). *C*, Far-UV circular dichroism analysis of 8 μM NS5A-D2 in 10 mM sodium phosphate, pH 7.4, 1 mM Tris(2-carboxyethyl) phosphine hydrochloride (*solid line*) complemented with 50% TFE (*dotted line*). The difference spectrum (*alternating dashed line*) was obtained by subtracting the latter spectrum from the former.

end to Fig. 1*A* for details). The degree of conservation among different genotypes was investigated by ClustalW alignment of 27 reference sequences representative for the major HCV genotypes and subtypes (see legend to Fig. 1*A*). The aa repertoire derived from this alignment revealed an apparent high level of variability, except for some well conserved positions that are

likely essential for the structure and/or function of NS5A-D2. However, despite this apparent variability, conservation of the hydropathic character at most positions indicates that the overall structure of NS5A-D2 is conserved among the different HCV genotypes. There are, however, some short variable stretches of sequences (*underlined* in Fig. 1*A*, *bottom*), which appear to be genotype-specific. Typically, a main sequence difference between genotypes is the four-aa deletion observed in genotype 2, including JFH1 (indicated by *hyphens* in Fig. 1*A*).

Molecular Characterization of HCV NS5A-D2 (JFH1)—NS5A-D2 is efficiently produced in a soluble form when recombinantly overexpressed in *E. coli* and could be purified to almost homogeneity (see supplemental Fig. 1). Despite an excellent agreement between expected (11,639 Da) and experimental mass as determined by mass spectroscopy, NS5A-D2 has an apparent molecular weight of ~18 kDa by SDS-PAGE. This discrepancy is probably due to the primary aa sequence of NS5A-D2, which includes many acidic residues and prolines (50, 51). In gel filtration chromatography, the protein elutes at a volume corresponding to a ~30-kDa globular protein, (Fig. 1*B*). Such a large apparent molecular weight in a gel filtration assay is commonly associated with natively unfolded proteins devoid of globular domain (52).

The structure of NS5A-D2 was further characterized by CD spectroscopy (Fig. 1*C*). In aqueous buffer, NS5A-D2 gave a complex spectrum with a large negative band around 198 nm and a shoulder in the 220–240 nm range, indicating a mixture of random coil structure with the presence of some poorly defined structures. To probe the potential conformational preference of NS5A-D2, we used TFE, which is known to stabilize the folding of peptidic sequences, especially those exhibiting an intrinsic propensity to adopt an α-helical structure (53). The addition of 50% TFE induced a limited structuration attributed to some α-helix formation. Indeed, the difference spectrum shown in Fig. 1*C* is consistent with a small amount of α-helical folding with a maximum at 192 nm and two minima at 208 and 222 nm. Assuming that the residue molar ellipticity at 222 nm is exclusively due to α-helix upon addition of TFE, a maximum of only about 6% α-helix content could be estimated, in agreement with the low level of α-helical structure predicted from aa sequence analysis (Fig. 1*A*).

The ^{1}H,^{15}N HSQC of NS5A-D2 (Fig. 2*A*) displays a narrow proton chemical shift range, limited to 1 ppm excluding three outlying peaks (Trp^{312}, Ala^{313}, and Arg^{326}; see below). This low level of dispersion again points to the nonstructured nature of the polypeptide, at least when isolated in solution. Using triple resonance NMR spectroscopy on a doubly labeled NS5A-D2 sample and an in-house developed product plane-based assignment procedure (44), all backbone amide proton resonances could be assigned except for the 15 proline residues. The outlying peaks were assigned to Trp^{312}, Ala^{313}, and Arg^{326} (Fig. 2). ^{13}CO, ^{13}Cα, and ^{13}Cβ resonances were assigned for 94 residues, and were used to probe the secondary structure content at a per-residue level in NS5A-D2. Carbon chemical shifts when compared with their values for the amino acid in a short unstructured peptide give a good indication of the secondary structure adopted by the amino acid in the full protein (54). Analysis of the chemical shift index (CSI) shows a majority of negative CSI values for ^{13}Cα and ^{13}CO, whereas the ^{13}Cβ CSI values are generally positive (Fig. 2*B*) (54). Although this hints at an extended structure, the CSI consensus values are zero all along the NS5A-D2 sequence, confirming the absence of stable secondary structure elements even at the local level.

Next to the assigned peaks, and despite the high level of purity obtained by our two-step purification procedure (supplemental Fig. 1), numerous, less intense peaks could be observed in the ^{1}H,^{15}N HSQC spectrum (Fig. 2*A*). Corresponding to residues in the vicinity of a proline in the *cis* conformation, 32 of these minor peaks could be assigned in the same triple resonance spectra used for the initial assignment (minor forms will be named *cis* forms in the following). Although the high content of proline residues (15 prolines in the 94-aa fragment of NS5A-D2) sometimes led to ambiguity regarding the identity of the *cis*-Pro at the origin of the chemical shift difference, the presence of several minor peaks corresponding to various residues around a given proline allowed the assignment and quantification of the *cis*/*trans* ratio for a major fraction of the prolyl bonds (supplemental Table 1).

Interaction between NS5A-D2 and Human Cyclophilins—As mutations in the C-terminal half of NS5A have been shown to confer CsA resistance for mutant HCV (26), we investigated the direct physical interaction between NS5A-D2 and cyclophilins. Although CypA is the prominent cytosolic isomerase (10, 25), the initial report of cyclophilins being involved in HCV replication suggests CypB as the corresponding partner (17). We therefore tested independently the interaction of NS5A-D2 with CypA and CypB. Finally, because we wanted to obtain, with a single sample, the chemical shift changes on both partners in order to map the mutual interaction surfaces, we mixed ^{15}N-labeled NS5A-D2 and ^{15}N,^{13}C-labeled CypA or CypB and used the planes from the HN(CO) and HN(noCO) experiments to obtain subspectra displaying only the one or the other molecular entity (47).

Comparing the Cyp subspectra in the absence and presence of an equimolar quantity of NS5A-D2, we noted that only a limited number of CypA or CypB resonances was affected (supplemental Fig. 2). Beyond proving the existence of a direct physical interaction between both partners, mapping the chemical shifts on the Cyp primary sequences and then on their respective three-dimensional structures allowed us to define precisely the interaction sites (Fig. 3). For both CypA and CypB, the interaction site is centered on the active site for their isomerase activity, which coincides with the CsA binding surface and even extends somewhat beyond this direct CsA binding surface (Fig. 3). In agreement with this, the interaction was completely abolished in the presence of CsA, as the spectra of Cyp/CsA with or without NS5A-D2 were strictly identical (data not shown). To quantify the interaction strength between both partners, we titrated increasing amounts of unlabeled NS5A-D2 into samples of ^{15}N-labeled CypA or CypB. Chemical shift changes of residues at the periphery of the binding site varied in a monotonous way from their free position toward the ligand saturated value, allowing the determination of K_D values of 64 and 90 μM for CypA and CypB, respectively (Fig. 4, *A*–*C*). However, residues in the active site of both cyclophilins broadened with increasing NS5A-D2 concentrations, as if multiple interactions

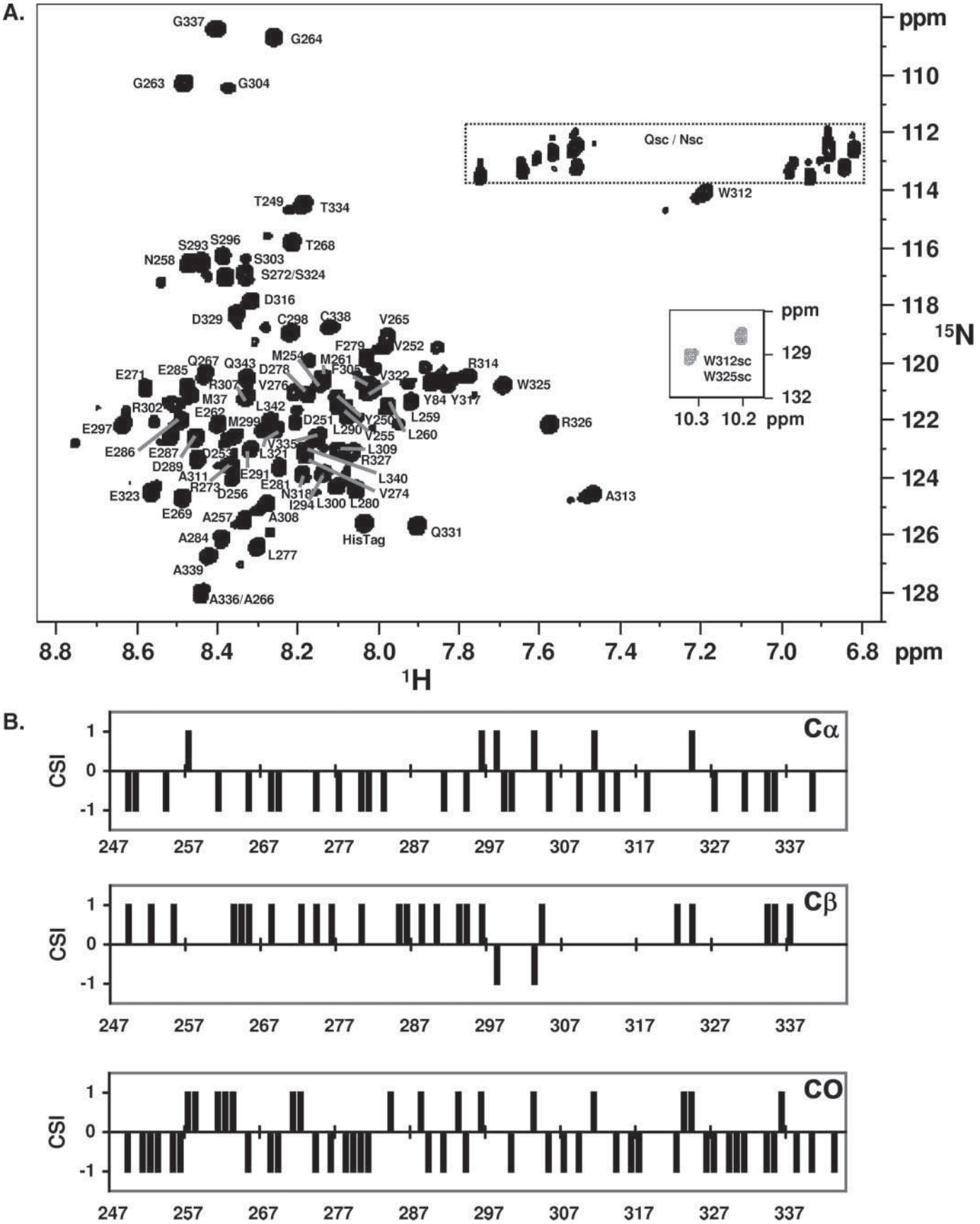


FIGURE 2. **NMR characterization of NS5A-D2 (JFH1).** *A*, assigned ^{1}H,^{15}N HSQC spectrum of domain 2 of NS5A (JFH1). The *smaller insert* shows the spectrum region corresponding to the two tryptophan side chains. *B*, CSI analysis of NS5A-D2. NS5A-D2 $^{13}C\alpha$, $^{13}C\beta$, and ^{13}CO chemical shifts were analyzed using CSI software (54). The consensus CSI is zero along the complete sequence and therefore is not shown.

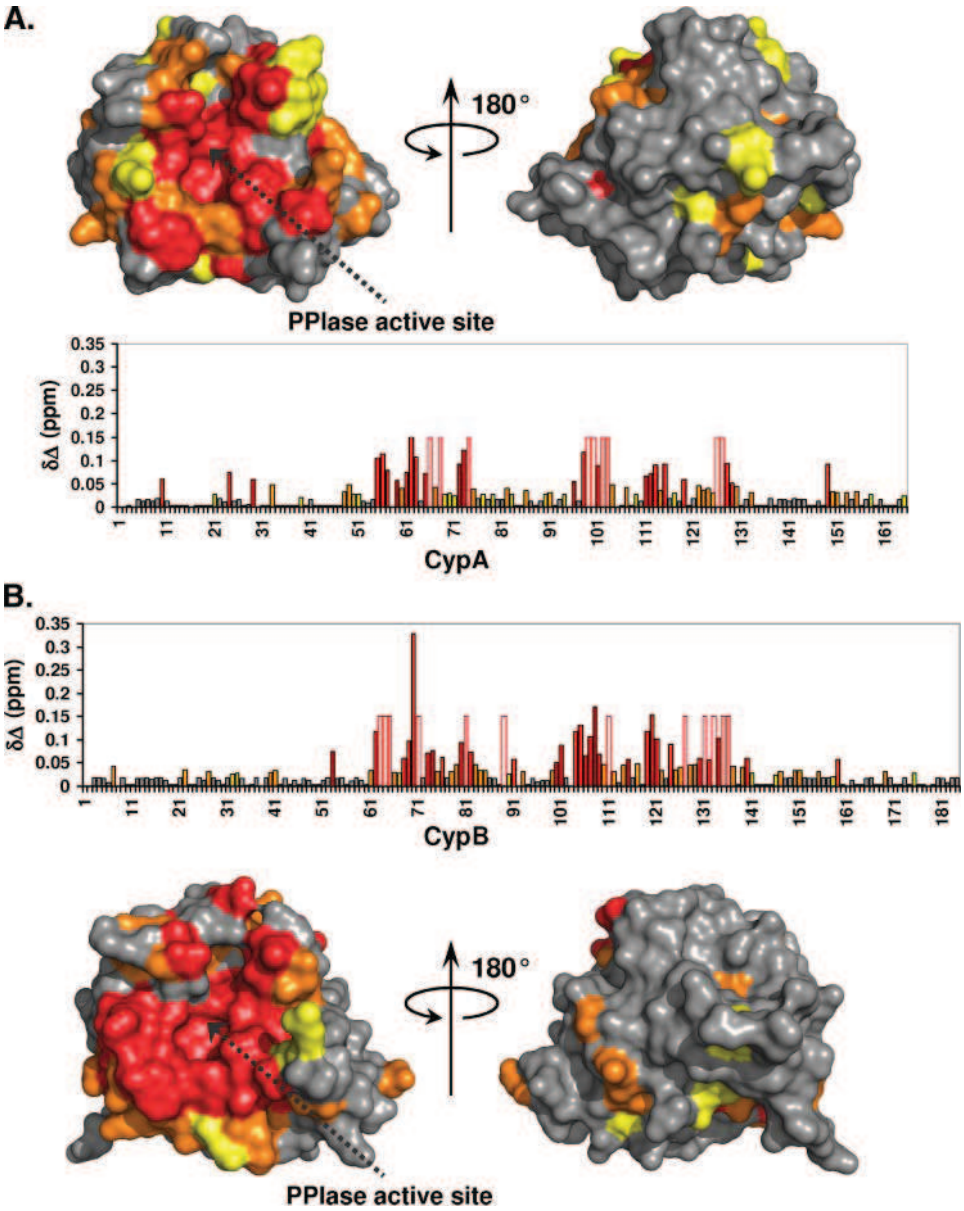


FIGURE 3. **Cyclophilin binding sites for NS5A-D2.** The ^{1}H and ^{15}N combined chemical shift perturbations ($\delta\Delta$) induced on the CypA (*A*) or CypB (*B*) spectra following NS5A-D2 addition in a 1:1 molar ratio were plotted along the cyclophilin primary sequences and on their respective three-dimensional molecular surfaces. Residues with combined chemical shift perturbations $0.02 \leq \delta\Delta \leq 0.03$ ppm are in *yellow*; $0.03 \leq \delta\Delta \leq 0.05$ ppm are in *orange*; and $\delta\Delta > 0.05$ ppm are in *red*. For cyclophilin residues for which the proton amide resonances disappear due to important line broadening in the presence of NS5A-D2, a fixed $\delta\Delta$ value of 0.15 ppm was set. These residues are depicted by an *open bar circled* in *red* in the diagrams. The PPIase active site of cyclophilins is indicated by a *dotted black arrow*.

were present simultaneously. To confirm this unexpected observation, we repeated the titration experiment with a synthetic peptide (PepD2, residues 304–323 of NS5A) corresponding to the best conserved region of NS5A-D2 that simultaneously contains the motif ^{310}PAWARP^{315} with the outlying ^{1}H,^{15}N chemical shift values (see above). With this peptide, the titration behavior, when monitored on exactly the same residues of the active site of CypA, did not show the broadening observed with the full NS5A-D2 domain. On the other hand, saturation was much slower to reach, and we derived a 10-fold weaker binding with a K_D value of 830 μM (Fig. 4, *D–F*). This all suggests that the D2 domain interacts in a distributed manner with the cyclophilin active site.

To confirm this by direct observation on the NS5A-D2 spectrum, we compared the HN(noCO) subspectra of ^{15}N-labeled NS5A-D2 alone with that of NS5A-D2 in the previous samples (supplemental Fig. 3). Concentrating first on the most intense peaks, which had all been mapped to their respective residue, next to a proline in the *trans* conformation, in the NS5A-D2 polypeptide we found a zone of significant spectral changes around the outlying peak of Trp^{312} (Fig. 5). Upon the addition of the cyclophilins, these peaks did not shift but rather broadened beyond detection. Line broadening occurs when the time scale of the exchange process is on the same order as that set by the frequency difference between the free and bound state. NMR line broadening of neighboring residues Arg^{302}, Ser^{303}, Ala^{311}, Ala^{313}, and Arg^{314} thereby was significantly more pronounced with CypA than with CypB (supplemental Figs. 3 and 5). Moreover, the amide proton resonances of residues Gly^{304}, Ala^{308}, Leu^{309} (see supplemental Fig. 3), Asp^{316}, Tyr^{317}, and Asn^{318} were unaffected in the presence of CypB, whereas they were no more detectable in the presence of CypA or broadened for Tyr^{317} (Figs. 5 and 6*C*). Among the numerous proline residues observed in NS5A-D2, only Pro^{310}, Pro^{315}, and Pro^{319}, which are in the direct vicinity of this interaction region, are fully conserved in any genotypes (Fig. 1*A*).

Other residues that had their peaks severely broadened upon the addition of the cyclophilins were Cys^{338} and Ala^{339}, but these C-terminal residues are just upstream of the C-terminal His tag, making an interpretation of this interaction in the isolated D2 domain more difficult. Importantly, however, the signals assigned to the minor *cis* forms for all prolines almost completely disappeared in the spectra of the complexes. This indicates that individual peptides containing *cis*-prolyl bonds interact via these *cis*-prolines with the cyclophilins. We thus next investigated the peptidyl-prolyl *cis*/*trans* isomerase activity of CypA and CypB.

Enzymatic Activities of Cyclophilins on Domain 2 of HCV NS5A Protein—We first characterized the cyclophilin catalyzed peptidyl-prolyl *cis*/*trans* isomerization by homonuclear EXSY (48, 55). In this experiment, one visualizes as off-diagonal peaks

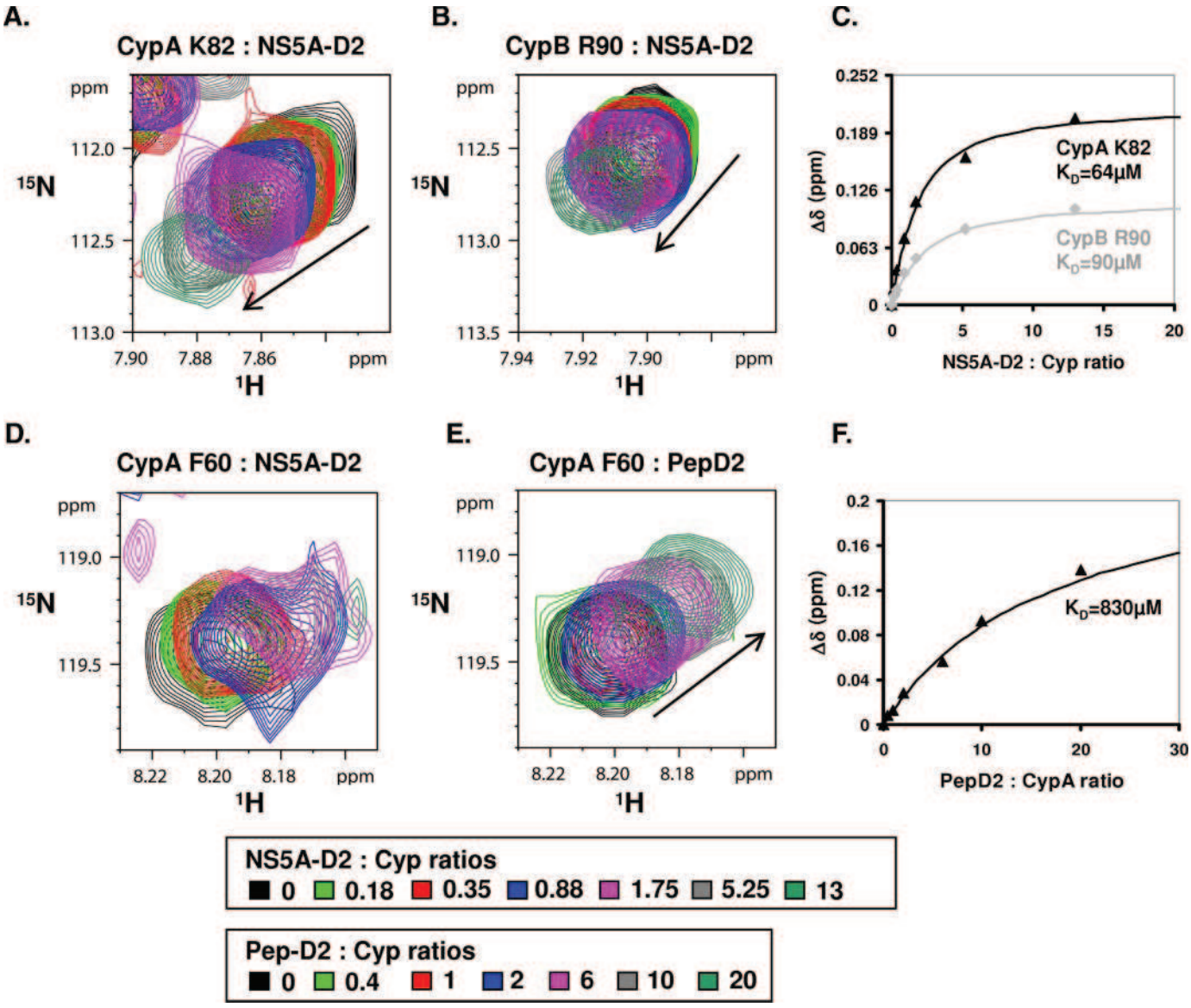


FIGURE 4. **Titration experiments between cyclophilins and NS5A-D2 or PepD2.** *Panels A*, *B*, *D*, and *E* correspond to the superimposition of the $^{1}H,^{15}N$ HSQC spectra of CypA (or CypB, in *B*) acquired in the presence of increasing amounts of unlabeled NS5A-D2 (*A*, *B*, and *D*) or Pep-D2 (*E*) (PepD2 corresponds to residues 304–323 of NS5A-D2: ^{304}GFPRALPAWARPDYNPPLVE^{323}). Lys^{82} in CypA is equivalent to Arg^{90} in CypB and is at the periphery of the NS5A-D2 binding site. *C*, titration curves corresponding to experiments in *A* (*black triangle*) and *B* (*gray diamonds*). The $^{1}H,^{15}N$ combined chemical shift perturbations $\Delta\delta$ (in ppm) ($\Delta\delta = (\delta(^{1}H)^{2} + 0.2.\delta(^{15}N)^{2})^{1/2}$) were plotted as a function of the NS5A-D2:cyclophilin molar ratios. The dissociation constants (K_D) were obtained by fitting the experimental data with the following equation: $K_D = [Cyp_{free}]\cdot[NS5A\text{-}D2_{free}]/[Cyp{:}NS5A\text{-}D2]$. *D* and *E*, Phe^{60} in CypA is directly in the binding site of NS5A-D2 and broadens when titrated with the D2 domain (*D*) but not with Pep-D2 (*E*). *F*, titration curve corresponding to experiments in *E* (*black triangles*).

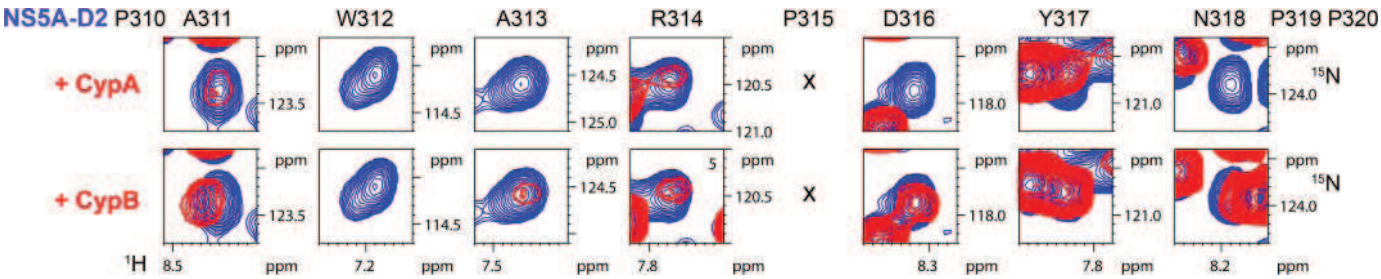


FIGURE 5. **A major interaction site of NS5A-D2 with CypA and CypB.** Each *panel* corresponds to the superposition of a $^{1}H,^{15}N$ HSQC spectrum acquired on NS5A-D2 alone (in *blue*) and of a $^{1}H,^{15}N$ plane obtained with the HNnoCO pulse sequence that specifically selects the NS5A-D2 subspectrum in a NS5A-D2/Cyp (1:1) sample (in *red*). The motif ^{310}PAWARPDYNPP^{320} of NS5A-D2 interacts with CypA.

those amide functions that have physically changed from the *cis* to *trans* (or vice versa) during the mixing delay (typically of the order of 100 ms). Without cyclophilins, the exchange rate of peptidyl-prolyl bonds, even in unstructured peptides, is too slow ($k_{exch} < 0.1\ s^{-1}$) to lead to detectable exchange peaks in EXSY spectra. When adding the cyclophilin in catalytic

amounts, we did detect several novel exchange peaks. However, the natively unfolded nature of NS5A-D2 and ensuing limited proton dispersion renders the assignment of these peaks extremely difficult on the sole basis of their proton chemical shift. Moreover, the proton chemical shift differences between *trans* and *cis* forms being often very limited, potential exchange peaks nearly coincide with the diagonal. The homonuclear EXSY experiments thus led us to conclude that both cyclophilins do isomerize distinct peptidyl-prolyl bonds within NS5A-D2 but without allowing the assignment of the peptidyl-prolyl bonds or evaluation of the catalytic efficacy.

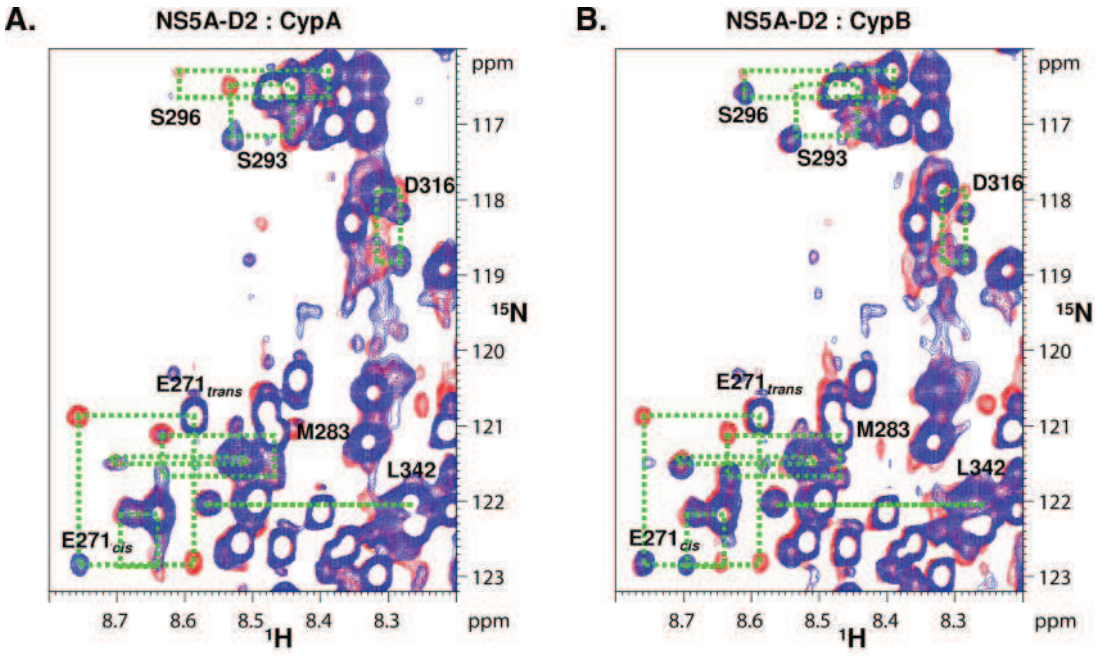
To increase resolution and allow assignment of the individual processes, we performed a series of ^{15}N z-exchange experiments (14, 15, 49, 56) on a ^{15}N-labeled NS5A-D2 sample in the presence of catalytic amounts of CypA or CypB (1:10). The exchange between two conformations is now monitored at the level of the amide function (characterized by a $^{1}H,^{15}N$ correlation peak in the HSQC spectrum) rather than for the sole amide proton frequency as in the EXSY experiment. Heteronuclear exchange spectra were acquired at different mixing times: 0.88, 25, 50, 100, 200, 300 and 400 ms. At the shortest mixing time (0.88 ms), no exchange peaks connecting the major and minor peak of a given residue were visible, but the minor peaks did already broaden (Fig. 6 and supplemental Fig. 4), in agreement with our previous results on the 1:1 complexes. Broadening was more severe for the complex with CypA than for the one with CypB, pointing toward an equally stronger binding of CypA to those alternative anchoring points formed by the *cis* prolines in the NS5A-D2 sequence. Upon increasing the exchange interval (mixing time), additional connecting peaks could be observed and assigned for several residues (Fig. 6 and supplemental Fig. 4), thereby confirming their assignment and excluding the possibility that these minor peaks come from a degradation product or other molecular entity. Comparing the exchange spectra obtained with CypA and CypB, we found roughly the same set of additional peaks, suggesting that both Cyp have a similar activity toward the peptidyl-prolyl bonds in NS5A-D2. Both Cyp, moreover, lack a clear specificity, as PPIase-catalyzed exchange peaks could be assigned for residues in the vicinity of almost all of the 15 proline residues in the NS5A-D2 sequence. For Pro^{306}, Pro^{319}, and Pro^{320} only, we did not detect exchange peaks for residues in their direct neighborhood, but spectral overlap clearly limited our analysis of the process around these prolines. What does distinguish both cyclophilins, however, is the catalytic efficacy of the isomerization. Even with careful normalization of the enzyme content in NMR samples, the CypA-catalyzed exchange peaks were generally more intense than those obtained with CypB. Extracting a rate constant (k_{exch}) from the buildup of the exchange peaks as a function of increasing exchange time (Fig. 6*D*), we found that the CypA-catalyzed exchange rates range from 14 s^{-1} for Gln^{331} to 61 s^{-1} for Met^{283}, with a mean value of 29 s^{-1} as calculated over the 10 residues for which a reliable rate constant could be extracted (Fig. 6*E*). CypB as an enzyme is less effective, with exchange rates ranging from 3 s^{-1} for Leu^{277} to 31 s^{-1} for Gln^{331}, and an average of 11 s^{-1} determined over the 14 NS5A-D2 residues for which the experimental data led to reliable curves. As was the case for the homonuclear EXSY spectra, the additional peaks connecting *cis* and *trans* conformers of a given residue disappeared upon the addition of CsA (supplemental Fig. 5).

DISCUSSION

NS5A is required in several steps of the HCV life cycle, including replication and infectious particle assembly (3, 27, 57), but its precise roles are still not known. Recent mutational analyses have shown that many residues of its D2 domain are essential for RNA replication (29), and several mutations in this domain were reported to confer resistance to CsA (26). However, the structural data required for further understanding of these observations are still limited.

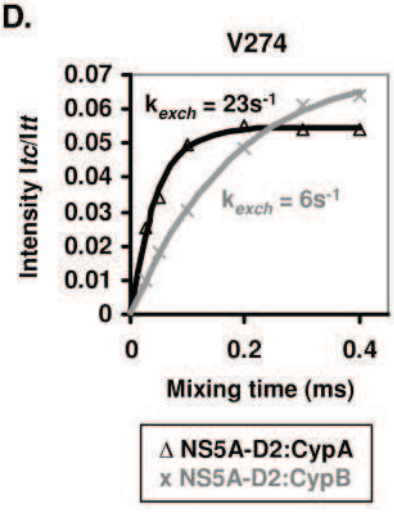
We have chosen here to study the D2 domain in the context of the HCV genotype 2a (JFH1). This clone, isolated from a Japanese patient with a fulminant hepatitis, allows for infectious virus propagation in cell culture (58–60). All biochemical and biophysical characterization methods indicate that the isolated domain 2 of NS5A (JFH-1), when overproduced recombinantly and purified to homogeneity, is unstructured (Figs. 1 and 2). A similar increase in apparent molecular weight, random coil CD spectrum, and limited dispersion for the amide proton chemical shifts in the NMR spectrum were described previously for the genotype 1a NS5A-D2 domain, although the two domains share only 48% sequence identity (36, 37) (See supplemental Fig. 6). The NS5A-D2 domain thus belongs to the growing group of natively unstructured proteins that gain function upon interaction with their molecular partners (61, 62). Furthermore, when we used the carbon chemical shifts to detect potential structure at the local level (CSI strategy) (54), we could not detect even small stretches of stable secondary structure that might have gone undetected by the macroscopic approaches described above. However, the amide resonances of the Trp^{312} and Ala^{313} in the most conserved ^{310}PAWARP^{315} motif resonate at an unusual proton and nitrogen frequency (Fig. 2*A*). As these anomalous chemical shifts were present equally in the genotype 1a NS5A-D2 domain (37), and as this Trp^{312}-Ala^{313} segment, as well as Pro^{310} and Pro^{315}, is fully conserved in all genotypes (Fig. 1*A*), we have synthesized a peptide centered on this motif and are currently pursuing a detailed NMR analysis to interpret the anomalous chemical shift values in structural terms.

Certain mutations in the D2 domain of NS5A confer resistance to CsA (26), a cyclic undecapeptide for which the primary target in the eukaryotic cell is members of the cyclophilin family (63). Cyclophilins are peptidy-prolyl *cis*/*trans* isomerases that are involved in the life cycle of several viruses. The best characterized example is CypA interacting with the capsid domain of the HIV Gag polyprotein precursor (12). For HCV, the implication of cyclophilins in the viral life cycle comes from the observation that the cyclophilin-specific inhibitor CsA has anti-HCV properties (18, 20, 23). In 2005, Watashi *et al.* (17) reported that CypB binds to the viral RdRp NS5B protein (genotype 1b) and regulates its RNA binding properties. However, Ishii *et al.* (22) reported that CypB does not regulate the RNA binding activity of NS5B in a JFH1 context (genotype 2a) and Robida *et al.* (24) have shown that there is no replication defect in a genotype 1b replicon system when CypB expression is abolished. Whereas these reports functionally link the cyclophilins



C.

```
          256        266        276        286        296        306
MNTYDVDMVD ANLLMEGGVA QTEPESRVPV LDFLEPMAEE ESDLEPSIPS ECMLPRSGFP
                                                                  *
          316        326        336        346
RALPAWARPD YNPPLVESWR RPDYQPPTVA GCALPLQHHH HHH
 **        *  **                              6xHis-tag
```



E.

NS5A-D2	k_{exch} (s^{-1})	
	CypA	CypB
E271	24	8
V274	23	6
L277	19	3
E281	59	6
M283	61	11
S293	17	10
I294	*ND**	6
S296	*ND**	6
L300	*ND**	18
W312	*ND**	17
D316	24	5
W325	17	17
Q331	14	31
L340	31	12

**ND, not determined (Exchange peaks were detected but no reliable k_{exch} values have been determined)*
Exchange peaks detected but not quantified for T268, F279, L280, E291, E297, A311, A339, and L342

to NS5B, a recent study indicates that the sensitivity of HCV for CsA depends not only on NS5B but equally (and even more) on NS5A (26). Finally, whereas the earlier reports mainly point to CypB as the modulator of the NS5A/B activity, recent results have questioned this, and the dependence of HCV replication on cyclophilin subtypes equally may vary with the genotype. Very recently, Yang *et al.* (25) have shown that CypA is an essential co-factor for numerous HCV genotypes, including genotypes 1a, 1b, and 2a (isolate JFH1).

In view of these conflicting reports, we used NMR spectroscopy to probe the interaction of NS5A-D2 (JFH1) with both CypB and CypA. Chemical shift perturbation experiments on both NS5A-D2·CypA and NS5A-D2·CypB complexes gave evidence for a direct physical interaction that is localized to the active site on the cyclophilins (Fig. 3). Important chemical shift perturbations have been measured for CypA Arg^{55}, Phe^{60}, Met^{61}, Asn^{102}, Phe^{113}, and His^{126} residues and the equivalent residues on CypB, all previously shown to interact directly with a peptide substrate (64, 65). Titration experiments with the NS5A-D2 domain against both cyclophilins allowed us to quantify the interaction with K_D values of 64 and 90 μM toward CypA and CypB, respectively. As the active site of the cyclophilins coincides with their CsA binding groove, we indeed found that CsA competes very efficiently with NS5A-D2 for binding to the cyclophilins. Its nanomolar affinity toward cyclophilins (66) causes a complete inhibition of the molecular interaction between NS5A-D2 and the cyclophilins.

Although the obtained K_D values are comparable to the 15 μM dissociation constant that has been measured for CypA toward HIV Capsid (14, 67), one fundamental difference became clear from the observed broadening of the Cyp active site resonances upon increasing NS5A-D2 concentration. The CypA/HIV capsid interaction indeed has been localized to a single Gly^{221}-Pro^{222} motif in the HIV capsid protein, whereas in the present case, it seems that many prolines can interact with the cyclophilins. When repeating the titration experiment with a peptide (PepD2, residues 304–323 of NS5A) containing only five out of the 15 proline residues in full-length NS5A-D2, the titration behavior proved more conventional but led to a 10-fold lower interaction strength. Other anchoring points thus contribute to the interaction with the intact D2 domain, and the line broadening observed for the *cis* proline-associated resonances, even upon addition of catalytic amounts of cyclophilins, suggests that the overall interaction strength comes from several anchorage points distributed over the NS5A-D2 sequence. The presence of multiple mutations in NS5A-D2 that confer CsA resistance to the HCV virus is in agreement with the absence of a single interaction hotspot on the D2 domain, but equally it suggests that a functional interaction requires a narrow window of Cyp concentration in the complex.

Both CypB and CypA bind a highly conserved motif in domain 2 of NS5A centered on the ^{310}PAWARP^{315} sequence in the JFH1 HCV clone. Whereas CypB solely interacts with this hexapeptide, the motif recognized by CypA is, however, larger and corresponds to ^{304}GFPRALPAWARPDYNPP^{320} (Figs. 5 and 6 and supplemental Fig. 3). Indeed NMR resonances of Gly^{304}, Ala^{308}, Leu^{309}, Asp^{316}, Tyr^{317}, and Asn^{318} are affected only following CypA addition, whereas Ala^{311}, Trp^{312}, Ala^{313}, and Arg^{314} resonances are perturbed in the presence of either cyclophilins. Although highly specific, this peptide does not contribute more than one-tenth of the interaction strength. The natural abundance $^{1}H,^{15}N$ HSQC spectrum acquired on this peptide mapped very well to the corresponding residues in the full-length NS5A-D2 sequence (data not shown), excluding a difference in structure and dynamics as the source of this discrepancy. Importantly, at least six residues in this NS5A-D2 motif recognized by CypA were shown previously to be essential for HCV replication in a subgenomic 1b replicon system (Con1 isolate) (29). In this genotype 1b isolate the motif is rather well conserved with only three amino acids substitution compared with genotype 2a (JFH1) (^{308}KFPRAMPIWARPDYNPP^{324}) (supplemental Fig. 6). The mutant M313A replicates with very low efficiency close to the detection limit, the mutant P314A was lethal, W316A was moderately impaired, mutant A317G yielded a small-colony phenotype, mutant Y321A was severely impaired in replication and also gave a small colony phenotype, and the P324A mutant was also lethal (residues highlighted in *black* in supplemental Fig. 6). These results from Tellinghuisen *et al.* (29), combined with ours showing a direct interaction of CypA with the corresponding region in NS5A-D2, and finally the finding that CypA is an essential co-factor for HCV replication (genotypes 1a, 1b, and 2a) (25) suggest that the replication defects of NS5A mutants might result from an altered interaction between the cyclophilin and NS5A-D2 in this zone.

Several groups have applied CsA treatment to virus-infected cell cultures to select for mutations that would confer resistance directly. A HCV replicon (genotype 1b) mutant bearing a mutation corresponding to Y317N in genotype 2a exhibited enhanced CsA resistance (26). This Tyr^{317} is just downstream of the identified motif and belongs to the binding site of CypA but not of CypB (Fig. 5). However, in the same study, six additional mutations were discovered in NS5A, of which four are located in domain 2, one in the low complexity sequence between D2 and D3, and another one in D3 (26). Mutations that have been identified in domain 2 (1b) (highlighted in *gray* in supplemental Fig. 6) correspond to the following residues in

FIGURE 6. ***Cis/trans* isomerization of HCV NS5A-D2 *X*-Pro peptide bonds catalyzed by CypA and CypB.** $^{1}H,^{15}N$ heteronuclear exchange spectra recorded at 800 MHz with mixing times of 0.88 ms (in *blue*) and 100 ms (in *red*) on [^{15}N]NS5A-D2 samples (220 μM) with catalytic amounts of either CypA (*A*) or CypB (*B*) (23 μM). The NMR resonances (*trans*, *cis*, and the two exchange peaks) of NS5A-D2 residues for which the PPIase activity of a cyclophilin can be evidenced are connected by *green dotted boxes*. *C*, amino acid sequence of NS5A-D2 (JFH1) construction. Residues on which PPIase activities of CypA and CypB have been monitored are bold and shown in *violet*. The 15 proline residues of NS5A-D2 are indicated on a *light gray* background. The previously defined binding site of cyclophilins is *boxed*, and residues that are affected only by CypA binding (see Fig. 4 and under "Results") are marked with *asterisks*. *D*, determination of the CypA (in *black* (Δ))- and CypB-catalyzed (in *gray* (×)) exchange rates toward the Val^{274}-Pro^{275} NS5A-D2 peptide bond. Experimental data measured on Val^{274} for (I_{tc}/I_{tt}) (CypA (Δ); CypB (×)) were fitted to the theoretical Equation 2 (see "Experimental Procedures") (CypA, *black line*; CypB, *gray line*). *E*, resulting exchange rates (k_{exch}) of the *cis/trans* isomerization processes in NS5A-D2 catalyzed by the addition of catalytic amounts of either CypA or CypB.

NS5A-D2 of genotype 2a (JFH1): Asp^{256}, Val^{276}, Leu^{280}, and Met^{299}. Therefore, they do not map directly to the above described conserved motif. These residues do however contribute to the interaction with the cyclophilins, as they are centered on the NS5A-D2 region for which highest efficiencies have been measured for the CypA-catalyzed *cis*/*trans* isomerization reactions (Fig. 6). Only Pro^{282} appears to be conserved in genotypes 1a and 1b (supplemental Fig. 6), arguing against the precise localization as the important factor for cyclophilin function in RNA replication. The presence of a well defined amount of cyclophilin at the NS5A-D2 surface seems to be required in order to confer functionality.

Our interaction experiments with a 1:1 molecular ratio between domain 2 of NS5A and cyclophilins showed a pronounced broadening of all resonances corresponding to residues in the vicinity of a *cis*-proline residue, leading us to investigate the peptidyl-prolyl *cis*/*trans* isomerase activity of the cyclophilins toward prolyl bonds in the NS5A-D2 domain. NMR exchange spectroscopy, previously used to characterize the CypA-catalyzed *cis*/*trans* isomerization of the Gly^{221}-Pro^{222} peptide bond in the HIV capsid (14, 15), indeed provided direct evidence for the catalytic activity of the cyclophilins and allowed us to assign the effect to individual prolyl bonds. Because of the unstructured nature of NS5A-D2 and the resulting low proton amide dispersion, ^{1}H,^{15}N heteronuclear z-exchange spectroscopy (49) proved to be superior to ^{1}H,^{1}H homonuclear EXSY spectroscopy and allowed us for the first time to prove *in vitro* that HCV NS5A-D2 is a substrate for the PPIase activity of at least two host cyclophilins (Fig. 6 and supplemental Fig. 4). Despite the fact that both CypA and CypB catalyze the *cis*/*trans* isomerization of the same NS5A-D2 *X*-Pro peptide bonds, they do not act with the same efficiency. Domain 2 of NS5A is a better substrate for CypA than for CypB, with a mean exchange rate (k_{exch}) of 28.9 s^{-1} for CypA and only 11.1 s^{-1} for CypB. Every enzyme equally has its preferred sites, which do not necessarily coincide. The highest enzymatic efficiencies have been measured in the Glu^{271}-Met^{283} region of NS5A-D2 for CypA, with maximal k_{exch} values of 59 and 61 s^{-1} for Glu^{281} and Met^{283}, respectively, which probably reflects the isomerization of the Glu^{281}-Pro^{282} peptidyl-prolyl bond. The maximal activity of CypA in the N-terminal region of NS5A-D2 coincides with the localization of the majority of resistance conferring mutations. Together with the stronger affinity, this supports the dominant role for CypA in the infection process. This conclusion has been confirmed by HCV infection and replication assays using cell lines with stable knockdown of CypA and CypB.[6] CypB displays more activity toward the C-terminal half of the NS5A-D2 domain, with an optimal activity toward the Gln^{331}-Pro^{332} peptidyl-prolyl bond (k_{exch} = 31 s^{-1}) (Fig. 6). For comparison, Bosco *et al.* (14, 15) have found that with an enzyme:substrate ratio comparable to that used here, the CypA-catalyzed *cis*/*trans* isomerization of the Gly^{221}-Pro^{222} HIV Capsid bond is characterized by a k_{exch} value around 10 s^{-1}. However, in their system, CypA specifically binds to and catalyzes *cis*/*trans* isomerization of Gly^{221}-Pro^{222} over other Gly-Pro motifs in the HIV capsid. We show here that CypA is enzymatically active on almost all *X*-Pro NS5A-D2 sites, albeit with different efficiencies. The absence of specificity of CypA toward NS5A-D2 sites is possibly related to the unstructured character of the protein, as cyclophilins lack specificity toward peptide substrates (68).

The present structure-function study provides the first molecular basis for the further understanding of the resistance of HCV replication to CsA and analogues. As CsA abolishes the interaction between NS5A-D2 and CypA but also the PPIase activity of CypA toward this domain, we cannot conclude whether it is the binding, the catalytic activity, or even both of these that are involved in the HCV replication process (69). Indeed, cyclophilins may play biological roles either by catalyzing the *cis*/*trans* isomerization of a peptide bond, as for the tyrosine kinase Itk (70), or by interacting with an *X*-Pro motif that is no longer available for interaction with others partners, as is the case with HIV capsid with TRIM5α and CypA (12, 13). Further studies with NS5A-D2 and the cyclophilins in the presence of an interacting partner such as NS5B and/or RNA will be necessary in order to evaluate their precise role in the HCV life cycle.

Acknowledgments—We gratefully acknowledge RD-Biotech (Besançon, France) for the cloning and initial expression and purification tests for NS5A-D2, Guillaume Blanc and Jennifer Molle for technical assistance, Michel Becchi for the mass spectroscopy measurements, Christophe Combet for bioinformatics support. CD experiments were performed on the platform "Production et Analyse de Protéines" of the IFR 128 BioSciences Gerland-Lyon Sud. The NMR facility used in this study was funded by the Région Nord-Pas de Calais (France), the CNRS, the Universities of Lille 1 and Lille 2, and the Institut Pasteur de Lille.

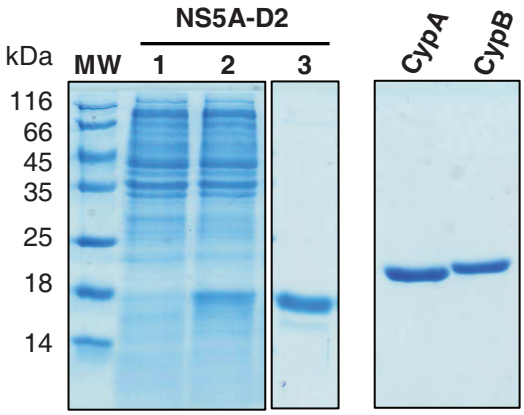
REFERENCES

1. National Institutes of Health (2002) *Hepatology* **36,** Suppl. 1, S2–S20
2. Appel, N., Schaller, T., Penin, F., and Bartenschlager, R. (2006) *J. Biol. Chem.* **281,** 9833–9836
3. Moradpour, D., Penin, F., and Rice, C. M. (2007) *Nat. Rev.* **5,** 453–463
4. Tellinghuisen, T. L., Evans, M. J., von Hahn, T., You, S., and Rice, C. M. (2007) *J. Virol.* **81,** 8853–8867
5. Lohmann, V., Korner, F., Koch, J., Herian, U., Theilmann, L., and Bartenschlager, R. (1999) *Science* **285,** 110–113
6. Handschumacher, R. E., Harding, M. W., Rice, J., Drugge, R. J., and Speicher, D. W. (1984) *Science* **226,** 544–547
7. Schreiber, S. L. (1991) *Science* **251,** 283–287
8. Barik, S. (2006) *Cell. Mol. Life Sci.* **63,** 2889–2900
9. Bergsma, D. J., Eder, C., Gross, M., Kersten, H., Sylvester, D., Appelbaum, E., Cusimano, D., Livi, G. P., McLaughlin, M. M., and Kasyan, K. (1991) *J. Biol. Chem.* **266,** 23204–23214
10. Watashi, K., and Shimotohno, K. (2007) *Drug Target Insights* **1,** 9–18
11. Luban, J., Bossolt, K. L., Franke, E. K., Kalpana, G. V., and Goff, S. P. (1993) *Cell* **73,** 1067–1078
12. Luban, J. (2007) *J. Virol.* **81,** 1054–1061
13. Sokolskaja, E., Berthoux, L., and Luban, J. (2006) *J. Virol.* **80,** 2855–2862
14. Bosco, D. A., Eisenmesser, E. Z., Pochapsky, S., Sundquist, W. I., and Kern, D. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5247–5252
15. Bosco, D. A., and Kern, D. (2004) *Biochemistry* **43,** 6110–6119
16. Bukovsky, A. A., Weimann, A., Accola, M. A., and Gottlinger, H. G. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 10943–10948
17. Watashi, K., Ishii, N., Hijikata, M., Inoue, D., Murata, T., Miyanari, Y., and Shimotohno, K. (2005) *Mol. Cell* **19,** 111–122

[6] A. Kaul and R. Bartenschlager, unpublished results.

