Universidad de Granada

Programa de Doctorado en Química

Tesis Doctoral

Thermodynamic and Structural Analysis of Ultrafast Folding Proteins

Celia Sánchez de Medina Revilla

Granada, 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Celia Sánchez de Medina Revilla
ISBN: 978-84-1306-216-7
URI: http://hdl.handle.net/10481/55984

ACKNOWLEDGEMENTS

Along the years in the laboratory, I have been lucky to count on the expertise of my PhD supervisors, Prof. Víctor Muñoz, who gave me the chance to do this Thesis and transmitted his enthusiasm about every aspect of protein folding, and Prof. Eva de Alba, guiding me on the amazing NMR world, and thank you specially for all your help throughout the writing process. During my time in the lab, all my colleagues have created a wonderful work atmosphere: Abhinav, Aga, Clara, Eduardo, Evelyn, Javier, Jörg, Lorenzo, Luis, Maggie, Malwina, Marta, Michele, Mila, Mourad, Nacho, Patricia, Rajendra, Ravi, Satish, Susana and Tanay. I will always be grateful to Mourad for directing me on my first steps into NMR; Malwina, for being a great labmate and teaching me about protein purification, thermodynamic techniques and for her friendship in and out the lab; Abhinav and Ravi, for their support along the writing period; and particularly to Clara, Marta and Michele, for their invaluable help.

I would like to specially thank my colleagues in Lewis Kay's lab for all their help, particularly to Ashok and Pramodh, who I had the pleasure to work with hand in hand and learn from them about the exciting Relaxation Dispersion NMR world under the attentive guidance of Prof. Lewis E. Kay. All the members made me feel part of the group from the beginning and were willing to solve any doubt I could have. I cannot forget all those who made me feel like home during my stay in Toronto: Angelo and his family, for trusting me even before we met and make me feel like part of their family; and Mike, for his friendship.

Obviously, a very special acknowledgement to my friends: Gustavo, Cristina, she knows this Thesis would have not even started without her, and Ángel, your friendship made me feel like home in Madrid, and I will never forget your constant backing and boosting me up during the writing process, with our talks and emails despite the distance.

I have many things to thank my family for, they always believed in me, specially my mother, who always encouraged me to fight for my dreams, and my father, insisting on never giving up. As well as my grandparents, specially my

grandfather Luis, who has always been an example on pursuing your goals. Definitely, I have to include my brother, Luis, and his wife, Helena, they were always there for me, as well as the Sánchez de Medina family with their interest on my Thesis even if they did not always understand what it was about, and to my uncle Félix, who kept me updated with his emails during my time in Canada.

Many others have also helped me throughout these years in the lab and after, on the writing period, that took longer than I expected, my gratitude to all of them. But finally, here it is!

"What we know is a drop,

what we don't know is an ocean"

Isaac Newton

ABSTRACT

Proteins are biopolymers involved in vast array of cellular processes performing very different roles such as catalysis (enzymes), regulation (hormones), structure (collagen...), defence (antibodies), etc. Until recently, most of the studied proteins showed a well-defined three-dimensional structure, encoded in its amino acid sequence that drives its protein folding. The disruption of the 3D structure (denaturation) of proteins results in the unfolding and even in the loss of the protein function. However, restoring the physiological conditions may allow for spontaneous folding back into the native conformation. The observation of the refolding towards the unique 3D structure after its unfolding is the basis of Anfinsen's hypothesis that the amino acid sequence folds in such a way that the free energy of the molecules reaches its minimum[1]. Nevertheless, sometimes proteins can fold into erroneous conformations, producing aberrant proteins that can cause aggregation and are pathological hallmarks of several degenerative diseases including Alzheimer's, Parkinson's, Huntington's and type II diabetes among others[2]. For years of research it has been attempted to figure out the intriguing paths that drive proteins into aggregates, and further, the protein folding process itself, which still remains relatively unknown, thus, becoming a major goal either in experimental, computational and theoretical research areas. It is therefore a mayor reason to focus every effort to deepen our understanding of the protein folding process, and this Thesis intends to add a drop in the bucket.

Although all proteins are driven towards their native conformation, this process is performed at different rates depending on the protein, on its energy barrier that separates folded and unfolded states. Slow folding proteins have been long studied, but on the other hand, fast folding proteins have been limited by the experimental techniques available. However, advances since mid 1990s have improved enormously the ultrafast folding approaches, with resolution in the nanosecond-to-millisecond range, giving a further boost to this research field through techniques like: equilibrium based on line shape analysis from NMR peaks and based on single molecule fluorescence, temperature relaxation methods based on lasers, pressure relaxation methods, optically induced relaxation techniques, ultrafast mixing techniques and even computational methods[3]. The fast folding

proteins expand the knowledge of the folding and might cover the gap between experimental and computational methods, since traditionally the last ones have been able to explore only processes on the sub-millisecond timescale.

The identification of the different folding scenarios predicted by the folding funnel hypothesis of the energy landscape theory (Bryngelson *et al.*[4]) is a fundamental step. In 1999 V. Muñoz and W. A. Eaton[5] proposed the selection of some candidates to search for downhill scenarios *vs*. the classical two-state folding mechanism. Fast folding proteins frequently show downhill folding since a small or inexistent energy barrier speeds up the folding process. For this Thesis project we focused on possible candidates of the downhill folding mechanism since the shallow energy barrier allows to study the barrier top populations[3] as later explained in Chapter 1.3.2. Therefore, two relatively small proteins with different secondary structure content were selected to study the folding patterns related to these structures: a WW domain, with all β-sheets; and a R3H domain, with β-sheet and α-helix secondary structures. These domains showed different thermodynamic behaviour related to their secondary structure content and hydrophobic forces among other features. Furthermore, two interesting proteins from a folding point of view are also studied with different techniques in the field of the Nuclear Magnetic Resonance (NMR). One of them is the frequently used for molecular simulation studies HP35 subdomain from Villin protein, focusing on the atomic unfolding of some residues with NMR, that shows less variation than would be expected for a fast folding protein with a tendency, in principle, towards the downhill scenario. The other, gpW protein, expected to be a two-state folder due to its size and $\alpha+\beta$ content, but with folding in the microsecond timescale[6], made it a very interesting case study, and for which the possible interconversion with an “invisible”, low populated, state was obtained, reflecting the non-static structure of proteins.

The thermodynamic characterization of the proteins was performed to obtain information about its cooperativity and behaviour looking at different probes within the protein, except in the cases where these studies were already performed. Later, taking as reference the atom-by-atom NMR study by Sadqi *et al.*[7] on the one-state downhill folder BBL domain (from the E2 subunit of the 2-oxoglutarate-dehydrogenase enzyme from *Escherichia coli*), we pretended to study the unfolding process of the selected domain. A similar analysis was performed by L. Sborgi on gpW protein[8], for which the atom-by-atom shows the behaviour of different NMR probes (amide ^{15}N and $^{1}H^{N}$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$), which can be complex if the process is

progressive like for downhill proteins[7] or more homogeneous like for two-state proteins[9] (even different tendencies may appear along the temperature range). Hence, the native state of gpW has been thoroughly studied along the thermal unfolding process, however, this native state may be a non-static structure since proteins change upon variations in their environment. To be able to deepen our understanding of the native state of gpW protein there was a barrier to overcome, since it has a folding rate too fast for the timescale affordable by Relaxation Dispersion (RD) Nuclear Magnetic Resonance (NMR). This technique allows the ambitious task of glimpsing into the dynamics of the native state of a protein in solution without perturbing it. It is therefore important to remember that the stable ground state can be in equilibrium with minimal populations of other meta-stable conformations at levels almost undetectable with regular techniques, that usually average all structures, making impossible to discern this very low populated states.

Protein folding, energy landscape, “invisible” states, and fast folding proteins are concepts related to protein function, a basic cornerstone for any living organism. The studies included in this Thesis will help in a better understanding of these concepts, and gave me the chance to explore different proteins with diverse techniques to understand the complex and fascinating world of protein folding.

RESUMEN

Las proteínas son biopolímeros implicados en gran diversidad de procesos celulares llevando a cabo distintas funciones: catalíticas (enzimas), de regulación (hormonas), estructurales (anticuerpos), etc. Hasta hace poco, la mayoría de las proteínas estudiadas mostraban una estructura tridimensional bien definida, codificada a través de su secuencia de aminoácidos, que dirige el plegamiento de la proteína. La alteración de la estructura tridimensional de la proteína (desnaturalización) conlleva el desplegamiento e incluso la pérdida de su función. Sin embargo, la vuelta a las condiciones fisiológicas puede permitir el plegamiento espontáneo de la proteína en su estructura nativa. Esta observación de retorno a la estructura tridimensional original tras su desplegado, es la base de la hipótesis de Anfinsen: la secuencia de aminoácidos se pliega de una forma concreta para llegar a un mínimo en la energía libre de la molécula[1]. No obstante, en ocasiones las proteínas se pliegan en conformaciones incorrectas, dando lugar a proteínas aberrantes que pueden producir agregación, y son indicadores patológicos de numerosas enfermedades degenerativas, incluyendo Alzheimer, Parkinson, Huntington, y diabetes tipo II entre otras[2]. Durante años de investigación se ha intentado descifrar los misteriosos caminos que llevan a las proteínas a agregar, y más aún, su proceso de plegamiento, que todavía permanece relativamente desconocido, convirtiéndose en un importante objetivo en distintas áreas, tanto experimental, como computacional y teórica. Por lo tanto, es importante centrar todo el esfuerzo posible en profundizar en el conocimiento del proceso de plegamiento proteico, y esta Tesis pretende añadir su granito de arena.

A pesar de que todas las proteínas tienden a su conformación nativa, la velocidad del proceso depende de la proteína, de la barrera energética que separa los estados plegado y desplegado. Las proteínas de plegamiento lento se han estudiado en profundidad, pero las técnicas experimentales han limitado el estudio de las proteínas con un plegamiento rápido. Sin embargo, desde mediados de la década de los 90 ha habido un gran avance en técnicas que estudian el plegamiento ultrarrápido, con resoluciones en tiempos de milisegundos a nanosegundos, impulsando este campo de conocimiento con técnicas como: equilibrio basado tanto en análisis de forma de línea por Resonancia Magnética Nuclear (RMN), como en fluorescencia de moléculas únicas, métodos basados en

relajación térmica con láser, técnicas de relajación por saltos de presión, relajación por inducción óptica, técnicas de mezclado ultrarrápido, y hasta métodos computacionales[3]. Las proteínas de plegamiento rápido amplían el conocimiento del plegamiento proteico y cubren la brecha que separa los métodos experimentales y los computacionales, ya que tradicionalmente estos últimos tan solo han sido capaces de investigar procesos por debajo de la escala de milisegundos.

La identificación de los distintos escenarios de plegamiento predichos por la hipótesis del embudo de plegamiento de la teoría de paisaje energético (Bryngelson *et al.*[4]) es una tarea primordial. En 1999 V. Muñoz y W. A. Eaton[5] propusieron la selección de proteínas candidatas para la búsqueda de escenarios de plegamiento *downhill* frente al clásico mecanismo de dos estados. Las proteínas de plegado rápido con frecuencia presentan un plegamiento *downhill* ya que su pequeña o nula barrera energética acelera el plegamiento. Para esta Tesis nos centramos en proteínas candidatas a un posible plegamiento tipo *downhill*, ya que su pequeña barrera energética permitiría el estudio de las poblaciones en lo alto de la barrera[3] como se explicará en el Capítulo 1.3.2. Por tanto, se seleccionaron dos proteínas relativamente pequeñas con distinto contenido de estructura secundaria: un dominio WW, formado por láminas β; y un dominio R3H, tanto con láminas β como hélice α. Estos dominios muestran distinto comportamiento termodinámico que se relaciona, entre otras características, con su contenido de estructura secundaria, así como con las fuerzas hidrófobicas internas. Además, otras dos proteínas interesantes desde el punto de vista de su plegamiento se han estudiado por RMN. Una de ellas se ha usado con frecuencia en estudios de simulación molecular, el subdominio HP35 de la proteína Vilina, en la que nos centramos en el desplegamiento a nivel atómico seguido por RMN, y que mostró menos variación a la esperada de una proteína de plegamiento rápido con tendencia, en principio, hacia un escenario tipo *downhill*. La otra, la proteína gpW, que se esperaba tuviese un plegamiento en dos estados por su tamaño y contenido en α+β, pero cuyo plegamiento se produce en microsegundos[6], la convierte en un interesante caso de estudio. Para esta última proteína se estudió la posible interconversión con un estado "invisible", poco poblado, el cual se consiguió determinar, reflejando la estructura no estática de las proteínas.

La caracterización termodinámica de estas proteínas se realizó para obtener información acerca de su cooperatividad y comportamiento a través de distintas sondas propias de la proteína, excepto en los casos en que esta caracterización ya

se había hecho con anterioridad. Posteriormente, tomando como referencia el estudio atómico (*atom-by-atom*) por RMN realizado por Sadqi *et al.*[7] sobre el dominio *downhill* de un estado BBL (procedente de la unidad E2 de la enzima 2-oxoglutarato-deshidrogenasa de *Escherichia coli*), se pretendía estudiar el desplegado de los dominios seleccionados. Un estudio parecido se llevó a cabo por L. Sborgi en la proteína gpW[8], para la cual el estudio *atom-by-atom* muestra el comportamiento de distintas sondas de RMN (^{15}N y $^{1}H^{N}$ amídicos, $^{13}C^{\alpha}$ y $^{13}C^{\beta}$), que puede ser complejo para procesos de plegamiento gradual como el caso de las proteínas tipo *downhill*[7] o más homogéneo para proteínas de dos estados[9] (a pesar de que pueden aparecer distintas tendencias a lo largo de rango de temperaturas). Por tanto, el estado nativo de gpW ha sido estudiado en profundidad a lo largo del desplegamiento térmico[8], sin embargo, este estado nativo puede ser una estructura no estática, ya que las proteínas cambian por alteraciones en su ambiente. Para ser capaces de profundizar en nuestro conocimiento del estado nativo de gpW hay que superar el obstáculo debido a la velocidad de plegado de esta proteína, muy rápida para las escalas de tiempo que abarca la Relajación Dispersiva (RD) por RMN. Esta técnica permite la ambiciosa tarea de vislumbrar la dinámica del estado nativo de la proteína en solución sin perturbarla. Es importante recordar que el estado estable fundamental puede estar en equilibrio con pequeñas poblaciones de otras conformaciones metaestables a niveles casi imperceptibles por técnicas habituales, que normalmente promedian todas las estructuras, imposibilitando la distinción de estados poco poblados.

Plegamiento proteico, paisajes energéticos, estados “invisibles”, y proteínas de plegamiento rápido, son conceptos relacionados con funciones proteicas, un pilar básico para cualquier organismo vivo. Los estudios incluidos en esta Tesis permitirán un mejor entendimiento de estos conceptos, y me han dado la oportunidad de explorar distintas proteínas a través de diversas técnicas para comprender el complejo y fascinante mundo del plegamiento de proteínas.

INDEX

ABBREVIATIONS

CD	Circular Dichroism
COSY	Correlation Spectroscopy
C_p	Heat Capacity
CPMG	Carr-Purcell-Meiboom-Gill
DEPT	Distortionless Enhancement by Polarization Transfer
DSC	Differential Scanning Calorimetry
DSS	2,2-Dimethyl-2-Silapentane-5-Sulfonate sodium salt
FES	Free Energy Surface
FRET	Fluorescence Resonance Energy Transfer
FTIR	Fourier Transform Infra-Red
H	Enthalpy
IDP	Intrinsically Disordered Protein
INEPT	Insensitive Nuclei Enhanced by Polarization Transfer
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Lysogeny Broth
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
PMSF	Phenylmethylsulfonyl Fluoride
RD	Relaxation Dispersion
RF	Radio Frequency
TCEP = TCEP·HCl	Tris(2-carboxyethyl)phosphine Hydrochloride
TFA	Trifluoroacetic Acid
T-jump	Temperature Jump
Tm	Apparent Melting Temperature
TOCSY	Total Correlation SpectroscopY
Tris	Tris(hydroxymethyl)aminomethane

PART I:

Introduction to Protein Folding

Chapter 1.- Protein Folding

1.1- From the amino acid sequence to the protein three-dimensional structure

Proteins are polypeptide chains with a defined three-dimensional structure. These structures are made of a sequence of amino acids that DNA encodes in the nucleotides, which transcription to RNA brings the information to the ribosomes, that synthetize the polypeptide chain with the combination of 20 amino acids (three nucleotides code for an amino acid) hold together by peptide bonds, creating the primary structure of the protein. This primary structure carries the information that will lead the protein folding process through different interactions (Figure 1.1). Some of these interactions are the hydrogen bonds between parts of the amino acid chain, that give rise to the basic elements of the secondary structure: the α-helix, which have hydrogen bonds joining the carbonyl (C=O) and the amide (NH) groups 4 residues away (i, i+4), originating a spring shape with 3.6 residues in each turn; and the β-sheet, in which hydrogen bonds connect amino acids from two β-strands that might be far in the protein sequence. To become active, a protein has to reach a defined three-dimensional structure in most of the cases, its tertiary structure – however, IDPs (Intrinsically Disordered Proteins) are an exception since they are active lacking a well defined structure in native conditions, in some cases, they may acquire the 3D structure upon interaction with a ligand becoming then active. This tertiary structure results from the interplay of the side chains of the amino acids belonging to different elements of the secondary structure through interactions like hydrophobic forces, Van der Waals, purely ionic, etc. Even if this tertiary structure might be already active, in some cases different protein monomers can interact, creating an oligomer, the higher step of organization – the quaternary structure.

However, this process is not as simple as it may look like. The steps undergone until the active tertiary structure is reached from its simple amino acidic chain constitute a very complex process, driven by an intricate net of weak forces, that altogether represent the *protein folding problem*[10].

The importance of understanding the protein folding process relies on the perspective of the 3D structure prediction only based on the amino acid sequence,

which would open up new opportunities to create synthetic proteins towards specific purposes.

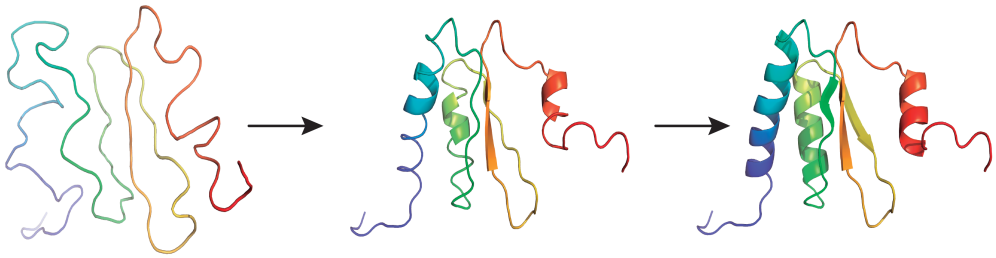


Figure 1.1. - Folding process, from amino acid sequence (left) to folded protein (right).

1.2.- What is protein folding?

For many decades, the way proteins fold to achieve its active conformation (also called native, on the right of Figure 1.1) has been a major research subject in protein science. In the late 1960s Cyrus Levinthal pointed out the very large number of degrees of freedom of the polypeptide chain that makes the molecule to have an astronomical number of possible conformations, making the folding process a 'find the needle-in-a-haystack' task, that is nonetheless found by the protein very quickly, the Levinthal's paradox[11]. Since then, there have been many advances in this field, but even now a comprehensive understanding of the protein folding problem, that would in turn allow to accurately predict the structure of the protein from its amino acid sequence, has not been achieved.

The thermodynamic hypothesis (Anfisen's dogma) postulates that the native structure is determined only by the protein's amino acid sequence, and it is the one in which the Gibbs free energy of the whole system is lowest[12]. The way proteins reach this minimum energy state was proposed to be following discrete folding pathways through a sequence of events taking place for the molecule, so that the protein reaches the correct metastable state[13]. However, during 1990s, a "New View" changed from the pathways towards a more complex scenario, a funnel-shape energy landscape, which drives the process while many elementary steps occur in parallel[4] (Figure 1.2).

1.3.- Energy Landscape and Protein Folding Scenarios

The Energy Landscape Theory, the "New View" of protein folding, is a statistical mechanics approach based on the principles of consistency and minimal frustration[4,14] that describes a rugged folding funnel-like shape as represented in Figure 1.2, biased towards the native state located at the bottom of the funnel. This funnelled scenario was compared by K. A. Dill *et al.*[15] to *a trickle of water down mountainsides*, in which the unfolded protein is located at the highest point of the funnel and rolls meandering down to reach its minimal energy conformation, hence the most thermodynamically stable state, the native conformation. However, it may find kinetic traps that slow down the folding or even hold the protein in non-native conformations, leading to a misfolded protein or aggregates. In 1995 J. D. Bryngelson *et al.* [4] proposed different folding scenarios based on the energy landscape. The two scenarios we focus on consider activation free-energy barriers between the folded (native) and the unfolded state to be over 3 *RT* or below 3 *RT*, for the two-state and the downhill scenarios, respectively; and thus, the cooperativity of the process from one state to another. The folding rate expression from this energy landscape approach is

$$k_{fold}(T) = \frac{1}{\tau_0} \exp\left(-\Delta G^{\dagger}/RT\right)$$

where $\Delta G^{\dagger}$ is the mentioned free energy barrier of the process, separating folded and unfolded conformations; τ_0, the rate of the protein motions to cross the energy barrier; *R*, the gas constant; and *T*, temperature.

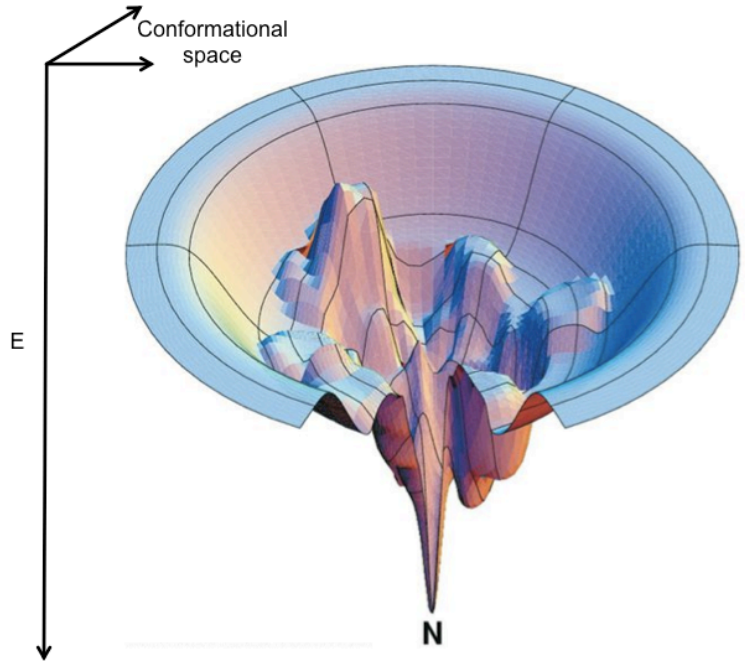


Figure 1.2. – Rugged energy landscape. The native folded conformation (N) is located at the bottom of the funnel, the lowest internal free energy location, while the unfolded or denatured is on the top edges. In the representation, the ups mean energy barriers, and the downs either favourable paths or kinetic traps if no way out.
Figure adapted from Dill *et al.*[15]

1.3.1.- Two-State Scenario

The classical view of protein folding is the two-state model, also called type 1 scenario[4] from the Energy Landscape Theory. This scenario shows a cooperative process that drives the molecule from unfolded to folded structure crossing an energy barrier over 3 *RT* higher than the free energy of the ground state[16], in an all-or-none mechanism[17], with an exponential folding kinetic similar to chemical reactions, implying the inexistence of observable folding intermediates between the native and the denatured structures. The unfolding process, as a function of temperature or denaturant concentration, produces a sigmoidal transition curve, characteristic of the two-state folding with sharper slopes being indicative of higher degree of cooperativity. This sigmoidal curve has clear flat pre- and post- baselines (indicative of no partially folded structures) determined by the folded and unfolded protein states, respectively, and characterized by the *Tm* (melting temperature, at which the protein is 50% folded and 50% unfolded) and the change on enthalpy, ΔH (determined by the width of the transition, and referenced at *Tm*). Since this folding occurs with no intermediates, all the possible probes followed by the different thermodynamical techniques provide the same *Tm* and ΔH values, which is another hallmark of the two-state folding. This scenario is typical of proteins with a nucleation core (usually hydrophobic) that drives the folding process since it is the rate limiting step, and folding rates in the millisecond range rate or slower.

Most of the studied proteins have been classified as two-state folders, either because the proteins were two-state folding proteins, or due to the limited resolution of the techniques available (mainly stopped-flow methods). Nowadays, the improvements and development of ultrafast folding approaches allow for a more detailed study of the folding process, avoiding the generalization of the two-states folding model[18,19].

1.3.2.- Downhill Scenario

The Energy Landscape Theory predicts also another scenario, in which folding can proceed without barriers or crossing small barriers (< 3 *RT*), therefore, the transition between folded and unfolded states is progressive and non-cooperative, it is the downhill, type 0[4] or one-state model, present in some fast

folding proteins (microsecond time-scale), and considered the most extreme manifestation of fast folding[3]. This scenario can be divided into three subgroups depending on the population of molecules folding downhill at denaturation midpoint (maximum height of the energy barrier, where the fraction of molecules folding downhill is minimal)[20]: the incipient downhill folders, with midpoint barriers between 1 and 3 *RT*; robust downhill folders, with marginal barriers (~1 *RT*); and the most extreme case, one-state proteins or proteins folding downhill globally, with insignificant or non-existing barriers.

The lack of a high energy barrier offers an exciting perspective where all the intermediate conformations are potentially detectable by experiments upon change on the external conditions along the broad transition of the folding process[5] even at the barrier top, for so long thought impossible to explore, 'bringing-to-light' features of the folding process 'hidden' until now. Therefore, this exciting theoretical prediction of downhill folding drove much of the experimental work on protein folding towards fast folding proteins, thus focusing on small proteins, since the folding rates scale up with size[21].

In 1993 Shlesinger and coworkers observed some "strange kinetics" (longer relaxation times)[22] in their experiments, as well as Sabelko *et al.*[23] in late 1990's by laser temperature-jump (T-jump) method. They noticed a continuous non-exponential unfolding, which could only be explained by a very small or almost non-existing energy barrier. These kinetics were in agreement with the results obtained in 2002 by Garcia-Mira *et al.*[24], who experimentally first proved the downhill folding scenario (though, it is not necessarily a non-exponential process[25,26]). They characterized a 40-residue downhill domain (BBL) by thermal unfolding with thermodynamical techniques, namely far-UV Circular Dichroism (CD), Fluorescence Resonance Free Energy (FRET), and Differential Scanning Calorimetry (DSC). Each technique probes a different feature of the protein structure, obtaining a different apparent melting temperature originated from a distinct unfolding transition, as predicted by statistical mechanical models[27]. The search for downhill scenarios then focused on the identification of quantitative parameters to define it, such as[5,28] I) probe-dependent equilibrium[24], II) complex coupling between denaturing agents[29], III) characteristic DSC thermogram[30], IV) gradual melting of secondary structure[31], V) heterogeneous process[7], and VI) generalized baseline crossing in fits to global two-state models[32] (which should always be explained along with the rest of the data). Thenceforth, other studies analysed the unfolding of BBL at atomic level via

NMR experiments, following different proton probes (at the backbone and side chains)[7]. Even though the sum of the individual NMR unfolding curves corresponds to the global behaviour observed by a low-resolution technique, the single probes show a quite heterogeneous behaviour, and an interaction map built on the patterns of the unfolding curves reveals the source of folding cooperativity. An analogous heterogeneity was also observed by nanosecond T-jump experiments, probing backbone structure (using IR absorption as a probe) and end-to-end distance (using FRET)[28]. However, the occurrence of non-native contacts with low free energy (with the possibility of folding intermediates)[33], and of other topological and energetic frustrations like native contacts between residues far in the sequence[34,35], may roughen the landscape, increasing the barrier height. The complete comprehension of the folding process would make it possible to sculpt the free energy surface (FES) via sequence (mutations) and environment (temperature)[36]. Studies on downhill folding proteins propose that the function of downhill folders might be modulated by changes on the distribution of the ensemble of structures, either by a gradation on the signal intensity or by changes on temperature or pH, suggesting a possible use as molecular rheostats[24,28,37]. In this regard, it has also been proposed an appealing connection between gradual unfolding and intrinsically disordered proteins (IDPs) that may explain the allosteric effects of folding coupled to binding[37], like the case of NCBD (nuclear co-activator-binding domain), with folding into different conformers depending on the binding partners[19,38-41].

1.4.- Fast Folding Proteins

The study of the possible downhill folding candidates arose in parallel to the development of techniques for the study of fast folding proteins, in the microsecond and sub-microsecond timescale. Some of these techniques that allow nowadays resolutions in such fast timescales are laser-induced temperature-jump, NMR line shape analysis, pressure relaxation methods, ultrafast mixing techniques, etc., which keep pushing towards time resolutions that enable to observe the folding speed limit. Proteins with fast folding and small size help covering the gap between experiments, that used to only reach milliseconds timescales until the development of the above mentioned techniques, and atomistic simulations, usually under the microseconds timescale. Nevertheless, detailed simulations of several microseconds have become possible, reaching up to 1 ms[42]. At the same time, these fast folding proteins also enable connecting both experiments and simulations with theory, offering a complete view of protein folding.

These fast folders are particularly interesting since they can probe some predictions from the Energy Landscape Theory, like the folding with very small or no barriers of the downhill folding proteins. This lack of barriers together with small protein size[21] and high content of α-helical secondary structure[43] may be some of the features to speed up the folding rate. Anyway, these features are just some starting points, since some proteins with different topologies can also show fast folding, even if less probable, like some WW domains, with all β-sheets[31,44,45], and gpW protein, with α+β topology[6]; at the time that some α-helical domains may have much slower folding times, like PDD, with ~10-fold slower than its structural homologue BBL[46]. Some proteins have been studied to increase its folding rate upon simple mutations, like the 35-residue headpiece subdomain of villin protein, a helical protein with microseconds folding that was mutated in two residues, increasing the folding rate exponentially[47].

The study of the kinetics and thermodynamics of these fast folding proteins, most of them highly stable, in comparison with other proteins with similar fold topologies constitutes a unique opportunity to design stable proteins with new functions[48], at the time that the study of these fast folding proteins and its possible downhill folding can also show the first glimpses of the folding mechanisms. Therefore, fast folding proteins offer new and very interesting perspectives in the protein folding knowledge.

PART II:

Material and Methods

Chapter 2.- Equilibrium Techniques for Protein Folding Analysis

Different spectroscopic techniques may help to deepen the preliminary protein characterization via equilibrium experiments. Even though these are low-resolution experiments, they allow discerning if the protein has a two-state or downhill folding mechanism by probing the different structural features of the proteins during its unfolding transition. In the case of the downhill folders (previously explained in Chapter 1.3.2), the protein loses structure gradually, changing the ensemble of conformations we can detect as unfolding proceeds, therefore, the different techniques highlighting specific structural features can show slightly different transitions, as in the downhill showcase protein BBL that exhibit large spread of *Tm* (from 295 to 335 K)[24]. For gpW protein (Chapter 8), this characterization was already performed by A. Fung *et al.*[6], observing the probe dependence typical of downhill proteins. Thus, only some experiments were performed as control instead of the full thermodynamical characterization.

2.1.- Circular Dichroism (CD) Spectroscopy

Circular dichroism is a spectroscopic technique based on the differential absorption of the left-handed and right-handed circularly polarised light, at different wavelengths, which happens when polarized light interacts with an optically active chiral centre. For protein studies, these optically active chiral centres are the backbone amide bonds of the protein, whose optical absorption lies in the far UV region (190-250 nm). The information obtained reports on the secondary structure content of the protein[49], the proportion of α-helix, β-sheet and random coil, that changes under denaturation conditions (pH, temperature, urea or other denaturants).

Each secondary structure shows characteristic positive and negative bands in a CD spectrum (see Figure 2.1):

- Alpha helix shows a positive band at 193 nm and two negative bands at 208 and 222 nm[50];
- Beta sheet has a positive band at 195 nm and a negative one at 218 nm[51];

- Random coil has a negative band at 195 nm and a slightly positive band around 212 nm[52].

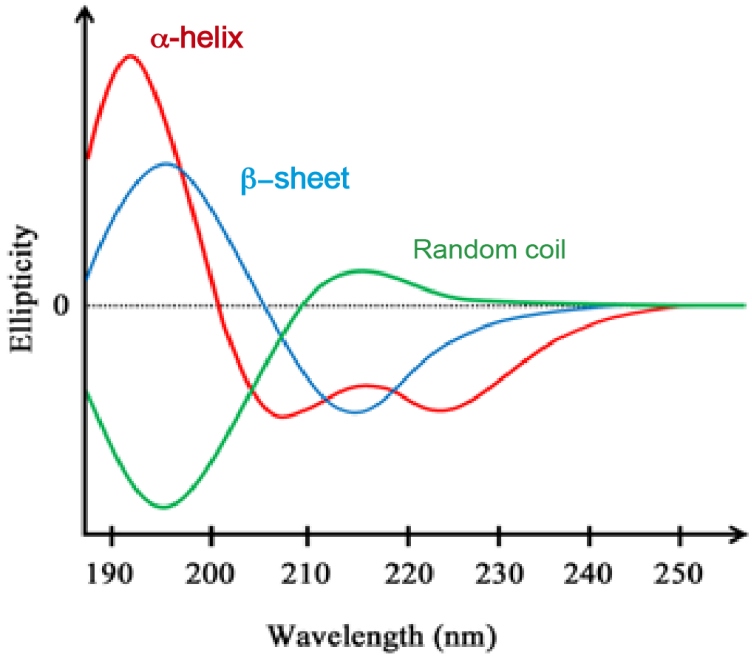


Figure 2.1. - Far-UV CD of the main secondary structures

However, some elements of the proteins can complicate the spectrum due to its dominance, like the aromatic residues Phenylalanine (Phe), Tyrosine (Tyr) and Tryptophan (Trp), as well as disulfide bonds. Furthermore, the presence of some chemical elements in the buffer (e.g. NaCl, GuHCl) can prevent the acquisition of the full CD spectrum in the far-UV region due to strong optical absorption of these components, therefore it is important to choose accurately the buffer composition when planning the experiment.

In order to be able to compare the spectra of different sample/proteins, the CD signal is expressed in mean residue ellipticity (MRE or $[\theta]_{MRE}$) which units are deg·cm^2·dmol^{-1}, obtained by:

$$[\theta]_{MRE} = \frac{\theta\,(\lambda)}{(10 \cdot l \cdot C \cdot n_{pb})} \qquad \textit{Equation 2.1}$$

where θ is the experimental ellipticity obtained from the protein (after subtracting the contribution arising from the buffer) at a specific wavelength (λ), *l* the cell path length (in cm), *C* the sample concentration (in mg/ml), and n_{pb} the number of peptide bonds.

To perform the experiments, the samples were dissolved in the appropriate buffer according to the desired pH – namely sodium phosphate 20 mM buffer for pH 7, used for WW and R3H domain, and pH 6, for gpW; and citrate-phosphate 20 mM buffer for low pH: pH 4.7, 3.5 and 3, for WW domain, gpW and HP35 subdomain, respectively. All samples had a protein concentration around 40-60 μM (measured by absorbance at 280 nm); the oscillation between the concentration of the different samples is corrected in Equation 2.1 for MRE. Finally a 1mm path quartz cuvette is loaded with the buffer (for baseline corrections) and later the sample. The spectropolarimeter used in this work was a Jasco J-815 coupled to a Peltier temperature controller (Jasco PTC-423S). For the temperature denaturation study, the temperature-ramp (T-ramp) started at 268 K with measurements every 5 K up to 368 K. In order to exclude possible protein loss due to precipitation in denaturing conditions, reference spectra at 298 K were recorded before and after the T-ramp.

2.2.- Fourier Transform Infra-Red (FTIR)

FTIR spectroscopy provides information about the secondary structure of the protein. The signal arises from the absorption of the infrared radiation (in the range from 1800 cm^{-1} to 1300 cm^{-1}) by the amide bonds that link the amino acids in the protein backbone, originating specific vibrational modes, with the Amide I and Amide II as the most prominent ones. The bands in the Amide I region (1700-1600 cm^{-1}) originate from the stretching vibrations of the C=O, while those in the Amide II region (1575-1480 cm^{-1}) derive from the bending vibrations of the N—H bond and the stretching vibrations of C—N. Since the hydrogen bonding between the C=O and the N—H groups drives the formation of the secondary structure, any variation on the secondary content modifies the amide bands, its spectral position and shape[53 54]. The most commonly used band to study proteins is the Amide I, because is the most sensitive to the secondary structure components and is correlated closely to the different elements (see Table 2.1), so henceforth the focus will be on this region.

	Amide I	
Conformation	**H_2O (cm^{-1})**	**D_2O (cm^{-1})**
α-helix	1653 (*s*) (1648-1655)	1650 (*s*) (1648-1655)
Antiparallel β-sheet	1632 (*s*) (1630-1636) 1690 (*w*) (1690-1693)	1632 (*s*) (1628-1638) 1675 (*w*) (1672-1678)
Parallel β-sheet	1630 (*s*) 1645 (*w*)	1632 (*s*) 1645 (*w*)
Random coil	1656 (*m, br*) (1656-1660)	1643 (*m, br*)

Table 2.1.- Frequency assignment (cm^{-1}) of the Amide I components to the different secondary structure of proteins. *S,m,w*,and *br* indicate that the band is strong, medium, weak or broad, respectively. Adapted from Arrondo *et al.*[55]

The samples were prepared at ~1 mM protein concentration with at least three deuteration cycles to exchange the amide protons with deuterium and to remove the H_2O solvent, avoiding its strong band at 1640 cm^{-1}, on the area of interest of the spectra. For the deuteration cycles the protein was dissolved in deuterated water, and after flash freezing in liquid nitrogen, lyophilized. Finally, for the experiment the lyophilized sample was dissolved in deuterated buffer and the pD adjusted according to the protein under study, correcting for the isotope effects on the glass electrode[56]. To perform the temperature-denaturation experiment, it is necessary to measure both the sample and buffer spectra at each temperature, in the range from 283 K to 368 K every 5 K, having a precise subtraction of the background signal. Spectra were measured on a thermostated JASCO FT/IR 4200 with TGS detector at 1 cm^{-1} resolution using MgF_2 windows and 50 mm Teflon spacer.

2.3.- Equilibrium Fluorescence

The fluorescence studies in this work are based on the intrinsic natural fluorophores of the proteins: the aromatic residues Trp, Tyr and Phe. This technique is useful to obtain information about the structure and dynamics of the protein, hence of any change. Trp has the strongest emission and usually predominates the spectra if present: it has a strong absorption band at ~ 280 nm and emits at ~ 350 nm in water, but its fluorescence is also very sensitive to its environment, changing its intensity and spectral position depending on the location in the protein structure (Trp is usually buried in the hydrophobic core of the protein when folded due to its

hydrophobicity, but is exposed when unfolded), that makes this amino acid quite useful indicator of the folding/unfolding transition. Tyr has the excitation wavelength ~274 nm and the emission ~303 nm. And for Phe excitation and emission wavelengths are around 257 and 282 nm, respectively. As Table 2.2 shows, Trp has the highest intensity out of these three amino acids, and Phe the weakest.

	Absorption		Fluorescence	
Amino Acid	λ_{max} (nm)	ε_{max} (M^{-1} cm^{-1})	λ_{max} (nm)	Quantum Yield
Trp	280	5600	348	0.2
Tyr	274	1400	303	0.14
Phe	257	200	282	0.04

Table 2.2.- Absorbance and fluorescence properties of the aromatic amino acids Trp, Tyr and Phe, in water at neutral pH. Figure adapted from Ghisaidoobe and Chung[57]

To obtain the fluorescence signal, the fluorophore is optically excited to a singlet state by absorbing a photon of energy corresponding to the excitation wavelength, undergoing a vertical transition according to the Frank-Condon principle; the excitation energy is subsequently partially dissipated through conformational changes and interactions with the environment, resulting in a relaxed singlet excited state, from which the emission transition takes place. As a consequence the fluorescence signal has a longer wavelength (lower energy) than the absorption wavelength.

To prepare the sample, it is necessary to take into account the nature and amount of the fluorophores present in each sequence, in order to choose the correct excitation and emission wavelengths and to have a concentration high enough to obtain spectra with a good signal to noise ratio, but not too high to saturate the detector. For R3H domain with 2 Tyr and 7 Phe, it was necessary to use a relative high concentration (~ 15 μM), while for WW domain, with 2 Trp and 3 Phe, a less concentrated sample was enough (~ 10 μM). pH was adjusted for each protein according to the required value. The measurements were done in a thermostated Flurolog FL3-11 spectrofluorometer (Jobin Yovin, Inc.) with 5 nm slit widths, and 0.55 s integration time at temperatures starting at 283 K and increasing every 5 K up to 368 K. The protein solution was equilibrated for 2 min at each temperature before data acquisition.

2.4.- Differential Scanning Calorimetry (DSC)

DSC is a powerful tool routinely used to characterize the stability of biomolecules, binding to other molecules, and for the purpose of the present study, folding of proteins through the measurement of the protein energetics by thermal denaturation. This technique has the potential to provide a complete description of the folding/unfolding equilibrium[58], thus identifying the cooperativity of the folding process, which makes DSC a very interesting technique to prove the small barrier or even barrierless feature of the downhill folding mechanism[6,24,59-61], moreover, it has been also applied to study intrinsically disordered proteins, nucleic acids and lipids[62-64].

The DSC thermogram presents a curve of the heat capacity protein *versus* the scanned temperatures, like the typical one shown in Figure 2.2. This curve results from the progressive heat of the protein sample that weakens the interactions within the protein molecule, which is an endothermic process and will require an increase of the energy supplied to the sample cell. From the heat absorption peak, after correcting the instrumental baseline and normalizing for the protein concentration, it is possible to obtain different thermodynamic parameters, like the melting temperature (*Tm*), the maximum height of the peak that represents the temperature transition midpoint between the folded and unfolded states, and the enthalpy (ΔH) from the peak integration.

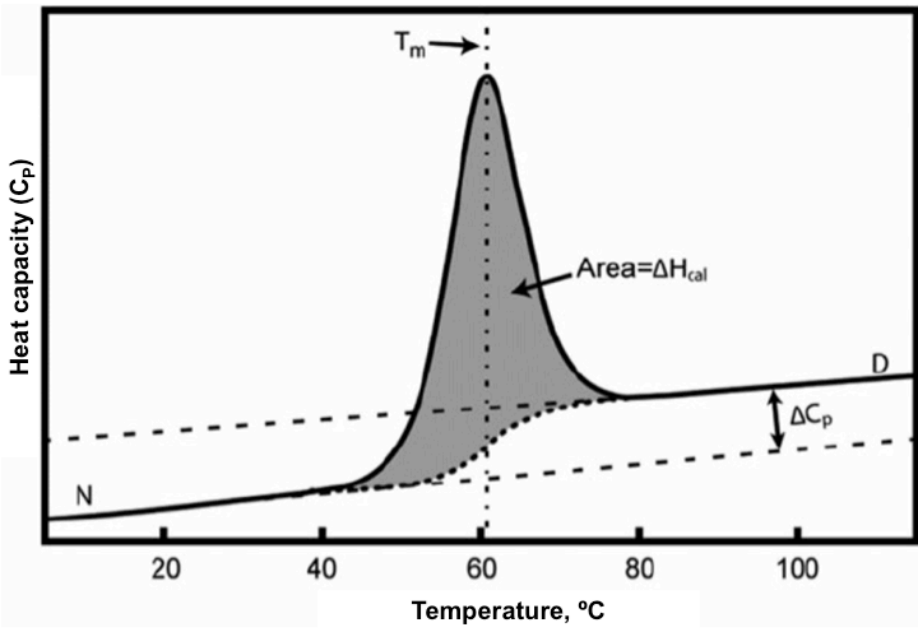


Figure 2.2. – Typical thermogram from a DSC experiment. *N* and *D* correspond to the native and denatured states, respectively. Figure adapted from Zhang and Ardejani[65]

Heat capacities for folded and unfolded states and their uncertainties were determined from the experimental thermograms, derived from the concentration dependence of the apparent heat capacity (ΔC_{app}), defined as the heat capacity of the protein solution taking the buffer-buffer baseline as reference[66]. For a protein solution it may be expressed as[67]:

$$\Delta C_{app} = \frac{C}{M} V_0 \, 10^{-3} \left(C_P - \frac{V_P}{V_W} C_{P,W} \right) \quad \textit{Equation 2.2}$$

where C is the protein concentration (mg/ml); M, the protein molecular mass (g/mol), V_0, the cell volume (ml); C_P and $C_{P,W}$, the heat capacities of the protein and the water (buffer), respectively; and the molar volumes of protein, V_P, and water, V_W. According to Equation 2.2, a plot of ΔC_{app} *versus* protein concentration (see Figure 5.5A) for a given temperature should be linear with zero intercept and with a value of the slope equal to $(V_0 / M) \times 10^{-3} (C_P - (V_P/V_W) C_{P,W})$, from which the absolute heat capacity of the protein can be calculated[67].

A DSC thermogram can show sharper (more cooperative)[68] or broader (progressive)[24] transitions, suggesting a higher or smaller energy barrier, respectively, or even a barrierless process. The heat capacity (C_p) baselines for folded (at low temperature) and unfolded (at high temperatures) states depend on the heat capacities of each state and can be theoretically estimated as proposed by Freire[69] (Freire's baselines, see Figures 2.2 and 6.6) for the folded state ($C_{P,N}$), as a function of temperature and depending only on the molecular weight of the protein, M_r (Equation 2.3), and for the fully solvated unfolded state ($C_{P,U}$), from the amino acid composition[69]. When the data is fitted to the traditional two-state model, the experimental baselines can indicate if the protein follows a two-state or a downhill behaviour: in case of a cooperative protein, they will show well defined native ($C_{P,N}$) and unfolded ($C_{P,U}$) baselines (similar to Freire's baselines) that provide a direct measure of ΔC_p[30,68]; however, if the protein has a small or negligible barrier the baselines can cross each other avoiding a reliable ΔC_p value[60] (see Figure 5.5B). This situation arises from the ensemble of conformations typical of a barrierless protein, even at low temperatures. In this later case, Muñoz & Sanchez-Ruiz proposed a variable-barrier model[30] for the DSC curve analysis based on Freire's baseline prediction[69] suitable for both, one- and two-state folders. The DSC

information has been previously applied to estimate free energy barriers for ultrafast folding proteins[70] and to develop multi-model Bayesian analysis[60].

$$C_{P,N} = \{(1.323 \pm 0.054) \pm (0.0067 \pm 0.0013)\,(T - 273.15)\}\,M_r \quad J\,K^{-1}mol^{-1}$$

Equation 2.3

DSC thermograms were measured by Prof. Beatriz Ibarra and Prof. Sanchez-Ruiz's laboratory, at the University of Granada (Spain), in a VP-DSC calorimeter, with a capillary-cell microcalorimeter from Microcal (Northampton, Massachusetts). The calorimeter holds two cells, sample (with the protein dissolved in the buffer) and reference buffer. The sample, with a concentration ~ 1 mM, was thoroughly dialyzed against the buffer, and the reference buffer was taken from the last dialysis step. Both were degassed before the measurements. Previous to the experiment, several buffer-buffer baselines were performed to confirm the proper equilibration of the instrument. During the experiment, sample and buffer cells are heated at a constant temperature-scanning rate, keeping both at the same temperature. When the protein begins to unfold, the calorimeter measures the difference of heat necessary for the protein sample cell (due to the denaturation process) compared with the heat applied to the buffer to avoid any temperature difference between both cells.

Chapter 3.- Nuclear Magnetic Resonance (NMR) Spectroscopy

3.1.- Historical Background

NMR has experienced an exponential development since the late 30's. C.J. Gorter and later I.I. Rabi, were pioneers in coining the term NMR. The latter developed a method to record the magnetic properties of atomic nuclei, for which he was awarded the Physics Nobel prize in 1937. Almost 10 years later, in 1946 F. Bloch and E. Purcell, independently, published studies on nuclei radiation phenomena of different substances (H_2O[71] and paraffin[72], respectively). These and further achievements resulted in both researchers sharing the Nobel Prize in Physics in 1952. These experiments were performed by Continuous Wave (CW) NMR, where a series of wavelengths excited progressively each different nuclear spin. The sequential excitation of each spin resulted in time-consuming experiments with low sensitivity and were also restricted to one dimension. To overcome these disadvantages, a new method was developed that is capable of exciting all nuclei simultaneously with short radiofrequency pulses. This method is known as Pulsed NMR. The excitation of the nuclear spin results in an oscillating magnetization that in turn produces an electric current oscillating with the corresponding resonance frequency. This signal, known as the Free Induction Decay (FID), varies with time and is detected in the receiver coil of the spectrometer. In 1966 R.R. Ernst and W.A. Anderson applied Fourier Transformation (FT) to Pulsed-NMR to obtain a frequency-domain spectrum, containing the pertinent resonance frequencies and signal intensities from the time-domain FID data[73]. Because of the application of FT to NMR and the development of multidimensional NMR (the first 2D NMR spectra were published in 1975[74]), Ernst was awarded in 1991 Nobel Prize in Chemistry. 2D NMR spectroscopy was the catalyzer for the rapid growth of NMR applications and methods to the structural study of biomacromolecules. The most prominent researchers in this field are A. Bax and K. Wüthrich, the latter was awarded the Nobel Prize in Chemistry in 2002, and determined the first protein structure in solution by NMR in 1980.

Nowadays, NMR is a powerful and versatile technique used in a variety of research areas providing different valuable information. As an analytical tool, NMR is widely applied in several fields of Chemistry and Physics, for example for the characterization, identification and quantification of organic and inorganic compounds, polymers, etc. Within the Life Sciences, NMR has applications in the study of biomacromolecular structure, function and assembly, together with drug screening and design[75,76]. NMR along with X-ray crystallography, are the only techniques capable of providing the three-dimensional structure of macromolecules at atomic resolution[77]. Besides its applications on these different fields, NMR is also an outstanding imaging technique used in the Medicinal Sciences, as Magnetic Resonance Imaging (MRI)[78] provides information on tissue and body structure and composition in a non-invasive manner. MRI mainly detects different proton relaxation rates and differences in the water content of tissues and organs, providing important disease-related information, such as the presence of infections, tumors, etc. MRI is also an essential technique to study soft tissues like brain, eyes, liver, vessels... where other techniques based on radioactivity are not appropriate[79].

3.2.- Basis of NMR

The atomic nucleus has a nuclear spin characterized by a nuclear spin quantum number (I) to which a magnetic momentum is associated. The nuclear spin determines the number of energy levels of the different nuclear states. When either only one or both the number of protons and neutrons in the nucleus are odd, the nuclear spin and the magnetic momentum are non-zero. Within this group there are nuclei with odd mass number resulting in half-integer spin (e.g. ^{1}H, ^{13}C, ^{15}N, ^{31}P with $I = ½$) and nuclei with even mass number resulting in integer spin (e.g. ^{2}H with $I = 1$). However, when both the number of protons and neutrons are even, the nuclear spin and magnetic momentum are zero (e.g. ^{12}C). Nuclei with $I = ½$ have two energy levels and are thus preferred in typical NMR studies of biomolecules for practical purposes. Specifically, H, N and C are the atoms most commonly found in proteins. Whereas magnetically active ^{1}H is the most abundant isotope for H, inactive ^{12}C and ^{14}N are preponderant at natural abundance. Therefore, in many instances it is necessary to artificially introduce magnetic ^{15}N and ^{13}C isotopes. This procedure, known as isotopic labeling, will be explained in detail in Material and Methods (Chapter 4).

The magnetic moment ($\vec{\mu}$) associated to the nuclear spin is defined by the expression:

$$\vec{\mu} = \gamma \vec{I} \qquad \textit{Equation 3.1}$$

where γ, the gyromagnetic ratio, is a constant value for each nucleus. In the absence of an external magnetic field the nuclear magnetic moments are randomly oriented. However, for non-zero nuclear spin I, an external magnetic field (B_0) affects the nuclear energy levels. The latter are characterized by the quantum number "*m*" (the possible directions of the spin angular momentum), which for spin-half nuclei can adopt the values –½ and +½. The number of possible energy levels is defined by $2I+1$. Therefore, for nuclei with I = ½, there are two possible energy levels: the magnetic moment of the lower energy level shares the same orientation as the external magnetic field, whereas it aligns in opposite direction in higher energy level (Figure 3.1). The positive and negative values of *m* are known in traditional NMR as α (lower energy or ground state) and β (excited state), respectively. The distribution of nuclei population into different energy levels is known as Zeeman Splitting (Figure 3.1). The energy associated to these levels in frequency units is given by:

$$E_m = m\nu \qquad \textit{Equation 3.2}$$

where ν is the Larmor frequency (Hz) that depends on the nucleus gyromagnetic ratio γ, the chemical shift δ, and the external magnetic field B_0:

$$\nu = -\tfrac{1}{2}\pi\,\gamma\,(1+\delta)B_0 \qquad \textit{Equation 3.3}$$

The nucleus can flip from one energy level to another, provided this transition is allowed. The frequency associated to this transition is:

$$\nu_{\alpha\beta} = E_\beta - E_\alpha = -\tfrac{1}{2}\nu - (+\tfrac{1}{2}\nu) = -\nu \qquad \textit{Equation 3.4}$$

Therefore, the transition will result in a spectrum with a single line at the Larmor frequency.

This energy flip can be achieved by applying a radiofrequency (RF) pulse. The intensity of this RF pulse depends on the strength of the external magnetic field

(Equations 3.2 and 3.3) as the gap between the energy levels depends on it. As it follows from Equation 3.3, it also depends on the characteristics of the nucleus such as the gyromagnetic ratio and the chemical shift. However, the gyromagnetic ratio is different for each nucleus, 4 times lower for ^{13}C than for ^{1}H, and 10 times for ^{15}N respect to ^{1}H, which decreases the signal intensity for these heteronuclei. Therefore, different pulse sequences have been developed to enhance the sensitivity for ^{13}C and ^{15}N in heteronuclear experiments. Using a vector representation of the nuclear magnetization, we can describe this energy transition as the movement of the magnetic vector originally aligned with the z-axis of the external field, to the XY plane. This vector will precess in the plane at the Larmor frequency. The magnetization is recorded in a coil and its intensity decreases with time, as the nucleus relaxes to its original state. We are referring to the FID mentioned earlier. What we really observe is a global FID that includes the response of all excited nuclei in the sample.

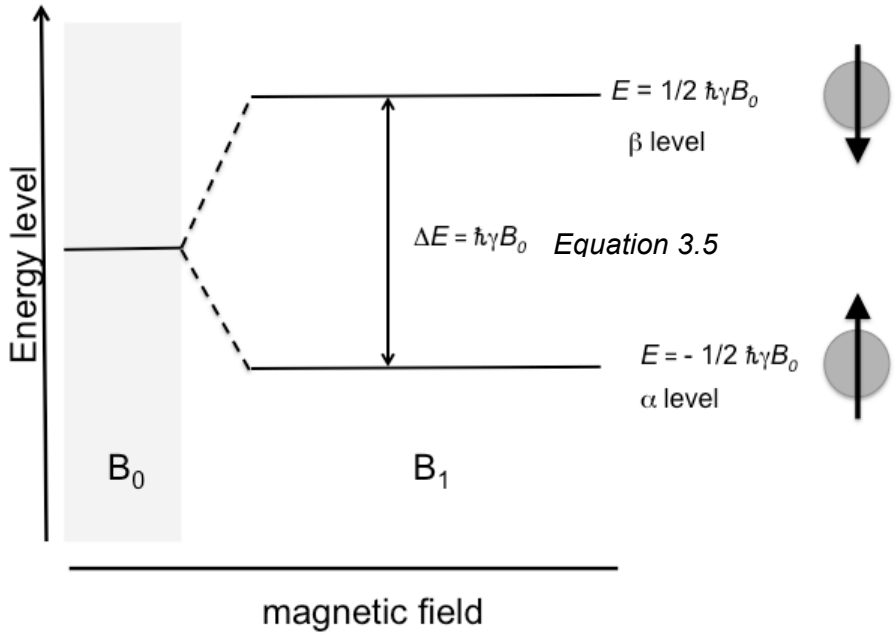


Figure 3.1.- Effect of a magnetic field on a ½ spin nuclei, generating in this case two energy levels, β when the spin is opposite to the direction of the magnetic field, and α when is in the same direction.

3.2.1.- Basic NMR Spectra Information

The typical NMR spectrum with signals in the frequency domain is obtained by Fourier transformation of the global FID. These signals carry useful information including:

- Intensity (I)
- Linewidth (λ)
- Chemical shift (δ)

3.2.1.1.- Intensity (*I*)

Signal intensity in NMR spectra is typically measured as the total area of the NMR peak in 1D spectra and the total peak volume in 2D spectra. This parameter varies with multiple factors, specifically with the magnetic field strength, sample concentration, sample relaxation properties, as well as other parameters that could affect NMR relaxation itself, such as sample size, temperature, solvent viscosity, etc.

3.2.1.2.- Linewidth (λ)

NMR linewidth is usually defined as the width of the peak at half of the maximum height. This parameter is directly related to the relaxation properties of the nuclei under study. The relation between the time decay of the FID and the Fourier-transformed NMR signal width is inverse, thus, when relaxation toward the original energy states happens fast (FID decays quickly), the resulting signal is broad, and the opposite happens for slow relaxation. The relaxation processes affecting signal linewidth are known in NMR as longitudinal (T1) and transverse (T2) relaxation. An in-depth description of the mechanisms involved in these two relaxation processes can be found in Chapter 3.5.

3.2.1.3.- Chemical Shifts (δ)

The chemical shift values coincide with the position of the peaks in the NMR spectra. This NMR parameter depends on the nuclear local environment, which includes the particular characteristics of each atom and the magnetic interaction

between different nuclei in the sample. An important factor affecting the value of the chemical shift is the shielding (σ) produced by the electron density around the nucleus, which effectively reduces the influence of the external magnetic field (B_0). This local magnetic field (B) is given by Equation 3.5:

$$B = (1 - \sigma) \cdot B_0 \qquad \textit{Equation 3.5}$$

The shielding modifies the resonance frequency. This modification is highly specific because it depends on the particular magnetic environment, thus NMR spectra are ideal for sample identification. The shielding effect is manifested in the presence of very different ranges of chemical shifts depending on the type of nuclei. For example, ^{1}H nuclei in proteins show substantially different chemical shifts depending on whether they are attached to N in the backbone or to C located in the amino acid side chain, and very importantly, they also show a strong dependence with protein secondary structure. Chemical shift values of nuclei are also affected by nearby nuclei to which they are connected through covalent chemical bonds. These effects are explained in Chapter 3.2.2.

For convenience, it is desirable to compare chemical shift values from different spectrometers, and thus independently of the external magnetic field strength. For this purpose, the signal of a reference substance is used and given the arbitrary value of 0 ppm. In this Thesis the compound DSS (2,2-Dimethyl-2-silapentane-5-sulfonate sodium salt)[80] was used as reference. The chemical shift, δ, is calculated as follows:

$$\delta = \frac{(\upsilon_{\text{observed}} - \upsilon_{\text{reference}})}{\upsilon_{\text{spectrometer}}} \cdot 10^6 \qquad \textit{Equation 3.6}$$

According to Equation 3.6 the chemical shift is, hence, independent on the spectrometer frequency and is expressed in *parts per million* (ppm) units.

As mentioned earlier, the chemical shifts of nuclei in the protein polypeptide backbone depend on the secondary structure among other factors. Therefore, it is possible to obtain important structural information by comparing observed chemical shifts to tabulated values of equivalent amino acids without defined structure (i.e. the random coil). In 1991, Wishart *et al.*[81] found a very strong correlation between the chemical shifts of the backbone atoms of the 20 natural amino acids and the

secondary structure of the protein. A specific nucleus shows different chemical shift value depending on its location in an α-helix or in a more expanded configuration, like the β-sheet. In particular, $^{1}H^{\alpha}$ chemical shifts move upfield when forming part of an α -helix, and shifts downfield with similar magnitude if located in a β-sheet[81]. The work by Wishart *et al.* is based on published chemical shifts of approximately 70 proteins[82], and more than 5000 residues. In addition to $^{1}H^{\alpha}$ protons, it also includes $^{1}H^{N}$ and ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C^{O}$. Furthermore, this work also finds that the α-helix dipole has an effect on the amide proton chemical shifts through hydrogen bond. Shortly after, several methods for secondary structure determination were proposed based on the chemical shifts[83]. One example is the widely used TALOS[84] (Torsion Angle Likelihood Obtained from Shift and sequence similarity) and its improved version TALOS+[85], which predicts the protein torsion angles (ϕ and ψ) from the observed chemical shifts of different nuclei (N, H^{N}, C^{α}, H^{α}, C^{β}, and C^{O}) and their comparison to empirical chemical shift values associated to protein structure databases.

3.2.2.- *J* coupling

J coupling arises from the spin-spin interactions between chemically bonded nuclei, transmitted through the electrons of those bonds. Scalar coupling, also known as spin-spin or *J* coupling, is expressed in Hertz, and depends on the product of the gyromagnetic ratios of the coupled spins, and the number and conformation of the involved bonds[86], but not on the external magnetic fields. *J* couplings are reported by including the bonded nuclei (for example *i* and *j*) and the number of bonds between them. For example, for *m* number of bonds the scalar coupling is denoted as: $^{m}J_{ij}$. These through-bond couplings arise from the effect that the different energy states of one nucleus has on another nucleus to which is chemically bonded. In the case of two bonded ½ spin nuclei, the energy levels of each one are split into two, giving rise to two Larmor frequencies (Figure 3.1). Generally, the multiplicity of the original NMR signal equals n+1 (n corresponds to the number of neighbouring nuclei). This splitting (expressed in Hz) into two (doublet), three (triplet), or multiple (multiplet) signals reduces considerably the signal-to-noise ratio and complicates the NMR analysis. However, this loss can be restored by inverting the spin-state of the coupled atom in a decoupling process.

J couplings provide very useful structural information. In particular, ^{3}J-coupling constants strongly depend on the ϕ and ψ torsion angles[87] characteristic of protein secondary structure (Chapter 1) as shown in Figure 3.3. The Karplus equation shown below (Equation 3.7) correlates these ^{3}J-coupling and the torsion angles:

$$^{3}J = A\cos^{2}(\theta) + B\cos(\theta) + C \qquad \textit{Equation 3.7}$$

where θ is the intervening dihedral angle between the vicinal nuclei in the sequence (ϕ and ψ for backbone, as well as side chain angles), and A, B and C are adjustable parameters obtained empirically that change upon the ^{3}J considered. However, this multi-valued equation may have several solutions in some cases, which is solved through simultaneous analysis of several *J*-couplings dependent on the same torsion angles. Furthermore, including NOE information enables to establish stereospecific assignments[88]. Thus, ^{3}J values obtained from Karplus Equation 3.7 provide information that can be translated into a structural restraint to define protein structures.

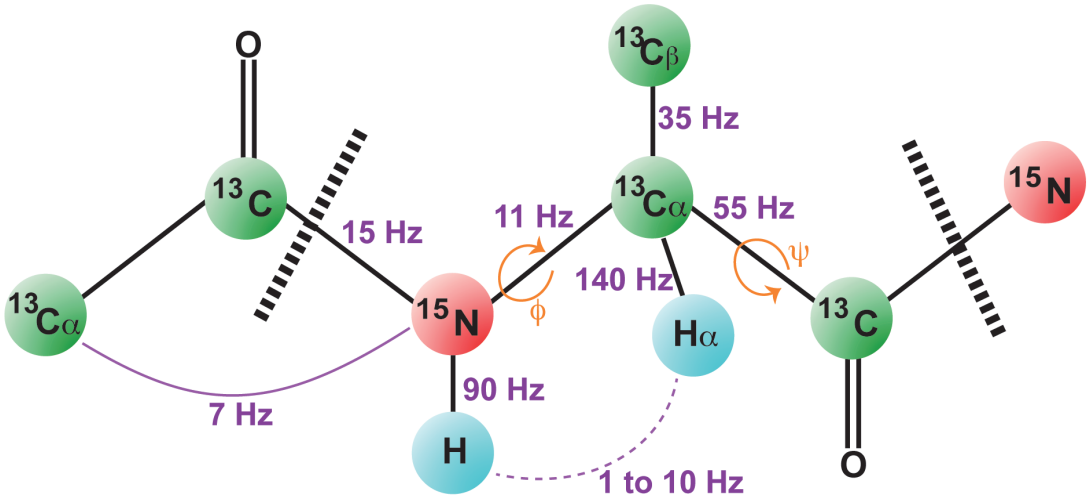


Figure 3.3.- Protein backbone with the main ^{1}J couplings. $^{2}J_{NC\alpha}$ has also been included to show the decrease in the coupling intensity when increasing the number of bonds, as well as a $^{3}J_{HNH\alpha}$ in dashed purple line. $^{1}J_{NC\alpha}$ has different value if the protein is denatured (11 Hz), or folded into a β-sheet (11,2 ± 0,5 Hz) or in an α-helix (9,9 ± 0,2 Hz) conformation. The dihedral angles ϕ and ψ are shown in orange. Black dashed lines delimit residues.

Some typical values for common scalar couplings in protein polypeptide backbone are shown in Figure 3.3. These values are important parameters in the design of NMR pulse program experiments to effectively transfer magnetization within the different nuclei. In addition, experiments designed to measure ^{1}J

couplings in proteins are particularly useful in the quantification of residual dipolar couplings, which have been proven to be of extraordinary importance in the improvement of macromolecular structural determination and interactions[89].

3.2.3.- Nuclear Overhauser Effect (NOE)

Nuclei that are distant in the protein amino acid sequence can be in close proximity when the protein folds adopting its native structure. These nuclei, now in close spatial proximity, can influence each other due to dipole-dipole relaxation (dipolar coupling). The interaction through space between magnetic nuclei arises from the local magnetic field created by a spin, affecting other nearby spins. It is the basis of the Nuclear Overhauser Effect (NOE), which is basically a change in the NMR signal intensity of a nucleus when nuclei participating in its dipole-dipole relaxation are excited by radiofrequency. The change in intensity of the signal is inversely proportional to the distance between the two nuclei ($\propto r^{-6}$), and is limited to approximately 5 Å. Because the NOE depends on the distance between nuclei, its measurement provides valuable data on interatomic distances that allows structural determination at atomic resolution. Another parameter influencing the NOE is the correlation time (τ_c), which is the average time needed for the molecule to rotate one radian[90]. Molecular weight, solvent viscosity and temperature are the main factors affecting τ_c. Larger molecular size, higher viscosity, and lower temperature values will result in an increase of the correlation time. The dependence of the NOE with τ_c results in positive values for small molecules (< 600 Da), negative for large molecules (> 1200 Da), and close to zero for middle-sized molecules.

To understand the origin of this NOE signal is important to have in mind that the origin of any NMR signal is the difference in population between a high (β) and a low (α) energy states. This NOE signal emerges from the polarization transfer via dipolar cross-relaxation originated by the perturbation of the energy level populations of the spin system. This effect is explained by the energy level diagram shown in Figure 3.4:

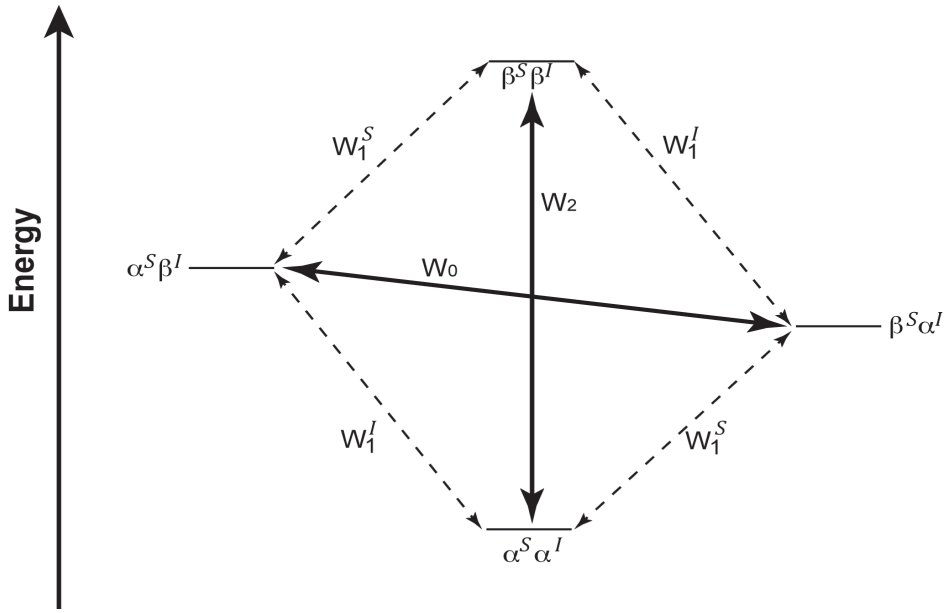


Figure 3.4.- Energy level diagram in a two-spin system (S,I) displaying the transition probabilities or transition rates (W), after the saturation of a nucleus S (Source) that interacts through space via dipole-dipole interaction with a nucleus I (Influenced). The NOE effect is originated from the zero quantum (W_0) and double quantum (W_2) transitions (solid arrow lines). Dashed lines indicate single quantum (W_1) transitions. α and β correspond to low and high energy levels, respectively.

When spin S (Source) receives a saturating radiofrequency pulse, the populations of the α and β energy levels of the S,I spin system become equal, thus creating a population difference relative to the ground state or Boltzmann Distribution. The dipolar coupled system tends to return to equilibrium via diverse relaxation processes. NOE signals are a manifestation of the system going back to the equilibrium state. The mechanisms to return to the equilibrium are as follows:

- W_0 mechanism is a Zero Order Quantum Transition as the two spins flip at the same time (I and S spins), thus decreasing the energy state of one of them ($\beta \rightarrow \alpha$) at the time the other one increases ($\alpha \rightarrow \beta$). Its frequency is small, resulting from the difference in frequency of the two nuclei, therefore the W_0 mechanism is favoured in slow tumbling molecules (large molecules and/or molecules in a viscous solution). In addition, the population difference across the I spin transitions decreases, thus decreasing its signal intensity and giving rise to a negative NOE.

- W_2 mechanism is a Double-Quantum Transition: Two aligned spins simultaneously flip, giving rise to a change in the magnetic quantum number of 2. The frequency of this mechanism corresponds to the sum of the frequencies of the

two nuclei and it will be preferred by fast tumbling molecules (small molecules with shorter τ_c). W_2 mechanism transfers spins to the $\alpha\alpha$ state or to the $\beta\beta$ state, increasing the population difference across the I spin transitions and thus, the signal intensity, producing a positive NOE.

- W_1 mechanism, Single Quantum Transition, involves the flip of only one spin, and corresponds to the T_1 relaxation process (or spin-lattice relaxation, see Chapter 3.5.1). W_1 transitions are induced by molecular motions that create magnetic fields oscillating at the Larmor frequencies of I or S and do not contribute to the appearance of the NOE. However, the W_1^S transition will reduce the saturation of S, decreasing the NOE magnitude.

A quantitative definition of NOE (Equation 3.8) can be derived from the Solomon equations and describes the sign and magnitude of the NOE signal:

$$\eta_I(S) = \frac{\gamma_I}{\gamma_S}\left[\frac{W_2 - W_0}{W_0 + 2W_1^I + W_2}\right] \qquad \textit{Equation 3.8}$$

where $\eta_I(S)$ is the NOE at I after saturation of S nucleus, γ_I and γ_S are the gyromagnetic ratios of I and S respectively, and W_0, W_1 and W_2 the rate constants for each relaxation process. This equation can also be written as Equation 3.9:

$$\eta_I(S) = \frac{\sigma_{IS}}{\rho_{IS}}\frac{\gamma_S}{\gamma_I} \qquad \textit{Equation 3.9}$$

Understanding that the cross relaxation rate is $\sigma_{IS} = W_2 - W_0$, and the total relaxation rate is $\rho_{IS} = W_0 + 2W_1^I + W_2$

3.3.- Atom-by-Atom NMR Assignment: 2D and 3D Experiments

As already mentioned, ^{1}H nuclei have a high natural abundance, but a 1-dimension ^{1}H NMR spectrum, even for small proteins, would be impossible to resolve due to signal crowding. Therefore, incorporation of other spectral dimensions provides additional information that facilitates protein spectra analysis. These multi-dimensional experiments (up to four dimensions in practical terms) can correlate different ½ spin nuclei, including different types of ^{1}H (homonuclear experiments) and ^{1}H with ^{13}C or ^{15}N (heteronuclear experiments). However, to compensate the natural low abundance of these heteronuclei (1.107% for ^{13}C and

0.366% for ^{15}N), the experimental time would have to increase excessively. To avoid these long experiments, the samples are enriched with ^{15}N and ^{13}C isotopes during protein synthesis (Chapter 4.4). Despite these improvements in protein NMR experimental techniques and protein enrichment in magnetically active isotopes, the molecular weight limitation is approximately 25-30 kDa[91]. The 2D heteronuclear experiments, correlating amide $^{1}H^{N}$ and amide ^{15}N, typically show a spread set of NMR signals for folded proteins (but collapsed for unfolded) that helps in the chemical shift assignment of the protein.

In a 1D NMR experiment the single frequency spectrum is acquired as a function of one time variable and then Fourier transformed. However, 2D experiments present two time variables, t_1 and t_2, with two Fourier transformations producing a spectrum with two frequencies (F1 and F2). A general 2D experiment can be defined by four basic elements (Figure 3.5):

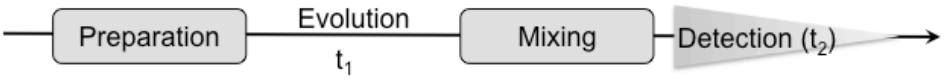


Figure 3.5.- 2D experiment scheme showing the four basic steps: preparation, evolution (t_1), mixing and detection (t_2,) time.

During the *preparation* time one or more pulses excite the sample, shifting the magnetization from the longitudinal to the transverse plane, where it evolves along the t_1 or *evolution* period. Afterwards, one or more pulses executed in the *mixing* time transfer the coherence from one spin to another (which will define the type of experiment). Finally, the generated FID signal is detected during t_2 (*detection*). This sequence of events is called pulse sequence.

The two-dimensional experiment is obtained by incrementing the t_1 time, starting with a t_1 value that is recorded at the end of the pulse sequence with the corresponding FID, and storing the data. After returning the system to equilibrium, t_1 is increased and the pulse sequence is performed again, and so on. The 2D experiment data is finally a series of t_1 values recorded as a function of t_2.

The key step that determines the type of correlation between both dimensions and thus, the information obtained from the final spectrum, is the mixing period. The coherence transfer can typically happen through-bond scalar coupling (*J* coupling) or through-space dipolar interactions (NOE)[92]. In the first type, the scalar

coupling interactions are mediated by covalent bonding, such as in the case of COSY (Correlation SpectroscopY) experiments and HSQC/HMQC experiments that contain INEPT[93] –Insensitive Nuclei Enhanced by Polarization Transfer– and DEPT[94] – Distortionless Enhancement by Polarization Transfer– for heteronuclear correlation or TOCSY[95] (Total Correlation SpectroscopY) type of coherence transfer. In the second type of experiment, magnetization mediated by dipolar cross-relaxation depends on the distance between the interacting spins, and generates cross-peaks in the NOE mixing process for both homonuclear and heteronuclear experiments[96,97].

An advantage of heteronuclear experiments is the possibility of selectively excite the proton or the heteronucleus independently in the same experiment by specific RF pulses that vary between them in hundreds of megahertzs (MHz), while in homonuclear experiments non-selective pulses typically excite all ^{1}H nuclei in the sample.

The scheme for 3D and 4D experiments is similar to that shown in Figure 3.5, where additional evolution periods are included. They are based on the same principles as 2D experiments and made by a combination of similar blocks of pulses. Multidimensional experiments may also combine both scalar and through-space magnetization transfer.

To understand how 2D and 3D experiments operate, one example of a 2D and two examples of 3D experiments are explained in more detail below.

3.3.1.- Two-Dimensional Experiment: [^{1}H-^{15}N]-HSQC

Heteronuclear 2D experiments correlate the chemical shifts of two different nuclei, like in the HSQC (Heteronuclear Single-Quantum Correlation)[98] experiment, where an heteronucleus X (X = ^{15}N or ^{13}C) and its directly bonded ^{1}H (^{1}J coupling), provide a single peak per ^{1}H-X pair. Thus, the [^{1}H-^{15}N]-HSQC experiment will produce a single peak per amino acid from its amide NH groups, providing information on the protein backbone and the side chain of some amino acids: Trp, Asn, Gln and Arg. This experiment is considered the fingerprint of the protein, and usually, one of the first experiments performed in any protein sample as well as

control between different samples of the same protein. This experiment has a similar sequence to HMQC (Heteronuclear Multiple-Quantum Correlation) experiment, which shows evolution during t_1 of multiple nuclei, ^{1}H and X, while during the same evolution period of the HSQC experiment, only X shows transversal magnetization. The evolution of only the X atoms avoids proton homonuclear coupling at t_1, increasing the f_1 dimension resolution for small molecules, which is the main advantage of the HSQC experiment versus HMQC, despite the first one implies a higher number of pulses that may decrease the signal-to-noise ratio if tuning or pulse calibration are not well performed. The differences between HSQC and HMQC experiments may be crucial when determining the sign of chemical shifts obtained for "invisible" states of proteins[99] as explained in Chapter 3.5.4.

In these experiments the heteronucleus is not directly detected like in other experiments, like for the 2D HETCOR (Heteronuclear Chemical-Shift Correlation) experiment, because the low gyromagnetic ratio produces low signal intensity. Thus, heteronuclear experiments usually have indirect or proton detection, where the ^{1}H nuclei (with high gyromagnetic ratio) is excited and its polarization is transferred to the heteronucleus, which is manipulated to transfer the magnetization back to the ^{1}H for detection (the reverse INEPT pulse sequence), thus increasing the sensitivity of the experiment almost 10 times (4 times for [^{1}H-^{13}C]-HSQC).

The [^{1}H-^{15}N]-HSQC experiment was used frequently during the research involved in this Thesis. The experiment, shown in Figure 3.6, can be explained in 5 main blocks, for which the product operator formalism[100] offers a clear physical meaning:

1.- The pulse sequence begins with an initial delay to allow the system to be at equilibrium.

2.- During the first INEPT block the polarization is transferred from the proton to the ^{15}N nuclei. This block consists on a first 90°_x pulse at the ^{1}H channel that places the magnetization of the proton on the *y* transverse plane, followed by a spin-echo sequence (τ – 180º – τ), where τ delay is $1/(4J)$ (J is the ^{1}H-^{15}N amide coupling constant). The time during the spin-echo allows the evolution of the magnetization with the coupling constant that is finally refocused by the two 180º pulses applied in both channels, resulting at the end of the second τ period a H_xN_z magnetization. The 90º pulses at the end of this block that transfers the magnetization from ^{1}H to ^{15}N,

constituting the real magnetization transfer. These two 90º pulses are not simultaneous to remove artefact by a gradient applied between the two pulses. The $90º_y$ proton pulse locates the magnetization in zz for both nucleus to allow for this gradient to be applied. Subsequently, the $90º_x$ ^{15}N pulse renders the transferred magnetization to the ^{15}N ($2H_zN_y$).

3.- The 180º pulse at the x-axis at half variable t_1 delay ($t_1/2$) is applied to refocus the ^1H-^{15}N coupling during ^{15}N chemical shift evolution.

4.- Reverse INEPT block brings the magnetization back to the sensitive ^1H nucleus, which is reached by applying $90º_x$ pulses in both channels, again shifted by a gradient as in the first INEPT block. In this case the first $90º_x$ pulse is on ^{15}N and after the gradient (removes artefacts), on ^1H. The magnetization is now placed in $-2H_yN_z$. τ delays allow the evolution of the coupling constant, and together with the $180º_x$ pulses in both channels to refocus ^1H chemical shift evolution constituting another spin-echo sequence. At the end of this period, the magnetization is ^1H_x and can now be detected.

5.- During the detection period (t_2) the in-phase ^1H magnetization is ^{15}N-decoupled with a GARP[101] or WALTZ[102] pulse sequence.

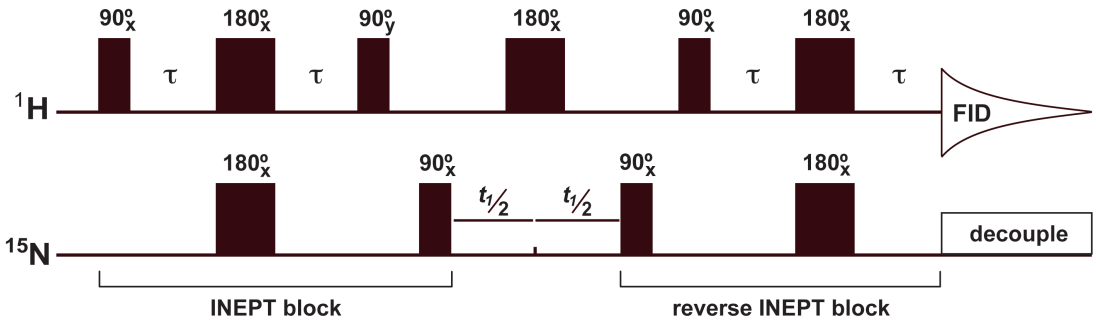
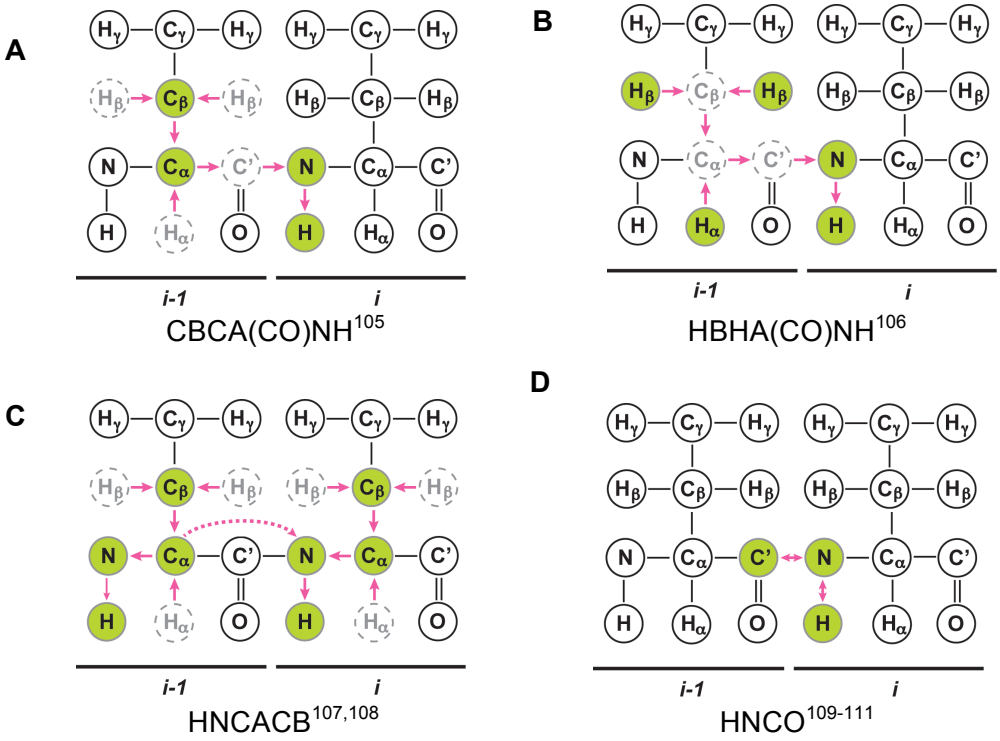


Figure 3.6.- [^1H-^{15}N]-HSQC pulse sequence. The upper set of pulses is applied to the ^1H and the lower to the heteronucleus, ^{15}N, in separate RF channels. The narrow and wider black bars represent the 90º pulses and the 180º pulses, respectively. The delay τ is set to $1/(4J_{1H\text{-}15N})$.

3.3.2.- Three-Dimensional Experiments

In most instances, 2D experiments still present a high degree of overlapping, rendering spectra considerably difficult to analyze. The increase in dimensionality facilitates signals separation and improves the resolution. Signal separation in 3D experiments is key for proteins full chemical shift assignment[103,104]. Typically, 3D heteronuclear experiments correlate ^{1}H, ^{15}N and ^{13}C nuclei. Heteronuclear 3D experiments help assign the backbone and the side chain chemical shifts, either aliphatic or aromatic, and can also provide stereospecific assignments. In these experiments different ^{1}J and ^{2}J couplings help in transferring the magnetization between the different nuclei along the protein sequence and through the amino acid side chain.