18. Nakagawa, M., Sakamoto, N., Enomoto, N., Tanabe, Y., Kanazawa, N., Koyama, T., Kurosaki, M., Maekawa, S., Yamashiro, T., Chen, C. H., Itsui, Y., Kakinuma, S., and Watanabe, M. (2004) *Biochem. Biophys. Res. Commun.* **313,** 42–47
19. Tanabe, Y., Sakamoto, N., Enomoto, N., Kurosaki, M., Ueda, E., Maekawa, S., Yamashiro, T., Nakagawa, M., Chen, C. H., Kanazawa, N., Kakinuma, S., and Watanabe, M. (2004) *J. Infect. Dis.* **189,** 1129–1139
20. Watashi, K., Hijikata, M., Hosaka, M., Yamaji, M., and Shimotohno, K. (2003) *Hepatology* **38,** 1282–1288
21. Flisiak, R., Horban, A., Gallay, P., Bobardt, M., Selvarajah, S., Wiercinska-Drapalo, A., Siwak, E., Cielniak, I., Higersberger, J., Kierkus, J., Aeschlimann, C., Grosgurin, P., Nicolas-Metral, V., Dumont, J. M., Porchet, H., Crabbe, R., and Scalfaro, P. (2008) *Hepatology* **47,** 817–826
22. Ishii, N., Watashi, K., Hishiki, T., Goto, K., Inoue, D., Hijikata, M., Wakita, T., Kato, N., and Shimotohno, K. (2006) *J. Virol.* **80,** 4510–4520
23. Nakagawa, M., Sakamoto, N., Tanabe, Y., Koyama, T., Itsui, Y., Takeda, Y., Chen, C. H., Kakinuma, S., Oooka, S., Maekawa, S., Enomoto, N., and Watanabe, M. (2005) *Gastroenterology* **129,** 1031–1041
24. Robida, J. M., Nelson, H. B., Liu, Z., and Tang, H. (2007) *J. Virol.* **81,** 5829–5840
25. Yang, F., Robotham, J. M., Nelson, H. B., Irsigler, A., Kenworthy, R., and Tang, H. (2008) *J. Virol.* **82,** 5269–5278
26. Fernandes, F., Poole, D. S., Hoover, S., Middleton, R., Andrei, A. C., Gerstner, J., and Striker, R. (2007) *Hepatology* **46,** 1026–1033
27. Appel, N., Zayas, M., Miller, S., Krijnse-Locker, J., Schaller, T., Friebe, P., Kallis, S., Engel, U., and Bartenschlager, R. (2008) *PLoS Pathog.* **4,** e1000035
28. Penin, F., Dubuisson, J., Rey, F. A., Moradpour, D., and Pawlotsky, J. M. (2004) *Hepatology* **39,** 5–19
29. Tellinghuisen, T. L., Foss, K. L., Treadaway, J. C., and Rice, C. M. (2008) *J. Virol.* **82,** 1073–1083
30. Brass, V., Bieck, E., Montserret, R., Wolk, B., Hellings, J. A., Blum, H. E., Penin, F., and Moradpour, D. (2002) *J. Biol. Chem.* **277,** 8130–8139
31. Penin, F., Brass, V., Appel, N., Ramboarina, S., Montserret, R., Ficheux, D., Blum, H. E., Bartenschlager, R., and Moradpour, D. (2004) *J. Biol. Chem.* **279,** 40835–40843
32. Tellinghuisen, T. L., Marcotrigiano, J., Gorbalenya, A. E., and Rice, C. M. (2004) *J. Biol. Chem.* **279,** 48576–48587
33. Huang, L., Hwang, J., Sharma, S. D., Hargittai, M. R., Chen, Y., Arnold, J. J., Raney, K. D., and Cameron, C. E. (2005) *J. Biol. Chem.* **280,** 36417–36428
34. Tellinghuisen, T. L., Marcotrigiano, J., and Rice, C. M. (2005) *Nature* **435,** 374–379
35. Tellinghuisen, T. L., Foss, K. L., and Treadaway, J. (2008) *PLoS Pathog.* **4,** e1000032
36. Liang, Y., Kang, C. B., and Yoon, H. S. (2006) *Mol. Cells* **22,** 13–20
37. Liang, Y., Ye, H., Kang, C. B., and Yoon, H. S. (2007) *Biochemistry* **46,** 11550–11558
38. Combet, C., Blanchet, C., Geourjon, C., and Deleage, G. (2000) *Trends Biochem. Sci.* **25,** 147–150
39. Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., Dehne-Garcia, A., Geourjon, C., Bettler, E., Hulo, C., Le Mercier, P., Bartenschlager, R., Diepolder, H., Moradpour, D., Pawlotsky, J. M., Rice, C. M., Trepo, C., Penin, F., and Deleage, G. (2007) *Nucleic Acids Res.* **35,** Database Suppl., D363–D366
40. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680
41. Cortay, J. C., Negre, D., Scarabel, M., Ramseier, T. M., Vartak, N. B., Reizer, J., Saier, M. H., Jr., and Cozzone, A. J. (1994) *J. Biol. Chem.* **269,** 14885–14891
42. Chen, Y. H., Yang, J. T., and Chau, K. H. (1974) *Biochemistry* **13,** 3350–3359
43. Hanoulle, X., Melchior, A., Sibille, N., Parent, B., Denys, A., Wieruszeski, J. M., Horvath, D., Allain, F., Lippens, G., and Landrieu, I. (2007) *J. Biol. Chem.* **282,** 34148–34158
44. Verdegem, D., Dijkstra, K., Hanoulle, X., and Lippens, G. (2008) *J. Biomol. NMR* **42,** 11–21
45. Grzesiek, S., Bax, A., Hu, J. S., Kaufman, J., Palmer, I., Stahl, S. J., Tjandra, N., and Wingfield, P. T. (1997) *Protein Sci.* **6,** 1248–1263
46. Ottiger, M., Zerbe, O., Guntert, P., and Wuthrich, K. (1997) *J. Mol. Biol.* **272,** 64–81
47. Golovanov, A. P., Blankley, R. T., Avis, J. M., and Bermel, W. (2007) *J. Am. Chem. Soc.* **129,** 6528–6535
48. Kaplan, J. L., and Fraenkel, G. (1980) *NMR of Chemically Exchanging Systems*, Academic Press, New York
49. Farrow, N. A., Zhang, O., Forman-Kay, J. D., and Kay, L. E. (1994) *J. Biomol. NMR* **4,** 727–734
50. Huang, L., Sineva, E. V., Hargittai, M. R., Sharma, S. D., Suthar, M., Raney, K. D., and Cameron, C. E. (2004) *Protein Expression Purif.* **37,** 144–153
51. Kieliszewski, M. J., Leykam, J. F., and Lamport, D. T. (1990) *Plant Physiol.* **92,** 316–326
52. Tompa, P. (2002) *Trends Biochem. Sci.* **27,** 527–533
53. Buck, M. (1998) *Q. Rev. Biophys.* **31,** 297–355
54. Wishart, D. S., and Sykes, B. D. (1994) *J. Biomol. NMR* **4,** 171–180
55. Kern, D., Drakenberg, T., Wikstrom, M., Forsen, S., Bang, H., and Fischer, G. (1993) *FEBS Lett.* **323,** 198–202
56. Kern, D., Eisenmesser, E. Z., and Wolf-Watz, M. (2005) *Methods Enzymol.* **394,** 507–524
57. Macdonald, A., and Harris, M. (2004) *J. Gen. Virol.* **85,** 2485–2502
58. Pietschmann, T., Kaul, A., Koutsoudakis, G., Shavinskaya, A., Kallis, S., Steinmann, E., Abid, K., Negro, F., Dreux, M., Cosset, F. L., and Bartenschlager, R. (2006) *Proc. Natl. Acad. Sci. U. S. A.* **103,** 7408–7413
59. Wakita, T., Pietschmann, T., Kato, T., Date, T., Miyamoto, M., Zhao, Z., Murthy, K., Habermann, A., Krausslich, H. G., Mizokami, M., Bartenschlager, R., and Liang, T. J. (2005) *Nat. Med.* **11,** 791–796
60. Zhong, J., Gastaminza, P., Cheng, G., Kapadia, S., Kato, T., Burton, D. R., Wieland, S. F., Uprichard, S. L., Wakita, T., and Chisari, F. V. (2005) *Proc. Natl. Acad. Sci. U. S. A.* **102,** 9294–9299
61. Dyson, H. J., and Wright, P. E. (2005) *Nat. Rev. Mol. Cell Biol.* **6,** 197–208
62. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D., and Nussinov, R. (2003) *Trends Biochem. Sci.* **28,** 81–85
63. Liu, J., Farmer, J. D., Jr., Lane, W. S., Friedman, J., Weissman, I., and Schreiber, S. L. (1991) *Cell* **66,** 807–815
64. Zhao, Y., and Ke, H. (1996) *Biochemistry* **35,** 7362–7368
65. Zhao, Y., and Ke, H. (1996) *Biochemistry* **35,** 7356–7361
66. Mikol, V., Kallen, J., and Walkinshaw, M. D. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5183–5186
67. Yoo, S., Myszka, D. G., Yeh, C., McMurray, M., Hill, C. P., and Sundquist, W. I. (1997) *J. Mol. Biol.* **269,** 780–795
68. Harrison, R. K., and Stein, R. L. (1990) *Biochemistry* **29,** 3813–3816
69. Fischer, G., Tradler, T., and Zarnt, T. (1998) *FEBS Lett.* **426,** 17–20
70. Brazin, K. N., Mallis, R. J., Fulton, D. B., and Andreotti, A. H. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 1899–1904
71. Simmonds, P., Bukh, J., Combet, C., Deleage, G., Enomoto, N., Feinstone, S., Halfon, P., Inchauspe, G., Kuiken, C., Maertens, G., Mizokami, M., Murphy, D. G., Okamoto, H., Pawlotsky, J. M., Penin, F., Sablon, E., Shin, I. T., Stuyver, L. J., Thiel, H. J., Viazov, S., Weiner, A. J., and Widell, A. (2005) *Hepatology* **42,** 962–973

Supplemental Figure 1



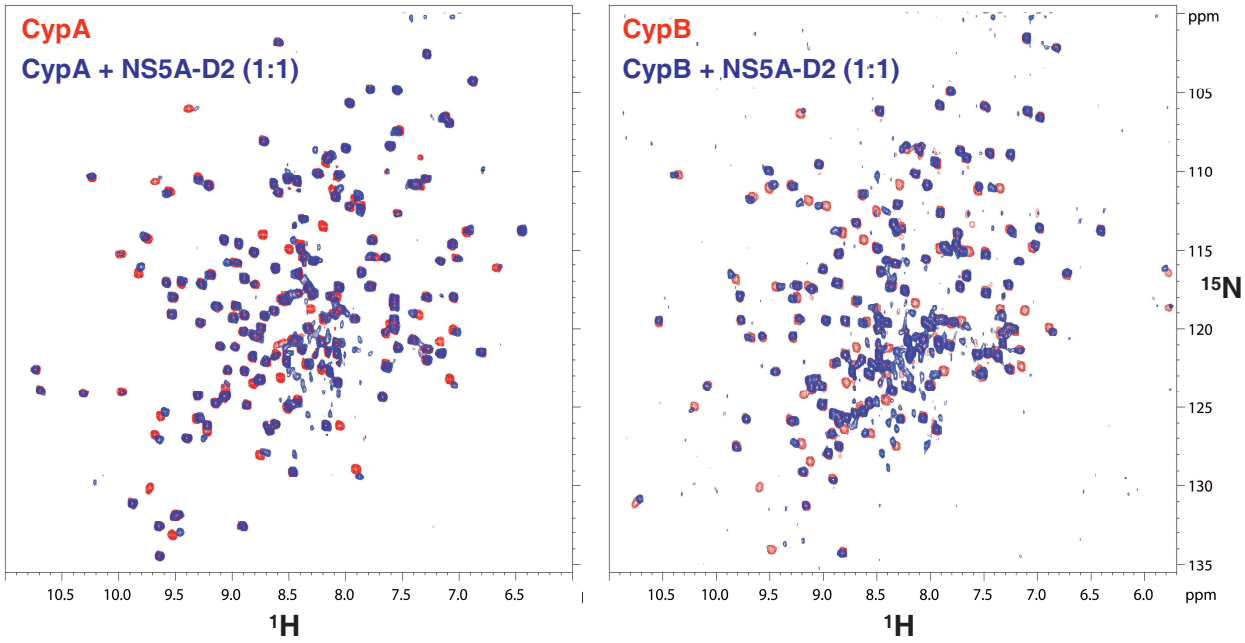
Expression and purification of domain 2 of NS5A and Cyclophilins. Recombinant proteins were analyzed by 15% SDS-PAGE and stained with Coomassie blue. Lane 1, total cell extract of *E. coli* electroporated with pT7-7-NS5A-D2 plasmid before and after, Lane 2, induction of expression. Lane 3, purified recombinant [^{15}N,^{13}C]-labelled NS5A-D2 (11.6kDa). The minor band in the SDS-PAGE is a proteolytic product of the D2 domain that we identified by mass spectroscopy (data not shown). CypA and CypB correspond to the purified recombinants Cyclophilin A (20.1kDa with HisTag) and Cyclophilin B (20.36kDa).

Supplemental Table 1.

NS5A-D2 min.	^{1}HN (ppm)	^{15}N (ppm)	^{13}Cα (ppm)	^{13}Cβ (ppm)	^{13}CO (ppm)	% min.
Q267	7.745	117.930	56.778	29.375	175.651	4.4
T268	8.289	115.570	61.704	70.269	173.558	8.1
E271	8.767	122.759	57.486	30.152	176.642	7.3
R273	8.309	124.049	56.204	30.760	174.869	6.4
V274	7.852	119.408	59.129	34.228	174.567	7.6
V276	8.168	124.467	62.714	32.525	176.289	7.5
L277	8.278	125.881	54.727	42.700	176.664	9.1
L280	8.087	123.768	54.955	42.933	175.787	8.4
E281	7.937	120.648	53.872	30.938	174.592	11.9
M283	8.624	121.590	55.888	33.251	175.883	7.1
L290	8.045	121.861	55.311	42.766	176.438	9.1
S293	8.546	117.165	58.402	63.817	174.356	9.8
I294	8.030	122.238	58.200	40.431	174.873	6.9
S296	8.617	116.570	58.896	64.190	175.089	4.7
E297	8.708	122.768	56.972	29.797	176.438	8.5
L300	7.958	122.093	52.756	43.206	175.522	13.6
R302	8.712	121.534	56.704	30.418	176.588	12.4
A311	8.629	124.146	53.376	18.507	177.463	8.3
W312	7.287	114.668	56.449	28.874	175.829	10.2
A313	7.518	124.740	51.887	19.351	176.622	11.6
D316	8.289	118.686	53.943	40.867	175.548	12.0
W325	7.770	121.081	57.184	29.198	175.840	6.4
R326	7.634	122.381	55.880	30.511	176.660	8.3
D329	8.316	119.212	53.473	40.750	174.477	6.9
Y330	7.836	120.158	57.832	38.815	174.565	5.5
Q331	7.986	124.351	52.965	29.280	172.528	6.5
A339	8.356	126.985	52.332	19.232	176.046	17.0
L340	7.894	120.056	52.725	42.943	175.945	8.2
L342	8.573	121.958	55.685	42.184	177.509	12.7

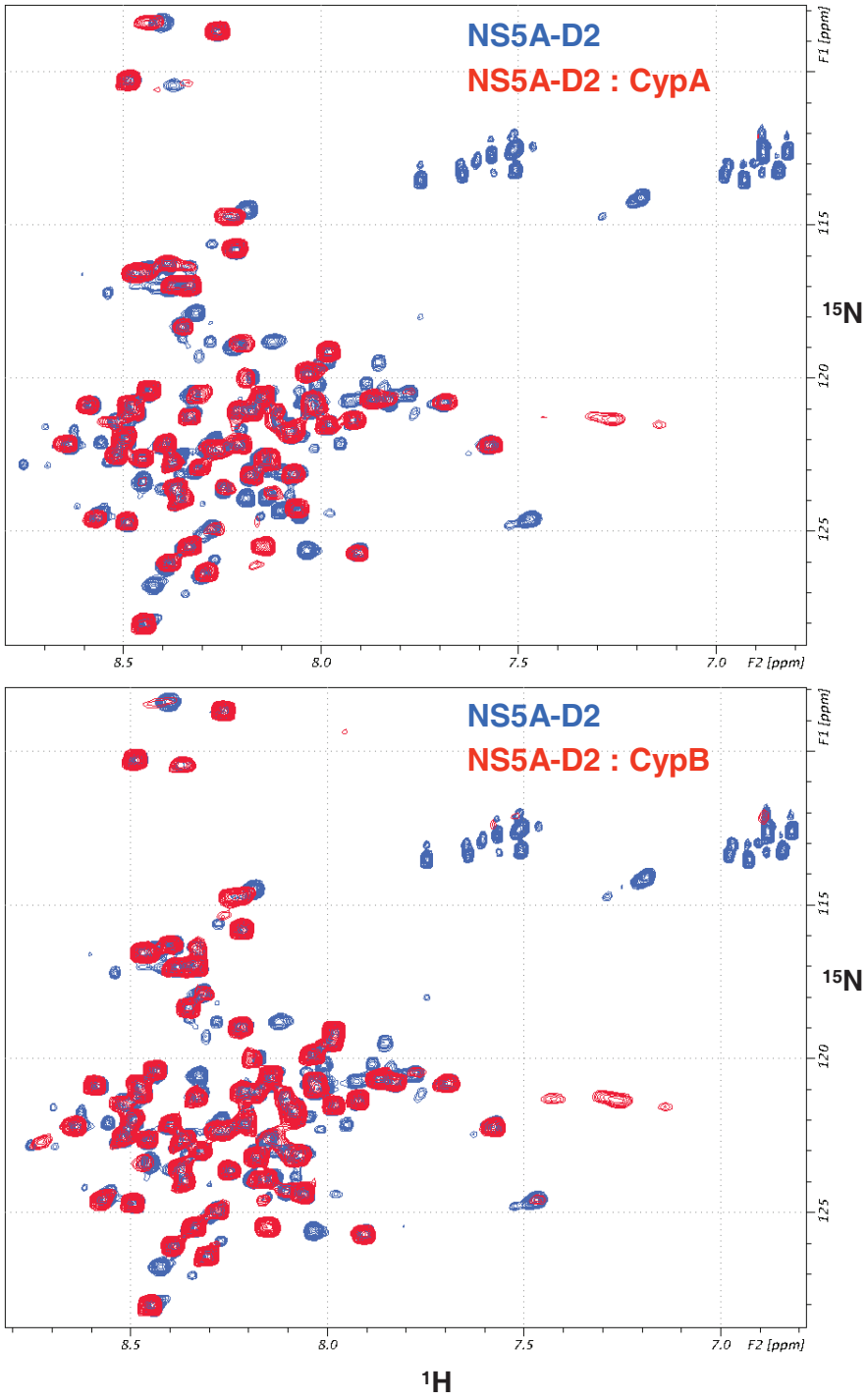
Assignments and quantifications of minor NMR resonances in NS5A-D2 ^{1}H-^{15}N-HSQC spectrum. Because of the low amide proton dispersion in the proton dimension, percentages for minor resonances compared to the corresponding major peaks were measured using maximal peak intensities rather peak integrals.

Supplemental Figure 2



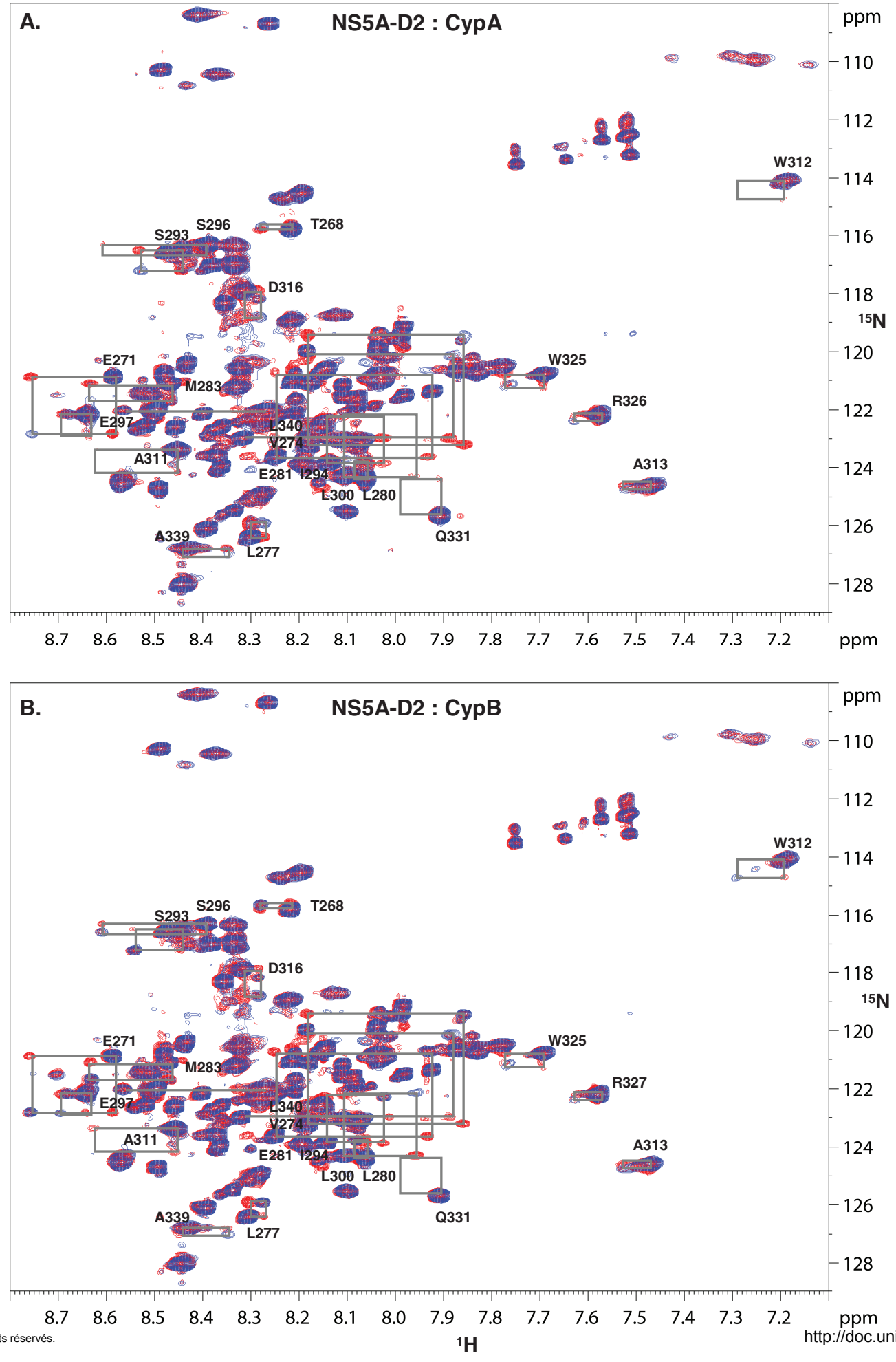
Interaction of human Cyclophilins with NS5A-D2. Each panel corresponds to the superimposition of two [^{1}H,^{15}N]-planes extracted from HNCO spectra. The first one, in red, was acquired on the [^{15}N,^{13}C]-cyclophilin alone, CypA on the left and CypB on the right, whereas the second one, in blue, was acquired on a mixture of [^{15}N,^{13}C]-Cyclophilin and [^{15}N]-NS5A-D2 in a molar ratio of 1:1. Addition of NS5A-D2 to CypB or CypA samples induces perturbations of chemical shifts and thus proving a direct interaction between these proteins.

Supplemental Figure 3



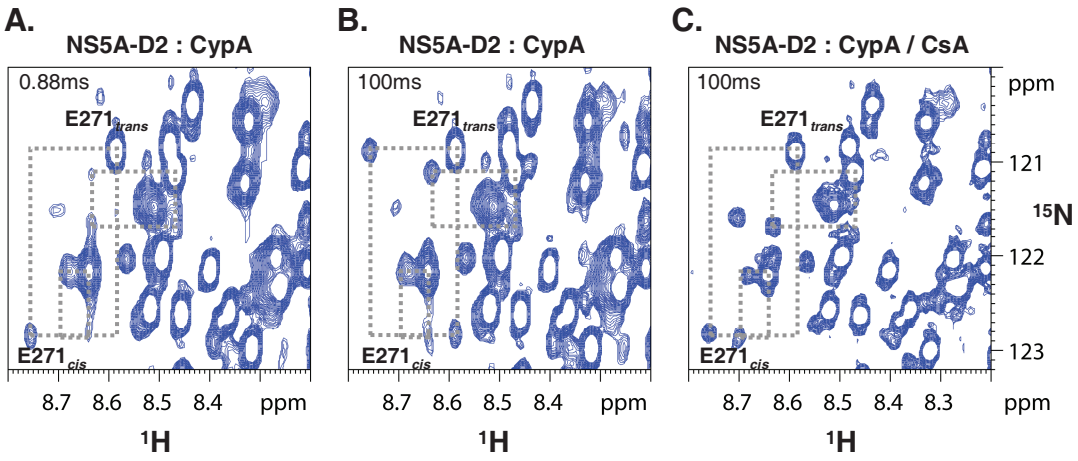
Interaction of NS5A-D2 with CypA and CypB. Each panel corresponds to the superimposition of a [^{1}H,^{15}N]-HSQC spectrum, in blue, acquired on NS5A-D2 alone and of a [^{1}H,^{15}N]-plane extracted from HNnoCO spectrum that specifically select the NS5A-D2 sub-spectrum in a NS5A-D2/Cyp (1:1) sample. Addition of Cyps causes intensive line broadening of resonances corresponding to NS5A-D2 populations where proline residues are in *cis* conformation. Furthermore several NS5A-D2 residues, in the trans populations, are differentially affected following addition of CypA or CypB.

Supplemental Figure 4



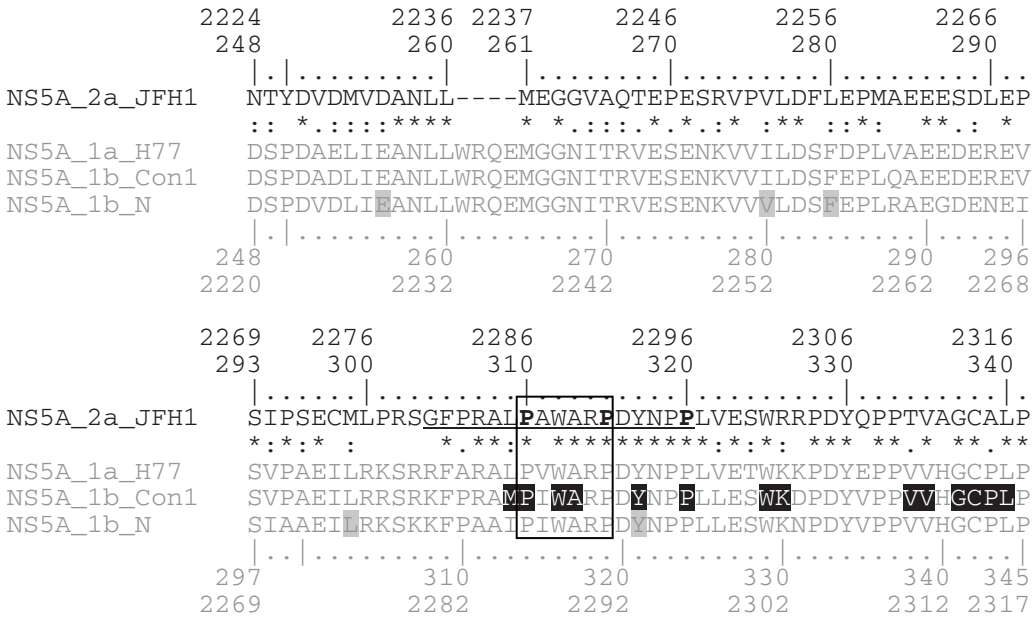
***Cis/trans* isomerization of NS5A-D2 X-Pro peptide bonds catalyzed by CypA and CypB.** $^{1}H,^{15}N$ heteronuclear exchange spectra recorded with mixing times of 0.88ms, in blue, and 100ms, in red, on [^{15}N]-NS5A-D2 samples (220µM) with catalytic amount of either CypA (**A.**) or CypB (**B.**) (23µM). The NMR resonances (*trans*, *cis* and the two exchange peaks) of NS5A-D2 residues for which the PPIase activity of a cyclophilin can be evidenced are connected by gray boxes. Exchange spectra were acquired on a 800MHz.

Supplemental Figure 5



Inhibition of PPIase activity of CypA on NS5A-D2. $^{1}H,^{15}N$ heteronuclear exchange spectra were acquired with mixing times of 0.88ms (**A.**) and 100ms (**B.**, **C.**) on a NS5A-D2:CypA (220µM:23µM) sample without (**B.**) or with an excess of CsA (**C.**). The NS5A-D2 exchange peaks due to CypA catalyzed isomerizations are no more detectable in presence of CsA, an inhibitor of cyclophilins PPIase activity.

Supplemental Figure 6



Sequences alignment of NS5A-D2 from different HCV strains. NS5A-D2 sequence of JFH1 strain (AB047639, genotype 2a) is shown in black, while the corresponding sequences of H77 (AF009606, genotype 1a), Con1 (AJ238799, genotype 1b) and N (AF139594, genotype 1b) strains are shown in grey. Sequences are numbered with respect to NS5A and the HCV polyproteins. The hyphens indicate the aa deletions comparatively to the sequence alignment. Identical, strongly conservative, and weakly conservative amino acid are indicated by asterisks, colons, and dots, respectively, accordingly to CLUSTALW convention. Prolines residues that are fully conserved in any genotypes are in bold in JFH1 sequence. The open box indicates a preferential CypA and CypB interacting region. The larger motif recognized by CypA is underlined (straight thin line).

NS5A-D2 study performed by Liang *et al.* in 2006 (37) was on genotype 1a (H77) (58.1% identity along the complete sequence and only 48.2% for domain 2). Residues reported by Tellinghuisen *et al.* in 2008 (29) to be essential for HCV replication in Con1 replicon are highlighted in black. Residues highlighted in grey correspond to those reported by Fernandez *et al.* (26) to be associated to HCV resistance to CsA in a HCV replicon of genotype 1b (AF139594). Note that the numbering of these residues in AF139594 strain is different because of a four aa insertion in the NS5A sequence.

3.2.2. Domain 3 of Non-Structural Protein 5A from Hepatitis C Virus is Natively Unfolded

Biochemical and Biophysical Research Communications 381 (2009) 634–638

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Domain 3 of non-structural protein 5A from hepatitis C virus is natively unfolded

Xavier Hanoulle [a,*], Dries Verdegem [a], Aurélie Badillo [b], Jean-Michel Wieruszeski [a], François Penin [b], Guy Lippens [a,*]

[a] UGSF, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France
[b] IBCP, UMR 5086 CNRS, Université de Lyon, IFR 128 BioSciences Gerland-Lyon-Sud, 69397 Lyon, France

ARTICLE INFO

Article history:
Received 13 February 2009
Available online 26 February 2009

Keywords:
Hepatitis C virus
NS5A
Domain 3
NMR
Unstructured
Circular dichroism

ABSTRACT

Hepatitis C virus (HCV) non-structural protein 5A (NS5A) is involved both in the viral replication and particle production. Its third domain (NS5A-D3), although not absolutely required for replication, is a key determinant for the production and assembly of novel HCV particles. As a prerequisite to elucidate the precise functions of this domain, we report here the first molecular characterization of purified recombinant HCV NS5A-D3. Sequence analysis indicates that NS5A-D3 is mostly unstructured but that short structural elements may exist at its N-terminus. Gel filtration chromatography, circular dichroism and finally NMR spectroscopy all point out the natively unfolded nature of purified recombinant NS5A-D3. This lack of stable folding is thought to be essential for primary interactions of NS5A-D3 domain with other viral or host proteins, which could stabilize some specific conformations conferring new functional features.

Introduction

Hepatitis C virus (HCV) is classified in the *Hepacivirus* genus within the *Flaviviridae* family. HCV infection often lead to chronic hepatitis, liver cirrhosis and hepatocellular carcinoma [1]. With 120–180 million people persistently infected worldwide, the absence of a vaccine and limited efficacy of current drug treatments turn HCV into a serious health challenge. HCV is a small (+)RNA enveloped virus. The HCV viral genome (~9.6 kb) is translated in a unique polyprotein of ~3000 amino acids [2]. Its processing by viral encoded or host proteases, in a co- and post-translational way, leads to at least 10 different proteins. These viral proteins are classified into structural (Core, E1 and E2) and non-structural (p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B) proteins. Non-structural HCV proteins are involved in the processing of the polyprotein precursor and in the viral replication. The minimal set of proteins required to achieve viral replication is NS3, NS4A, NS4B, NS5A and NS5B [3]. NS5A is a large (49 kDa) phospho-protein absolutely required for HCV replication and particle assembly [4–7] but for which the precise function(s) remains to be elucidated. Up to date, no enzymatic activity has been detected for this viral protein. Recently, de Chassey et al. have reported a protein interaction network during HCV infection [8]. NS5A is the viral protein, following NS3, displaying the highest number of interactions with human proteins. NS5A is anchored to the ER membrane on its cytoplasmic side via an amphipathic N-terminal α-helix [5,9]. Its cytoplasmic part is constituted by three domains, NS5A-D1, -D2 and -D3, that are connected by low-complexity sequences [10]. NS5A-D1, for which a X-ray structure has been solved [11], is a zinc-binding domain with RNA binding activity [12]. Sequences of NS5A-D2 and -D3 are significantly less conserved among the HCV genotypes than for -D1. Domain 2 of NS5A is required for HCV replication [7,13] and was shown to be natively unfolded [14]. NS5A-D3 (residues 356–447) is dispensable for HCV RNA replication, but has an essential role for viral particle production and assembly [4,15]. It has been proposed to be equally natively unfolded [6,10], but despite its key role in HCV infection, no biochemical and structural data have been presented to underscore this feature. This study combines bioinformatics with experimental biochemical and biophysical tools to characterize, for the very first time, HCV NS5A-D3 at a molecular level. We report direct experimental evidence showing the disordered nature of isolated NS5A-D3.

Materials and methods

Sequence analysis. Sequence analyses were performed using tools available at the Institut de Biologie et Chimie des Protéines (IBCP) Network Protein Sequence Analysis (NPSA) website

* Corresponding authors. Fax: +33 (0)3 20 43 65 55.
E-mail addresses: xavier.hanoulle@univ-lille1.fr (X. Hanoulle), guy.lippens@univ-lille1.fr (G. Lippens).

0006-291X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2009.02.108

(http://npsa-pbil.ibcp.fr) [16]. HCV NS5A sequences were retrieved from the European HCV Database (http://euhcvdb.ibcp.fr/) [17]. Multiple-sequence alignments were performed with Clustal W. The repertoire of residues at each aa position and their frequencies observed in natural sequence variants were computed by the use of a program developed at the IBCP (F. Dorkeld, C. Combet, F. Penin, and G. Deléage, unpublished data). The disorder prediction was done using the PONDR software [18–20]. The mean net charge/mean hydrophobicity analysis was done using same parameters as in [21].

Cloning, expression and purification of NS5A-D3. The synthetic sequence coding for domain 3 of NS5A from HCV Con1 strain (euHCVdb [17]; #AJ238799, genotype 1b) was cloned in the plasmid pT7.7 [22]. Expression of the recombinant domain was done in *Escherichia coli* BL21(DE3) Star strain. The resulting domain 3 of HCV NS5A (NS5A-D3) (residues 359–445) has extra M- and -LQHHHHHH extensions at N- and C-termini. Bacteria were grown in Luria–Bertani medium for non-labeled protein or in M9-based semi rich medium (M9 medium supplemented with [^{15}N]-NH_4Cl (1 g/L), [$^{13}C_6$]-D-glucose (2 g/L) (when ^{13}C labeling required), Isogro ^{13}C,^{15}N powder growth medium (0.7 g/L, Sigma) and ampicillin (100 μg/ml). When OD_{600nm} reached ~0.8 induction was carried out at 37 °C with 0.3 mM isopropyl-β-D-galactopyranoside (IPTG) for 4 h. Harvested cells were lyzed using lysozyme and sonication. NS5A-D3 was first purified on a HisTrap column (1 ml, GE Healthcare). Selected fractions were then pooled, dialyzed again 20 mM Tris–Cl, pH 7.4, 2 mM EDTA overnight and loaded on a anion exchange column (ResourceQ 1 ml, GE Healthcare). NS5A-D3 containing fractions were selected, using SDS–PAGE analysis, pooled and dialyzed overnight against 20 mM NaH_2PO_4/Na_2HPO_4 pH 6.4, 30 mM NaCl, 0.02% NaN_3, 0.2 mM THP buffer. NS5A-D3 was then concentrated up to 440 μM with a Vivaspin 15 concentrator (cut-off 5 kDa) (Satorius Stedim Biotech). Following filtration at 0.2 μm, NS5A-D3 aliquots were frozen at −80 °C with few Chelex beads (Sigma).

Circular dichroism (CD). CD spectra were recorded on a Chirascan dichrographe (Applied Photophysics, Surrey, UK) calibrated with 1*S*-(+)-10-camphorsulfonic acid. Measurements were carried out at room temperature in a 0.1-cm path length quartz cuvette, with protein concentrations at 8 μM. Spectra were recorded in the 185–260 nm wavelength range with a 0.5 nm increment and a 1 s integration time. Spectra were processed, baseline corrected, and smoothed using Chirascan software. Spectral units were expressed as the molar ellipticity per residue by using protein concentrations determined by measuring the UV light absorbance of tyrosine and tryptophane at 280 nm.

NMR data collection and assignments. NMR spectra were acquired on a 440 μM [^{15}N,^{13}C]-labeled NS5A-D3 sample using a Bruker Avance 800 MHz spectrometer with a standard triple resonance probe (Bruker Biospin, Karlsruhe, Germany). The proton chemical shifts were referenced using the methyl signal of trimethyl silyl propionate (TMSP). Spectra were processed with the Bruker TopSpin package. Spectra were analyzed using the product-plane method developed in our laboratory [23]. NS5A-D3 backbone assignments were realized using 2D ^{1}H,^{15}N HSQC and triple resonance 3D HNCO, HN(CA)CO, HNCACB, HN(CO)CACB and HN(CA)NNH spectra. Assignments were deposited to the BMRB, accession number 16166.

Results and discussion

Sequence analyses of NS5A domain 3

Sequence analyses and structure predictions were performed to assess the degree of conservation of the NS5A-D3 domain and to identify potential essential amino acids (aa) and motifs. The aa repertoire deduced from the analysis of 363 HCV isolates of genotype 1b revealed some aa strictly conserved in 50% of sequence positions (denoted by asterisks in Fig. 1), mainly in the 359–380 and 408–439 regions. The apparent variability of the central 381–407 region is however not that important at many positions, as the observed residues exhibit similar physicochemical properties. Both by the similarity patterns (dots) as well as the hydropathic patterns in this region (see legend to Fig. 1 for details) remain well conserved. Remarkably, the strict conservation of residues at positions 389, 390 and 393 in this variable region likely points to an essential role of these residues for the structure and/or function of NS5A-D3. According to all secondary structure prediction methods tested and summarized in Fig. 1 (bottom), the NS5A-D3 sequence does not seem to display regular secondary structure elements, except in the N-terminal region 359–378 where some

Fig. 1. Sequence analysis and biochemical characterization of NS5A domain 3 (NS5A-D3). Amino acid repertoires. The NS5A-D3 amino acids (aa) 359–445 sequence from the HCV Con1 strain of genotype 1b (GenBank Accession No. AJ238799), which was used in this study, is indicated. Residues are numbered with respect to NS5A and the HCV Con1 polyprotein (top row). The aa repertoire deduced from the Clustal W multiple alignments of 363 NS5A sequences of genotype 1b is reported. Amino acids observed at a given position in fewer than three distinct sequences (<1%) were not included. The degree of aa and physicochemical conservation at each position can be inferred from the extent of variability (with the observed aa listed in decreasing order of frequency from top to bottom) together with the similarity index according to Clustal W convention (asterisk, invariant; colon, highly similar; dot, similar), and the consensus hydropathic pattern deduced from the consensus aa repertoire. o, hydrophobic position (F, I, W, Y, L, V, M, P, C); n, neutral position (G, A, T, S); i, hydrophilic position (K, Q, N, H, E, D, R); v, variable position (i.e., when both hydrophobic and hydrophilic residues are observed at a given position). Secondary structure predictions are indicated as helical (h), extended (e), undetermined (coil, c), or ambiguous (?). Second Cons., consensus of protein secondary structures predictions for NS5A-D3 from Con1 strain deduced from the set of prediction methods available at the NPSA website, including DPM, DSC, HNNC, MLRC, PHD, Predator, and SOPM (see http://npsa-pbil.ibcp.fr [16] and references therein).

extended and α-helix stretches are predicted by most methods. In addition, the PONDR software (www.pondr.com) was used to predict disorder in NS5A-D3. The analysis performed with the VL-XT algorithm [18–20] indicates that, with exception of the first 16 residues, almost all the NS5A-D3 sequence is predicted to be disordered. Analysis of the amino acids content of NS5A-D3 reveals that this domain is highly acidic (Asp + Glu = 21 residues, 21.8%), serine/threonine-rich (Ser + Thr = 27 residues, 31%), proline-rich (6 residues, 6.9%). The calculated p*I* for NS5A-D3 (Con1) is 3.4 and it is consistent with the absence of structure as most of the natively unfolded proteins display extreme values for the calculated isoelectric point (p*I*), above 9 or below 5 [21]. A charge/hydrophobicity analysis, which has been shown to be efficient in the prediction of the folded/unfolded state of a protein [21], was done on NS5A-D3 (Supplementary Fig. 1). This domain has a mean net charge and a mean hydrophobicity, respectively, of 0.15 and 0.44 and hence corresponds to the unfolded protein group. All the bioinformatics tools that have been used thus predict that NS5A-D3 is mainly unfolded but do hint to a potential α-helix at its N-terminus.

Expression, purification and molecular characterization of NS5A-D3 from HCV strain Con1

Domain 3 of NS5A protein from HCV Con1 strain (ID AJ238799, genotype 1b) was efficiently overexpressed in *E. coli* BL21(DE3) Star cells (Invitrogen) (Fig. 2A, lanes 1 and 2). The resulting recombinant domain 3 (denoted NS5A-D3, residue 359–445) was well soluble and was purified to homogeneity, as shown in Fig. 2A (lane 3), following a two-step protocol. The purified protein exhibited the expected mass (9925 Da) as determined by mass spectroscopy (not shown). By SDS–PAGE, NS5A-D3 has an apparent molecular weight (MW) of ~20 kDa (Fig. 2A). This discrepancy is probably due to the primary sequence of NS5A-D3, which includes many acidic residues and prolines. In gel filtration chromatography, the protein elutes at a volume corresponding to a mass of ~45 kDa for a globular protein (Fig. 2B). Such a large apparent MW in a gel filtration assay is commonly associated with natively unfolded proteins devoid of globular domain [24].

The circular dichroism (CD) spectrum of NS5A-D3 exhibits a single negative peak centered at 198 nm (Fig. 2C) that is typical of unfolded polypeptide. The lack of characteristic minima around 208 and 222 nm indicative of α-helix, or the minimum around 215 nm that points to β-sheet elements, confirms the absence of regular secondary structure in NS5A-D3. Uversky has proposed that a plot representing the molar ellipticity at 200 nm versus 222 nm may be used to classify unfolded proteins in two groups: coil-like or pre-molten globular proteins [25]. According to this criterium, NS5A-D3 belongs to the coil-like group (Supplementary Fig. 2). In order to confirm these predictions at the per-residue level, we used NMR spectroscopy to gain structural data at atomic level.

NMR spectroscopy on HCV NS5A-D3

The ^{1}H,^{15}N HSQC spectrum of NS5A-D3 (Fig. 3A) exhibits a rather narrow ^{1}H chemical shift dispersion limited to 0.75 ppm. NS5A-D3 resonances are clustered in three regions of the NMR spectrum corresponding to, respectively, the glycines, the serine and threonine and finally the other residues. This clustering and the limited dispersion in proton dimension are typical of non-structured polypeptides. Three dimensional NMR experiments were recorded on a ^{15}N,^{13}C-labeled NS5A-D3 sample. During our NMR work on the neuronal Tau protein [26–28], we have developed a product-plane based assignment method [23] allowing to cope with the large degree of overlap both for the proton and carbon frequencies in unfolded proteins. The same strategy proved efficient with the spectra of NS5A-D3, as all the non-prolyl backbone proton amide resonances could be assigned (Fig. 3A), with the exception of the first Met residue from the cloning strategy. All the $^{13}C_O$, $^{13}C_\alpha$ and $^{13}C_\beta$ resonances from NS5A-D3 were assigned except for the Pro417 which is in a 417-PP-418 motif. Using the chemical shift index (CSI) method [29], these ^{13}C values were used to probe the secondary structure content in NS5A-D3 at a *per residue* level (Fig. 3B). The analysis revealed that $^{13}C_\beta$ CSI values are mostly positive whereas $^{13}C_\alpha$ and $^{13}C_O$ ones are almost equally divided between positives and negatives. The resulting CSI consensus values are zero (Fig 3B, bottom), confirming the absence in NS5A-D3 of stable secondary structure elements even at the local level. Nevertheless, careful inspection of CSI data in the region 369–379 of NS5A-D3 shows that this region, with positives CSI values for $^{13}C_\alpha$ and $^{13}C_O$, may have some tendency to α-helical character. Interestingly, this N-terminal region was predicted to contain α-helical elements by most of the prediction methods used and also corresponds to a non-disordered region in the PONDR analysis. NS5A-D3 is thus natively unfolded, at least when isolated from the other NS5A domains, although its N-terminus displays a certain propensity to α-helical structuration that is not stable enough to be characterized by CD or NMR spectroscopy.

Next to the assigned NS5A-D3 resonances, numerous less intense peaks could be observed in the ^{1}H,^{15}N HSQC spectrum (Fig. 3A). These peaks correspond to NS5A-D3 residues for which

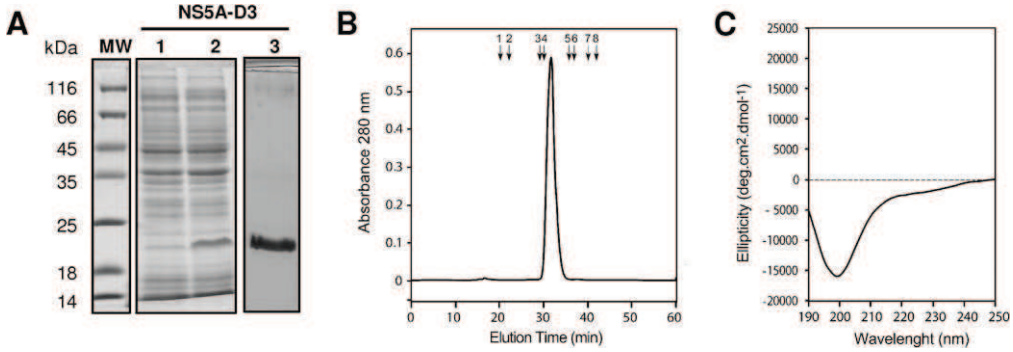


Fig. 2. (A) Expression and purification of NS5A domain 3 (HCV Con1 strain). Recombinant proteins were analyzed by 15% SDS–PAGE and stained with Coomassie blue. Lane 1, total cell extract of *E. coli* BL21(DE3) Star electroporated with pT7-7-NS5A-D3 plasmid before and, lane 2, after induction of expression. Lane 3, purified recombinant [^{15}N,^{13}C]-labeled NS5A-D3 (9.9 kDa). (B) Gel filtration analysis of NS5A-D3 was performed on a Superdex S200 column equilibrated in sodium phosphate buffer, pH 7.4, with a flow rate of 0.5 ml/min. Elution volumes of globular protein standards are indicated by black arrows with the following corresponding molecular weights: 1, thyroglobulin (669,000 Da); 2, ferritin (440,000 Da); 3, aldolase (158,000 Da); 4, conalbumin (75,000 Da); 5, ovalbumin (43,000 Da); 6, chymotrypsin (25,000 Da); 7, ribonuclease (13,700 Da); 8, vit B12 (1355 Da). (C) Far-UV circular dichroism analysis of 8 μM NS5A-D3 in 10 mM sodium phosphate, pH 7.4.

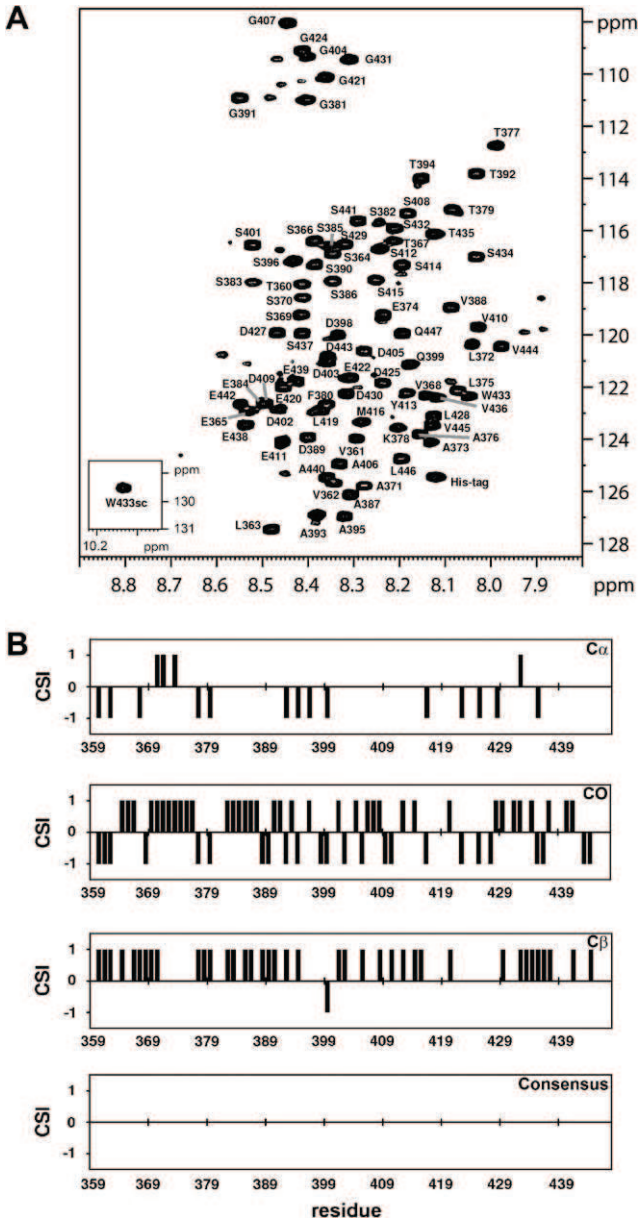


Fig. 3. (A) Assigned ^{1}H,^{15}N HSQC NMR spectrum of domain 3 of NS5A (HCV Con1 strain). The small insert show the spectrum region corresponding to the Trp433 side chain. The spectrum was recorded at 800 MHz. Assignments of minor peaks, resulting from the *cis/trans* equilibriums of the six proline residues in NS5A-D3, are available as Supplementary data. (B) Chemical shift index (CSI) analysis of NS5A-D3. $^{13}C_{\alpha}$, $^{13}C_{\beta}$ and $^{13}C_{O}$ chemical shifts were analyzed using the CSI software [29]. The resulting consensus CSI values are zero all along the NS5A-D3 sequence.

their proton amide resonances are sensitive to the *cis/trans* equilibrium of a proline in their direct neighborhood. Twenty of these minor peaks have been assigned to different residues (see Supplementary Table 1) located in the close proximity of one of the six prolines in NS5A-D3 (see Fig. 1).

In summary, we report here the first molecular characterization of the domain 3 of HCV NS5A protein. We show efficient recombinant NS5A-D3 expression in *E. coli* and purification. Using a combination of sequence analysis tools and experimental measurements we demonstrate that NS5A-D3 is natively unfolded in solution. Our results provide the first molecular basis for further understanding of NS5A-D3 interactions with a variety of biological partners and the essential functional role of this domain in HCV particle formation.

Acknowledgments

This work was supported by the French Centre National de la Recherche Scientifique and Universities of Lille and Lyon and a grant from the French National Agency for Research on AIDS and viral Hepatitis (ANRS). X. Hanoulle was supported by a fellowship from the ANRS. The NMR facility used in this study was funded by the Région Nord-Pas-de-Calais (France), the CNRS, the Universities of Lille1 and Lille2 and the Institut Pasteur de Lille.

The authors gratefully acknowledge RD-Biotech (Besançon, France) for the cloning and the initial expression and purification tests for NS5A-D3, Guillaume Blanc and Jennifer Molle for technical assistance, Michel Becchi for the mass spectroscopy measurements, Christophe Combet for bioinformatics support, and Ralf Bartenschlager for valuable discussions and help.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2009.02.108.

References

[1] NIH, Consensus Development Conference statement. Management of hepatitis C, Hepatology 36 (2002) S2–S20.
[2] D. Moradpour, F. Penin, C.M. Rice, Replication of hepatitis C virus, Nat. Rev. Microbiol. 5 (2007) 453–463.
[3] V. Lohmann, F. Korner, J. Koch, U. Herian, L. Theilmann, R. Bartenschlager, Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line, Science 285 (1999) 110–113.
[4] N. Appel, M. Zayas, S. Miller, J. Krijnse-Locker, T. Schaller, P. Friebe, S. Kallis, U. Engel, R. Bartenschlager, Essential role of domain III of nonstructural protein 5A for hepatitis C virus infectious particle assembly, PLoS Pathog. 4 (2008) e1000035.
[5] F. Penin, V. Brass, N. Appel, S. Ramboarina, R. Montserret, D. Ficheux, H.E. Blum, R. Bartenschlager, D. Moradpour, Structure and function of the membrane anchor domain of hepatitis C virus nonstructural protein 5A, J. Biol. Chem. 279 (2004) 40835–40843.
[6] F. Penin, J. Dubuisson, F.A. Rey, D. Moradpour, J.M. Pawlotsky, Structural biology of hepatitis C virus, Hepatology 39 (2004) 5–19.
[7] T.L. Tellinghuisen, K.L. Foss, J.C. Treadaway, C.M. Rice, Identification of residues required for RNA replication in domains II and III of the hepatitis C virus NS5A protein, J. Virol. 82 (2008) 1073–1083.
[8] B. de Chassey, V. Navratil, L. Tafforeau, M.S. Hiet, A. Aublin-Gex, S. Agaugue, G. Meiffren, F. Pradezynski, B.F. Faria, T. Chantier, M. Le Breton, J. Pellet, N. Davoust, P.E. Mangeot, A. Chaboud, F. Penin, Y. Jacob, P.O. Vidalain, M. Vidal, P. Andre, C. Rabourdin-Combe, V. Lotteau, Hepatitis C virus infection protein network, Mol. Syst. Biol. 4 (2008) 230.
[9] V. Brass, E. Bieck, R. Montserret, B. Wolk, J.A. Hellings, H.E. Blum, F. Penin, D. Moradpour, An amino-terminal amphipathic alpha-helix mediates membrane association of the hepatitis C virus nonstructural protein 5A, J. Biol. Chem. 277 (2002) 8130–8139.
[10] T.L. Tellinghuisen, J. Marcotrigiano, A.E. Gorbalenya, C.M. Rice, The NS5A protein of hepatitis C virus is a zinc metalloprotein, J. Biol. Chem. 279 (2004) 48576–48587.
[11] T.L. Tellinghuisen, J. Marcotrigiano, C.M. Rice, Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase, Nature 435 (2005) 374–379.
[12] L. Huang, J. Hwang, S.D. Sharma, M.R. Hargittai, Y. Chen, J.J. Arnold, K.D. Raney, C.E. Cameron, Hepatitis C virus nonstructural protein 5A (NS5A) is an RNA-binding protein, J. Biol. Chem. 280 (2005) 36417–36428.
[13] S. Liu, I.H. Ansari, S.C. Das, A.K. Pattnaik, Insertion and deletion analyses identify regions of non-structural protein 5A of hepatitis C virus that are dispensable for viral genome replication, J. Gen. Virol. 87 (2006) 323–327.
[14] Y. Liang, H. Ye, C.B. Kang, H.S. Yoon, Domain 2 of nonstructural protein 5A (NS5A) of hepatitis C virus is natively unfolded, Biochemistry 46 (2007) 11550–11558.
[15] T.L. Tellinghuisen, K.L. Foss, J. Treadaway, Regulation of hepatitis C virion production via phosphorylation of the NS5A protein, PLoS Pathog. 4 (2008) e1000032.
[16] C. Combet, C. Blanchet, C. Geourjon, G. Deleage, NPS@: network protein sequence analysis, Trends Biochem. Sci. 25 (2000) 147–150.
[17] C. Combet, N. Garnier, C. Charavay, D. Grando, D. Crisan, J. Lopez, A. Dehne-Garcia, C. Geourjon, E. Bettler, C. Hulo, P. Le Mercier, R. Bartenschlager, H. Diepolder, D. Moradpour, J.M. Pawlotsky, C.M. Rice, C. Trepo, F. Penin, G. Deleage, euHCVdb: the European hepatitis C virus database, Nucleic Acids Res. 35 (2007) D363–D366.

[18] X. Li, P. Romero, M. Rani, A.K. Dunker, Z. Obradovic, Predicting protein disorder for N-, C-, and internal regions, Genome Inform. Ser. Workshop Genome Inform. 10 (1999) 30–40.
[19] P. Romero, Z. Obradovic, A.K. Dunker, Sequence data analysis for long disordered regions prediction in the Calcineurin family, Genome Inform. Ser. Workshop Genome Inform. 8 (1997) 110–124.
[20] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins 42 (2001) 38–48.
[21] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (2000) 415–427.
[22] J.C. Cortay, D. Negre, M. Scarabel, T.M. Ramseier, N.B. Vartak, J. Reizer, M.H. Saier Jr., A.J. Cozzone, In vitro asymmetric binding of the pleiotropic regulatory protein, FruR, to the ace operator controlling glyoxylate shunt enzyme synthesis, J. Biol. Chem. 269 (1994) 14885–14891.
[23] D. Verdegem, K. Dijkstra, X. Hanoulle, G. Lippens, Graphical interpretation of Boolean operators for protein NMR assignments, J. Biomol. NMR 42 (2008) 11–21.
[24] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (2002) 527–533.
[25] V.N. Uversky, Natively unfolded proteins: a point where biology waits for physics, Protein Sci. 11 (2002) 739–756.
[26] G. Lippens, A. Sillen, C. Smet, J.M. Wieruszeski, A. Leroy, L. Buee, I. Landrieu, Studying the natively unfolded neuronal Tau protein by solution NMR spectroscopy, Protein Pept. Lett. 13 (2006) 235–246.
[27] G. Lippens, J.M. Wieruszeski, A. Leroy, C. Smet, A. Sillen, L. Buee, I. Landrieu, Proline-directed random-coil chemical shift values as a tool for the NMR assignment of the tau phosphorylation sites, Chembiochem 5 (2004) 73–78.
[28] C. Smet, A. Leroy, A. Sillen, J.M. Wieruszeski, I. Landrieu, G. Lippens, Accepting its random coil nature allows a partial NMR assignment of the neuronal Tau protein, Chembiochem 5 (2004) 1639–1646.
[29] D.S. Wishart, B.D. Sykes, The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data, J. Biomol. NMR 4 (1994) 171–180.

3.2.3. Investigation of the Structural Properties of the Third Domain of HCV's NS5A and its Interaction with Cyclophilin A

We have taken the research on NS5A-D3 one step further by performing a more detailed assay of its structural propensities and, just as we did for NS5A-D2, by studying the interaction with the cyclophilins. These investigations were performed both on the Con1 and JFH1 Hepatitis C virus strain, contrarily to the previous work where we have only considered the third domain of the Con1 strain. A manuscript describing the following results is in preparation.

Where the work presented in the paper above ("Domain 3 of non-structural protein 5A from hepatitis C virus is natively unfolded") mostly indicates that the third domain of the non-structural protein 5A (NS5A-D3) belongs the class of intrinsically disordered proteins, we here show the presence of residual α-helical structure towards either end of the amino acid sequence of this domain in both mentioned strains (Con1 and JFH1) of the Hepatitis C virus. Increased levels of α-helical secondary structure in these regions of NS5A-D3 were inducible in 50% trifluoroethanol (TFE). We have used multidimensional NMR spectroscopy, more specifically $^3J_{HN,H\alpha}$ and NOE measurements, to characterise the induced structuration.

An extended sequence analysis is a first indicator of tendencies towards secondary structure elements. We have analysed the amino acid sequences of the Con1 and JFH1 versions of NS5A-D3 using a few disorder/residual structure predictors: PONDR [326, 236, 328, 301, 289, 300] and the metapredictor metaPrDOS [50]. MetaPrDOS employs a series of individual disorder predictors and subsequently delivers a consensus of the different results obtained. Its output is represented in fig. 3.3. The results indicate that the sequence of the third domain of NS5A in both strains of the virus implements a mostly disordered protein (the curves are mostly above the threshold of 0.5). However, at the N-terminus end, a dip in the consensus curve is clearly visible, which indicates a lesser degree of disorder. In both cases, the curve also declines towards the very end of the C-terminus, which is also indicative for a higher degree of structure. The results of the PONDR predictor, together with the same metapredictor results and some NMR derived information are further shown in fig. 3.6.

In addition to this initial analysis, the NMR sequential assignment was performed on four samples using the classical triple resonance assignment strategy and the assignment tool described in chapter 2. The four samples correspond to (*a*) NS5A-D3-Con1 in aqueous solution, (*b*) NS5A-D3-JFH1 in aqueous solution, (*c*) NS5A-D3-Con1 in 50% TFE and (*d*) NS5A-D3-JFH1 in 50% TFE. The samples conditions were (*a*) 440 μM of protein in 600 μL of sample, 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 1 mM THP, 0.02% NaN_3, 5% D_2O, pH 6.4 at 298 K. (*b*) 500 μM of protein in 350 μL of sample, 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 1 mM THP, 0.02% NaN_3, 5% D_2O, pH 6.3 at 298 K. (*c*) 360 μM of protein in 600 μL of sample, 50% deuterated TFE, 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 1 mM THP, 0.02% NaN_3, 5% D_2O, pH 6.4 at 298 K and (*d*) 500 μM of protein in 600 μL of sample, 50% of deuterated TFE, 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 2 mM THP, 0.02% NaN_3, 5% D_2O, pH 6.35 at 298 K. The assignment spectra (HNCO, HN(CA)CO, HNCACB, HN(CO)CACB and HN(CA)NNH) were in all cases recorded on a Bruker Avance 800 MHz spectrometer (Bruker

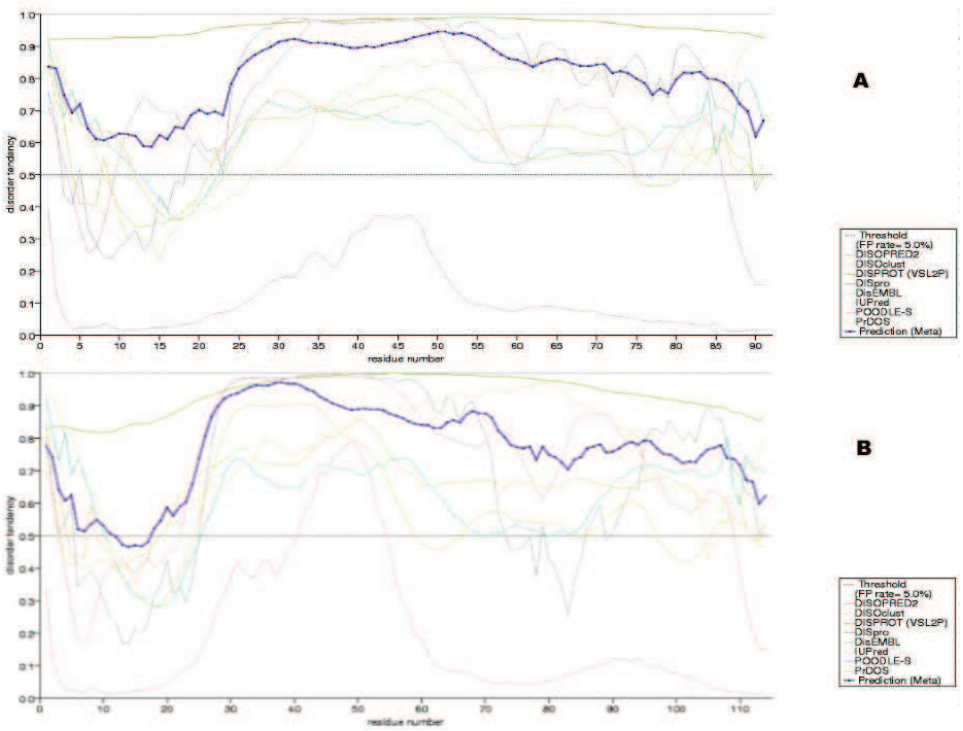


Figure 3.3. Visualisation of the disorder-predictor output for (A) NS5A-D3-Con1 and (B) NS5A-D3-JFH1. The outputs of the individual predictors are drawn in light colours (colour code explained in the legend). The consensus calculated by MetaPrDOS is drawn in intense blue. The threshold is at 0.5. Above this value, the protein is considered disordered; below 0.5, it is considered structured.

Biospin, Karlsruhe, Germany). HSQCs of all four protein/solvent combinations were completely assigned including most minor peaks corresponding to residues (sequentially) not far from proline residues in the cis conformation. The assigned HSQCs are shown in figs. 3.4 and 3.5.

The narrow proton chemical shift dispersion these HSQC spectra exhibit, again points out the random coil character of the proteins. The spectra recorded in 50% TFE contain some peaks that have shifted downfield compared to situation in aqueous solution. These peaks correspond to residues that appear in zones undergoing structuration in the TFE solution. Indeed, TFE has been demonstrated to stabilise secondary structure elements, especially α-helical structure. The assigned triple resonance spectra allowed for a further chemical shift analysis. C_α, C_β and CO chemical shifts are known to be quite sensitive to structuration of the backbone chain. We have first performed an SSP-analysis of the chemical shifts obtained for the proteins in aqueous solution. As mentioned previously the SSP-tool provides a quantitative measure for the fractional secondary structure propensity. The results of this analysis (presented in fig. 3.6) mainly confirms the general picture obtained from the sequence analysis, i.e. both strains of the virus have a NS5A-D3 protein domain with clear residual α-helical structure near the N-terminus and (although less clear because it concerns a shorter stretch) the C-terminus.

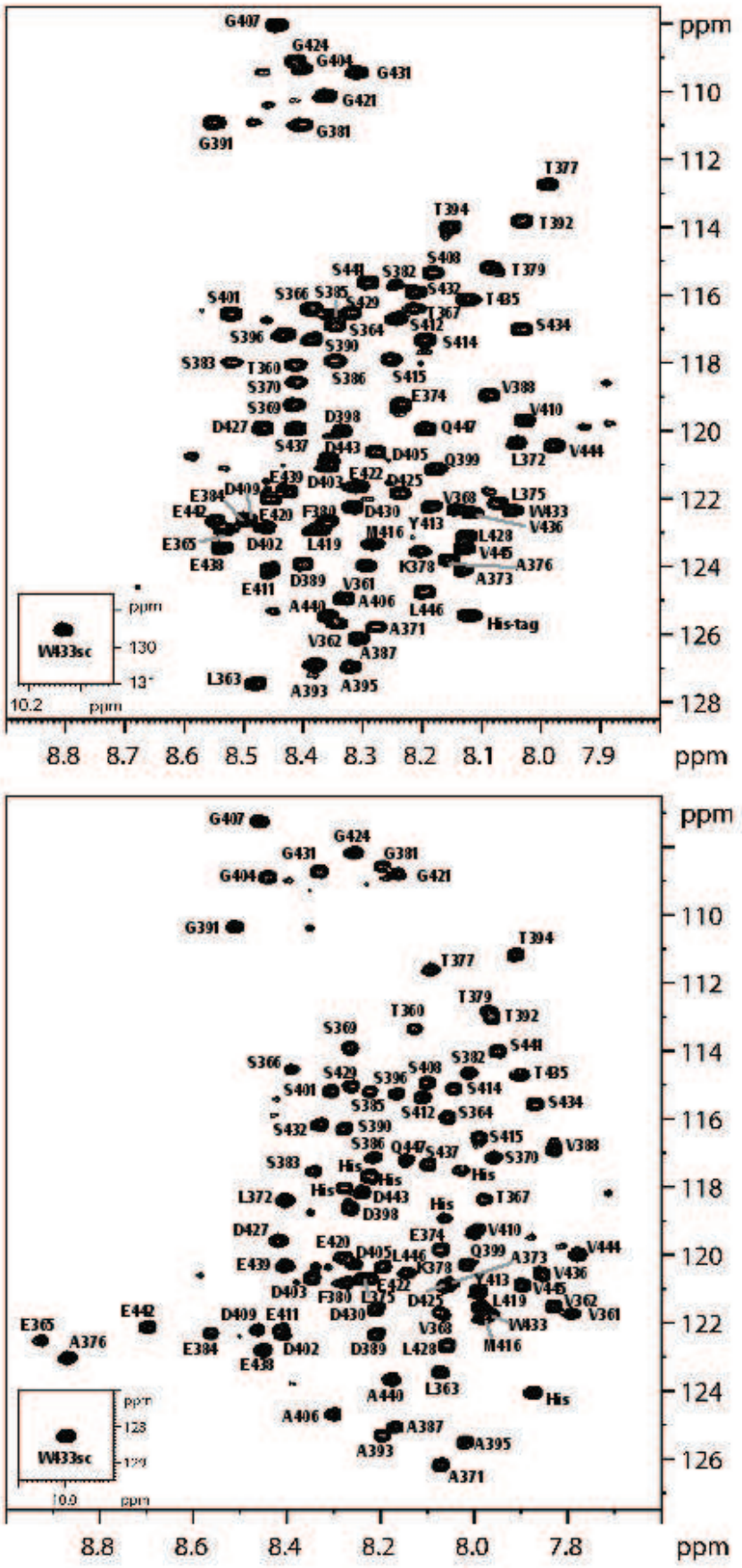


Figure 3.4. Annotated HSQC of NS5A D3 (Con1) in H_2O (top) and in 50% TFE (bottom). Only the labels of major peaks are given. Spectra recorded at 800 MHz.

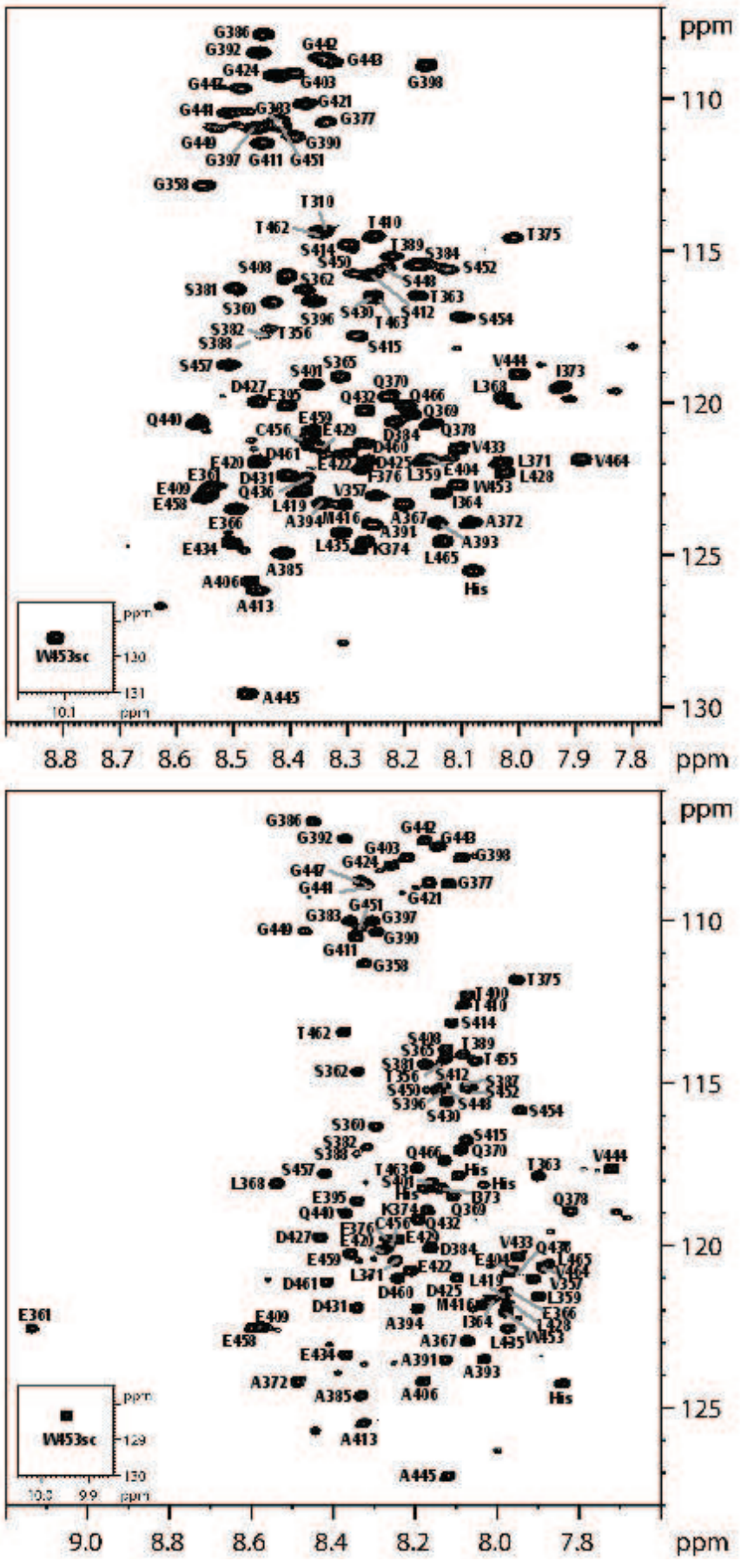


Figure 3.5. Annotated HSQC of NS5A D3 (JFH1) in H_2O (top) and in 50% TFE (bottom). Only the labels of major peaks are given. Spectra recorded at 800 MHz.

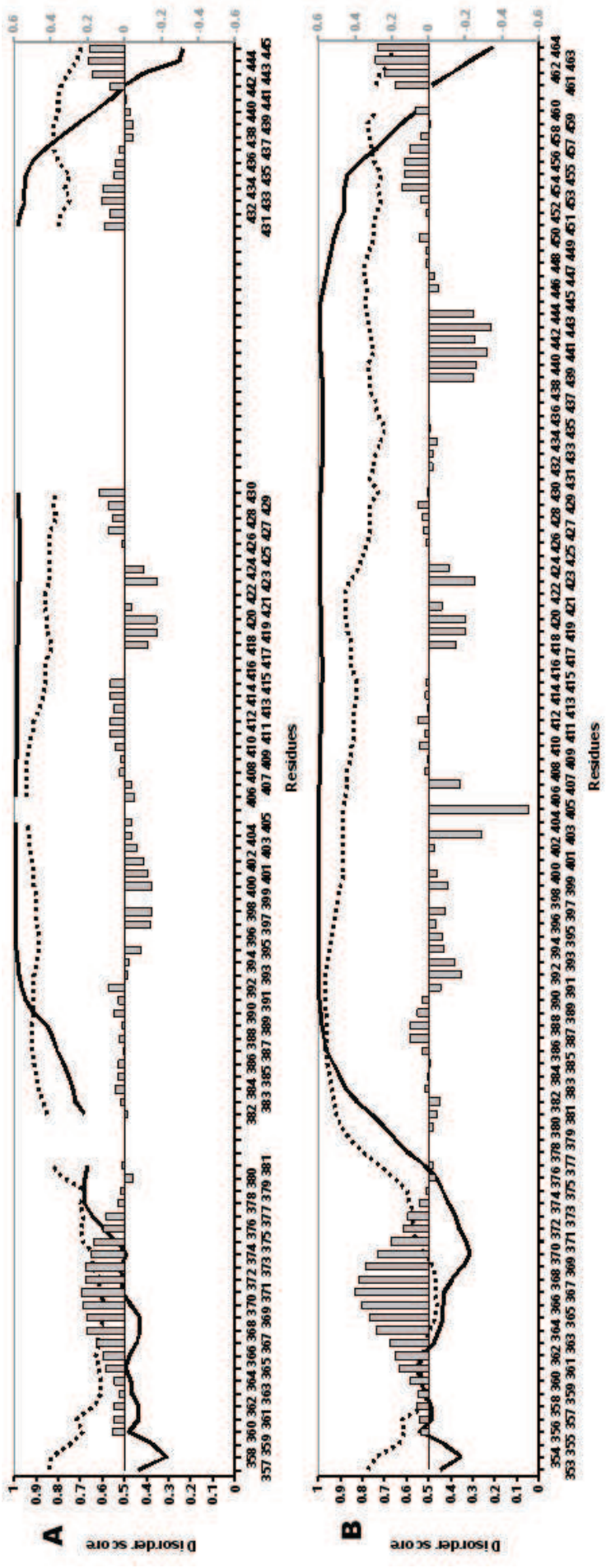


Figure 3.6. Results of the PONDR (bold line), metaPrDOS (dotted line) and SSP (grey columns) analysis for (A) NS5A D3 (Con1) and (B) NS5A D3 (JFH1) (both in aqueous solutions). The axes on the left hand are valid for the sequence predictors. The values of 0.5 separates between disorder (>0.5) and order (<0.5). Right hand side axes refer to the SSP values. Positive values represent α-structure propensity and negative values represent β-structure propensity. Gaps in the sequences are used in order to obtain a maximum amount of congruence between the two sequences (i.e. equivalent amino acid stretches at the same positions).

The informative carbon chemical shifts have also been analysed in the situation where the tendencies towards α-helical structure were reinforced by the presence of TFE. If amino acid regions genuinely have an urging towards helical structure, deviations from random coil chemical shifts are in this case expected to be larger. TFE induced ΔC_α, ΔC_β and $\Delta C'$ values are shown in fig. 3.8 The found back pattern is now even more unambiguous than in the previous analyses. The ΔC_α and $\Delta C'$ values clearly demonstrate the existence of two α-helical structured zones in NS5A-D3 of each of the virus strains. These zones span the regions of roughly residues 363-381 and 440-445 in NS5A-D3-Con1 and 359-377 and 460-464 in NS5A-D3-JFH1.

Additional evidence came from NMR experiments that are actually probing for structure. The HNHA experiment and NOESY-HSQC spectrum of the two D3 domains in a 50% TFE solution were acquired for this purpose. The same samples were used as for recording the assignment spectra, however, these spectra were obtained on a Bruker Avance 600 MHz spectrometer (Bruker Biospin, Karlsruhe, Germany). The NOESY-HSQC spectra were recorded with $\tau_m = 200$ ms. The ${}^3J_{HN,H\alpha}$ coupling constants obtained from the HNHA spectra are given in fig. 3.8. The lower J-values (3-5 Hz) observed in the same zones identified previously to contain secondary structure, confirm the α-helical structuration for the residues in these zones. Finally, the NOE patterns found back for both protein domains in a 50% TFE solution are reproduced in fig. 3.9. These NOE patterns again reflect sampling of α regions of ϕ, ψ conformational space for the previously determined zones of both proteins. In these regions the α-helix typical $d_{\alpha N}$(i,i+3) NOEs and even an occasional $d_{\alpha N}$(i,i+4) NOE appear. Also d_{NN}(i,i+1) and especially d_{NN}(i,i+2) NOEs are indicative for the α-helical structuration.

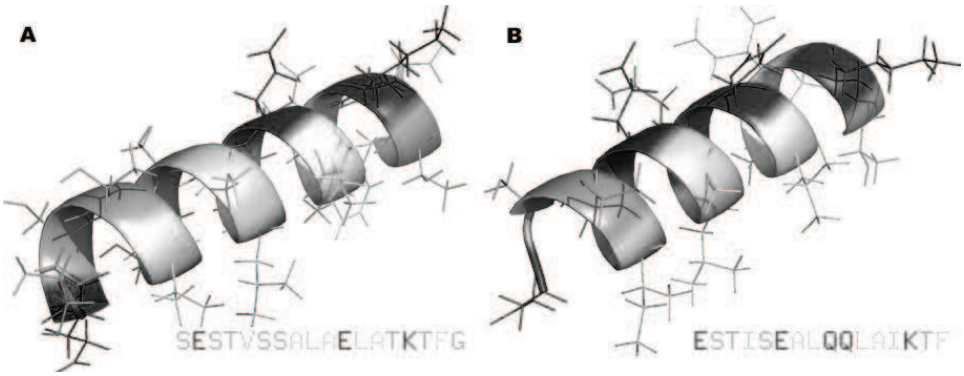


Figure 3.7. Model of the N-terminal sequence stretches ^{364}SESTVSSALAELATKTFG^{381} of NS5A-D3-Con1 (A) and ^{361}ESTISEALQQLAIKT^{375} of NS5A-D3-JFH1 (B) forced in an α-helical structuration. (these stretches being the ones predicted by CSI as adopting an α-helix conformation; see fig. 3.9) The darkness of the residue refers to the hydropathy of the residues. Residues in white are hydrophobic, those in black are hydrophilic, while residues in the two grey scales have properties in between. The helices have a clear amphipathic character.

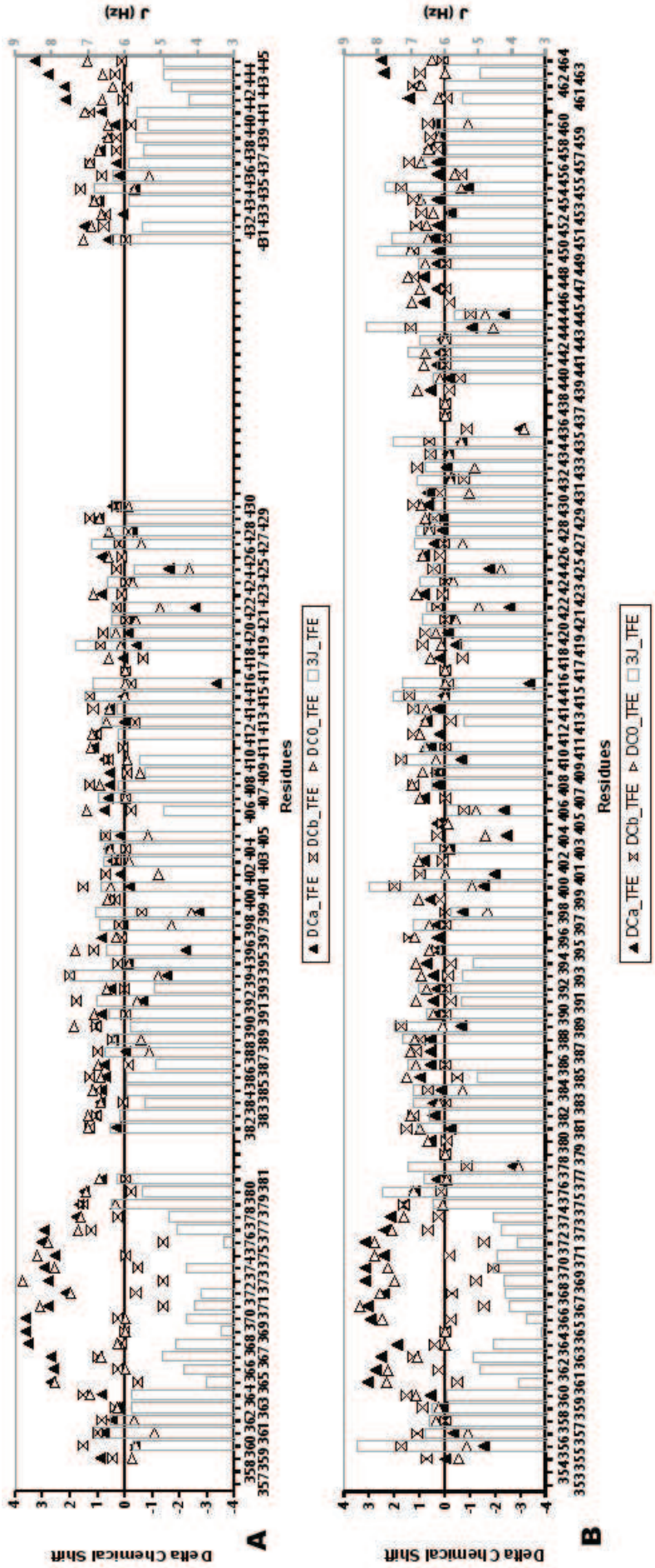


Figure 3.8. NMR parameters obtained on NS5A-D3-Con1 (A) and NS5A-D3-JFH1 (B) in a 50% TFE solution. Values on the left are chemical shift differences between measured and random coil chemical shifts (taken from [433]) in ppm. The C_α-deviation per residue is indicated by the hollow left pointing triangle symbol, the C_β-deviation by the double triangle symbol and the CO-deviation by the filled up pointing triangle symbol. Successive positive secondary C_α and C shifts can be interpreted in terms of populations of α-helical segments. The horizontal bars give the backbone $^3J_{HN.H\alpha}$ coupling constants measured in the HNHA experiment [414] for each residue. The corresponding J-values (in Hz) can be obtained from the scale on the right. Values around 6 Hz are characteristic for random coil behaviour. Larger values are obtain in the case of β-strands, smaller values in the case of α-helices. Gaps in the chemical shift information are explained by the maximum sequence similarity overlap between Con1 and JFH1 versions that we have applied throughout this discussion. Additional gaps in the coupling constant information are caused by uninterpretable HNHA signals because of overlap.

Figure 3.9. The NOE patterns observed from the NOESY-HSQC spectra recorded on NS5A-D3-Con1 (A) and NS5A-D3-JFH1 (B). NOEs exceeding residue 445 and 464 in A and B respectively refer to NOE contacts with residues of the His-tag (LQHHHHHHH) which are not represented here. The bottom line of each figure part represents the CSI (Chemical Shift Index) [433] consensus on the secondary structure of the chain based on the carbon chemical shifts.

Res.	R55	I57	C62	Q63	S77	K82	S99	M100	A103	F113	W121(sc)	L122
K_D (μM)	644	600	527	375	650	518	464	393	539	477	349	625

Table 3.1. CypA/NS5A-D3-JFH1 K_D values obtained from the different CypA residues.

To conclude, the third domains of the NS5A protein of both the Con1 and JFH1 strain of the Hepatitis C virus, appear to contain similar amino acid stretches, one towards the N-terminus of the polypeptide chain and another towards the C-terminus, that tend to populate α regions of ϕ, ψ space. However, the helical regions towards the C-terminus are first of all shorter and are composed for relatively large parts of residues from the His-tag, which probably have an influence on this helix propensity. Hence, the structural character of these few C-terminal residues is not necessarily of biological importance. On the other hand, the well-defined α-helical region in the N-terminal moiety of both proteins could well be a molecular recognition element (α-MoRE) for interaction with other viral or host molecular agents. Whether this is indeed the case and what might be the interaction partner remains to be determined. In fig. 3.7, we present these zones of interest for both the Con1 and JFH1 viral strain forced in a α-helical conformation. These images reveal that the considered region in both strains contains four hydrophobic residues that mostly fall on one side of the helix. These findings are in agreement with the fact the regions could operate as interaction element with a hydrophobic molecular partner.

Similar to what we did previously for the second domain (D2) of NS5A, again since mutations in this D3 domain (although less abundant; there is the V444A mutation in the Con1 strain and another mutation just outside the domain in the flexible linker between D2 and D3 [130]) have been associated with CsA resistance, we have first of all examined the interaction between CypA and NS5A-D3 by means of a titration experiment. For this purpose, unlabelled NS5A-D3 was gradually added to an ^{15}N-labelled CypA sample and HSQC spectra were successively recorded in order to follow the CypA peak behaviour. It was started with 700 μL solution of 100 μM cyclophilin A (50 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 2 mM EDTA, 2 mM DTT, 5% D_2O, pH 6.2 and T = 298 K). Gradually, freeze-dried NS5A-D3 was added to this solution, giving rise to samples of 100 μM CypA and 0, 0.03, 0.1, 0.25, 0.5, 1 and 1.5 mM NS5A-D3. CypA residues that are directly involved in the NS5A-D3 interaction will be characterised by a corresponding peak displacement in the subsequent ^{1}H,^{15}N-HSQC spectra. These shifts in chemical shift were measured using eq. 2.14 and the calculation of the dissociation constant (K_D) of the interaction complex was done by applying eq. 2.18. The cypA residues that undergo signal shifts upon interaction with NS5A-D3 (both the Con1 and JFH1 strains) are R55, I57, C62, Q63, S77, K82, S99, M100, A103, F113, W121(sc), L122. As an example, the shifts observed for the K82 residue of CypA during the titration experiment are shown in fig. 3.10. The shifts observed for Fitting eq. 2.18 to the chemical shift displacements observed for these peaks in the NS5A-D3-JFH1 spectra lead to the K_D values given in table 3.1 and an average K_D of 514$\pm$100.

It proved more difficult to extract this kind of informations from the same experiment performed with cyclophilin A and NS5A-D3 (Con1). Increasing the IUP concentration resulted very quickly (as of 0.5 mM) in a gel-like

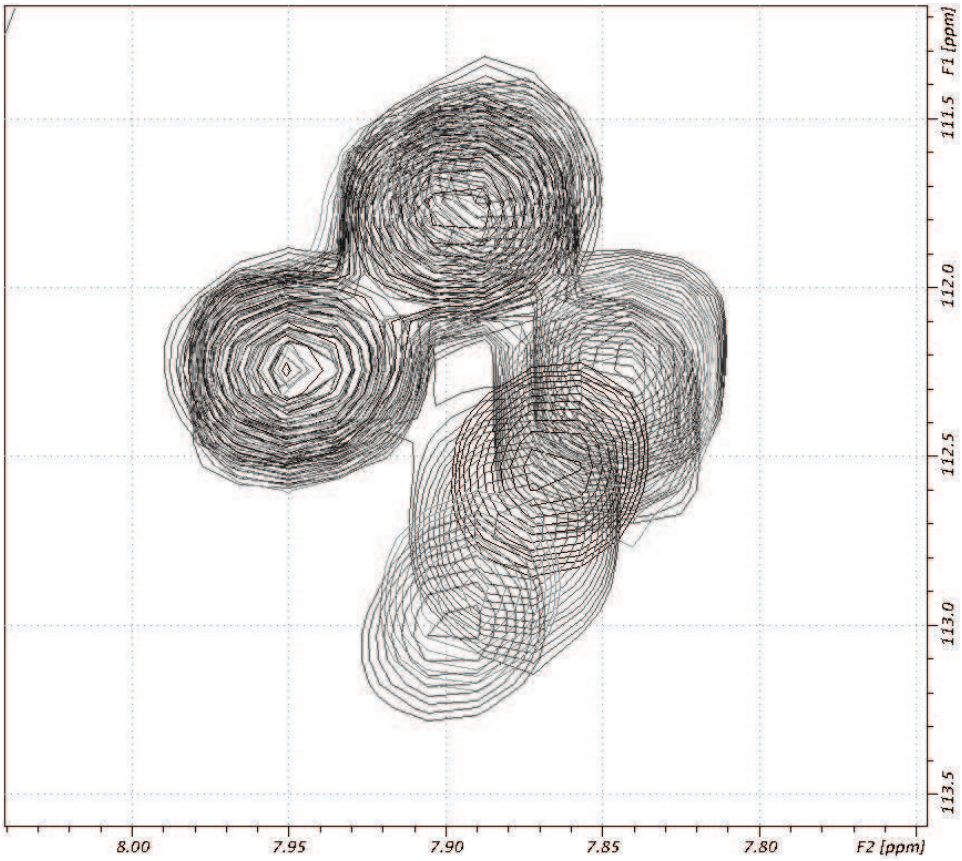


Figure 3.10. Shifts of the cyclophilin A K82 residue signal upon titration with NS5A-D3 (JFH1). The peak shifts downwards with increasing amounts of the interacting IUP. The consecutive peaks correspond to NS5A-D3:CypA ratios of respectively 0, 0.3, 1, 2.5, 5, 10 and 15. The two other peaks at 7.95,112.2 ppm and 7.89,111.7 ppm are the Thr73 and Asn149 HSQC signals resp.

sample, which resulted in the broadening of several of the followed-up peaks. However, the initial points of the several titration curves we were able to obtain, are characterised by a steeper slope than was the case for NS5A-D3 (JFH1) which indicates a smaller (in the range of 2 to 20 times) K_D value. It could hence be concluded that the Con1 strain version of NS5A-D3 form a somewhat tighter complex with cyclophilin than the JFH1 strain version.

Besides the affinity, we have also tried to study the PPIase activity of CypA on both NS5A-D3s using a series of ^{15}N z-exchange experiments [126]. In these experiments, the NS5A-D3 (Con1) sample contained 30 μM non-labelled cyclophilin A and 300 μM ^{15}N-labelled NS5A-D3 in a 600 μL sample (and further: 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 1 mM THP and 0.02% NaN_3 at pH 6.4 and 298K). The NS5A-D3 (JFH1) sample contained 23 μM of non-labelled cyclophilin A and 220 μM ^{15}N-labelled NS5A-D3 in a 435 μL sample (also containing 20 mM Na_2HPO_4/NaH_2PO_4, 30 mM NaCl, 1 mM THP and 0.02% NaN_3 at pH 6.5 and 298.7K). Although the two series of experiments were not recorded with the same protein concentrations, the relative amounts are the same (1:10) (catalytic amounts of cypA). The different mixing times (τ_m) applied were 50, 100, 200, 300 and 400 ms for the NS5A-D3 (Con1) sample and 0.88, 25, 50, 100, 200, 300 and 400 ms for the NS5A-D3 (JFH1) sample. All spectra were recorded at 800 MHz (Bruker Avance).

For NS5A-D3 (Con1) minor cis-prolyl induced signals were assigned for a total of 16 residues (A395, Q399, S401, S414, S415, M416, L419, E420, G421, E422, G424, D425, D427, L428, S429 and D430). However, exchange peaks were identified only for Q399, S401 and L419. These exchange peaks are represented in fig. 3.11, together with the position of these residues in the NS5A-D3 (Con1) sequence.

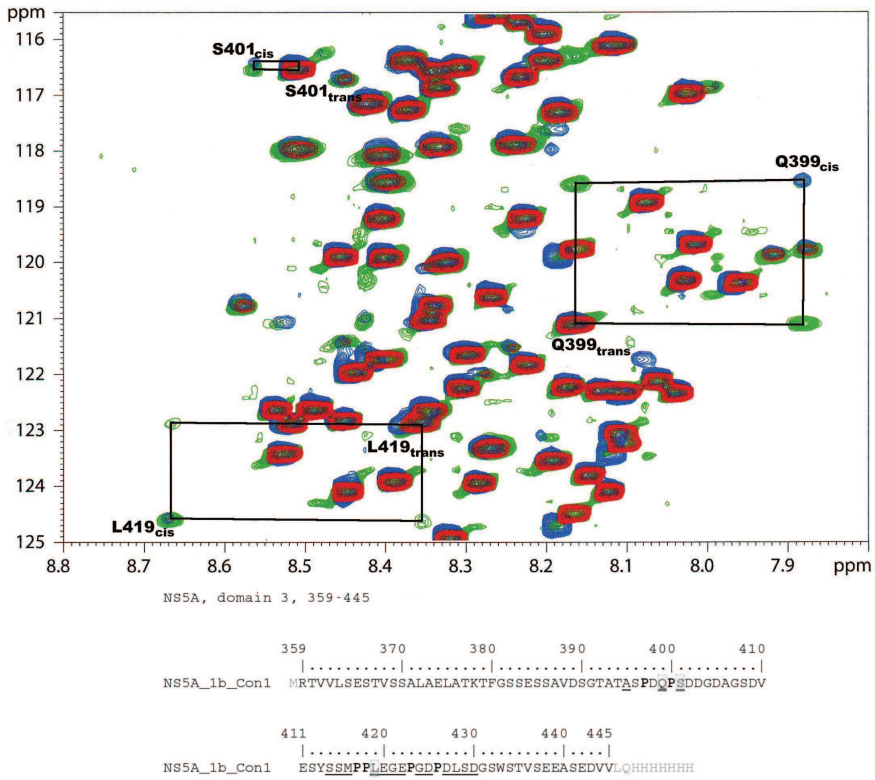


Figure 3.11. Zoom of the $^{1}H,^{15}N$ correlation spectral zone comprising the Q399, S401 and L419 signals. The spectrum in blue is the HSQC of NS5A-D3 (Con1) in absence of cypA. Represented in red is the HSQC of the mixture NS5A-D3 (Con1)/cypA (10:1). Green signals are a superposition of five recorded z-exchange experiments. Cyclophilin A was not labelled, so no cypA peaks appear in these spectra. The bottom part of the image shows the sequential position of the three residues of interest. Residues are underlined if a cis signal for them was assigned.

The L419 residue is neighbour to different prolines than Q399 and S401, which indicates that NS5A-D3 (Con1) interacts with CypA in a distributed manner, i.e. at least two prolyl bonds are substrate to the cypA PPIase activity (see bottom part of fig. 3.11). Of these three residues, only the Q399 system contains four isolated peaks which allow k_{exch} calculations in two different ways, applying eqs. 2.12 and 2.13. We obtained a k_{exch} value for this residue 20 s^{-1} and 23 s^{-1} resp (21 s^{-1} on average). The S401 system suffers from severe peak overlap hence no exchange rate constants were deduced for this residue. The I_{ct} and I_{cc} peaks of L419 are well resolved and using eq. 2.13 we calculated a k_{exch} of 5 s^{-1}.

In the NS5A-D3 (JFH1) HSQC, 26 minor peaks (assigned to residues in the vicinity of prolines undergoing peptidyl-prolyl cis/trans isomerisation) were assigned. These are T375, G377, Q378, T400, S401, E404, A406, E409, S415, M416, L419, E420, G421, E422, G424, D425, D427, L428, E429, S430, E434, L435, Q436, V444, A445 and G447. Isomerisation induced exchange

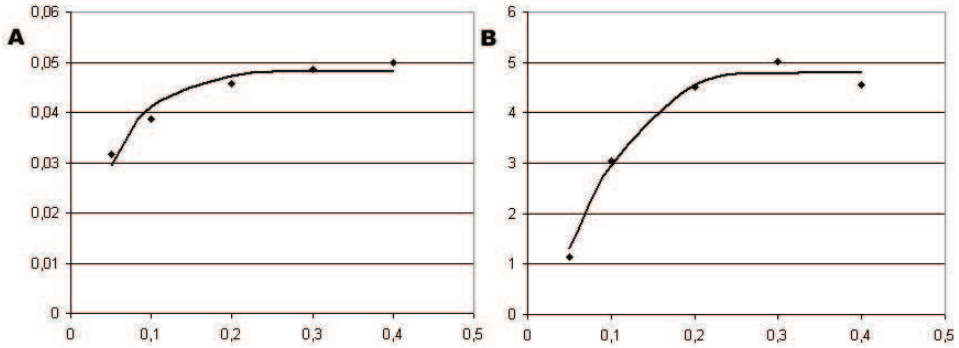


Figure 3.12. The experimental (discrete points) and fitted theoretical (curve) behaviour of (A) I_{tc}/I_{tt} and (B) I_{ct}/I_{cc} observed for residue Q399. The horizontal axes represent the mixing time duration (τ_m) expressed in seconds. The vertical axes are unitless since they embody an intensity ratio. The theoretical curves are obtain using a k_{exch} of 20 s^{-1} and 23 s^{-1} resp.