Assignment of the full protein can be achieved through sets of 2D and 3D experiments (even 4D if necessary). Figure 3.7 shows a summary of the three-dimensional experiments performed during this Thesis:



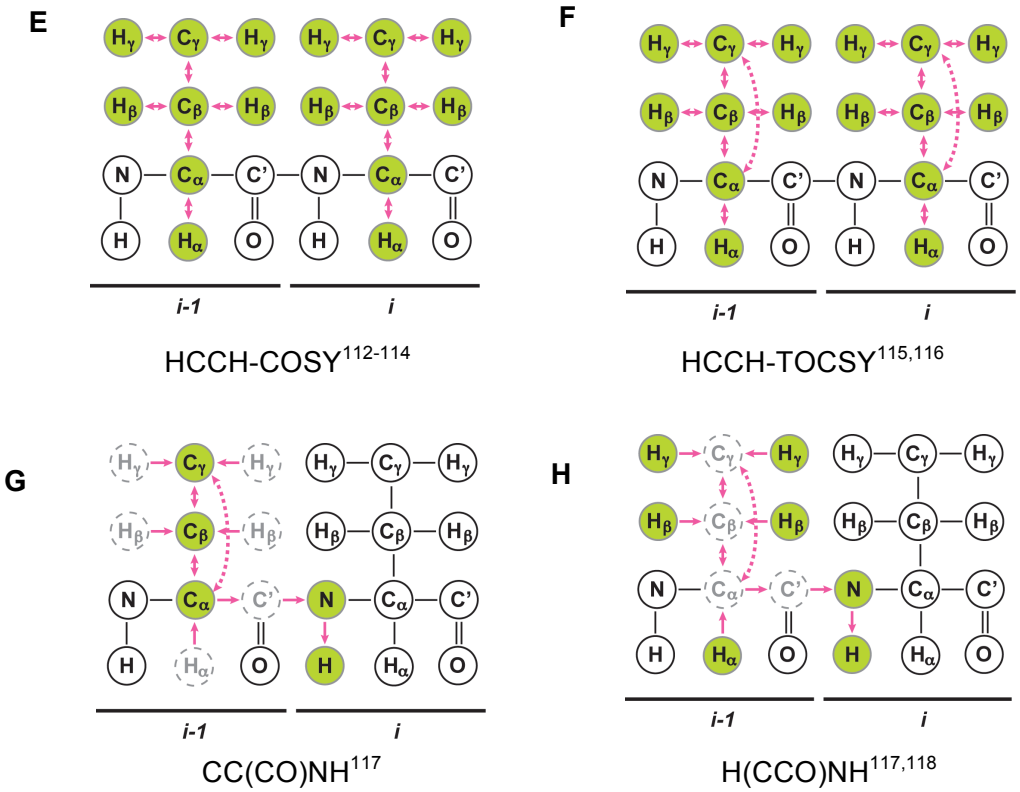


Figure 3.7.- Three dimensional heteronuclear experiments performed in this Thesis. Gray dashed circles and pink arrows indicate the atoms involved in the magnetization flow, while the green filled circles are the atoms observed in the experiment. Continuous and dashed arrows represent ^{1}J and ^{2}J couplings, respectively. Figure adapted from Cavanagh *et al.*[119]

These experiments are based on two-dimensional experiments including additional evolution and mixing times (see Figure 3.8). The 3D experiments have cubical representation, with one dimension per NMR parameter labelled during each evolution and detection period, with individual 2D planes. Thus, 3D experiments are based on two magnetization transfer steps and two evolution variable times (t_1 and t_2), whose values are incremented to obtain the second and third dimensions. However, to have a reasonable acquisition time with a good resolution, the number of data points has to be small, especially for t_1 and t_2 evolution times, at the time the data processing methods are designed to compensate for the poor digital resolution associated with the acquisition of a small data set[120]. The increment on the number of evolution periods also increases the loss of signal by relaxation processes. Therefore, to avoid large signal loss by relaxation, it is interesting to keep the mixing times short, easier in heteronuclear experiments because the larger coupling constants result in more efficient transfer of magnetization. Most of the experiments use indirect detection to increase the sensitivity for the heteronuclei as explained for 2D experiments.

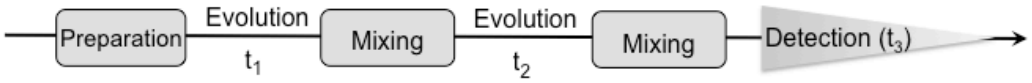


Figure 3.8.- 3D experiment scheme showing the extra evolution (t_2) and mixing time before the detection that differentiate the structure of 2D and 3D experiments.

To obtain the assignment of the full protein, it is important to get the chemical shifts of the different atoms through a set of 2D and 3D experiments. The diverse 3D experiments provide information about different atoms, and the complementarity of all the information from inter- and intra-residue experiments helps to obtain the full assignment. As examples of three-dimensional experiments, CBCA(CO)NH and HCCH-COSY experiments are going to be described in some detail, the first one as inter-residue experiment to obtain the protein backbone and the second as intra-residue experiment for the side chain.

3.3.2.1.- 3D Inter-Residue Experiment: CBCA(CO)NH

The 3D experiments CBCA(CO)NH and HNCACB are the first steps to obtain the protein backbone assignment. CBCA(CO)NH experiment provides correlations between the amide H and N resonances of residue *i*, with the C^{α} and C^{β} from the preceding (*i-1*) residue, bearing out the assignments of these peaks, at the time that HNCACB also gives the C^{α} and C^{β} chemical shifts of *i* residue as shown in Figures 3.7A and 3.7C. Once the HNCACB experiment shows the C^{α} and C^{β} peaks of *i* residue, this allows moving to the next residue, becoming this one the i-1 residue, leading to the walk through the protein backbone. The distinctive side chain chemical shifts of the different amino acids help connecting the residues within the protein amino acid sequence.

CBCA(CO)NH experiment establishes inter-residue correlation by transferring coherence via ^1J couplings: $^1J_{CH}$, $^1J_{CC}$, $^1J_{CON}$ and $^1J_{NH}$. The magnetization flow is indicated in Figure 3.9:

$$H\beta, H\alpha \Rightarrow C\beta, C\alpha\ (t_1) \Rightarrow C\alpha \Rightarrow CO \Rightarrow N\ (t_2) \Rightarrow H^N\ (t_3)$$

Figure 3.9.- Magnetization transfer during 3D CBCA(CO)NH experiment begins on the α and β protons and is transmitted to the directly bound ^{13}C (*i-1* residue), where it evolves during t_1. These carbons will locate the magnetization on $^{13}C^{\alpha}$ from where it will be transfer to $^{13}C^{O}$ (*i-1* residue), ^{15}N (t_2 evolution time) and finally to $^1H^N$ that is acquired along t_3 (^{15}N and $^1H^N$ corresponding to *i* residue). Figure adapted from Sattler *et al.*[121]

The experiment shows an initial INEPT pulse sequence to transfer from the $^{1}H^{\alpha}$ and $^{1}H^{\beta}$ to their directly bound $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ ($^{1}J_{CH}$), respectively, followed by a constant-time chemical shift evolution t_1, corresponding to $1/(4*J_{CC})$. At $t_1/2$ period, a shaped 180º pulse on $^{13}C^{O}$ removes scalar coupling (this type of pulse keeps $^{13}C^{\alpha}$ spins unperturbed). Then, a 90º carbon pulse refocuses the magnetization and transfers it from $^{13}C^{\beta}$ to $^{13}C^{\alpha}$. Afterwards, there are two blocks with a fixed evolution delay (constant-time) to obtain antiphase magnetization followed by two simultaneous 90º pulses on different nuclei to transfer magnetization. The first block, via $^{1}J_{C\alpha CO}$, creates antiphase magnetization of $^{13}C^{\alpha}$ with respect to $^{13}C^{O}$, and the 90º pulses in $^{13}C^{\alpha}$ and $^{13}C^{O}$ transfer magnetization to $^{13}C^{O}$; therefore, on the second block, via $^{1}J_{CON}$, the antiphase is created between $^{13}C^{O}$ and ^{15}N, and the pulses on $^{13}C^{O}$ and ^{15}N transfer the magnetization to the latter. At the same time, H^{α} and H^{β} are decoupled while carbon spins are in transverse plane, and H^{N} while in ^{15}N. Finally, the magnetization on ^{15}N evolves along the t_2 period, to be transferred to $^{1}H^{N}$ by a invert-INEPT, as seen for [$^{1}H^{N}$-^{15}N]-HSQC (Chapter 3.3.1), for proton acquisition (during t_3) with ^{15}N decoupling[105,121,122].

In this experiment, the carbon pulses are centred on the appropriate frequency depending on where the excitation is needed ($^{13}C^{\alpha}$, or $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ simultaneously), to maximize their excitation, and the pulse power adjusted to decrease the effect on $^{13}C^{O}$ [119].

In this Thesis, this experiment has been performed to obtain the backbone of the uniformly labelled protein gpW together with other 3D experiments like HBHA(CO)HN, HNCO and the side chain assignment experiment HCCH-TOCSY.

3.3.2.2.- 3D Intra-Residue Experiment: HCCH-COSY

3D COSY (COrrelation SpectroscopY) experiment improves substantially the classical 2D COSY, where the overlapping of the chemical shifts hampers the assignment. In the 3D experiment HCCH-COSY, the exchange of magnetization between all spins along the aliphatic side chains allows the assignment of its proton and carbon atoms (Figure 3.7E), but not for the amino acids with aromatic side chains. Within a spin system, the aliphatic proton chemical shifts are found at the carbon frequencies of the system, at the ^{13}C chemical shifts. In a typical HCCH

experiment[112] (either 2D or 3D versions) magnetization is transferred rapidly via ^1J coupling from the side chain protons to the directly attached carbon ($^1J_{HC} \approx 140$ Hz) and then to the neighbouring carbon ($^1J_{CC} \approx$ 33-45 Hz), to finally return to the attached protons ($^1J_{HC}$)[112,113]. The large *J*-couplings between ^{13}C atoms and between ^{13}C and its directly attached ^1H (Figure 3.3) allows faster magnetization transfer among the spins, consequently, signal loss due to relaxation is minimized. Magnetization transfer in the HCCH-COSY is illustrated in Figure 3.10:

$$^1H_{i-1}\ (t_1) \Rightarrow {}^{13}C_{i-1}\ (t_2) \Rightarrow {}^{13}C_i \Rightarrow {}^1H_i\ (t_3)$$

Figure 3.10.- Magnetization transfer during 3D HCCH-COSY experiment along the aliphatic side chain of the amino acids. The *i-1* and *i* subscripts refer to a couple H-C and the next one along the side chain, respectively. As for other 3D experiments, t_1, t_2 and t_3 are the magnetization evolution periods. Figure adapted from Rule *et al.*[86]

The pulse sequence of HCCH-COSY experiment begins with a 90º pulse on ^1H, and the evolution of the proton chemical shifts during the variable t_1, that includes a 180º ^{13}C decoupling pulse in the middle of the period to ensure the carbon decoupling. The magnetization is then transferred to the attached ^{13}C via an INEPT sequence, where a delay helps reaching the antiphase respect to the directly attached ^{13}C, and the magnetization is transmitted by 90º pulses on ^1H and ^{13}C at the same time. Subsequently, ^{13}C chemical shift evolves during the variable t_2 time and a constant delay generates ^{13}C magnetization antiphase with neighbouring ^{13}C spins. Two 180º pulses in t_2 period, one on ^1H and another one on ^{13}C, decouples proton and $^{13}C^{\alpha}$ from $^{13}C^{O}$. Later, a 90º carbon pulse transfers the magnetization to the next ^{13}C, and finally, a inverse-INEPT pulse sequence sends the magnetization back to the protons for acquisition (t_3) while ^{13}C spins are decoupled[114,119]. This experiment produces in-phase and anti-phase peaks, the corrections to get everything in-phase decrease the resolution of the experiment. However, in the HCCH-TOCSY experiment (Figure 3.7F) the resulting spectra in all in-phase, improves the resolution of the experiment.

The spectrum resulting from the experiment contains the side chain proton chemical shifts of ^1H *versus* ^1H in F1 and F2 dimensions, and the ^{13}C chemical shifts in the F3 dimension. Thus, it allows us to obtain the side chain assignment of the residue without correlation with the previous and/or next residue, which is convenient for the studied villin HP35 variants because the labelling is not uniform, only some residues are ^{13}C and ^{15}N labelled, hence, inter-residue magnetization

transfer is not possible. For the villin HP35 samples, the 3D HNCACB experiment (Figure 3.7C) has also been performed to obtain the $^{1}H^{N}$ and ^{15}N chemical shifts of the labelled residues. The inter-residue magnetization transfer sometimes observed in this experiment did not occur due to lack of labelled neighbouring residues.

3.4.- Chemical Exchange in NMR

Detection of subtle changes in the chemical environment of atoms is possible with very few techniques. NMR spectroscopy can detect the small effects caused by the variations of pH, solvent, conformational exchange processes, etc. that do not affect the chemical state of the molecules. Chemical exchange processes typically alter the environment of a chemical group, which in turn, might modify the resonance signals of an NMR spectrum. For example, a molecule with a –COOH group in equilibrium with $–COO^{-}$ will affect the chemical shift of the carbonyl itself and those of other nearby nuclei. In this case the molecule is changing between protonated and deprotonated. Therefore, under certain conditions, this change in the spectra provides information on the forward and backward constant rates of the exchange process, as well as about the populations of the intervening species[86].

The different chemical environments caused by chemical exchange (k_{ex}) can affect several NMR parameters, including chemical shift, relaxation time, coupling constants, etc., and it usually has a measurable effect in signal intensity and linewidth[123]. Chemical exchange includes intra- and inter-molecular processes. Complex formation such as protein-protein interaction and protein-ligand binding, are inter-molecular exchange processes, whereas side chain protein motions, protein folding and dynamics are examples of intra-molecular conformational exchange. In the latter case the following exchange rates are considered as Equation 3.10 shows:

$$A \underset{k_{BA}}{\overset{k_{AB}}{\rightleftharpoons}} B \quad \text{Where} \quad k_{ex} = k_{AB} + k_{BA} \qquad \textit{Equation 3.10}$$

where k_{AB} and k_{BA} are the forward and reverse rate-constants between *A* and *B*, and k_{ex} their sum.

Species A and B exist at equilibrium with the populations, P_A and P_B, respectively ($P_A + P_B = 1$), and they show frequencies, υ_A and υ_B, with a chemical shift difference: $\Delta\upsilon = |\upsilon_A - \upsilon_B|$.

Depending on the rate of exchange, Figure 3.11 shows the three main regimes that can be distinguished:

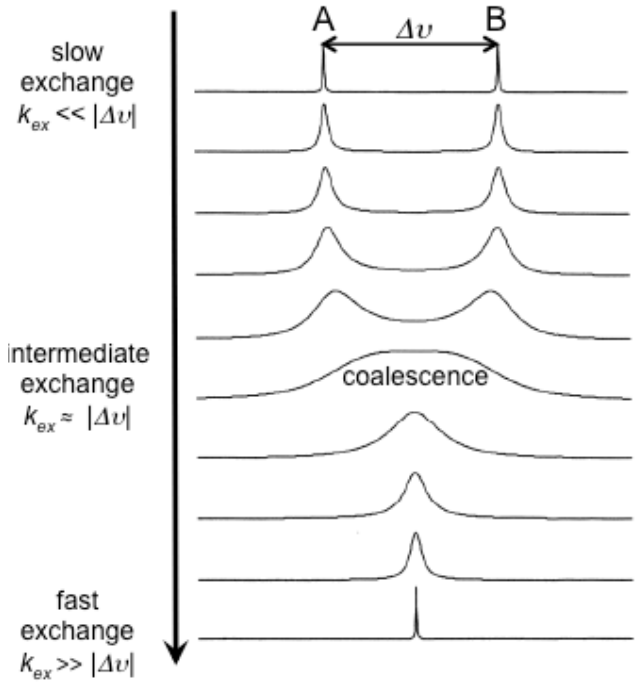


Figure 3.11.- Chemical exchange for a two-site system. At the top of the figure the exchange is very slow on the NMR chemical shift timescale, and two peaks corresponding to the two states are discernible. As the exchange rate increases, the signal intensity decreases, ultimately resulting in a single signal with a population-weighted average chemical shift in the fast exchange regime. In the intermediate exchange regime the peaks produce a single broad peak known as coalescence.

- In the slow exchange regime the rate of exchange is slower than the frequency difference of the spin in each environment ($k_{ex} << |\Delta\upsilon|$). Two signals can be observed in the spectrum.

- In the intermediate exchange regime ($k_{ex} \approx |\Delta\upsilon|$) or coalescence, one broad peak shows in the spectrum. Its resonance frequency depends on the relative populations of A and B, and shifts towards the prevalent one. The broadening of the NMR signal depends on the relaxation processes involved. The chemical exchange process of the system can be obtained with a lineshape analysis. In fact, anomalous

peak broadening (i.e. R_{ex} > 0) is a hallmark of dynamics in the intermediate exchange regime and is exploited extensively in Relaxation Dispersion NMR[123] .

- In the fast exchange regime (k_{ex} >> $|\Delta\upsilon|$), one signal is observed in the spectrum with a population-weighted averaged resonance frequency: $\upsilon^{Obs} = P_A\ \upsilon_A + P_B\ \upsilon_B$. The averaging occurs because the spins do not remain in either environment long enough to detect the corresponding resonance frequency.

The sample temperature and the strength of the external field affect the chemical exchange regime. Increasing temperature typically favors the fast exchange regime, and stronger magnetic field shifts the equilibrium towards the slow exchange regime. In some NMR relaxation experiments, such as Relaxation Dispersion, the comparison of peak broadening in different fields can provide very relevant information on the exchange process (see Chapter 3.5.3).

3.5.- Relaxation Phenomena in NMR

Proteins are flexible molecules that undergo multiple structural and conformational changes, occurring both at the atomic level and at the level of the whole polypeptide chain. Examples of these processes include folding, misfolding, chemical exchange, molecular binding and recognition, etc. NMR relaxation phenomena are exquisitely sensitive to molecular dynamics and motion, and thus NMR is a powerful technique for the accurate analysis of the processes mentioned above, even to the point of detecting protein conformations as significantly low populated as 0,5%[124].

Protein function is highly dependent on dynamic processes associated to protein flexibility and molecular recognition. Therefore, the study of relaxation phenomena by NMR can unveil very important information on protein function. Different NMR techniques can be applied to analyse relaxation phenomena that range from picoseconds to seconds, including internal motions (backbone motion and side chain rotations), rotational diffusion, different processes of conformational exchange, such as ligand binding, allosteric regulation, and also chemical exchange, for example amide proton exchange with water.

It is extremely important to estimate or measure the relaxation rates of the molecule subject of our NMR study, as basic parameters of all NMR experiments will heavily depend on these rates. For example, the recycle delay between acquisitions (the repetition rate), as well as the signal linewidth of the NMR spectra, and thus the overall sensitivity of the experiment strongly depend on these relaxation rates. The relaxation rate is based on the recovery rate of the equilibrium magnetization, always influenced by the surrounding molecules. The return to equilibrium of the spins implies the recovery of the populations predicted by the Boltzmann distribution (states α and β, ground and excited, respectively), and disappearance of the transverse magnetization (in the *x* and *y* axes, M_{xy}), only longitudinal magnetization (along the *z* axis, parallel to the static magnetic field) should be present. This longitudinal magnetization, represented as M_z, is proportional to the natural difference of population between the α and β states, n_α and n_β, respectively, also written as Equation 3.11:

$$M_z \propto (n_\alpha - n_\beta) \qquad \text{Equation 3.11}$$

To understand different concepts along this NMR section we are going to focus on two of the main sources of NMR relaxation, those resulting from spin interactions with the surroundings (the lattice), known as T_1, spin-lattice or longitudinal relaxation, and those caused by the loss of coherence of the transverse magnetization, called T_2, spin-spin or transverse relaxation. These interactions are essential because magnetic relaxation occurs spontaneously at a negligible rate, after for example a 90º pulse, which shifts the magnetization from the *z*-axis to the *xy*-plane (M_z shifts to M_{xy}) and alters n_α and n_β ratios. At the end of this Relaxation NMR section, the Carr-Purcell-Meiboom-Gill (CPMG) Relaxation Dispersion (RD) is explained, as it provides extremely detailed information on protein dynamics involving interconversion between different conformations.

3.5.1.- Longitudinal (T_1) Relaxation

T_1 relaxation corresponds to the recovery of the equilibrium magnetization on the *z*-axis (M_z^0) that leads to the restoration of the Boltzmann distribution of populations. It is originated mainly by the dipolar interactions between the spin system within the molecule due to translations, rotations and internal motions of the

molecules (also referred to as the lattice), causing the nearby spins to create transient, fluctuating magnetic fields at the Larmor frequency of the spin system. Therefore, T_1 relaxation is a process of intramolecular energy transfer as well as from the nuclear spin to surrounding nuclei.

The T_1 relaxation value determines the recycle delay at the beginning of any NMR experiment, which can be measured by an inversion-recovery pulse sequence. As shown in Figure 3.12A, the sequence begins with a recycle delay long enough to ensure all magnetization is at equilibrium along the *z*-axis (M_z^0). The initial magnetization is inverted by a 180º pulse, shifting it to –*z* axis ($-M_z^0$), and followed by different values of a recovery delay (of the magnetization back to *z*, $M_z(t)$), *t*. Finally, a 90º pulse (Figure 3.12B) shifts the recovered M_z from the *z*-axis to the *xy*-plane, converting it into observable transverse magnetization. An exponential curve is obtained (Figure 3.12C) as the magnetization returns to equilibrium in the *z*-axis, which depends on the different values of the previous recovery delay *t* (a longer delay allows more magnetization to return to *z* before the 90º pulse, thus the intensity obtained gets closer to M_0). Therefore, the magnetization in the z-axis is used to obtain the relaxation rate constant (R_z) or T_1 relaxation: $T_1 = {}^1/_{R_z}$. The full mathematical expression of the evolution of the magnetization can be found in Keeler[125], and it shows that this experiment predicts the time evolution as:

$$ln\left[\frac{M_z(t) - M_z^0}{-2M_z^0}\right] = -R_z t \qquad \textit{Equation 3.12}$$

If the signal intensity is considered proportional to the *z*-magnetization, Equation 3.12 can also be expressed as:

$$ln\left[\frac{S(t) - S^0}{-2S^0}\right] = -R_z t \qquad \textit{Equation 3.13}$$

where the signal intensity at the moment of magnetization equilibrium (M_z^0) is S^0, and the signal after the recovery delay is $S(t)$. Plotting the natural logarith *versus t* in Equation 3.13 results in a straight line of slope $-R_z$.

A quick estimate of the T_1 value is possible from this inversion recovery experiment, where instead of changing *t* with random values, the value resulting in

no signal due to equilibrium of the magnetization (t_{null}) turns $S(t)$ to zero, thus modifying Equation 3.13 into:

$$\ln\left[\frac{1}{2}\right] = -R_z t_{null} \quad \text{or} \quad R_z = \frac{\ln 2}{t_{null}} \quad \text{or} \quad T_1 = \frac{t_{null}}{\ln 2}$$

This latter experiment is the easiest to use because $T_1 \approx 1.4\ t_{null}$.

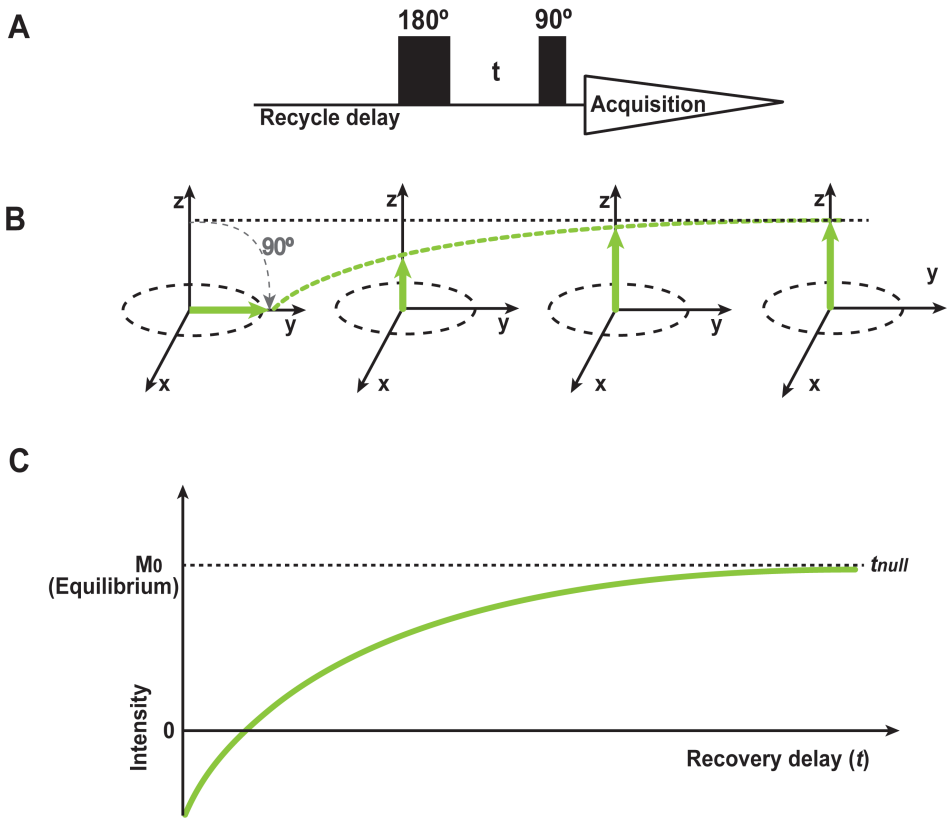


Figure 3.12. – Inversion-recovery experiment to obtain T_1 relaxation. **A)** The experiment begins with a recycle delay so all the magnetization is at equilibrium (*z*-axis, M_z), then a 180º pulse inverts the magnetization, followed by a recovery delay (*t*) of different lengths and finally a 90º pulse (magnetization is shifted to the transversal *xy*-plane) so the magnetization becomes measurable. **B)** The figure shows the recovery of the magnetization after the 90º pulse, with different values due to the change on the recovery delay *t*. **C)** Exponential curve obtained due to the return of the magnetization to equilibrium (M_0), at time t_{null}.

3.5.2.- Transverse (T_2) Relaxation

The other main source of relaxation, the T_2, transverse or spin-spin relaxation, results in a decrease of the transverse magnetization in the *xy* plane to zero, by the loss of coherence of the sample in the *xy* plane that was created by the RF pulse. There are two main sources of T_2 relaxation, the first one is the same

local magnetic fields that can originate the T_1 relaxation, flipping the spins from the *xy* to the *z* plane and recovering the longitudinal relaxation. Thus, anything causing T_1 relaxation may also contribute to the T_2 relaxation by depleting the magnetization on the *xy* plane. The second source requires no transitions; it is due to differences on the local magnetic fields at different locations in the sample, which would vary the Larmor frequency of the spins, and so, decrease the coherence.

Both T_1 and T_2 relaxation depend directly on the correlation time (τ_c) as shown in Figure 3.13 . τ_c was defined in Chapter 3.2.3 as a factor dependent on the size of the molecules in solution and the solvent viscosity (which would involve temperature and solutes concentration). Small molecules (such as free water molecules) have low values of τ_c, and follow fast motions, therefore relaxation processes are inefficient by the mismatch with the spin Larmor frequency (ω_0), the energy of the motion is spread over a wide frequency range. The dependence of T_1 with τ_c, is V-shaped, showing the fastest relaxation at $\tau_c = 1/\omega_0$. For values of the correlation time that are larger or smaller than $1/\omega_0$, T_1 increases because of the mismatch with the Larmor frequency, both for small molecules and large molecules, and/or more viscous solutions. However, T_2 relaxation time decreases as τ_c increases with no minimum. Therefore, transverse relaxation processes are more effective for larger molecules.

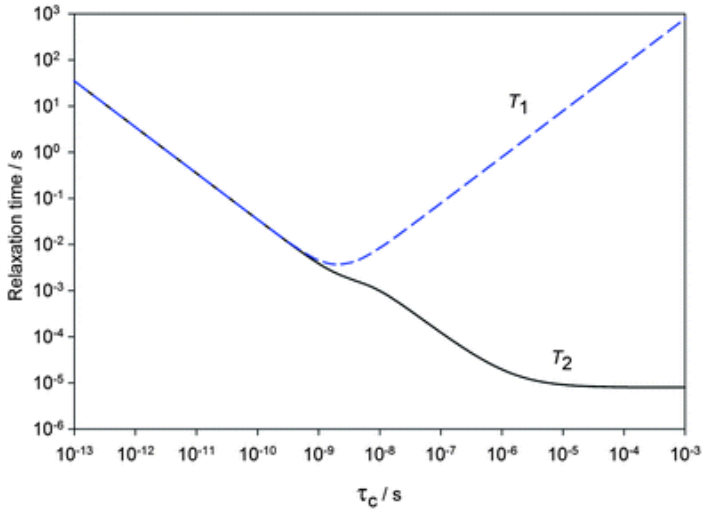


Figure 3.13. – NMR relaxation times T_1 and T_2 as a function of the correlation time (τ_c). The lower values of τ_c correspond to small molecules with fast motion in a non-viscous solution, and the higher τ_c values correspond to larger molecules (such as proteins) and/or viscous solution. The inflection point of T_1 corresponds to $\tau_c = 1/\omega_0$. Figure adapted from Bloembergen et *al*.[90]

T_2 relaxation is usually measured by the classical spin-echo experiment with the Carr-Purcell-Meiboom-Gill (CPMG) sequence[126,127]. This sequence consists of an initial 90º pulse that places the magnetization from *z* (M_z) to *xy* (M_{xy}) plane, where it evolves during the delay period, and is shifted to its mirror image by a 180º pulse, followed by another delay. The delays should be short to avoid broadening of the peak signal by molecular diffusion. This "delay-180º-delay" block of the experiment (also known as spin-echo) is used to determine the M_{xy} decay, and is repeated several times. T_2 relaxation arises from the exponential decay of the amplitude as the number of spin-echoes increases.

3.5.3.- Carr-Purcell-Meiboom-Gill (CPMG) Relaxation Dispersion (RD)

X-Ray crystallography and NMR are powerful techniques that provide detailed protein structural information. However, molecular dynamics and protein internal motions are also intimately involved in protein function, and dynamics can only be revealed through meticulous NMR analysis. Molecular motion can lead to protein conformations different from the native state, often scarcely populated, which may be crucial for protein function by, for example, opening to the solvent a protein region usually closed in the main state, or transmitting a signal upon binding to another molecule, etc. Several NMR techniques have recently been developed to perform detailed characterization of protein dynamics spanning different timescales – from picoseconds-nanoseconds, like Residual Dipolar Coupling (RDC) and Nuclear Spin Relaxation (NSR), to microseconds-miliseconds, such as Carr-Purcell-Meiboom-Gill relaxation dispersion (CPMG-RD), chemical exchange saturation transfer (CEST) and lineshape analysis, and even longer events, that may take seconds, like Real-Time (RT) NMR. The work in this Thesis is mainly focused on CPMG-RD experiments, a technique particularly valuable to characterize dynamics in the microsecond to millisecond timescale. CPMG-RD is sensitive to conformations present at very low levels (> 0,5%) and can hence detect sparsely populated conformations that are "invisible" to most other biophysical techniques. CPMG-RD was first described in 1950's[128,129], and its extension to protein dynamics began in the 1990's in conjunction with pulse sequence development[130] and isotopic labelling strategies[131].

Some biochemical processes that may involve the participation of “invisible” states are ligand binding and release[132-135], folding and unfolding events[59,136-138], allostery[139] and catalysis[140-142]; thus, the need to detect and characterize these excited states becomes key to obtain mechanistic knowledge. When these processes fall into the millisecond-microseconds timescale, CPMG-RD experiments are able to access these “invisible” sparsely populated structures. These invisible states (also called minor or excited states) should have a population (p_B) greater than 0.5%[124,143] of the total molecules to be detected, exchanging with a major and visible state (ground), p_A[124,144]. The sum of both populations is the total population

$$p_A + p_B = 1$$

The populations of the interconverting states can be rewritten in terms of the exchange rate constants (Equations 3.14A and B)[145], k_{AB} and k_{BA} as

$$p_A = \frac{k_{BA}}{k_{AB} + k_{BA}} \quad Equation\ 3.14A \quad ; \quad p_B = \frac{k_{AB}}{k_{AB} + k_{BA}} \quad Equation\ 3.14B$$

These constants have already been explained in Chapter 3.4 and the definitions of A and B states applied in that Chapter will correspond to the major and minor state, respectively. Furthermore, it is especially interesting that the chemical shifts of the excited state can be determined using CPMG-RD[144], opening up an avenue for the three-dimensional structure determination of these “invisible” states.

Intra- or inter-molecular exchange processes (Chapter 3.4) have to be studied in detail through the information contained in the increased linewidth (often by hundreds of Hz) of the cross-peaks of the spectra, to obtain information on the minor state. In addition to regular protein dynamics[146], the exchange process contributes to the increased linewidth, which determines the effective relaxation rate. Therefore, the linewidth provides information on the rates of interconversion between major and minor states (kinetics), chemical shift differences (structure) and the population of the exchanging states (thermodynamics)[124].

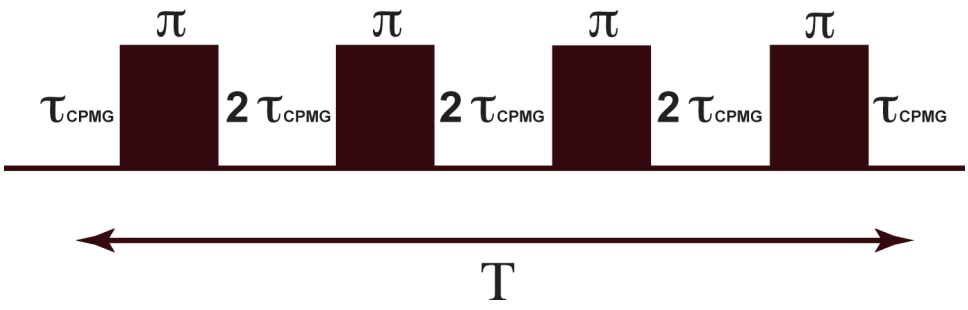


Figure 3.14. – A basic CPMG-RD pulse scheme employing 180º (π) pulses, where $2\tau_{CPMG}$ is the delay between pulses. In this fragment of the pulse sequence, an *n* number of spin-echo modules $(\tau_{CPMG} - 180º - \tau_{CPMG} - \tau_{CPMG} - 180º - \tau_{CPMG})_n$ applied during a constant time (T) on the transverse plane results in the refocusing of the magnetization.

CPMG-RD experiments are based on spin-echo experiments as shown in Figure 3.14. In the simplest case, after an initial $90º_y$ pulse to generate transverse magnetization, a series of π pulses ($180º_x$) applied during a constant time (T) with the rate $\nu_{CPMG} = 1/(4\tau_{CPMG})$ (where $2\tau_{CPMG}$ is the delay between the 180º pulses) is used to refocus the magnetization, partially or completely quenching the exchange effects[147]. To understand the refocusing effect, we can picture a system with two conformations, A and B, and follow an NMR-active nucleus in both. For each conformation in the absence of exchange, the magnetization will precess with different frequencies during τ_{CPMG}, accumulating a certain phase; the application of a single 180º pulse will invert this accrued phase, completely refocusing the magnetization after the same time τ_{CPMG} for which it had been evolving prior the pulse. When the two conformations are in exchange process, the nucleus will jump stochastically between the different frequencies of evolution of magnetization. Because of the stochastic nature of the jump, each molecule will accrue a different phase during the evolution periods before and after the 180º pulse, resulting in incomplete refocusing across the ensemble and a higher effective relaxation rate. As ν_{CPMG} increases (by the increase of the 180º (π) pulsing rate and the decrease of the $2\tau_{CPMG}$ delay while the total exchange time T is constant, as shown in Figure 3.15, the time τ_{CPMG} available for the frequency difference between the two conformers to evolve is shorter, suppressing the dephasing, and therefore, improving the refocusing effect and increasing peak intensity. The 180º pulses are always applied in even numbers to compensate for pulse imperfections. Thus, this procedure allows to monitor the signal intensity as a function of the repetition rate of ν_{CPMG}, $/(\nu_{CPMG})$, recorded during the constant time (T) block, compared with a reference signal, I_0 (intensity of a peak in the reference spectrum recorded without

the constant time period T), providing an effective transverse relaxation rate ($R_{2,eff}$) that quantifies the signal decay rate according to the relation[148]

$$R_{2,eff} = -\frac{1}{T} ln \frac{I(\nu_{CPMG})}{I_0} \qquad Equation\ 3.15$$

A typical relaxation dispersion profile of $R_{2,eff}(\nu_{CPMG})$ obtained from Equation 3.15 is shown in Figure 3.15 with a decrease of the effective transverse relaxation rate as the rate of π pulse application increases. These profiles are modelled according to the Bloch-McConnell equations[149], and depend non-linearly on the exchange parameters p_B, k_{ex} and on the absolute chemical shift differences $|\Delta\omega_{AB}| = |\omega_B - \omega_A|$, where ω_A and ω_B are the chemical shifts of states A (major) and B (minor), respectively, expressed in ppm units, or in rad/s as $|\Delta\varpi_{AB}| = |\varpi_B - \varpi_A|$. Nevertheless, a flat profile indicates non-exchanging nuclei.

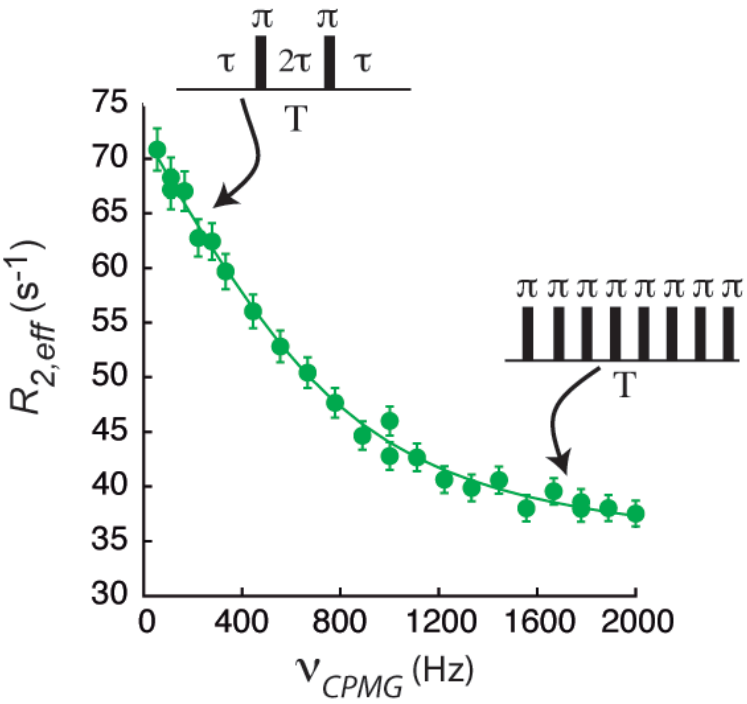


Figure 3.15.- Relaxation dispersion profile obtained from a CPMG-RD experiment where the increase on the application rate of π pulses (ν_{CPMG}) during a constant time (T) decreases the dephasing resulting from exchange between two populations, reflected in the decrease of $R_{2,eff}$.

The CPMG-RD experiments are executed as 2D HSQC experiments of the different nuclei where each 2D is acquired with a different ν_{CPMG} value. Usually it is

possible to reach frequencies up to ν_{CPMG} = 1000 Hz for ^{15}N, 1500 Hz for ^{13}C and 2000 Hz for $^{1}H^{N}$ with the limits defined by hardware conditions[150]. At least three duplicate points at selected ν_{CPMG} values are included to evaluate errors in $R_{2,eff}$ for all nuclei. After analysing the NMR data with the NMRPipe suite of programs[151], and assigning the peaks with NMRDraw[151] and Sparky[152], the peak intensities are quantified using the program FuDA[153], which fits the lineshape across ν_{CPMG} values globally. Intensities are converted to $R_{2,eff}$ using Equation 3.15. Uncertainties in $R_{2,eff}$ ($\Delta R_{2,eff}$) are estimated from the measured peak intensity as shown in Equation 3.16[124]

$$\Delta R_{2,\mathrm{eff}} = \frac{1}{\mathrm{T}} \frac{< \Delta I >}{I(\nu_{CPMG})} \quad Equation\ 3.16$$

where <ΔI> provides the average standard deviation from duplicate points.

The profiles obtained from FuDA[153] are visualized to select nuclei with exchange contributions to the transverse relaxation rate of $R_{ex} > 5\ s^{-1}$, where $R_{ex} = R_{2,eff}(\nu_{CPMG} = \nu_{min}) - R_{2,eff}(\nu_{CPMG} = \infty)$, similar to Figure 3.15. These selected atoms are used as input for the program ChemEX[154,155], that simulates the evolution of magnetization during the CPMG periods by numerically propagating the Bloch-McConnell equation[156] and extracts the exchange parameters ζ (ζ = {k_{ex}, p_B and $|\Delta\omega|$}), by minimization of X^2 (ζ)[124] (Equation 3.17)

$$\mathrm{X}^2(\zeta) = \sum \frac{(\mathrm{R}_{2,\mathrm{eff}}^{\mathrm{clc}}(\zeta) - \mathrm{R}_{2,\mathrm{eff}})^2}{(\Delta \mathrm{R}_{2,\mathrm{eff}})^2} \quad Equation\ 3.17$$

where $R_{2,eff}^{clc}$ corresponds to the theoretical data, which considers the initial magnetization of the ground state and the magnetization after the CPMG sequence as explained in detail elsewhere[124].

This procedure is performed for each CPMG experiment for each nucleus at each magnetic field strength. Once the k_{ex} and p_B parameters show a clear minimum in the X^2 distribution, these values are fixed to obtain the $|\Delta\omega|$ values for every single atom.

Nevertheless, the exchange rate accessible to CPMG-RD experiments is in the 100 $s^{-1} \leq k_{ex} \leq 2000\ s^{-1}$ range, and for faster processes the dispersion profiles alone are not enough to obtain confident exchange parameters and chemical shift differences due to the correlation between p_B and $\Delta\omega$ in Equation 3.18, even when reasonably accurate k_{ex} values can be fit[157].

$$R_{2,eff}(\nu_{CPMG}) = \left(1 - 4\,\frac{\nu_{CPMG}}{k_{ex}}\right)\frac{p_A p_B(\Delta\omega)^2}{k_{ex}} + R_{2,eff}\,(\nu_{CPMG} \to \infty) \quad Equation\ 3.18$$

However, it has been recently shown that faster exchanging processes up to 4000 – 6000 s^{-1}, can be studied including a combined quantitative analysis of relaxation dispersion profiles at different magnetic fields and ground state chemical shifts measured either as a function of different static magnetic fields or as a comparison of peak positions recorded in HSQC/HMQC experiments[157]. The exchange broadening and the exchange induced shift of the chemical shifts on the magnetic field depend on the exchange parameters in different ways, therefore, their global fitting including the relaxation dispersion profiles at different fields with ChemEX breaks the correlation between p_B and $|\Delta\omega|$ in Equation 3.18. A detailed explanation of the different influence of the chemical exchange on the chemical shifts and the relaxation rates is provided by Vallurupalli *et al.*[157]. This facilitates the accurate extraction of both p_B and so $|\Delta\omega|$ values, as well as the peak positions of the invisible state (ω_B).

For gpW protein (see Chapter 8) the chemical shifts were obtained from the assignment of triple resonance experiments at 500 and 600 MHz magnets at 1ºC. At the same temperature - checked with a thermocouple inserted into an NMR tube - the Relaxation Dispersion experiments for the different nuclei ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$,$^{13}C^{\alpha}$, and methyl-^{13}C were performed at 500 and 800 MHz fields according to pulse schemes previously described[148,158-162]. The only modifications on the pulse sequences were to remove the side-chain correlations of the Arg residues in the ^{15}N and $^{1}H^{N}$ experiments through a (backbone-amide-specific) REBURP pulse of 2.00 or 1.25 ms for the 500 and 800 MHz fields, respectively, on the ^{15}N channel during the first INEPT transfer period; and in the $^{1}H^{N}$ experiments, the amide magnetization at the beginning of the constant time (T) CPMG element was antiphase with respect to ^{15}N. Any NMR experiment has a prescan delay previous to the pulse scheme to allow the system to return to equilibrium, this delay is 2.5 s in all RD experiments. The constant-time CPMG relaxation delays for the RD profile experiments for each

nucleus were 30, 18, 24,16 and 30 ms for ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and methyl-13C, respectively. For the CPMG-RD experiments dispersion data were collected for each nucleus with 17 CPMG fields ranging from 33.3 to 1000 Hz for ^{15}N (500 and 800 MHz), 21 CPMG fields ranging from 55.6 to 2000 Hz for $^{1}H^{N}$ (500 and 800 MHz), 10 (12) CPMG fields from 83.3 to 833.3 (1000) Hz for $^{13}C^{O}$ at 500 MHz (800 MHz), 10 (13) CPMG fields from 62.5 to 1000 Hz for $^{13}C^{\alpha}$ at 500 MHz (800 MHz), and 17 CPMG fields ranging from 66.7 to 2000 Hz for $^{13}CH_3$ (500 and 800 MHz). A minimum of three duplicates on certain ν_{CPMG} values were acquired to estimate the $R_{2,eff}$ errors.

3.5.4.- Determination of the "Invisible" State Chemical Shifts

The experiments that enable us to break the correlation between p_B and $|\Delta\omega|$ also allow to obtain the signs of these chemical shifts differences for the different nuclei by comparison of the peak position in different experiments:

- The signs for ^{15}N can be obtained from [$^{1}H^{N}$-^{15}N]-HSQC experiments at different fields (also referred as H(S/S)QC experiments), and from [$^{1}H^{N}$-^{15}N]-HMQC and [$^{1}H^{N}$-^{15}N]-HSQC data sets at the same field (H(S/M)QC experiments), both HSQC and HMQC experiments are explained in Chapter 3.3.1. The signs from H(S/S)QC experiments arise from the difference in the field strength, which results in a small shift, σ (Hz) or $\overline{\sigma}$ (ppm) as shown in Figure 3.16. In the case of H(S/M)QC experiments, arises from the different manifold of exchange averaging in the single- and multiple-quantum experiments, which is non-linear, giving rise to the small shift Ω (Hz) or $\overline{\Omega}$ (ppm)[99] (Figure 3.16).

- The shifts for $^{1}H^{N}$ can be also obtained from the $^{1}H^{N}/^{15}N$ zero- and double-quantum coherence relaxation dispersion profiles[163], and from a comparison of the direct-dimension chemical shifts from regular HSQC experiments acquired at different fields[164].

- In a similar way, the signs of the $\Delta\omega$ of $^{13}C^{O}$, $^{13}C^{\alpha}$ and $^{1}H^{\alpha}$ can be determined from a comparison of resonance positions, for $^{13}C^{O}$, through $^{1}H^{N}$ -$^{13}C^{O}$ data sets recorded at different fields where the $^{13}C^{O}$ chemical shifts derives from evolution of either single- or multiple- quantum ($^{1}H^{N}$-$^{13}C^{O}$ or ^{15}N-$^{13}C^{O}$) coherence[143].

In the same way, a series of 2D ($^{13}C^{\alpha}$-$^{1}H^{\alpha}$) correlation maps recorded at different magnetic fields provide the signs for $^{13}C^{\alpha}$ and $^{1}H^{\alpha}$, which requires excellent solvent suppression since these protons may be very close to the water line[143].

For the studies on gpW protein (see Chapter 8), the fields with different strengths were 500 and 800 MHz for the sign determination experiments.

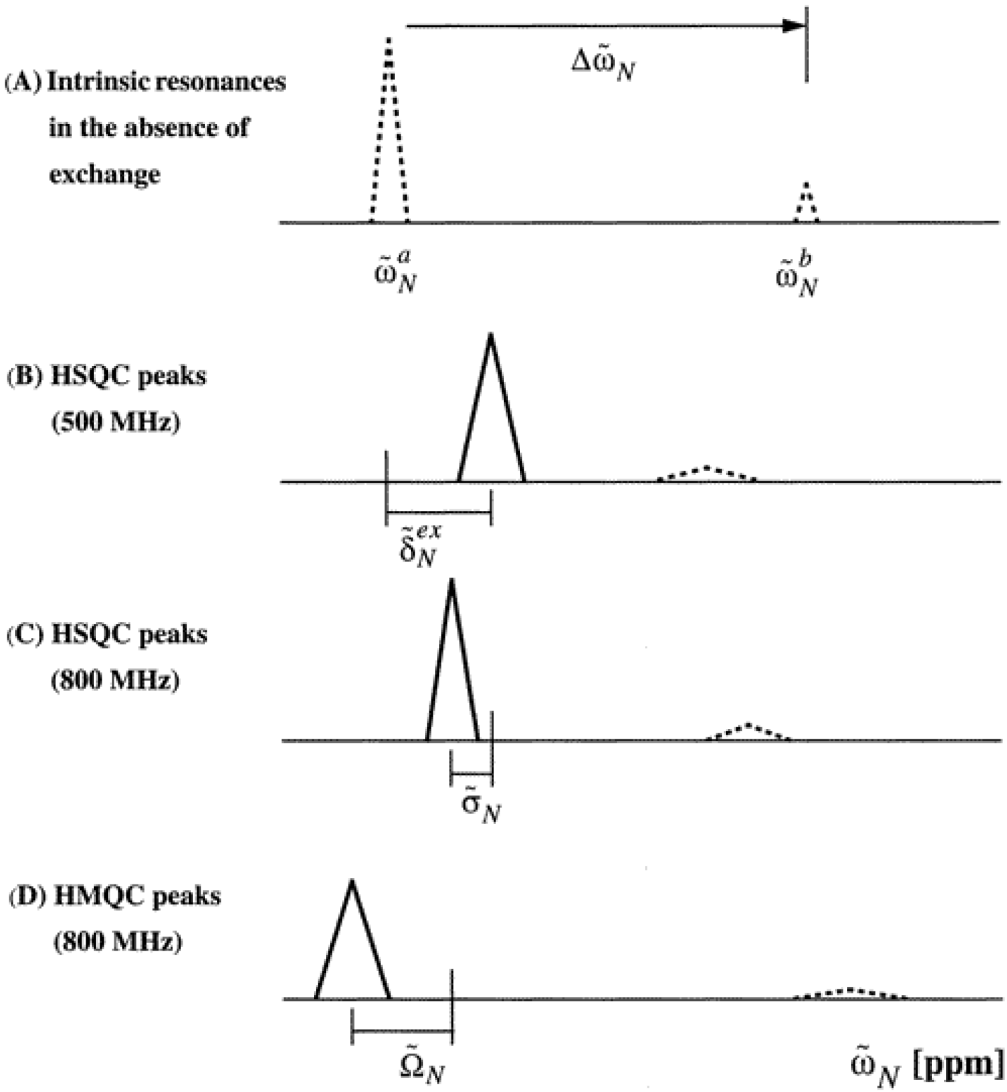


Figure 3.16.- (A) Representation of the peaks without chemical exchange and (B, C, D) relative shifts upon undergoing exchange processes in different NMR experiments and fields. Solid lines represent observable peaks, whereas dashed lines correspond to non-detectable peaks. $\overline{\delta}$ cannot be determined due to the lack of a reference peak, but $\overline{\sigma}$ and $\overline{\Omega}$ are experimentally measurable. The peaks in the HMQC (D) can be shifted to the left or to the right depending on the parameters that determine $\overline{\Omega}$. Figure adapted from Skrynnikov *et al.*[99]

The determination of the sign and the previously obtained $\Delta\omega$ values enables the exact determination of the peak positions for the excited or "invisible" state, impossible to unveil in any other way.

3.6.- Sample Preparation for NMR Experiments

3.6.1.- WW Domain, R3H Domain and HP35 from Villin Samples

The experiments for these samples were performed on a 600 MHz spectrometer with a triple resonance 5 mm TXI-probe and z-axis gradient, which temperature is controlled by a BTO temperature control unit. Since some of the experiments (the different samples of HP35 subdomain from villin) were performed at different temperatures to monitor the thermal unfolding at atomic level, a precise control of each temperature was greatly important. Thus, calibration of the probe was performed in a 5 mm NMR tube with an ethylene glycol sample, valid for temperature calibrations between 273 and 416 K. Each temperature was stabilized for 30 minutes with a constant air flow and a 1D ^{1}H experiment run to obtain the chemical shifts, which difference at the maximum height of the peaks can provide the real temperature of the probe following Amman *et al.*[165].

The samples for the different proteins were prepared as:

- WW domain was prepared at 0.5 mM protein concentration in H_2O with 0.05 mM DSS and 10% of D_2O at pH 4.7 in pressured tube. The low concentration of this protein for NMR experiments is due to solubility problems at high concentrations.

- R3H domain samples had a protein concentration of 0.8 mM and were dissolved in 20 mM phosphate buffer with 10% D_2O, 0.05 mM DSS, 1 mM TCEP and 0.1 mM sodium azide (NaN_3) at pH 7.0 in a pressurized tube at 8 bar.

- The variants of HP35 subdomain from villin have different labelling patterns (see Table 7.1). All samples had a protein concentration of 1 mM in 30 mM glycine buffer (with glycine-HCl D5) with 0.1 mM DSS, 0.1 mM NaN_3 and 10% D_2O / 90% H_2O at pH 3.0. Tubes were pressurized at 8 bar to avoid boiling issues at high temperatures.

3.6.2.- Samples for gpW Protein Experiments

The gpW samples for NMR had a concentration of 1 mM of protein. For pH 3.5 the protein was dissolved in 20 mM glycine buffer with 1 mM NaN_3 and 10% D_2O / 90% H_2O, while for pH 6.5 the buffer was 25 mM sodium phosphate, 50 mM NaCl, 0.5 mM EDTA, 0.5 mM NaN_3 and 10% D_2O/ 90% H_2O.

To slow down the folding rate of gpW protein (1.7μs at 355 K), some of the first tests consisted on checking different temperatures (measured with a thermocouple inserted into an NMR tube as previously mentioned) or adding substances. The tested substances were:

A) Urea:

The concentrations of 1 M, 1.5 M and 2 M were checked. The urea was previously purified to remove ionic species using a mixed bead resin, which avoids interactions of the urea with the positively charged residues. The urea concentration in the buffers was obtained through the refractive index and a calculator (http://sosnick.uchicago.edu/gdmcl.html).

B) Glycerol:

The glycerol samples were prepared by adding buffers with the desired glycerol concentration (29% and 42% were tested), determined through the refractive index.

For both, urea and glycerol samples, the buffers contained the same components at pH 3.5, with the addition of the corresponding concentration of one of them.

Chapter 4.- General Expression and Purification Methods

All proteins were purified in the laboratory except for HP35, produced by synthesis and provided by Dr. Robert Tycko's laboratory. Therefore, this Chapter concerns exclusively to the other studied proteins excluding this synthetized HP35 subdomain from villin protein.

The studied proteins have been cloned into the expression vector pSUMO (Small Ubiquitin-like MOdifier) by an external biotechnological company (TOP Gene Technologies, Inc.), producing a fusion protein (SUMO+protein), that was expressed in *Escherichia coli* BL21(DE3) cells, a commonly used strain of cells for high protein expression. The addition of SUMO is important due to the small size of the studied proteins because it improves protein stability and solubility[166], and allows an easier purification protocol through its Histidine-Tag. To improve the low yield of the WW domain, BL21(DE3) pLys competent cells were transformed to incorporate this protein fused to SUMO. This cell strain provides an improved control in case the protein is toxic for the cell, a possible cause for the destruction of the fused protein, that may have produced the low yield.

4.1.- Cell Growth

The cell culture started with an overnight preinocule in a LB (Luria-Bertani) broth with 0,27 mM ampicillin (the plasmid is ampicillin resistant) and 0,1 mM chloramphenicol for the pLys strain (this strain is resistant to both antibiotics), at 37 ºC. Next day the preinocule was added with a 1:1000 scale to LB with the same concentration of antibiotics. The culture was incubated and stirred until the desired Optical Density (OD) of cells at 600 nm for each protein was reached, inducing then with isopropyl β-D-thiogalactopyranoside (IPTG) to increase the protein production while stirring at the same temperature of 37 ºC or lower in the case of the WW domain (check Table 4.1 for OD and IPTG concentration). A few hours after induction, depending on the protein (Table 4.1), the cells were harvested by centrifugation at 5000 rpm for 40 minutes at 4 ºC, for gpW the speed was 8000 rpm for 30 minutes.

	WW	R3H	gpW
Induction OD	1,6	1,3-1,9	0,8
IPTG concentration (mM)	0,42	0,42	0,42
Temperature after induction (ºC)	30	37	37
Time to harvest after induction (h)	Overnight	Overnight	4

Table 4.1.- Comparative of the OD, IPTG concentration and temperature after induction for each purified protein.

4.2.- Cell Lysate

This process is slightly different for each protein:

a) **WW and R3H:** The pellet cells are resuspended in 10 mM Tris, 150 mM NaCl and 10 mM Imidazole buffer with 1 mM PMSF (phenylmethylsulfonyl fluoride, as serine protease inhibitor) and lysed by sonication on ice for 6 times during 10 seconds, waiting 1 minutes each time to avoid heating. Then, ultracentrifugation at 30000 rpm for 1 hour separated the soluble proteins from the cell debris.

b) **gpW:** For this protein, the cells lysis was not complete with the buffer used for the previous proteins, so the pellet was resuspended in the same lysis buffer adding 8 M urea and stirred until homogeneous, then ultracentrifuged at 18000 rpm for 45 minutes. The supernatant is diluted to half (to decrease the urea concentration) and sonicated because its high density (5 times for 10 seconds with 1 minute waiting time).

4.3.- HPLC Purification

The supernatant containing the soluble proteins was filtered and injected in affinity nickel column in the HPCL. The proteins in the lysate were washed with a 20 mM Tris, 150 mM NaCl and 10 mM imidazole buffer except for SUMO fused proteins that contain a His-tag that attaches to the column. The fused-protein is eluted afterwards with a imidazole gradient[167,168] from 10 mM up to 500 mM in

buffers with the same content on Tris and NaCl. The fused-proteins come out around 50-60% of the gradient.

A three steps dialysis in the same buffer without imidazole is then performed to remove it before a second nickel column step. Previous to the second column step, the fusion protein is cut with ~0,1 mg of ULP-1 protease per 1 mg of fused protein sample that releases the protein from the SUMO. In the nickel column this time the protein is washed and the SUMO with the His-Tag remains attached.

The free protein (without the tag) is now dialyzed against water and after lyophilized. However, for gpW the procedure included a third HPLC step with a C4 reverse phase column, with a gradient from 5% to 95% acetonitrile (ACN) in water and 0,1% trifluoroacetic acid (TFA), to remove salt and protein impurities that may appear, and finally lyophilized as the other proteins.

After purification, the adequate protein size is checked by MALDI-TOF mass spectroscopy.

4.4.- Isotope labelling for NMR experiments

To obtain the samples to perform the different NMR experiments described in Chapter 3, the procedure was similar except for the use of minimal media with $^{15}NH_4Cl$ and $^{13}C_6$-glucose (from Cambridge Isotope Laboratories, Inc.) as the only sources of nitrogen and carbon to get uniformly ^{15}N and/or ^{13}C labelling. To perform some of the Relaxation Dispersion NMR experiments for certain probes specific labelling schemes were necessary, therefore, selective labelling was obtained by adding [1-^{13}C]glucose for methyl-^{13}C-labeling[169], or [2-^{13}C]glucose for $^{13}C^{\alpha}$-labelling[169].

4.5.- Protein Concentration and Molecular Weights

The protein concentration for the different experiments was obtained according to the Beer-Lambert law, by absorbance at 280 nm with the molar extinction coefficients (ε): 11380, 2560, 5500 and 1490 $M^{-1}cm^{-1}$, for WW domain, R3H domain, HP35 villin subdomain and gpW protein, respectively.

The molecular weight for the different proteins was: 4284,8 g/mol for WW domain, 10143,3 g/mol for R3H domain, and 6976 g/mol for gpW protein. These values were corrected for the different isotopic labelling. The molecular weight of HP35 villin subdomain is a particular case since the different samples had different labelling patterns, thus the molecular weight was slightly different for each one: 4093,44, 4088,48, 4094,44, 4078,55 and 4097,8 g/mol for variant 1, 2, 3, 4 and 5, respectively.

PART III:

Proteins Under Study

During this Thesis different proteins and domains with diverse content on α-helix and β-sheet have been studied. In all cases, the small size favours its study as models for the different secondary structure content. These proteins/domains are either candidates of downhill folding (WW and R3H domains) or already included as downhill folders (HP35 and gpW protein) according to the criteria mentioned in Chapter 1.3.2. The study of the different cases begins with a thermodynamical study including the techniques explained along Chapter 2, unless previously characterized, and after detailed NMR study.

Chapter 5.- WW Domain (WW3 from Nedd-2)

5.1.- General Features

WW domains, with about 40 residues or even less, are considered the smallest independently folding domains made of β-sheet, with three antiparallel strands[170]. These domains are named after the two conserved Tryptophan (W) residues included in their compact core located at about 20-22 residues distance. The function of these domains is to mediate protein-protein interactions through binding to ligands with proline-rich motifs. These domains operate in a variety of signalling pathways, involving membrane, cytoplasm and nucleus, like the transport of proteins, electrolytes and signal receptors, as well as in DNA transcription and RNA processing[171], and also in the interaction with proteins involved in important neurodegenerative diseases, like the huntingtin of Huntington disease[172], and the AAP (Alzheimer's disease amyloid precursor protein) of Alzheimer's disease[173]. Furthermore, WW domains have been widely studied since some WW domain-ligand complexes were suggested as amenable for gene and drug therapy[174,175].

Classically, folding studies have been focused on α-helical structures, however, during the late '90s and the first decade of the XXI century, numerous studies on β structures were performed highlighting the different folding speed between these two kind of structures, up to 30 times slower for these last ones[176,177]. The huge difference on the folding rate arises from the proximity in the helix of the residues participating in the folding through the creation of hydrogen bonds every three residues, and the large number of

productive ways of building it up, *versus* the sensitivity of the β-hairpin to the position of the stabilizing residues, which makes this critical interaction less likely to occur and therefore substantially decreases its folding rate[178]. Nevertheless, WW domains were the first fast β-sheet folders under study[179], showing the fastest folding rates for natural beta-structures[180], and their engineering to speed the folding rate up has been an important research topic in both experimental[181] and simulation[182] areas, for example, molecular dynamic simulations that found a WW domain folding in about 20 μs[42] and mutants as fast as 4.3 μs[182]. It has been proven that the first steps in the folding of hairpins are different than those typical for the helices: at physiological temperatures is triggered by the β-turn[178] (usually made of 4 or 5 residues connecting the β-strands, described in more details in the next paragraphs) by bringing together via hydrogen bonding the residues of the different strands involved on the stabilization of the structure. Hence, any mutation on this region may accelerate the folding rate and/or increase the thermodynamical stability, but also it may eliminate the function of the WW domain[183]. At higher temperature, or upon mutational destabilization, the hydrophobic residues come to play for these hairpins, even if they do not fully dominate the rate[184]. Experiments on larger β-sheet proteins have shown separate phases for fast hydrophobic collapse and strand formation[185]. Thus, WW domains represent small, fast folding and mainly β-sheet structures, an important piece in the puzzle of the folding knowledge, and interesting candidates of the downhill scenario (Chapter 1.3.2), with some domains already included as downhill folders[45,186].

For this work the domain under study is the third WW domain from Ubiquitin-protein ligase Nedd4-2 (see Figures 5.1 for location, Figure 5.2 for sequence and Figure 5.3 for structure). As Figure 5.1 shows, the Nedd4-2 is located at the cellular cytoplasm and down-regulates the epithelial Na^+ channels (ENaC) by its ubiquitin-protein ligase activity in response to increased intracellular Na^+ concentration. Nedd4-2 controls cell surface stability of the channels and its ubiquitylation[187]. It has 4 WW domains, the WW3 domain has a high-affinity for the PY (Proline-Tyrosine) motifs present in the three subunits (αβγ) constituting the ENaC. This interaction triggers the ubiquitylation of the α and γ subunits at the NH_2 termini and therefore targets the protein for lysosomal degradation[188]. As consequence, the concentration of ENaC at the membrane level drops leading to a decreased Na^+ flow into the cell.

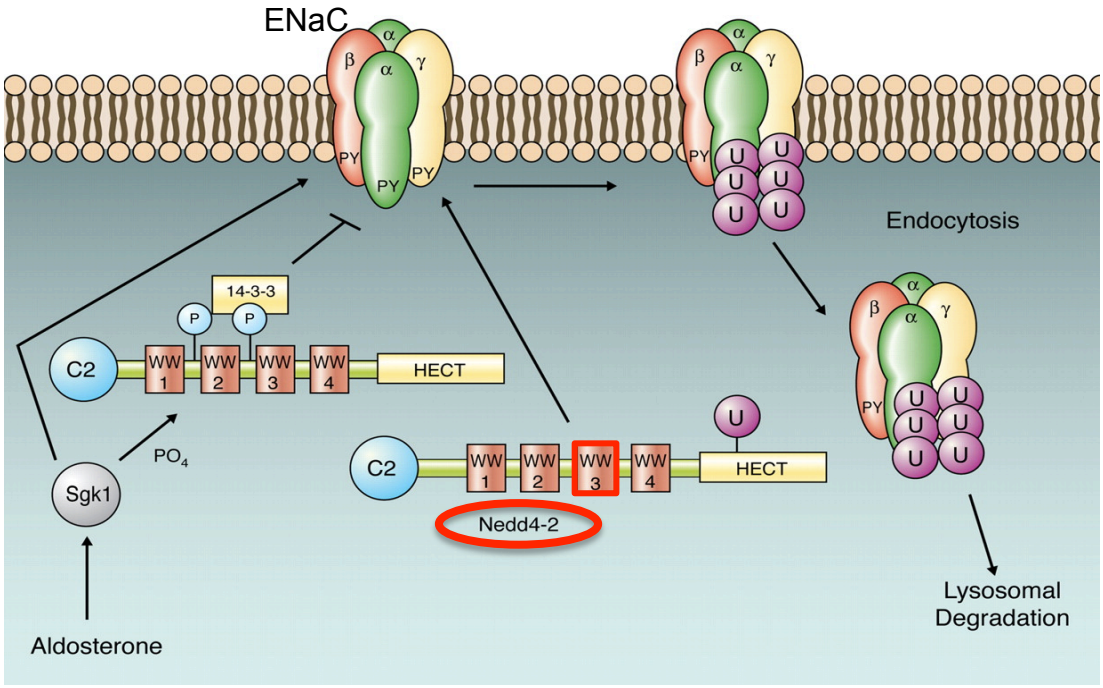


Figure 5.1.- The WW domain under study, the WW3 of Ubiquitin-protein ligase Nedd4-2 (highlighted in red), is involved in the regulation of the epithelial Na^+ channel (ENaC). The Nedd4-2 is formed by different domains including 4 WW domains that interact with the Proline-Tyrosine (PY) motifs of the ENaC subunits, ubiquitylating the NH_2 termini of α- and γ-ENaC subunits, and resulting in its lysosomal degradation. However, if decrease of ENaC is not necessary, phosphorilation of the Nedd4-2 allows the accumulation of the channel at the plasma membrane. Figure adapted from Staub, O. and Rotin, D.[188]

The sequence of this 36-residue domain WW3 of Nedd4-2 from house mouse (Figure 5.2) for this work corresponds to the PDB 1WR7, without the 5 initial residues.

Figure 5.2.- Sequence of WW3 from Nedd4-2. Black lines represent residues without secondary structure assignment; yellow arrows, β-strand; red, α-helix; and green for turns.

This domain has the two typical Trp residues located at positions 8 and 30 and together with two Phe residues in the central β-sheet, constitutes a hydrophobic minicore. The secondary structure content is 31% of the β-sheet characteristic of WW domains, but also a minor content of α-helix (7%) as shown in Figure 5.3.

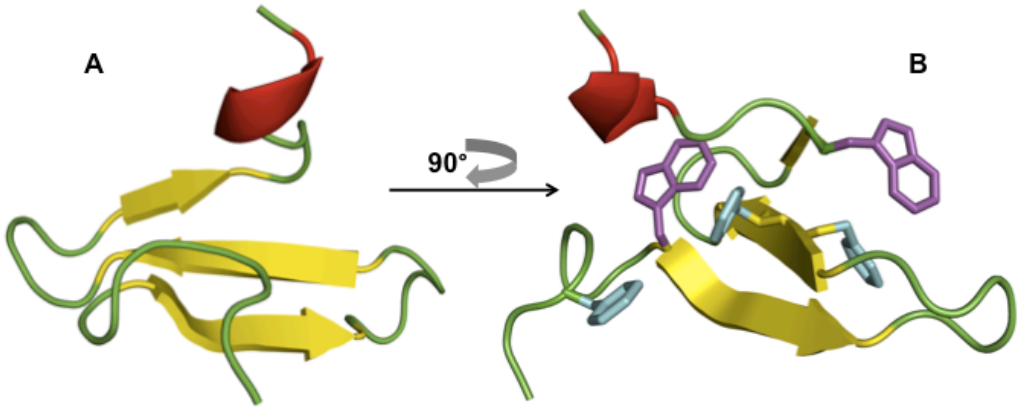


Figure 5.3.- Structure of WW3 from Nedd4-2. **A)** For simplicity only the β-sheets and the α-helix have been coloured, in yellow and red, respectively. The β-turns (in green) connect the pairs of β-strands. **B)** Residues comprising the hydrophobic core are highlighted, in magenta for Trp residues at positions 8 and 30 and in cyan the Phe residues at positions 3, 19 and 20.

5.2.- Thermodynamic Characterization

The purpose of the thermodynamical characterization is to provide information about the thermal unfolding process of the protein. Different experimental techniques were used, allowing to observe diverse probes depending on the technique since each one reports on different structural aspects, namely (i) fluorescence spectroscopy that reports on the packing of the hydrophobic core, thus the tertiary structure; (ii) far-UV CD and FTIR that provides information on the amount of secondary structure; (iii) DSC that allows to characterize the overall thermodynamics of the process. The curves obtained from the thermal unfolding of the protein are fitted to the two-state model, with *Tm*, ΔH (enthalpy of unfolding at the *Tm*), and slopes and intercepts for both baselines information (for folded and unfolded states) as floating parameters as previously described[46,189]. The *Tm* values obtained from the different techniques are expected to be very similar if the unfolding is a cooperative process, or to vary within a range if the process is non-cooperative, like for downhill proteins. These experiments were performed in collaboration with Dr. Malwina Szczepaniak and the DSC performed by Prof. Beatriz Ibarra, from the University of Granada.

The far-UV CD experiment in native conditions of this WW3 domain of Nedd4-2 shows a spectrum different from the expected from a pure only β-sheet structure (i.e., a positive band at 196 nm and a negative one at 218 nm, as explained in Chapter 2.1), with a maximum change in ellipticity appearing at 229 nm with a positive band and a negative

band around 205 nm (see Figure 5.4A). These values had been previously observed for other WW domains[189,190] arising the positive band around 230 nm from packing of the aromatic residues[191] (two Trp and three Phe residues located as shown in Figure 5.3B) at the hydrophobic core, and the negative peak around 205 nm, consistent with the spectra exhibited by several small β-sheet structures, suggesting that the WW domain is properly folded under these conditions[192]. Thus, the far-UV CD of this protein may reflect the integrity of the hydrophobic core as well as the stability of the hydrogen bonds in the β-hairpins. The far-UV CD experiment was performed in the range from 268 K up to 368 K showing the high stability of this WW domain even at the slightly acidic pH the experiments were performed, pH 4.7, with a *Tm* of 337.9 K and *ΔH* of 103 kJ/mol (Figure 5.4D) obtained from a two-state fitting of the signal followed at 229 nm *vs* temperature.

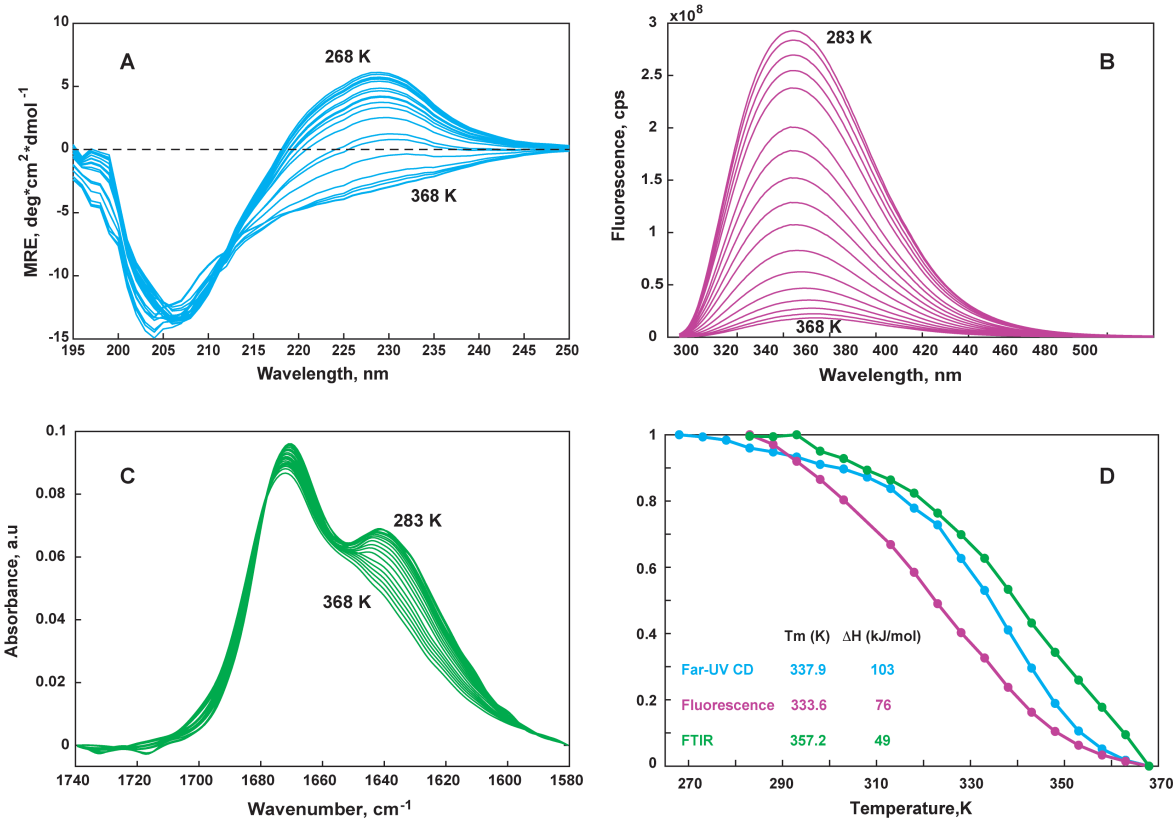


Figure 5.4.- Equilibrium thermal unfolding of WW3 domain from Nedd4-2 by different experimental techniques. The temperature increments were 5 K for all techniques. **A)** Far-UV CD spectrum from 195 to 250 nm at different temperatures starting at 268 K up to 368 K. **B)** Fluorescence spectrum showing the emission wavelength range from 290 to 500 nm after exciting at 280 nm since the main natural fluorophores are the two tryptophan residues present, common to other WW domains. **C)** Steady-state FTIR spectrum at the Amide I region, with the main absorbance variation at 1640 cm^{-1}. **D)** Normalized thermal unfolding curves followed at the maximum of the signal variation as temperature increases: 229 nm for far-UV CD, 344 nm for fluorescence, and 1640 cm^{-1} for FTIR, corresponding to the colours cyan, magenta and green, respectively. For each experiment the corresponding curve has been fitted independently to the two-state model to obtain *Tm* and *ΔH* values (parameters shown in the inset).

Equilibrium fluorescence spectroscopy follows the unfolding through the fluorophores of the protein. The fluorescence for this WW domain is dominated by the two Trp residues displayed in magenta in Figure 5.3B, the main natural fluorophore of WW domains, thus the excitation wavelength was set at 280 nm, and the detection range was extending from 300 nm to 600 nm. The fluorescence plot shown in Figure 5.4B presents a maximum at 344 nm, which is followed in Figure 5.4D producing a very broad transition between folded and unfolded states. The quantum yield of the fluorescence of the Trp decreases as the temperature increases due to the unpacking of the hydrophobic core, but instead of being a cooperative process as expected for a protein with a folding centered on a hydrophobic nucleus, it appears as a broad unfolding. A two-state fitting of the unfolding curve at 344 nm gives a *Tm* of 333.6 K and ΔH of 76 kJ/mol (Figure 5.4D).

The unfolding curves of both far-UV CD and fluorescence produce a different melting temperature after fitting to the two-state model (ΔTm ~4 K), however to obtain further information about the local secondary structure of the protein, a steady-state FTIR in D_2O was performed. This thermal denaturation performed in the range from 283 up to 368 K (shown in Figure 5.4C) focuses on the Amide I region (1600 – 1700 cm^{-1}), the most sensitive spectral region to the secondary structure of the protein due to the C=O stretch vibrations, with characteristic peaks slightly different for each topology. In the case of this WW domain, the main bands appear at ~ 1670 cm^{-1}, corresponding to theoretical calculation of β-sheets[193], and at 1640 cm^{-1}, also typical of β-sheets. Within the range of 1620 and 1640 cm^{-1} also peaks characteristic of unordered structures can be observed, which could indicate that this WW domain have different degrees of conformational stability, with regions presenting conformational exchange[194,195]. The thermal unfolding curve is followed at 1640 cm^{-1}, the peak giving the stronger change in absorbance, which after the fitting to the two-state model provides the highest *Tm* value, 357.2 K, pointing out the high stability of the secondary structure elements, and the lowest ΔH, 49 kJ/mol, indicative of the fact that the transition is quite broad, more than observed using the other spectroscopic techniques.

Preliminary Differential Scanning Calorimetry (DSC) experiments (see DSC basis on Chapter 2.4) were performed on our WW domain. The apparent heat capacity at different protein concentrations (Figure 5.5A) was obtained at different temperatures (298 K, 313 K, 333 K and 373 K) and produced the absolute heat capacity (C_p) with Equation 2.2. This absolute heat capacity is plotted *versus* the temperature, giving the thermogram

of the protein (Figure 5.5B) that represents the native and unfolded states of the protein at low and high temperatures, respectively, and the transition between them. The thermogram exhibits a broad asymmetric profile, with a progressive increase on C_p until reaching the maximum, followed by a faster decrease. The pre-transition baseline, usually parallel or overlapping Freire's native baseline for two-state proteins, is very steep, indicating a progressive loss of structure even at low temperatures, without a clear baseline for the native state of the protein. This sharp pre-transition, represented as a dotted line in the figure for visual purposes, points out the lack of a hard-breaking core in the protein structure. Moreover, the protein is close to Freire native baseline (Equation 2.3) only at the lowest temperature (281 K), revealing the flexibility of the protein structure even at low temperatures. The protein unfolds gradually until it reaches a maximum of the absolute heat capacity around 340 K, then its C_p decreases in a faster mode because the little remaining structure. This continuous unfolding of the protein is characteristic of downhill folding, due to the lack of a stable core, since the cooperative break up of a hard-core would triggers the unfolding otherwise, without a progressive process. The fitting of the DSC profile to a two-state model would produce crossing native and unfolded baselines denoting the inability of this model to fit the data from proteins lacking a static native state (linear extrapolation of the native baseline has been represented in Figure 5.5B as visual reference), indicating the physical impossibility of the application of the two-state model to this domain. This effect has been previously observed for different proteins like a leucine zipper[196], the global downhill folder Naf-BBL[30,197], and proteins folding over a small free energy barrier like PDD[60]. However, aggregation problems due to the high concentration of protein required for these experiments made the results less reliable, even if they provide an initial view of the DSC profile.

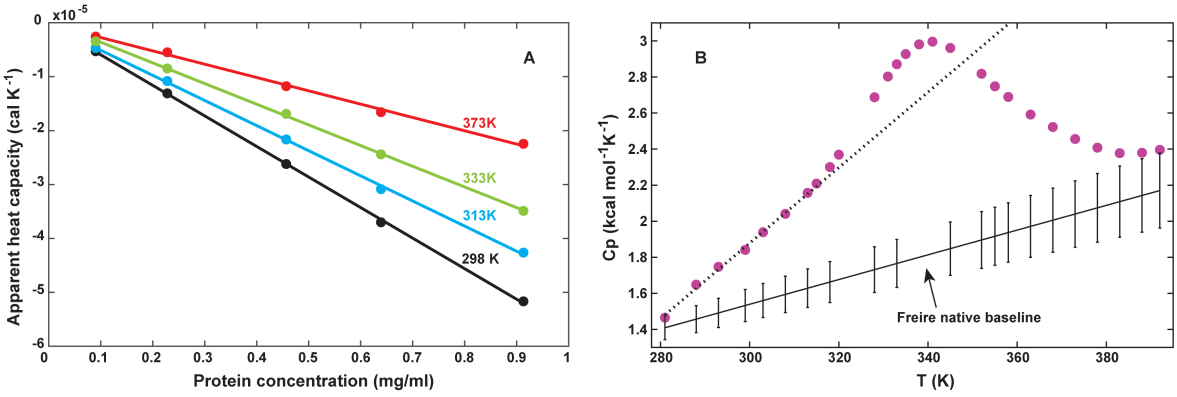


Figure 5.5.- Preliminary calorimetry experiments on WW3 domain from Nedd4-2. **A)** Apparent heat capacity (ΔC_{app}) *versus* protein concentration at different temperatures. **B)** The thermogram profile shows the absolute heat capacity during the thermal denaturation (pink circles). The solid black like represents Freire native baseline (Equation 2.3) with error bars, and the dotted line represents the linear extrapolation of the native state.

5.3.- NMR Studies

An atom-by-atom analysis was performed in this Nedd4-2 WW3 domain with the idea of going forward with a full analysis of this β-sheet protein to compare the results with previous studies on downhill proteins with different topologies. However, aggregation issues together with low yield of the protein purification paused the project after the assignment of most of the chemical shifts at 293 K shown in Table 5.1. This assignment was performed in collaboration with Dr. Mourad Sadqi. Interesting NMR studies have been published regarding other WW domains interacting with ENaC subunits by CPMG-RD analysis, producing very interesting results by the identification of two conformations exchanging on the μs-ms timescale[194]. The conformations in equilibrium are a natively folded peptide binding-competent state and a random coil like denatured state, with the latter populated to ~20% at 37 ºC[198].

Different 2D ([^{1}H-^{13}C]-HSQC and [^{1}H-^{15}N]-HSQC) and 3D NMR (CBCA(CO)NH, HNCACB, CCONH, HCCONH and HAHB(CO)NH) experiments were performed for the assignment of this WW domain. Within the 3D experiments, CBCA(CO)NH, HNCACB and HAHB(CO)NH provide the backbone of the protein, while CCONH and HCCONH not only the backbone, but also the side chain. Chemical shifts from H^{N}, N, H^{α}, C^{α}, H^{β} and C^{β} are shown in Table 5.1.

Residue		HN	N	Hα	Cα	Hβ	Cβ
1	Gln	a	a	4,07	55,09	2,20	31,45
2	Ser	8,80	118,59	4,50	58,02	3,80	63,80
3	Phe	8,23	121,12	4,75	56,93	2,99/3,25	39,48
4	Leu	8,67	122,39	4,28	53,63	1,67 / 1,83	41,77
5	Pro	-	-	a	57,42	a	33,37
6	Pro	-	-	4,50	63,77	2,39/1,98	31,54
7	Gly	9,04	112,41	4,36/3,87	45,57		
8	Trp	7,90	118,74	5,85	56,58	3,38/3,09	31,39
9	Glu	9,41	121,12	4,81	54,87	2,15/2,09	34,85
10	Met	8,99	124,91	4,98	54,23	2,11/1,84	35,98
11	Arg	8,47	126,25	4,41	53,52	1,36/1,06	34,05
12	Ile	8,26	119,18	4,38	59,38	1,86	37,15
13	Ala	8,95	131,38	4,73	51,75	a	18,02
14	Pro	-	-	4,41	65,38	2,45/1,98	31,45
15	Asn	7,59	111,54	4,65	52,35	3,28/2,87	37,01
16	Gly	8,66	108,59	4,33/3,60	45,06		
17	Arg	7,81	122,49	4,86	53,78	2,13/1,93	31,08
18	Pro	-	-	5,17	62,34	2,13/1,73	32,56
19	Phe	8,85	118,12	4,77	55,61	2,56/2,35	40,35
20	Phe	8,84	117,56	5,36	56,93	3,24	41,60
21	Ile	9,22	122,05	4,33	59,97	1,34	41,15
22	Asp	8,29	124,70	3,65	51,19	2,09/0,14	38,74
23	His	8,85	121,64	4,29	57,83	3,57/3,23	27,58
24	Asn	8,45	115,85	4,51	55,78	3,29/2,79	38,45
25	Thr	6,68	104,41	4,34	60,90	4,37	70,47
26	Lys	7,94	119,13	3,80	57,34	2,16/1,89	28,81
27	Thr	7,45	110,79	4,65	60,70	4,04	71,86
28	Thr	8,34	115,44	5,24	60,29	3,81	71,38
29	Thr	9,27	117,97	4,88	59,19	4,38	69,64
30	Trp	8,73	124,94	5,18	57,79	3,81/3,21	30,20
31	Glu	8,79	121,81	4,40	55,73	2,23/1,98	29,64
32	Asp	8,26	126,36	4,73	50,36	3,26/2,97	41,80
33	Pro	-	-	4,04	62,16	1,06	31,10
34	Arg	8,29	118,72	3,98	56,87	1,89/1,67	29,79
35	Leu	7,36	118,98	4,32	54,40	1,65	41,62
36	Lys	7,31	125,51	4,11	57,47	1,81	33,32

(a) Unidentified chemical shifts

Table 5.1.- Backbone chemical shifts of WW3 domain from Nedd4-2 at pH 4.7 and 293 K.