Res.	Q378	S401	G403	E404	A406 (1)	A406 (2)
k_{exch} (I_{tc}/I_{tt}) (s^{-1})	(-)	16	14	1	2	21 (-)
k_{exch} (I_{ct}/I_{cc}) (s^{-1})	2 (-)	27 (-)	(-)	3	2	(-)
Res.	E409	M416	V444	A445	G447	S448
k_{exch} (I_{tc}/I_{tt}) (s^{-1})	30	(-)	(-)	30	39 (-)	20
k_{exch} (I_{ct}/I_{cc}) (s^{-1})	(-)	(-)	55 (-)	45 (-)	(-)	(-)

Table 3.2. Obtained k_{exch} values for all residues identified to possess exchange peaks, determined in two different ways (eqs. 2.12 and 2.13). If we were unable to fit a theoretical curve to a set of observed peak intensity ratios, (-) is written. A number followed by (-) is given for a lazy fit, i.e. a fitting curve could be calculated but the experimental values differ largely from this curve.

peaks were identified with certainty for Q378, S401, E404, A406, E409, M416, V444, A445 and G447 and possibly also for G403 and S448. A treating of the z-exchange peak intensities provided a k_{exch}-value for some of the residues (see table 3.3). However, due to peak overlap and/or low intensity peaks, no appropriate theoretical fit could be found for the experimental peak intensity progression of others (see table 3.2).

As for NS5A-D2, we observe a non-specific interaction of NS5A-D3 with cyclophilin A. For the Con1 genotype, chemical exchange in the vicinity of 4 out of 6 prolines (P397, P400, 417, 418) is detected when the CypA is added in catalytic amounts. Due to their pairwise sequential proximity, these observations could, however, also be explained by only two proline residues (e.g. P400 and P418) being substrate to the CypA PPIase activity. As indicated in fig. 3.13, 9 out of 14 proline residues of the NS5A-D3 in the JFH1 genotype are possibly substrate to the CypA PPIase activity. Again, this number could be smaller in reality as the observed chemical exchange for some residues can be caused by the peptidyl-prolyl cis/trans isomerisation

Res.	S401	G403	E404	A406 (1)	E409	A445	S448
k_{exch} (s^{-1})	16	14	2	2	30	30	20

Table 3.3. k_{exch} values that could be assigned with more or less certainty.

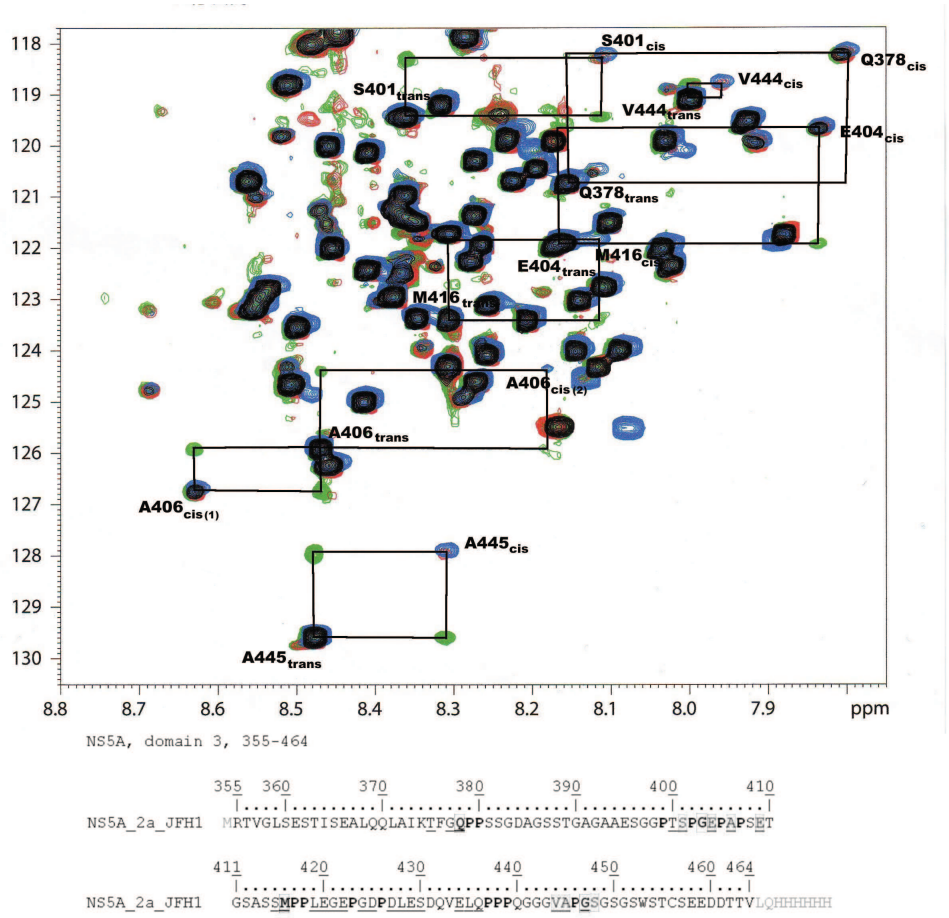


Figure 3.13. Zone of the NS5A-D3 ^{1}H,^{15}N correlation space comprising most of the identified exchange peaks. The black spectrum is the HSQC of the mixture NS5A-D3 (JFH1)/cypA (10:1). The blue spectrum is the HSQC of NS5A-D3 (JHF1) in absence of cypA. Represented in red is the z-exchange spectrum of the mixture recorded with a τ_c of 0.88 ms. The spectrum in green is a superposition of the other recorded z-exchange spectra (τ_c = 25, 50, 100, 200, 300 and 400 ms). Residues underlined in black in the bottom part of the figure represent residues for which a minor cis conformation peak was assigned. For residues represented on a grey background corresponding exchange peaks were identified in the exchange spectra indicating that neighbouring prolines residues might be subject to peptidyl-prolyl cis/trans isomerisation.

of more than one neighbouring proline residue. NS5A-D3 (JFH1) proline residues that seem unaffected by cypA are P423, P426, P437, P438 and P439.

A comparison of the CypA activity on domain 3 of NS5A of the Con1 and JFH1 strains is given in fig. 3.15. This figure indicates that the proline residues unaffected by CypA in the Con1 strain of the Hepatitis C virus are conservedly unaffected in the JFH1 strain. The additional prolines of NS5A-D3 (JFH1) not interacting with CypA (P437, P438 and P439) are unique to this sequence, although other JFH1-unique prolines do seem to interact with the PPIase (P379 and/or P380). Of the prolines contained in well preserved sequence stretches, only the cis/trans isomerisation of the peptide bonds preceding P417 and/or P418 appear to be accelerated by CypA in both the Con1 and JFH1 version of NS5A-D3.

Concerning the activity, the exchange rate values (k_{exch}) of ∼2 to ∼30

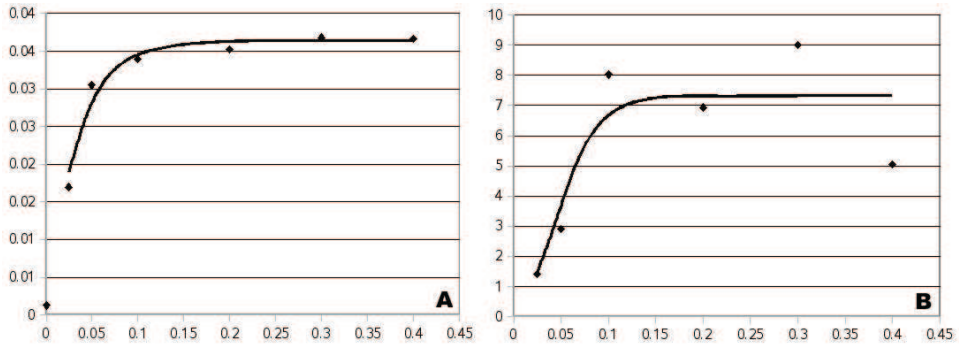


Figure 3.14. The k_{exch} calculation for NS5A-D3 (JFH1) residue A445. (A) represents the calculation based on the I_{tc} and I_{tt} intensities. The good fit of the theoretical curve with the experimental values hints that found back k_{exch} value of 30 s^{-1} must be close to the exact exchange rate. (B) The value of k_{exch} determined based on the I_{ct} and I_{cc} intensities (45 s^{-1}) is less confident. As in fig. 3.12, vertical axes are unitless while horizontal axes are expressed in units of seconds.

```
D3_Con1(1b)   RKRTVVLSESTVSSALAELATKTFG---SSESSAVDSGTATASPDQPSDDGD-AGSDVES 412
D3_JFH1(2a)   RRRTVGLSESTISEALQQLAIKTFGQPPSSGDAGSSTGAGAAESGGPTSPGEPAPSETGS 412
              *:*** *::**:*_.* :** **** ** :. .:*:.:* .. .:. *  *:. *

D3_Con1(1b)   YSSMPPLEGEPGDPDLSD-------------------GSWSTVSEEASEDVVCC 447
D3_JFH1(2a)   ASSMPPLEGEPGDPDLESDQVELQPPPQGGGVAPGSGSGSWSTCSEED-DTTVCC 466
               ***************..                 ***** ***  : .***
```

Figure 3.15. Comparison of residues affected by the peptidyl-prolyl cis/trans isomerisation of some proline residues. The symbols beneath the sequence indicate the extent of sequence invariability expressed as the similarity index according to Clustal W convention (asterisk, invariant; colon, highly similar; dot, similar). [385]

s^{-1} are smaller than what was found for the activity of CypA on NS5A-D2 (14 s^{-1} - 61 s^{-1}) in our previous study [176] (see above).

As was the case for the Cyclophilin interaction with NS5A-D2, we presently cannot conclude whether the actual PPIase activity of CypA or rather the mere binding of NS5A-D3 to CypA is the crucial factor in the HCV replication process.

Chapter 4

Towards Higher Levels of Structuration

4.1. Chemical Shift Predictions for NMR Assignments

Most of the published scientific output can be assigned to the fact that *work grows out of other work, and there are very few eureka moments*[1]. This common truth equally explains the work presented in this chapter.

In the preface of this text, I stressed on the importance of physical simulation methods in structural biology. This not only refers to simulations of the actual structural and dynamical behaviour of molecules in their biological context, but could also include simulations of results produced by biophysical experiments. Although experimental results contain only indirect information on the system under study, trying to predict these results can be useful to validate a proposed model system. In the case of NMR spectra, the emphasis of prediction methods has been on the chemical shift. Until now, these predictions have been performed in three different ways: (*a*) *ab initio*, quantum mechanically based [64, and refs. therein], (*b*) semi-empirically [192, and refs. therein] and (*c*) homology-based (e.g. [272, 346]). The first approach is nowadays only applicable for small molecules. The second can be applied for proteins and uses a set of equations to describe the through bond and through space contributions of the magnetic field at the nucleus. The third approach uses chemical shift databases of previously assigned homologous molecules to predict chemical shifts of new systems.

The different chemical shift prediction algorithms and tools have been developed more out of scientific curiosity than for the immediate practical applications (perhaps apart from the previously mentioned model validation). More interesting must be that the knowledge acquired while developing chemical shift prediction methods has encouraged the reverse search for applications in quantitative structure determination based on chemical shifts [66, 347, 348].

Predicted chemical shifts have (apparently) strangely enough not been extensively used for NMR assignments. Their non-perfect character is to blame for this, as slight prediction errors easily become uninterpretable in the case of equally small chemical shift differences between different spins in the system. This means that moderate to good chemical shift predictions of a structure could indicate that the general influences on the chemical shifts were well understood, but at the same time be useless for NMR assignments of that system. The practical knowledge acquired during the development of the assignment procedure reported in chapter 2, has led very naturally to a predicted-chemical-shift-based assignment tool for protein with known structure. It makes again use of the product- and sum-plane principle and is probably one of most efficient ways to go from non-perfect predictions to accurate assignments. Obviously such a tool is mainly of interest for pro-

[1] quotation from contemporary artist Anish Kapoor

teins with a well-defined, previously elucidated structure. Although NMR chemical shift predictions can in principle also be performed on an ensemble of structures, representing an intrinsically unstructured protein, the small difference between the averaged shift values typical for such IUPs, would probably not lead to many assignments.

The paper describing this functionality has recently been submitted to the Journal of Biomolecular NMR and is included hereafter on the following few pages.

Noname manuscript No.
(will be inserted by the editor)

Towards an automated assignment of proteins of known structure

Dries Verdegem Jean-Michel Wieruszeski Guy Lippens

Received: date / Accepted: date

Abstract NMR spectra of proteins with a known structure should be assigned more rapidly because of the ability to back-calculate some of the protein properties that can guide the assignment process. Here, we report an assignment method based on the matching of experimental and predicted C_α, C_β and N chemical shifts. We propose the use of a 2D C_α/C_β correlation spectrum such as the CC COSY or CC TOCSY to match predicted and experimental peaks. Confronting the carbon shift predictions with the experimental spectra allows a detailed examination of those predictions, thus offering a decent guide during the matching process. The graphical interpretation of Boolean logic functionality for NMR is used to convert identified C_α,C_β correlations in HSQC assignments. Both a semi- and fully automated execution mode are presented. Results on four test proteins demonstrate that the method is capable of assigning very large parts of the sequence in the semi-automatic mode and that many starting points for a more classical assignment can be obtained effortlessly in the automatic mode. Chemical shift predictions as implemented in presently available programs hence can form the basis for routinely assigning proteins with known structure.

1 Introduction

NMR has emerged as one of the methods of choice for the study of protein-protein interactions or general protein-ligand binding. With the protein structure in hand, be it by NMR, X-ray crystallography or homology modelling, the perturbation of the chemical shifts in the ^{1}H,^{15}N HSQC spectrum (Bodenhausen and Ruben, 1980) does not only give the binary answer whether the interaction exists or not, but equally informs about the precise interaction surface, and can give through titration experiments the dissociation constants over a wide range of affinities. Although this is generally recognized in the wider (structural) biology community at present, extending definitely beyond the NMR community, one problem remains the assignment of the HSQC spectrum. The recording of the NMR spectra thereby can more and more easily be outsourced, be it to a structural genomics center or some central NMR facility where the standard battery of 3D spectra are run at high field in a routine mode. However, the assignment still asks for manual intervention and thus expertise, which forms a hurdle for many biologists that would like to use NMR as an alternative and even more powerful method than surface Plasmon resonance or fluorescence spectroscopy. For these non-NMR specialists, it would seem obvious that the assignment should not require as much effort when structural information on the protein is available. Such is the case, and methods have been reported that rely on a matching between measured and back-calculated RDCs or on NOE-connectivity information to obtain a structure guided assignment (Al-Hashimi and Patel, 2002; Al-Hashimi et al, 2002; Apaydın et al, 2008; Bailey-Kellogg et al, 2000; Bartels et al, 1996; Delaglio et al, 2000; Dobson et al, 1984; Erdmann and Rule, 2002; Hus et al, 2002; Jung and Zweckstetter, 2004b,a; Langmead and Donald, 2003, 2004; Langmead et al, 2004; Pristovšek et al, 2002; Pristovšek and Franzoni, 2006; Stratmann et al, 2009; Tian et al, 2001; Xiong and Bailey-Kellogg, 2007; Xiong et al, 2008; Zweckstetter and Bax, 2001). However, because individual proteins will behave differently when using a given alignment medium, these methods require experimental testing of e.g. the absence of physical interaction between the

Dries Verdegem Jean-Michel Wieruszeski Guy Lippens
Unité de Glycobiologie Structurale et Fonctionelle, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France.
Tel.: +330320337241
Fax.: +330320436555
E-mail: guy.lippens@univ-lille1.fr

protein and the orienting medium (bicelles, phages, ...), and thereby again place an additional constraint on the user or NMR center. An assignment strategy based on chemical shift predictions alone therefore seems the best option to go from the known 3D structure and the standard set of 3D spectra to the automatic assignment of the HSQC. Although some efforts have been made to expedite the assignment process using residue type characteristic chemical shifts (Atreya et al, 2000, 2002; Grzesiek and Bax, 1993; Marin et al, 2004; Verdegem et al, 2008; Xu et al, 2006), to our knowledge only one study was based on pure per-residue predicted-shifts-based assignments (Gronwald et al, 1998). By matching a set of predicted ^{15}N and backbone and side-chain ^{1}H shifts to a set of corresponding observed peaks, this study attempted the residue-specific assignments. However, several of its algorithm design decisions thwarted an efficient applicability. A tough peak picking phase was required and the predictions, for which it relies on prediction method ORB (Gronwald et al, 1997), asked for an extensive set of assigned homologous proteins.

In our previous paper on protein ^{1}H,^{15}N HSQC assignments (Verdegem et al, 2008) describing the use of product and sum planes —aptly outlined as originating from the application of Boolean operators— we introduced post-MUSIC as a computational utility capable of a fast recovery of the amino acid type and even the sequential position of HSQC signals of unstructured proteins. Indeed, for those unfolded proteins, all C_α and C_β chemical shifts are known to resonate very close to the typical random coil chemical shifts. Thus, defining small windows around the amino acid typical C_α and C_β random coil chemical shifts, summing all HNCACB (Wittekind and Mueller, 1993) ^{1}H,^{15}N spectral planes falling in such a window and finally multiplying the resulting C_α and C_β sum planes provides one with a ^{1}H,^{15}N product plane containing intensity only at the positions of that amino acid type in the HSQC. The additional use of the HN(CO)CACB spectrum (Grzesiek and Bax, 1992) furthermore allows for the unambiguous assignment of the second residue in unique residue pairs in the sequence. The main advantage of this technique is that assignments can be made within just a few seconds following the transformation of the 3D spectra, since no peak picking is required.

For structured proteins, the situation is slightly different. The C_α and C_β chemical shifts tend to be spread much wider because of the structural elements that force certain bond angles and influence the local magnetic environment. Applying post-MUSIC with random coil chemical shifts therefore would ask for much wider window sizes that inevitably become totally unselective. If the structure of the protein under study is known, the output of a chemical shift predictor (Arun and Langmead, 2006; Gronwald et al, 1997; Luman et al, 2001; Meiler, 2003; Neal et al, 2003; Shen and Bax, 2007; Wishart et al, 1997; Xu and Case, 2001) could fulfill the role of window center to turn post-MUSIC into a tool for the non-sequential assignment of the NMR spectra. Unfortunately, the quality of the predictions limits the use of such an algorithm. If the C_α and C_β chemical shift predictions of a given residue are far off from the experimental values, using them as the center of small windows leads to no or, even worse, wrong predictions. As there is no prior information on the quality of the predictions, such mistakes would go by completely unnoticed and ultimately the confidence in the HSQC assignments made this way would suffer.

Increasing the confidence in the predicted values is the central theme of this paper. For this, we confront the predicted values with the experimental C_α and C_β correlation peaks as recorded in a carbon detected CC COSY or TOCSY spectrum. A 2-dimensional comparison of both the experimental and predicted C_α and C_β chemical shifts not only allows for an immediate inspection of the quality of the predictions, but also allows to distinguish those that are reliable from those that seem distant of any experimental peak and thereby can be eliminated from the further procedure. We demonstrate the newly developed post-MUSIC upgrade on four structured proteins that have both been studied by X-ray crystallography and NMR, and discuss its implementation as a way to go to structure-based assignment of the HSQC spectrum.

2 Theory and Methods

Before addressing the automated version, we want to illustrate the method on a semi-automated version. This proves the usefulness of the principle since it allows for the assignment of large parts of protein sequences. These numerous non-sequential assignments can be used as punctual markers for interaction mapping studies right after the transformation of the 3D spectra. Assignments in the automated mode, that we discuss in a second phase, result in less annotated HSQC peaks, but they are obtained instantaneously. This automated mode can thus be used to generate rapidly and without ambiguity a number of starting points for the more traditional sequential assignment procedure, based on a walk through the spectra to connect carbon shifts of neighbouring residues.

2.1 Mode 1: Semi-automatic: The non-sequential assignment

Whereas biological NMR traditionally has relied on pulse sequences whereby both the excitation and observation nucleus was the abundant and sensitive proton, recent progress in the development of increasingly high magnetic fields and the introduction of novel cryogenically cooled probeheads (Kovacs et al, 2005) has extended the applicability to carbon detected experiments (Bermel et al, 2006). The 2D di-

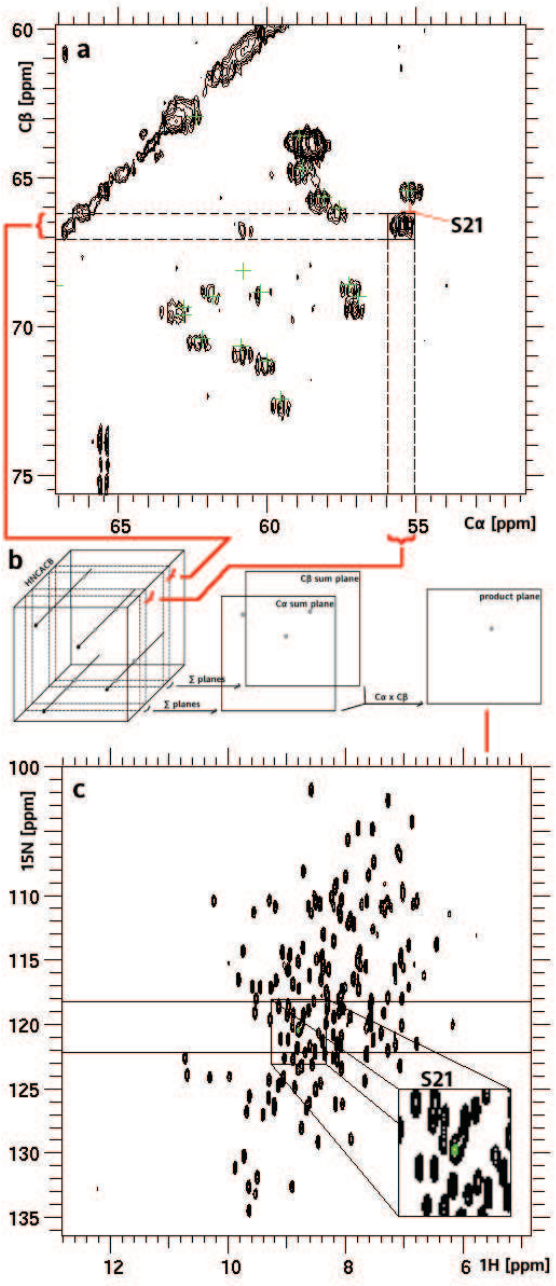


Fig. 1 The assignment principle in Mode 1. In a zoom of the CypA CC TOCSY comprising the serine and threonine residue signals (**a**), a visual inspection of the relative positions of real peaks and predicted peaks (green crosses) allows to define two serine 21 selective carbon chemical shift windows by drawing a rectangular box around the signal. In **b**, a simplified version of the CypA HNCACB spectrum joined with the HSQC spectrum in front can be imagined. Summing all ^{1}H,^{15}N spectral planes in both of the defined windows will yield sum planes with a selectively increased intensity at the HSQC coordinates of Ser21. Even more selectivity is obtained by multiplying the two sum planes to one product plane, that, when finally presented on top of the HSQC, enables an easy assignment (**c**). The predicted N chemical shift of Ser21 (horizontal lines are drawn at $\pm$2 ppm this value) confirms the assignment proposed by the product plane.

rectly carbon detected CC TOCSY (Eletsky et al, 2003) offers a clarifying 2D view of all C_α and C_β chemical shift correlations in the protein under study, and consequently plays a pivotal role in the setup of our method. With a proton, carbon dually cooled cryoprobe installed on a 700MHz Bruker AvanceIII spectrometer, the experiment took a mere 13 hours on a 340μM CypA sample, and can therefore be added to the traditional series of 3D spectra without an excessive increase in time.

In the first step of the assignment method (depicted in fig. 1 for the concrete example of residue Ser21 of Cyclophilin A), the chemical shift predictions of the SPARTA (Shen and Bax, 2007) program is graphically superposed on the CC TOCSY as crosses at the calculated C_α and C_β combinations. Such superposition immediately allows to spot potentially useful predictions as those exactly coinciding with a peak in the spectrum and/or those occurring in more sparsely populated regions. Secondly, one defines the smallest possible zone in the C_α, C_β domain that should, according to the visual inspection, contain the carbon chemical shifts of the residue of interest. This zone defining step, which is done by drawing a box around it, is the only manual step in the procedure, and is therefore equally the step that seemingly requires most experience of the user (see below). The dimensions of the box define two windows that are applied on the HNCACB spectrum in a post-MUSIC-like manner. I.e., all ^{1}H,^{15}N HNCACB spectral planes falling within the window borders are summed up point-by-point, resulting in a C_α and C_β sum plane. The point-by-point multiplication of these two planes finally results in product plane that is superposed on the HSQC. Inherent to its definition, the product plane contains, in the ideal case, one single peak, that corresponds to the HSQC peak position of the residue if the C_α , C_β peak indeed was the correct one. To validate the assignment, we found it useful to consider also the nitrogen chemical shift prediction. An additional N-window with the predicted value as centre is in that case drawn as two extra lines on the HSQC.

Although this method works well for truly isolated peaks, the assignment method becomes more robust if one simultaneously exploits the HN(CO)CACB spectrum, because this leads to the identification of a pair of residues rather than one isolated one. The same sum-and-product plane derived from the HN(CO)CACB spectrum around the experimental peak closest to the predicted value of the (i-1) residue should indeed give the same peak as identified on the basis of i residue and the HNCACB spectrum. This not only increases the confidence in the original assignment, but also permits to enlarge the window widths to some extent in the case of doubt because of signal overlap, while still preserving enough selectivity for an unambiguous assignment.

A third option is to use only the HN(CO)CACB spectrum. This is advantageous for example for the assignment

of glycine residues. Since glycines lack C_β atoms, they are present in the CC TOCSY only as diagonal resonance peaks around 45ppm and thus are often very little dispersed. As a consequence, it can be hard to decide on a selective C_α window. When considering the full Glycine C_α window, the resulting Glycine sum plane generally lacks selectivity. Better results can then be expected from the sole use of the i-1 plane.

As before (Verdegem et al, 2008), all calculated spectra, predestined for presentation on the screen are first normalized and then multiplied by a constant factor to obtain natural intensities. The proposed method in this mode also again adopts the process-and-run principle, linking the selection of a given peak directly to the extraction and manipulations of the corresponding planes in the 3D spectra. Thereby, the assignment effort can be limited to certain residues of interest, which strongly enhances its speed.

2.2 Mode 2: Automation - Divide and Conquer: Starting point generation

For a full assignment of the HSQC spectrum, the sequential assignment walk in its numerical or graphical version could certainly benefit from a robust starting point generating step. The fact that some predicted TOCSY peaks can be unambiguously matched to the corresponding real peak raises the possibility of an automation of this matching process, providing a starting point generating step in a very fast manner. Obviously, this does require a peak-picking step of the C_α, C_β peaks, which can be done in the CC TOCSY spectrum, but equally from an automated peak-picking of the HNCACB 3D spectrum in the carbon dimension.

For the actual matching of real and predicted peaks, we developed a divide-and-conquer algorithm. The problem of matching experimental data to information of a structural model is usually solved as a bipartite matching problem of a weighted bipartite graph. Such a graph is formed by two sets of vertices and a value associated to each of the edges (that connect two vertices, one from each set). In this case, one set would be formed by the real C_α,C_β peaks (R), the other by the predicted C_α,C_β peaks (P). The edge values indicate how well the two connected vertices would match. The challenge of a maximum weighted bipartite matching then implies the maximising of the sum of a subset of edge weights so that the corresponding edges connect every vertex in the graph just once. The obtained subset of edges yields the optimal set of links between predicted and real peaks. The divide-and-conquer algorithm is based on this principle.

The basic idea of the divide part is to create subproblems (S_i) in the automated matching procedure. If a group of real and predicted C_α,C_β peaks forms an isolated "island" in the 2D carbon carbon space, optimal matches can be searched among them, without interference of other, more distant peaks. In order to accomplish this functionality, the user is asked for a critical distance (in ppm), D_c. If two peaks, real or predicted, lie separated from each other at a distance greater than D_c, they belong to a different subproblem S_i, unless there exists one or more other peak(s) through which both former peaks can be connected by dis-

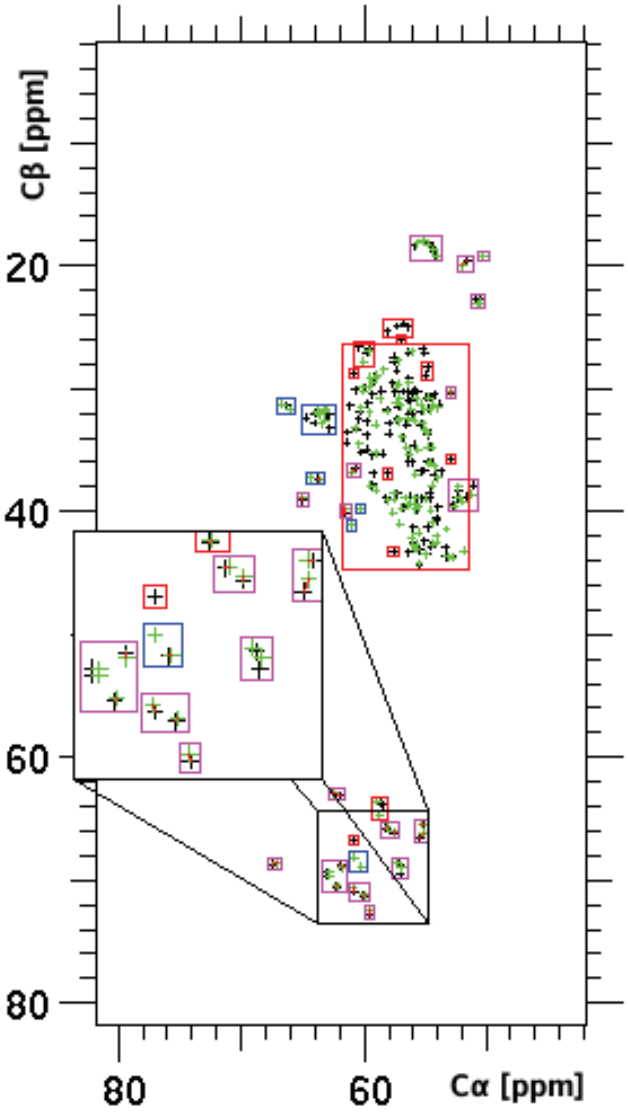


Fig. 2 The divide-and-conquer algorithm applied to cyclophilin A. The black crosses represent the peak-picked real CC TOCSY peaks. The green cross positions indicate the per-residue C_α and C_β chemical shift predictions. The different boxes drawn on screen are the generated subproblems S_i after defining the critical distance D_c as 1.1ppm. The colour code is: pink for a subproblem with equal amounts of real and predicted peaks; blue and red indicate a surplus of predicted peaks and real peaks respectively. The red lines visible between some real and predicted peaks in certain subproblems indicate the matches the conquer part of the algorithm has made.

tances smaller than D_c. Distances between two peaks 1 and 2 are calculated as Euclidean distances:

$$d(1\ 2) = \sqrt{(C_{\alpha\,1} - C_{\alpha\,2})^2 + (C_{\beta\,1} - C_{\beta\,2})^2} \quad (1)$$

The actual matching functionality is implemented in the subsequent conquer part. Two adjustments to the matching principle described above are needed for it to elegantly address the present assignment problem. First, we redefine the problem as to try to minimise a penalty that corresponds to every real peak - predicted peak link distribution in each of the subproblems S_i. It follows that weight attached to each edge should decrease as the two vertices it connects (a real peak and a predicted peak) form a better match. This allows us to define the edge weight simply as the Euclidean distance between the two vertices concerned (eq. 1). Consequently, the penalty for a subproblem is calculated as an index of dissimilarity given by

$$D_x(P^{S_i}\ R^{S_i}) = \sum_{j=1}^{n} d(p_{j\,x}\ r_{j\,x}) \quad (2)$$

where the sum is taken over every linked predicted peak - real peak pair (j) of a certain peak pairing distribution (x) in the subproblem S_i. The number of pairs, n, equals $min(N_p^{S_i}\ N_r^{S_i})$, being the smallest of the number of predicted peaks and the number of real peaks of the subproblem. The penalty D_x is thus given by the mere summation of distances in ppm between every predicted peak and the real peak to which it is matched. Distributions (x) that minimise eq. 2 more than others are more likely to hold the correct match. However, due to the imperfect nature of the chemical shift predictions, a pure minimum weighted bipartite matching algorithm cannot be used. This would come down to the assumption that the closest real C_α,C_β peak to any predicted C_α,C_β peak is the correct matching partner for that predicted peak. In real situations, with non-zero prediction errors, that property never holds. It all brings us to the second algorithm adjustment, described in the following three conquer rules:

1. As the correct matches might correspond to a D_x slightly higher than its smallest possible value D_{min} in cases where prediction errors and inter peak distances have the same order of magnitude, a range of distributions x has to be considered. We define a distribution as interesting if $D_{min} \quad D_x \quad 2 \times D_{min}$ holds and decisively only the matches occurring in every one of these interesting combinations are accepted as correct. The existence of this rule entails the necessity of knowing D_x for every x. In order to explore the entire dissimilarity index space, an exhaustive search is performed. The number of possible combinations (C) between $N_p^{S_i}$ predicted peaks and $N_r^{S_i}$ real peaks equals:
$$C = \frac{\left(max\left(N_p^{S_i}\ N_r^{S_i}\right)\right)!}{|N_p^{S_i} - N_r^{S_i}|!} \quad (3)$$
2. Analogous to D_{min}, we define D_{max} as the maximum value of D_x for a certain subproblem S_i. Then, in the case of $max(N_p^{S_i}\ N_r^{S_i}) > 1$, if $D_{max}/min(N_p^{S_i}\ N_r^{S_i}) < D_{msd}$, the corresponding subproblem S_i is not treated. Here, D_{msd} is the median of the shortest distances array, an array containing the distances between every predicted C_α,C_β peak and its closest-by real C_α,C_β peak of the entire spectrum. This rule was created to prevent that, in a situation where several predicted and real peaks lie very close together (closer than the typical error in the chemical shift prediction), wrong assignments would be made by applying rule 1. The median is chosen instead of the mean value since the shortest distances distribution can be quite heavy-tailed due to missing or accidentally coinciding real peaks.
3. Finally, it is fair to state that in subproblems with large amounts of peaks, the mesh of peaks is getting more complex and chances that rule 1 succeeds in making many assignments are reduced drastically. For this reason, subproblems for which $max(N_p^{S_i}\ N_r^{S_i}) > 10$ are not considered in the conquer part of our method. This also implies that the algorithm, that scales as $O(n!)$ for each subproblem (see eq. 3) and therefore is NP-hard, is prevented from evoking runs of more than a few minutes time.

The output of the algorithm for cyclophilin A is presented in fig. 2. Once a set predicted peak - real peak links has been automatically created, the graphical plane manipulation functionality is again summoned to generate the residue indicating ^{1}H,^{15}N product planes. Since the exact C_α and C_β chemical shift of the desired residue are in this case exactly pinpointed, the use of C_α and C_β windows with a specific non-zero width as employed in the previous section (mode 1) is now unnecessary. The final product plane is obtained by a multiplication of two planes directly retracted from the HNCACB at the chemical shift coordinates of the real CC peak and is again, possibly in combination with the predicted N chemical shift information as two horizontal lines, superimposed on the HSQC.

The divide-and-conquer algorithm is defensive in its nature, in that it only tries to make those matches that may be held for almost certainly correct. As a consequence, we cannot intend entire assignments in the automated mode. It should rather be seen as a method capable of assigning any number from a few to dozens of HSQC peaks (depending on the quality of the predictions and the crowdedness of the spectrum) almost instantly, which can be used as starting points to speed up more conventional assignment tools or as initial assignments to start an interaction study.

3 Results

We have tested the semi-automatic version of our method (mode 1) on the structured protein Cyclophilin A (CypA) and the automated version (mode 2) on CypA and three other proteins ranging in length from 76 to 238 residues (see table 1). The best set of predicted C_α,C_β peaks were generated by SPARTA (Shen and Bax, 2007) for all four proteins and we have thus consistently used it to obtain all results described hereafter. In order to test the divide-and-conquer algorithm, no actual NMR spectra are required as only the correctness of the calculated matches between real and predicted chemical shifts, both of which can, either directly or indirectly, be found in the literature, needs to be verified. We therefore present experimental data only for the Cyclophilin A protein. The HSQC, HNCACB and HN(CO)CACB spectra were acquired on a Bruker Avance 600 MHz equipped with a proton only cryogenic triple resonance probe head using standard Bruker pulse programs. The CC TOCSY was acquired on 700 MHz Bruker Avance spectrometer with a proton/carbon cryogenic triple resonance probe. The acquisition parameters were: 1024 (1H) and 256 (15N) complex points at 16 scans per increment for the HSQC, 1024 (1H), 104 (15N) and 192 (13C) complex points and 8 scans per increment for the HNCACB and HN(CO)CACB and finally 2048 direct and 304 indirect complex points and 128 scans per increment for the CC TOCSY. The CypA sample (with a total volume of 0.7 mL) contained 340 μM of the protein in an aqueous buffer with 50 mM Na_2HPO_4/NaH_2PO_4, 40 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 6.3 and was studied at 298 K.

Two examples of the semi-automatic assignment module are shown in Figure 3. All predicted C_α/C_β correlations are drawn on the experimental CC spectrum as green crosses, with the one under study in blue and its (i-1) neighbour in red. For the L122, a zoom of the spectrum shows that the predicted Leu122 correlation almost coincides with a unique peak at 54.07, 39.65 ppm. Manual drawing of the blue box thereby defines the limits of the sum planes in both the C_α and C_β dimension of the HNCACB spectrum. Although this information on itself is already sufficient to generate a product plane corresponding to this unique peak, we equally predict the C_α,C_β values of N121, its upstream neighbour. Although this peak again coincides with a rather isolated zone of the spectrum, it is not clear whether one or two experimental peaks are present in this zone. Drawing the red box around the complete zone avoids the need to decide, even though it leads to summation of some more planes in both dimensions. As the resulting product of product planes indicates a single peak, that moreover falls within the zone of possible nitrogen chemical shifts, the user can assign without ambiguity the experimental peak at 7.04, 120.08 ppm to L122.

The second panel of Figure 3 shows an at first sight more complicated situation. The predicted values for F113 lead no longer to an isolated cross peak, but rather to a blue cross in a crowded region, and things are even worse for F112. Boxes were drawn considering the zone of closest peaks that are not heavily occupied by other predicted peaks (as green crosses), but still letting enough margin not to exclude true correlations. For the L111 peak, this leads to a rather large (red) box with a width of over 1.5 ppm in both the C_α and C_β dimension, because no obvious other candidates for the peaks in the lower left and upper right corner could be detected. The product of product planes again yields a single peak in the ^{1}H,^{15}N HSQC spectrum, that falls within the zone of predicted nitrogen values.

Subjecting the cyclophilin A spectra to this procedure allowed to assign 122 peaks on a total of 158 assignable HSQC peaks (residue 1 and the six prolines have no corresponding resonance peaks in the HSQC). For this, mostly both the HNCACB and HN(CO)CACB spectrum were employed, except in the case of the glycines where only the HN(CO)CACB spectrum was used for reasons explained earlier. Obviously, by analogy, residues following a glycine were assigned with just the HNCACB. Of all 122 correctly assigned HSQC peaks, 94 were indicated by a final product plane containing only a single peak. For the remaining 28 peaks, two or sometimes more product plane peaks appeared (at reasonable contour levels) but only one of them fell in the zone determined by the predicted N chemical shift $\pm$ 2 ppm and thus nevertheless was accepted as the correct assignment. Importantly, not once did a product plane containing one or a few peaks in combination with the N-shift prediction zone indicate a wrong assignment. The residues the method was unable assign are V2, F8, D9, V20, E23, E34, S40, G42, K44, G45, F46, K49, G50, H54, Q63, G65, D66, F67, T68, H70, N71, G72, G75, K76, E81, E86, G94, G109, I114, G124, K125, V128, M136, I138, Q163, E165. For the first and last residue SPARTA gave no predictions. Furthermore, this list contains 9 glycine residues, for which the lack of C_β carbon reduces the selectivity of the product plane, but also 6 further residues that have a glycine in the (i-1) position. Pairwise information indeed is essential for the method to work reliably. For S40, C_α and C_β are predicted to resonate at almost the same value of 62.5 ppm, and such is also the case experimentally. The sum planes thereby almost cancel one another, leading to a product plane barely above noise level. For all other residues, their own C_α,C_β or their i-1 neighbour values were badly predicted, in the sense that no obvious box could be traced around a reasonable zone. The consequence of the in this case unavoidable necessity to apply large window widths is that product planes are obtained with meaningless information, leading to the appearance of no or more than one peak in the predicted N zone.

		NMR chemical shift info		*X-ray crystallography info*		
Protein	Length (res)	BMRB Entry	NMR ref.	X-ray PDB ID	Resolution (Å)	X-ray ref.
Human Ubiquitin	76	15410	(Jaravine et al, 2008)	1ubq	1.80	(Vijay-kumar et al, 1987)
Alpha-domain of ATPase	94	5107	(Smith et al, 2001)	1qzm	1.90	(Botos et al, 2004)
Cyclophilin A	165	/	(Ottiger et al, 1997)	2cpl	1.63	(Ke, 1992)
Alpha-Adaptin appendage domain	238	6034	(Denisov et al, 2004)	1b9k	1.90	(Owen et al, 1999)

Table 1 The protein test set. For all proteins the experimental NMR backbone assignments were recovered from the BMRB (Ulrich et al, 2008), except for cyclophilin A, for which the chemical shift list was obtained directly from the authors. The PBD (Berman et al, 2000) delivered the X-ray structures.

As a curiosity, we might note that we started the assignment with the residues described in (Ke et al, 1993) as those forming the active site of the cypA protein; R55, F60, M61, Q63, A101, N102, F113, L122, H126. All except Q63 were easily assignable within minutes after the processing of the NMR spectra, and could therefore be used for a definition of the interaction zone with CsA or a chemically related compound.

Where for assignments in mode 1, less perfect C_α and C_β predictions can still result in unambiguous assignments because of the selective power of the product plane principle, a different truth holds for the automated assignments in mode 2. As the divide-and-conquer algorithm tries to match every predicted peak to a unique real peak, the number of proposed correct matches will largely depend on the prediction quality. Obviously, the determining factor will be the quality of the chemical shift predictions, but other factors such as the crowdedness of certain spectral zones, related to the protein size, structure and amino acid composition, will all play an important role as well.

For Cyclophilin A, a critical distance (D_c) of 1.1 ppm was used, resulting in the divide pattern represented in fig. 2. This pattern and the subsequent conquer algorithm as outlined above led to an output of twenty matches, all of them correct. They allowed the immediate assignment of the residues T5, E15, S21, T32, F36, S40, S51, T68, T73, I78, I89, S99, A101, A103, T107, S110, S147, T152, T157 and I158. Not surprisingly, Serine and Threonine residues are overrepresented in the successful automatic assignment, as their C_α/C_β correlations are in a rather isolated zone. The automated assignment of the alpha-domain of ATPase and the 'appendage' domain of the alpha subunit of the endocytotic AP2 adaptor complex yielded only six and five automated assignments, again all correct. The former protein, although only half the size of Cyclophilin A, contains only α-helices, leading to a more crowded spectrum. The limited success of the second automatic prediction is mainly due to the increased size of the appendage domain, leading to lesser isolated zones that could be identified by the divide part of the algorithm. Moreover, for both proteins, SPARTA clearly performed less accurately compared to CypA. Critical distances (D_c) of 1.3 and 1 ppm resp. were used. These critical distance values are easily decided upon after a quick inspection of the boxy output of the divide part of the algorithm and are mainly dependent on the mean C_α,C_β prediction error.

As for the human ubiquitin protein, a D_c of 1 ppm resulted in 21 matches. Due to some unfortunate peak positions versus peak predictions, two of these were wrong. The correct matches are those of V17, E18, P19, S20, T22, V26, A28, K33, P38, F45, A46, T55, S57, N60, I61, Q62, K63, E64 and S65. The C_α,C_β signal of T7 and L43 were wrongly matched to T14 and L15 respectively. Of these incorrect matches, only the second would lead to a wrong HSQC assignment if we had included the nitrogen prediction in our automated algorithm. Indeed, using the nitrogen chemical shift prediction as a validation for the product plane outcome, the residues T7 and T14 differing over six ppm in this dimension, the predicted N zone would indicate the wrong match. L43 is thus the only incorrectly assigned residue, as the corresponding predicted N zone would be unable to differentiate between 124.41 ppm (L43) and 125.16 ppm (L15). Some details on correct, incorrect and non- assignments of human ubiquitin are presented in fig. 4.

4 Discussion

The sequential assignment of protein NMR spectra, based on the matching of the C_α, C_β and CO chemical shifts between successive residues (Ikura et al, 1990; Kay et al, 1990; Montelione and Wagner, 1990), have been the rule for almost twenty years now. Although enormously powerful, this method requires entire sequence stretches to be assigned before chemical shift information becomes available for any of the residues in such a stretch. For one, this makes the process of assigning a subset of residues of particular interest (ligand binding regions, post-translational modification sites, ...) more labour intensive than necessary. Ideally, the assignment effort would be deliberately restricted to those residues of interest and their behaviour followed-up in a subsequent set of HSQC spectra. Secondly, in the case where a selective backbone labelling was applied during the recombinant protein production in order to obtain

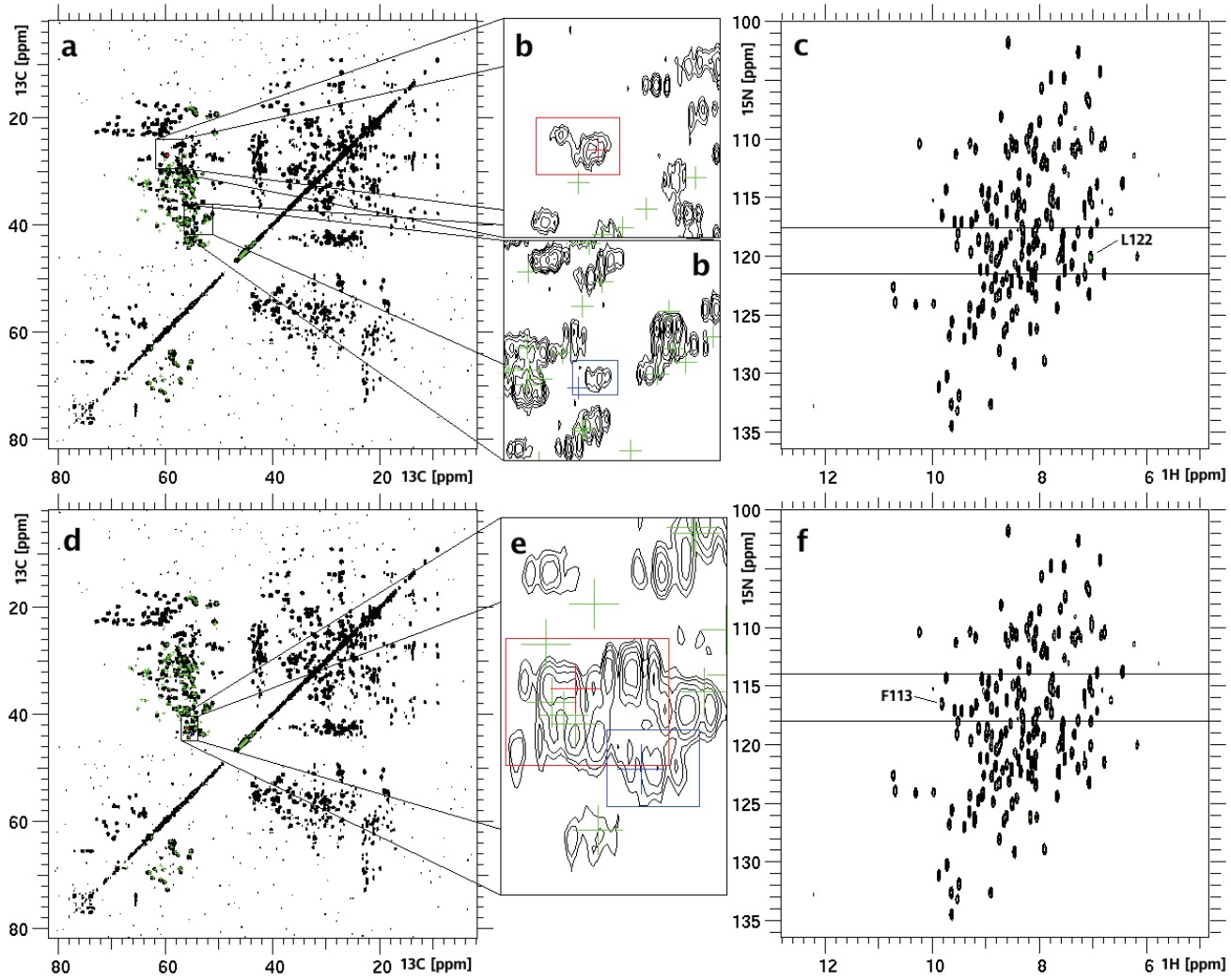


Fig. 3 Two examples of the assigning power of our semi-automatic non-sequential method. Among the SPARTA predicted CC TOCSY peaks of cyclophilin A in **a**, one belongs to the N121 prediction (red) and another to the L122 prediction (blue) of the protein. Zooming in on these different parts of the spectrum (**b**), allows to decide quite easily on the necessary window defining boxes. Executing the necessary plane summations and multiplications results in a final product plane presented on top of the cypA HSQC in **c**. A similar reasoning holds for the assignment of the F113 residue presented in **d**, **e** and **f**. **e** clearly reveals the moderate quality of these particular predictions. However, taken everything in consideration, relatively small C_α and C_β windows for application on both the HNCACB and HN(CO)CACB can be defined. Containing just two peaks (green signals), the product plane (**f**), together with predicted N shift information, unambiguously reveals the true F113 HSQC position.

less crowded NMR spectra, the sequential assignment method is not even an option, forcing researchers to reach for more exotic ways of assigning. Three more or less recent developments in NMR are the basis for the present work that deals with the chemical shift assignment of proteins of known structure. These are first of all the chemical shift predictors performing especially well for C_α, C_β and backbone amide nitrogen chemical shifts. Secondly, there is the possibility of recording spectra with direct carbon detection at acceptable S/N allows the use a CC TOCSY for the immediate inspection of the predictions. Finally, our own graphical interpretation of boolean logic readily allows to translate the carbon prediction into HSQC assignments. The presented results show that semi-automatically, large parts of protein sequences can be assigned without ever having peak picked a single spectrum.