5.4.- Discussion

The thermodynamical studies of WW3 from Nedd4-2 produced thermal unfolding curves (Figure 5.4D) showing differences not only in the cooperativity, but also on the *Tm* and ΔH values originated from the fitting of the curves to the two-state model as observed previously for proteins in the downhill regime with different free energy barriers: non-existing for the protein BBL[24], and marginal for PDD[20] (a structural and functional homologue of BBL), gpW[6] and λ-repressor[26]. All curves show an overall sigmoidal shape, with broad transitions, and steep pre- and post- transition baselines indicating the possibility of molecules undergoing structural changes, this temperature dependence may indicate a folding mechanism close to the downhill scenario[199]. In the case of the FTIR experiment a clear unfolded baseline is lacking as result of the high stability of the secondary structure of the domain and a probable steep post-transition baseline. The *Tm*

values expand from the 333.6 K of fluorescence to 357.2 K of FTIR, a maximum difference of about 24 K. This variation in the same parameter reflects the heterogeneity on the thermal unfolding of the protein depending on the probe followed by each technique. Since the fluorescence reports on the tertiary structure, the fact that its thermal unfolding shows a lower value of *Tm* with respect to the other techniques indicates the breakdown of the hydrogen contacts, that hold together the tertiary structure as a trigger of the unfolding mechanism of the protein. The far-UV CD reports on the hydrophobic core and the secondary structure; since in this case it is followed at 229 nm, an area that is influenced as well by the aromatic Trp residues, the weight of the hydrophobic core on the overall signal is prevalent, resulting in a thermal unfolding curve that is intermediate between the one observed by fluorescence and the one observed by FTIR. The higher *Tm* value from FTIR spectroscopy arises from the Amide–I band, indicating the increased stability of the hydrogen bonds in the structure of the antiparallel β-sheet (secondary structure). The differences on the broadness of the unfolding curves, reflected on ΔH, appear not only in one-state folders, but also in other proteins with small energy barriers[20], and in this case, expands from 103 kJ/mol from the far-UV CD to 49 kJ/mol from FTIR. The higher ΔH value from far-UV CD arise from the unpacking of the hydrophobic core followed at 230 nm, process that is typically more cooperative than the loss of local secondary structures within the protein, however, it does not behave in a cooperative fashion as could be expected for two-state proteins. The lower ΔH value from the equilibrium FTIR is produced by the slow progressive loss of secondary structure, consistent with its high *Tm*, pointing out the high stability of the β-sheet structure by the gradual breakage of the hydrogen bonds between the antiparallel sheets. The progressive unfolding is also confirmed by the preliminary DSC data, with a broad unfolding curve lacking of a flat pre-transition baseline that would be expected for a protein with a cooperative unfolding. Moreover, this pre-transition indicates a continuous and uniform loss of the protein structure as the temperature increases, typical of a downhill folding.

5.5.- Conclusions

The analysis of the different thermodynamical experiments performed in this Nedd4-WW3 domain shows a gradual unfolding of the structure and provides the typical dispersion on *Tm* and ΔH characteristic of the downhill folding, hallmark of structural flexibility in the native state. Consequently, this domain could be included towards a

folding downhill-like scenario, even though further studies could provide more detailed information.

Similar behaviour has been observed recently in the comparison of different WW domains, with low sequence identity, providing useful insights on the folding cooperativity and the comparison, as a group, *versus* the α-helical proteins classically studied[45]. It has been proposed that this conformational heterogeneity may be involved in the recognition of peptide ligands in coupled folding/binding equilibria[45]. Therefore, WW domains are interesting models of β-sheet proteins to study in detail the folding process and their tendency towards the two-state or the downhill folding scenario.

Chapter 6.- R3H Domain

6.1.- General Features

Different nucleic-acid-binding proteins have been studied, like the leucine-zipper, zinc-finger, etc.[200,201]. Here we will focus on a R3H domain, a motif involved in polynucleotide-binding, including DNA, RNA and single-stranded nucleic acids[202], usually found in proteins together with nucleic acid-binding domains. R3H domains have a characteristic motif, consisting of two highly conserved residues, an arginine and a histidine, spaced by three residues (...-Arg-X-X-X-His-...), and a pretty stable pattern of hydrophobic residues, prolines and glycines[203].

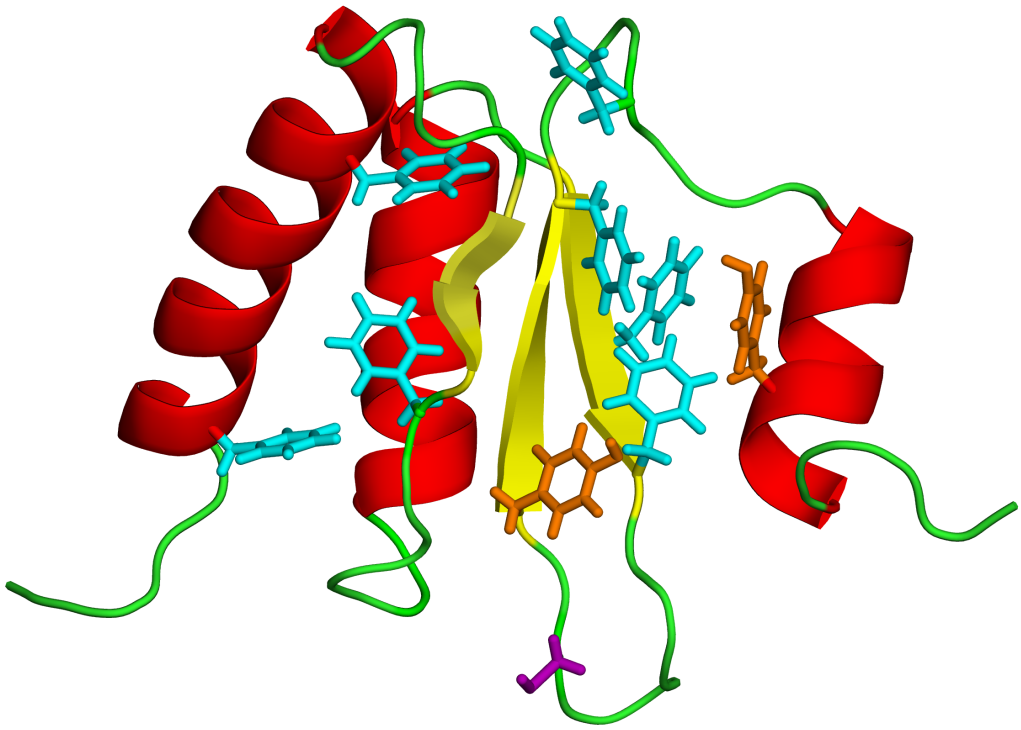


Figure 6.1.- Three-dimensional structure of the R3H domain of the human sperm-associated antigen 7 (PDB 2CPM), consisting of three antiparallel β-sheets (coloured in yellow) and three α-helices (coloured in red). The 7 Phe residues are highlighted in cyan, both Tyr residues in orange and the Cys in purple.

The R3H domain in this study belongs to the human sperm-associated antigen 7, which PDB structure is 2CPM. Its fold consists of a three-stranded β-sheet and three α-helices as shown in Figure 6.1.The amino acid sequence (Figure 6.2) of this 88-residue domain presents the conserved Arg and the His residues,

typical of these domains, at positions 38 and 41, respectively, at the beginning of the second α-helix. The hydrophobic residues are located in the interior of the domain, protected from solvent exposure, except for V6, V47 and I36, placed in the outside of the domain facing the solvent. However, very little is known about this domain despite its interesting binding to nucleic acids.

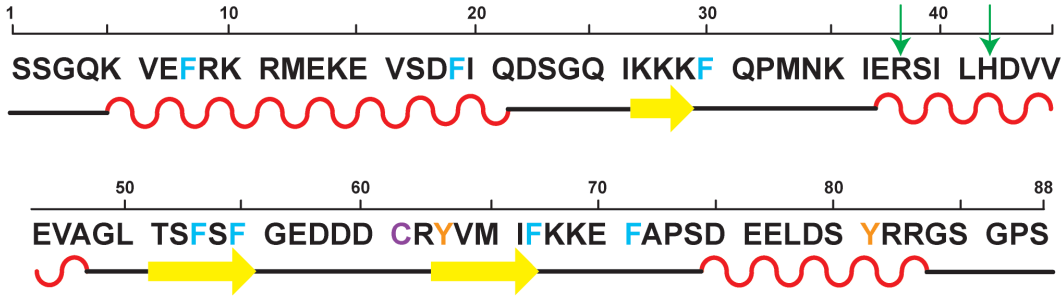


Figure 6.2.- Amino acid sequence of the 88-residue R3H domain with the secondary structure displayed underneath. The colour coding is the same followed in Figure 6.1, and the green arrows point the conserved Arg and His separated by three residues in all R3H domains.

The secondary structure of this domain, including α-helices and β-sheets, together with its bigger size, could be an interesting candidate for folding studies and to compare with the previously studied small all-beta WW domain. At the same time, the comparison with a smaller α/β protein, gpW, previously studied in our group[6,8,60], with a downhill folding, could point out some differences due to the size, amino acidic content and secondary structure ratios (R3H has a 43% α-helix and 13% β-sheet while gpW, 54% and 16%, respectively) as these factors affect the folding and unfolding rates[204].

As presented here, this R3H domain has been characterized thermodynamically. However, preliminary NMR experiments on a ^{15}N labelled sample at different temperatures revealed the loss of NMR signal at temperatures around the *Tm* of the protein.

6.2.- Thermodynamic Characterization

The different techniques applied to the unfolding study of the R3H domain provide a general view of the process. These studies were performed in collaboration with Dr. Malwina Szczepaniak. In this case study, the different

topology and amino acidic content of the domain produce very different profiles compared to the other studied proteins. The experiments were performed in phosphate buffer at pH 7.0, and 1 mM TCEP (tris(2-carboxyethyl)phosphine hydrochloride) was added to reduce the Cys residue present in the domain, avoiding disulfide bridges.

In the far-UV CD experiment of the R3H domain (Figure 6.3), the spectrum at low temperatures is dominated by the α-helices of the structure with their typical double minima around 208 nm and a bit more intense at 222 nm (see Figure 2.1), indicating the presence of more than one α-helix in the structure (a more intense peak at 208 would indicate the presence of a single α-helix), and a maximum around 196 nm. This native helical structure decreases as the thermal unfolding proceeds from 268 K to 368 K (in 5 K steps) towards the random coil conformation observed at high temperatures, beginning with a progressive fray of the α-helices and its typical steep pre-transition, observable in the temperature dependence curve at 222 nm (Figure 6.3 right). The unfolding curve presents a cooperative process with a pretty fast transition from folded to unfolded state, which fitting to the two-state model gives a *Tm* of 335 K and *ΔH* of 238 kJ/mol. Another far-UV CD experiment was performed at pH 2 to obtain the stability of the protein at low pH, however, even at low temperatures the protein lacks any secondary structure peaks, with a feature-less profile, consistent with that recorded at high temperatures and pH 7, towards the random coil.

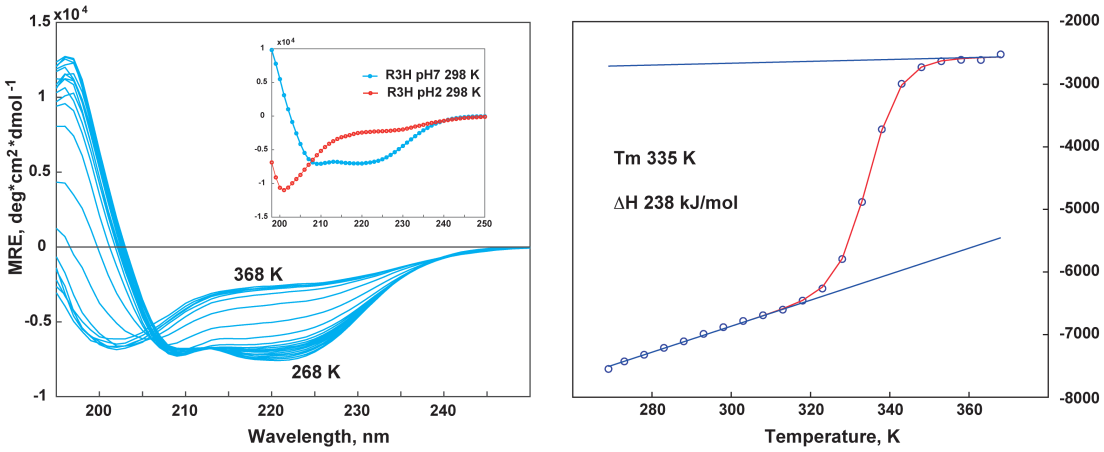


Figure 6.3.- Thermal unfolding by far-UV CD of R3H domain at pH 7. On the left figure the native state at 268 K progressively unfolds by increasing the temperature every 5 K up to 368 K, where the protein has no secondary structure. The inset shows the protein ellipticity at 298 K for pH 7 and pH 2, unfolded in the last one, between 197 – 250 nm. On the right, sigmoidal unfolding curve followed at 222 nm indicating a cooperative process, its fitting to the two-state model provides a *Tm* of 335 K and a *ΔH* of 238 kJ/mol.

The fluorescence spectrum (Figure 6.4) is dominated by the aromatic residues, in the case of the R3H domain, 2 Tyr and 7 Phe residues, and as mentioned in Chapter 2.3, the concentration for the experiment is higher (~ 15 μM), due to the lower intensity of these fluorophores compared to Trp. Both Tyr residues dominate the emission spectrum, with a more intense signal than the Phe residues, with the maximum at 303 nm. The loss in the intensity while increasing the temperature is due to the increase of quenching processes, not to the unfolding of the protein. This quenching in the fluorescence along the different temperatures, even at the beginning of the ramp, could indicate the exposition of the fluorophores to the solvent also at the native state despite the general layout of the hydrophobic residues towards a core. This progressive decrease of the fluorescence intensity is almost linear, making impossible a two-state fitting that would provide interesting information about the *Tm* and ΔH values to compare with other thermodynamic experiments.

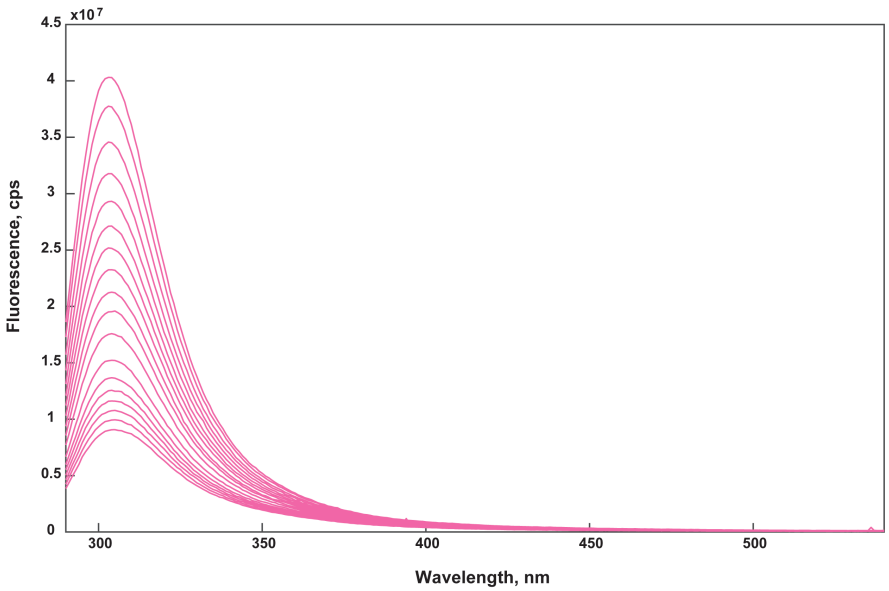


Figure 6.4.- Thermal unfolding by fluorescence of R3H domain at pH 7. The emission signal at 303 nm corresponds to both Tyr residues of the domain.

The monitoring of the secondary structure of the R3H domain by FTIR displays in the Amide I region a complex temperature dependence of the α-helical spectral region. For proteins with a high helical content, there is two bands at 1632 and 1652 cm^{-1} corresponding to the helical carbonyl groups exposed and non-exposed, respectively, and also include weighted contributions from hydrogen-bonded and non-hydrogen-bonded amides. In the case of our domain, only the

contribution at 1652 cm^{-1} is appreciated (Figure 6.5), indicating the predominance of solvent-buried groups. Another band in this spectrum, at 1660 cm^{-1}, corresponds to the random coil[18,178,179]. An SVD analysis was performed on the data to dissociate different contributions to the Amide I band as a function of temperature. The two-state fitting of the second V component provides a *Tm* of 330 K, and a *ΔH* of 225 kJ/mol, slightly lower compared to the previous experiment of far-UV CD.

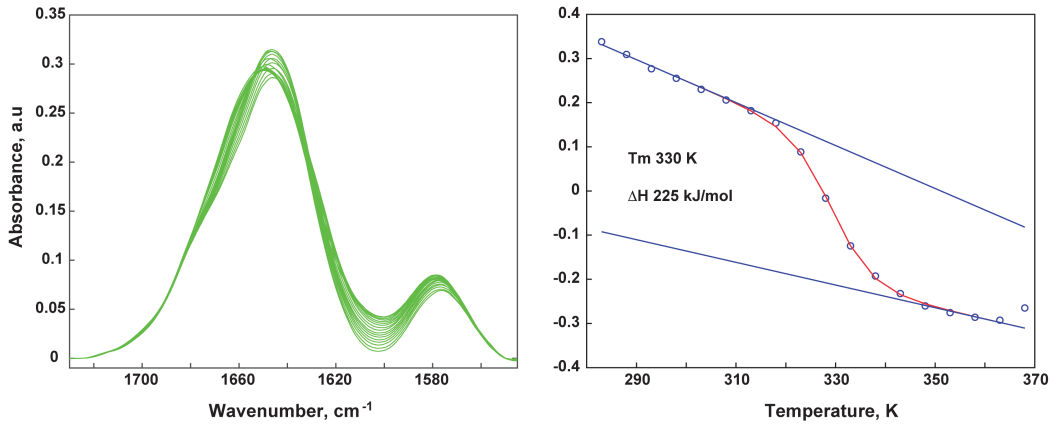


Figure 6.5.- Thermal unfolding by FTIR of R3H domain at pH 7. On the left, unfolding spectrum from 283 to 368 K. On the right, fitting to the two-state model of the V2 component from the SVD analysis.

The DSC thermogram of this domain (Figure 6.6) was performed at neutral pH 7 at different scan rates of 90 K/h (the usual scan rate) and 200 K/h, producing very similar transitions with a slightly higher apparent *Tm* for the last one, due to a deficient equilibration time, too short for an uniform temperature in the protein sample. At this pH the pre-transition baseline is higher than the expected for the native state regarding the empirical correlation derived by Freire and co-workers[205], named here as native and unfolded Freire's predictions. This increased heat capacity for the native state could suggest a "native state" undergoing large-scale enthalpy fluctuations possibly originated by the presence of structural changes as previously observed for PDD[20], a protein with an incipient free energy barrier, however, in our case the pretty flat region may also suggest an increased solvation[20,206], with only a slight fluctuation as visible for the low-rate temperature scan. The transition-less profile at low pH 2 matches Freire's unfolded state, and almost overlaps with the pre-transition at pH 7, therefore, the similar shape dependent on temperature suggests a solvation effect. At high temperature the protein is completely unfolded at both pH values as both post-transition regions overlap with the unfolded state expected by Freire's predictions.

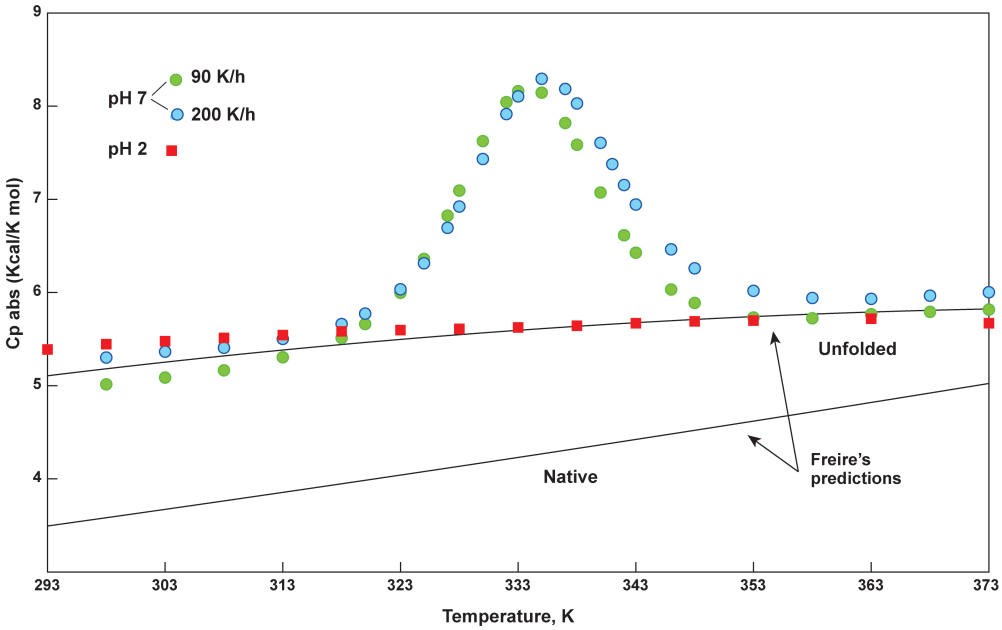


Figure 6.6.- DSC thermogram of R3H domain. Green and blue circles represent the experimental data of the DSC at pH 7 with different scan rates, while red squares display the experimental data at pH 2. The black lines indicate the native and unfolded baselines according to Freire's predictions.

The fitting of the thermogram to the two-state model generates a *Tm* of 335 K, similar to the values obtained from the previously explained thermodynamic experiments, and a Van't Hoff enthalpy (ΔH_{VH}) of 239 kJ/mol (which depends on the shape of the endotherm and on the two-state model used for the fitting), indicating a pretty cooperative process. However, the calorimetric enthalpy (ΔH_{cal}), which is model independent, has a value of 177 kJ/mol. The difference between both enthalpy values indicates some disagreement with the two-state model, despite the protein cooperativity.

Additional unfolding experiments were performed by chemical denaturation followed by far-UV CD to obtain information about conformational protein stability with a denaturant other than only temperature. In our case, for these double-perturbation experiments two different denaturants were selected, urea and guanidinium hydrochloride (GuHCl), with different denaturing intensities, weaker and stronger, respectively. The denaturants report on the importance of the electrostatic interactions for the protein, as the GuHCl salt has a higher electrostatic charge than urea, however, the urea has a non-electrostatic mechanism, and its unfolding process is driven dominantly by its interaction with the polar amide surface of proteins[207]. The denaturant concentration of each stock solution was checked by

refractive index measurements, and similarly, the different protein concentration samples were prepared as separate solutions by the mix of a protein stock solution and a high concentration solution of denaturant with the protein, yielding different denaturant concentrations. Full far-UV CD experiments were performed at each denaturant concentration. The unfolding curves due to the action of the denaturant are followed at 222 nm, where the loss of the α-helix structure allow to monitor the change on the protein conformation, and the comparison of the curves from both agents (Figure 6.7), GuHCl (black curve) and urea (red curve), show different shape and melting concentration, C_m (concentration with half the population of the protein folded and the other half unfolded). The unfolding happens at low concentration of both denaturants, 1.3 M of GuHCl and 3.7 M of urea, indicating the low stability of the secondary structure. The process is cooperative, steeper for GuHCl, which even lacks the pre-transition baseline. The broader unfolding curve produced by the urea was straightforward explained with the Muñoz-Eaton (ME) model[208], an statistical mechanical model previously applied for the global downhill folder BBL[24], and which variant has also been applied to enlighten the folding mechanism of other proteins like gpW[209] and villin[210]. For PDD, structural homologue of the global downhill folder BBL, that folds in microseconds crossing an incipient free energy barrier, the broader shape arises from a urea-temperature coupling that decreases the thermal unfolding cooperativity and progressively lowers the height of the free energy barrier[20], resulting in a less cooperative behaviour. The global fit to the two-state model produces a free energy change for the folding/unfolding reaction (ΔG) of 14.8 kJ/mol.

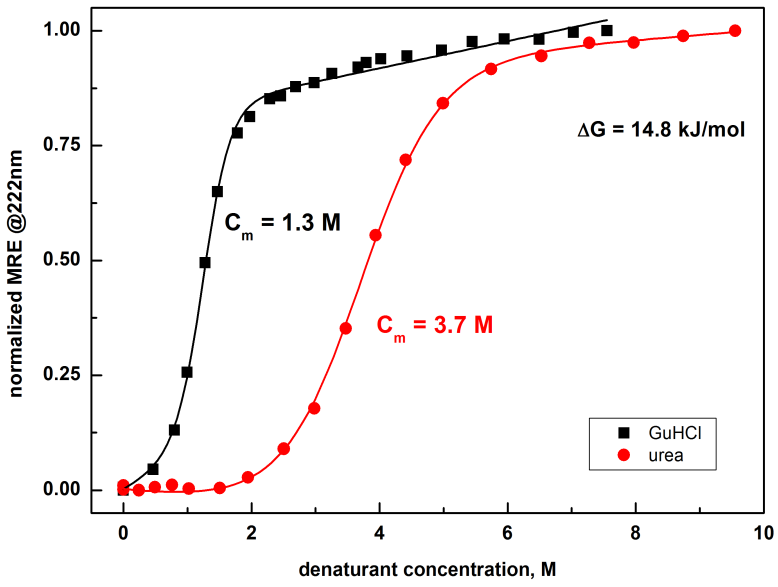


Figure 6.7.- Double perturbation experiment of the R3H domain with the denaturing agents GuHCl, black squares, and urea, red circles. Global fit to the two-state model follows the same colour coding. The secondary structure is monitored at 222 nm at different denaturant concentrations. Both curves are normalized for comparison.

6.3.- NMR studies

Up to date, thorough studies on protein unfolding at atomic level are not an usual practice, even the level of detail obtained can bring to light many interesting features possibly involved on the protein function. Our idea was to provide that level of detail on the R3H domain under study to observe how much detail is possible to obtain from a protein with a highly cooperative folding since the intermediate species may be more unstable and difficult to obtain. Prior to our studies, the NMR structure of a different R3H domain at 298 K was solved[202], and in 2005 *Nagata & co*. deposited the structure of the R3H domain under study in the Protein Data Bank. However, a detailed study of the unfolding of an R3H domain had never been performed, and very few examples of two-state folders or closely behaviour have been followed, like the α-spectrin SH3[9]. The only α/β protein with the thermal unfolding followed at atomic level was gpW, but its size is smaller and the unfolding, towards a downhill behaviour.

Unfortunately, initial NMR experiments at pH 7 showed the loss of the signals around the *Tm* by the coalescence effect (see Chapter 3.4) as Figure 6.8 shows, due to the k_{ex} in the intermediate regime, precluding the chance of further studies.

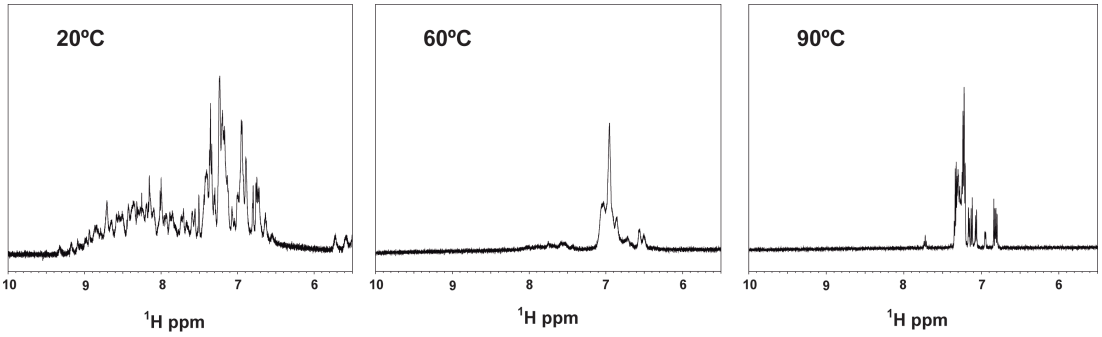


Figure 6.8.- 1D NMR spectra of the amide region of the R3H domain at different temperatures.

6.4.- Discussion

The different experiments performed on the R3H domain show behaviour towards the two-state folding. The hydrophobic residues of the domain are facing the interior of the protein, except for the Tyr residues, that may have a more exposed site as proved by fluorescence, as well as the three residues Val 6, Val 47 and Ile 36, as mentioned at the beginning of this Chapter, maybe due to its involvement in the function of the domain. From the far-UV CD and DSC experiments the *Tm* of the protein is around 335 K, possibly due to the disruption of the hydrophobic core, which would explain the high ΔH value indicating a cooperative process since once the hydrophobic core loses its packing, the protein lacks its stability element and the unfolding proceeds with less obstacles. The slightly lower *Tm* and ΔH values from the FTIR points out the melting of the secondary structure elements a bit earlier than the break up of the hydrophobic core. This R3H domain also has a quick loss of structure upon denaturants and low pH values.

6.5.- Conclusions

In general, this domain can be excluded as a downhill folder, even slight loss of structure occurs in the pre-transition, mainly by the fraying of the α-helices. This kind of gradual native disorder has been previously observed for two-state folders, like SH3 domains[9,211].

Chapter 7.- HP35 Subdomain of Villin Protein

7.1.- General Features

Villin protein is a ~95 kDa protein comprising 7 domains (see Figure 7.1), member of the gelsolin superfamily[212] with actin regulatory functions[213]. Its crosslinking with actin filaments allows the movement of the microvilli at the terminal region of epithelial cells where it is located[214]. Out of the seven domains, the 76-residue "headpiece" domain is located at the C-terminus, where we will study the last 35 residues of this domain, the HP35 subdomain.

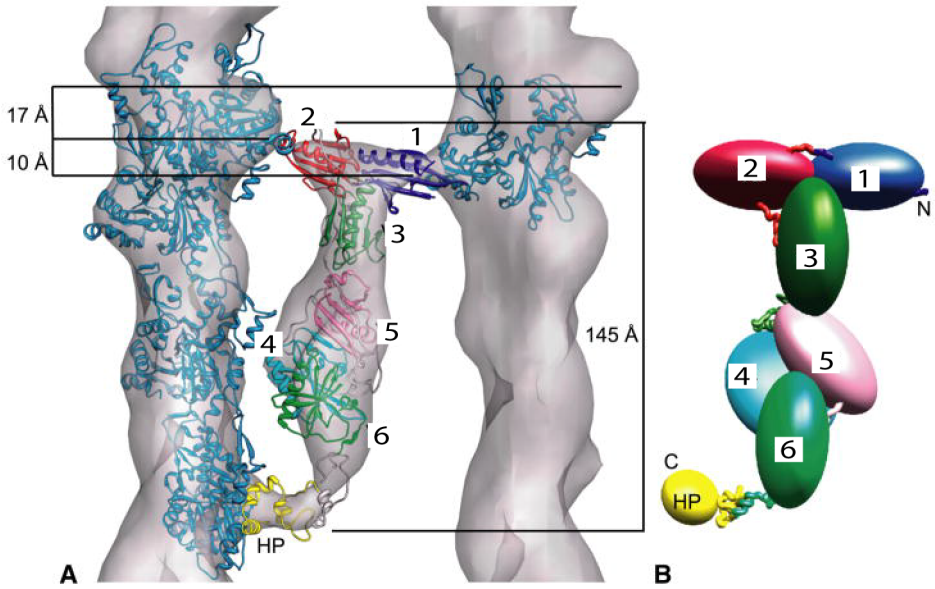


Figure 7.1.- Model of villin protein acting as crosslink between the actin fibres (**A**) and topology of the villin domain (**B**). The "headpiece" (HP) domain is coloured in yellow. Figure adapted from Hampton *et al.*[214]

This villin HP35 subdomain, despite its small size, retains high thermostability as shown by McKnight *et al.* (*Tm* ~343 K at pH 7)[215]. The structure made of a three α-helix bundle constitutes an autonomously folding monomer, which folding is not based on the hydrophobic core of Phe residues as could be expected, or on disulfide bonds (it lacks Cys residues). The stability of HP35 is distributed along the sequence, and only the lack of Phe 58 causes the major destabilization[216,217]. Thermodynamic analysis by differential scanning calorimetry

indicates that the villin headpiece subdomain crosses a relatively small free energy barrier to fold into its native state[218]. This barrier is, however, close to the two-state folding limit and much higher than it would be expected for its fast folding kinetics (~5 μs[70]). The barrier also presents independence to temperature[218]. The small size and relatively simple structure of this subdomain has turned it into a favourite folding model, resulting on an enormous variety of studies, like molecular dynamic (MD) simulations[42,219-224] or experimental, such as mutations to study the interactions that stabilize the structure[210,220,221,225-230] or to create an ultrafast folding protein[47,226,231].

Thus, HP35 represents an interesting case study in this work due to its fast folding, with rate similar to that of the one-state downhill folder BBL, i.e. ~4-5 μs at 320 K[46]. However, BBL lacks an energy barrier altogether[24] whereas DSC experiments have determined a significant free energy barrier for HP35[70]. Moreover, the thermodynamic barrier estimation from DSC measurements (reports the population at the barrier top in equilibrium conditions) usually underestimates the kinetic barrier experienced by the folding protein (determined by the kinetic flux through all the microscopic barrier crossing paths). Therefore, it is likely that the kinetic folding barrier of HP35 is even higher, thereby reaching the two-state folding scenario. The implication is that its thermodynamic unfolding behaviour looked at the atomic level (using chemical shift perturbations) could more closely resemble a typical two-state, or all-or-none, transition instead of the gradual heterogeneous transitions of fast folders that do not cross a significant barrier, such as the global downhill folder BBL. In this way, HP35 could be a perfect counterexample of near two-state folding on small proteins with simple α-helical topology and very fast folding kinetics.

This study is a collaboration with Dr. Robert Tycko (National Institutes of Health, NIH, Bethesda, USA), who provided synthetic samples of HP35 isotopically labelled in specific residue positions. Chemical synthesis was convenient as it avoids the problems of producing small unstable proteins by recombinant means. However, uniform isotopic labelling of the full-length protein was not realistic due to the extremely high cost of some of the Fmoc-derivatives of isotopically labelled amino acids. Moreover, signal overlap could be significant in this case due to the relative homogeneity of HP35 sequence. The plan in this case was to chemically synthesize by standard solid-phase methods a series of HP35 variants identical in sequence, but each one with a different small subset of residues isotopically labelled. The labelling would permit to explore the folding-unfolding thermodynamics

at the level of individual atoms, looking at both the backbone and the side-chains, and use carbon and nitrogen chemical shifts rather than protons, which are less reliable probes of protein conformation. In HP35 the interactions stabilizing the subdomain arise from aliphatic side chains, thus, labelling them will allow to obtain information on these connections. Therefore, five samples were synthesized, called V1, V2, V3, V4, and V5, identical in sequence and with various subsets of isotopically labelled residues (from a minimum of 4 to a maximum of 6 residues). Another important factor in the design was to include one particular labelled residue in common among samples to operate as an internal reference given that, in contrast to previous atom-by-atom analysis of folding, here the data will be obtained using different samples measured in different days. The labelling pattern is shown in Table 7.1, and it is spread along the protein sequence to cover all the structure (see Figure 7.2) as already mentioned. The amino acid sequence starts in the residue Leu 42 and ends in Phe 76, preserving the numbering corresponding to the complete headpiece domain sequence.

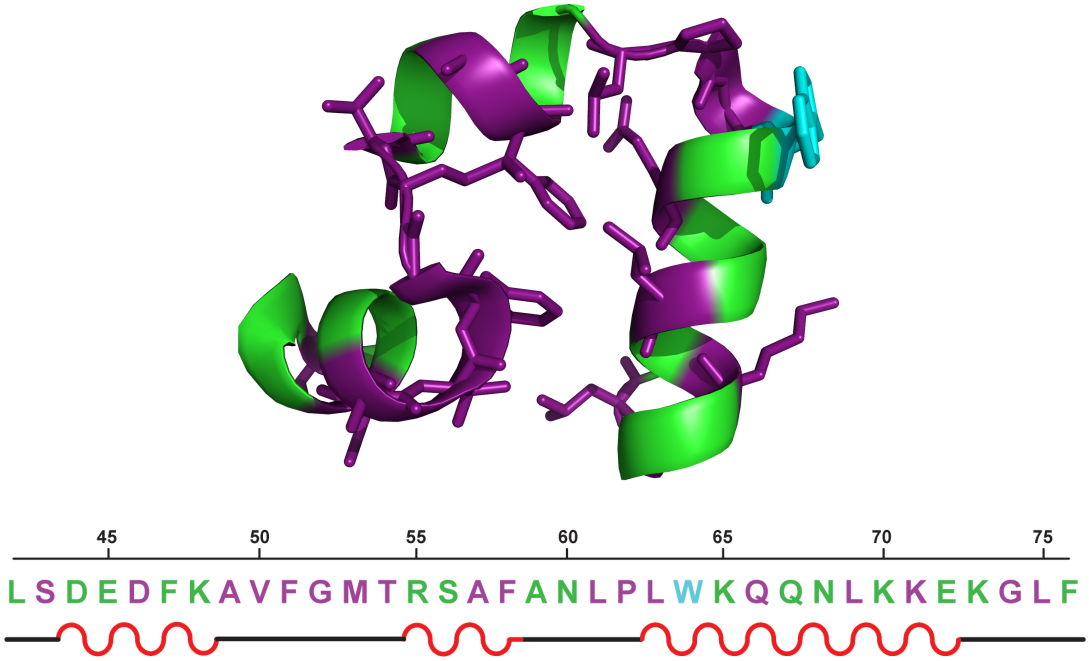


Figure 7.2.- HP35 subdomain of villin protein with the isotopically labelled residues from the different variants (adding up V1, V2, V3, V4 and V5) in purple, and the tryptophan residue (not labelled) in cyan. The secondary structure is represented under the amino acid protein sequence with the α-helices in red. The structure corresponds to the PDB 3MYA.

	V1	V2	V3	V4	V5
S43					X
D46					X
A49			X		
V50	X	X		X	X
F51		X			
G52	X	X	X		
M53			X		
T54	X				
A57	X	X		X	
F58	X		X		
L61		X			
P62			X		
L63	X				
Q66					X
L69			X	X	
K71					X
G74				X	
L75					X

Table 7.1.- Labelled residues in each HP35 variant. All samples have at least one residue repeated in the other variants as internal reference.

7.2.- Thermodynamic Characterization

HP35 has already been thoroughly studied from a thermodynamical viewpoint in different papers[70,210,215,221,232], so the general characterization for this protein was not repeated. However, far-UV CD experiments were performed at pH 3.0 on the different samples to obtain information on the thermal stability of the protein under conditions of fast conformational exchange and slow proton exchange with water. Also the mass of all villin samples was confirmed by mass spectrometry (MALDI-TOF), with the expected molecular weight for each sample with an average ±4,9 Da error (the different molecular weight according to the isotopic labelling was taken into account).

The far-UV CD experiments were performed at ~ 55 μM protein in citrate-phosphate 20 mM buffer at pH 3.0. The low pH was used to slightly destabilize the structure and decrease the *Tm* (343 K at pH 7.0[215]) in order to obtain baselines for both folded and unfolded states avoiding temperatures over 373 K in the NMR instrument. At least two experiments per sample have been performed.

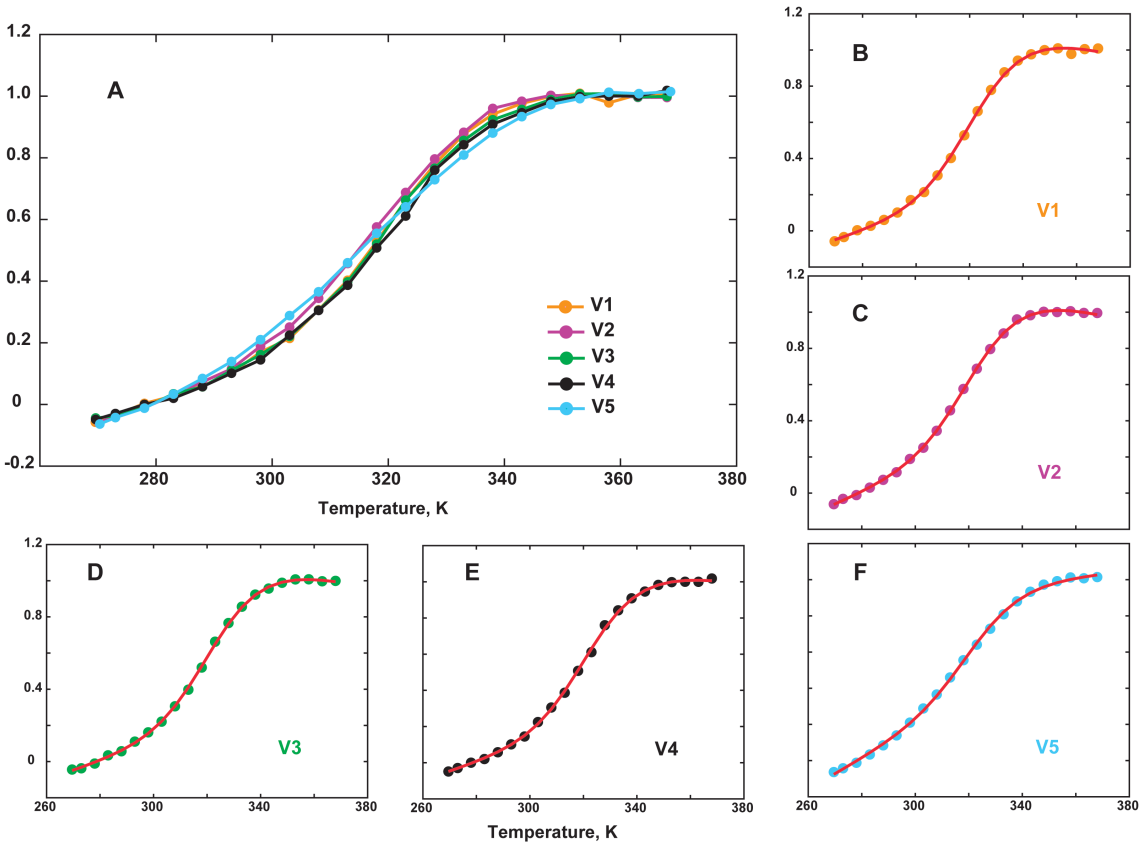


Figure 7.3.- Normalized unfolding curves from far-UV CD at 222 nm of HP35 subdomain at pH 3.0. In all cases, the data for each variant corresponds to the average of three experiments. **A)** Superimposed unfolding curves from each HP35 variant (V1, V2, V3, V4 and V5). **B**, **C**, **D**, **E**, and **F** show the global fit (red lines) of the experimental unfolding data of V1 (orange), V2 (magenta), V3 (green), V4 (black) and V5 (cyan) samples, respectively.

The helical structure of the native state of HP35 generates spectra with a more pronounced minimum at 222 nm than at 208 nm indicative of a large α-helix fraction in the native structure. As we can observe in Figure 7.3A the unfolding curves followed at 222 nm *versus* the temperature from the different HP35 variants show essentially the same overall behaviour, but there are some slight differences between variants. These differences are real and repeatedly observed in systematic experiments performed independently by two researchers (myself and Dr. Matija Popovic, postdoc at the Muñoz lab). The origin for such differences may be their separate chemical synthesis process (for example, a varying amount of salts in the freeze-dried powder), but it remains undetermined. In these experiments is also noticeable that the pre-transition baseline is largely sloped, indicating the loss of the helical CD signal due to fraying of the α-helices, similarly to what we have seen in other proteins like the R3H domain included in this Thesis (Chapter 6). The unfolding process continues smoothly towards the transition region, showing a continuous loss of structure with a steeper slope. The flat post-transition baseline at

high temperatures suggests a completed unfolded transition in this temperature range, which could be due to the stabilization of a partially structured denatured state ensemble by clustering of the hydrophobic residues, as it has been previously proposed[233-235]. Therefore, despite the significant free energy barrier estimated by DSC for the protein, there are signs in its CD melting curve of certain degree of gradual unfolding, agreeing with previous observations that not even the hydrophobic core (formed by Phe 47, Phe 51 and Phe 58 residues) is required for the specific fold of the subdomain[216,217]. The fitting of the curves at 222 nm to a two-state unfolding model outputs the characteristic *Tm* and (*ΔH)* parameters, which have average values for each variant of 322 K (121 kJ/mol), 322 K (93 kJ/mol), 322 K (104 kJ/mol), 323 K (104 kJ/mol), and 320 K (101 kJ/mol), for V1, V2, V3, V4 and V5, respectively.

7.3.- NMR Studies

The five HP35 subdomain variants have different labelling patterns as previously mentioned (see Table 7.1), with only a few residues labelled in each one. In the labelled residues all carbon and nitrogen atoms are isotopically labelled and thus we performed a full chemical shift assignment of not only the backbone atoms, but also those of the side chain for each labelled residue. The carbon chemical shifts of the side-chains of a protein have never been investigated using the NMR atom-by-atom analysis. Previous studies have in fact either restricted the analysis to proton signals[7], which are very sensitive to the local electronic environment and not only conformation, or to isotopically labelled backbone atoms without side chain information (^{15}N, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$)[8].

The samples were prepared without solubility issues as explained in Chapter 3.6.1. HP35 variants did not give the type problems arising from severe line broadening near the denaturation midpoint, as we observed in the R3H domain, presumably because the folding kinetics of this protein is too fast to result in significant broadening. Therefore, HP35, as BBL, is a perfect model system for the atom-by-atom analysis of folding interaction networks. Initial 3D experiments for full assignment were performed for each sample: NHCACB for backbone, and HCCH-COSY for side chain assignments. The unfolding process was monitored by [^{1}H-^{15}N] and [^{1}H-^{13}C]-HSQC experiments from 273 to 358 K, with HCCH-TOCSY in case of ambiguity. The chemical shifts at 293 K are indicated in Table 7.2.

Residue		HN	N	Hα	Cα	Hβ	Cβ	Side Chain
43	Ser	9,23	121,45	4,61	57,48	4,07/4,31	64,53	
46	Asp	8,10	120,94	4,60	55,26	2,85	38,77	
49	Ala	7,72	122,04	4,11	54,48	1,49	18,21	
50	Val	7,88	117,46	3,58	65,15	1,51	32,46	Hγ 0,80/-0,03; Cγ 21,95/20,40
51	Phe	8,37	113,85	4,29	59,73	2,47/2,99	39,93	
52	Gly	8,18	108,38	3,95/4,07	46,05			
53	Met	7,64	114,27	*	*	2,14/2,36	35,12	Hγ2,37/2,8; Cγ 29,76
54	Thr	8,25	107,78	4,50	60,88	*	*	Hγ 1,38; Cγ 22,02
57	Ala	7,68	125,21	4,11	54,72	1,48	18,59	
58	Phe	8,41	121,14	4,13	60,90	3,01/3,16	39,94	
61	Leu	7,42	121,88	4,34	53,04	0,85/1,61	41,10	Hγ 1,85; Cγ 25,68
								Hδ 0,50/0,73; Cδ 26,07/22,19
62	Pro			4,37	62,37	1,11/2,37	32,19	Hγ 1,64/2,08; Cγ 27,80
								Hδ 3,12/3,81; Cδ 49,79
63	Leu	8,87	125,54	3,79	58,66	1,73	41,02	Hγ 1,62; Cγ 26,95
								Hδ 0,92/0,95; Cδ 24,27/23,74
66	Gln	7,61	118,91	3,46	59,22	1,94/2,06	28,70	Hγ 0,99; Cγ 33,46
69	Leu	8,34	120,88	4,24	58,11	1,69/2,14	42,31	Hγ 1,87; Cγ 27,57
								Hδ 0,96/1,06; Cδ 25,96/23,38
71	Lys	8,07	118,95	4,13	59,05	2,01	32,30	Hγ 1,47/1,57; Cγ 25,03
								Hδ 1,72; Cδ 29,22
								Hε 2,96; Cε 42,01
74	Gly	7,88	108,02	46,02	3,99/3,86			
75	Leu	7,88	120,43	4,44	54,25	1,48/1,61	42,69	Hγ 1,51; Cγ 26,81
								Hδ 0,87/0,81; Cδ 24,87/25,5

* Non Determined Peak

Table 7.2.- Chemical shifts of the labelled residues the HP35 sample at pH 3.0 and 293 K. In the cases of the residues are repeated in more than one sample, the values have been averaged.

The global fitting to the two-state unfolding model for all of the individual chemical shifts *versus* temperature unfolding curves for each variant produced the *Tm* and ΔH as shown in Table 7.3. The variations between each sample are larger than from the CD experiment. This is probably because, in contrast to the CD experiments, the parameters derived from the NMR data for each variant report on averages from different residues, thus adding to the chemical heterogeneity that we observed by CD. This result already hints that HP35 may experience a certain degree of atomic heterogeneity in unfolding even though it folds over a significant free energy barrier. Table 7.3. also includes the error estimation (95% confidence) based on the global fit for each variant, thus confirming that the reported differences are statistically significant.

Sample	***Tm* (K)**	**ΔH (kJ/mol)**
Villin 1 (V1)	322,7 ± 0,5	102 ± 4
Villin 2 (V2)	320,7 ± 0,7	107 ± 7
Villin 3 (V3)	320,0 ± 0,3	98 ± 3
Villin 4 (V4)	326,4 ± 0,7	106 ± 5
Villin 5 (V5)	321,5 ± 0,6	107 ± 6

Table 7.3.- *Tm* and ΔH from the global fit to the two-state model of all unfolding curves for each villin variant.

However, inspection of the *Tm* and ΔH obtained from independent fits to each of the individual atoms of the HP35 variants (shown in Appendix A with the error estimations, 95% confidence) indicate that the differences between unfolding patterns of various atoms is significant but rather limited in range. This result contrast with the vast heterogeneity in unfolding behaviours observed previously on downhill folding proteins such as gpW protein[8] and BBL[7]. Moreover, the heterogeneity is also lower in qualitative terms as all the chemical shift curves we analysed (a total of 229 different probes for the 5 variants) exhibited typical two-state-like sigmoidal curves or featureless linear transitions (i.e. when the chemical shift was not affected by the folding status of the protein). No signs of three-state like (double transitions) or even more complex curves were found here, again in contrast to previous behaviour in gpW[8] and BBL[7]. This limited variability in atomic unfolding behaviours is direct indication of a protein with an unfolding that is more cooperative, and thus two-state-like, than what is expected for a downhill folding protein. This result is very interesting as HP35 is a very small subdomain, folding

into one of the simplest topologies and it does it in just a few microseconds. It is also interesting to note that residues conforming the hydrophobic core of the protein, and known to play an important role in the stabilization of the native 3D structure, such as for example Phe 58, show more consistency in atomic unfolding behaviours, with sigmoidal curves for almost all of its atoms. Phe 58 does indeed form interactions with the other two core aromatic residues in HP35 subdomain (Phe 47 and Phe 51), establishing the protein hydrophobic core together with other 7 residues (Leu 42, Val 50, Met 53, Leu 61, Lys 65, Leu 69 and Lys 70)[225].

7.4.- Discussion

The general behaviour obtained by far-UV CD and NMR experiments shows a HP35 subdomain from villin protein with a gradual loss of structure under thermal unfolding, but with *Tm* and ΔH differences among atomic probes that are much less marked than expected for downhill folding protein. The reason for the little differences may be the energy barrier of this subdomain, that limits the complexity of behaviours, and probably creates a single but overall behaviour (due to the stability spread along the amino acid sequence instead of focused on a hydrophobic core[217,225]), without meta-stable conformations along the free energy surface, dragging this protein closer to a two-state behaviour. The differences between the villin samples by far-UV CD and NMR experiments have been obtained also by Dr. Matija Popovic, assuring that they are caused by the samples and not an artefact.

7.5.- Conclusions

Therefore, our data confirms that the equilibrium unfolding of HP35 is not pure two-state-like, but it is rather close, making this protein the fastest known example of two-state folding protein.

Chapter 8.- gpW Protein

This Chapter on gpW protein has been published[236] and is included in the Thesis as Appendix B. Therefore, the references to the paper within the Chapter are omitted.

8.1.- General Features

The gpW protein is located on the head to tail connector of the bacteriophage λ of *Escherichia coli*, and it has been suggested that six molecules may polymerize on the connector for later joining six molecules of other protein (gpFII) to build the linking collar before the tail coupling[237]. This small monomeric 68-residue protein has a hydrophobic region at the C terminus, which decreases solubility, crucial for its biological activity but not for the thermodynamic stability[238], therefore, the work will proceed without this hydrophobic tail, as previous thermodynamic[6] and structural[239] studies. The conservative T2V mutation is also conserved (Figure 8.1).

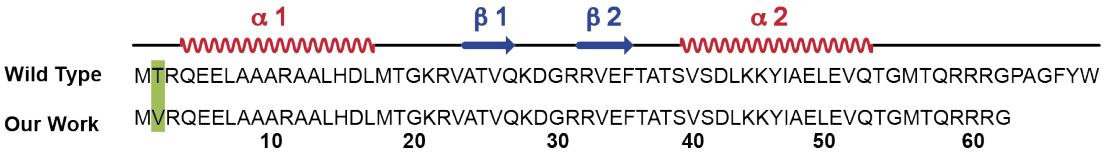


Figure 8.1. – gpW protein sequence of the wild type 68-residue and the 62-residue gpW used in this work, which corresponds to both PDB 2L6R (at pH 3.5) and 2L6Q (at pH 6.5). T2→V mutation is preserved as in previous works[238,239]. The sequence lacks the hydrophobic tail at the C-terminus present in the wild type. On top of the sequences both α-helices and β-sheets are shown in red and blue, respectively.

The 62-residue gpW protein used for the study was produced as a SUMO-fused protein and purified as previously explained in Chapter 4. The tertiary structure of this α/β protein is constituted by two α-helices at C- and N- terminal ends connected by a β-hairpin (with two antiparallel β-sheets) orthogonally disposed to them as displayed in Figure 8.2. The protein lacks Trp residues, but other residues like Val, Leu and Ile create a hydrophobic interface between the helices and the hairpin, as highlighted in the figure.

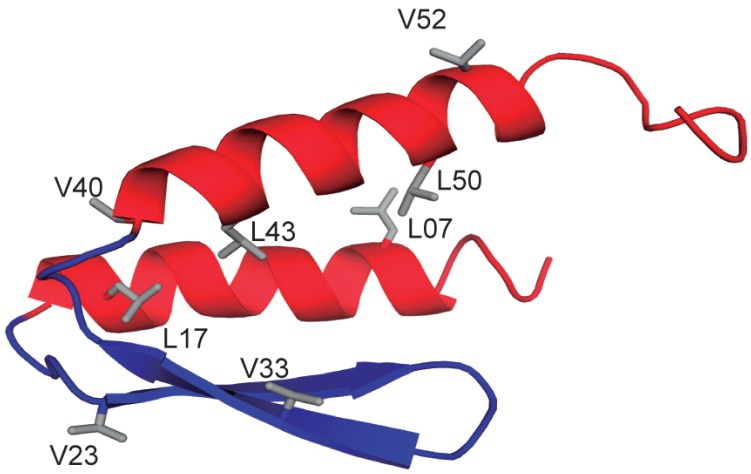


Figure 8.2. – 62-residue gpW protein structure at 293 K and pH 3.5 (PDB 2L6R). The secondary structure elements are represented in red and blue for the α-helices and the β-hairpin, respectively. In gray some interesting Val and Leu residues, some of them involved in contacts between the helices and the sheets.

The gpW study case is particularly interesting from a folding viewpoint because it is a natural single domain protein, that despite its midsize and combined topology, folds in the microseconds timescale, with a small energy barrier and a downhill-like behaviour[6] even it is not a one-state folder and exhibits some cooperativity. However, previously small energy barriers had been associated with small mainly α-helical domains, many of them with artificial origins. All these interesting gpW features led to exhaustive experiments at atomic-resolution on the interaction network by NMR upon unfolding, in parallel with long-timescale MD simulations at different temperatures, to uncover the details of the complex interaction network at every moment of the unfolding process[8].

Therefore, the steady state of gpW and its (un)folding has been thoroughly studied, even the conformation dynamics remained unclear. This missing piece of information, essential for the function of numerous proteins, is guided by the free energy landscape. Hence, an important goal on the thorough knowledge of a protein and its roles (specially those that may behave as molecular rheostat), should lead us towards the unveiling of its free energy landscape. The purpose of the study performed on gpW is to deep into the understanding of the dynamics of fast-folding proteins by pushing towards the limits of Relaxation Dispersion NMR, pursuing the possible “invisible” conformations of the gpW protein that shows a stepwise folding manner[42].

Within protein dynamics, the function may not arise from the more populated protein conformation (obtained from conventional methods), but from a sparsely populated conformation, or from the interconversion between them. Subsequently, methods to discern these sparsely populated conformations, "invisible" states, become a commitment, and Relaxation Dispersion NMR (RD-NMR) has been proved to be a crucial technique that unveils these interesting low-populated protein conformations located at local minima of the free energy landscape[144,147,240].

8.2.- Thermodynamical Characterization

Thermodynamics of gpW have been previously studied by Fung *et al.* 2008[6] at pH 6.0 revealing many of the downhill folder features, like probe dependent thermal denaturation on equilibrium experiments (with a 5 K mismatch in T_m), complex coupling between two denaturing agents, and negligible folding free energy barrier. The different thermodynamic experiments, together with the fast folding kinetics obtained with T-jump (33 μs and 1.7 μs, at 310 K and 355 K, respectively) and a folding barrier ≤3 *RT* (~2.8 kJ/mol) at T_m, characteristic all of them of a downhill folding behaviour, confirmed gpW to fold within the downhill folding regime. Therefore, as control, only the routinely mass spectrometry, and far-UV CD experiments were performed since expression and purification methods have changed regarding previous studies (see Chapter 4).

The far-UV CD experiment was performed as previously explained (Chapter 2.1) at pH 3.5 to slightly destabilize the native structure of the protein to increase the possible additional states. However, despite the weak destabilization, no other structural or mayor stability changes were desired, thus, far-UV CD experiments at pH 3.5 compared to the previous thermodynamical studies at pH 6.0[6] were performed. This lower pH presents very similar cooperativity as pH 6.0 as shown in the Figure 8.3A with only a minor decrease on the ΔH value of 10 kJ/mol (from 150 kJ/mol at pH 6.0, to 140 kJ/mol at pH 3.5), a small variation compared to the important change in pH, suggesting an overall high structural stability. The higher decrease on *Tm* of 11 K (from 341 K to 330 K at pH 6.0 and 3.5, respectively), indicates the desired slight destabilization of the protein. The comparison between the available NMR structures at pH 3.5 and pH 6.5 shows no mayor changes as shown in Figure 8.3B (the 0.5 of pH difference is not considered relevant considering the protein stability).

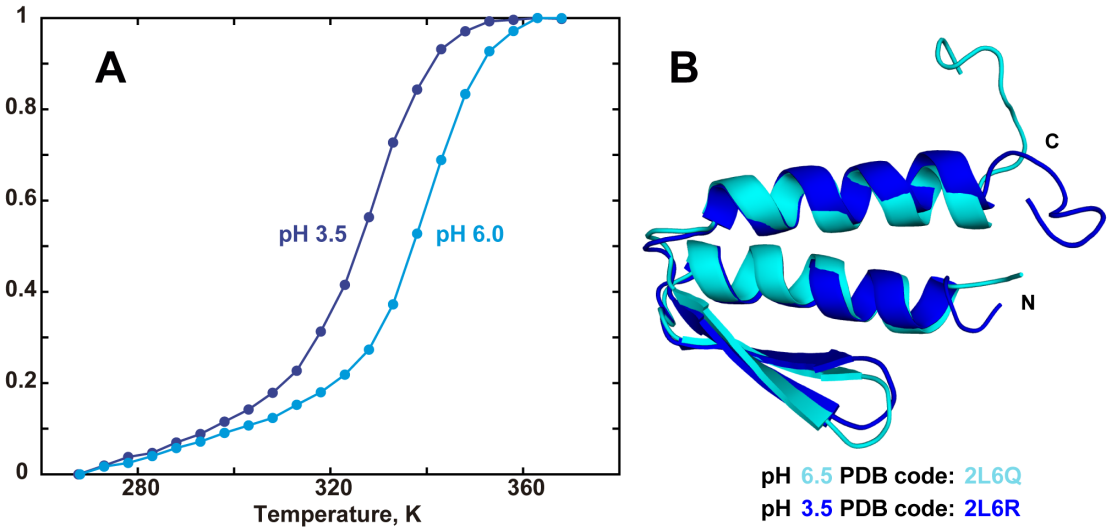


Figure 8.3. – Comparison of gpW at pH 3.5 (dark blue) and a more neutral pH (cyan). **A)** Normalized far-UV CD data at pH 3.5 and 6.0 followed at 222 nm. **B)** Superimposed NMR structures of gpW at pH 3.5 (PDB 2L6R) and pH 6.5 (PDB 2L6Q), using backbone heavy atoms (excluding C^{O}).

8.3.- NMR Studies: Relaxation Dispersion (RD) NMR

The studies at atomic level on the protein gpW with NMR experiments were performed in order to bring to light the possible existence of another conformation or conformations in equilibrium with the detected conformation. These "invisible" conformations even undetectable under regular NMR experiments cause the broadening of the native state resonances due to the exchange with the observable state. When the interconversion rate between both conformations, the observable ("ground") and the invisible ("conformationally excited") states (from now on called *X*), is in the milliseconds timescale, the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence can refocus the chemical shifts and decrease the broadening through pulses applied at different rates (υ_{CPMG}) as previously explained in Chapter 3.5.3.

To reach an increased population of additional states the pH was decreases to 3.5 as previously mentioned, without change on the folding rate of this fast folding protein. To reach the goal to study the protein with RD-NMR and inquire its free energy surface, the folding of gpW in the microsecond timescale (1.7 μs at 355 K[6]) was out of the range for RD-NMR experiments. Therefore, different methods to slow down the folding process were tested. Initially, different concentrations of urea (1 M, 1.5 M and 2 M) and glycerol (29% and 42%) were examined at different temperatures (293 K and 283 K), the sample preparation can be found in Chapter

3.6.2. However, the addition of these substances may create undesired effects and distort the native state of the protein, like the shifting towards the unfolded state for the urea samples (destabilizing the protein) or towards the native state for the glycerol ones (stabilizing the structure). Finally, to reach the k_{ex} within the RD-NMR range (100 s^{-1} ≤ k_{ex} ≤ 2000 s^{-1}) the temperature was decreased to 1 ºC. At this low temperature the [^{1}H-^{15}N]-HSQC spectrum obtained (Figure 8.4) is a very similar to the 25 ºC. Therefore, this low temperature of 1 ºC and pH 3.5 are reliable for the RD-NMR studies on gpW because even though it slows down the protein folding rate and slightly destabilizes the protein, the overall structure remains the same. This low temperature broadens the cross-peaks, like V33 or A37 (see bottom left of Figure 8.4) and in the [^{1}H-^{15}N]-HSQC spectrum three out of the 61 expected peaks are beyond detection. This broadening is caused by the conformational exchange between the ground and the conformationally excited populations of the protein in the micro- or milli-second time scale. Therefore, the magnitude of the broadening can be altered by a CPMG pulse train, as explained previously in detail in Chapter 3.5.3, that refocus the magnetization, decreasing the effective transverse relaxation rate ($R_{2,eff}$) as the refocusing pulses frequency (ν_{CPMG}) increases, originating the typical relaxation dispersion profile (see Figure 3.15) with the decay in the $R_{2,eff}$ for the atoms with conformational exchange.

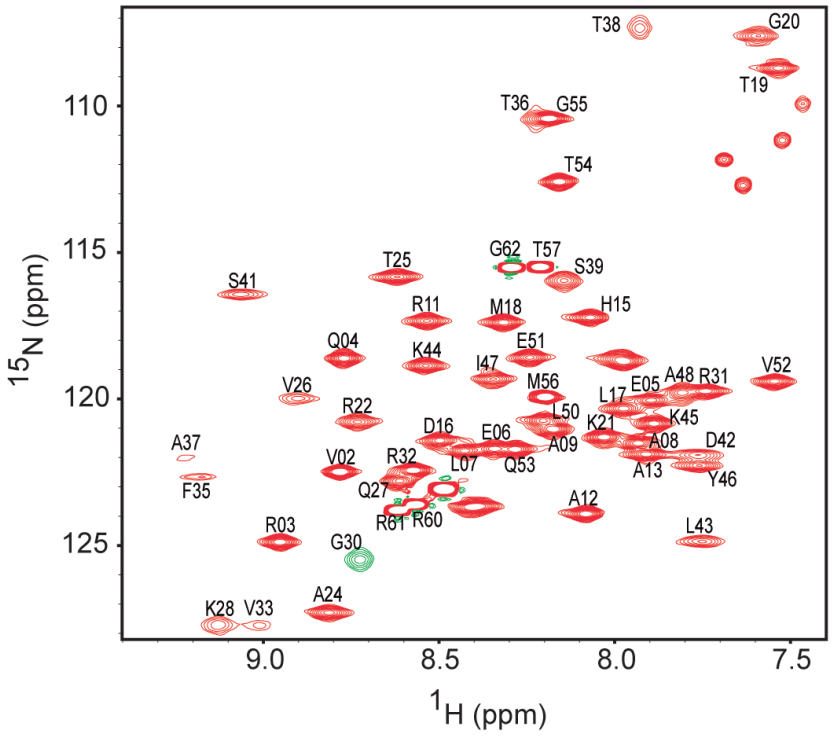


Figure 8.4. – [^{1}H–^{15}N]-HSQC of gpW at 1 ºC and pH 3.5 performed in a 500 MHz magnet. The assignment is indicated for each peak. In green the aliased cross-peak.

The presence of a change on $R_{2,eff}$ indicates the existence of conformational exchange. Therefore, to obtain detailed information about the whole protein RD-NMR experiments were performed on different type of nuclei: on the backbone ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$ and $^{13}C\alpha$ to find out about the secondary structure of the protein; while the methyl-^{13}C reports on the different hydrophobic packing of each conformation. As previously explained in Chapter 3.5.3, CPMG-RD experiments consist of series of 2D HSQC experiments acquired with a different number of π pulses (180º pulses) within the same T time, the CPMG fields. The analysis of all the experiments for the same nuclei creates the relaxation dispersion profiles.

Dispersion profiles with $R_{ex} > 5\ s^{-1}$ reveal the presence of exchange between the conformations of the native state and one or more excited states, and were found in: 22 residues for ^{15}N, 32 for $^{1}H^{N}$, 24 for $^{13}C^{\alpha}$, and 25 for $^{13}C^{O}$. This R_{ex} is the difference: $R_{2,eff}(\nu_{CPMG} = \nu_{min}) - R_{2,eff}(\nu_{CPMG} = \infty)$, where ν_{min} is the minimum CPMG pulsing frequency for each nuclei at each field: 33.3, 55.6, 83.3, 62.5 and 66.7 Hz for ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and $^{13}CH_3$, respectively. Some of these dispersion profiles are represented in Figure 8.5, with different spin probes from α-helices (in orange and red colours) and β-sheets (in cyan and blue). In general, backbone spin probes (A, B, C and D of Figure 8.5) located in the β-sheets show dispersion profiles (G30, K28, V26, R22, etc.), denoting the presence of conformational exchange between ground and excited states, while those in the α-helices (M56, K45, D42, A48, etc.) are relatively flat, meaning very little or no involvement in the exchange. Thus, the major conformational exchange is found in the β-hairpin, with this area changing between the ground and the excited states, at the time that both helices remain almost identical, except in some residues closer in the amino acid sequence to the hairpin, like V40 located at the beginning of α2-helix, adjacent to β2-sheet, with relevant R_{ex} contributions. However, some of the spin probes within the helices also present dispersion profiles, it is the case of the side chain methyl groups facing the hairpin (Figure 8.5 E and F), like L43 from the α2 helix that faces V33 in the β2 sheet (see Figure 8.2 for locations), both with pronounced dispersion profiles, indicating a change in the hydrophobic interface of the excited state. Side chain and backbone probes from the α-helical regions pointing out of the interface lack the exchange profile, like L07 residue, with flat profiles for backbone ^{15}N and methyl groups (Figure 8.5 A and E).

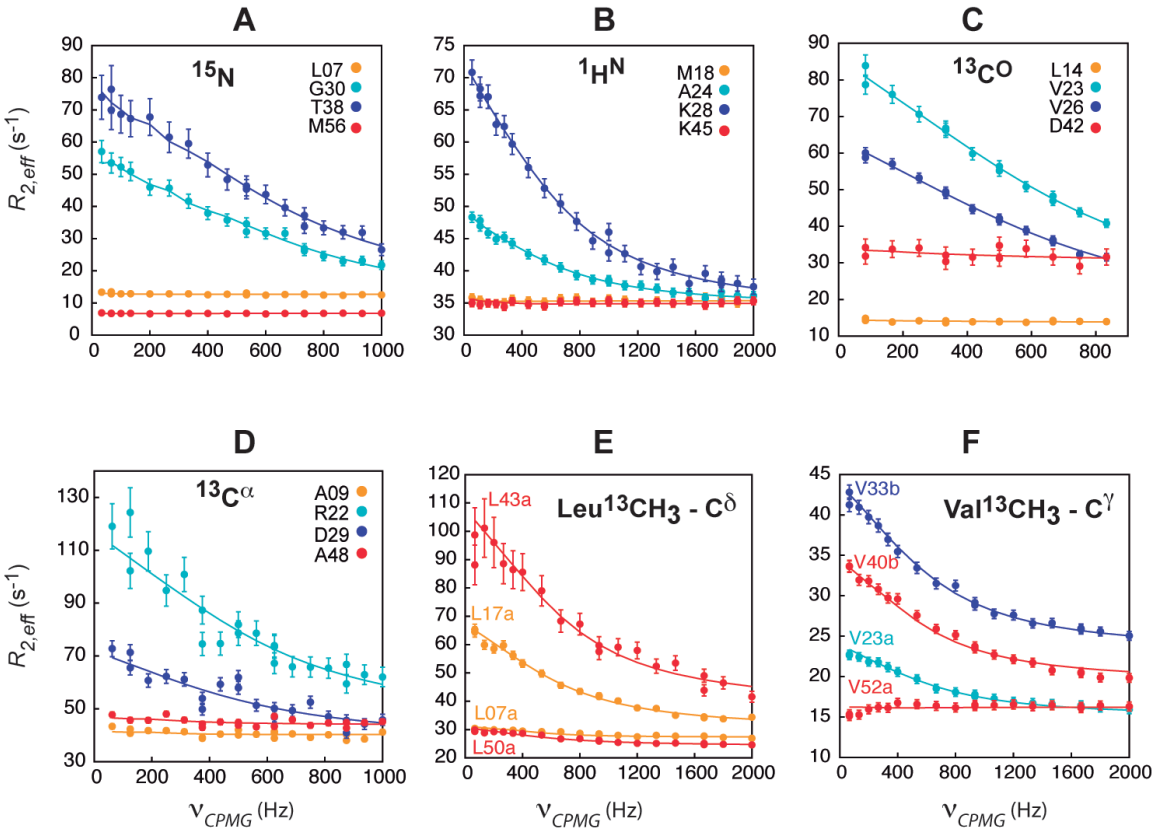


Figure 8.5.– Relaxation dispersion profiles at 1 ºC of ^{15}N (**A**), ^{1}H^{N} (**B**), ^{13}C^{O} (**C**), ^{13}C$^{\alpha}$ (**D**), and methyl-^{13}C from Leu residues (**E**) and Val residues (**F**), recorded at 500 MHz. Circles represent experimental data with its corresponding error bars, and the solid lines are the best fit to a two-state exchange model. Cyan and blue represent residues located in the β-hairpin, while orange and red residues located in the α-helices. Stereospecific assignments of the prochiral methyl groups have not been performed, thus methyl groups from the same residue are distinguished with the letters 'a' or 'b' after the residue, corresponding 'a' with the most upfield shifted in the ^{13}C dimension.

The dispersion profiles from backbone (^{15}N and ^{1}H^{N}) and side chain (methyl-^{13}C) with significant R_{ex} contributions (R_{ex} > 5 s^{-1}) were fitted as explained in Chapter 3.5.3 to the Bloch-McConnell equations[149] for two-site exchange processes, obtaining a k_{ex} of 4087 ± 42 s^{-1}. The k_{ex} value obtained was confirmed with the minimum in the graph of reduced χ^2 (χ^2_{red}) *versus* different k_{ex} values (Figure 8.6 A and B). This k_{ex} implies that the process is in the moderately fast regime, where the extraction of a reliable p_X (population of the excited, *X*, state, called in Chapter 3.5.3 as p_B) becomes a challenge because it remains coupled in the product p_X (1 − p_X) $\Delta\omega^2$ when the RD profiles are fitted. This coupling creates a situation with a flat profile in the plot of χ^2_{red} *versus* different values of p_X (Figure 8.6C), which would virtually admit any p_X value from 5 % to 50 % (over this value would become the ground state) and the associated adjustment of $\Delta\varpi$. To obtain the structural information on the *X* state, reliable $\Delta\varpi$ are crucial, otherwise, the chemical shifts of

this excited state can not be obtained. Vallurupalli *et al.*[157] got to break this correlation for k_{ex} values up to 6000 s^{-1} with the peak positions of a combination of experiments at different field strengths (500 and 800 MHz) and with different quantum correlations (HSQC and HMQC experiments) – further explanations in Chapters 3.5.3 and 3.5.4. For gpW, 31, 26, and 25 signs were obtained for the 53 $\Delta\varpi$ values measured for ^{15}N, 52 for $^{1}H^{N}$, and 47 for $^{13}C^{O}$, respectively. However, for the $^{13}C^{\alpha}$ shifts was not possible to determine the signs due to the limited quality of $^{1}H^{\alpha}$-$^{13}C^{\alpha}$ correlation maps. The inclusion of the peak position shifts with the previous RD profile information broke the correlation between p_X and $\Delta\omega$, and allowed a minimum in the χ^2_{red} *versus* p_X graph at 9.0 ± 0.3 % of population of the excited state (Figure 8.6D).

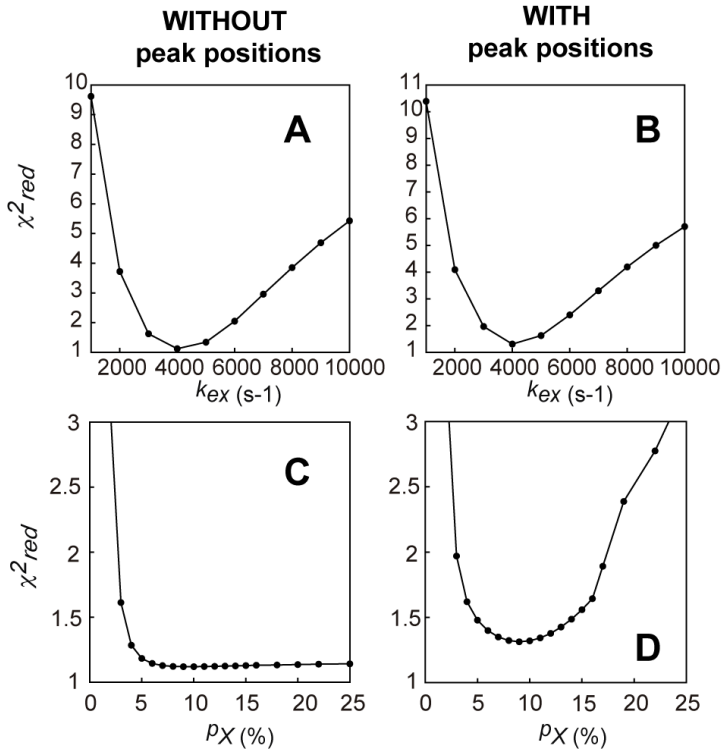


Figure 8.6.– Plots of χ^2_{red} *versus* k_{ex} (**A** and **B**) and p_X (**C** and **D**) from the fitting of the RD profiles of ^{15}N, $^{1}H^{N}$ and methyl-^{13}C at 1 ºC in 500 and 800 MHz fields. For k_{ex} a minimum in the graph is obtained with (**A**) and without (**B**) the information about the position of the peaks obtained from HSQC and HMQC experiments at both fields, while for p_X only the addition of the peak positions allows to break the coupling of p_X with $\Delta\omega$ and generate a minimum that confirms the reliability of the value (**D**).

Once reliable k_{ex} and p_X values are obtained, fixing them to a two-state exchange model and fitting the RD profiles, allows to get chemical shifts for the different nuclei for which we have performed CPMG experiments: ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$, and methyl-^{13}C. Some large chemical shifts are observed for the nuclei, feature

of a structural change. To reveal if the change leads towards a random coil structure, the chemical shift differences experimentally obtained ($\Delta\varpi_{exp} = \Delta\varpi_{XN} = \varpi_X - \varpi_N$) were compared with the values corresponding to a folding-unfolding transition ($\Delta\varpi_{pred} = \varpi_U - \varpi_N$) and plotted in Figure 8.7. All chemical shifts for the native state (ϖ_N), chemical shift differences ($\Delta\varpi_{XN}$), and chemical shifts expected for the random coil (ϖ_{RC}, predicted using Neighbor Corrected Intrinsically Disordered Protein Library[241]) are included for ^{15}N, ^{1}H^{N}, ^{13}C^{O}, ^{13}C$^{\alpha}$ and methyl-^{13}C (except the ϖ_{RC} for this last one) in the Supplementary Information of the paper included as Appendix B.

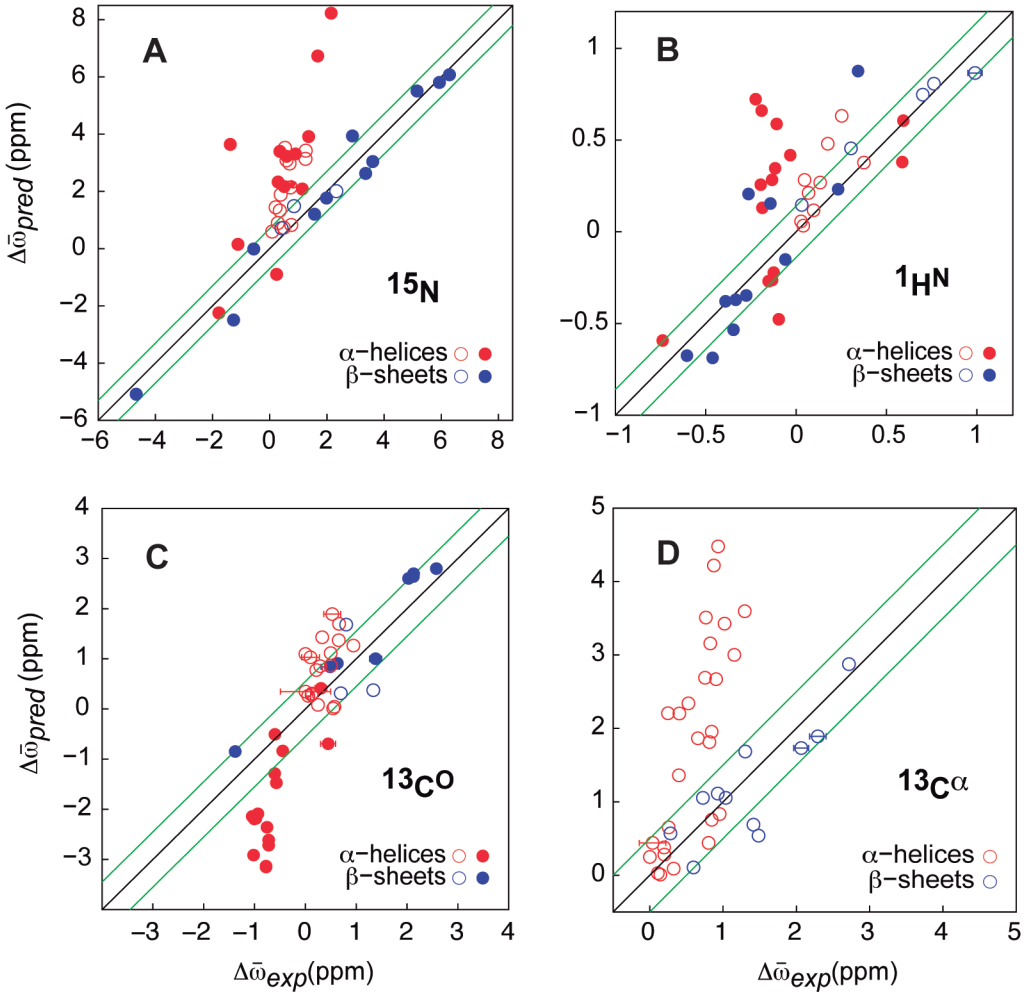


Figure 8.7.– Chemical shifts differences ($\Delta\varpi$, in ppm) between native and excited states from the experiments on gpW at 1 ºC ($\Delta\varpi_{exp}$) *versus* the predicted chemical shifts if the excited state were a random coil ($\Delta\varpi_{pred}$). The colour code of the circles corresponds to red for the α-helical nuclei, and blue for those located in the β-hairpin. Nuclei for which the sign of the $\Delta\varpi$ was obtained are plotted as filled circles, while the cases with no available sign are indicated as open circles and the $|\Delta\varpi_{pred}|$ *versus* $|\Delta\varpi_{exp}|$ were plotted. Error bars smaller than the circle are omitted. The black line is the reference $y = x$, and the green lines the uncertainties in $\Delta\varpi_{pred}$ based on the error in the predicted chemical shift of the unfolded values according to Tamiola *et al.*[241] of 0.65, 0.14, 0.5 and 0.41 ppm for ^{15}N, ^{1}H^{N}, ^{13}C^{O} and ^{13}C$^{\alpha}$, respectively.

For all the backbone nuclei a general trend is followed as can be seen in Figure 8.7, with the β-sheet nuclei (in blue) correlating with the $\Delta\varpi$ expected for an unfolding process, while the nuclei in the α-helices (in red) show no correlation, which is especially remarkable for the ^{15}N nuclei (Figure 8.7A). Both ^{13}C plots ($^{13}C^{O}$ and $^{13}C^{\alpha}$ in Figures 8.7 C and D), very sensitive to changes in the secondary structure, follow the same patterns, consistent with unfolded β-sheets and intact α-helices in the invisible state. These results agree with the information obtained from TALOS+ (Torsion Angle Likeliness Obtained from Shift and Sequence Similarity)[85] plotted in Figure 8.8, with the α-helices in the excited state mainly intact, showing a high confidence even when the missing signs of $\Delta\varpi_{XN}$ from many ^{13}C are chosen as farther from the random coil values (Figure 8.8C). Moreover, when the missing signs are chosen as closer to the random coil values (Figure 8.8B), they remain highly conserved. For TALOS+ to provide the information is required the chemical shifts from different backbone nuclei of the conformation, however some values were missing for the experiments at 1 ºC, lacking that residue information in the plots of Figure 8.8.

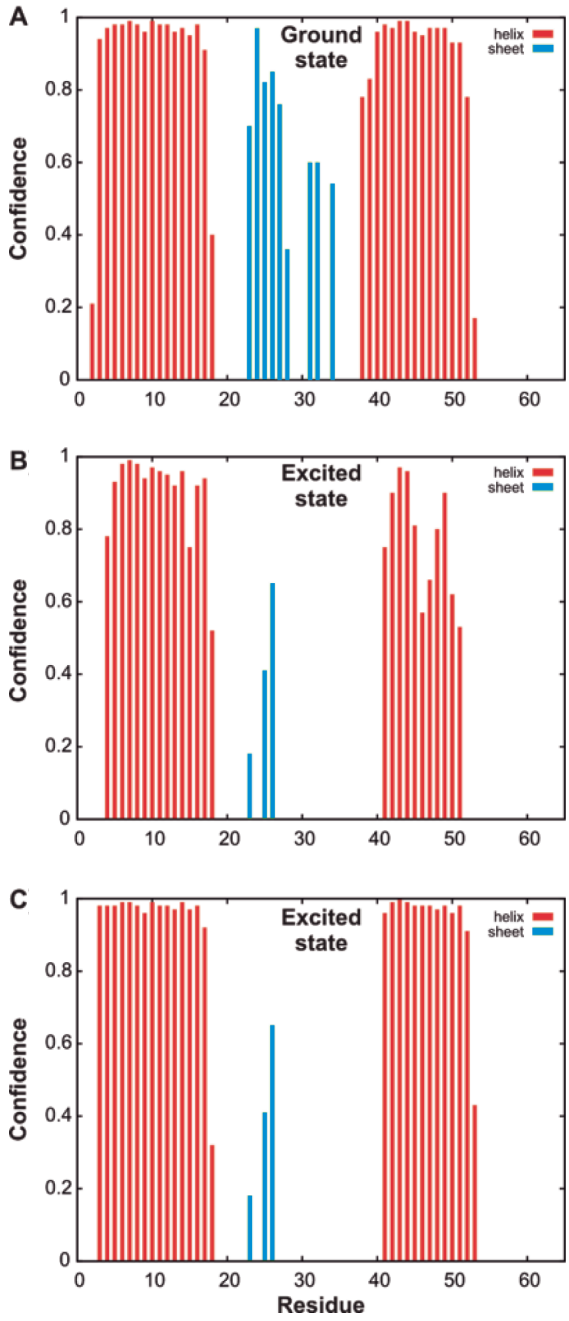


Figure 8.8.– Confidence of the residues of gpW in the secondary structure of the invisible state at 1 ºC as predicted by TALOS+[85]. **A)** Predicted secondary structure for the native or ground state. **B and C)** Predicted secondary structure for the excited state. In both plots for residues located in the β-hairpin (in blue) where the signs of $\Delta\varpi_{XN} = \varpi_X - \varpi_N$ were not available, they were chosen so the ϖ_X is moving towards the random coil chemical shifts. The difference between these last two plots is when the signs of ϖ_X in the α-helices (in red) were not available, they were chosen as closer and farther to the random coil chemical shifts for **B** and **C**, respectively.