The fast assignments that come along with the application of the proposed divide-and-conquer algorithm serve as ideal starting points for the semi-automatic sequential assignment method we reported previously. The algorithm offers the maximum amount of output in a minimum amount of time and with enough reliability. Although more efficient algorithms, such as the Kuhn-Munkres algorithm (Kuhn, 1955, 1956; Munkres, 1957), could improve the search for D_{min} (see Theory and Methods, mode 2) and the corresponding optimal matches, the introduction of certainty values for each

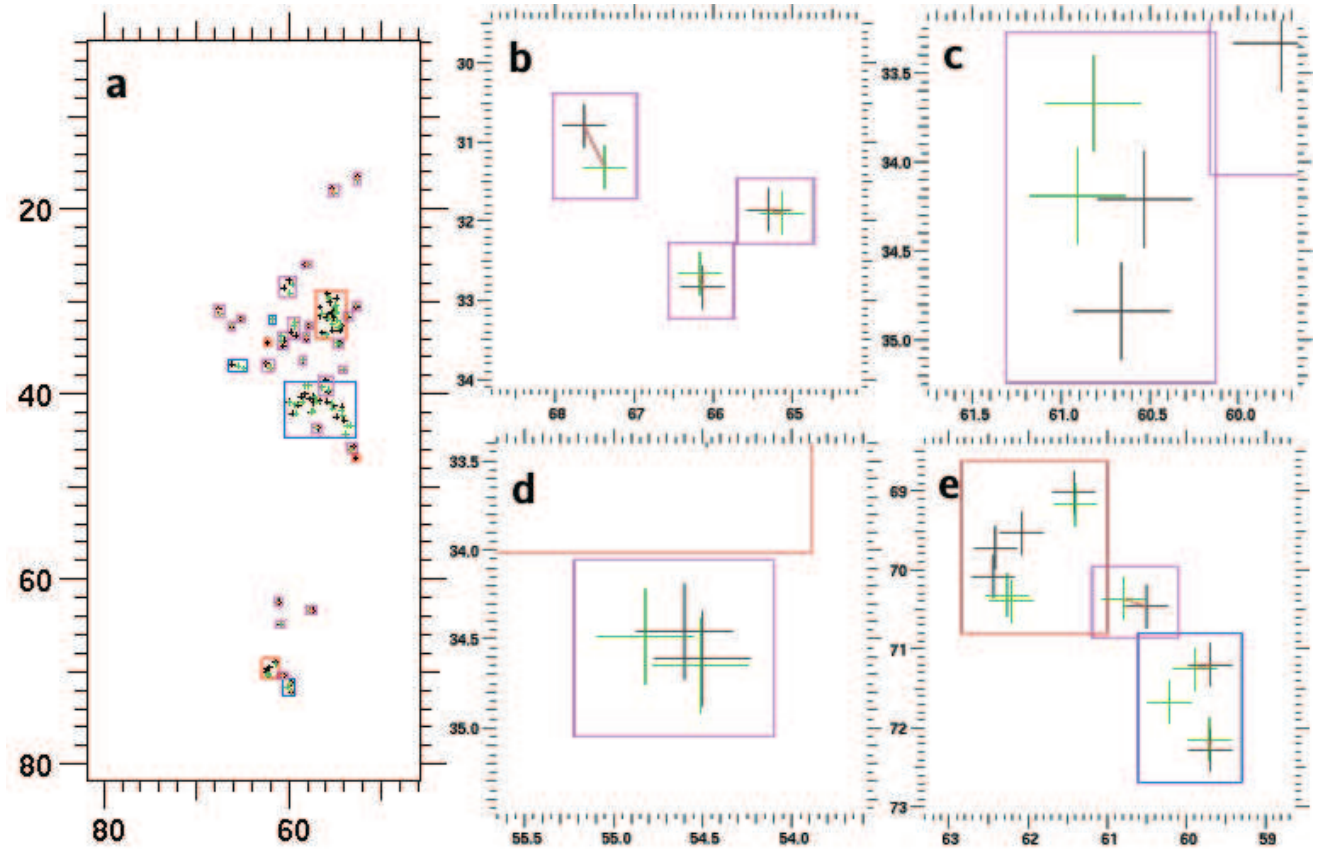


Fig. 4 **b**, **c**, **d** and **e** are selected zooms of the entire region spanned by the real and predicted C_α,C_β peaks of human ubiquitin presented in **a**. The divide boxes, for which the same colour code as in fig. 2 applies, are obtained after defining a critical distance D_c of 1 ppm. Matches between real (black) and predicted (green) C_α,C_β peaks are indicated by a red line. If a subproblem S_i is constituted of only one real and one predicted peak (**b**), a match is obvious. **c** shows a subproblem containing two real and two predicted peaks. From eq. 3 it follows that two match patterns are possible. However, since in this case the largest of the dissimilarity indices D_x is not greater than two times the smallest index, conquer rule 1 prevents an assignment. The situation presented in **d** would lead to two assignments according to rule 1, but is made impossible by conquer rule 2. What could be seen as an obvious matching problem might just be the result of unfortunate predicting errors, as the two real peaks are separated so poorly in the carbon carbon space. **e** shows the seven threonine signals and three matches that were made for these residues. However, one of the three matches (the one in the pink box) is incorrect. The green prediction peak to which it is matched (T14) really belongs to one of the real peaks in the red box and the true prediction of this T7 signal is the only free green peak in the blue box.

match made would be indispensable. However, not only are non-certain assignments in general pointless in the sense that they should be verified by some other means, but if, as in this case, the assignments are only intended as starting points for a more elaborate method, non-certain assignments are to be avoided altogether. Ultimately, more powerful algorithms should be developed, capable of surpassing the simple one-to-one matches done by our divide-and-conquer algorithm and by being able to decide on C_α and C_β windows as a human operator in the semi-automatic execution mode.

5 Acknowledgements

D.V. received a predoctoral grant of the French Ministère de la Recherche. We thank Drs. X. Hanoulle and I. Landrieu (Lille) for sample preparation and careful reading of the manuscript and Dr. R. Kuemmerle (Etlingen, Switzerland) for recording the CypA CC spectrum. The NMR facilities used in this study were funded by the FEDER, Ministère de la Recherche, Région Nord-Pas de Calais (France), the CNRS, the University of Lille 1 and the Institut Pasteur de Lille. Financial support from the TGIR-TGE RMN for conducting research is gratefully acknowledged.

References

Al-Hashimi HM, Patel DJ (2002) Residual dipolar couplings: Synergy between NMR and structural genomics. J Biomol NMR

Al-Hashimi HM, Gorin A, Majumdar A, Gosser A, Patel DJ (2002) Towards structural genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. J Mol Biol

Apaydın MS, Conitzer V, Donald BR (2008) Structure-based protein NMR assignments using native structural ensembles

Arun K, Langmead CJ (2006) Structure based chemical shift prediction using Random Forests non-linear regression. In: Proceedings of the 4th Asia-Pacific Bioinformatics Conference, pp 317–326

Atreya H, Sahu S, Chary K, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17:125–136

Atreya H, Chary K, Govil G (2002) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. Curr Sci 83:1372–1376

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparce unassigned NMR data. J Comput Biol 7:537–558

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res

Bermel W, Bertini I, Felli IC, Piccioli M, Pierattelli R (2006) ^{13}C-detected protonless NMR spectroscopy of proteins in solution. Prog NMR Spectrosc

Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett

Botos I, Melnikov EE, Cherry S, Khalatova AG, Rasulova FS, Tropea JE, Maurizi MR, Rotanova TV, Gustchina A, Wlodawer A (2004) Crystal structure of the AAA+ α domain of E. coli Lon protease at 1.9 Å resolution. J Struct Biol

Delaglio F, Kontaxis G, Bax A (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. J Am Chem Soc

Denisov AY, Ritter B, McPherson PS, Gehring K (2004) Letter to the Editor: 1H, 15N, 13C resonance assignments and 15N-1H residual dipolar couplings for the alpha-adaptin ear-domain. J Biomol NMR

Dobson CM, MA H, C R (1984) Nuclear overhauser effects and the assignment of the proton NMR spectra of proteins. FEBS Lett

Eletsky A, Moreira O, Kovacs H, Pervushin K (2003) A novel strategy for the assignment of side-chain resonances in completely deuterated large proteins using ^{13}C spectroscopy. J Biomol NMR

Erdmann MA, Rule GS (2002) Rapid protein structure detection and assignment using residual dipolar couplings. Tech. rep., School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

Gronwald W, Boyko RF, Snnichsen FD, Wishart DS, Sykes BD (1997) ORB, a homology-based program for the prediction of protein NMR chemical shifts. J Biomol NMR

Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA: Chemical shift based computer aided protein NMR assignments. J Biomol NMR 12:395–405

Grzesiek S, Bax A (1992) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson

Grzesiek S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly ^{13}C/^{15}N-enriched proteins. J Biomol NMR 3:185–204

Hus JC, Prompers JJ, Brüschweiler R (2002) Assignment strategy for proteins with known structure. J Magn Reson

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667

Jaravine VA, Zhuravleva A, Permi P, Ibraghimov I, Orekhov VY (2008) Hyperdimensional NMR spectroscopy with nonlinear sampling. J Am Chem Soc

Jung YS, Zweckstetter M (2004a) Backbone assignment of proteins with known structure using residual dipolar couplings. J Biomol NMR

Jung YS, Zweckstetter M (2004b) Mars - robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Ke H (1992) Similarities and differences between human cyclophilin A and other beta-barrel structures. Structural refinement at 1.63 Å resolution

Ke H, Mayrose D, Cao W (1993) Crystal structure of cyclophilin A complexed with substrate Ala-Pro suggests a solvent-assisted mechanism of cis-trans isomerization. Proc Natl Acad Sci USA

Kovacs H, Moskau D, Spraul M (2005) Crygenically cooled probes—a leap in NMR technology. Prog NMR Spectrosc

Kuhn H (1955) The Hungarian method for the assignment problem. Nav Res Logist Quarterly

Kuhn H (1956) Variants of the Hungarian method for assignment problems. Nav Res Logist Quarterly

Langmead CJ, Donald BR (2003) An improved nuclear vector replacement algorithm for nuclear magnetic resonance assignment. Tech. Rep. TR2004-494, Dartmouth College, Computer Science, Hanover, NH, URL http://www.cs.dartmouth.edu/reports/TR2004-494.pdf

Langmead CJ, Donald BR (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR assignments. J Biomol NMR 29:111–138

Langmead CJ, Yan A, Lilien R, Wang L, Donald BR (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. J Comput Biol 11:277–298

Luman NR, King MP, Augspurger JD (2001) Predicting ^{15}N amide chemical shifts in proteins. I. An additive model for the backbone contribution. J Comput Chem

Marin A, Malliavin TE, Nicolas P, Delsuc MA (2004) From NMR chemical shifts to amino acid types: Investigation of the predictive power carried by nuclei. J Biomol NMR 30:47–60

Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR

Montelione GT, Wagner G (1990) Conformation-independent sequential NMR connections in isotope-enriched polypeptides by ^{1}H-^{13}C-^{15}N triple resonance experiments. J Magn Reson 87:183–188

Munkres J (1957) Algorithms for the assignment and transportation problems. J Soc Indust and Appl Math

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. J Biomol NMR

Ottiger M, Zerbe O, Güntert P, Wüthrich K (1997) The NMR solution conformation of unligated human cyclophilin A. J Mol Biol

Owen D, Vallis Y, Noble M, Hunter J, Dafforn T, Evans P, McMahon H (1999) A structural explanation for the binding of multiple ligands by the alpha-adaptin appendage domain. Cell

Pristovšek P, Franzoni L (2006) Stereospecific assignments of protein NMR resonances based on the tertiary structure and 2D/3D NOE data. J Comput Chem

Pristovšek P, Rüterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program *st2nmr*. J Comput Chem 23:335–340

Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38:289–302

Smith CK, Wohnert J, Sauer RT, Schwalbe H (2001) Letter to the editor: Assignments of the 1H, 13C, and 15N resonances of the substrate-binding SSD domain from Lon protease. J Biomol NMR

Stratmann D, van Heijenoort C, Guittet E (2009) NOEnet—Use of NOE networks for NMR resonance assignment of proteins with

known 3D structure. Bioinformatics

Tian F, Valafar H, Prestegard J (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 123:11,791–11,796

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res

Verdegem D, Dijkstra K, Hanoulle X, Lippens G (2008) Graphical interpretation of boolean operators for protein NMR assignments. J Biomol NMR

Vijay-kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol

Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated ^{1}H and ^{13}C chemical shift prediction using the BioMagResBank. J Biomol NMR

Wittekind M, Mueller L (1993) HNCACB, A high sensitivity 3D NMR experiment to correlate amide proton and nitrogen resonance with the α-carbon and β-carbon resonances in proteins. J Magn Reson B 101:201–205

Xiong F, Bailey-Kellogg C (2007) A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, pp 403–410

Xiong F, Pandurangan G, Bailey-Kellogg C (2008) Contact replacement for nmr resonance assignment. Bioinformatics

Xu XP, Case DA (2001) Automated prediction of ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and ^{13}C' chemical shifts in proteins using a density functional database. J Biomol NMR

Xu Y, Wang X, Yang J, Vaynberg J, Qin J (2006) PASA - a program for automated protein NMR backbone signal assignment by pattern-filtering approach. J Biomol NMR 34:41–56

Zweckstetter M, Bax A (2001) Single-step determination of protein substructures using dipolar couplings: Aid to structural genomics. J Am Chem Soc

4.2. NMR Tool Development with NMRpython

All software tools described in this thesis have been written in python and use the NMRpython library. As this library is not documented, nor publically available, it is probably a good moment to discuss it into some more detail. The NMRpython library (NMRpy) is written at the university of Groningen (Netherlands) and was on ongoing development until its creator, engineer Klaas Dijkstra, retired somewhere in 2008. The Molecular Dynamics and NMR group of the Groningen University used NMRpy to execute more exotic manipulation on raw and processed NMR data, that are not implemented in standard NMR tools such as NMRpipe [92], Sparky [159], NMRview [205], ccpNMR [410] or whatever. Consecutive uses of their software some of the time revealed novel bugs that were then corrected. Because of the fact that the authors never considered NMRpy as finished, it was never described in a publication and never put online for download. Despite this, Klaas was kind enough to provide a copy of (the still buggy, but mainly well functioning) NMRpy to a very small amount of external scientist (I do not know the exact number however), including me.

The main advantage of NMRpy is that it allows the manipulation of NMR spectra as matrices. For example, the first of the following commands sets all directly acquired data points corresponding to the first data point in the indirect dimension of an HSQC spectrum to zero. The second command puts the point corresponding to the first data point in both direct and indirect dimensions equal to the second data point in both dimensions.

```
hsqc.array[0,:]=0.
hsqc.array[0,0]=hsqc.array[1,1]
```

(where one must know that counting in python start at 0). The resulting new HSQC spectra can afterwards be saved to disc or plotted on screen. These simple commands illustrate to what extent the matrix-interpretation of NMR spectra can lead to powerful applications. Besides this functionality, NMRpy contains numerous NMR-related functions for performing e.g. Fourier transformations, phase corrections, peak pickings, and so on.

The downside of NMRpy is the graphical part. For plotting one- or two-di-
mensional spectra on screen, it makes use of the (now) rather archaic GIST library [52]. This latter is seemingly incompatible with the Tcl/Tk language commonly applied by python to generate graphical user interface (GUI) programs. This means that in a in python written GUI program, if the NMRpy library is summoned, the (GIST) plot window blocks and does not respond to plotting or contouring commands. Hence, the only possible way to use NMRpy is in the command line interface mode, also supported by python. To obtain NMR related functionality, command after command has to by typed in the command line interface (see fig. 4.1).

It should be clear that the way of using NMRpy as demonstrated in fig. 4.1 impedes lengthy routine applications that differ slightly for every new supplied set of NMR spectra. Such routine applications are only possible if the spectroscopist is provided an additional GUI window wherein the different buttons to click represent different sets of python commands. I have been able to circumvent the GIST-Tcl/Tk incompatibility without the need of extensive programming, by using a concept from operating system and high-level program development: multithreading. A thread of execution

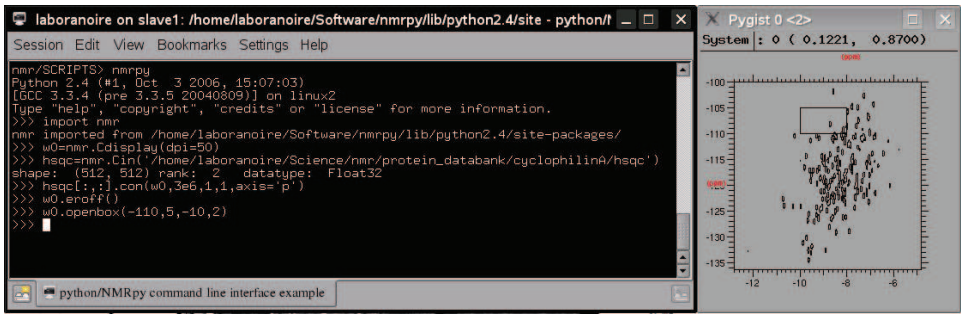


Figure 4.1. Demonstration of the use of the NMRpython library in the command line interface of python. The first "nmrpy" command opens the python interpreter (typing "nmrpy" is equivalent to typing "python"). Python commands can then be given one by one. "import nmr" allows the subsequent use of NMRpy. "w0=nmr.Cdisplay()" opens the GIST plot window which is seen at the right. "hsqc=nmr.Cin('someHSQC')" imports some HSQC spectrum from the hard disc. "hsqc[:,:].con(w0,3e6,1,1,axis='p')" contours this spectrum in window w0. "w0.eroff()" permits other spectra/things to be drawn on top of what is already present in w0. And finally, w0.openbox(-110,5,-10,2) puts a square box on screen with lower left coordinates (-10,-110), width 2 ppm and height 5 ppm. NMRpy plots spectra in ppm values on negative axes to obtain the declining-values-from-left-to-right-and-from-top-to-bottom effect.

allows for two or more concurrently running tasks in the same computer program. Threads within a process share state as well as memory and other resources.

Concretely, to obtain the combination of windows previously presented in fig. 2.15, which allow to make sequential NMR assignments, the following commands have to be given.

```
/home/me> python
>>> import nmr
>>> w0=nmr.Cdisplay()
>>> execfile('import.py')
>>> spectrum['hsqc'][:,:].con(w0,3e6,1,1,axis='p')
>>> execfile('verlip_1.5_t.py')
>>>
```

The first three lines function as described in fig. 4.1. `execfile('import.py')` replaces the earlier mentioned `nmr.Cin()` command and opens a window that allows to graphically import many different spectra (HSQC, HNCACB, HN(CO)
CACB, ...) at once. While this import window exists, the `w0` display becomes unusable for reasons explained above. Once closed, `w0` is again available, and an array variable named spectrum is created containing a reference to the different imported spectra. The command `spectrum['hsqc'][:,:].con(w0, 3e6,1,1,axis='p')` is arbitrary. The point is to fill the `w0` display with something before calling the final window, which seemed obligatory to prevent a crash of the application. The command `execfile('verlip_1.5_t.py')` finally opens the assignment GUI tool in a thread. The file verlip_1.5_t.py is actually a small script looking as follows:

```
import os
import __main__
import thread
def doit(program):
```

```
  orig_path=os.getcwd()
  os.chdir('/home/me/Software/nmrpy/lib/python2.4/site-packages/nmr/SCRIPTS')
  execfile(program)
  os.chdir(orig_path)
thread.start_new(doit,('verlip_1.5.py',))
```

This script basically summons the actual assignment code (written in file verlip_1.5.py and reproduced in Appendix A) and opens the GUI window that can now perfectly coexist with the `w0` window, and both are functional. The code in verlip_1.5.py shares the same memory and thus the same variables with the main program. The main program corresponds with the python command line interface which in this case also stays unlocked.

The way of working with NMRpy proposed here, seems to solve the GIST-
Tcl/Tk reconciliation problem rather effectively. Going deeper into the code to find a cleaner solution would perhaps prevent the rare occasional crash that characterises the system at this moment. Besides the tools described in this text, a series of other python programs was further developed. The complete list of implemented functionalities (representing more than 25 000 rules of code in total) is the following. There are tools for:

— making sequential NMR assignments (see above)
— performing the post-MUSIC functionality (see above)
— making NMR assignment of proteins with known structure (see above)
— qualitatively comparing two 1D spectral lines pulled out of two different 3D spectra
— determining NMR peak intensities in a few different manners (maximum intensity, integrated intensity, with or without baseline correction)
— rapidly determining the $^3J_{HN,H\alpha}$ coupling constant from peak intensities in the HNHA spectrum
— for converting between the bruker and NMRpy spectral format and vice versa
— a few other functionalities

NMRpy deserves to be further maintained and developed, so it could ultimately be made publically available and provide an easy means of software tool development for many NMR groups. Although large community tools such as CcpNMR aim, among other things, to provide a platform for NMR tool development, this latter demands an extensive knowledge of their data system and object oriented programming in python. The NMRpython library can be effectively used with only a basic knowledge of python.

Conclusion and Perspectives

The work described in this thesis is first of all of a technical nature. Because of the main interest of the research group in a few intrinsically unstructured proteins, the assignment of their NMR spectra was a continuous struggle. In order to facilitate this task, a graphical, semi-automatic assignment tool has been developed, that is greatly adapted to handling spectra with large amounts of peak overlap. A second assignment principle, allowing the straightforward assignment of structured proteins based on chemical shift predictions came as a logic consequence of the first tool.

With their NMR spectra assignment becoming a matter of only a few days, effort can be concentrated on those NMR experiments digging for the biologically relevant characteristics of the IUP, which can be done either in free solution, or with interacting partners.

The intrinsically disordered neuronal protein Tau is involved in tubulin polymerisation into microtubules during a process that seems highly regulated by different levels of Tau phosphorylation and the differential expression of the six Tau isoforms. However, the protein is invariably found back hyperphosphorylated when aggregated in straight or in paired helical filaments in Alzheimer disease-affected neurons. Despite substantial research efforts, most of the structural details this proteins exhibits during its different modes of action, are unknown. Future experiments, conducted with powerful techniques will have to fill this void. Hence, we have for example planned an NMR study of the interaction between Tau and T2R, that could instil us with additional information on Tau's behaviour. The numerous assigned Tau resonances will prove useful in all future experiments.

A second IUP studied in some more detail was the non-structural 5A (NS5A) protein of the Hepatitis C virus. We have investigated the structural behaviour of this protein in its free solution state. The protein's third domain (D3) is shown to contain at least one clear region with residual α-helical structure. It now seems highly likely that the D3 domain of this protein interacts with a cellular partner through a coupled folding upon binding process. The next challenge will be to identify the exact interacting partner. We are therefore also considering a study that could reveal interactions, if these exist, between NS5A and NS5B, since both of these proteins are reported to be involved in the viral RNA replication. The possible influence of phosphorylation of this protein, which is suggested to modulate the efficiency of the RNA replication, on its structural behaviour/interactions can also be examined. Another protein seemingly playing a vital role in the HCV RNA replication is the host's cyclophilin, indicated by the observed cyclosporin A inhibition of the replication process. In this work, the interactions between cyclophilin and NS5A are investigated. We found that NS5A indeed interacts with cyclophilin, but this at numerous places (prolines) in the NS5A protein chain. The current observations do not allow us to decide whether the PPIase activity of cyclophilin on NS5A is crucial in this matter

or whether the importance of the cyclophilin interaction for RNA replication is due to other aspects. This latter issue also remains to be investigated.

Appendix A

Python Code of the Assignment Tool

```
__main__.VERSION_num="1.5"

from Tkinter import *
__main__.Toplevel=Toplevel
__main__.Frame=Frame
__main__.Label=Label
__main__.Button=Button
__main__.Entry=Entry
__main__.Text=Text
__main__.Scrollbar=Scrollbar
__main__.Listbox=Listbox
__main__.RIDGE=RIDGE
__main__.E=E
__main__.W=W
__main__.Y=Y
__main__.SUNKEN=SUNKEN
__main__.VERTICAL=VERTICAL
__main__.LEFT=LEFT
__main__.RIGHT=RIGHT
__main__.BOTH=BOTH

import os

__main__.root=Tk()

# Initialising some parameters first
#-------------------------------------------------------------------------------
windowwidth=32
__main__.wiwi=windowwidth
#-------------------------------------------------------------------------------
__main__.current_spectrum='hsqc'
#-------------------------------------------------------------------------------
__main__.hsqc=None
__main__.hncacb=None
__main__.cbcanh=None
__main__.hncocacb=None
__main__.cbcaconh=None
__main__.hncaco=None
__main__.hnco=None
__main__.hnn=None
__main__.hncacb_slice=None
__main__.hncocacb_slice=None
__main__.hncaco_slice=None
__main__.hnco_slice=None
__main__.hnn_slice=None
__main__.plane1=None
__main__.plane2=None
__main__.plane3=None
__main__.selection=None
#-------------------------------------------------------------------------------
__main__.peaksigns_window=None
__main__.levels_window=None
__main__.info_window=None
__main__.about_window=None
#-------------------------------------------------------------------------------
__main__.result_hsqc=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_ca_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_cb_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_ca_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_cb_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_co_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
```

```
__main__.result_co_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_n1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
__main__.result_n2=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
#-------------------------------------------------------------------------------------
__main__.level_hsqc=3e6
__main__.numlines_hsqc=10
__main__.interline_hsqc=1.5
__main__.level_hncacb=1e7
__main__.numlines_hncacb=5
__main__.interline_hncacb=2
__main__.level_hncocacb=1e7
__main__.numlines_hncocacb=5
__main__.interline_hncocacb=2
__main__.level_hncaco=1e7
__main__.numlines_hncaco=5
__main__.interline_hncaco=2
__main__.level_hnco=3e6
__main__.numlines_hnco=5
__main__.interline_hnco=2
__main__.level_hnn=1e7
__main__.numlines_hnn=5
__main__.interline_hnn=1.5
__main__.level_selection=1e8
__main__.numlines_selection=5
__main__.interline_selection=2
#-------------------------------------------------------------------------------------
__main__.levelfactor=1.
__main__.oldvalue_scale=0.
__main__.oldvalue_N=0.
direction=IntVar()
__main__.cavar=IntVar()
__main__.cbvar=IntVar()
__main__.covar=IntVar()
__main__.planedir=StringVar()
#-------------------------------------------------------------------------------------
# some variables concerning peak signs
__main__.peaksign_ca_i=''
__main__.peaksign_cb_i=''
__main__.peaksign_ca_gly_i=''
__main__.peaksign_ca_imin1=''
__main__.peaksign_cb_imin1=''
__main__.peaksign_ca_gly_imin1=''
__main__.peaksign_co_i=''
__main__.peaksign_co_imin1=''
__main__.peaksign_hnn=''
#-------------------------------------------------------------------------------------
# some important chemical shift values
__main__.bmrb_ca={"Ala":53.15,"Cys":57.96,"Asp":54.65,"Glu":57.36,
                  "Phe":58.16,"Gly":45.33,"His":56.47,"Ile":61.58,
                  "Lys":56.95,"Leu":55.65,"Met":56.15,"Asn":53.52,
                  "Pro":63.32,"Gln":56.57,"Arg":56.81,"Ser":58.70,
                  "Thr":62.17,"Val":62.43,"Trp":57.64,"Tyr":58.09}
__main__.bmrb_cb={"Ala":18.95,"Cys":33.36,"Asp":40.82,"Glu":29.97,
                  "Phe":39.89,"Gly":45.33,"His":30.19,"Ile":38.58,
                  "Lys":32.74,"Leu":42.24,"Met":32.99,"Asn":38.64,
                  "Pro":31.81,"Gln":29.13,"Arg":30.63,"Ser":63.78,
                  "Thr":69.64,"Val":32.68,"Trp":30.00,"Tyr":39.26}
__main__.rancoil_ca={"Ala":52.5,"Cys":55.4,"Asp":54.2,"Glu":56.6,
                     "Phe":57.7,"Gly":45.1,"His":55.0,"Ile":61.1,
                     "Lys":56.2,"Leu":55.1,"Met":55.4,"Asn":53.1,
                     "Pro":63.3,"Gln":55.7,"Arg":56.0,"Ser":58.3,
                     "Thr":61.8,"Val":62.2,"Trp":57.5,"Tyr":57.9}
__main__.rancoil_cb={"Ala":19.1,"Cys":41.1,"Asp":41.1,"Glu":29.9,
                     "Phe":39.6,"Gly":45.1,"His":29.0,"Ile":38.8,
                     "Lys":33.1,"Leu":42.4,"Met":32.9,"Asn":38.9,
                     "Pro":32.1,"Gln":29.4,"Arg":30.9,"Ser":63.8,
                     "Thr":69.8,"Val":32.9,"Trp":29.6,"Tyr":38.8}
__main__.aacode={"Ala":"A","Cys":"C","Asp":"D","Glu":"E","Phe":"F",
                 "Gly":"G","His":"H","Ile":"I","Lys":"K","Leu":"L",
                 "Met":"M","Asn":"N","Pro":"P","Gln":"Q","Arg":"R",
                 "Ser":"S","Thr":"T","Val":"V","Trp":"W","Tyr":"Y",}
#-------------------------------------------------------------------------------------

#----------------------#
```

```
# Initialising the GUI |
#----------------------#
bigframe=Frame(root)
bigframe.grid()

#------------------------------#
# Import the different spectra |
#------------------------------#
def simport():
# first rename variables
 try: __main__.hsqc=__main__.spectrum['hsqc']
 except KeyError: pass
 try: __main__.hncacb=__main__.spectrum['hncacb']
 except KeyError: pass
 try: __main__.cbcanh=__main__.spectrum['cbcanh']
 except KeyError: pass
 try: __main__.hncocacb=__main__.spectrum['hncocacb']
 except KeyError: pass
 try: __main__.cbcaconh=__main__.spectrum['cbcaconh']
 except KeyError: pass
 try: __main__.hncaco=__main__.spectrum['hncaco']
 except KeyError: pass
 try: __main__.hnco=__main__.spectrum['hnco']
 except KeyError: pass
 try: __main__.hnn=__main__.spectrum['hnn']
 except KeyError: pass

# now some 'spectrum present?' checks
 if __main__.hsqc==None:
  print 'Please, import a hsqc spectrum as well'
  print 'use the command: -- execfile("import.py") -- to do this'
  return None
 if __main__.hncacb==None and __main__.cbcanh==None:
  print 'This algorithm requires at least a hncacb or cbcanh spectrum'
  print 'use the command: -- execfile("import.py") -- to import one'
  return None
 if __main__.hncocacb==None and __main__.cbcaconh==None:
  __main__.button_iplus1.config(state=DISABLED)
 if __main__.hncaco==None or __main__.hnco==None:
  __main__.button_co.config(state=DISABLED)
  __main__.button_co_i.config(state=DISABLED)
  __main__.button_co_imin1.config(state=DISABLED)
  __main__.checkbutton_co.config(state=DISABLED)
 if __main__.hnn==None:
  __main__.button_hnn.config(state=DISABLED)
  __main__.button_n1.config(state=DISABLED)
  __main__.button_n2.config(state=DISABLED)
 w0.er()
 __main__.current_spectrum='hsqc'
 __main__.hsqc[:,:].con(w0,__main__.level_hsqc,__main__.numlines_hsqc,__main__.interline_hsqc,axis='p')
 __main__.size()
 if __main__.cbcanh!=None and __main__.hncacb==None:
  __main__.peaksign_ca_i='-'
  __main__.peaksign_cb_i='+'
  __main__.peaksign_ca_gly_i='+'
  __main__.hncacb=__main__.cbcanh
 else:
  __main__.peaksign_ca_i='+'
  __main__.peaksign_cb_i='-'
  __main__.peaksign_ca_gly_i='+'
 if __main__.cbcaconh!=None and __main__.hncocacb==None:
  __main__.peaksign_ca_imin1='+'
  __main__.peaksign_cb_imin1='+'
  __main__.peaksign_ca_gly_imin1='+'
  __main__.hncocacb=__main__.cbcaconh
 else:
  __main__.peaksign_ca_imin1='+'
  __main__.peaksign_cb_imin1='-'
  __main__.peaksign_ca_gly_imin1='+'
 __main__.peaksign_co_i='+'
 __main__.peaksign_co_imin1='+'
 __main__.peaksign_hnn='+'

def draw_line(downfrom,upto,leftfrom,rightto):
```

```
 for i in range(5):
  w0.plyx((downfrom,upto),(leftfrom,rightto))
__main__.draw_line=draw_line

#---------------------------#
# Contour the HSQC onscreen |
#---------------------------#
def goto_hsqc():
 w0.er()
 __main__.current_spectrum='hsqc'
 __main__.hsqc[:,:].con(w0,__main__.level_hsqc,__main__.numlines_hsqc,__main__.interline_hsqc,axis='p')
 __main__.size()
__main__.goto_hsqc=goto_hsqc

#--------------------#
# Click an HSQC peak |
#--------------------#
def pick_hsqc():
 __main__.result_hsqc=w0.mouse(-1,0,"Click on a hsqc signal")
 cross_up=__main__.result_hsqc[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_hsqc[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_hsqc[0],__main__.result_hsqc[0])
 __main__.draw_line(__main__.result_hsqc[1],__main__.result_hsqc[1],cross_left,cross_right)

#---------------------------------------------#
# Contour the HNCACB/HN(CO)CACB slices onscreen |
#---------------------------------------------#
def goto_cacb():
 if __main__.planedir.get() not in ('H3D','N3D'):print 'Select an extract direction first';return None
 if result_hsqc[1]!=0.:
  w0.er()
  w0.eroff()
  if __main__.planedir.get()=='H3D':
   __main__.current_spectrum='cacb_h'
   val=__main__.hncacb.ppm_to_ch(1,-result_hsqc[1],1)[1]
   __main__.hncacb_slice=nmr.mk_nmr(__main__.hncacb.array[:,val,:])
   __main__.hncacb_slice.par.sfx=__main__.hncacb.par.sfx
   __main__.hncacb_slice.par.sfy=__main__.hncacb.par.sfz
   __main__.hncacb_slice.par.tofpx=__main__.hncacb.par.tofpx
   __main__.hncacb_slice.par.tofpy=__main__.hncacb.par.tofpz
   __main__.hncacb_slice.par.swx=__main__.hncacb.par.swx
   __main__.hncacb_slice.par.swy=__main__.hncacb.par.swz
  elif __main__.planedir.get()=='N3D':
   __main__.current_spectrum='cacb_n'
   val=__main__.hncacb.ppm_to_ch(1,1,-result_hsqc[0])[2]
   __main__.hncacb_slice=nmr.mk_nmr(__main__.hncacb.array[:,:,val])
   __main__.hncacb_slice.par.sfx=__main__.hncacb.par.sfy
   __main__.hncacb_slice.par.sfy=__main__.hncacb.par.sfz
   __main__.hncacb_slice.par.tofpx=__main__.hncacb.par.tofpy
   __main__.hncacb_slice.par.tofpy=__main__.hncacb.par.tofpz
   __main__.hncacb_slice.par.swx=__main__.hncacb.par.swy
   __main__.hncacb_slice.par.swy=__main__.hncacb.par.swz

  if (__main__.peaksign_ca_i=='+' or __main__.peaksign_ca_gly_i=='+') and __main__.peaksign_cb_i!='+':
   __main__.hncacb_slice[:,:].con(w0,__main__.level_hncacb,__main__.numlines_hncacb,
                                  __main__.interline_hncacb,axis='p',color='blue')
  if (__main__.peaksign_ca_i=='-' or __main__.peaksign_ca_gly_i=='-') and __main__.peaksign_cb_i!='-':
   __main__.hncacb_slice[:,:].con(w0,-__main__.level_hncacb,__main__.numlines_hncacb,
                                  __main__.interline_hncacb,axis='p',color='blue')
  if __main__.peaksign_cb_i=='+':
   __main__.hncacb_slice[:,:].con(w0,__main__.level_hncacb,__main__.numlines_hncacb,
                                  __main__.interline_hncacb,axis='p',color='red')
  elif __main__.peaksign_cb_i=='-':
   __main__.hncacb_slice[:,:].con(w0,-__main__.level_hncacb,__main__.numlines_hncacb,
                                  __main__.interline_hncacb,axis='p',color='red')
  __main__.size()
  if __main__.hncocacb!=None:
   if __main__.planedir.get()=='H3D':
    val=__main__.hncocacb.ppm_to_ch(1,-result_hsqc[1],1)[1]
    __main__.hncocacb_slice=nmr.mk_nmr(__main__.hncocacb.array[:,val,:])
    __main__.hncocacb_slice.par.sfx=__main__.hncocacb.par.sfx
    __main__.hncocacb_slice.par.sfy=__main__.hncocacb.par.sfz
```

```
    __main__.hncocacb_slice.par.tofpx=__main__.hncocacb.par.tofpx
    __main__.hncocacb_slice.par.tofpy=__main__.hncocacb.par.tofpz
    __main__.hncocacb_slice.par.swx=__main__.hncocacb.par.swx
    __main__.hncocacb_slice.par.swy=__main__.hncocacb.par.swz
   elif __main__.planedir.get()=='N3D':
    val=__main__.hncocacb.ppm_to_ch(1,1,-result_hsqc[0])[2]
    __main__.hncocacb_slice=nmr.mk_nmr(__main__.hncocacb.array[:,:,val])
    __main__.hncocacb_slice.par.sfx=__main__.hncocacb.par.sfy
    __main__.hncocacb_slice.par.sfy=__main__.hncocacb.par.sfz
    __main__.hncocacb_slice.par.tofpx=__main__.hncocacb.par.tofpy
    __main__.hncocacb_slice.par.tofpy=__main__.hncocacb.par.tofpz
    __main__.hncocacb_slice.par.swx=__main__.hncocacb.par.swy
    __main__.hncocacb_slice.par.swy=__main__.hncocacb.par.swz

   if __main__.peaksign_ca_imin1=='+' or __main__.peaksign_ca_gly_imin1=='+' or
    __main__.peaksign_cb_imin1=='+':
    __main__.hncocacb_slice[:,:].con(w0,__main__.level_hncocacb,__main__.numlines_hncocacb,
                                     __main__.interline_hncocacb,axis='p')
   if __main__.peaksign_ca_imin1=='-' or __main__.peaksign_ca_gly_imin1=='-' or
    __main__.peaksign_cb_imin1=='-':
    __main__.hncocacb_slice[:,:].con(w0,-__main__.level_hncocacb,__main__.numlines_hncocacb,
                                     __main__.interline_hncocacb,axis='p')

  if __main__.planedir.get()=='H3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[0],__main__.result_hsqc[0])
   w0.color('magenta')
   if __main__.result_ca_i[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_ca_i[1],__main__.result_ca_i[1],cross_left,cross_right)
   if __main__.result_cb_i[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_cb_i[1],__main__.result_cb_i[1],cross_left,cross_right)
   w0.color('green')
   if __main__.result_ca_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_ca_imin1[1],__main__.result_ca_imin1[1],cross_left,cross_right)
   if __main__.result_ca_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_cb_imin1[1],__main__.result_cb_imin1[1],cross_left,cross_right)
  elif __main__.planedir.get()=='N3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[1],__main__.result_hsqc[1])
   w0.color('magenta')
   if __main__.result_ca_i[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_ca_i[1],__main__.result_ca_i[1],cross_left,cross_right)
   if __main__.result_cb_i[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_cb_i[1],__main__.result_cb_i[1],cross_left,cross_right)
   w0.color('green')
   if __main__.result_ca_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_ca_imin1[1],__main__.result_ca_imin1[1],cross_left,cross_right)
   if __main__.result_ca_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_cb_imin1[1],__main__.result_cb_imin1[1],cross_left,cross_right)
  w0.color('black')

__main__.goto_cacb=goto_cacb

#---------------------#
# Click the Ca(i) peak |
#---------------------#
def pick_ca_i():
 __main__.result_ca_i=w0.mouse(-1,0,"Click on the Ca (i) signal")
 cross_up=__main__.result_ca_i[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_ca_i[1]-abs(__main__.limit1-__main__.limit2)/30
```

```
 cross_left=__main__.result_ca_i[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_ca_i[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_ca_i[0],__main__.result_ca_i[0])
 __main__.draw_line(__main__.result_ca_i[1],__main__.result_ca_i[1],cross_left,cross_right)

#----------------------#
# Click the Cb(i) peak |
#----------------------#
def pick_cb_i():
 __main__.result_cb_i=w0.mouse(-1,0,"Click on the Cb (i) signal")
 cross_up=__main__.result_cb_i[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_cb_i[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_cb_i[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_cb_i[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_cb_i[0],__main__.result_cb_i[0])
 __main__.draw_line(__main__.result_cb_i[1],__main__.result_cb_i[1],cross_left,cross_right)

#------------------------#
# Click the Ca(i-1) peak |
#------------------------#
def pick_ca_imin1():
 __main__.result_ca_imin1=w0.mouse(-1,0,"Click on the Ca (i-1) signal")
 cross_up=__main__.result_ca_imin1[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_ca_imin1[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_ca_imin1[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_ca_imin1[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_ca_imin1[0],__main__.result_ca_imin1[0])
 __main__.draw_line(__main__.result_ca_imin1[1],__main__.result_ca_imin1[1],cross_left,cross_right)

#------------------------#
# Click the Cb(i-1) peak |
#------------------------#
def pick_cb_imin1():
 __main__.result_cb_imin1=w0.mouse(-1,0,"Click on the Cb (i-1) signal")
 cross_up=__main__.result_cb_imin1[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_cb_imin1[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_cb_imin1[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_cb_imin1[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_cb_imin1[0],__main__.result_cb_imin1[0])
 __main__.draw_line(__main__.result_cb_imin1[1],__main__.result_cb_imin1[1],cross_left,cross_right)

#-----------------------------------------#
# Contour the HNCO/HN(CA)CO slices onscreen |
#-----------------------------------------#
def goto_co():
 if __main__.planedir.get() not in ('H3D','N3D'):print 'Select an extract direction first';return None
 if result_hsqc[1]!=0.:
  w0.er()
  w0.eroff()
  if __main__.planedir.get()=='H3D':
   __main__.current_spectrum='co_h'
   val=__main__.hncaco.ppm_to_ch(1,-result_hsqc[1],1)[1]
   __main__.hncaco_slice=nmr.mk_nmr(__main__.hncaco.array[:,val,:])
   __main__.hncaco_slice.par.sfx=__main__.hncaco.par.sfx
   __main__.hncaco_slice.par.sfy=__main__.hncaco.par.sfz
   __main__.hncaco_slice.par.tofpx=__main__.hncaco.par.tofpx
   __main__.hncaco_slice.par.tofpy=__main__.hncaco.par.tofpz
   __main__.hncaco_slice.par.swx=__main__.hncaco.par.swx
   __main__.hncaco_slice.par.swy=__main__.hncaco.par.swz
  elif __main__.planedir.get()=='N3D':
   __main__.current_spectrum='co_n'
   val=__main__.hncaco.ppm_to_ch(1,1,-result_hsqc[0])[2]
   __main__.hncaco_slice=nmr.mk_nmr(__main__.hncaco.array[:,:,val])
   __main__.hncaco_slice.par.sfx=__main__.hncaco.par.sfy
   __main__.hncaco_slice.par.sfy=__main__.hncaco.par.sfz
   __main__.hncaco_slice.par.tofpx=__main__.hncaco.par.tofpy
   __main__.hncaco_slice.par.tofpy=__main__.hncaco.par.tofpz
   __main__.hncaco_slice.par.swx=__main__.hncaco.par.swy
   __main__.hncaco_slice.par.swy=__main__.hncaco.par.swz

  if __main__.peaksign_co_i=='+':
   __main__.hncaco_slice[:,:].con(w0,__main__.level_hncaco,__main__.numlines_hncaco,
                                  __main__.interline_hncaco,axis='p',color='blue')
  elif __main__.peaksign_co_i=='-':
```

```
   __main__.hncaco_slice[:,:].con(w0,-__main__.level_hncaco,__main__.numlines_hncaco,
                                  __main__.interline_hncaco,axis='p',color='blue')
  __main__.size()

  if __main__.planedir.get()=='H3D':
   val=__main__.hncaco.ppm_to_ch(1,-result_hsqc[1],1)[1]
   __main__.hnco_slice=nmr.mk_nmr(__main__.hnco.array[:,val,:])
   __main__.hnco_slice.par.sfx=__main__.hnco.par.sfx
   __main__.hnco_slice.par.sfy=__main__.hnco.par.sfz
   __main__.hnco_slice.par.tofpx=__main__.hnco.par.tofpx
   __main__.hnco_slice.par.tofpy=__main__.hnco.par.tofpz
   __main__.hnco_slice.par.swx=__main__.hnco.par.swx
   __main__.hnco_slice.par.swy=__main__.hnco.par.swz
  elif __main__.planedir.get()=='N3D':
   val=__main__.hncaco.ppm_to_ch(1,1,-result_hsqc[0])[2]
   __main__.hnco_slice=nmr.mk_nmr(__main__.hnco.array[:,:,val])
   __main__.hnco_slice.par.sfx=__main__.hnco.par.sfy
   __main__.hnco_slice.par.sfy=__main__.hnco.par.sfz
   __main__.hnco_slice.par.tofpx=__main__.hnco.par.tofpy
   __main__.hnco_slice.par.tofpy=__main__.hnco.par.tofpz
   __main__.hnco_slice.par.swx=__main__.hnco.par.swy
   __main__.hnco_slice.par.swy=__main__.hnco.par.swz

  if __main__.peaksign_co_imin1=='+':
   __main__.hnco_slice[:,:].con(w0,__main__.level_hnco,__main__.numlines_hnco,
                                __main__.interline_hnco,axis='p')
  elif __main__.peaksign_co_imin1=='-':
   __main__.hnco_slice[:,:].con(w0,-__main__.level_hnco,__main__.numlines_hnco,
                                __main__.interline_hnco,axis='p')

  if __main__.planedir.get()=='H3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[0],__main__.result_hsqc[0])
   w0.color('magenta')
   if __main__.result_co_i[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_co_i[1],__main__.result_co_i[1],cross_left,cross_right)
   w0.color('green')
   if __main__.result_co_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[0]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[0]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_co_imin1[1],__main__.result_co_imin1[1],cross_left,cross_right)
  elif __main__.planedir.get()=='N3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[1],__main__.result_hsqc[1])
   w0.color('magenta')
   if __main__.result_co_i[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_co_i[1],__main__.result_co_i[1],cross_left,cross_right)
   w0.color('green')
   if __main__.result_co_imin1[1]!=0.:
    cross_left=__main__.result_hsqc[1]+abs(__main__.limit3-__main__.limit4)/30
    cross_right=__main__.result_hsqc[1]-abs(__main__.limit3-__main__.limit4)/30
    __main__.draw_line(__main__.result_co_imin1[1],__main__.result_co_imin1[1],cross_left,cross_right)
  w0.color('black')

__main__.goto_co=goto_co

#----------------------#
# Click the CO(i) peak |
#----------------------#
def pick_co_i():
 __main__.result_co_i=w0.mouse(-1,0,"Click on the CO (i) signal")
 cross_up=__main__.result_co_i[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_co_i[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_co_i[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_co_i[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_co_i[0],__main__.result_co_i[0])
 __main__.draw_line(__main__.result_co_i[1],__main__.result_co_i[1],cross_left,cross_right)

#------------------------#
# Click the CO(i-1) peak |
#------------------------#
def pick_co_imin1():
```

```
 __main__.result_co_imin1=w0.mouse(-1,0,"Click on the CO (i-1) signal")
 cross_up=__main__.result_co_imin1[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_co_imin1[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_co_imin1[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_co_imin1[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_co_imin1[0],__main__.result_co_imin1[0])
 __main__.draw_line(__main__.result_co_imin1[1],__main__.result_co_imin1[1],cross_left,cross_right)

#--------------------------------#
# Contour the HNN slice onscreen |
#--------------------------------#
def goto_hnn():
 if __main__.planedir.get() not in ('H3D','N3D'):print 'Select an extract direction first';return None
 if result_hsqc[1]!=0.:
  w0.er()
  w0.eroff()
  if __main__.planedir.get()=='H3D':
   __main__.current_spectrum='hnn_h'
   val=__main__.hnn.ppm_to_ch(1,-result_hsqc[1],1)[1]
   __main__.hnn_slice=nmr.mk_nmr(__main__.hnn.array[:,val,:])
   __main__.hnn_slice.par.sfx=__main__.hnn.par.sfx
   __main__.hnn_slice.par.sfy=__main__.hnn.par.sfz
   __main__.hnn_slice.par.tofpx=__main__.hnn.par.tofpx
   __main__.hnn_slice.par.tofpy=__main__.hnn.par.tofpz
   __main__.hnn_slice.par.swx=__main__.hnn.par.swx
   __main__.hnn_slice.par.swy=__main__.hnn.par.swz
  elif __main__.planedir.get()=='N3D':
   __main__.current_spectrum='hnn_n'
   val=__main__.hnn.ppm_to_ch(1,1,-result_hsqc[0])[2]
   __main__.hnn_slice=nmr.mk_nmr(__main__.hnn.array[:,:,val])
   __main__.hnn_slice.par.sfx=__main__.hnn.par.sfy
   __main__.hnn_slice.par.sfy=__main__.hnn.par.sfz
   __main__.hnn_slice.par.tofpx=__main__.hnn.par.tofpy
   __main__.hnn_slice.par.tofpy=__main__.hnn.par.tofpz
   __main__.hnn_slice.par.swx=__main__.hnn.par.swy
   __main__.hnn_slice.par.swy=__main__.hnn.par.swz

  if __main__.peaksign_hnn=='+':
   __main__.hnn_slice[:,:].con(w0,__main__.level_hnn,__main__.numlines_hnn,
                               __main__.interline_hnn,axis='p')
  if __main__.peaksign_hnn=='-':
   __main__.hnn_slice[:,:].con(w0,-__main__.level_hnn,__main__.numlines_hnn,
                               __main__.interline_hnn,axis='p')
  __main__.size()

  if __main__.planedir.get()=='H3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[0],__main__.result_hsqc[0])
   __main__.draw_line(__main__.result_hsqc[1],__main__.result_hsqc[1],__main__.limit3,__main__.limit4)
  elif __main__.planedir.get()=='N3D':
   __main__.draw_line(__main__.limit1,__main__.limit2,__main__.result_hsqc[1],__main__.result_hsqc[1])
   __main__.draw_line(__main__.result_hsqc[1],__main__.result_hsqc[1],__main__.limit3,__main__.limit4)

__main__.goto_hnn=goto_hnn

#-----------------------------------#
# Click the first N(i-1)/N(i+1) peak |
#-----------------------------------#
def pick_n1():
 __main__.result_n1=w0.mouse(-1,0,"Click on a N(i-1)/N(i+1) signal")
 cross_up=__main__.result_n1[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_n1[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_n1[0]+abs(__main__.limit3-__main__.limit4)/30
 cross_right=__main__.result_n1[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_n1[0],__main__.result_n1[0])
 __main__.draw_line(__main__.result_n1[1],__main__.result_n1[1],cross_left,cross_right)

#-----------------------------------#
# Click the second N(i-1)/N(i+1) peak |
#-----------------------------------#
def pick_n2():
 __main__.result_n2=w0.mouse(-1,0,"Click on a N(i-1)/N(i+1) signal")
 cross_up=__main__.result_n2[1]+abs(__main__.limit1-__main__.limit2)/30
 cross_down=__main__.result_n2[1]-abs(__main__.limit1-__main__.limit2)/30
 cross_left=__main__.result_n2[0]+abs(__main__.limit3-__main__.limit4)/30
```

```
 cross_right=__main__.result_n2[0]-abs(__main__.limit3-__main__.limit4)/30
 __main__.draw_line(cross_up,cross_down,__main__.result_n2[0],__main__.result_n2[0])
 __main__.draw_line(__main__.result_n2[1],__main__.result_n2[1],cross_left,cross_right)

#-------------------------------------------------------------------------#
# Defining the set of GUI widgets corresponding to the previous functions |
#-------------------------------------------------------------------------#
#-----------------------------------------------------#
# Widgets for the goto_hsqc() and pick_hsqc() functions |
#-----------------------------------------------------#
smallframe_hsqc=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
smallerframe_hsqc=Frame(smallframe_hsqc)
label_hsqc=Label(smallerframe_hsqc,text="--- HSQC ---                          ")
label_hsqc.pack(side=LEFT)
button_hsqc=Button(smallerframe_hsqc,width=1,text="GO",command=goto_hsqc)
button_hsqc.pack(side=RIGHT)
smallerframe_hsqc.grid(row=0,sticky=W)
button_hn=Button(smallframe_hsqc,width=(windowwidth-4),text="HN pick",command=pick_hsqc)
button_hn.grid(row=1,sticky=W)
smallframe_hsqc.grid(row=2,sticky=W)
#-------------------------------------------------#
# Widgets for obtaining the plane direction info |
#-------------------------------------------------#
smallframe_planedir=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
label_planedir1=Label(smallframe_planedir,text="Extract plane: H-3thD")
label_planedir1.grid(row=0,column=0,sticky=W)
__main__.radiobutton_H3D=Radiobutton(smallframe_planedir,variable=__main__.planedir,value='H3D')
__main__.radiobutton_H3D.grid(row=0,column=1,sticky=W)
label_planedir2=Label(smallframe_planedir,text="N-3thD")
label_planedir2.grid(row=0,column=2,sticky=W)
__main__.radiobutton_N3D=Radiobutton(smallframe_planedir,variable=__main__.planedir,value='N3D')
__main__.radiobutton_N3D.grid(row=0,column=3,sticky=W)
smallframe_planedir.grid(row=3,sticky=W)
#-----------------------------------------------------------------------------------------------------------------#
# Widgets for the goto_cacb(), pick_ca_i(), pick_cb_i(), pick_ca_imin1() and pick_cb_imin1() functions |
#-----------------------------------------------------------------------------------------------------------------#
middleframe_cacb=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
smallframe_cacb=Frame(middleframe_cacb)
smallerframe_cacb=Frame(smallframe_cacb)
label_cacb1=Label(smallerframe_cacb,text="--- HN")
label_cacb1.pack(side=LEFT)
label_cacb2=Label(smallerframe_cacb,text="CA",fg="blue")
label_cacb2.pack(side=LEFT)
label_cacb3=Label(smallerframe_cacb,text="CB",fg="red")
label_cacb3.pack(side=LEFT)
label_cacb4=Label(smallerframe_cacb,text="/ HNCOCACB ---")
label_cacb4.pack(side=LEFT)
smallerframe_cacb.grid(row=0,column=0,sticky=W)
button_cacb=Button(smallframe_cacb,width=1,text="GO",command=goto_cacb)
button_cacb.grid(row=0,column=1,sticky=W)
smallframe_cacb.grid(row=0,columnspan=2,sticky=W)
button_ca_i=Button(middleframe_cacb,width=(windowwidth-8)/2,text="CA(i) pick",command=pick_ca_i)
button_ca_i.grid(row=1,column=0,sticky=W)
button_cb_i=Button(middleframe_cacb,width=(windowwidth-8)/2,text="CB(i) pick",command=pick_cb_i)
button_cb_i.grid(row=2,column=0,sticky=W)
button_ca_imin1=Button(middleframe_cacb,width=(windowwidth-8)/2,text="CA(i-1) pick",command=pick_ca_imin1)
button_ca_imin1.grid(row=1,column=1,sticky=W)
button_cb_imin1=Button(middleframe_cacb,width=(windowwidth-8)/2,text="CB(i-1) pick",command=pick_cb_imin1)
button_cb_imin1.grid(row=2,column=1,sticky=W)
middleframe_cacb.grid(row=4,sticky=W)
#-------------------------------------------------------------------#
# Widgets for the goto_co(), pick_co_i() and pick_co_imin1() functions |
#-------------------------------------------------------------------#
middleframe_co=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
smallframe_co=Frame(middleframe_co)
smallerframe_co=Frame(smallframe_co)
label_co1=Label(smallerframe_co,text="--- HNCA")
label_co1.pack(side=LEFT)
label_co2=Label(smallerframe_co,text="CO",fg="blue")
label_co2.pack(side=LEFT)
label_co3=Label(smallerframe_co,text="/ HNCO ---          ")
label_co3.pack(side=LEFT)
smallerframe_co.grid(row=0,column=0,sticky=W)
__main__.button_co=Button(smallframe_co,width=1,text="GO",command=goto_co)
```

```
__main__.button_co.grid(row=0,column=1,sticky=W)
smallframe_co.grid(row=0,columnspan=2,sticky=W)
__main__.button_co_i=Button(middleframe_co,width=(windowwidth-8)/2,text="CO(i) pick",command=pick_co_i)
__main__.button_co_i.grid(row=1,column=0,sticky=W)
__main__.button_co_imin1=Button(middleframe_co,width=(windowwidth-8)/2,text="CO(i-1) pick",
 command=pick_co_imin1)
__main__.button_co_imin1.grid(row=1,column=1,sticky=W)
middleframe_co.grid(row=5,sticky=W)
#-------------------------------------------------------------#
# Widgets for the goto_hnn(), pick_n1() and pick_n2() functions |
#-------------------------------------------------------------#
middleframe_hnn=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
smallframe_hnn=Frame(middleframe_hnn)
label_hnn1=Label(smallframe_hnn,text="--- HNN ---                                ")
label_hnn1.pack(side=LEFT)
__main__.button_hnn=Button(smallframe_hnn,width=1,text="GO",command=goto_hnn)
__main__.button_hnn.pack(side=RIGHT)
smallframe_hnn.grid(row=0,columnspan=2,sticky=W)
__main__.button_n1=Button(middleframe_hnn,width=(windowwidth-8)/2,text="N1 pick",command=pick_n1)
__main__.button_n1.grid(row=1,column=0,sticky=W)
__main__.button_n2=Button(middleframe_hnn,width=(windowwidth-8)/2,text="N2 pick",command=pick_n2)
__main__.button_n2.grid(row=1,column=1,sticky=W)
middleframe_hnn.grid(row=6,sticky=W)

#-----------------------------------------------------#
# In/ex-clude the Ca plane in product plane calculation |
#-----------------------------------------------------#
def not_ca():
 if __main__.cavar.get()==0:
  print 'Ca plane will not be included in product plane calculation'
 elif __main__.cavar.get()==1:
  print 'Ca plane will also be included in product plane calculation'

#-----------------------------------------------------#
# In/ex-clude the Cb plane in product plane calculation |
#-----------------------------------------------------#
def not_cb():
 if __main__.cbvar.get()==0:
  print 'Cb plane will not be included in product plane calculation'
 elif __main__.cbvar.get()==1:
  print 'Cb plane will also be included in product plane calculation'

#-----------------------------------------------------#
# In/ex-clude the CO plane in product plane calculation |
#-----------------------------------------------------#
def not_co():
 if __main__.covar.get()==0:
  print 'CO plane will not be included in product plane calculation'
 elif __main__.covar.get()==1:
  print 'CO plane will also be included in product plane calculation'

#-------------------------------------------#
# Calculate and Contour the product i-1 plane |
#-------------------------------------------#
def select_imin1():
 if __main__.cavar.get()==0 and __main__.cbvar.get()==0 and __main__.covar.get()==0:
  print 'include at least one of three planes (Ca,Cb,CO)';return
 __main__.selection=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
 __main__.selection.par.sfx=__main__.hncacb.par.sfx
 __main__.selection.par.sfy=__main__.hncacb.par.sfy
 __main__.selection.par.tofpx=__main__.hncacb.par.tofpx
 __main__.selection.par.tofpy=__main__.hncacb.par.tofpy
 __main__.selection.par.swx=__main__.hncacb.par.swx
 __main__.selection.par.swy=__main__.hncacb.par.swy
 __main__.selection.array[:,:]=1
 peaksign=1
 if __main__.result_ca_imin1[1]!=0. and __main__.cavar.get()==1:
  val1=__main__.hncacb.ppm_to_ch(-__main__.result_ca_imin1[1],1,1)[0]
  __main__.plane1=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
  __main__.plane1.array[:,:]=__main__.hncacb.array[val1,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane1.array[:,:]
  if (-__main__.result_ca_imin1[1])<47.5:
   if __main__.peaksign_ca_gly_i=='-':peaksign=peaksign*-1
  else:
```

```
   if __main__.peaksign_ca_i=='-':peaksign=peaksign*-1
 if __main__.result_cb_imin1[1]!=0. and __main__.cbvar.get()==1:
  val2=__main__.hncacb.ppm_to_ch(-__main__.result_cb_imin1[1],1,1)[0]
  __main__.plane2=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
  __main__.plane2.array[:,:]=__main__.hncacb.array[val2,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane2.array[:,:]
  if __main__.peaksign_cb_i=='-':peaksign=peaksign*-1
 if __main__.result_co_imin1[1]!=0. and __main__.covar.get()==1:
  val3=__main__.hncaco.ppm_to_ch(-__main__.result_co_imin1[1],1,1)[0]
  __main__.plane3=nmr.mk_nmr(__main__.hncaco.array[1,:,:])
  __main__.plane3.array[:,:]=__main__.hncaco.array[val3,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane3.array[:,:]
  if __main__.peaksign_co_i=='-':peaksign=peaksign*-1
 max=__main__.selection.maxarg()[0]
 min=__main__.selection.minarg()[0]
 if peaksign==1 and abs(max)<abs(min):
  print 'WARNING : negative noise peak larger in absolute value than positive signal peak'
 if peaksign==-1 and abs(max)>abs(min):
  print 'WARNING : positive noise peak larger than negative signal peak in absolute value'
 if peaksign==-1:max=min
 __main__.max=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
 __main__.max.array[:,:]=max/1e9
 __main__.selection.array[:,:]=__main__.selection.array[:,:]/__main__.max.array[:,:]
 __main__.goto_hsqc()
 w0.eroff()
 __main__.selection[:,:].con(w0,__main__.level_selection,__main__.numlines_selection,
  __main__.interline_selection,axis='p',color='green')
 if __main__.hnn!=None:
  __main__.draw_line(__main__.result_n1[1],__main__.result_n1[1],__main__.limit3,__main__.limit4)
  __main__.draw_line(__main__.result_n2[1],__main__.result_n2[1],__main__.limit3,__main__.limit4)

#-------------------------------------------#
# Calculate and Contour the product i+1 plane |
#-------------------------------------------#
def select_iplus1():
 if __main__.cavar.get()==0 and __main__.cbvar.get()==0 and __main__.covar.get()==0:
  print 'include at least one of three planes (Ca,Cb,CO)';return
 __main__.selection=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
 __main__.selection.par.sfx=__main__.hncacb.par.sfx
 __main__.selection.par.sfy=__main__.hncacb.par.sfy
 __main__.selection.par.tofpx=__main__.hncacb.par.tofpx
 __main__.selection.par.tofpy=__main__.hncacb.par.tofpy
 __main__.selection.par.swx=__main__.hncacb.par.swx
 __main__.selection.par.swy=__main__.hncacb.par.swy
 __main__.selection.array[:,:]=1
 peaksign=1
 if __main__.result_ca_i[1]!=0. and __main__.cavar.get()==1:
  val1=__main__.hncocacb.ppm_to_ch(-__main__.result_ca_i[1],1,1)[0]
  __main__.plane1=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
  __main__.plane1.array[:,:]=__main__.hncocacb.array[val1,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane1.array[:,:]
  if (47.5>(-__main__.result_ca_i[1])):
   if __main__.peaksign_ca_gly_imin1=='-':peaksign=peaksign*-1
  else:
   if __main__.peaksign_ca_imin1=='-':peaksign=peaksign*-1
 if __main__.result_cb_i[1]!=0. and __main__.cbvar.get()==1:
  val2=__main__.hncocacb.ppm_to_ch(-__main__.result_cb_i[1],1,1)[0]
  __main__.plane2=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
  __main__.plane2.array[:,:]=__main__.hncocacb.array[val2,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane2.array[:,:]
  if __main__.peaksign_cb_imin1=='-':peaksign=peaksign*-1
 if __main__.result_co_i[1]!=0. and __main__.covar.get()==1:
  val3=__main__.hnco.ppm_to_ch(-__main__.result_co_i[1],1,1)[0]
  __main__.plane3=nmr.mk_nmr(__main__.hnco.array[1,:,:])
  __main__.plane3.array[:,:]=__main__.hnco.array[val3,:,:]
  __main__.selection.array[:,:]=__main__.selection.array[:,:]*__main__.plane3.array[:,:]
  if __main__.peaksign_co_imin1=='-':peaksign=peaksign*-1
 max=__main__.selection.maxarg()[0]
 min=__main__.selection.minarg()[0]
 if peaksign==1 and abs(max)<abs(min):
  print 'WARNING : negative noise peak larger in absolute value than positive signal peak'
 if peaksign==-1 and abs(max)>abs(min):
  print 'WARNING : positive noise peak larger than negative signal peak in absolute value'
 if peaksign==-1:max=min
```

```
 __main__.max=nmr.mk_nmr(__main__.hncacb.array[1,:,:])
 __main__.max.array[:,:]=max/1e9
 __main__.selection.array[:,:]=__main__.selection.array[:,:]/__main__.max.array[:,:]
 __main__.goto_hsqc()
 w0.eroff()
 __main__.selection[:,:].con(w0,__main__.level_selection,__main__.numlines_selection,
  __main__.interline_selection,axis='p',color='magenta')
 if __main__.hnn!=None:
  __main__.draw_line(__main__.result_n1[1],__main__.result_n1[1],__main__.limit3,__main__.limit4)
  __main__.draw_line(__main__.result_n2[1],__main__.result_n2[1],__main__.limit3,__main__.limit4)

#-------------------------------------------------------------------------#
# Defining the set of GUI widgets corresponding to the previous functions |
#-------------------------------------------------------------------------#
#-----------------------------------------------------------#
# Widgets for the not_ca(), not_cb() and not_co() functions |
#-----------------------------------------------------------#
smallframe_walkto=Frame(bigframe,relief=RIDGE,width=windowwidth,borderwidth=4)
label_walkto=Label(smallframe_walkto,text="--- Walk to...: ---")
label_walkto.grid(row=0,columnspan=2,sticky=W)
smallerframe_walkto=Frame(smallframe_walkto)
label_include_1=Label(smallerframe_walkto,text="Include: Ca")
label_include_1.grid(row=0,column=0,sticky=W)
__main__.checkbutton_ca=Checkbutton(smallerframe_walkto,variable=__main__.cavar,command=not_ca)
__main__.checkbutton_ca.grid(row=0,column=1,sticky=W)
label_include_2=Label(smallerframe_walkto,text="Cb")
label_include_2.grid(row=0,column=2,sticky=W)
__main__.checkbutton_cb=Checkbutton(smallerframe_walkto,variable=__main__.cbvar,command=not_cb)
__main__.checkbutton_cb.grid(row=0,column=3,sticky=W)
label_include_3=Label(smallerframe_walkto,text="CO")
label_include_3.grid(row=0,column=4,sticky=W)
__main__.checkbutton_co=Checkbutton(smallerframe_walkto,variable=__main__.covar,command=not_co)
__main__.checkbutton_co.grid(row=0,column=5,sticky=W)
label_include_4=Label(smallerframe_walkto,text="plane")
label_include_4.grid(row=0,column=6,sticky=W)
smallerframe_walkto.grid(row=1,columnspan=2,sticky=W)
#-------------------------------------------------------------#
# Widgets for the select_imin1() and select_iplus1() functions |
#-------------------------------------------------------------#
button_imin1=Button(smallframe_walkto,width=(windowwidth-8)/2,text="Select i-1",command=select_imin1)
button_imin1.grid(row=2,column=0,sticky=W)
__main__.button_iplus1=Button(smallframe_walkto,width=(windowwidth-8)/2,
 text="Select i+1",command=select_iplus1)
__main__.button_iplus1.grid(row=2,column=1,sticky=W)
smallframe_walkto.grid(row=7,sticky=W)

#---------------------------------------------#
# Fit the spectrum or spectral slice to window |
#---------------------------------------------#
def size():
 if __main__.current_spectrum=='hsqc':
  __main__.limit1=-__main__.hsqc.par.tofpy-(__main__.hsqc.par.swy/__main__.hsqc.par.sfy)/2.
  __main__.limit2=-__main__.hsqc.par.tofpy+(__main__.hsqc.par.swy/__main__.hsqc.par.sfy)/2.
  __main__.limit3=-__main__.hsqc.par.tofpx-(__main__.hsqc.par.swx/__main__.hsqc.par.sfx)/2.
  __main__.limit4=-__main__.hsqc.par.tofpx+(__main__.hsqc.par.swx/__main__.hsqc.par.sfx)/2.
 if __main__.current_spectrum=='cacb_h':
  __main__.limit1=-__main__.hncacb.par.tofpz-(__main__.hncacb.par.swz/__main__.hncacb.par.sfz)/2.
  __main__.limit2=-__main__.hncacb.par.tofpz+(__main__.hncacb.par.swz/__main__.hncacb.par.sfz)/2.
  __main__.limit3=-__main__.hncacb.par.tofpx-(__main__.hncacb.par.swx/__main__.hncacb.par.sfx)/2.
  __main__.limit4=-__main__.hncacb.par.tofpx+(__main__.hncacb.par.swx/__main__.hncacb.par.sfx)/2.
 if __main__.current_spectrum=='cacb_n':
  __main__.limit1=-__main__.hncacb.par.tofpz-(__main__.hncacb.par.swz/__main__.hncacb.par.sfz)/2.
  __main__.limit2=-__main__.hncacb.par.tofpz+(__main__.hncacb.par.swz/__main__.hncacb.par.sfz)/2.
  __main__.limit3=-__main__.hncacb.par.tofpy-(__main__.hncacb.par.swy/__main__.hncacb.par.sfy)/2.
  __main__.limit4=-__main__.hncacb.par.tofpy+(__main__.hncacb.par.swy/__main__.hncacb.par.sfy)/2.
 if __main__.current_spectrum=='co_h':
  __main__.limit1=-__main__.hncaco.par.tofpz-(__main__.hncaco.par.swz/__main__.hncaco.par.sfz)/2.
  __main__.limit2=-__main__.hncaco.par.tofpz+(__main__.hncaco.par.swz/__main__.hncaco.par.sfz)/2.
  __main__.limit3=-__main__.hncaco.par.tofpx-(__main__.hncaco.par.swx/__main__.hncaco.par.sfx)/2.
  __main__.limit4=-__main__.hncaco.par.tofpx+(__main__.hncaco.par.swx/__main__.hncaco.par.sfx)/2.
 if __main__.current_spectrum=='co_n':
  __main__.limit1=-__main__.hncaco.par.tofpz-(__main__.hncaco.par.swz/__main__.hncaco.par.sfz)/2.
  __main__.limit2=-__main__.hncaco.par.tofpz+(__main__.hncaco.par.swz/__main__.hncaco.par.sfz)/2.
  __main__.limit3=-__main__.hncaco.par.tofpy-(__main__.hncaco.par.swy/__main__.hncaco.par.sfy)/2.
```

```
  __main__.limit4=-__main__.hncaco.par.tofpy+(__main__.hncaco.par.swy/__main__.hncaco.par.sfy)/2.
 if __main__.current_spectrum==’hnn_h’:
  __main__.limit1=-__main__.hnn.par.tofpz-(__main__.hnn.par.swz/__main__.hnn.par.sfz)/2.
  __main__.limit2=-__main__.hnn.par.tofpz+(__main__.hnn.par.swz/__main__.hnn.par.sfz)/2.
  __main__.limit3=-__main__.hnn.par.tofpx-(__main__.hnn.par.swx/__main__.hnn.par.sfx)/2.
  __main__.limit4=-__main__.hnn.par.tofpx+(__main__.hnn.par.swx/__main__.hnn.par.sfx)/2.
 if __main__.current_spectrum==’hnn_n’:
  __main__.limit1=-__main__.hnn.par.tofpz-(__main__.hnn.par.swz/__main__.hnn.par.sfz)/2.
  __main__.limit2=-__main__.hnn.par.tofpz+(__main__.hnn.par.swz/__main__.hnn.par.sfz)/2.
  __main__.limit3=-__main__.hnn.par.tofpy-(__main__.hnn.par.swy/__main__.hnn.par.sfy)/2.
  __main__.limit4=-__main__.hnn.par.tofpy+(__main__.hnn.par.swy/__main__.hnn.par.sfy)/2.
 w0.limits(limit1,limit2,limit3,limit4)

__main__.size=size

#-------------------------------------------------------------------------#
# Defining the set of GUI widgets corresponding to the previous function |
#-------------------------------------------------------------------------#
#--------------------------------#
# Widgets for the size() function |
#--------------------------------#
smallframe_size=Frame(bigframe)
button_size=Button(smallframe_size,width=(windowwidth-2),text="size",command=size)
button_size.grid(row=0,column=0,sticky=W)
smallframe_size.grid(row=10,sticky=W)

#----------------------------------------#
# Most Of The Menubar Functions Start Here |
#----------------------------------------#

#--------------------------#
# Clear the plotting screen |
#--------------------------#
def erase():
 w0.er()

#----------------------------------------------------------#
# Clear the plotting screen + reinitialise some parameters |
#----------------------------------------------------------#
def clearmem():
 w0.er()
 __main__.hsqc=None
 __main__.hncacb=None
 __main__.cbcanh=None
 __main__.hncocacb=None
 __main__.cbcaconh=None
 __main__.hncaco=None
 __main__.hnco=None
 __main__.hnn=None
 __main__.hncacb_slice=None
 __main__.hncocacb_slice=None
 __main__.hncaco_slice=None
 __main__.hnco_slice=None
 __main__.hnn_slice=None
 __main__.plane1=None
 __main__.plane2=None
 __main__.plane3=None
 __main__.selection=None
 __main__.result_hsqc=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_ca_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_cb_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_ca_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_cb_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_co_i=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_co_imin1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_n1=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.result_n2=(0.,0.,0.,0.,0.,0.,0.,0.,0,0,0)
 __main__.i_type=’hn’
 __main__.imin1_type=’hn’
 if __main__.info_window!=None:
  __main__.text_info.delete(’1.0’,’50.0’)

#---------------------#
# The PeakSigns Window |
```

```
#---------------------#
def set_peaksigns():
 if __main__.entry_ca_i.get() in ('+','-',''):__main__.peaksign_ca_i=__main__.entry_ca_i.get()
 else: print "type '+' or '-' as Ca (i) sign";__main__.entry_ca_i.delete(0,10);return None
 if __main__.entry_cb_i.get() in ('+','-',''):__main__.peaksign_cb_i=__main__.entry_cb_i.get()
 else: print "type '+' or '-' as Cb (i) sign";__main__.entry_cb_i.delete(0,10);return None
 if __main__.entry_ca_gly_i.get() in ('+','-',''):__main__.peaksign_ca_gly_i=__main__.entry_ca_gly_i.get()
 else: print "type '+' or '-' as Ca (gly) (i) sign";__main__.entry_ca_gly_i.delete(0,10);return None
 if __main__.entry_ca_imin1.get() in ('+','-',''):__main__.peaksign_ca_imin1=__main__.entry_ca_imin1.get()
 else: print "type '+' or '-' as Ca (i-1) sign";__main__.entry_ca_imin1.delete(0,10);return None
 if __main__.entry_cb_imin1.get() in ('+','-',''):__main__.peaksign_cb_imin1=__main__.entry_cb_imin1.get()
 else: print "type '+' or '-' as Cb (i-1) sign";__main__.entry_cb_imin1.delete(0,10);return None
 if __main__.entry_ca_gly_imin1.get() in
  ('+','-',''):__main__.peaksign_ca_gly_imin1=__main__.entry_ca_gly_imin1.get()
 else: print "type '+' or '-' as Ca (gly) (i-1) sign";__main__.entry_ca_gly_imin1.delete(0,10);return None
 if __main__.entry_co_i.get() in ('+','-',''):__main__.peaksign_co_i=__main__.entry_co_i.get()
 else: print "type '+' or '-' as Co (i) sign";__main__.entry_co_i.delete(0,10);return None
 if __main__.entry_co_imin1.get() in ('+','-',''):__main__.peaksign_co_imin1=__main__.entry_co_imin1.get()
 else: print "type '+' or '-' as Co (i-1) sign";__main__.entry_co_imin1.delete(0,10);return None
 if __main__.entry_hnn.get() in ('+','-',''):__main__.peaksign_hnn=__main__.entry_hnn.get()
 else: print "type '+' or '-' as HNN sign";__main__.entry_hnn.delete(0,10);return None
 __main__.peaksigns_window.destroy()

__main__.set_peaksigns=set_peaksigns

def cancel_peaksigns():
 __main__.peaksigns_window.destroy()

__main__.cancel_peaksigns=cancel_peaksigns

#-----------------------------------#
# The PeakSigns Window: the GUI part |
#-----------------------------------#
def go_peaksigns():
 if __main__.peaksigns_window!=None:__main__.peaksigns_window.destroy()
 __main__.peaksigns_window=__main__.Toplevel(__main__.root)
 __main__.peaksigns_window.title("--- Peak Signs ---")
 smallframe_peaksigns=__main__.Frame(__main__.peaksigns_window,relief=__main__.RIDGE,
  width=__main__.wiwi/2,borderwidth=4)
 label_peak_ca_i=__main__.Label(smallframe_peaksigns,text='Ca (i)')
 label_peak_ca_i.grid(row=0,column=0,sticky=__main__.W)
 __main__.entry_ca_i=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_ca_i.grid(row=0,column=1,sticky=__main__.E)
 label_peak_cb_i=__main__.Label(smallframe_peaksigns,text='Cb (i)')
 label_peak_cb_i.grid(row=1,column=0,sticky=__main__.W)
 __main__.entry_cb_i=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_cb_i.grid(row=1,column=1,sticky=__main__.E)
 label_peak_ca_gly_i=__main__.Label(smallframe_peaksigns,text='Ca (gly) (i)')
 label_peak_ca_gly_i.grid(row=2,column=0,sticky=__main__.W)
 __main__.entry_ca_gly_i=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_ca_gly_i.grid(row=2,column=1,sticky=__main__.E)
 label_peak_ca_imin1=__main__.Label(smallframe_peaksigns,text='Ca (i-1)')
 label_peak_ca_imin1.grid(row=3,column=0,sticky=__main__.W)
 __main__.entry_ca_imin1=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_ca_imin1.grid(row=3,column=1,sticky=__main__.E)
 label_peak_cb_imin1=__main__.Label(smallframe_peaksigns,text='Cb (i-1)')
 label_peak_cb_imin1.grid(row=4,column=0,sticky=__main__.W)
 __main__.entry_cb_imin1=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_cb_imin1.grid(row=4,column=1,sticky=__main__.E)
 label_peak_ca_gly_imin1=__main__.Label(smallframe_peaksigns,text='Ca (gly) (i-1)')
 label_peak_ca_gly_imin1.grid(row=5,column=0,sticky=__main__.W)
 __main__.entry_ca_gly_imin1=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_ca_gly_imin1.grid(row=5,column=1,sticky=__main__.E)
 label_peak_co_i=__main__.Label(smallframe_peaksigns,text='Co (i)')
 label_peak_co_i.grid(row=6,column=0,sticky=__main__.W)
 __main__.entry_co_i=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_co_i.grid(row=6,column=1,sticky=__main__.E)
 label_peak_co_imin1=__main__.Label(smallframe_peaksigns,text='Co (i-1)')
 label_peak_co_imin1.grid(row=7,column=0,sticky=__main__.W)
 __main__.entry_co_imin1=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
 __main__.entry_co_imin1.grid(row=7,column=1,sticky=__main__.E)
 label_peak_hnn=__main__.Label(smallframe_peaksigns,text='HNN')
 label_peak_hnn.grid(row=8,column=0,sticky=__main__.W)
 __main__.entry_hnn=__main__.Entry(smallframe_peaksigns,relief=__main__.SUNKEN,width=2)
```

```
 __main__.entry_hnn.grid(row=8,column=1,sticky=__main__.E)
 smallframe_peaksigns.grid(row=0,sticky=__main__.W)
 smallframe_peaksigns2=__main__.Frame(__main__.peaksigns_window)
 button_set_peaksigns=__main__.Button(smallframe_peaksigns2,width=4,text="OK",command=__main__.set_peaksigns)
 button_set_peaksigns.grid(row=10,column=0,sticky=__main__.W)
 button_cancel_peaksigns=__main__.Button(smallframe_peaksigns2,width=4,text="Cancel",
  command=__main__.cancel_peaksigns)
 button_cancel_peaksigns.grid(row=10,column=1,sticky=__main__.W)
 smallframe_peaksigns2.grid(row=1,sticky=__main__.W)

 __main__.entry_ca_i.delete(0,10)
 __main__.entry_ca_i.insert(0,__main__.peaksign_ca_i)
 __main__.entry_cb_i.delete(0,10)
 __main__.entry_cb_i.insert(0,__main__.peaksign_cb_i)
 __main__.entry_ca_gly_i.delete(0,10)
 __main__.entry_ca_gly_i.insert(0,__main__.peaksign_ca_gly_i)
 __main__.entry_ca_imin1.delete(0,10)
 __main__.entry_ca_imin1.insert(0,__main__.peaksign_ca_imin1)
 __main__.entry_cb_imin1.delete(0,10)
 __main__.entry_cb_imin1.insert(0,__main__.peaksign_cb_imin1)
 __main__.entry_ca_gly_imin1.delete(0,10)
 __main__.entry_ca_gly_imin1.insert(0,__main__.peaksign_ca_gly_imin1)
 __main__.entry_co_i.delete(0,10)
 __main__.entry_co_i.insert(0,__main__.peaksign_co_i)
 __main__.entry_co_imin1.delete(0,10)
 __main__.entry_co_imin1.insert(0,__main__.peaksign_co_imin1)
 __main__.entry_hnn.delete(0,10)
 __main__.entry_hnn.insert(0,__main__.peaksign_hnn)

#-----------------#
# The Info Window |
#-----------------#
#-------------------------------------------------------------#
# The Info Window: Obtaining info on clicked peak positions |
#-------------------------------------------------------------#
def get_info():
 try: h=round(-__main__.result_hsqc[0],4)
 except AttributeError: h="???"
 if h==0.: h="???"
 try: n=round(-__main__.result_hsqc[1],4)
 except AttributeError: n="???"
 if n==0.: n="???"
 try: cai=round(-__main__.result_ca_i[1],4)
 except AttributeError: cai="???"
 if cai==0.: cai="???"
 try: cbi=round(-__main__.result_cb_i[1],4)
 except AttributeError: cbi="???"
 if cbi==0.: cbi="???"
 try: coi=round(-__main__.result_co_i[1],4)
 except AttributeError: coi="???"
 if coi==0.: coi="???"
 try: caimin1=round(-__main__.result_ca_imin1[1],4)
 except AttributeError: caimin1="???"
 if caimin1==0.: caimin1="???"
 try:  cbimin1=round(-__main__.result_cb_imin1[1],4)
 except AttributeError: cbimin1="???"
 if cbimin1==0.: cbimin1="???"
 try: coimin1=round(-__main__.result_co_imin1[1],4)
 except AttributeError: coimin1="???"
 if coimin1==0.: coimin1="???"
 i_type=''
 imin1_type=''
 if cai!="???" and cbi!="???":
  for i in __main__.aacode.keys():
   if abs(__main__.rancoil_ca[i]-cai)<3 and abs(__main__.rancoil_cb[i]-cbi)<3:
    i_type=i_type+__main__.aacode[i]
 else: i_type='???'
 if caimin1!="???" and cbimin1!="???":
  for i in __main__.aacode.keys():
   if abs(__main__.rancoil_ca[i]-caimin1)<3 and abs(__main__.rancoil_cb[i]-cbimin1)<3:
    imin1_type=imin1_type+__main__.aacode[i]
 else: imin1_type='???'
 text="residue\t\t???\nH(i)\t\t"+str(h)+"\nN(i)\t\t"+str(n)+"\nCa(i)\t\t"+str(cai)+"\nCb(i)\t\t"+str(cbi)
 text=text+"\nCO(i)\t\t"+str(coi)+"\nCa(i-1)\t\t"+str(caimin1)+"\nCb(i-1)\t\t"+str(cbimin1)
```

```
 text=text+"\nCO(i-1)\t\t"+str(coimin1)+"\npos i types\t"+i_type+"\npos i-1 types\t"+imin1_type
 __main__.text_info.delete('1.0','50.0')
 __main__.text_info.insert('1.0',text)

__main__.get_info=get_info

#-----------------------------------------------------#
# The Info Window: Writing back obtained info to file |
#-----------------------------------------------------#
def write_info():
# We determine in what file the info should be written
 if os.path.isfile(__main__.entry_choosedir.get()):
  pathname=os.path.split(__main__.entry_choosedir.get())[0]
  filename=os.path.split(__main__.entry_choosedir.get())[1]
 elif os.path.isdir(__main__.entry_choosedir.get()):
  pathname=__main__.entry_choosedir.get()
  filename='verlip_out.txt'
  os.system('touch '+pathname+'/'+filename)
 elif os.path.isdir(os.path.split(__main__.entry_choosedir.get())[0]):
  pathname=os.path.split(__main__.entry_choosedir.get())[0]
  filename=os.path.split(__main__.entry_choosedir.get())[1]
  os.system('touch '+pathname+'/'+filename)
 else:
  print 'Non-existing path'
  return None
# read the info from the info text and put in the right format
 input=open(pathname+'/'+filename,'r')
 firstline=input.readline()[:-1]
 input.close()
 titles_file_temp=firstline.split('\t')
 titles_file=[]
 values_info={}
 for title in titles_file_temp:
  if title!='':
   titles_file.append(title)
   values_info[title]=''

 titles_info_temp=__main__.text_info.get('1.0','50.0').split('\n')
 titles_info=[]
 for title in titles_info_temp:
  if title.split('\t')[0]!='':
   titles_info.append(title.split('\t')[0])
   if title.split('\t')[0]!=title.split('\t')[-1]:
    values_info[title.split('\t')[0]]=title.split('\t')[-1]
   else: values_info[title.split('\t')[0]]=''

 new_titles_file=titles_file[:]
 for title in titles_info:
  if title not in titles_file:
   new_titles_file.append(title)

 num_behindtabs={}
 for i in range(len(new_titles_file)):
  num_behindtabs[new_titles_file[i]]=2-(len(new_titles_file[i])//8)
  num_behindtabs[new_titles_file[i]+'_value']=2-(len(str(values_info[new_titles_file[i]]))//8)

 title_line=''
 info_line=''
 for i in range(len(new_titles_file)):
  title_line=title_line+new_titles_file[i]+num_behindtabs[new_titles_file[i]]*'\t'
  info_line=info_line+values_info[new_titles_file[i]]+num_behindtabs[new_titles_file[i]+'_value']*'\t'
 print 'Info written to file',filename
 print title_line
 print info_line
# write the info to the file
 input=open(pathname+'/'+filename,'r')
 content=input.readlines()
 input.close()
 new_content=[title_line+'\n']
 new_content.extend(content[1:])
 new_content.append(info_line+'\n')
 output=open(pathname+'/'+filename,'w')
 output.writelines(new_content)
 output.close()
```

```
__main__.write_info=write_info

#-------------------------------#
# The Info Window: the GUI part |
#-------------------------------#
def go_info():
 if __main__.info_window!=None:__main__.info_window.destroy()
 __main__.info_window=__main__.Toplevel(__main__.root)
 __main__.info_window.title("--- Info ---")
 smallframe_info=__main__.Frame(__main__.info_window,relief=RIDGE,width=__main__.wiwi,borderwidth=4)
 button_getinfo=__main__.Button(smallframe_info,width=(__main__.wiwi-12)/3,
  text="Get Info",command=__main__.get_info)
 button_getinfo.grid(row=1,column=0,sticky=__main__.W)
 button_writeinfo=__main__.Button(smallframe_info,width=(__main__.wiwi-12)/3,
  text="Write Info",command=__main__.write_info)
 button_writeinfo.grid(row=1,column=1,sticky=__main__.W)
 __main__.entry_choosedir=__main__.Entry(smallframe_info,width=(__main__.wiwi)/3,relief=__main__.SUNKEN)
 __main__.entry_choosedir.grid(row=1,column=2,sticky=__main__.W)
 __main__.entry_choosedir.insert(0,'/home/laboranoire/Science/nmr/protein_databank')
 smallerframe_info=__main__.Frame(smallframe_info)
 scrollbar_info=__main__.Scrollbar(smallerframe_info, orient=__main__.VERTICAL)
 __main__.text_info=__main__.Text(smallerframe_info,relief=__main__.SUNKEN,yscrollcommand=scrollbar_info.set,
  width=(__main__.wiwi-4),height=12)
 scrollbar_info.config(command=__main__.text_info.yview)
 scrollbar_info.pack(side=__main__.RIGHT, fill=__main__.Y)
 __main__.text_info.pack(side=LEFT, fill=__main__.BOTH,expand=1)
 smallerframe_info.grid(row=2,columnspan=3,sticky=__main__.W)
 smallframe_info.grid(row=0,sticky=W)

#-------------------#
# The Levels Window |
#-------------------#
def level_up():
 try: __main__.levelfactor=float(__main__.entry_factor.get())
 except ValueError: __main__.levelfactor=1.0
 if __main__.listbox_levels.get(ACTIVE)=='HSQC':
  product=__main__.level_hsqc*__main__.levelfactor
  print 'level hsqc =',__main__.level_hsqc,'*',__main__.levelfactor,'=',product
  __main__.level_hsqc=product
 elif __main__.listbox_levels.get(ACTIVE)=='HNCACB':
  product=__main__.level_hncacb*__main__.levelfactor
  print 'level hncacb =',__main__.level_hncacb,'*',__main__.levelfactor,'=',product
  __main__.level_hncacb=product
 elif __main__.listbox_levels.get(ACTIVE)=='HNCOCACB':
  product=__main__.level_hncocacb*__main__.levelfactor
  print 'level hncocacb =',__main__.level_hncocacb,'*',__main__.levelfactor,'=',product
  __main__.level_hncocacb=product
 elif __main__.listbox_levels.get(ACTIVE)=='HNCACO':
  product=__main__.level_hncaco*__main__.levelfactor
  print 'level hncaco =',__main__.level_hncaco,'*',__main__.levelfactor,'=',product
  __main__.level_hncaco=product
 elif __main__.listbox_levels.get(ACTIVE)=='HNCO':
  product=__main__.level_hnco*__main__.levelfactor
  print 'level hnco =',__main__.level_hnco,'*',__main__.levelfactor,'=',product
  __main__.level_hnco=product
 elif __main__.listbox_levels.get(ACTIVE)=='HNN':
  product=__main__.level_hnn*__main__.levelfactor
  print 'level hnn =',__main__.level_hnn,'*',__main__.levelfactor,'=',product
  __main__.level_hnn=product
 elif __main__.listbox_levels.get(ACTIVE)=='Selection':
  product=__main__.level_selection*__main__.levelfactor
  print 'level selection =',__main__.level_selection,'*',__main__.levelfactor,'=',product
  __main__.level_selection=product
 else: print 'Something went very badly wrong!!'

__main__.level_up=level_up

def level_down():
 try: __main__.levelfactor=float(__main__.entry_factor.get())
 except ValueError: __main__.levelfactor=1.0
 if __main__.listbox_levels.get(ACTIVE)=='HSQC':
  quotient=__main__.level_hsqc/__main__.levelfactor
  print 'level hsqc =',__main__.level_hsqc,'/',__main__.levelfactor,'=',quotient
```

```
  __main__.level_hsqc=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='HNCACB':
  quotient=__main__.level_hncacb/__main__.levelfactor
  print 'level hncacb =',__main__.level_hncacb,'/',__main__.levelfactor,'=',quotient
  __main__.level_hncacb=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='HNCOCACB':
  quotient=__main__.level_hncocacb/__main__.levelfactor
  print 'level hncocacb =',__main__.level_hncocacb,'/',__main__.levelfactor,'=',quotient
  __main__.level_hncocacb=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='HNCACO':
  quotient=__main__.level_hncaco/__main__.levelfactor
  print 'level hncaco =',__main__.level_hncaco,'/',__main__.levelfactor,'=',quotient
  __main__.level_hncaco=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='HNCO':
  quotient=__main__.level_hnco/__main__.levelfactor
  print 'level hnco =',__main__.level_hnco,'/',__main__.levelfactor,'=',quotient
  __main__.level_hnco=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='HNN':
  quotient=__main__.level_hnn/__main__.levelfactor
  print 'level hnn =',__main__.level_hnn,'/',__main__.levelfactor,'=',quotient
  __main__.level_hnn=quotient
 elif __main__.listbox_levels.get(ACTIVE)=='Selection':
  quotient=__main__.level_selection/__main__.levelfactor
  print 'level selection =',__main__.level_selection,'/',__main__.levelfactor,'=',quotient
  __main__.level_selection=quotient
 else: print 'Something went very badly wrong!!'

__main__.level_down=level_down

#---------------------------------#
# The Levels Window: the GUI part |
#---------------------------------#
def go_levels():
 if __main__.levels_window!=None:__main__.levels_window.destroy()
 __main__.levels_window=__main__.Toplevel(__main__.root)
 __main__.levels_window.title("--- Levels ---")
 smallframe_levels=__main__.Frame(__main__.levels_window,relief=RIDGE,width=__main__.wiwi,borderwidth=4)
 smallerframe_levels=__main__.Frame(smallframe_levels)
 scrollbar_levels=__main__.Scrollbar(smallerframe_levels, orient=__main__.VERTICAL)
 __main__.listbox_levels=__main__.Listbox(smallerframe_levels,yscrollcommand=scrollbar_levels.set,
 width=(__main__.wiwi/3),height=1)
 scrollbar_levels.config(command=__main__.listbox_levels.yview)
 scrollbar_levels.pack(side=__main__.RIGHT, fill=__main__.Y)
 for item in ["HSQC","HNCACB","HNCOCACB","HNCACO","HNCO","HNN","Selection"]:
  __main__.listbox_levels.insert(END, item)
 __main__.listbox_levels.pack(side=__main__.LEFT, fill=__main__.BOTH,expand=1)
 smallerframe_levels.grid(row=1,column=0,sticky=W)

 smallerframe_levels2=__main__.Frame(smallframe_levels)
 label_factor=__main__.Label(smallerframe_levels2,text="     factor:")
 label_factor.grid(row=0,column=0,sticky=__main__.W)
 __main__.entry_factor=__main__.Entry(smallerframe_levels2,relief=SUNKEN,width=(8))
 __main__.entry_factor.grid(row=0,column=1,sticky=__main__.W)
 button_factorup=__main__.Button(smallerframe_levels2,width=5,text="UP",command=__main__.level_up)
 button_factorup.grid(row=1,column=0,sticky=__main__.W)
 button_factordown=__main__.Button(smallerframe_levels2,width=5,text="DOWN",command=__main__.level_down)
 button_factordown.grid(row=1,column=1,sticky=__main__.W)
 smallerframe_levels2.grid(row=1,column=1,sticky=__main__.W)
 smallframe_levels.grid(row=0,sticky=__main__.W)

#------------------#
# The About Window |
#------------------#
def go_about():
 if __main__.about_window!=None:__main__.about_window.destroy()
 __main__.about_window=__main__.Toplevel(__main__.root)
 __main__.about_window.title("--- About ---")
 about="This is the Protein NMR Assignment Program:\nVerLipAssign %s" % __main__.VERSION_num
 about=about+"\nWritten by Dries Verdegem\nIf any questions or problems,
  mail to:\ndries.verdegem@ed.univ-lille1.fr"
 label_about=__main__.Label(__main__.about_window,text=about)
 label_about.pack()

top=Menu(__main__.root)
```

```
__main__.root.config(menu=top)

mainmenu=Menu(top,tearoff=0)
mainmenu.add_command(label='Import',command=simport,underline=0)
mainmenu.add_command(label='Erase',command=erase,underline=0)
mainmenu.add_command(label='Clear Memory',command=clearmem,underline=0)
top.add_cascade(label='Main',menu=mainmenu,underline=0)

windowsmenu=Menu(top)
windowsmenu.add_command(label='PeakSigns Window', command=go_peaksigns,underline=0)
windowsmenu.add_command(label='Levels Window', command=go_levels,underline=0)
windowsmenu.add_command(label='Info Window', command=go_info,underline=0)
top.add_cascade(label='Windows',menu=windowsmenu,underline=0)

helpmenu=Menu(top,tearoff=0)
helpmenu.add_command(label='About',command=go_about,underline=0)
top.add_cascade(label='Help',menu=helpmenu,underline=0)

__main__.root.title("VerLipAssign %s" % __main__.VERSION_num)
__main__.root.wait_window()
```

Appendix B

Tau F3 and F5 Chemical Shifts

Tau F3

residue	HN	N	Ca	Cb	CO
S208					
R209					
S210					
R211					
T212					
P213			63.2911	32.1664	176.794
S214	8.4674	116.8568	58.1822	63.8772	174.0858
L215	8.3601	125.6492	53.0049	41.8738	175.2278
P216			62.9711	32.037	176.8292
T217	8.3866	118.1086	59.9942	69.9605	172.5734
P218					
P219			62.9711	32.037	177.0754
T220	8.3098	115.1469	61.9357	69.9605	174.4551
R221	8.4653	123.86	55.8524	30.9344	175.8796
E222	8.5069	124.0596	54.4286	29.7673	174.4375
P223			63.0334	32.1664	176.794
K224	8.4564	121.9267	56.3701	33.0074	176.6533
K225	8.4525	124.0318	56.2407	33.1363	176.4423
V226	8.2696	122.7792	62.0651	32.943	175.6685
A227	8.4656	128.8586	52.2889	19.3526	177.4271
V228	8.2511	121.2128	62.324	32.8135	176.0709
V229	8.4028	126.0715	62.324	32.6841	175.9499
R230	8.6035	126.4999	55.8524	31.0015	176.0554
T231	8.3562	119.2981	59.8648	69.831	172.3272
P232					
P233			62.8414	32.0991	176.8109
K234	8.4853	121.9059	56.2407	33.2018	176.6181
S235	8.4994	118.996	56.3701	63.3594	172.89
P236			63.3595	32.0994	177.2161
S237	8.4841	116.1069	58.5076	63.6805	174.8948
S238	8.3915	118.0832	58.3116	63.8772	174.5079
A239	8.3534	125.9544	53.0049	19.0938	178.0781
K240	8.2919	120.2398	56.629	32.943	176.9699
S241	8.2779	116.7759	58.5705	63.8772	174.7016
R242	8.3971	123.0973	56.3701	30.7426	176.3016
L243	8.2374	122.9896	55.3347	42.3915	177.3568
Q244	8.4471	121.7468	55.8524	29.4483	175.9851
T245	8.1875	115.9292	61.6768	69.9605	173.9099
A246	8.3713	128.1117	50.5457	18.1877	175.3872
P247			62.8417	32.037	176.6885
V248	8.3011	122.2571	59.8648	32.6841	174.6134
P249			62.9711	32.1663	176.6352
M250	8.4919	122.193	53.2638	32.7453	174.543
P251			63.23	32.1664	176.3368
D252	8.4841	120.729	54.0403	41.0972	176.583
L253	8.3256	123.8207	55.5935	42.0032	177.814
K254	8.3483	120.5226	56.8879	32.5547	176.6337
N255	8.3211	118.6919	53.3274	38.7674	175.1761
V256	8.0257	120.5087	62.7123	32.6186	176.372
K257	8.473	124.9921	56.4291	32.9429	176.7588
S258	8.3058	117.1114	58.441	63.8772	174.5247
K259	8.4522	123.8051	56.3701	33.0724	176.6707
I260	8.2267	122.4536	61.5474	38.5719	176.8815
G261	8.6013	113.5404	45.239	45.239	174.1913
S262	8.2948	115.7369	58.3116	64.0066	175.3168
T263	8.3937	115.8475	62.1945	69.5721	174.9651
E264	8.4587	123.0721	57.0816	30.2249	176.3016
N265	8.4994	119.4193	53.4549	38.638	175.3168
L266	8.1905	122.3626	55.4641	42.2621	175.3168
K267	8.2221	121.0987	56.4996	32.943	176.3368
H268	8.2305	120.1512	55.9818	30.7426	174.9651