The change on the central region of gpW protein, in the β-hairpin, is very striking in the Figure 8.9 by the high $\Delta\varpi_{RMS}$ values in that area when all residues are evaluated. $\Delta\varpi_{RMS}$ is obtained from $\Delta\varpi_{RMS} = [(1/N)\Sigma_i(\Delta\varpi_i/\Delta\varpi_{i,STD})^2]^{1/2}$, where *N* is the number of nuclei included in the sum (≤4, because are included ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$ and $^{13}C^{\alpha}$), $\Delta\varpi_i$ is the difference between the chemical shifts in the native and invisible state for one of the nucleus in a given residue, and $\Delta\varpi_{i,STD}$ corresponds to the standard deviation of the distribution of chemical shifts for that nucleus in the context of the amino acid under consideration, which is obtained from the BMRB (Biological Magnetic Resonance Data Bank)[242]. Therefore, the evaluation of the backbone chemical shifts at each residue in Figure 8.9 stands out the significant variations within the central, β-sheets region ($\langle \Delta\varpi_{RMS} \rangle$ = 0.7 ± 0.3), while in both side helices the changes are smaller ($\langle \Delta\varpi_{RMS} \rangle$ = 0.2 ± 0.1, and 0.2 ± 0.2 for α1 and α2 helices, respectively).

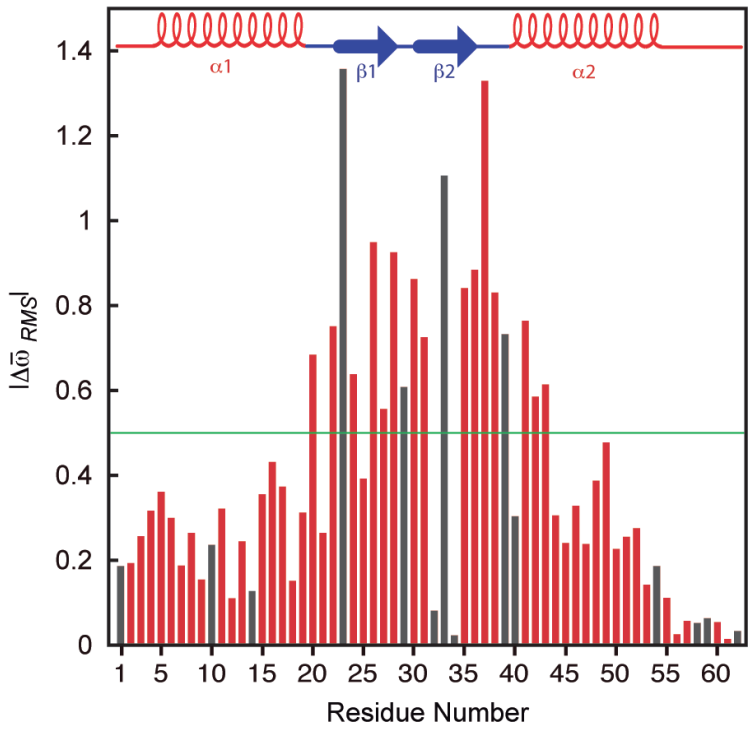


Figure 8.9.– Distribution of $|\Delta\varpi_{RMS}|$ values along the protein sequence. Red bars indicate values obtained when three or more $\Delta\varpi$ values are considered, while in the gray bars, only two or fewer, which can happen when the residue is overlapped in the [$^{1}H^{N}$-^{15}N]-HSQC (like the cases of A10, L14, E34, E49, Q58 and R59) or when the peaks are too broad due to exchange (V23, D29, V33, and V40), precluding the acquisition of reliable chemical shifts. The green line at 0.5 is used as threshold for visual purposes.

8.4.- Kinetic Experiments: Infrared Temperature-Jump

Kinetic experiments measure the protein relaxation rate after an initial perturbation, in our study a temperature jump (T-jump) with infrared absorption as a probe. Experiments on gpW were performed by Dr. Michele Cerminara on a custom-built apparatus[243] in which an infrared laser pulse induces a temperature jump of 10 – 12 ºC when impinging on the sample and creates the perturbation. The excitation beam originates from a Nd:YAG laser (Continuum Surelite I) operating at 4 Hz, whose wavelength is shifted from the fundamental at 1064 nm to 1907 nm by using a 1m-long Raman cell (Light Age Inc) filled with a high-pressure mixture of H_2 and Ar in order to match the energy of the vibrational overtones of D_2O that is used as solvent for these experiments. The resulting pulses have an energy of about 30 mJ/pulse and a duration of 5-10 ns. The probe beam is a far-IR Quantum cascade laser (Daylight Solutions) that is tuneable in the range between 1605 – 1690 cm^{-1} range to match the maximum absorption of the protein in the α-helix Amide I band, for the experiments on gpW, it was set at 1632 cm^{-1}. The light transmitted from the sample was recorded by a fast mercury cadmium telluride (MCT) detector (Kolmar Technologies) coupled to an oscilloscope (Tektronix DPO4032). The samples were deuterated with the same procedure used for equilibrium FTIR (see Chapter 2.2) through series of deuteration cycles to exchange the amide protons with deuterium. After last cycle the lyophilized sample was dissolved in deuterated 20 mM glycine buffer with a protein concentration of 4 mg/ml at pD 3.5 and placed in a thermostated custom-built sample holder between two MgF_2 windows separated by a 50 μm Teflon spacer. The experimental data from the IR T-jump (relaxation rate and amplitude as a function of temperature) are fitted to a model of four parameters (ΔH_{res}, x, k_0 and E_a), obtaining from the fit the overall parameters for gpW at pH 3.5 of ΔH_{res} = 4.23 kJ/mol, x = 0.94, k_0 = 1.37 $\cdot 10^{12}$ s^{-1} and E_a = 0.877 kJ/mol.

Temperature jump techniques are off-equilibrium experiments; they measure the relaxation from a perturbed state back to the base temperature at which the sample is thermostated. Thus, these measurements can be performed only in a range of temperatures in which there is a relevant change of the protein conformation between the perturbed state and the base temperature, which means in a rage of ± 20 K around the *Tm* of the protein. Therefore, the experiments were designed at a value of pH that would allow to have a *Tm* low enough to get closer to

the values of temperature imposed by the NMR experiments (that required the temperature to be low enough to slow down the folding process in order to have enough sampling to measure a decay). Putting together the requirements of the two techniques, the NMR experiments were performed at 274 K (very far from the *Tm* of the protein that at this pH is 330 K) while the T-jump experiments showed a measurable amplitude only above 295K. Therefore, T-jump experiments were performed every 5 K between 295 and 340 K and then fitted to a 1D Free Energy Surface (FES) model to extrapolate the value of the rate at lower temperatures (Figure 8.10) obtaining a rate constant of 5000 ± 100 s^{-1} at 274 K used by NMR (Figure 8.10 inset). The uncertainty of ± 100 s^{-1} is one standard deviation in the rates obtained from jackknife analysis[244]. Considering the extrapolation to values far from the T-jump data and the fixed $\Delta C_{p,res}$ (the relevance of this parameter is explained in detail in the next section), this estimation is very close to the k_{ex} experimentally obtained at the same temperature from RD-NMR, 4087 ± 42 s^{-1}. The similar rates suggest that *X* is in the unfolding pathway of the protein, probably as rate-limiting step, even if the RD-NMR data could not assure it due to the lack of a detectable unfolded conformation at this low temperature.

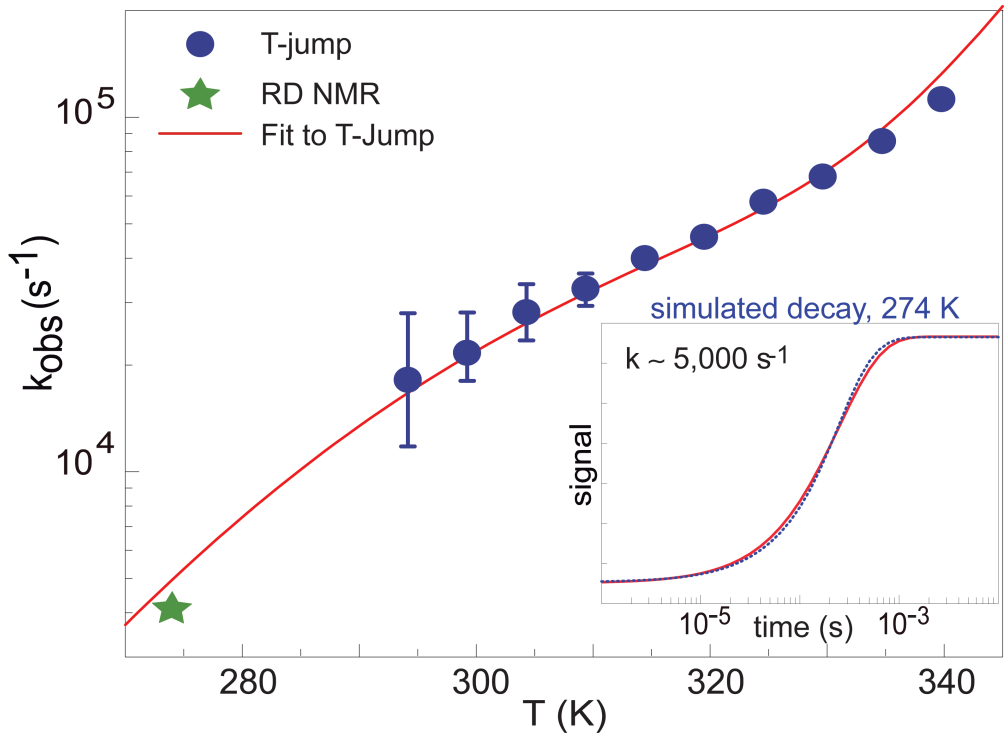


Figure 8.10. – Infrared relaxation kinetics of gpW at pD 3.5. Blue circles represent the experimental T-jump values obtained at different temperatures, the red curve their fitting to a 1D FES model and the green star is the k_{ex} value obtained from RD NMR at 274 K. Inset shows a decay at 274 K simulated by the FES model (blue dashed line) and its single exponential fit (red line).

8.5.- Free Energy Surface (FES) Model

The model used to analyze the IR T-jump data is based on a mean-field approximation and represents the energy landscape as a one-dimensional FES resultant from the order parameter nativeness of the protein[31,204]. The explanation of the model follows the detailed description of DeSancho *et al.*[204]. The nativeness (*n*) is measured as the probability of finding the different residues in native conformation by looking at the dihedral angles (φ and ψ), thus, it can range from 0 for the fully unfolded protein (with none of the angles with native values) to 1 for the native conformation (with all the angles in folded native conformation). This nativeness will define the changes in conformational entropy of the protein (ΔS_{conf}) according to Equation 8.1:

$$\Delta S_{\mathrm{conf}}(\mathrm{n}) = \mathrm{N}\,(-\mathrm{R}\,[\mathrm{n}\ln(\mathrm{n}) + (1-\mathrm{n})\ln(1-\mathrm{n})] + (1-\mathrm{n})\Delta S_{\mathrm{conf,res}})$$

Equation 8.1

where *N* is the number of residues in the protein and $\Delta S_{conf,res}$ the difference in the conformational entropy between the folded and the unfolded conformation of a protein residue (by fixing it with native angle values). In this work, $\Delta S_{conf,res}$ is going to be fixed to 16.5 J/(mol·K), generated by calculating the average value obtained from the Robertson and Murphy analysis of a database of calorimetric data[245].

The change in stabilization enthalpy is taken also as a function of the nativeness ($\Delta H(n)$) - which excludes the solvation free energy - and is modelled as a probabilistic Markov chain, depending on the breaking of the native interactions as the protein unfolds, with constant differential changes in the decrease of nativeness as Equation 8.2 shows:

$$\Delta\mathrm{H}(\mathrm{n}) = \mathrm{N}\,\Delta\mathrm{H}_{\mathrm{res}}[(1-\mathrm{x}^{(1-\mathrm{n})})/(1-\mathrm{x})]$$

Equation 8.2

where ΔH_{res} is the total stabilization energy per residue, and *x* determines the rate of breaking native interactions, which is a function of *n*. Thus, the term within square brackets in Equation 8.2 depending on *x* indicates the fraction of remaining native stabilization energy, with the same values as previously mentioned for *n*, ranging from 1 for the native conformation to 0 for the fully unfolded protein, and will be

affected by local and non-local stabilization interactions. Therefore, from Equation 8.2 we can infer that the decay in stabilization enthalpy is exponential as the protein loses its native structure.

With a parallel treatment as a Markov chain, the change in heat capacity is also determined as a function of nativeness ($\Delta C_p(n)$) as Equation 8.3 presents

$$\Delta C_p(n) = N\,\Delta C_{p,res}[(1 - x_C^{(1-n)})/(1 - x_C)]$$

Equation 8.3

where $\Delta C_{p,res}$ is the change in heat capacity per residue and x_C the characteristic rate of change of the heat capacity depending on n. For simplicity, $\Delta C_{p,res}$ is fixed to the average value of 53.5 J/mol K, very close to 52.5 J/mol K obtained empirically by Robertson and Murphy[245] and x_C to 0.0136, previously used by our group[31].

Therefore, the free energy surface at any given temperature (T) is given by Equation 8.4 where we can obtain the $\Delta S(n)$, $\Delta H(n)$ and $\Delta C_p(n)$ from Equations 8.1, 8.2 and 8.3, respectively.

$$\Delta G(n) = \Delta H(n) - T\,\Delta S(n) + \Delta C_p(n)[(T - T_m) + T\ln(T/T_{ref})]$$

Equation 8.4

where T_m is the denaturation midpoint temperature of the protein and T_{ref} is 385 K, the convergence temperature for the folding entropy where the solvation free energy cancels out[31].

The relaxation kinetics as consequence of a perturbation on the FES are treated as a simple diffusion model with a constant diffusion coefficient as function of nativeness, but dependent on the solvent viscosity and temperature according to Equation 8.5

$$D = \frac{k_0}{\eta(T)}\exp(-N\,E_a/RT)$$

Equation 8.5

where η is the viscosity of the water as a function of temperature, k_0 is the pre-factor and E_a is the activation energy per residue (constitutes the temperature dependence

of *D*). Consequently, both, *D* (Equation 8.5) and any change in the FES shape (Equation 8.4), determine the temperature dependence of the relaxation rate.

8.6.- Discussion

The protein folding is a complex process that involves interactions within the amino acid chain towards the native conformation of the protein. Underneath, the process is driven by the free energy surface that may involve local minima, producing transient conformations, and energy barriers, hindering the exchange between states. The conformations along the folding process or as result of kinetic traps sometimes are low populated, and only some techniques are able to detect them, like DSC[246] (appearing as a complex unfolding behaviour) or kinetic experiments[28]. RD-NMR has been proven as a powerful technique to structurally characterize these "invisible" conformations located at local minima of the FES. However, RD-NMR has certain timescale limitations and fast-folding proteins are in principle out-of-range. For folding to proceed in the microseconds timescale or faster, the underlying FES requires low barriers and shallow minima, favouring the appearance of partially folded conformations. Atomic-level molecular dynamics simulations have been able to fold and unfold multiple times 12 such fast-folding proteins[42], noticing that the native structure is reached in a stepwise fashion, with some secondary elements appearing at an early stage of the folding, and acting as template for other native contacts.

The protein gpW with a folding relaxation rate of 125000 s^{-1} at 67 ºC[6] was way out of the window for RD-NMR. However, the folding rates can be altered, for example by changing the temperature. In the case of gpW, the temperature was decreased to 1 ºC, where the extrapolation from T-jump experiments produces a folding rate 5000 ± 100 s^{-1} (Figure 8.10). In fact, this temperature slowed down the folding rate enough to acquire RD profiles with significant R_{ex} for ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$, and methyl-^{13}C nuclei, that when fitted to a two-site exchange model, produced a k_{ex} of 4087 ± 42 s^{-1} (with a of χ^2_{red} ~ 1, Figure 8.6A and B), in reasonable agreement with the result from T-jump experiments considering the 20 ºC extrapolation. However, the relatively fast exchange could not break the correlation between p_X and $\Delta\omega$, until the peak position was included to the profiles for the fitting, obtaining then a p_X of 9.0 ± 0.3% (Figure 8.6C and D). This information of the peak positions also allowed to get the chemical shifts of the excited state *X*, which comparison with

the values of the random coil values (Figure 8.7) points towards the unfolding of the β-hairpin structure, while both α-helices remain pretty intact (Figure 8.9). Thus, even at the low temperature of 1 ºC, the secondary structure of gpW appears with a higher stability of the helices compared with the hairpin. This hierarchical structural stability has been observed in long-time-scale molecular dynamic simulations for the folding and unfolding of gpW, indicating that near the denaturation midpoint a similar structure (with well defined α-helices and unstructured β-hairpin) is formed[8]. Furthermore, simulations and NMR experiments were both consistent with the appearance of the helices prior to the hairpin during the folding process[8].

This *X* state obtained with a population close to 10% can be considered a rate-limiting step in the unfolding of gpW due to the similar rates from the T-jump experiments and the fitting of the RD-NMR data. The usual chemical kinetics interpretation of the RD-NMR data, would be to locate this *X* state as a local minimum in the FES as on-pathway intermediate separated by barriers of undetermined height as shown on the top of Figure 8.11. However, another interpretation arises from the FES analysis of the T-jump data, with *X* state as an ensemble of partly folded conformations at the top of the barrier observed by IR spectroscopy near the *Tm* represented at the bottom Figure 8.11. Under highly stabilizing conditions, like the 1 ºC of the RD-NMR experiments, the FES of gpW is predicted to become downhill (green surface in Figure 8.11), and the low barrier at the *Tm* (in magenta in Figure 8.11) changes into a saddle point where the “invisible” *X* state interconverts with the native state with almost single-exponential kinetics (Figure 8.10 inset) and a probability of ~10% of the population (Figure 8.11 inset), very close to the value obtained from RD-NMR. Both interpretations agree with the RD-NMR data in which *N* and *X* states interconvert with a 9:1 ratio.

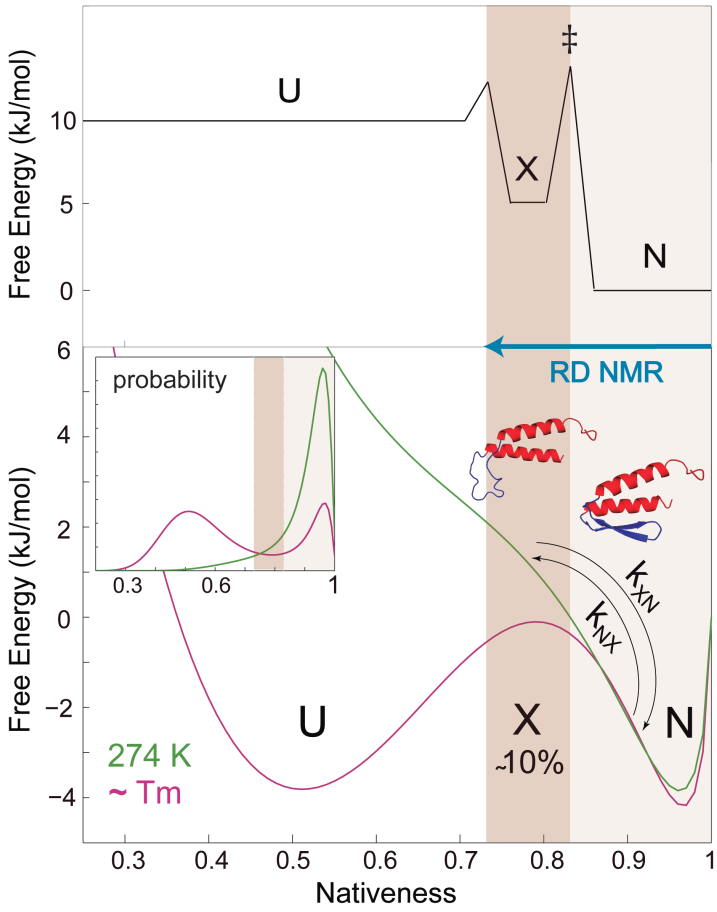


Figure 8.11. – The top FES figure is a schematic representation of the native (N), invisible (X) and unfolded (U) conformations in terms of a chemical kinetic model, and separated by energy barriers. The bottom figure represents the FES obtained from Equation 8.4 for the *Tm* (magenta) and 1 ºC (green) as a function of the nativeness of the protein. The inset in the bottom figure shows the probability distribution at *Tm* and 1 ºC with the different shadowing indicating the different conformations. The blue arrow, between the upper and bottom figure, indicates the structures reached with the RD NMR experiments performed on gpW.

Even gpW has been studied at the low pH 3.5, the *X* state would be comparable with the intermediate obtained at a more neutral pH of 6.5, as Figure 8.12 shows comparing the $|\Delta\varpi_{XN}|$ data of ^{15}N RD experiments at 1 ºC and both pH values. The strong correlation supports the existence of the *X* state over different pH values, from acid to neutral, coherent with the also similar native structures obtained at both pH (Figure 8.3B).

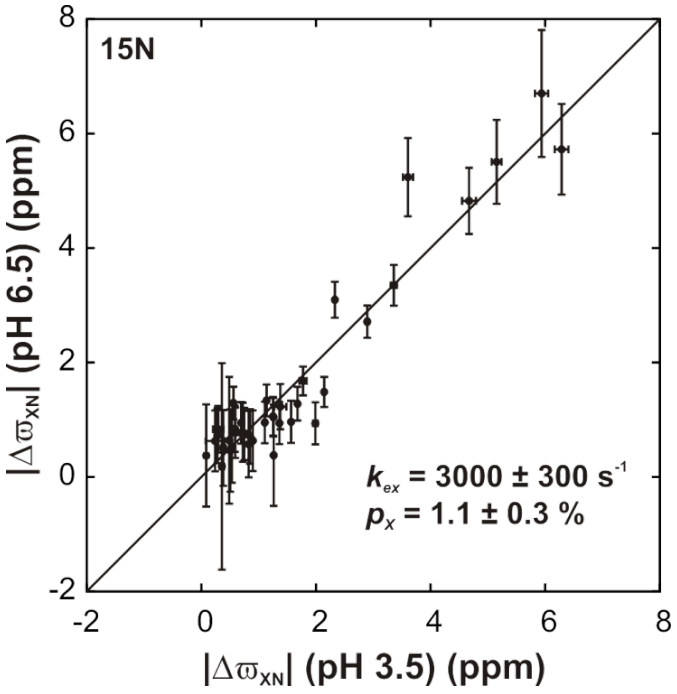


Figure 8.12. – Comparison of the $\Delta\varpi_{XN}$ at the pH 3.5 used for all the experiments on gpW *versus* pH 6.5, all performed at 1 ºC.

8.7.- Conclusions

In the detailed analysis of the atomic behaviour obtained by RD-NMR, *X* state appears in the folding process of gpW protein, even at low temperatures, as a productive intermediate at the end of the kinetic folding pathway. In this *X* state, the α-helices are formed while the β-hairpin is still unfolded, probably because the hairpin may be the last step in the folding process. The lower stability of the hairpin in gpW protein had been observed previously with low-resolution techniques, that detected a decreased stability in the tertiary contacts with the β-hairpin than within the α-helices[6], and with long-time-scale simulations[8].

This low stability of the β-hairpin was observed in other systems, and has been proposed to play a possible role as conformational switch[247]. Considering the location of gpW protein in the collar of the λ bacteriophage, the functions of the full-length gpW protein as structural protein or as interacting with the DNA (through direct binding or as a plug to prevent the ejection of the DNA[248]) have been suggested[4]. Maxwell *et al.*[248] proposed the hypothesis that the protein (monomeric in solution) polymerizes into a ring with a six-fold rotational symmetry creating a gate for the DNA flow during infection, with a structure comparable to the DNA

sliding clamps[248]. Maybe the features of gpW with the hydrophobic tail at the C-terminus, where the last three residues are crucial for its activity (its lack impedes the assembly)[238], and the flexible hairpin, exchanging between folded and unfolded structures, can involve a double function for this protein. Therefore, gpW could be anchored to other proteins by hydrophobic interactions arising specially from the C-terminal tail, at the time that the hairpin, with several basic residues (pI of 10.8 for the full length protein[249]), may be interacting with the DNA, allowing or blocking its flow across the collar of the bacteriophage by the interconversion of both conformations.

BIBLIOGRAPHY

BIBLIOGRAPHY

1 Anfinsen, C. The formation and stabilization of protein structure. *Biochemical Journal* **128**, 737 (1972).

2 Reynaud, E. Protein misfolding and degenerative diseases. *Nature Education* **3**, 28 (2010).

3 Gruebele, M. in *Protein Folding, Misfolding and Aggregation* 106-138 (RSC Publishing: London, 2008).

4 Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein-Folding: A Synthesis *Proteins-Structure Function and Genetics* **21**, 167-195 (1995).

5 Eaton, W. A. Searching for "downhill scenarios" in protein folding. *Proceedings of the National Academy of Sciences* **96**, 5897-5899, doi:10.1073/pnas.96.11.5897 (1999).

6 Fung, A., Li, P., Godoy-Ruiz, R., Sanchez-Ruiz, J. M. & Munoz, V. Expanding the realm of ultrafast protein folding: gpW, a midsize natural single-domain with alpha+beta topology that folds downhill. *J. Am. Chem. Soc.* **130**, 7489-7495, doi:10.1021/jaB01401a (2008).

7 Sadqi, M., Fushman, D. & Munoz, V. Atom-by-atom analysis of global downhill protein folding. *Nature* **442**, 317-321, doi:10.1038/nature04859 (2006).

8 Sborgi, L. *et al.* Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *J. Am. Chem. Soc.* **137**, 6506-6516 (2015).

9 Campos, L. A. *et al.* Gradual Disordering of the Native State on a Slow Two-State Folding Protein Monitored by Single-Molecule Fluorescence Spectroscopy and NMR. *The Journal of Physical Chemistry B*, doi:10.1021/jp403051k (2013).

10 Anfinsen, C. B., Haber, E., Sela, M. & White, F. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences* **47**, 1309-1314 (1961).

11 Levinthal, C. in *Proceedings of a meeting held at Allerton House.* 22-24 (University of Illinois Press: Urbana, IL, USA).

12 Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223-230, doi:10.1126/science.181.4096.223 (1973).

13 Levinthal, C. Are there pathways for protein folding. *J. Chim. phys* **65**, 44-45 (1968).

14 Go, N. Theoretical studies of protein folding. *Annual review of biophysics and bioengineering* **12**, 183-210 (1983).

15 Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural Biology* **4**, 10-19 (1997).

16 Munoz, V., Sadqi, M., Naganathan, A. N. & de Sancho, D. Exploiting the downhill folding regime via experiment. *Hfsp Journal* **2**, 342-353, doi:10.2976/1.2988030 (2008).

17 Tanford, C. Protein denaturation. *Adv Protein Chem* **23**, 121-282 (1968).

18 Jackson, S. E. How do small single-domain proteins fold? *Folding & Design* **3**, R81-R91 (1998).

19 Muñoz, V. & Cerminara, M. When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. *Biochemical Journal* **473**, 2545-2559 (2016).

20 Naganathan, A. N. & Muñoz, V. Thermodynamics of downhill folding: multi-probe analysis of PDD, a protein that folds over a marginal free energy barrier. *The Journal of Physical Chemistry B* **118**, 8982-8994 (2014).

21 Naganathan, A. N. & Munoz, V. Scaling of folding times with protein size. *J. Am. Chem. Soc.* **127**, 480-481, doi:10.1021/ja044449u (2005).

22 Shlesinger, M. F., Zaslavsky, G. M. & Klafter, J. Strange kinetics. *Nature* **363**, 31-37 (1993).

23 Sabelko, J., Ervin, J. & Gruebele, M. Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences* **96**, 6031-6036 (1999).

24 Garcia-Mira, M. M., Sadqi, M., Fischer, N., Sanchez-Ruiz, J. M. & Munoz, V. Experimental identification of downhill protein folding. *Science* **298**, 2191-2195 (2002).

25 Hagen, S. J. Exponential decay kinetics in "downhill" protein folding. *Proteins: Structure, Function, and Bioinformatics* **50**, 1-4 (2003).

26 Liu, F. & Gruebele, M. Tuning λ6-85 towards downhill folding at its melting temperature. *Journal of Molecular Biology* **370**, 574-584 (2007).

27 Munoz, V. Thermodynamics and kinetics of downhill protein folding investigated with a simple statistical mechanical model. *International Journal of Quantum Chemistry* **90**, 1522-1528, doi:10.1002/qua.10384 (2002).

28 Li, P., Oliva, F. Y., Naganathan, A. N. & Munoz, V. Dynamics of one-state downhill protein folding. *Proc Natl Acad Sci U S A* **106**, 103-108 (2009).

29 Oliva, F. Y. & Muñoz, V. A Simple Thermodynamic Test To Discriminate between Two-State and Downhill Folding. *J. Am. Chem. Soc.* **126**, 8596-8597, doi:10.1021/ja048309w (2004).

30 Muñoz, V. & Sanchez-Ruiz, J. M. Exploring protein-folding ensembles: a variable-barrier model for the analysis of equilibrium unfolding experiments. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 17646-17651 (2004).

31 Naganathan, A. N., Doshi, U. & Muñoz, V. Protein folding kinetics: Barrier effects in chemical and thermal denaturation experiments. *J. Am. Chem. Soc.* **129**, 5673-5682, doi:10.1021/ja0689740 (2007).

32 Sadqi, M., Fushman, D. & Muñoz, V. Structural Biology: Analysis of'downhill'protein folding; Analysis of protein-folding cooperativity (Reply). *Nature* **445**, 17-18 (2007).

33 Capaldi, A. P., Shastry, M. R., Kleanthous, C., Roder, H. & Radford, S. E. Ultrarapid mixing experiments reveal that Im7 folds via an on-pathway intermediate. *Nature Structural & Molecular Biology* **8**, 68-72 (2001).

34 Alm, E. & Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 11305 (1999).

35 Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39-42 (2000).

36 Gruebele, M. Protein folding: the free energy surface. *Current Opinion in Structural Biology* **12**, 161-168, doi:http://dx.doi.org/10.1016/S0959-440X(02)00304-4 (2002).

37 Muñoz, V., Campos, L. A. & Sadqi, M. Limited cooperativity in protein folding. *Current Opinion in Structural Biology* **36**, 58-66 (2016).

38 Kjaergaard, M., Teilum, K. & Poulsen, F. M. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12535-12540, doi:10.1073/pnas.1001693107 (2010).

39 Demarest, S. J. *et al.* Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549-553 (2002).

40 Qin, B. Y. *et al.* Crystal structure of IRF-3 in complex with CBP. *Structure* **13**, 1269-1277, doi:10.1016/j.str.2005.06.011 (2005).

41 Ferreon, J. C. *et al.* Cooperative regulation of p53 by modulation of ternary complex formation with CBP/p300 and HDM2. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6591-6596, doi:10.1073/pnas.0811023106 (2009).

42 Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **334**, 517-520, doi:10.1126/science.1208351 (2011).

43 Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins1. *Journal of Molecular Biology* **277**, 985-994 (1998).

44 Freddolino, P. L., Liu, F., Gruebele, M. & Schulten, K. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.* **94**, L75-L77 (2008).

45 Iglesias-Bexiga, M. *et al.* Protein Folding Cooperativity and Thermodynamic Barriers of the Simplest β-Sheet Fold: A Survey on WW Domains. *The Journal of Physical Chemistry B*, doi:10.1021/acs.jpcb.8b05198 (2018).

46 Naganathan, A. N., Li, P., Perez-Jimenez, R., Sanchez-Ruiz, J. M. & Munoz, V. Navigating the Downhill Protein Folding Regime via Structural Homologues. *J. Am. Chem. Soc.* **132**, 11183-11190, doi:10.1021/ja103612q (2010).

47 Kubelka, J., Chiu, T. K., Davies, D. R., Eaton, W. A. & Hofrichter, J. Sub-microsecond Protein Folding. *Journal of Molecular Biology* **359**, 546-553, doi:10.1016/j.jmb.2006.03.034 (2006).

48 Gruebele, M. Downhill protein folding: evolution meets physics. *Comptes rendus biologies* **328**, 701-712 (2005).

49 Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols* **1**, 2876-2890 (2007).

50 Holzwarth, G. & Doty, P. The ultraviolet circular dichroism of polypeptides1. *J. Am. Chem. Soc.* **87**, 218-228 (1965).

51 Greenfield, N. J. & Fasman, G. D. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* **8**, 4108-4116 (1969).

52 Venyaminov, S. Y., Baikalov, I., Shen, Z. M., Wu, C.-S. C. & Yang, J. Circular dichroic analysis of denatured proteins: inclusion of denatured proteins in the reference set. *Analytical Biochemistry* **214**, 17-24 (1993).

53 Byler, D. M. & Susi, H. Examination of the secondary structure of proteins by deconvolved FTIR spectra. *Biopolymers* **25**, 469-487 (1986).

54 Surewicz, W. K. & Mantsch, H. H. New insight into protein secondary structure from resolution-enhanced infrared spectra. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* **952**, 115-130 (1988).

55 Arrondo, J. L. R., Muga, A., Castresana, J. & Goñi, F. M. Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy. *Progress in biophysics and molecular biology* **59**, 23-56 (1993).

56 Glasoe, P. K. & Long, F. Use of glass electrodes to measure acidities in deuterium oxide. *The Journal of Physical Chemistry* **64**, 188-190 (1960).

57 Ghisaidoobe, A. & Chung, S. *Intrinsic Tryptophan Fluorescence in the Detection and Analysis of Proteins: A Focus on F√ ∂rster Resonance Energy Transfer Techniques*. Vol. 15 (2014).

58 Sanchez-Ruiz, J. M. Probing free-energy surfaces with differential scanning calorimetry. *Annual review of physical chemistry* **62**, 231-255 (2011).

59 Farber, P., Darmawan, H., Sprules, T. & Mittermaier, A. Analyzing protein folding cooperativity by differential scanning calorimetry and NMR spectroscopy. *J. Am. Chem. Soc.* **132**, 6214-6222 (2010).

60 Naganathan, A. N., Perez-Jimenez, R., Muñoz, V. & Sanchez-Ruiz, J. M. Estimation of protein folding free energy barriers from calorimetric data by

multi-model Bayesian analysis. *Physical Chemistry Chemical Physics* **13**, 17064-17076 (2011).

61 Johnson, C. M. Differential scanning calorimetry as a tool for protein folding and stability. *Archives of Biochemistry and Biophysics* **531**, 100-109 (2013).

62 Spink, C. H. Differential scanning calorimetry. *Methods in cell biology* **84**, 115-141 (2008).

63 Demetzos, C. Differential scanning calorimetry (DSC): a tool to study the thermal behavior of lipid bilayers and liposomal stability. *Journal of liposome research* **18**, 159-173 (2008).

64 Permyakov, S. E. Differential scanning microcalorimetry of intrinsically disordered proteins. *Intrinsically Disordered Protein Analysis: Volume 2, Methods and Experimental Tools*, 283-296 (2012).

65 Zhang, Y. & Ardejani, M. S. *Differential Scanning Calorimetry to Quantify the Stability of Protein Cages*. Vol. 1252 (2015).

66 Georgescu, R. E., Garcia-Mira, M. M., Tasayco, M. L. & Sanchez-Ruiz, J. M. Heat capacity analysis of oxidized Escherichia coli thioredoxin fragments (1-73, 74-108) and their noncovalent complex: Evidence for the burial of apolar surface in protein unfolded states. *European Journal of Biochemistry* **268**, 1477-1485 (2001).

67 Guzman-Casado, M., Parody-Morreale, A., Robic, S., Marqusee, S. & Sanchez-Ruiz, J. M. Energetic evidence for formation of a pH-dependent hydrophobic cluster in the denatured state of Thermus thermophilus ribonuclease H. *Journal of Molecular Biology* **329**, 731-743 (2003).

68 Perez-Jimenez, R., Godoy-Ruiz, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. The efficiency of different salts to screen charge interactions in proteins: a Hofmeister effect? *Biophys. J.* **86**, 2414-2429 (2004).

69 Freire, E. Differential scanning calorimetry. *Protein Stability and Folding: Theory and Practice*, 191-218 (1995).

70 Godoy-Ruiz, R. *et al.* Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data. *Journal of Physical Chemistry B* **112**, 5938-5949, doi:10.1021/jp0757715 (2008).

71 Bloch, F. Nuclear induction. *Physical Review* **70**, 460 (1946).

72 Purcell, E. M., Torrey, H. & Pound, R. V. Resonance absorption by nuclear magnetic moments in a solid. *Physical Review* **69**, 37 (1946).

73 Ernst, R. & Anderson, W. Application of Fourier transform spectroscopy to magnetic resonance. *Review of Scientific Instruments* **37**, 93-102 (1966).

74 Aue, W., Bartholdi, E. & Ernst, R. R. Two-dimensional spectroscopy. Application to nuclear magnetic resonance. *The Journal of Chemical Physics* **64**, 2229-2246 (1976).

75 Roberts, G. C. Applications of NMR in drug discovery. *Drug Discovery Today* **5**, 230-240 (2000).

76 van Beusekom, B. Applications of NMR in Drug Discovery. (2014).

77 Cavanagh, J., Fairbrother, W. J., Palmer Iii, A. G., Rance, M. & Skelton, N. J. in *Protein NMR Spectroscopy (Second Edition)* 1-28 (Academic Press, 2007).

78 Edelman, R. R. & Warach, S. Magnetic resonance imaging. *New England Journal of Medicine* **328**, 708-716 (1993).

79 Berger, A. How does it work?: Magnetic resonance imaging. *BMJ: British Medical Journal* **324**, 35 (2002).

80 Tiers, G. & Coon, R. Notes-Preparation of Sodium 2, 2-Dimethyl-2-silapentane-5-sulfonate, a Useful Internal Reference for NSR Spectroscopy in Aqueous and Ionic Solutions. *The Journal of Organic Chemistry* **26**, 2097-2098 (1961).

81 Wishart, D. S., Sykes, B. D. & Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of Molecular Biology* **222**, 311-333 (1991).

82 Markley, J. L. Two-dimensional nuclear magnetic resonance spectroscopy of proteins: An overview. *Methods in enzymology* **176**, 12-64 (1989).

83 Wishart, D., Sykes, B. & Richards, F. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647-1651 (1992).

84 Cornilescu, G., Delaglio, F. & Bax, A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR* **13**, 289-302 (1999).

85 Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **44**, 213-223, doi:10.1007/s10858-009-9333-z (2009).

86 Rule, G. S. & Hitchens, T. K. *Fundamentals of protein NMR spectroscopy*. Vol. 5 (Springer Science & Business Media, 2006).

87 Karplus, M. Contact electron-spin coupling of nuclear magnetic moments. *The Journal of Chemical Physics* **30**, 11-15 (1959).

88 Vuister, G. W. J-couplings. Measurement and Usage in Structure Determination.

89 de Alba, E. & Tjandra, N. NMR dipolar couplings for the structure determination of biopolymers in solution. *Progress In Nuclear Magnetic Resonance Spectroscopy* **40**, 175-197 (2002).

90 Bloembergen, N., Purcell, E. M. & Pound, R. V. Relaxation effects in nuclear magnetic resonance absorption. *Physical Review* **73**, 679 (1948).

91 Remerowski, M. *et al.* 1H, 13C and 15N NMR backbone assignments and secondary structure of the 269-residue protease subtilisin 309 from Bacillus lentus. *Journal of Biomolecular NMR* **4**, 257-278 (1994).

92 Cavanagh, J., Fairbrother, W. J., Palmer III, A. G. & Skelton, N. J. *Protein NMR spectroscopy: principles and practice*. (Academic Press, 1995).

93 Burum, D. & Ernst, R. Net polarization transfer via a J-ordered state for signal enhancement of low-sensitivity nuclei. *Journal of Magnetic Resonance (1969)* **39**, 163-168 (1980).

94 Robin Bendall, M. & Pegg, D. T. Complete accurate editing of decoupled 13C spectra using DEPT and a quaternary-only sequence. *Journal of Magnetic Resonance (1969)* **53**, 272-296, doi:http://dx.doi.org/10.1016/0022-2364(83)90032-X (1983).

95 Braunschweiler, L. & Ernst, R. Coherence transfer by isotropic mixing: application to proton correlation spectroscopy. *Journal of Magnetic Resonance (1969)* **53**, 521-528 (1983).

96 Fischer, M. W., Zeng, L. & Zuiderweg, E. R. Use of 13C-13C NOE for the Assignment of NMR Lines of Larger Labeled Proteins at Larger Magnetic Fields. *J. Am. Chem. Soc.* **118**, 12457-12458 (1996).

97 Kövér, K. E. & Batta, G. Theoretical and practical aspects of one-and two-dimensional heteronuclear overhauser experiments and selective 13C T1-determinations of heteronuclear distances. *Progress In Nuclear Magnetic Resonance Spectroscopy* **19**, 223-266 (1987).

98 Bodenhausen, G. & Ruben, D. J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chemical Physics Letters* **69**, 185-189 (1980).

99 Skrynnikov, N. R., Dahlquist, F. W. & Kay, L. E. Reconstructing NMR Spectra of "Invisible" Excited Protein States Using HSQC and HMQC Experiments. *J. Am. Chem. Soc.* **124**, 12352-12360, doi:10.1021/ja0207089 (2002).

100 Ernst, R. R., Bodenhausen, G. & Wokaun, A. *Principles of nuclear magnetic resonance in one and two dimensions*. Vol. 14 (Clarendon Press Oxford, 1987).

101 Shaka, A., Barker, P. & Freeman, R. Computer-optimized decoupling scheme for wideband applications and low-level operation. *Journal of Magnetic Resonance (1969)* **64**, 547-552 (1985).

102 Shaka, A., Keeler, J. & Freeman, R. Evaluation of a new broadband decoupling sequence: WALTZ-16. *Journal of Magnetic Resonance (1969)* **53**, 313-340 (1983).

103 Fesik, S. W. & Zuiderweg, E. R. Heteronuclear three-dimensional NMR spectroscopy. A strategy for the simplification of homonuclear two-dimensional NMR spectra. *Journal of Magnetic Resonance (1969)* **78**, 588-593 (1988).

104 Clore, G. M. & Gronenborn, A. M. Determination of Three-Dimensional Structures of Proteins and Nucleic Acids in Solution by Nuclear Magnetic Resonance Spectroscop. *Critical reviews in biochemistry and molecular biology* **24**, 479-564 (1989).

105 Grzesiek, S. & Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **114**, 6291-6293, doi:10.1021/ja00042a003 (1992).

106 Grzesiek, S. & Bax, A. Amino acid type determination in the sequential assignment procedure of uniformly 13 C/15 N-enriched proteins. *Journal of Biomolecular NMR* **3**, 185-204 (1993).

107 Grzesiek, S. & Bax, A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **99**, 201-207 (1992).

108 Wittekind, M. & Mueller, L. HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha-and beta-carbon resonances in proteins. *Journal of Magnetic Resonance, Series B* **101**, 201-205 (1993).

109 Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **89**, 496-514 (1990).

110 Grzesiek, S. & Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *Journal of Magnetic Resonance (1969)* **96**, 432-440 (1992).

111 Muhandiram, D. & Kay, L. E. Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. *Journal of Magnetic Resonance, Series B* **103**, 203-216 (1994).

112 Bax, A. *et al.* Practical aspects of proton-carbon-carbon-proton three-dimensional correlation spectroscopy of 13C-labeled proteins. *Journal of Magnetic Resonance (1969)* **87**, 620-627 (1990).