Q269	8.3653	123.1457	53.5226	28.9306	173.9099
P270			63.6183	32.0369	177.7085
G271	8.6795	110.1627	45.3684	45.3684	175.0173
G272	8.3895	108.7682	45.3684		174.0506
G273	8.382	108.7419	45.1096		174.0506
K274	8.2193	120.9394	56.3701	33.0724	176.7045
V275	8.2415	122.1985	62.3841	32.8135	175.9499
Q276	8.5709	125.3765	55.5935	29.639	175.5278
I277	8.3991	124.2488	61.0297	38.5085	176.0906
I278	8.3752	126.1864	60.9002	38.638	175.7037
N279	8.6223	124.0588	53.0673	38.8968	175.0003
K280	8.4111	123.0537	56.5612	33.0724	176.372
K281	8.3765	122.7774	56.4995	32.8135	176.4775
L282	8.2438	123.8229	55.0758	42.5209	176.7588
D283	8.4146	122.17	54.0404	41.0972	176.4231
L284	8.4205	124.0674	55.3346	41.8085	177.9011
S285	8.4244	116.1015	59.4765	63.6773	174.6486
N286	8.3598	120.4289	53.3932	38.8306	175.352
V287	8.0113	120.1724	62.9711	32.5547	176.4775
Q288	8.5069	123.7168	56.1113	29.3189	176.2664
S289	8.3783	117.2353	58.5705	63.8079	174.7892
K290	8.4632	123.4206	56.4996	32.943	176.7053
C291	8.416	120.2055	58.8293	28.0246	175.2113
G292	8.5726	111.7998	45.3684		174.1913
S293	8.2988	115.771	58.441	63.8772	175.0512
K294	8.5399	123.2476	56.629	32.8135	176.4247
D295	8.27	120.4691	54.6196	41.2266	175.9147
N296	8.3563	118.7927	53.3932	38.7674	175.0873
I297	8.0357	120.9976	61.2885	38.5085	176.0906
K298	8.3785	125.3194	56.1113	32.943	176.0202
H299	8.4095	122.0617	56.1113	30.7426	174.8947
V300	8.213	124.1896	59.7353	32.6841	174.2617
P301			63.6183	32.037	177.6733
G302	8.6557	110.4102	45.3684		175.0003
G303	8.4147	108.9253	45.3684		174.9614
G304	8.4146	108.9354	45.239		174.1913
S305	8.2957	115.7295	58.3116	64.0066	174.5782
V306	8.2335	121.9374	62.324	32.8135	175.9499
Q307	8.5054	124.8884	55.5935	29.4483	175.563
I308	8.3342	124.282	61.0297	38.638	175.7553
V309	8.2572	125.7354	62.0002	32.943	175.4927
Y310	8.5048	126.3645	57.9233	39.0263	175.0706
K311	8.2204	126.0992	53.6521	33.0054	173.5744
P312			62.9711	32.037	176.8995
V313	8.2462	120.6254	62.4534	32.943	175.6685
D314	8.4428	124.4278	53.7137	41.032	178.0602
L315	8.5822	125.3856	55.5935	41.7443	178.0602
S316	8.4792	116.2231	59.7935	63.7477	174.9299
K317	8.0171	122.1639	56.1113	32.873	176.6356
V318	8.0018	121.137	62.7123	32.6841	176.5478
T319	8.3224	118.1495	61.8063	69.9605	174.5591
S320	8.3679	118.5886	58.2447	64.0066	
K321	8.5066	123.7342	56.4347	33.0031	176.7236
S322	8.4655	117.379	58.5705	63.8772	175.0706
G323	8.4928	111.8541	45.3684		173.4175
S234	7.9742	121.3219	59.9942	64.9751	178.7803

Tau F5

residue	HN	N	Ca	Cb	CO
N167			53.2638	39.0262	174.8947
A168	8.5856	125.3368	52.6166	19.2875	177.7964
T169	8.2683	114.656	62.1296	69.8309	174.3496
R170	8.4347	124.3847	55.9818	30.8075	175.8796
I171	8.378	124.9741	58.6353	38.5731	174.6134
P172			63.1651	32.1664	176.4775
A173	8.434	124.8201	52.3577	19.3526	177.8139
K174	8.4199	121.2263	56.1759	33.1368	176.583
T175	8.3007	118.9332	59.9941	69.831	172.3448
P176					
P177			62.7124	31.9722	176.2313
A178	8.4352	125.8728	50.4163	17.9936	175.686
P179			62.9068	32.1664	176.8996
K180	8.5401	122.0619	56.3051	33.0078	176.7059
T181	8.2633	118.4655	59.8648	69.831	172.3623
P182					
P183			63.1007	32.1017	177.0754
S184	8.5358	116.4096	58.3114	63.8769	174.8595
S185	8.489	117.9162	58.5704	64.0066	174.9124
G186	8.4219	110.7527	45.045		173.734
E187	8.2737	121.956	54.2993	29.8367	174.2617
P188					
P189			62.9063	32.102	177.0754
K190	8.5838	122.1717	56.2407	33.0724	176.9347
S191	8.4282	116.9468	58.5705	64.0066	175.1585
G192	8.5307	111.0094	45.3037		173.9451
D193	8.2666	120.5381	54.5581	41.3563	176.7237
R194	8.5361	122.118	56.1113	30.2894	176.7236
S195	8.4613	117.1102	59.3473	63.8769	175.2465
G196	8.4819	110.738	45.3684		173.91
Y197	8.0647	120.2939	58.0528	39.0263	175.8092
S198	8.1898	118.2996	57.9233	64.136	173.6637
S199	8.3612	119.0354	56.4997	63.3595	172.9251
P200			63.812	32.0369	177.5502
G201	8.5007	109.4658	45.1095		173.9451
S202	8.1952	116.8948	56.6289	63.3593	172.8548
P203			63.8124	32.037	177.5326
G204	8.5022	109.205	45.1095		174.0507
T205	8.1058	115.6827	59.8648	69.8309	173.2416
P206			64.0066	32.0369	177.7788
G207	8.6089	109.597	45.3683		174.4902
S208	8.1949	115.6692	58.6999	64.0066	174.9299
R209	8.4547	122.9237	56.2407	30.7427	176.4775
S210	8.3339	116.9015	58.441	63.8771	174.4902
R211	8.4701	123.1867	56.1113	30.8719	176.2664
T212	8.3225	118.4205	59.994	69.7017	172.7844
P213			63.2301	32.2309	176.7941
S214	8.4717	116.8544	58.2468	63.8772	174.1033
L215	8.3661	125.6663	53.0049	41.8739	175.2289
P216			62.9711	32.0369	176.8292
T217	8.3929	118.1406	60.0582	69.8317	172.5734
P218					
P219			63.036	32.1018	177.0754
T220	8.3108	115.1538	62.0003	69.8956	174.4551
R221	8.466	123.8584	55.7878	30.9363	175.8619
E222	8.5066	124.0727	54.4286	29.7716	174.455
P223			63.1005	32.1664	176.7765
K224	8.4548	121.9517	56.3702	33.0076	176.6533
K225	8.4543	124.0479	56.1759	33.2018	176.4247
V226	8.2714	122.7682	62.0651	32.943	175.6685
A227	8.4655	128.8576	52.2928	19.2879	177.4095

V228	8.2503	121.2002	62.324	32.8783	176.0905
V229	8.3999	126.0711	62.324	32.7489	175.9499
R230	8.6018	126.4909	55.8524	30.9369	176.0554
T231	8.3515	119.3021	59.8649	69.7015	172.3272
P232					
P233			62.8417	32.1017	176.8116
K234	8.4884	121.8663	56.2407	33.137	176.6358
S235	8.4899	118.9402	56.3701	63.3593	172.9075
P236			63.3594	32.1015	177.2161
S237	8.4776	115.9901	58.5704	63.8772	174.8596
S238	8.3802	118.0954	58.4411	64.0065	174.4727
A239	8.3430	126.0403	52.9402	19.0938	178.0426
K240	8.2934	120.2493	56.5645	32.9429	176.9699
S241	8.2789	116.79	58.5705	63.878	174.6837
R242	8.4019	123.0573	56.3701	30.6774	176.2664
L243	8.2496	122.8412	55.3345	42.3267	177.3041
Q244	8.4227	121.543	55.9169	29.5131	175.985
T245	8.1993	115.7601	61.6768	70.0254	173.4175
A246	8.0579	131.8882	54.0404	20.1292	182.5622

Bibliography

[1] M. S. Ackerman and D. Shortle. Molecular alignment of denatured states of staphylococcal nuclease with strained polyacrylamide gels and surfactant liquid crystalline phases. *Biochemistry*, 41(9):3089–3095, 2002.

[2] M. S. Ackerman and D. Shortle. Robustness of the long-range structure in denatured staphylococcal nuclease to changes in amino acid sequence. *Biochemistry*, 41(46):13791–13797, 2002.

[3] A. Adler, N. Greenfield, and G. Fasman. Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol.*, 27:675–735, 1973.

[4] A. Alexandrescu and D. Shortle. Backbone dynamics of a highly disordered 131 residue fragment of staphylococcal nuclease. *J. Mol. Biol.*, 242(4):527–546, 1994.

[5] F. H.-T. Allain, C. C. Gubser, P. W. Howe, K. Nagai, D. Neuhaus, and G. Varani. Specificity of ribonucleoprotein interaction determined by RNA folding during complex formation. *Nature*, 380(6575):646–650, 1996.

[6] A. d. C. Alonso, T. Zaidi, M. Novak, I. Grundke-Iqbal, and K. Iqbal. Hyperphosphorylation induces self-assembly of τ into tangles of paired helical filaments/straight filaments. *Proc. Natl. Acad. Sci. USA*, 98(12):6923–6928, 2001.

[7] A. S. Altieri, D. P. Hinton, and R. A. Byrd. Association of biomolecular systems via pulsed field gradient NMR self-diffusion measurements. *J. Am. Chem. Soc.*, 117(28):7566–7567, 1995.

[8] P. Amayed, D. Pantaloni, and M.-F. Carlier. The effect of stathmin phosphorylation on microtubule assembly depends on tubulin critical concentration. *J. Biol. Chem.*, 277(25):22718–22724, 2002.

[9] L. Amniai, P. Barbier, A. Sillen, J.-M. Wieruszeski, V. Peyrot, G. Lippens, and I. Landrieu. Alzheimer disease specific phosphoepitopes of tau interfere with assembly of tubulin but not binding to microtubules. *FASEB J.*, 23(4):1146–1152, 2009.

[10] N. Appel, T. Pietschmann, and R. Bartenschlager. Mutational analysis of hepatitis C virus nonstructural protein 5A: Potential role of differential phosphorylation in RNA replication and identification of a genetically flexible domain. *J. Virol.*, 79(5):3187–3194, 2005.

[11] M. R. Arkin and J. A. Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.*, 3(4):301–317, 2004.

[12] F. Avbelj and R. L. Baldwin. Origin of the neighboring residue effect on peptide backbone conformation. *Proc. Natl. Acad. Sci. USA*, 101(30):10967–10972, 2004.

[13] Y. Bai, J. Chung, H. J. Dyson, and P. E. Wright. Structural and dynamic characterization of an unfolded state of poplar apo-plastocyanin formed under nondenaturing conditions. *Protein Sci.*, 10(5):1056–1066, 2001.

[14] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database : its relevance to human molecular medical research. 75(5):312–316, 1997.

[15] A. J. Baldwin, S. J. Anthony-Cahill, T. P. Knowles, G. Lippens, J. Christodoulou, P. D. Barker, and C. M. Dobson. Measurement of amyloid fibril length distributions by inclusion of rotational motion in solution NMR diffusion measurements. *Angew. Chem. Int. Ed. Engl.*, 47(18):3385–3387, 2008.

[16] A. J. Baldwin, J. Christodoulou, P. D. Barker, C. M. Dobson, and G. Lip-

pens. Contribution of rotational diffusion to pulsed field gradient diffusion measurements. *J. Chem. Phys.*, 127(11):114505, 2007.

[17] R. L. Baldwin and B. H. Zimm. Are denatured proteins ever random coils? *Proc. Natl. Acad. Sci. USA*, 97(23):12391–12392, 2000.

[18] W. C. Barker, J. S. Garavelli, D. H. Haft, L. T. Hunt, C. R. Marzec, B. C. Orcutt, G. Y. Srinivasarao, L.-S. L. Yeh, R. S. Ledley, H.-W. Mewes, F. Pfeiffer, and A. Tsugita. The PIR-international protein sequence database. *Nucleic Acids Res.*, 26(1):27–32, 1998.

[19] I. Baskakov and D. W. Bolen. Forcing thermodynamically unfolded proteins to fold. *J. Biol. Chem.*, 273(9):4831–4834, 1998.

[20] I. Baskakov, A. Wang, and D. Bolen. Trimethylamine-N-oxide counteracts urea effects on rabbit muscle lactate dehydrogenase function: A test of the counteraction hypothesis. *Biophys. J.*, 74(5):2666–2673, 1998.

[21] I. V. Baskakov, R. Kumar, G. Srinivasan, Y.-s. Ji, D. W. Bolen, and E. B. Thompson. Trimethylamine N-oxide-induced cooperative folding of an intrinsically unfolded transcription-activating fragment of human glucocorticoid receptor. *J. Biol. Chem.*, 274(16):10693–10696, 1999.

[22] J. Baum, C. M. Dobson, P. A. Evans, and C. Hanley. Characterization of a partly folded protein by NMR methods: studies on the molten globule state of guinea pig α-lactalbumin. *Biochemistry*, 28(1):7–13, 1989.

[23] A. Bax. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.*, 12(1):1–16, 2003.

[24] A. Bax and R. Freeman. Investigation of complex networks of spin-spin coupling by two-dimensional NMR. *J. Magn. Reson.*, 44(3):542–561, 1981.

[25] A. Bax and M. Ikura. An efficient 3D NMR technique for correlating the proton and ^{15}N backbone amide resonances with the α-carbon of the preceding residue in uniformly ^{15}N/^{13}C enriched proteins. *J. Biomol. NMR*, 1(1):99–104, 1991.

[26] L. D. Belmont and T. J. Mitchison. Identification of a protein that interacts with tubulin dimers and increases the catastrophe rate of microtubules. *Cell*, 84(4):623–631, 1996.

[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[28] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA*, 102(47):17002–17007, 2005.

[29] C. W. Bertoncini, R. M. Rasia, G. R. Lamberto, A. Binolfi, M. Zweckstetter, C. Griesinger, and C. O. Fernandez. Structural characterization of the intrinsically unfolded protein β-synuclein, a natural negative regulator of α-synuclein aggregation. *J. Mol. Biol.*, 372(3):708–722, 2007.

[30] E. A. Bienkiewicz, J. N. Adkins, and K. J. Lumb. Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27^{Kip1}. *Biochemistry*, 41(3):752–759, 2002.

[31] J. Biernat, N. Gustke, D. G., E.-M. Mandelkow, and E. Mandelkow. Phosphorylation of ser^{262} strongly reduces binding of tau to microtubules : distinction between PHF-like immunoreactivity and microtubule binding. *Neuron*, 11(1):153–163, 1993.

[32] A. Bloomer, J. Champness, G. Bricogne, R. Staden, and A. Klug. Protein disk of tobacco mosaic virus at 2.8Å resolution showing the interactions within and between subunits. *Nature*, 276(5686):362–368, 1978.

[33] W. Bode, P. Schwager, and R. Huber. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding : The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9Å resolution. *J. Mol. Biol.*, 118(1):99–112, 1978.

[34] G. Bodenhausen and D. J. Ruben. Natural abundance nitrogen-15 NMR

by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.*, 69(1):185–189, 1980.

[35] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280(1):1–9, 1998.

[36] D. Bolen and I. V. Baskakov. The osmophobic effect: natural selection of a thermodynamic force in protein folding. *J. Mol. Biol.*, 310(5):955–963, 2001.

[37] A. M. Bonvin, R. Boelens, and R. Kaptein. NMR analysis of protein interactions. *Curr. Opin. Chem. Biol.*, 9(5):501–508, 2005.

[38] H. S. Bose, R. M. Whittal, M. A. Baldwin, and W. L. Miller. The active form of the steroidogenic acute regulatory protein, StAR, appears to be a molten globule. *Proc. Natl. Acad. Sci. USA*, 96(13):7250–7255, 1999.

[39] J. M. Bourhis, B. Canard, and S. Longhi. Predicting protein disorder and induced folding: From theoretical principles to practical applications. *Curr. Protein Pept. Sci.*, 8(2):135–149, 2007.

[40] J.-M. Bourhis, V. Receveur-Bréchot, M. Oglesbee, X. Zhang, M. Buccellato, H. Darbon, B. Canard, S. Finet, and S. Longhi. The intrinsically disordered C-terminal domain of the measles virus nucleoprotein interacts with the C-terminal domain of the phosphoprotein via two distinct sites and remains predominantly unfolded. *Protein Sci.*, 14(8):1975–1992, 2005.

[41] C. Bracken. NMR spin relaxation methods for characterization of disorder and folding in proteins. *J. Mol. Graph. Model.*, 19(1):3–12, 2001.

[42] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.*, 14(5):570–576, 2004.

[43] D. Braun, G. Wider, and K. Wüthrich. Sequence-corrected 15N "random coil" chemical shifts. *J. Am. Chem. Soc.*, 116(19):8466–8469, 1994.

[44] C. J. Brown, S. Takayama, A. M. Campen, P. Vise, T. W. Marshall, C. J. Oldfield, C. J. Williams, and A. K. Dunker. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.*, 55(1):104–110, 2002.

[45] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones. Protein structure prediction servers at University College London. *Nucleic Acids Res.*, 33:w36–W38, 2005.

[46] M. Buck. Crystallography: Embracing conformational flexibility in proteins. *Structure*, 11(7):735–736, 2003.

[47] M. Buck, H. Schwalbe, and C. M. Dobson. Characterization of conformational preferences in a partly folded protein by heteronuclear NMR spectroscopy: Assignment and secondary structure analysis of hen egg-white lysozyme in trifluoroethanol. *Biochemistry*, 34(40):13219–13232, 1995.

[48] M. Buck, H. Schwalbe, and C. M. Dobson. Main-chain dynamics of a partially folded protein: 15N NMR relaxation measurements of hen egg white lysozyme denatured in trifluoroethanol. *J. Mol. Biol.*, 257(3):669–683, 1996.

[49] A. V. Buevich, U. P. Shinde, M. Inouye, and J. Baum. Backbone dynamics of the natively unfolded pro-peptide of subtilisin by heteronuclear NMR relaxation studies. *J. Biomol. NMR*, 20(3):233–249, 2001.

[50] A. Bulashevska and R. Eils. Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered. *J. Theor. Biol.*, 254(4):799–803, 2008.

[51] A. Bundi and K. Wüthrich. ^{1}H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers*, 18(2):285–297, 1979.

[52] L. Busby. Gist: A scientific graphics package for python. In *Conference: 4. international Python workshop*, 1996.

[53] G. W. Bushnell, G. V. Louie, and G. D. Brayer. High-resolution three-dimensional structure of horse heart cytochrome c. *J. Mol. Biol.*, 214(2):585–595, 1990.

[54] K. Butner and M. Kirschner. Tau protein binds to microtubules through a flexible array of distributed weak sites. *J. Cell Biol.*, 115(3):717–730, 1999.

[55] V. Bychkova, R. Berni, G. L. Rossi, V. Kutyshenko, and O. Ptitsyn.

Retinol-binding protein is in the molten globule state at low pH. *Biochemistry*, 31(33):7566–7571, 1992.

[56] V. E. Bychkova, A. E. Dujsekina, A. Fantuzzi, O. B. Ptitsyn, and G.-L. Rossi. Release of retinol and denaturation of its plasma carrier, retinol-binding protein. *Fold. Design*, 3(4):285–291, 1998.

[57] V. E. Bychkova, R. H. Pain, and O. B. Ptitsyn. The 'molten globule' state is involved in the translocation of proteins across membranes? *FEBS Lett.*, 238(2):231–234, 1988.

[58] V. Bystrov. Spin-spin coupling and the conformational states of peptide systems. *Prog. NMR Spectrosc.*, 10(2):41–82, 1976.

[59] K. Cai, R. Langen, W. L. Hubbell, and H. G. Khorana. Structure and function in rhodopsin: Topology of the C-terminal polypeptide chain in relation to the cytoplasmicloops. *Proc. Natl. Acad. Sci. USA*, 94(26):14267–14272, 1997.

[60] I. Callebaut, G. Labesse, P. Durand, A. Poupon, L. Canard, J. Chomilier, B. Henrissat, and J. P. Mornon. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol. Life Sci.*, 53(8):621–645, 1997.

[61] A. Cammers-Goodwin, T. J. Allen, S. L. Oslick, K. F. McClure, J. H. Lee, and D. Kemp. Mechanism of stabilization of helical conformations of polypeptides by water containing trifluoroethanol. *J. Am. Chem. Soc.*, 118(13):3082–3090, 1996.

[62] W. Cao, C. Bracken, N. R. Kallenbach, and M. Lu. Helix formation and the unfolded state of a 52-residue helical protein. *Protein Sci.*, 13(1):177–189, 2004.

[63] A. S. Carroll, D. E. Gilbert, X. Liu, J. W. Cheung, J. E. Michnowicz, G. Wagner, T. E. Ellenberger, and T. K. Blackwell. Skn-1.

[64] L. B. Casabianca and A. C. de Dios. Ab initio calculations of NMR chemical shifts. *J. Chem. Phys.*, 128(5):052201.1–052201.10, 2008.

[65] L. Cassimeris. The oncoprotein 18/stathmin family of microtubule destabilizers. *Curr. Opin. Cell Biol.*, 14(1):18–24, 2002.

[66] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA*, 104(23):9615–9620, 2007.

[67] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, N. J. Skelton, and M. Rance. *Protein NMR Spectroscopy: Principles and Practice (Second Edition)*. Academic Press, second edition edition, 2007.

[68] J. Cheng, M. J. Sweredoski, and P. Baldi. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery*, 11(3):213–222, 2005.

[69] W.-Y. Choy and J. D. Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.*, 308(5):1011–1032, 2001.

[70] C. L. Chyan, C. Wormald, C. M. Dobson, P. A. Evans, and J. Baum. Structure and stability of the molten globule state of guinea pig α-lactalbumin: A hydrogen exchange study. *Biochemistry*, 32(21):5681–5691, 1993.

[71] M.-J. Clément, I. Jourdain, S. Lachkar, P. Savarin, B. Gigant, M. Knossow, F. Toma, A. Sobel, and P. A. Curmi. N-terminal stathmin-like peptides bind tubulin and impede microtubule assembly. *Biochemistry*, 44(44):14616–14625, 2005.

[72] D. W. Cleveland, S.-Y. Hwo, and M. W. Kirschner. Physical and chemical properties of purified tau factor and the role of tau in microtubule assembly. *J. Mol. Biol.*, 116(2):227–247, 1977.

[73] G. M. Clore, A. Szabo, A. Bax, L. E. Kay, P. C. Driscoll, and A. M. Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *J. Am. Chem. Soc.*, 112(12):4989–4991, 1990.

[74] R. T. Clubb, V. Thanabal, and G. Wagner. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue $^{1}H^{n}$, ^{15}N, and $^{13}C'$

chemical shifts in ^{15}N-^{13}C-labelled proteins. *J. Magn. Reson.*, 97(1):213–217, 1992.

[75] K. Coeytaux and A. Poupon. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, 21(9):1891–1900, 2005.

[76] G. O. Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11(8):1425–1433, 2001.

[77] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, 13(3):289–302, 1999.

[78] M. S. Cortese, J. P. Baird, V. N. Uversky, and A. K. Dunker. Uncovering the unfoldome: Enriching cell extracts for unstructured proteins by acid treatment. *J. Proteome Res.*, 4(5):1610–1618, 2005.

[79] T. Craig, T. Veenstra, S. Naylor, A. Tomlinson, K. Johnson, S. Macura, N. Juranić, and R. Kumar. Zinc binding properties of the DNA binding domain of the 1,25-dihydroxyvitamin D3 receptor. *Biochemistry*, 36(34):10482–10491, 1997.

[80] V. Csizmók, M. Bokor, P. Bánki, E. Klement, K. F. Medzihradszky, P. Friedrich, K. Tompa, and P. Tompa. Primary contact sites in intrinsically unstructured proteins: The case of calpastatin and microtubule-associated protein 2. *Biochemistry*, 44(10):3955–3964, 2005.

[81] P. A. Curmi, S. r. S. Andersen, S. Lachkar, O. Gavet, E. Karsenti, M. Knossow, and A. Sobel. The stathmin/tubulin interaction [**?**]. *J. Biol. Chem.*, 272(40):25029–25036, 1997.

[82] J. Danielsson, J. Jarvet, P. Damberg, and A. Gräslund. Translational diffusion measured by PFG-NMR on full length and fragments of the alzheimer Aβ(1-40) peptide. Determination of hydrodynamic radii of random coil peptides of varying length. *Magn. Reson. Chem.*, 40(13):S89–S97, 2002.

[83] G. W. Daughdrill, L. J. Hanely, and F. W. Dahlquist. The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations. *Biochemistry*, 37(4):1076–1082, 1998.

[84] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker. *Protein Folding Handbook*, chapter Natively disordered proteins. Wiley-VCH Verlag GmbH & Co, 2005.

[85] D. C. David, R. Layfield, L. Serpell, Y. Narain, M. Goedert, and M. G. Spillantini. Proteasomal degradation of tau protein. *J. Neurochem.*, 83(1):176–185, 2002.

[86] K. J. A. Davies. Degradation of oxidized proteins by the 20S proteasome. *Biochimie*, 83(3-4):301–310, 2001.

[87] R. Dayanandan, M. Van Slegtenhorst, T. Mack, L. Ko, S.-H. Yen, K. Leroy, J.-P. Brion, B. Anderton, M. Hutton, and S. Lovestone. Mutations in tau reduce its microtubule binding properties in intact cells and affect its phosphorylation. *FEBS Lett.*, 446(2-3):228–232, 1999.

[88] K. T. Dayie, G. Wagner, and J.-F. o. Lefèvre. Theory and practice of nuclear spin relaxation in proteins. *Annu. Rev. Phys. Chem.*, 47:243–282, 1996.

[89] M. M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in α-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.*, 172(2):476–477, 2005.

[90] M. M. Dedmon, C. N. Patel, G. B. Young, and G. J. Pielak. FlgM gains structure in living cells. *Proc. Natl. Acad. Sci. USA*, 99(20):12681–12684, 2002.

[91] A. Dehner and H. Kessler. Diffusion NMR spectroscopy: Folding and aggregation of domains in p53. *ChemBioChem*, 6(9):1550–1565, 2005.

[92] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, 6(3):277–293, 1995.

[93] E. D. Demaine, S. Langerman, and J. ORourke. Geometric restrictions on producible polygonal protein chains. *Algorithmica*, 44(2):167–181, 2006.

[94] S. J. Demarest, S.-Q. Zhou, J. Robblee, R. Fairman, B. Chu, and D. P. Raleigh. A comparative study of peptide models of the α-domain of α-lactalbumin, lysozyme, and α-lactalbumin/lysozyme chimeras allows the elucidation of critical factors that contribute to the ability to form stable partially folded states. *Biochemistry*, 40(7):2138–2147, 2001.

[95] G. Di Paolo, B. Antonsson, D. Kassel, B. M. Riederer, and G. Grenningloh. Phosphorylation regulates the microtubule-destabilizing activity of stathmin and its interaction with tubulin. *FEBS Lett.*, 416(2):149–152, 1997.

[96] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

[97] K. A. Dill and D. Shortle. Denatured states of proteins. *Annu. Rev. Biochem.*, 60:795–825, 1991.

[98] D. Dolgikh, R. Gilmanshin, E. Brazhnikov, V. Bychkova, G. Semisotnov, S. Venyaminov, and O. Ptitsyn. Alpha-Lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett.*, 136(2):311–315, 1981.

[99] A. Dorléans, B. Gigant, R. B. Ravelli, P. Mailliet, V. Mikol, and M. Knossow. Variations in the colchicine-binding domain provide insight into the structural switch of tubulin. *Proc. Natl. Acad. Sci. USA*, 106(33):13775–13779, 2009.

[100] Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.

[101] Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, 347(4):827–839, 2005.

[102] Z. Dosztányi, M. Sandor, P. Tompa, and I. Simon. Prediction of protein disorder at the domain level. *Curr. Protein Pept. Sci.*, 8(2):161–171, 2007.

[103] P. Douzou and G. Petsko. Proteins at work: "stop-action" pictures at subzero temperatures. *Adv. Protein Chem.*, 36:245–361, 1984.

[104] A. N. Drozdov, A. Grossfield, and R. V. Pappu. Role of solvent in determining conformational preferences of alanine dipeptide in water. *J. Am. Chem. Soc.*, 126(8):2574–2581, 2004.

[105] L. D'Silva, P. Ozdowy, M. Krajewski, U. Rothweiler, M. Singh, and T. A. Holak. Monitoring the effects of antagonists on protein-protein interactions with nmr spectroscopy. *J. Am. Chem. Soc.*, 127(38):13220–13226, 2005.

[106] C. Dumanchin, A. Camuzat, D. Campion, P. Verpillat, D. Hannequin, B. Dubois, P. Saugier-Veber, C. Martin, C. Penet, F. Charbonnier, Y. Agid, T. Frebourg, and A. Brice. Segregation of a missense mutation in the microtubule-associated protein tau gene with familial frontotemporal dementia and parkinsonism. *Hum. Mol. Genet.*, 7(11):1825–1829, 1998.

[107] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002.

[108] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered proteins. *J. Mol. Graph. Model.*, 19(1):26–59, 2001.

[109] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown. Intrinsic protein disorder in complete genomes. In *Genome Informatics 11*, pages 161–171, 2000.

[110] A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. Walton Chen, V. Vacic, Z. Obradovic, and V. N. Uversky. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9:S1, 2008.

[111] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethu-

raman, S. Weng, D. Botstein, and J. M. Cherry. *Saccharomyces* genome database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucl. Acids Res.*, 30(1):69–72, 2002.

[112] H. J. Dyson and P. E. Wright. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Meth. Enzymol.*, 339:258–270, 2001.

[113] H. J. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by NMR. *Chem. Rev.*, 104(8):3607–3622, 2004.

[114] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, 2005.

[115] A. Einstein. On a heuristic point of view about the creation and conversion of light. *Ann. Physik*, 17:132–148, 1905.

[116] D. Eliezer. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 19(1):23–30, 2009.

[117] D. Eliezer, P. Barré, M. Kobaslija, D. Chan, X. Li, and L. Heend. Residual structure in the repeat domain of tau: Echoes of microtubule binding and paired helical filament formation. *Biochemistry*, 44(3):1026–1036, 2005.

[118] D. Eliezer, K. Chiba, H. Tsuruta, S. Doniach, K. O. Hodgson, and H. Kihara. Evidence of an associative intermediate on the myoglobin refolding pathway. *Biophys. J.*, 65(2):912–917, 1993.

[119] D. Eliezer, J. Yao, H. J. Dyson, and P. E. Wright. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nat. Struct. Biol.*, 5(2):148–155, 1998.

[120] R. J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26(10):597–604, 2001.

[121] J. Emsley and J. Lindon. *NMR spectroscopy using liquid crystal solvents*. Oxford; New York: Pergamon Press, 1975.

[122] M. J. Evans, C. M. Rice, and S. P. Goff. Phosphorylation of hepatitis C virus nonstructural protein 5A modulates its protein interactions and viral RNA replication. *Proc. Natl. Acad. Sci. USA*, 101(35):13038–13043, 2004.

[123] P. Fan, C. Bracken, and J. Baum. Structural characterization of monellin in the alcohol-denatured state by NMR: Evidence for β-sheet to α-helix conversion. *Biochemistry*, 32(6):1573–1582, 1993.

[124] B. Farmer, II, R. Venters, L. Spicer, M. Wittekind, and L. Müller. A refocused and optimized HNCA: Increased sensitivity and resolution in large macromolecules. *J. Biomol. NMR*, 2(2):195–202, 1992.

[125] N. A. Farrow, R. Muhandiram, A. U. Singer, S. M. Pascal, C. M. Kay, G. Gish, S. E. Shoelson, T. Pawson, J. D. Forman-Kay, and L. E. Kay. Backbone dynamics of a free and a phosphopeptide-complexed src homology 2 domain studied by 15N NMR relaxation. *Biochemistry*, 33(19):5984–6003, 1994.

[126] N. A. Farrow, O. Zhang, J. D. Forman-Kay, and L. E. Kay. A heteronuclear correlation experiment for simultaneous determination of ^{15}N longitudinal decay and chemical exchange rates of systems in slow equilibrium. *J. Biomol. NMR*, 4(5):727–734, 1994.

[127] N. A. Farrow, O. Zhang, A. Szabo, D. A. Torchia, and L. E. Kay. Spectral density function mapping using ^{15}N relaxation data exclusively. *J. Biomol. NMR*, 6(2):153–162, 1995.

[128] G. D. Fasman. *Circular Dichroism and the Conformational Analysis of Biomolecules*. Plenum Press, New York, 1996.

[129] L. Feigin and D. Svergun. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*. Plenum Press, New York, 1987.

[130] F. Fernandes, D. S. Poole, S. Hoover, R. Middleton, A.-C. Andrei, J. Gerstner, and R. Striker. Sensitivity of hepatitis C virus to cyclosporine A depends on nonstructural proteins NS5A and NS5B. *Hepatology*, 46(4):1026–1033, 2007.

[131] J. C. Ferreon and V. J. Hilser. The effect of the polyproline II (PPII) conformation on the denatured state entropy. *Protein Sci.*, 12(3):447–457, 2003.

[132] F. o. Ferron, S. Longhi, B. Canard, and D. Karlin. A practical overview of protein disorder prediction methods. *Proteins*, 65(1):1–14, 2006.

[133] K. M. Fiebig, H. Schwalbe, M. Buck, L. J. Smith, and C. M. Dobson. Toward a description of the conformations of denatured states of proteins. comparison of a random coil model with NMR measurements. *J. Phys. Chem.*, 100(7):2661–2666, 1996.

[134] E. Fischer. Einfluss der configuration auf die wirkung der enzyme. *Ber. Deutsch. Chem. Ges.*, 27(3):2985–2993, 1894.

[135] G. Fischer, H. Bang, and C. Mech. Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides. *Biomed. Biochim. Acta*, 43(10):1101–1111, 1984.

[136] G. Fischer, T. Tradler, and T. Zarnt. The mode of action of peptidyl prolyl cis/trans isomerases in vivo: binding vs. catalysis. *FEBS Lett.*, 426(1):17–20, 1998.

[137] N. C. Fitzkee and G. D. Rose. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. USA*, 101(34):12497–12502, 2004.

[138] S. L. Flaugh and K. J. Lumb. Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27^{Kip1}. *Biomacromolecules*, 2(2):538–540, 2001.

[139] P. J. Flory. *Statistical Mechanics Of Chain Molecules*. Wiley, New York, 1969.

[140] A. Fontana, P. Polverino de Laureto, V. De Filippis, E. Scaramella, and M. Zambonin. Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.*, 2(2):R17–R26, 1997.

[141] T. Förster. Intermolecular energy migration and fluorescence. *Ann. Phys.*, 437(1-2):55–75, 1948.

[142] K. Fredriksson, M. Louhivuori, P. Permi, and A. Annila. On the interpretation of residual dipolar couplings as reporters of molecular dynamics. *J. Am. Chem. Soc.*, 126(39):12646–12650, 2004.

[143] T. Frenkiel, C. Bauer, M. Carr, B. Birdsall, and J. Feeney. HMQC-NOESY-HMQC, a three-dimensional NMR experiment which allows detection of nuclear overhauser effects between protons with overlapping signals. *J. Magn. Reson.*, 90(2):420–425, 1990.

[144] I. Furó and S. V. Dvinskikh. NMR methods applied to anisotropic diffusion. *Magn. Reson. Chem.*, 40(13):S3–S14, 2002.

[145] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.*, 338(5):1015–1026, 2004.

[146] C. A. Galea, A. A. High, J. C. Obenauer, A. Mishra, C.-G. Park, M. Punta, A. Schlessinger, J. Ma, B. Rost, C. A. Slaughter, and R. W. Kriwacki. Large-scale analysis of thermostable, mammalian proteins provides insights into the intrinsically disordered proteome. *J. Proteome Res.*, 8(1):211–226, 2009.

[147] C. A. Galea, Y. Wang, S. G. Sivakolundu, and R. W. Kriwacki. Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits. *Biochemistry*, 47(29):7598–7609, 2008.

[148] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, 22(23):2948–2949, 2006.

[149] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. Expected packing density allows prediction of both amyloidogenic and disordered regions in protein chains. *J. Phys.: Condens. Matter*, 19(28):285225.1–285225.15, 2007.

[150] S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya. To be folded or to be unfolded? *Protein Sci.*, 13(11):2871–2877, 2004.

[151] A. E. Garcia. Characterization of non-alpha helical conformations in ala peptides. *Polymer*, 45(2):669–676, 2004.

[152] E. Garner, P. Cannon, P. Romero, Z. Obradovic, and A. K. Dunker. Predicting disordered regions from amino acid sequence: Common themes despite

differing structural characterization. In *Genome Informatics 9*, volume 9, pages 201–213, 1998.

[153] K. Gast, H. Damaschun, K. Eckert, K. Schulze-Forster, H. R. Maurer, M. Mueller-Frohne, D. Zirwer, J. Czarnecki, and G. Damaschun. Prothymosin α: A biologically active protein with random coil conformationt. *Biochemistry*, 34(40):13211–13218, 1995.

[154] B. Gigant, C. Wang, R. B. Ravelli, F. Roussi, M. O. Steinmetz, P. A. Curmi, A. Sobel, and M. Knossow. Structural basis for the regulation of tubulin by vinblastine. *Nature*, 435(7041):519–522, 2005.

[155] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.*, 268(1):158–169, 1997.

[156] J. R. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. II. distance restraints from paramagnetic relaxation and calculation of an ensemble of structures. *J. Mol. Biol.*, 268(1):170–184, 1997.

[157] O. Glatter and O. Kratky. *Small angle X-ray scattering*. Academic Press, London, 1982.

[158] J. Glushka, M. Lee, S. Coffin, and D. Cowburn. Nitrogen-15 chemical shifts of backbone amides in bovine pancreatic trypsin inhibitor and apamin. *J. Am. Chem. Soc.*, 111(20):7716–7722, 1989.

[159] T. Goddard and D. Kneller. Sparky 3. University of California, San Francisco, 1989.

[160] A. P. Golovanov, R. T. Blankley, J. M. Avis, and W. Bermel. Isotopically discriminated NMR spectroscopy: A tool for investigating complex protein interactions in vitro. *J. Am. Chem. Soc.*, 129(20):6528–6535, 2007.

[161] B. Goode, P. Denis, D. Panda, M. Radeke, H. Miller, L. Wilson, and S. Feinstein. Functional interactions between the proline-rich and repeat regions of tau enhance microtubule binding and assembly. *Mol. Biol. Cell*, 8(2):353–365, 1997.

[162] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe. Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study. *J. Am. Chem. Soc.*, 129(5):1179–1189, 2007.

[163] M. J. Grey, Y. Tang, E. Alexov, C. J. McKnight, D. P. Raleigh, and A. G. Palmer III. Characterizing a partially folded intermediate of the villin headpiece domain under non-denaturing conditions: Contribution of His41 to the pH-dependent stability of the N-terminal subdomain. *J. Mol. Biol.*, 355(5):1078–1094, 2006.

[164] K.-H. Großand H. R. Kalbitzer. Distribution of chemical shifts in ^{1}H nuclear magnetic resonance spectra of proteins. *J. Magn. Reson.*, 76(1):87–99, 1988.

[165] S. Grzesiek, J. Anglister, and A. Bax. Correlation of backbone amide and aliphatic side-chain resonances in ^{13}C/^{15}N-enriched proteins by isotropic mixing of ^{13}C magnetization. *J. Magn. Reson. B*, 101(1):114–119, 1993.

[166] S. Grzesiek and A. Bax. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.*, 114(16):6291–6293, 1992.

[167] S. Grzesiek and A. Bax. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.*, 99(1):201–207, 1992.

[168] S. Grzesiek and A. Bax. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.*, 96(2):432–440, 1992.

[169] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu. Tight regulation of unstructured proteins: From transcript synthesis to protein degradation. *Science*, 322(5906):1365–1368, 2008.

[170] K. Gunasekaran, C.-J. Tsai, S. Kumar, D. Zanuy, and R. Nussinov. Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.*, 28(2):81–85, 2003.

[171] O. Gursky and D. Atkinson. Thermal unfolding of human high-density

apolipoprotein A-1: implications for a lipid-free molten globular state. *PNAS*, 93(7):2991–2995, 1996.

[172] H. Gutowsky, D. McCall, and C. Slichter. Nuclear magnetic resonance multiplets in liquids. *J. Chem. Phys.*, 21(2):279–292, 1953.

[173] E. Hahn and D. Maxwell. Spin echo measurements of nuclear spin coupling in molecules. *Phys. Rev.*, 88(5):1070–1084, 1952.

[174] B. Halle. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA*, 99(3):1274–1279, 2002.

[175] P. Hammarström and U. Carlsson. Is the unfolded state the rosetta stone of the protein folding problem? *Biochem. Biophys. Res. Commun.*, 276(2):393–398, 2000.

[176] X. Hanoulle, A. Badillo, J.-M. Wieruszeski, D. Verdegem, I. Landrieu, R. Bartenschlager, F. Penin, and G. Lippens. Hepatitis C virus NS5A protein is a substrate for the peptidyl-prolyl cis/trans isomerase activity of cyclophilins A and B. *J. Biol. Chem.*, 284(20):13589–13601, 2009.

[177] M. R. Hansen, L. Mueller, and A. Pardi. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Biol.*, 5(12):1065–1074, 1998.

[178] M. Hasegawa, M. Morishima-Kawashima, K. Takio, M. Suzuki, K. Titani, and Y. Ihara. Protein sequence and mass spectrometric analyses of tau in the Alzheimer's disease brain. *J. Biol. Chem.*, 267(24):17047–17054, 1992.

[179] B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, and A. K. Dunker. Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, 19(8):929–949, 2009.

[180] H. C. Hemmings, Jr., A. C. Nairn, D. W. Aswad, and P. Greengard. DARPP-32, a dopamine- and adenosine 3':5'-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. II. purification and characterization of the phosphoprotein from bovine caudate nucleus. *J. Neurosci.*, 4(1):99–110, 1984.

[181] M. Hennig, W. Bermel, A. Spencer, C. M. Dobson, L. J. Smith, and H. Schwalbe. Side-chain conformations in an unfolded protein: $\chi 1$ distributions in denatured hen lysozyme determined by heteronuclear ^{13}C, ^{15}N NMR spectroscopy. *J. Mol. Biol.*, 288(4):705–723, 1999.

[182] S. Honnappa, W. Jahnke, J. Seelig, and M. O. Steinmetz. Control of intrinsically disordered stathmin by multisite phosphorylation. *J. Biol. Chem.*, 281(23):16078–16083, 2006.

[183] S. B. Horwitz, H.-J. Shen, L. He, P. Dittmar, R. Neef, J. Chen, and U. K. Schubart. The microtubule-destabilizing activity of metablastin (p19) is controlled by phosphorylation. *J. Biol. Chem.*, 272(13):8129–8132, 1997.

[184] M. Hoshino, H. Katou, Y. Hagihara, K. Hasegawa, H. Naiki, and Y. Goto. Mapping the core of the bold β2-microglobulin amyloid fibril by h/d exchange. *Nat. Struct. Biol.*, 9(5):332–336, 2002.

[185] B. Howell, N. Larsson, M. Gullberg, and L. Cassimeris. Dissociation of the tubulin-sequestering and microtubule catastrophe-promoting activities of Oncoprotein 18/stathmin. *Mol. Biol. Cell*, 10(1):105–118, 1999.

[186] Q.-X. Hua, W.-h. Jia, B. P. Bullock, J. F. Habener, and M. A. Weiss. Transcriptional activator-coactivator recognition: Nascent folding of a kinase-inducible transactivation domain predicts its structure on coactivator binding. *Biochemistry*, 37(17):5858–5866.

[187] L. Huang, J. Hwang, S. D. Sharma, M. R. Hargittai, Y. Chen, J. J. Arnold, K. D. Raney, and C. E. Cameron. Hepatitis C virus nonstructural protein 5A (NS5A) is an RNA-binding protein. *J. Biol. Chem.*, 280(43):36417–36428, 2005.

[188] W. L. Hubbell, A. Gross, R. Langen, and M. A. Lietzow. Recent advances in site-directed spin labeling of proteins. *Curr. Opin. Struct. Biol.*, 8(5):649–656, 1998.

[189] R. Huber and W. Bennett. Conformational flexibility and its functional significance in some proteins. *Trends Biochem. Sci.*, 4:271–283, 1979.

[190] R. Huber and W. S. Bennett, Jr. Functional significance of flexibility in proteins. *Biopolymers*, 22(1):261–279, 1983.

[191] W. E. Hull and B. D. Sykes. Dipolar nuclear spin relaxation of ^{19}F in multispin systems. application to ^{19}F labeled proteins. *J. Chem. Phys.*, 63(2):867–880, 1975.

[192] C. A. Hunter, M. J. Packer, and C. Zonta. From structure to chemical shift and vice-versa. *Prog. NMR Spectrosc.*, 47(1-2):27–39, 2005.

[193] L. M. Iakoucheva, C. J. Brown, J. Lawson, Z. Obradović, and A. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 323(3):573–584, 2002.

[194] L. M. Iakoucheva, P. Radivojac, C. J. Brown, T. R. OConnor, J. G. Sikes, Z. Obradovic, and A. K. Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, 32(3):1037–1049, 2004.

[195] M. Ikura, A. Bax, G. M. Clore, and A. M. Gronenborn. Detection of nuclear Overhauser effects between degenerate amide proton resonances by heteronuclear three-dimensional NMR spectroscopy. *J. Am. Chem. Soc.*, 112(24):9020–9022, 1990.

[196] T. Ishida and K. Kinoshita. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.*, 35:W460–W464, 2007.

[197] R. Ishima and K. Nagayama. Quasi-spectral-density function analysis for nitrogen-15 nuclei in proteins. *J. Magn. Reson. B*, 108(1):73–76, 1995.

[198] D. M. Jacobs, A. S. Lipton, N. G. Isern, G. W. Daughdrill, D. F. Lowry, X. Gomes, and M. S. Wold. Human replication protein A: Global fold of the N-terminal RPA-70 domain reveals a basic cleft and flexible C-terminal linker. *J. Biomol. NMR*, 14(4):321–331, 1999.

[199] M. F. Jeng, S. W. Englander, G. A. Elove, H. Roder, and A. J. Wand. Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry*, 29(46):10433–10437, 1990.

[200] M. R. Jensen and M. Blackledge. On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J. Am. Chem. Soc.*, 130(34):11266–11267, 2008.

[201] M. R. Jensen, P. R. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, and M. Blackledge. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure*, 17(9):1169–1185, 2009.

[202] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. USA*, 102(37):13099–13104, 2005.

[203] A. K. Jha, A. Colubri, M. H. Zaman, S. Koide, T. R. Sosnick, and K. F. Freed. Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry*, 44(28):9691–9702, 2005.

[204] M. Jimenez, J. Nieto, M. Rico, J. Santoro, J. Herranz, and F. Bermejo. A study of the NH NMR signals of Gly-Gly-X-Ala tetrapeptides in H_2O at low temperature. *J. Mol. Struct.*, 143(1-2):435–438, 1986.

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

[206] D. T. Jones and J. J. Ward. Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, 53:573–578, 2003.

[207] E. Josefsson, D. OConnell, T. J. Foster, I. Durussel, and J. A. Cox. The binding of calcium to the B-repeat segment of SdrD, a cell surface protein of *Staphylococcus aureus*. *J. Biol. Chem.*, 273(47):31145–31152, 1998.

[208] L. Jourdain, P. Curmi, A. Sobel, D. Pantaloni, and M.-F. Carlier. Stathmin: A tubulin-sequestering protein which forms a ternary T_2S complex with two tubulin molecules. *Biochemistry*, 36(36):10817–10821, 1997.

[209] S. Kar, K. Sakaguchi, Y. Shimohigashi, S. Samaddar, R. Banerjee, G. Basu, V. Swaminathan, T. K. Kundu, and S. Roy. Effect of phosphorylation on

the structure and fold of transactivation domain of p53. *J. Biol. Chem.*, 277(18):15579–15585, 2002.

[210] D. Karlin, S. Longhi, V. Receveur, and B. Canard. The N-terminal domain of the phosphoprotein of morbilliviruses belongs to the natively unfolded class of proteins. *Virology*, 296(2):251–262, 2002.

[211] M. Karplus. Contact electron-spin coupling of nuclear magnetic moments. *J. Chem. Phys.*, 30(1):11–15, 1959.

[212] M. Kataoka, Y. Hagihara, K. Mihara, and Y. Goto. Molten globule of cytochrome c studied by small angle X-ray scattering. *J. Mol. Biol.*, 229(3):591–596, 1993.

[213] M. Kataoka, F. Tokunaga, K. Kuwajima, and Y. Goto. Structural characterization of the molten globule of α-lactalbumin by solution X-ray scattering. *Protein Sci.*, 6(2):422–430, 1997.

[214] L. E. Kay. Protein dynamics from NMR. *Nat. Struct. Biol.*, 5(7):513–517, 1998.

[215] L. E. Kay, M. Ikura, R. Tschudin, and A. Bax. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.*, 89(3):496–514, 1990.

[216] M. S. Z. Kellermayer, S. B. Smith, C. Bustamante, and H. L. Granzier. Complete unfolding of the titin molecule under external force. *J. Struct. Biol.*, 122(1-2):197–205, 1998.

[217] A. Kentsis, M. Mezei, T. Gindin, and R. Osman. Unfolded state of polyalanine is a segmented polyproline II helix. *Proteins Struct. Funct. Bioinform.*, 55(3):493–501, 2004.

[218] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA*, 101(34):12491–12496, 2004.

[219] D. M. Korzhnev, K. Kloiber, V. Kanelis, V. Tugarinov, and L. E. Kay. Probing slow dynamics in high molecular weight proteins by methyl-TROSY NMR spectroscopy: Application to a 723-residue enzyme. *J. Am. Chem. Soc.*, 126(12):3964–3973, 2004.

[220] D. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44(2):98–104, 1958.

[221] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. USA*, 93(21):11504–11509, 1996.

[222] A. Kryshtafovych, K. Fidelis, and J. Moult. Progress from CASP6 to CASP7. *Proteins*, 69(S8):194–207.

[223] W. Kuhn. über die gestalt fadenförmiger moleküle in lösungen. *Colloid & Polymer Science*, 68(1):2–15, 1934.

[224] S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips Jr. Dynamics of proteins in crystals: Comparison of experiment with simple models. *Biophys. J.*, 83(2):723–732, 2002.

[225] T. Küntziger, O. Gavet, A. Sobel, and M. Bornens. Differential effect of two stathmin/Op18 phosphorylation mutants on *Xenopus* embryo development. *J. Biol. Chem.*, 276(25):22979–22984, 2001.

[226] K. Kuwajima. A folding model of alpha-lactalbumin deduced from the three-state denaturation mechanism. *J. Mol. Biol.*, 114(2):241–258, 1977.

[227] S. Labeit and B. Kolmerer. Titins: Giant proteins in charge of muscle ultrastructure and elasticity. *Science*, 270(5234):293–296, 1995.

[228] J. Lackowicz. *Principles of Fluorescence Spectroscopy*. Kluwer Academic/Plenum Publishers, second edition, 1999.

[229] E. R. Lacy, I. Filippov, W. S. Lewis, S. Otieno, L. Xiao, S. Weiss, L. Hengst, and R. W. Kriwacki. p27 binds cyclinCDK complexes through a sequential

mechanism involving binding-induced protein folding. *Nat. Struct. Mol. Biol.*, 11(4):358–364, 2004.

[230] I. Landrieu, L. Lacosse, A. Leroy, J.-M. Wieruszeski, X. Trivelli, A. Sillen, N. Sibille, H. Schwalbe, K. Saxena, T. Langer, and G. Lippens. NMR analysis of a tau phosphorylation pattern. *J. Am. Chem. Soc.*, 128(11):3575–3583, 2006.

[231] N. E. LaPointe, G. Morfini, G. Pigino, I. N. Gaisina, A. P. Kozikowski, L. I. Binder, and S. T. Brady. The amino terminus of tau inhibits kinesin-dependent axonal transport: Implications for filament toxicity. *J. Neurosci. Res.*, 87(2):440–451, 2009.

[232] N. Larsson, U. Marklund, H. M. Gradin, G. Brattsand, and M. Gullberg. Control of microtubule dynamics by oncoprotein 18: dissection of the regulatory role of multisite phosphorylation during mitosis. *Mol. Cell. Biol.*, 17(9):5530–5539, 1997.

[233] J.-F. Lefèvre, K. Dayie, J. Peng, and G. Wagner. Internal mobility in the partially folded DNA binding and dimerization domains of GAL4: NMR analysis of the N-H spectral density functions. *Biochemistry*, 35(8):2674–2686, 1996.

[234] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys. PCB*, 65:44–45, 1968.

[235] M. H. Levitt. *Spin Dynamics*. John Wiley & Sons Ltd, 2001.

[236] X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic. Predicting protein disorder for N-, C- and internal regions. In *Genome Informatics 10*, pages 30–40, 1999.

[237] M. A. Lietzow, M. Jamin, H. J. Dyson, and P. E. Wright. Mapping long-range contacts in a highly unfolded protein. *J. Mol. Biol.*, 322(4):655–662, 2002.

[238] P. Lieutaud, B. Canard, and S. Longhi. MeDor: a metaserver for predicting protein disorder. *BMC Genomics*, 9:S25.1–S25.5, 2008.

[239] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. Protein disorder prediction: Implications for structural proteomics. *Structure*, 11(11):1453–1459, 2003.

[240] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson. Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, 31(13):3701–3708, 2003.

[241] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme A binding protein. *J. Am. Chem. Soc.*, 126(10):3291–3299, 2004.

[242] G. Lindwall and R. D. Cole. Phosphorylation affects the ability of tau protein to promote microtubule assembly. *J. Biol Chem.*, 259(8):5301–5305, 1984.

[243] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *J. Am. Chem. Soc.*, 104(17):4546–4559, 1982.

[244] G. Lippens, J.-M. Wieruszeski, A. Leroy, C. Smet, A. Sillen, L. Buée, and I. Landrieu. Proline-directed random-coil chemical shift values as a tool for the NMR assignment of the tau phosphorylation sites. *ChemBioChem*, 5(1):73–78, 2004.

[245] A. Liu, R. Riek, G. Wider, C. von Schroetter, R. Zahn, and K. Wüthrich. NMR experiments for resonance assignments of ^{13}c, ^{15}n doubly-labeled flexible polypeptides: Application to the human prion protein hPrP(23-230). *J. Biomol. NMR*, 16(2):127–138, 2000.

[246] J. Liu and B. Rost. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.*, 31(13):3833–3835, 2003.

[247] T. M. Logan, Y. Thériault, and S. W. Fesik. Structural characterization of the FK506 binding protein unfolded in urea and guanidine hydrochloride. *J. Mol. Biol.*, 236(2):637–648, 1994.

[248] F. Löhr and H. Rüterjans. A new triple-resonance experiment for the sequential assignment of backbone resonances in proteins. *J. Biomol. NMR*, 6(2):189–197, 1995.

[249] S. Longhi, V. Receveur-Bréchot, D. Karlin, K. Johansson, H. Darbon, D. Bhella, R. Yeo, S. Finet, and B. Canard. The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J. Biol. Chem.*, 278(20):18638–18648, 2003.

[250] R. Loris, I. Marianovsky, J. Lah, T. Laeremans, H. Engelberg-Kulka, G. Glaser, S. Muyldermans, and L. Wyns. Crystal structure of the intrinsically flexible addiction antidote MazE. *J. Biol. Chem.*, 278(30):28252–28257, 2003.

[251] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, and A. Annila. On the origin of residual dipolar couplings from denatured proteins. *J. Am. Chem. Soc.*, 125(50):15647–15650, 2003.

[252] S. C. Lovell, I. W. Davis, W. B. Arendall III, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by cα geometry: ϕ,ψ and cβ deviation. *Proteins*, 50(3):437–450, 2003.

[253] D. F. Lowry, A. Stancik, R. M. Shrestha, and G. W. Daughdrill. Modeling the accessible conformations of the intrinsically unstructured transactivation domain of p53. *Proteins Struct. Funct. Bioinform.*, 71(2):587–598, 2008.

[254] K. P. Lu, G. Finn, T. H. Lee, and L. K. Nicholson. Prolyl cis-trans isomerization as a molecular timer. *Nat. Chem. Biol.*, 3(10):619–629, 2007.

[255] J. Luban, K. L. Bossol, E. K. Franke, G. V. Kalpana, and S. P. Goff. Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell*, 73(6):1067–1078, 1993.

[256] T. Lührs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert, and R. Riek. 3D structure of [a.

[257] P. Luo and R. L. Baldwin. Mechanism of helix induction by trifluoroethanol: A framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. *Biochemistry*, 36(27):8413–8421, 1997.

[258] Y. Luo and R. L. Baldwin. Trifluoroethanol stabilizes the pH 4 folding intermediate of sperm whale apomyoglobin. *J. Mol. Biol.*, 279(1):49–57, 1998.

[259] A. N. Lupas and M. Gruber. The structure of alpha-helical coiled coils. *Adv. Protein Chem.*, 70:37–78, 2005.

[260] R. M. MacCallum. Order/disorder prediction with self organising maps. available from: http://www.forcasp.org/paper2127.html.

[261] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res.*, 33:D54–D58, 2005.

[262] E. Mandelkow, J. Biernat, G. Drewes, N. Gustke, B. Trinczek, and E. Mandelkow. Tau domains, phosphorylation, and interactions with microtubules. *Neurobiol. Aging*, 16(3):355–263, 1995.

[263] T. Manna, D. Thrower, H. P. Miller, P. Curmi, and L. Wilson. Stathmin strongly increases the minus end catastrophe frequency and induces rapid treadmilling of bovine brain microtubules at steady state *in Vitro*. *J. Biol. Chem.*, 281(4):2071–2078, 2006.

[264] W.-Y. Mark, J. C. Liao, Y. Lu, A. Ayed, R. Laister, B. Szymczyna, A. Chakrabartty, and C. H. Arrowsmith. Characterization of segments from the central region of BRCA1: An intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? *J. Mol. Biol.*, 345(2):275–287, 2005.

[265] J. A. Marsh, V. K. Singh, Z. Jia, and J. D. Forman-Kay. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. *Protein Sci.*, 15(12):2795–2804, 2006.

[266] S. B. Marston and C. S. Redwood. The molecular anatomy of caldesmon. *Biochem. J.*, 279:1, 1991.

[267] E. R. McCarney, J. E. Kohn, and K. W. Plaxco. Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins. *Crit. Rev. Biochem. Mol. Biol.*, 40(4):181–189, 2005.

[268] H. M. McConnell. Reaction rates by nuclear magnetic resonance. *J. Chem. Phys.*, 28(3):430–431, 1958.

[269] S. Meier, M. Blackledge, and S. Grzesiek. Conformational distributions of unfolded polypeptides from novel NMR techniques. *J. Chem. Phys.*, 128(5):052204.1–052204.14, 2008.

[270] S. Meier, D. Häussinger, P. Jensen, M. Rogowski, and S. Grzesiek. High-accuracy residual $^{1}H^{n}$-^{13}C and $^{1}H^{n}$-$^{1}H^{n}$ dipolar couplings in perdeuterated proteins. *J. Am. Chem. Soc.*, 125(1):44–45, 2003.

[271] S. Meier, M. Strohmeier, M. Blackledge, and S. Grzesiek. Direct observation of dipolar couplings and hydrogen bonds across a β-hairpin in 8 M urea. *J. Am. Chem. Soc.*, 129(4):754–755, 2007.

[272] J. Meiler. PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 26(1):25–37, 2003.

[273] A. Merrill, F. Cohen, and W. Cramer. On the nature of the structural change of the colicin E1 channel peptide necessary for its translocation-competent state. *Biochemistry*, 29(24):5829–5836, 1990.

[274] G. Merutka, H. J. Dyson, and P. E. Wright. 'random coil' ^{1}H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J. Biomol. NMR*, 5(1):14–24, 1995.

[275] M. Mezei, P. J. Fleming, R. Srinivasan, and G. D. Rose. Polyproline II helix is the preferred conformation for unfolded polyalanine in water. *Proteins*, 55(3):502–507, 2004.

[276] R. Mohana-Borges, N. K. Goto, G. J. Kroon, H. J. Dyson, and P. E. Wright. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.*, 340(5):1131–1142, 2004.

[277] Y.-K. Mok, C. M. Kay, L. E. Kay, and J. Forman-Kay. NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.*, 289(3):619–638, 1999.

[278] D. Moradpour, F. Penin, and C. M. Rice. Replication of hepatitis C virus. *Nat Rev. Microbiol.*, 5(6):453–463, 2007.

[279] A. S. Morar, A. Olteanu, G. B. Young, and G. J. Pielak. Solvent-induced collapse of α-synuclein and acid-denatured cytochrome c. *Protein Sci.*, 10(11):2195–2199, 2001.

[280] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton. Stereochemical quality of protein structure coordinates. *Proteins*, 12(4):345–364, 1992.

[281] Y. Mu and G. Stock. Conformational dynamics of trialanine in water: A molecular dynamics study. *J. Phys. Chem. B*, 106(20):5294–5301, 2002.

[282] D. R. Muhandiram and L. E. Kay. Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. *J. Magn. Reson. B*, 103(3):203–216, 1994.

[283] R. Mukhopadhyay and J. H. Hoh. AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. *FEBS Lett.*, 505(3):374–378, 2001.

[284] M. D. Mukrasch, S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow, and M. Zweckstetter. Structural polymorphism of 441-residue tau at single residue resolution. *PLoS Biol.*, 7(2):399–414, 2009.

[285] M. D. Mukrasch, J. Biernat, M. von Bergen, C. Griesinger, E. Mandelkow, and M. Zweckstetter. Sites of tau important for aggregation populate β-structure and bind to microtubules and polyanions. *J. Biol. Chem.*, 280(26):24978–24986, 2005.

[286] P. Neddermann, M. Quintavalle, C. Di Pietro, A. Clementi, M. Cerretani, S. Altamura, L. Bartholomew, and R. De Francesco. Reduction of hepatitis C virus NS5A hyperphosphorylation by selective inhibition of cellular kinases activates viral RNA replication in cell culture. *J. Virol.*, 78(23):13306–13314, 2004.

[287] P. Neudecker, A. Zarrine-Afsar, W.-Y. Choy, D. R. Muhandiram, A. R. Davidson, and L. E. Kay. Identification of a collapsed intermediate with

non-native long-range interactions on the folding pathway of a pair of Fyn SH3 domain mutants by NMR relaxation dispersion spectroscopy. *J. Mol. Biol.*, 363(5):958–976, 2006.

[288] O. Obolensky, K. Schlepckow, H. Schwalbe, and A. Solovyov. Theoretical framework for NMR residual dipolar couplings in unfolded proteins. *J. Biomol. NMR*, 39(1):1–16, 2007.

[289] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, 61:176–182, 2005.

[290] F. Ochsenbein, R. Guerois, J.-M. Neumann, A. Sanson, E. Guittet, and C. van Heijenoort. 15N NMR relaxation as a probe for helical intrinsic propensity: The case of the unfolded D2 domain of annexin I. *J. Biomol. NMR*, 19(1):3–18, 2001.

[291] M. Ohgushi and A. Wada. "molten-globule state": a compact form of globular proteins with mobile side-chains. *FEBS Lett.*, 164(1), 1983.

[292] N. Olashaw, T. Bagui, and W. Pledger. Cell cycle control: a complex issue. *Cell Cycle*, 3(3):263–264, 2004.

[293] S. B. Ozkan, G. A. Wu, J. D. Chodera, and K. A. Dill. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA*, 104(29):11987–11992, 2007.

[294] A. G. Palmer III. Probing molecular motion by NMR. *Curr. Opin. Struct. Biol.*, 7(5):732–737, 1997.

[295] A. G. Palmer III, J. Cavanagh, P. E. Wright, and M. Rance. Sensitivity improvement in proton-detected two-dimensional heteronuclear correlation NMR spectroscopy. *J. Magn. Reson.*, 93(1):151–170, 1991.

[296] H. Pan, G. Barany, and C. Woodward. Reduced BPTI is collapsed. a pulsed field gradient NMR study of unfolded and partially folded bovine pancreatic trypsin inhibitor. *Protein Science*, 6(9):1985–1992, 1997.

[297] R. V. Pappu and G. D. Rose. A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.*, 11(10):2437–2455, 2002.

[298] A. Pardi, M. Billeter, and K. Wthrich. Calibration of the angular dependence of the amide proton-cα proton coupling constants, $^3j_{HN\alpha}$, in a globular protein: Use of $^3j_{HN\alpha}$ for identification of helical secondary structure. *J. Mol. Biol.*, 180(3):741–751, 1984.

[299] A. Pastore and V. Saudek. The relationship between chemical shift and secondary structure in proteins. *J. Magn. Reson.*, 90(1):165–176, 1990.

[300] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7:176–182, 2006.

[301] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.*, 3(1):35–60, 2005.

[302] C. J. Penkett, C. Redfield, I. Dodd, J. Hubbard, D. L. McBay, D. E. Mossakowska, R. A. G. Smith, C. M. Dobson, and L. J. Smith. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J. Mol. Biol.*, 274(2):152–159, 1997.

[303] S. E. Permyakov, I. S. Millett, S. Doniach, E. A. Permyakov, and V. N. Uversky. Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins Struct. Funct. Bioinform.*, 53(4):855–862, 2003.

[304] M. Persson, P. Hammarström, M. Lindgren, B.-H. Jonsson, M. Svensson, and U. Carlsson. EPR mapping of interactions between spin-labeled variants of human carbonic anhydrase II and GroEL: Evidence for increased flexibility of the hydrophobic core by the interaction. *Biochemistry*, 38(1):432–441, 1999.

[305] W. Peti, L. J. Smith, C. Redfield, and H. Schwalbe. Chemical shifts in denatured proteins: Resonance assignments for denatured ubiquitin and comparisons with other denatured proteins. *J. Biomol. NMR*, 19(2):153–165, 2001.

[306] K. W. Plaxco, C. J. Morton, S. B. Grimshaw, J. A. Jones, M. Pitkeathly, I. D. Campbell, and C. M. Dobson. The effects of guanidine hydrochloride on the 'random coil' conformations and NMR chemical shifts of the peptide series GGXGG. *J. Biomol. NMR*, 10(3):221–230, 1997.

[307] J. Prestegard, H. Al-Hashimi, and J. Tolman. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Quart. Rev. Biophys.*, 33(4):371–424, 2000.

[308] U. Preuss, J. Biernat, E. Mandelkow, and E. Mandelkow. The 'jaws' model of tau-microtubule interaction examined in CHO cells. *J. Cell Sci.*, 110(6):789–800, 1997.

[309] J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman. Foldindex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16):3435–3438, 2005.

[310] O. Ptitsyn. Molten globule and protein folding. *Adv. Protein Chem.*, 47:83–229, 1995.

[311] Y. Qu and D. Bolen. Efficacy of macromolecular crowding in forcing proteins to fold. *Biophys. Chem.*, 101-102:155–165, 2002.

[312] A. Quintas, M. J. a. M. Saraiva, and R. M. Brito. The tetrameric protein transthyretin dissociates to a non-native monomer in solution. a novel model for amyloidogenesis. *J. Biol. Chem.*, 274(46):32943–32949, 1999.

[313] R. Rademakers and C. Cruts, M. van Broeckhoven. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. *Hum. Mutat.*, 24(4):277–295.

[314] P. Radivojac, Z. Obradovic, C. J. Brown, and A. K. Dunker. Improving sequence alignments for intrinsically disordered proteins. In *Biocomputing 2002: Proceedings of the Pacific Symposium*, pages 589–600, 2002.

[315] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker. Protein flexibility and intrinsic disorder. *Protein Sci.*, 13(1):71–80, 2004.

[316] D. Rajamani, S. Thiel, S. Vajda, and C. J. Camacho. Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 101(31):11287–11292, 2004.

[317] R. C. Ramachandran, G.N. and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963.

[318] M. Rance, G. Wagner, O. Sørensen, K. Wüthrich, and R. Ernst. Application of ω1-decoupled 2D correlation spectra to the study of proteins. *J. Magn. Reson.*, 59(2):250–261, 1984.

[319] R. B. Ravelli, B. Gigant, P. A. Curmi, I. Jourdain, S. Lachkar, A. Sobel, and M. Knossow. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature*, 428(6979):198–202, 2004.

[320] V. Redeker, S. Lachkar, S. Siavoshian, E. Charbaut, J. Rossier, A. Sobel, and P. A. Curmi. Probing the native structure of stathmin and its interaction domains with tubulin. Combined use of limited proteolysis, size exclusion chromatography, and mass spectrometry. *J. Biol. Chem.*, 275(10):6841–6849, 2000.

[321] U. Reimer, G. Scherer, M. Drewello, S. Kruber, M. Schutkowski, and G. Fischer. Side-chain effects on peptidyl-prolyl cis/trans isomerisation. *J. Mol. Biol.*, 279(2):449–460, 1998.

[322] J. P. Richards, H. P. Bächinger, R. H. Goodman, and R. G. Brennan. Analysis of the structural properties of cAMP-responsive element-binding protein (CREB) and phosphorylated CREB. *J. Biol. Chem.*, 271(23):13716–13723, 1996.

[323] R. Richarz and K. Wüthrich. Carbon-13 NMR chemical shifts of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers*, 17(9):2133–2141, 1978.

[324] C. Rischel and F. M. Poulsen. Modification of a specific tyrosine enables

tracing of the end-to-end distance during apomyoglobin folding. *FEBS Lett.*, 374(1):105–109, 1995.

[325] P. Rizzu, J. C. Van Swieten, M. Joosse, M. Hasegawa, M. Stevens, A. Tibben, M. F. Niermeijer, M. Hillebrand, R. Ravid, B. A. Oostra, M. Goedert, C. M. van Duijn, and P. Heutink. High prevalence of mutations in the microtubule-associated protein tau in a population study of frontotemporal dementia in the netherlands. *Am. J. Hum. Genet*, 64(2):414–421, 1999.

[326] P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, and A. K. Dunker. Identifying disordered regions in proteins from amino acid sequence. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 90–95, 1997.

[327] P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot, and A. K. Dunker. Thousands of proteins likely to have long disordered regions. In *Biocomputing 1998: Proceedings of the Pacific Symposium*, pages 437–448, 1998.

[328] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins*, 42(1):38–48, 2001.

[329] C. I. Rubin and G. F. Atweh. The role of stathmin in the regulation of the cell cycle. *J. Cell. Biochem.*, 93(2):242–250, 2004.

[330] A. L. Rucker and T. P. Creamer. Polyproline II helical structure in protein unfolded states: Lysine peptides revisited. *Protein Sci.*, 11(4):980–985, 2002.

[331] P. Salamon and A. Konopka. A maximum entropy principle for the distribution of local complexity in naturally occurring nucleotide sequences. *Computers Chem.*, 16(2):117–124, 1992.

[332] P. Salamon, J. Wootton, A. Konopka, and L. Hansen. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. *Computers Chem.*, 17(2):135–148, 1993.

[333] J. Santoro and G. C. King. A constant-time 2D overbodenhausen experiment for inverse correlation of isotopically enriched species. *J. Magn. Reson.*, 97(1):202–207, 1992.

[334] H.-J. Sass, G. Musco, S. J. Stahl, P. T. Wingfield, and S. Grzesiek. Solution NMR of proteins within polyacrylamide gels: Diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J. Biomol. NMR*, 18(4):303–309, 2000.

[335] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost. Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE*, 4(2):e4433, 2009.

[336] S. L. Schreiber and G. R. Crabtree. The mechanism of action of cyclosporin A and FK506. *Immunol. Today*, 13(4):136–142, 1992.

[337] E. Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Phys. Rev.*, 28(6):1049–1070, 1926.

[338] H. Schwalbe, K. M. Fiebig, M. Buck, J. A. Jones, S. B. Grimshaw, A. Spencer, S. J. Glaser, L. J. Smith, and C. M. Dobson. Structural and dynamical properties of a denatured protein. heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea. *Biochemistry*, 36(29):8977–8991, 1997.

[339] S. Schwarzinger, G. J. Kroon, T. R. Foss, J. Chung, P. E. Wright, and H. J. Dyson. Sequence-dependent correction of random coil NMR chemical shifts. *J. Am. Chem. Soc.*, 123(13):2970–2978, 2001.

[340] S. Schwarzinger, G. J. Kroon, T. R. Foss, P. E. Wright, and H. J. Dyson. Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView. *J. Biomol. NMR*, 18(1):43–48, 2000.

[341] O. Schweers, E. Schönbrunn-Hanebeck, A. Marx, and E. Mandelkow. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J. Biol. Chem.*, 269(39):24290–24297, 1994.

[342] S. K. Seeley, R. M. Weis, and L. K. Thompson. The cytoplasmic fragment of the aspartate receptor displays globally dynamic behavior. *Biochemistry*, 35(16):5199–5206, 1996.

[343] G. Semisotnov, H. Kihara, N. Kotova, K. Kimura, Y. Amemiya, K. Wakabayashi, I. Serdyuk, A. Timchenko, K. Chiba, K. Nikaido, T. Ikura, and K. Kuwajima. Protein globularization during folding. a study by synchrotron small-angle x-ray scattering. *J. Mol. Biol.*, 262:559–574, 1996.

[344] L. Serrano. Comparison between the ϕ distribution of the amino acids in the protein database and NMR data indicates that amino acids have various ϕ propensities in the random coil conformation. *J. Mol. Biol.*, 254(2):322–333, 1995.

[345] R. J. Sheaff, J. D. Singer, J. Swanger, M. Smitherman, J. M. Roberts, and B. E. Clurman. Proteasomal turnover of $p21^{Cip1}$ does not require $p21^{Cip1}$ ubiquitination. *Mol. Cell*, 5(2):403–410, 2000.

[346] Y. Shen and A. Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38(4):289–302, 2007.

[347] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker, and A. Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA*, 105(12):4685–4690, 2008.

[348] Y. Shen, R. Vernon, D. Baker, and A. Bax. De novo protein structure generation from incomplete chemical shift assignments. *J. Biomol. NMR*, 43(2):63–78, 2009.

[349] Z. Shi, A. Olson, G. D. Rose, R. L. Baldwin, and N. R. Kallenbach. Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl. Acad. Sci. USA*, 99(14):9190–9195, 2002.

[350] Z. Shi, R. Woody, and N. Kallenbach. Is polyproline II a major backbone conformation in unfolded proteins? *Adv. Protein Chem.*, 62:163–240, 2002.

[351] I. Shimada. NMR techniques for identifying the interface of a larger protein-protein complex: Cross-saturation and transferred cross-saturation experiments. *Methods Enzymol.*, 394:483–506, 2005.

[352] S. Shimizu and H. S. Chan. Origins of protein denatured state compactness and hydrophobic clustering in aqueous urea: inferences from nonpolar potentials of mean force. *Proteins*, 49(4):560–566, 2002.

[353] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, 97(16):8868–8873, 2000.

[354] D. Shortle and M. S. Ackerman. Persistence of native-like topology in a denatured protein in 8 M urea. *Science*, 293(5529):487–489, 2001.

[355] D. R. Shortle. Structural analysis of non-native states of proteins by NMR methods. *Curr. Opin. Struct. Biol.*, 6(1):24–30, 1996.

[356] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274(5292):1531–1534, 1996.

[357] N. Sibille, A. Sillen, A. Leroy, J.-M. Wieruszeski, B. Mulloy, I. Landrieu, and G. Lippens. Structural impact of heparin binding to full-length tau as studied by NMR spectroscopy. *Biochemistry*, 45(41):12560–12572, 2006.

[358] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker. DisProt: the database of disordered proteins. *Nucleic Acids Res.*, 35:D786–D793, 2007.

[359] A. Sillen, P. Barbier, I. Landrieu, S. Lefebvre, J.-M. Wieruszeski, A. Leroy, V. Peyrot, and G. Lippens. NMR investigation of the interaction between the neuronal protein tau and the microtubules. *Biochemistry*, 46(11):3055–3064, 2007.

[360] A. Sillen, A. Leroy, J.-M. Wieruszeski, A. Loyens, J.-C. Beauvillain, L. Buée, I. Landrieu, and G. Lippens. Regions of tau implicated in the paired helical fragment core as defined by NMR. *Chembiochem.*, 6(10):1849–1856, 2005.

[361] A. Sillen, J.-M. Wieruszeski, A. Leroy, A. Ben Younes, I. Landrieu, and

G. Lippens. High-resolution magic angle spinning NMR of the neuronal tau protein integrated in Alzheimer's-like paired helical fragments. *J. Am. Chem. Soc.*, 127(29):10138–10139, 2005.

[362] P. Simmonds, J. Bukh, C. Combet, G. Deléage, N. Enomoto, S. Feinstone, P. Halfon, G. Inchauspé, C. Kuiken, G. Maertens, M. Mizokami, D. G. Murphy, H. Okamoto, J.-M. Pawlotsky, F. Penin, E. Sablon, T. Shin-I, L. J. Stuyver, H.-J. Thiel, S. Viazov, A. J. Weiner, and A. Widell. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*, 42(4):962–973, 2005.

[363] N. Skelton, A. Palmer, M. Akke, J. Kordel, M. Rance, and W. Chazin. Practical aspects of two-dimensional proton-detected ^{15}N spin relaxation measurements. *J. Magn. Reson. B*, 102(3):253–264, 1993.

[364] C. Smet, A. Leroy, A. Sillen, J.-M. Wieruszeski, I. Landrieu, and G. Lippens. Accepting its random coil nature allows a partial NMR assignment of the neuronal tau protein. *ChemBioChem.*, 5(12):1639–1646, 2004.

[365] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, and C. M. Dobson. Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, 255(3):494–506, 1996.

[366] L. J. Smith, K. M. Fiebig, H. Schwalbe, and C. M. Dobson. The concept of a random coil: Residual structure in peptides and denatured proteins. *Fold. Design*, 1(5):R95–R106, 1996.

[367] E. Sontag, V. Nunbhakdi-Craig, G. Lee, R. Brandt, C. Kamibayashi, J. Kuret, C. L. White III, M. C. Mumby, and G. S. Bloom. Molecular interactions among protein phosphatase 2A, tau, and microtubules. *J. Biol. Chem.*, 274(36):25490–25498, 1999.

[368] S. Spera and A. Bax. Emperical correlation between protein backbone conformation and cα and cβ ^{13}c nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc.*, 113(14):5490–5492, 1991.

[369] M. O. Steinmetz, R. A. Kammerer, W. Jahnke, K. N. Goldie, A. Lustig, and J. van Oostrum. Op18/stathmin caps a kinked protofilament-like tubulin tetramer. *EMBO J.*, 19(4):572–580, 2000.

[370] E. Stejskal and J. Tanner. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.*, 42(1):288–292, 1965.

[371] S. Stenholm and K.-A. Suominen. *Quantum Approach to Informatics*. Wiley & Sons, Inc., New York, 2005.

[372] C.-T. Su, C.-Y. Chen, and C.-M. Hsu. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res.*, 35:W465–W472, 2007.

[373] C.-T. Su, C.-Y. Chen, and Y.-Y. Ou. Protein disorder prediction by condensed pssm considering propensity for order or disorder. *BMC Bioinformatics*, 7:319.1–319.16, 2006.

[374] K. Sugase, H. J. Dyson, and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, 447(7147):1021–1025, 2007.

[375] L. Szilágyi and O. Jardetzky. α-proton chemical shifts and secondary structure in proteins. *J. Magn. Reson.*, 83(3):441–449, 1989.

[376] H. Tanaka, T. Yamashita, M. Asada, S. Mizutani, H. Yoshikawa, and M. Tohyama. Cytoplasmic p21$^{Cip1/WAF1}$ regulates neurite remodeling by inhibiting Rho-kinase activity. *J. Cell Biol.*, 158(2):321–329, 2002.

[377] C. Tanford. Protein denaturation. *Adv. Protein Chem.*, 23:121–282, 1968.

[378] J. Tanner. Use of the stimulated echo in NMR diffusion studies. *J. Chem. Phys.*, 52(5):2523–2526, 1970.

[379] O. Tcherkasskaya and O. B. Ptitsyn. Direct energy transfer to study the 3D structure of non-native proteins: AGH complex in molten globule state of apomyoglobin. *Protein Eng.*, 12(6):485–490, 1999.

[380] O. Tcherkasskaya and O. B. Ptitsyn. Molten globule versus variety of inter-

mediates: influence of anions on ph-denatured apomyoglobin. *FEBS Lett.*, 455(3):325–331, 1999.

[381] K. Teilum, B. B. Kragelund, and F. M. Poulsen. Transient structure formation in unfolded acyl-coenzyme A-binding protein observed by site-directed spin labelling. *J. Mol. Biol.*, 324(2):349–357, 2002.

[382] T. L. Tellinghuisen, J. Marcotrigiano, A. E. Gorbalenya, and C. M. Rice. The NS5A protein of hepatitis C virus is a zinc metalloprotein. *J. Biol. Chem.*, 279(47):48576–48587, 2004.

[383] T. L. Tellinghuisen, J. Marcotrigiano, and C. M. Rice. Structure of the zinc-binding domain of an essential component of the hepatitis C virus replicase. *Nature*, 435(7040):374–379, 2005.

[384] V. Thanabal, D. O. Omecinsky, M. D. Reily, and W. L. Cody. The ^{13}C chemical shifts of amino acids in aqueous solution containing organic solvents: Application to the secondary structure characterization of peptides in aqueous trifluoroethanol solution. *J. Biomol. NMR*, 4(1):47–59, 1994.

[385] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[386] M. L. Tiffany and S. Krimm. Circular dichroism of poly-L-proline in an unordered conformation. *Biopolymers*, 6(12):1767–1770, 1968.

[387] N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278(5340):1111–1114, 1997.

[388] M. Tollinger, K. Kloiber, B. Ágoston, C. Dorigoni, R. Lichtenecker, W. Schmid, and R. Konrat. An isolated helix persists in a sparsely populated form of KIX under native conditions. *Biochemistry*, 45(29):8885–8893, 2006.

[389] P. Tompa. Intrinsically unstructured proteins. *Trends Biochem. Sci.*, 27(10):527–533, 2002.

[390] P. Tompa. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, 579(15):3346–3354, 2005.

[391] P. Tompa and P. Csermely. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.*, 18(11):1169–1175, 2004.

[392] P. Tompa, J. Prilusky, I. Silman, and J. L. Sussman. Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins*, 71(2):903–909, 2008.

[393] H. Torrey. Bloch equations with diffusion terms. *Phys. Rev.*, 104(3):563–565, 1956.

[394] K. Trombitás, M. Greaser, S. Labeit, J.-P. Jin, M. Kellermayer, M. Helmes, and H. Granzier. Titin extensibility in situ: Entropic elasticity of permanently folded and permanently unfolded molecular segments. *J. Cell Biol.*, 140(4):853–859, 1998.

[395] V. Tugarinov and L. E. Kay. Quantitative ^{13}C and ^{2}H NMR relaxation studies of the 723-residue enzyme malate synthase G reveal a dynamic binding interface. *Biochemistry*, 44(49):15970–15977, 2005.

[396] R. Tycko, F. J. Blanco, and Y. Ishii. Alignment of biopolymers in strained gels: A new way to create detectable dipole-dipole couplings in high-resolution biomolecular NMR. *J. Am. Chem. Soc.*, 122(38):9340–9341, 2000.

[397] V. N. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, 11(4):739–756, 2002.

[398] V. N. Uversky. What does it mean to be natively unfolded? *Eur. J. Biochem.*, 269(1):2–12, 2002.

[399] V. N. Uversky and A. L. Fink. Do protein molecules have a native-like topology in the pre-molten globule state? *Biochemistry (Mosc)*, 64(5):552–555, 1999.

[400] V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, 41(3):415–427, 2000.

[401] V. N. Uversky, A. S. Karnoup, D. J. Segel, S. Seshadri, S. Doniach, and A. L. Fink. Anion-induced folding of *Staphylococcal* nuclease: characterization of multiple equilibrium partially folded intermediates. *J. Mol. Biol.*, 278(4):879–894, 1998.

[402] V. N. Uversky, C. J. Oldeld, and A. K. Dunker. Intrinsically disordered proteins in human diseases: Introducing the d^2 concept. *Annu. Rev. Biophys.*, 37:215–246, 2008.

[403] V. N. Uversky, N. Y. Protasova, V. V. Rogov, K. S. Vassilenko, A. T. Gudkov, and V. P. Kutyshenko. Circularly permuted dihydrofolate reductase possesses all the properties of the molten globule state, but can resume functional tertiary structure by interaction with its ligands. *Protein Sci.*, 5(9):1844–1851, 1996.

[404] V. N. Uversky and O. B. Ptitsyn. "partly folded" state, a new equilibrium state of protein molecules: Four-state guanidinium chloride-induced unfolding of β-lactamase at low temperature. *Biochemistry*, 33(10):2782–2791, 1994.

[405] V. N. Uversky and O. B. Ptitsyn. Further evidence on the equilibrium "pre-molten globule state": Four-state guanidinium chloride-induced unfolding of carbonic anhydrase B at low temperature. *J. Mol. Biol.*, 255(1):215–228, 1996.

[406] M. Van Hoy, K. K. Leuther, T. Kodadek, and S. A. Johnston. The acidic activation domains of the GCN4 and GAL4 proteins are not α-helical but form β-sheets. *Cell*, 72(4):587–594, 1993.

[407] T. Vavouri, J. I. Semple, R. Garcia-Verdugo, and B. Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138(1):198–208, 2009.

[408] J. A. Vila, H. A. Baldoni, D. R. Ripoll, A. Ghosh, and H. A. Scheraga. Polyproline II helix conformation in a proline-rich environment: A theoretical study. *Biophys. J.*, 86(2):731–742, 2004.

[409] M. von Bergen, P. Friedhoff, J. Biernat, J. Heberle, E.-M. Mandelkow, and E. Mandelkow. Assembly of τ protein into Alzheimer paired helical filaments depends on a local sequence motif (^{306}VQIVYK^{311}) forming β structure. *Proc. Natl. Acad. Sci. USA*, 97(10):5129–5134, 2000.

[410] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinás, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins.*, 59:687–696, 2005.

[411] S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton, and A. K. Dunker. DisProt: a database of protein disorder. *Bioinformatics*, 21(1):137–140, 2005.

[412] L. Vugmeyster and C. J. McKnight. Phosphorylation-induced changes in backbone dynamics of the dematin headpiece C-terminal domain. *J. Biomol. NMR*, 43(1):39–50, 2009.

[413] G. W. Vuister and A. Bax. Resolution enhancement and spectral editing of uniformly ^{13}C-enriched proteins by homonuclear broadband ^{13}C decoupling. *J. Magn. Reson.*, 98(2):428–435, 1992.

[414] G. W. Vuister and A. Bax. Quantitative J correlation: a new approach for measuring homonuclear three-bond $J(H^N H_\alpha)$ coupling constants in ^{15}N-enriched proteins. *J. Am. Chem. Soc.*, 115(17):7772–7777, 1993.

[415] A. Vullo, O. Bortolami, G. Pollastri, and S. C. E. Tosatto. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, 34:W164–W168, 2006.

[416] A. Wang and D. Bolen. A naturally occurring protective system in urea-rich cells: Mechanism of osmolyte protection of proteins against urea denaturation. *Biochemistry*, 36(30):9101–9108, 1997.

[417] A. C. Wang and A. Bax. Reparametrization of the karplus relation for $^3J(H\alpha - N)$ and $^3J(HN - C')$ in peptides from uniformly 13C/15N-enriched human ubiquitin. *J. Am. Chem. Soc.*, 117(6):1810–1813, 1995.

[418] L. Wang and B. R. Donald. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *J. Biomol. NMR*, 29(3):223–242, 2004.

[419] L. Wang and B. R. Donald. A data-driven, systematic search algorithm for structure determination of denatured or disordered proteins. In *Comput. Syst Bioinformatics Conf.*, pages 67–78, 2006.

[420] L. Wang, H. R. Eghbalnia, A. Bahrami, and J. L. Markley. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J. Biomol. NMR*, 32(1):13–22, 2005.

[421] Y. Wang and O. Jardetzky. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci.*, 11(4):852–861, 2002.

[422] Y. Wang and D. S. Wishart. A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins. *J. Biomol. NMR*, 31(2):143–148, 2005.

[423] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20(13):2138–2139, 2004.

[424] J. J. Ward, L. J. Sodhi, Jaspreet S. ; McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, 337(3):635–645, 2004.

[425] K. Watt, T. J. Jess, S. M. Kelly, N. C. Price, and I. J. McEwan. Induced α-helix structure in the aryl hydrocarbon receptor transactivation domain modulates protein-protein interactions. *Biochemistry*, 44(2):734–743, 2005.

[426] P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and P. T. Lansbury, Jr. NACP, a protein implicated in alzheimer's disease and learning, is natively unfolded. *Biochemistry*, 35(43):13709–13715, 1996.

[427] R. Weisemann, H. Rüterjans, and W. Bermel. 3D triple-resonance NMR techniques for the sequential assignment of NH and ^{15}N resonances in ^{15}N- and ^{13}C-labelled proteins. *J. Biomol. NMR*, 3(1):113–120, 1993.

[428] D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, and L. J. Smith. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry*, 38(50):16424–16431, 1999.

[429] R. Williams. The conformational properties of proteins in solution. *Biol. Rev. Camb. Philos Soc.*, 54(4):389–437, 1979.

[430] D. Wishart, B. Sykes, and F. Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.*, 222(2):311–333, 1991.

[431] D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, and B. D. Sykes. ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigation of nearest-neighbour effects. *J. Biomol. NMR*, 5(1):67–81, 1995.

[432] D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, and B. D. Sykes. ^{1}H, ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, 6(2):135–140, 1995.

[433] D. S. Wishart and B. D. Sykes. The ^{13}c Chemical-Shift Index: A simple method for the identification of protein secondary structure using ^{13}c chemical-shift data. *J. Biomol. NMR*, 4(2):852–861, 1994.

[434] D. S. Wishart and B. D. Sykes. Chemical shifts as a tool for structure determination. *Methods Enzymol.*, 239:363–392, 1994.

[435] D. S. Wishart, B. D. Sykes, and F. Richards. The chemical shift index: A fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651, 1992.

[436] G. Witman, D. Cleveland, M. Weingarten, and M. Kirschner. Tubulin requires tau for growth onto microtubule initiating sites. *Proc. Natl. Acad. Sci. USA*, 73(11):4070–4074, 1976.

[437] M. Wittekind and L. Mueller. HNCACB, A high sensitivity 3D NMR experiment to correlate amide proton and nitrogen resonance with the α-carbon and β-carbon resonances in proteins. *J. Magn. Reson. B*, 101(2):201–205, 1993.

[438] T. Wittmann, G. M. Bokoch, and C. M. Waterman-Storer. Regulation of microtubule destabilizing activity of Op18/stathmin downstream of Rac1. *J. Biol. Chem.*, 279(7):6196–6203, 2004.

[439] J. C. Wootton. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, 18(3):269–285, 1994.

[440] J. C. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, 17(2):149–163, 1993.

[441] S. Woutersen and P. Hamm. Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J. Phys. Chem. B*, 104 (47)(47):11316–11320, 2000.

[442] J. O. Wrabl, D. Shortle, and T. B. Woolf. Correlation between changes in nuclear magnetic resonance order parameters and conformational entropy: Molecular dynamics simulations of native and denatured staphylococcal nuclease. *Proteins*, 38(2):123–133, 2000.

[443] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293(2):321–331, 1999.

[444] L. C. Wu, P. B. Laub, G. A. Elove, J. Carey, and H. Roder. A noncovalent peptide complex as a model for an early folding intermediate of cytochrome c. *Biochemistry*, 32(38):10271–10276, 1993.

[445] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic, and V. N. Uversky. Functional anthology of intrinsic disorder. III. ligands, postranslational modifications and diseases associated with intrinsically disordered proteins. *J. Proteome Res.*, 6(5):1917–1932, 2007.

[446] T. Yamazaki, W. Lee, C. H. Arrowsmith, D. Muhandiram, and L. E. Kay. A suite of triple resonance NMR experiments for the backbone assignment of 15N, 13C, 2H labeled proteins with high sensitivity. *J. Am. Chem. Soc.*, 116(26):11655–11666, 1994.

[447] D. Yang and L. E. Kay. Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: Application to protein folding. *J. Mol. Biol.*, 263(2):369–382, 1996.

[448] D. Yang, Y.-K. Mok, J. D. Forman-Kay, N. A. Farrow, and L. E. Kay. Contributions to protein entropy and heat capacity from bond vector motions measured by NMR spin relaxation. *J. Mol. Biol.*, 272(5):790–804, 1997.

[449] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21(16):3369–3376, 2005.

[450] J. Yao, J. Chung, D. Eliezer, P. E. Wright, and H. J. Dyson. NMR structural and dynamic characterization of the acid-unfolded state of apomyoglobin provides insights into the early events in protein folding. *Biochemistry*, 40(12):3561–3571, 2001.

[451] J. Yao, V. A. Feher, B. F. Espejo, M. T. Reymond, P. E. Wright, and H. J. Dyson. Stabilization of a type VI turn in a family of linear peptides in water solution. *J. Mol. Biol.*, 243(4):736–753, 1994.

[452] Q. Yi, M. L. Scalley-Kim, E. J. Alm, and D. Baker. NMR characterization of residual structure in the denatured state of protein L. *J. Mol. Biol.*, 299(5):1341–1351, 2000.

[453] Y. Zhan, X. Song, and G. W. Zhou. Structural analysis of regulatory protein domains using GST-fusion proteins. *Gene*, 281(1-2):1–9, 2001.

[454] H. Zhang, S. Neal, and D. S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, 25(3):173–195, 2003.

[455] J. Zhang and C. R. Matthews. Ligand binding is the principal determinant of stability for the p21^{H-ras} protein. *Biochemistry*, 37(42):14881–14890, 1998.

[456] O. Zhang, J. D. Forman-Kay, D. Shortle, and L. E. Kay. Triple-resonance NOESY-based experiments with improved spectral resolution: Applications

to structural characterization of unfolded, partially folded and folded proteins. *J. Biomol. NMR*, 9(2):181–200, 1997.

[457] O. Zhang, L. E. Kay, J. P. Olivier, and J. D. Forman-Kay. Backbone ^{1}H and ^{15}N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. *J. Biomol. NMR*, 4(6):845–858, 1994.

[458] O. Zhang, L. E. Kay, D. Shortle, and J. D. Forman-Kay. Comprehensive NOE characterization of a partially folded large fragment of staphylococcal nuclease Δ131Δ, using NMR methods with improved resolution. *J. Mol. Biol.*, 272(1):9–20, 1997.

[459] E. R. Zuiderweg. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, 41(1):1–7, 2002.

[460] J. Zurdo, J. M. Sanz, C. González, M. Rico, and J. P. Ballesta. The exchangeable yeast ribosomal acidic protein YP2β shows characteristics of a partly folded state under physiological conditions. *Biochemistry*, 36(31):9625–9635, 1997.

Abstract

Many proteins and protein regions have been shown be intrinsically unstructured/disordered (IUPs) and still carry out diverse and important functions *in vivo*. The technique of choice for studying IUPs is Nuclear Magnetic Resonance (NMR). However, to be able to obtain information of many possible NMR spectra, these must first be assigned. This process is complicated in the case of IUPs by the increased amount of signal overlap. To facilitate the assignment, a graphical semi-automatic assignment tool using the concept of product and sum planes was developed. Using this tool, the study of individual IUPs by NMR became conceivable. A first considered IUP is human Tau. The backbone and C_β resonances have been fully assigned for two Tau fragments (F3 and F5) and partially assigned for full-length Tau P301L. These NMR assignments of Tau could eventually lead to more insight in the structural behaviour of the protein upon its binding to or polymerisation of microtubules, and in its aggregated form which is observed to be one of the hallmarks of Alzheimer's disease. Secondly, the Hepatitis C virus (HCV) non-structural protein 5A (NS5A) was considered. The structural properties of both the second and third domain (out of three) of this protein have been assessed and some residual α-helical structure was observed, which could be indicative of regions prone to interaction with other cellular partners. We have also examined the interaction between both CypA and CypB and the domains D2 and D3 of NS5A, as these PPIases might be involved in HCV replication.

Résumé

De nombreuses protéines ou domaines de protéines sont intrinsèquement non structurés/désordonnés (IUPs), mais possèdent néanmoins des fonctions diverses et importantes *in vivo*. La technique de choix pour étudier les IUPs est la Résonance Magnétique Nucléaire (RMN). Cependant, pour obtenir de l'information à partir des différentes expériences de RMN possibles, les spectres doivent d'abord être attribués. Pour faciliter cette attribution, un outil graphique, semi-automatique qui utilise le concept des plans produit et somme a été développé. Cet outil a permis l'étude de deux IUPs individuelles, Tau et NS5A VHC. Les résonances RMN du squelette et des C_β ont été attribuées entièrement pour deux fragments de Tau (F3 et F5) et partiellement pour Tau entier P301L. Ces attributions RMN de Tau pourraient mener à davantage de compréhension sur le comportement structural de cette protéine quand elle se lie à, ou polymérise des microtubules. Aussi la formation des agrégés de Tau, qui est une des caractéristiques de la maladie d'Alzheimer, pourrait être étudiée plus en détail. Dans un deuxième temps, la protéine non structurale 5A (NS5A) du virus de l'Hépatite C (VHC) a été étudiée. Les propriétés structurales du deuxième et troisième domaine (sur trois) de cette protéine ont été évaluées. De la structure hélice α résiduelle a été observée, ce qui pourrait indiquer des régions prédisposées à interagir avec d'autres partenaires cellulaires. On a également examiné l'interaction entre CypA et CypB et les domaines D2 et D3 de NS5A, car ces PPIases pourraient jouer un rôle dans la réplication de VHC.

Keywords

Intrinsically Unstructured Proteins (IUPs), human Tau, HCV NS5A, NMR, Sequential NMR Assignment of IUPs, Non-Sequential NMR Assignment of Structured Proteins

Thesis Lab

Université des Sciences et Technologies de Lille (USTL Lille 1)
CNRS UMR 8576 - Unité de Glycobiologie Structurale et Fonctionnelle
Groupe de RMN - Bâtiment C9, Cité Scientifique
59655, Villeneuve d'Ascq cedex, France