113 Kay, L. E., Ikura, M. & Bax, A. Proton-proton correlation via carbon-carbon couplings: A three-dimensional NMR approach for the assignment of aliphatic resonances in proteins labeled with carbon-13. *J. Am. Chem. Soc.* **112**, 888-889 (1990).

114 Ikura, M., Kay, L. E. & Bax, A. Improved three-dimensional 1H-13C-1H correlation spectroscopy of a 13C-labeled protein using constant-time evolution. *Journal of Biomolecular NMR* **1**, 299-304 (1991).

115 Bax, A., Clore, G. M. & Gronenborn, A. M. 1H-1H correlation via isotropic mixing of 13C magnetization, a new three-dimensional approach for assigning 1H and 13C spectra of 13C-enriched proteins. *Journal of Magnetic Resonance (1969)* **88**, 425-431 (1990).

116 Olejniczak, E. T., Xu, R. X. & Fesik, S. W. A 4D HCCH-TOCSY experiment for assigning the side chain 1 H and 13 C resonances of proteins. *Journal of Biomolecular NMR* **2**, 655-659 (1992).

117 Grzesiek, S., Anglister, J. & Bax, A. Correlation of backbone amide and aliphatic side-chain resonances in 13C/15N-enriched proteins by isotropic mixing of 13C magnetization. *Journal of Magnetic Resonance, Series B* **101**, 114-119 (1993).

118 Montelione, G. T., Lyons, B. A., Emerson, S. D. & Tashiro, M. An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. *J. Am. Chem. Soc.* **114**, 10974-10975 (1992).

119 Cavanagh, J., Fairbrother, W. J., Palmer Iii, A. G., Rance, M. & Skelton, N. J. in *Protein NMR Spectroscopy (Second Edition)* 533-678 (Academic Press, 2007).

120 Bruch, M. D. Multidimensional NMR spectroscopy of liquids. *Practical Spectroscopy Series* **21**, 145-238 (1996).

121 Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution. *Progress In Nuclear Magnetic Resonance Spectroscopy* **34**, 93-158 (1999).

122 Parella, T. NMRGuide3. 5. *BRUKER Biospin: Billerica, MA, USA* (1998).

123 Kleckner, I. R. & Foster, M. P. An introduction to NMR-based approaches for measuring protein dynamics. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1814**, 942-968 (2011).

124 Korzhnev, D. M. *et al.* Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* **430**, 586-590 (2004).

125 Keeler, J. *Understanding NMR spectroscopy*. (John Wiley & Sons, 2013).

126 Kay, L. E., Torchia, D. A. & Bax, A. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28**, 8972-8979, doi:10.1021/bi00449a003 (1989).

127 Kay, L. E., Nicholson, L. K., Delaglio, F., Bax, A. & Torchia, D. Pulse sequences for removal of the effects of cross correlation between dipolar and chemical-shift anisotropy relaxation mechanisms on the measurement of heteronuclear T1 and T2 values in proteins. *Journal of Magnetic Resonance (1969)* **97**, 359-375 (1992).

128 Carr, H. Y. & Purcell, E. M. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. *Physical Review* **94**, 630-638 (1954).

129 Meiboom, S. & Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Review of Scientific Instruments* **29**, 688-691, doi:10.1063/1.1716296 (1958).

130 Loria, J. P., Rance, M. & Palmer, A. G. A relaxation-compensated Carr-Purcell-Meiboom-Gill sequence for characterizing chemical exchange by NMR spectroscopy. *J. Am. Chem. Soc.* **121**, 2331-2332 (1999).

131 Lundström, P., Vallurupalli, P., Hansen, D. F. & Kay, L. E. Isotope labeling methods for studies of excited protein states by relaxation dispersion NMR spectroscopy. *Nature Protocols* **4**, 1641-1648, doi:10.1038/nprot.2009.118 (2009).

132 Mulder, F. A., Mittermaier, A., Hon, B., Dahlquist, F. W. & Kay, L. E. Studying excited states of proteins by NMR spectroscopy. *Nature Structural & Molecular Biology* **8**, 932-935 (2001).

133 Bouvignies, G. *et al.* Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **477**, 111-U134, doi:10.1038/nature10349 (2011).

134 Sugase, K., Dyson, H. J. & Wright, P. E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**, 1021-U1011, doi:10.1038/nature05858 (2007).

135 Korzhnev, D. M., Bezsonova, I., Lee, S., Chalikian, T. V. & Kay, L. E. Alternate binding modes for a ubiquitin-SH3 domain interaction studied by NMR spectroscopy. *Journal of Molecular Biology* **386**, 391-405 (2009).

136 Schanda, P., Brutscher, B., Konrat, R. & Tollinger, M. Folding of the KIX domain: characterization of the equilibrium analog of a folding intermediate using 15 N/13 C relaxation dispersion and fast 1 H/2 H amide exchange NMR spectroscopy. *Journal of Molecular Biology* **380**, 726-741 (2008).

137 Neudecker, P., Lundström, P. & Kay, L. E. Relaxation Dispersion NMR Spectroscopy as a Tool for Detailed Studies of Protein Folding. *Biophys. J.* **96**, 2045-2054, doi:10.1016/j.bpj.2008.12.3907 (2009).

138 Korzhnev, D. M., Religa, T. L., Banachewicz, W., Fersht, A. R. & Kay, L. E. A Transient and Low-Populated Protein-Folding Intermediate at Atomic Resolution. *Science* **329**, 1312-1316, doi:10.1126/science.1191723 (2010).

139 Farber, P. J. & Mittermaier, A. Relaxation dispersion NMR spectroscopy for the study of protein allostery. *Biophysical Reviews* **7**, 191-200 (2015).

140 Eisenmesser, E. Z. *et al.* Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117-121 (2005).

141 Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638-1642 (2006).

142 Kovrigin, E. L. & Loria, J. P. Enzyme dynamics along the reaction coordinate: critical role of a conserved residue. *Biochemistry* **45**, 2636-2647 (2006).

143 Hansen, D. F., Vallurupalli, P., Lundström, P., Neudecker, P. & Kay, L. E. Probing chemical shifts of invisible states of proteins with relaxation dispersion NMR spectroscopy: How well can we do? *J. Am. Chem. Soc.* **130**, 2667-2675, doi:10.1021/ja078337p (2008).

144 Palmer, A., Kroenke, C. D. & Loria, J. P. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods in enzymology* **339**, 204-238 (2001).

145 Hansen, D. F., Vallurupalli, P. & Kay, L. E. Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. *Journal of Biomolecular NMR* **41**, 113-120 (2008).

146 Millet, O., Loria, J. P., Kroenke, C. D., Pons, M. & Palmer, A. G. The static magnetic field dependence of chemical exchange linebroadening defines the NMR chemical shift time scale. *J. Am. Chem. Soc.* **122**, 2867-2877 (2000).

147 Mittermaier, A. & Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* **312**, 224-228 (2006).

148 Mulder, F. A., Skrynnikov, N. R., Hon, B., Dahlquist, F. W. & Kay, L. E. Measurement of slow (μs-ms) time scale dynamics in protein side chains by 15N relaxation dispersion NMR spectroscopy: application to Asn and Gln residues in a cavity mutant of T4 lysozyme. *J. Am. Chem. Soc.* **123**, 967-975 (2001).

149 McConnell, H. M. Reaction rates by nuclear magnetic resonance. *The Journal of Chemical Physics* **28**, 430-431 (1958).

150 Vallurupalli, P. Chemical Exchange. *TIFR Mumbai Workshop Series on NMR and Related Topics* (2009).

151 Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**, 277-293, doi:10.1007/bf00197809 (1995).

152 Kneller, D. & Kuntz, I. in *Journal of Cellular Biochemistry.* 254-254 (WILEY-LISS DIV JOHN WILEY & SONS INC 605 THIRD AVE, NEW YORK, NY 10158-0012).

153 Hansen, D. F. http://www.biochem.ucl.ac.uk/hansen/fuda/.

154 Bouvignies, G. & Kay, L. E. Measurement of proton chemical shifts in invisible states of slowly exchanging protein systems by chemical exchange saturation transfer. *The Journal of Physical Chemistry B* **116**, 14311-14317 (2012).

155 Bouvignies, G. https://github.com/gbouvignies/chemex. (2012).

156 Charlier, C. *et al.* Structure and dynamics of an intrinsically disordered protein region that partially folds upon binding by chemical-exchange NMR. *J. Am. Chem. Soc.* (2017).

157 Vallurupalli, P., Bouvignies, G. & Kay, L. E. Increasing the exchange time-scale that can be probed by CPMG relaxation dispersion NMR. *The Journal of Physical Chemistry B* **115**, 14891-14900 (2011).

158 Hansen, D. F., Vallurupalli, P. & Kay, L. E. An improved 15N relaxation dispersion experiment for the measurement of millisecond time-scale dynamics in proteins. *The Journal of Physical Chemistry B* **112**, 5898-5904 (2008).

159 Ishima, R. & Torchia, D. Extending the range of amide proton relaxation dispersion experiments in proteins using a constant-time relaxation-compensated CPMG approach. *Journal of Biomolecular NMR* **25**, 243-248, doi:10.1023/a:1022851228405 (2003).

160 Vallurupalli, P., Hansen, D. F., Lundström, P. & Kay, L. E. CPMG relaxation dispersion NMR experiments measuring glycine 1Hα and 13Cα chemical

shifts in the 'invisible' excited states of proteins. *Journal of Biomolecular NMR* **45**, 45-55 (2009).

161 Lundström, P., Hansen, D. F. & Kay, L. E. Measurement of carbonyl chemical shifts of excited protein states by relaxation dispersion NMR spectroscopy: comparison between uniformly and selectively 13C labeled samples. *Journal of Biomolecular NMR* **42**, 35-47, doi:10.1007/s10858-008-9260-4 (2008).

162 Lundström, P., Vallurupalli, P., Religa, T. L., Dahlquist, F. W. & Kay, L. E. A single-quantum methyl 13C-relaxation dispersion experiment with improved sensitivity. *Journal of Biomolecular NMR* **38**, 79-88 (2007).

163 Orekhov, V. Y., Korzhnev, D. M. & Kay, L. E. Double- and Zero-Quantum NMR Relaxation Dispersion Experiments Sampling Millisecond Time Scale Dynamics in Proteins. *J. Am. Chem. Soc.* **126**, 1886-1891, doi:10.1021/ja038620y (2004).

164 Bouvignies, G. *et al.* A simple method for measuring signs of 1HN chemical shift differences between ground and excited protein states. *Journal of Biomolecular NMR* **47**, 135-141, doi:10.1007/s10858-010-9418-8 (2010).

165 Ammann, C., Meier, P. & Merbach, A. E. A simple multinuclear NMR thermometer. *Journal of Magnetic Resonance (1969)* **46**, 319-321 (1982).

166 Panavas, T., Sanders, C. & Butt, T. R. SUMO fusion technology for enhanced protein production in prokaryotic and eukaryotic expression systems. *SUMO protocols*, 303-317 (2009).

167 Yip, T.-T., Nakagawa, Y. & Porath, J. Evaluation of the interaction of peptides with Cu (II), Ni (II), and Zn (II) by high-performance immobilized metal ion affinity chromatography. *Analytical Biochemistry* **183**, 159-171 (1989).

168 Hutchens, T. W. & Yip, T.-T. Protein interactions with immobilized transition metal ions: Quantitative evaluations of variations in affinity and binding capacity. *Analytical Biochemistry* **191**, 160-168 (1990).

169 Lundström, P. *et al.* Fractional C-13 enrichment of isolated carbons using 1-C-13 - or 2-C-13 -glucose facilitates the accurate measurement of dynamics at backbone C-alpha and side-chain methyl positions in proteins. *Journal of Biomolecular NMR* **38**, 199-212, doi:10.1007/s10858-007-9158-6 (2007).

170 Macias, M. J. *et al.* Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature* **382**, 646-649 (1996).

171 Sudol, M. & Hunter, T. NeW wrinkles for an old domain. *Cell* **103**, 1001-1004 (2000).

172 Passani, L. A. *et al.* Huntingtin's WW domain partners in Huntington's disease post-mortem brain fulfill genetic criteria for direct involvement in Huntington's disease pathogenesis. *Human molecular genetics* **9**, 2175-2182, doi:10.1093/hmg/9.14.2175 (2000).

173 McLoughlin, D. M. & Miller, C. C. The FE65 proteins and Alzheimer's disease. *Journal of neuroscience research* **86**, 744-754 (2008).

174 Sudol, M. The WW Domain and its Proline-Rich Ligand in Alzheimer's Disease and Muscular Dystrophy. *Emerging therapeutic targets* **1**, 81-84 (1997).

175 Sudol, M., Sliwa, K. & Russo, T. Functions of WW domains in the nucleus. *FEBS Lett.* **490**, 190-195 (2001).

176 Williams, S. *et al.* Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* **35**, 691-697 (1996).

177 Thompson, P. A., Eaton, W. A. & Hofrichter, J. Laser temperature jump study of the helix⇌coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. *Biochemistry* **36**, 9200-9210 (1997).

178 Munoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. Folding dynamics and mechanism of beta-hairpin formation. *Nature* **390**, 196-199 (1997).

179 Crane, J. C., Koepf, E. K., Kelly, J. W. & Gruebele, M. Mapping the transition state of the WW domain β-sheet. *Journal of Molecular Biology* **298**, 283-292 (2000).

180 Kubelka, J., Hofrichter, J. & Eaton, W. A. The protein folding 'speed limit'. *Current Opinion in Structural Biology* **14**, 76-88, doi:10.1016/j.sbi.2004.01.013 (2004).

181 Nguyen, H., Jäger, M., Moretto, A., Gruebele, M. & Kelly, J. W. Tuning the free-energy landscape of a WW domain by temperature, mutation, and truncation. *Proceedings of the National Academy of Sciences* **100**, 3948-3953 (2003).

182 Piana, S. *et al.* Computational design and experimental testing of the fastest-folding β-sheet protein. *Journal of Molecular Biology* **405**, 43-48 (2011).

183 Jäger, M. *et al.* Structure-function-folding relationship in a WW domain. *Proceedings of the National Academy of Sciences* **103**, 10648-10653 (2006).

184 Jäger, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The folding mechanism of a beta-sheet: the WW domain. *Journal of Molecular Biology* **311**, 373-393 (2001).

185 Yeh, S.-R., Ropson, I. J. & Rousseau, D. L. Hierarchical folding of intestinal fatty acid binding protein. *Biochemistry* **40**, 4205-4210 (2001).

186 Gopi, S. *et al.* Toward a quantitative description of microscopic pathway heterogeneity in protein folding. *Physical Chemistry Chemical Physics* **19**, 20891-20903 (2017).

187 Fotia, A. B. *et al.* The role of individual Nedd4-2 (KIAA0439) WW domains in binding and regulating epithelial sodium channels. *Faseb Journal* **16**, 70-+, doi:10.1096/fj.02-0497fje (2002).

188 Staub, O. & Rotin, D. Role of ubiquitylation in cellular membrane transport. *Physiological Reviews* **86**, 669-707 (2006).

189 Cobos, E. S. *et al.* Thermodynamic characterization of the folding equilibrium of the human Nedd4-WW4 domain: at the frontiers of cooperative folding. *Biochemistry* **48**, 8712-8720 (2009).

190 Macias, M. J., Gervais, V., Civera, C. & Oschkinat, H. Structural analysis of WW domains and design of a WW prototype. *Nature Structural & Molecular Biology* **7**, 375-379 (2000).

191 Woody, R. W. Contributions of tryptophan side chains to the far-ultraviolet circular dichroism of proteins. *European biophysics journal* **23**, 253-262 (1994).

192 Koepf, E. K., Petrassi, H. M., Sudol, M. & Kelly, J. W. WW: An isolated three-stranded antiparallel β-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Science* **8**, 841-853 (1999).

193 Krimm, S. & Bandekar, J. in *Advances in protein chemistry* Vol. 38 181-364 (Elsevier, 1986).

194 Bobby, R. *et al.* Structure and dynamics of human Nedd4-1 WW3 in complex with the αENaC PY motif. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1834**, 1632-1641 (2013).

195 Pucheta-Martinez, E. *et al.* Changes in the folding landscape of the WW domain provide a molecular mechanism for an inherited genetic syndrome. *Scientific reports* **6** (2016).

196 Dragan, A. I. & Privalov, P. L. Unfolding of a leucine zipper is not a simple two-state transition. *Journal of Molecular Biology* **321**, 891-908, doi:10.1016/S0022-2836(02)00699-X (2002).

197 Naganathan, A. N., Perez-Jimenez, R., Sanchez-Ruiz, J. M. & Munoz, V. Robustness of downhill folding: Guidelines for the analysis of equilibrium folding experiments on small proteins. *Biochemistry* **44**, 7435-7449, doi:10.1021/bi050118y (2005).

198 Panwalkar, V. *et al.* The Nedd4-1 WW Domain Recognizes the PY Motif Peptide through Coupled Folding and Binding Equilibria. *Biochemistry* **55**, 659-674, doi:10.1021/acs.biochem.5b01028 (2016).

199 Naganathan, A. N. & Muñoz, V. Determining denaturation midpoints in multiprobe equilibrium protein folding experiments. *Biochemistry* **47**, 6752-6761 (2008).

200 Harrison, S. C. A structural taxonomy of DNA-binding domains. *Nature* **353**, 715-719 (1991).

201 Pabo, C. O. & Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry* **61**, 1053-1095 (1992).

202 Liepinsh, E. L., A.; Sharipo, A.; Guignard, L.; Otting, G. Solution Structure of the R3H Domain from Human S[mu]bp-2. *Journal of Molecular Biology* **326**, doi:10.1016/S0022-2836(02)01381-5 (2003).

203 Grishin, N. V. The R3H motif: A domain that binds single-stranded nucleic acids. *Trends in Biochemical Sciences* **23**, 329-330 (1998).

204 De Sancho, D. & Muñoz, V. Integrated prediction of protein folding and unfolding rates from only size and structural class. *Physical Chemistry Chemical Physics* **13**, 17030-17043 (2011).

205 Gomez, J., Hilser, V. J., Xie, D. & Freire, E. The heat capacity of proteins. *Proteins: Structure, Function, and Bioinformatics* **22**, 404-412 (1995).

206 Cooper, A. Thermodynamic fluctuations in protein molecules. *Proceedings of the National Academy of Sciences* **73**, 2740-2741 (1976).

207 Auton, M., Holthauzen, L. M. F. & Bolen, D. W. Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proceedings of the National Academy of Sciences* **104**, 15317-15322 (2007).

208 Munoz, V. & Eaton, W. A. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 11311-11316 (1999).

209 Bruscolini, P. & Naganathan, A. N. Quantitative prediction of protein folding behaviors from a simple statistical model. *J. Am. Chem. Soc.* **133**, 5372-5379 (2011).

210 Kubelka, J., Henry, E. R., Cellmer, T., Hofrichter, J. & Eaton, W. A. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proceedings of the National Academy of Sciences* **105**, 18655-18662, doi:10.1073/pnas.0808600105 (2008).

211 Wani, A. H. & Udgaonkar, J. B. Native state dynamics drive the unfolding of the SH3 domain of PI3 kinase at high denaturant concentration. *Proceedings of the National Academy of Sciences*, pnas. 0908617106 (2009).

212 Bretscher, A. & Weber, K. Villin is a major protein of the microvillus cystoskeleton which binds both G and F actin in a calcium-dependent manner. *Cell* **20**, 839-847 (1980).

213 Khurana, S. & George, S. P. Regulation of cell structure and function by actin-binding proteins: villin's perspective. *FEBS Lett.* **582**, 2128-2139 (2008).

214 Hampton, C. M., Liu, J., Taylor, D. W., DeRosier, D. J. & Taylor, K. A. The 3D structure of villin as an unusual F-Actin crosslinker. *Structure* **16**, 1882-1891 (2008).

215 McKnight, C. J., Doering, D. S., Matsudaira, P. T. & Kim, P. S. A thermostable 35-residue subdomain within villin headpiece. *Journal of Molecular Biology* **260**, 126-134 (1996).

216 Frank, B. S., Vardar, D., Buckley, D. A. & McKnight, C. J. The role of aromatic residues in the hydrophobic core of the villin headpiece subdomain. *Protein Science* **11**, 680-687, doi:10.1110/ps.22202 (2002).

217 Xiao, S., Bi, Y., Shan, B. & Raleigh, D. P. Analysis of core packing in a cooperatively folded miniature protein: the ultrafast folding villin headpiece helical subdomain. *Biochemistry* **48**, 4607-4616 (2009).

218 Wang, M. *et al.* Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J. Am. Chem. Soc.* **125**, 6032-6033 (2003).

219 Mittal, J. & Best, R. B. Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water. *Biophys. J.* **99**, L26-L28, doi:10.1016/j.bpj.2010.05.005 (2010).

220 Saladino, G., Marenchino, M. & Gervasio, F. L. Bridging the Gap between Folding Simulations and Experiments: The Case of the Villin Headpiece. *Journal of chemical theory and computation* **7**, 2675-2680, doi:10.1021/ct2002489 (2011).

221 Lv, C. *et al.* Low Folding Cooperativity of Hp35 Revealed by Single-Molecule Force Spectroscopy and Molecular Dynamics Simulation. *Biophys. J.* **102**, 1944-1951, doi:10.1016/j.bpj.2012.03.028 (2012).

222 Tusell, J. R. & Callis, P. R. Simulations of tryptophan fluorescence dynamics during folding of the villin headpiece. *The Journal of Physical Chemistry B* **116**, 2586-2594 (2012).

223 Harada, R. & Kitao, A. The Fast-Folding Mechanism of Villin Headpiece Subdomain Studied by Multiscale Distributed Computing. *Journal of chemical theory and computation* **8**, 290-299, doi:10.1021/ct200363h (2012).

224 Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences* **110**, 17874-17879, doi:10.1073/pnas.1311599110 (2013).

225 Vermeulen, W. *et al.* Solution structures of the C-terminal headpiece subdomains of human villin and advillin, evaluation of headpiece F-actin-binding requirements. *Protein Science* **13**, 1276-1287, doi:10.1110/ps.03518104 (2004).

226 Chiu, T. K. *et al.* High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7517-7522, doi:10.1073/pnas.0502495102 (2005).

227 Tang, Y., Grey, M. J., McKnight, J., Palmer, A. G. & Raleigh, D. P. Multistate folding of the villin headpiece domain. *Journal of Molecular Biology* **355**, 1066-1077 (2006).

228 Lei, H., Chen, C., Xiao, Y. & Duan, Y. The protein folding network indicates that the ultrafast folding mutant of villin headpiece subdomain has a deeper folding funnel. *The Journal of Chemical Physics* **134**, 205104 (2011).

229 Havlin, R. H. & Tycko, R. Probing site-specific conformational distributions in protein folding with solid-state NMR. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3284-3289 (2005).

230 Hu, K.-N., Havlin, R. H., Yau, W.-M. & Tycko, R. Quantitative determination of site-specific conformational distributions in an unfolded protein by solid-state nuclear magnetic resonance. *Journal of Molecular Biology* **392**, 1055-1073 (2009).

231 Cellmer, T., Buscaglia, M., Henry, E. R., Hofrichter, J. & Eaton, W. A. Making connections between ultrafast protein folding kinetics and molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **108**, 6103-6108 (2011).

232 Jean K. Chung, M. C. T., and Michael D. Fayer. Dynamics of the folded and unfolded villin headpiece (HP35) measured with ultrafast 2D IR vibrational echo spectroscopy. *PNAS* **108**, 6, doi:10.1073/pnas.1100587108/-/DCSupplemental (2011).

233 Brewer, S. H. *et al.* Effect of modulating unfolded state structure on the folding kinetics of the villin headpiece subdomain. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16662-16667 (2005).

234 Tang, Y., Rigotti, D. J., Fairman, R. & Raleigh, D. P. Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain. *Biochemistry* **43**, 3264-3272 (2004).

235 Tang, Y., Goger, M. J. & Raleigh, D. P. NMR characterization of a peptide model provides evidence for significant structure in the unfolded state of the villin headpiece helical subdomain. *Biochemistry* **45**, 6940-6946 (2006).

236 Sanchez-Medina, C. *et al.* Probing the free energy landscape of the fast folding gpW protein by relaxation dispersion NMR. *J. Am. Chem. Soc.* (2014).

237 Murialdo, H., Xing, X., Tzamtzis, D., Haddad, A. & Gold, M. The product of the bacteriophage lambda W gene: purification and properties. *Biochemistry and cell biology* **81**, 307-315 (2003).

238 Maxwell KL, D. A., Murialdo H, Gold M. Thermodynamic and functional characterization of protein W from bacteriophage lambda. The three C-terminal residues are critical for activity. *J Biol Chem* **Jun 23;275(25)**, 18879-18886 (2000).

239 Sborgi, L., Verma, A., Munoz, V. & de Alba, E. Revisiting the NMR Structure of the Ultrafast Downhill Folding Protein gpW from Bacteriophage lambda. *PLoS One* **6**, doi:e2640910.1371/journal.pone.0026409 (2011).

240 Palmer, A. G. & Massi, F. Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy. *Chemical Reviews* **106**, 1700-1719 (2006).

241 Tamiola, K., Acar, B. & Mulder, F. A. A. Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **132**, 18000-18003, doi:10.1021/ja105656t (2010).

242 University of Wisconsin System. *Biological Magnetic Resonance Data Bank*, <www.bmrb.wisc.edu> (2014).

243 Desai, T. M., Cerminara, M., Sadqi, M. & Munoz, V. The Effect of Electrostatics on the Marginal Cooperativity of an Ultrafast Folding Protein. *Journal of Biological Chemistry* **285**, 34549-34556, doi:10.1074/jbc.M110.154021 (2010).

244 Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. Numerical recipes in C. *Cambridge University Press* **1**, 3 (1988).

245 Robertson, A. D. & Murphy, K. P. Protein structure and the energetics of protein stability. *Chemical Reviews* **97**, 1251-1268 (1997).

246 Naganathan, A. N., Sanchez-Ruiz, J. M. & Munoz, V. Direct measurement of barrier heights in protein folding. *J. Am. Chem. Soc.* **127**, 17970-17971 (2005).

247 Balusek, C. *et al.* in *Quantitative Models for Microscopic to Macroscopic Biological Macromolecules and Tissues* 1-20 (Springer, 2018).

248 Maxwell, K. L. *et al.* The solution structure of bacteriophage λ protein W, a small morphogenetic protein possessing a novel fold1. *Journal of Molecular Biology* **308**, 9-14 (2001).

249 Smith, M. P. & Feiss, M. Sequence analysis of the phage 21 genes for prohead assembly and head completion. *Gene* **126**, 1-7 (1993).

APPENDIX A:

T_m and ΔH Values from Fitting the Chemical Shift Unfolding Curves of Villin Variants to the Two-State Model

Chemical shifts fitted to a two-state model for V1 variant

		Tm (K)	Tm error (±)§	ΔH (kJ/mol)	ΔH error (±)§
Val 50					
	N	322,8	0,4	114	4
	HN	*	*	*	*
	Ca	322	2	93	11
	Ha	322,3	0,8	94	6
	Cb	*	*	*	*
	Hb	322,8	0,7	102	5
	Cg1	*	*	*	*
	Cg2	320	1	101	12
	Hg1	321,8	0,6	101	5
	Hg2	323	4	121	34
Gly 52					
	N	323,4	0,3	106	3
	HN	*	*	*	*
	Ca	319,6	0,7	111	8
	Ha1	322	3	115	38
	Ha2	325	8	89	36
Thr 54					
	N	322,4	0,4	102	3
	HN	323	3	98	26
	Ca	328	7	129	82
	Ha	326	2	116	25
	Cb	#	#	#	#
	Hb	#	#	#	#
	Cg	323	3	88	13
	Hg	323	2	109	16
Ala 57					
	N	325	1	154	23
	HN	322,7	0,6	99	5
	Ca	322,4	0,5	105	4
	Ha	322	4	89	20
	Cb	*	*	*	*
	Hb	324	3	101	21
Phe 58					
	N	323,3	0,4	96	3
	HN	323,4	0,9	98	6
	Ca	322,1	0,5	108	4
	Ha	323	1	105	7
	Cb	319,6	0,9	102	8
	Hb1	323	5	135	72
	Hb2	325	9	107	39
Leu 63					
	N	323,5	0,4	99	3
	HN	324	1	101	7
	Ca	322,8	0,7	102	5
	Ha	323	1	96	7
	Cb	323,0	0,9	109	9
	Hb	320	2	174	53
	Cg	*	*	*	*
	Hg	323	5	147	70
	Cd1	322	2	107	15
	Cd2	324,7	0,8	112	8
	Hd1	323	3	128	42
	Hd2	323	5	142	74

§ Fitting errors at 95% confidence (two standard deviations)
* Straight lines that do not fit to the two-state model
\# Water overlap

Chemical shifts fitted to a two-state model for V2 variant

		Tm (K)	Tm error (±)§	ΔH (kJ/mol)	ΔH error (±)§
Val 50					
	N	322,2	0,7	125	9
	HN	*	*	*	*
	Ca	320	1	116	14
	Ha	320	1	98	12
	Cb	*	*	*	*
	Hb	321	2	101	12
	Cg1	*	*	*	*
	Cg2	320	1	107	12
	Hg1	320	1	108	9
	Hg2	323	2	126	25
Phe 51					
	N	320,8	0,7	108	7
	HN	322	4	82	15
	Ca	321,1	0,7	97	6
	Ha	321	1	91	7
	Cb	322	2	108	13
	Hb1	320	1	106	11
	Hb2	318,4	0,9	124	12
Gly 52					
	N	322,9	0,8	108	8
	HN	*	*	*	*
	Ca	320	2	123	26
	Ha1	323	5	97	40
	Ha2	318	2	120	28
Ala 57					
	N	322	1	184	37
	HN	321,2	0,6	108	6
	Ca	322,1	0,8	106	7
	Ha	319	2	112	19
	Cb	*	*	*	*
	Hb	323	2	107	17
Leu 61					
	N	321,4	0,9	117	10
	HN	322,0	0,7	112	7
	Ca	*	*	*	*
	Ha	322	1	104	10
	Cb	318	3	100	24
	Hb1	321	2	101	12
	Hb2	*	*	*	*
	Cg	323	11	85	57
	Hg	324	4	106	30
	Cd1	322	2	94	10
	Cd2	322	1	111	12
	Hd1	321	1	100	11
	Hd2	322	1	111	11

§ Fitting errors at 95% confidence (two standard deviations)
* Straight lines that do not fit to the two-state model

Chemical shifts fitted to a two-state model for V3 variant

		Tm (K)	Tm error (±)§	ΔH (kJ/mol)	ΔH error (±)§
Ala 49					
	N	320,4	0,3	105	2
	HN	320,6	0,9	103	9
	Ca	319	1	100	9
	Ha	319	2	108	15
	Cb	316	6	124	95
	Hb	318	2	94	21
Gly 52					
	N	320,7	0,3	105	3
	HN	*	*	*	*
	Ca	318	1	121	15
	Ha1	316	4	107	50
	Ha2	328	11	68	30
Met 53					
	N	319,9	0,5	96	4
	HN	320,4	0,5	102	5
	Ca	#	#	#	#
	Ha	#	#	#	#
	Cb	320,1	0,7	107	6
	Hb1	318	3	97	30
	Hb2	321	2	101	15
	Cg	320	2	89	12
	Hg1	320	3	117	38
	Hg2	318	4	74	23
Phe 58					
	N	320,5	0,3	96	2
	HN	320,7	0,9	95	7
	Ca	321	1	92	8
	Ha	320	1	103	9
	Cb	317,8	0,8	112	10
	Hb1	321	5	91	37
	Hb2	320	7	154	99
Pro 62					
	Ca	316	2	119	28
	Ha	315	13	69	69
	Cb	318	1	96	10
	Hb1	320	1	98	11
	Hb2	321	2	99	15
	Cg	317	5	103	47
	Hg1	320	4	109	40
	Hg2	324	6	83	33
	Cd	320	2	115	21
	Hd1	320	1	100	8
	Hd2	*	*	*	*
Leu 69					
	N	321,2	0,3	99	2
	HN	320	1	97	8
	Ca	319,9	0,6	101	5
	Ha	323	8	107	48
	Cb	314	3	107	20
	Hb1	318	3	113	32
	Hb2	317	2	101	20
	Cg	320	2	178	60
	Hg	321	2	88	16
	Cd1	318,1	0,9	101	9
	Cd2	*	*	*	*
	Hd1	316	4	107	38
	Hd2	319	2	99	14

§ Fitting errors at 95% confidence (two standard deviations)
* Straight lines that do not fit to the two-state model
\# Water overlap

Chemical shifts fitted to a two-state model for V4 variant

		Tm (K)	Tm error (±)§	ΔH (kJ/mol)	ΔH error (±)§
Val 50					
	N	328	2	119	14
	HN	*	*	*	*
	Ca	326	1	98	7
	Ha	327	2	100	9
	Cb	*	*	*	*
	Hb	327	1	109	7
	Cg1	*	*	*	*
	Cg2	326	1	99	8
	Hg1	327	2	115	20
	Hg2	325	4	150	50
Ala 57					
	N	325	1	228	58
	HN	325,5	0,5	109	4
	Ca	326,9	0,7	104	4
	Ha	326	3	102	17
	Cb	*	*	*	*
	Hb	329	2	95	12
Leu 69					
	N	325,3	0,8	113	6
	HN	328	1	98	6
	Ca	325,7	0,7	105	5
	Ha	322	9	113	46
	Cb	314	4	84	24
	Hb1	327	2	101	20
	Hb2	324	2	109	17
	Cg	324	3	114	25
	Hg	330	3	88	12
	Cd1	325	2	106	12
	Cd2	*	*	*	*
	Hd1	326	3	89	22
	Hd2	328	2	95	9
Gly 74					
	N	324	1	121	11
	HN	325,8	0,6	105	5
	Ca	324	1	112	11
	Ha1	326	3	173	62
	Ha2	329	6	95	30

§ Fitting errors at 95% confidence (two standard deviations)
* Straight lines that do not fit to the two-state model

Chemical shifts fitted to a two-state model for V5 variant

		Tm (K)	Tm error (±)§	ΔH (kJ/mol)	ΔH error (±)§
Ser 43					
	N	318	2	96	13
	HN	324	4	79	20
	Ca	*	*	*	*
	Ha	*	*	*	*
	Cb	321	3	104	23
	Hb1	319	3	106	35
	Hb2	325	5	79	25
Asp 46					
	N	322,2	0,9	90	7
	HN	323	2	111	19
	Ca	323	3	170	62
	Ha	*	*	*	*
	Cb	286	10	159	110
	Hb	*	*	*	*
Val 50					
	N	320,8	0,9	134	13
	HN	*	*	*	*
	Ca	319	2	86	13
	Ha	322	3	87	16
	Cb	*	*	*	*
	Hb	322	2	189	50
	Cg1	*	*	*	*
	Cg2	322,2	0,8	107	7
	Hg1	321	2	106	18
	Hg2	325	5	178	120
Gln 66					
	N	322,4	0,7	103	5
	HN	324	1	106	9
	Ca	324,7	0,9	94	5
	Ha	325	1	98	7
	Cb	314	2	215	93
	Hb1	*	*	*	*
	Hb2	*	*	*	*
	Cg	320	3	138	55
	Hg	323,2	0,8	105	6
Lys 71					
	N	320,3	0,8	114	9
	HN	315	3	100	26
	Ca	319,9	0,7	121	9
	Ha	319	3	146	63
	Cb	317	3	98	29
	Hb	314	3	131	44
	Cg	*	*	*	*
	Hg1	*	*	*	*
	Hg2	*	*	*	*
	Cd	315	5	107	62
	Hd	311	7	92	58
	Ce	*	*	*	*
	He	*	*	*	*
Leu 75					
	N	324,4	0,9	120	10
	HN	323	8	118	91
	Ca	*	*	*	*
	Ha	319	3	172	92
	Cb	*	*	*	*
	Hb1	*	*	*	*
	Hb2	319	3	177	80
	Cg	*	*	*	*
	Hg	*	*	*	*
	Cd1	*	*	*	*
	Cd2	*	*	*	*
	Hd1	*	*	*	*
	Hd2	326	6	139	81

§ Fitting errors at 95% confidence (two standard deviations)
* Straight lines that do not fit to the two-state model

APPENDIX B: Published Paper

Published as:

Sanchez-Medina, C., Sekhar, A., Vallurupalli, P., Cerminara, M., Munoz, V. & Kay, L.E. Probing the free energy landscape of the fast folding gpW protein by relaxation dispersion NMR. *J. Am. Chem. Soc.* (2014). **URL:** https://doi.org/10.1021/ja502705y

Probing the free energy landscape of the fast folding gpW protein by relaxation dispersion NMR

Celia Sanchez-Medina[1], Ashok Sekhar[2], Pramodh Vallurupalli[2], Michele Cerminara[1], Victor Muñoz[1] and Lewis E. Kay[2,3,*]

[1]Centro de Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Darwin 3, Campus de Cantoblanco, 28049 Madrid, Spain.

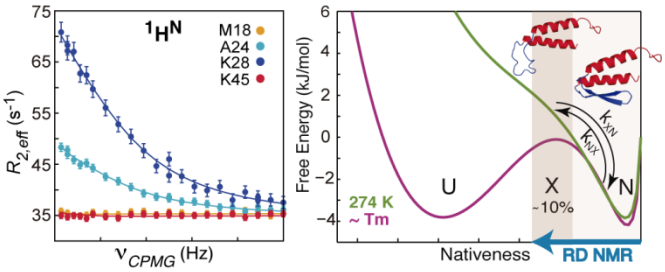
[2]Departments of Molecular Genetics, Biochemistry and Chemistry, The University of Toronto, Toronto, Ontario, Canada, M5S 1A8

[3]Program in Molecular Structure and Function, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, Canada M5G 1X8

*Address correspondence to Prof. Lewis E Kay, kay@pound.med.utoronto.ca

Abstract

The topographic features of the free energy landscapes that govern the thermodynamics and kinetics of conformational transitions in proteins, that in turn are integral for function, are not well understood. This reflects the experimental challenges with characterizing these multidimensional surfaces, even for small proteins. Here we focus on a 62 residue protein, gpW, that folds very rapidly into a native structure with an α/β topology in which α-helices are at the N- and C-terminal ends of the molecule, with a central β-hairpin positioned orthogonally to the helices. Using relaxation dispersion NMR spectroscopy to probe the conformational fluctuations in gpW at 1°C it is found that the native state interconverts with a transiently formed, sparsely populated second state with a lifetime of 250 μs, consistent with the global folding-unfolding rate under these conditions. In this low populated state the β-hairpin is unfolded, whereas the α-helices remain predominantly formed. Our results argue for a hierarchical stability of secondary structural elements and demonstrate the existence of a complex free energy landscape even in this small, fast folding single domain protein.



Introduction

Conformational fluctuations direct the folding of proteins into complex, functional three-dimensional structures, and often play important roles in molecular recognition and enzyme catalysis [1]. Conformational changes are usually described in terms of diffusive motions of protein molecules on a free energy surface of reduced dimensionality [2,3]. Minima on this landscape correspond to conformational ensembles whose relative free energies determine their populations, whereas barriers separating the states and the intra-molecular diffusion coefficient establish the rates of their interconversion. Experimental characterization of the conformational free energy surface has proven to be difficult because of its underlying complexity and several aspects including the magnitudes of the free energy barriers separating states, barrier transition paths, and the identification and structural elucidation of sparsely populated on and off-pathway folding intermediates remain challenging to experiment.

Native state hydrogen-deuterium exchange (HDX) experiments, in concert with NMR [4,5] or mass spectrometric detection [6], have provided fundamental insights into the free energy surface of proteins that fold slowly with complex kinetics [7], revealing the presence of subglobal cooperative folding units referred to as foldons [8]. The examination of kinetic folding pathways in a number of such proteins using HDX pulse labeling shows a sequence of events that follows the order of stability of individual foldons [9,10]. This observation has led to the suggestion that protein folding proceeds via the ordered assembly of foldons with the most stable foldons folding first [8]. Small single-domain proteins (less than approximately 100 residues), on the other hand, often fold without populating intermediates, indicating that their energy landscapes have simpler topography [11]. It has been traditionally assumed that small single-domain proteins fold in a cooperative two-state fashion [12] owing to a high free energy barrier separating their native and unfolded states. However, ultrafast kinetic methods have led to the discovery of many single-domain proteins that fold in only microseconds [13]. Such fast folding proteins approach the folding speed limit, which has been estimated to be around 1 μs by many procedures, including estimates from the kinetics of secondary structure formation [14], from the emergence of a second diffusive kinetic phase in some fast-folding proteins [15], from the scaling of folding rates with protein size [16], or more recently from the experimental study of folding transition path times [17]. For some fast-folding proteins the free energy barrier is so low that the conformers at the barrier top are populated [13], and thus their equilibrium unfolding appear complex and probe dependent, even though the underlying free energy landscape is essentially smooth [18]. Additionally, the recent implementation of sub-

millisecond molecular dynamics simulations has indicated that such small fast-folding proteins also assemble native structure in a stepwise manner [19].

Fast-folding proteins are thus attractive targets for atomic-resolution experimental methods [13]. In this regard, relaxation dispersion NMR (RD NMR) has evolved into an important technique for the structural characterization of local minima in the protein free energy landscape [20-22]. While such rare states do not give rise to detectable cross-peaks in NMR spectra their presence causes broadening of resonances derived from the native state. In cases where the rates of interconversion between observable ('ground') and invisible ('conformationally excited') states are on the ms timescale, the magnitude of exchange broadening can be systematically altered via a Carr-Purcell-Meiboom-Gill (CPMG) pulse train consisting of chemical shift refocusing pulses applied at varying frequencies (ν_{CPMG}). The graph of the effective transverse relaxation rate ($R_{2,eff}$) of a particular nucleus as a function of ν_{CPMG} is referred to as the relaxation dispersion profile and can be fit to a model of conformational exchange to extract chemical shift differences of the nucleus between the native and invisible states. The chemical shifts so obtained have been used to solve the three-dimensional atomic resolution structures of invisible states of proteins by incorporating them into chemical shift-based structure prediction algorithms [23,24]. Structures of on-pathway folding intermediates of the FF module [25,26] and the Fyn SH3 domain [27], as well as a sparsely populated conformer of T4 lysozyme [28] and an Abp1p SH3/Ark peptide complex [29] have been obtained in this manner.

Here, we employ RD NMR methodology to explore the energy landscape of the 62 residue protein, gpW [30], from λ bacteriophage, that adopts an α+β fold in the native state with N- and C-terminal α helices and an orthogonal β hairpin (Fig. 1A) [31]. gpW is an ultrafast folding protein with a folding time constant of ~8 μs at its melting temperature (T_m = 67º C at pH 6.0)[30]. RD NMR experiments reveal a sparsely and transiently populated invisible state (referred to as *X* in what follows) in conformational exchange with native gpW. Chemical shifts of *X* obtained from the analysis of the RD NMR data show that the β hairpin is locally unfolded in *X* whereas the two α helices are intact, clearly establishing a hierarchy in the stability of structural units in this small protein. The detection and characterization of a low-lying thermally accessible partially unfolded state in gpW provides experimental evidence for the existence of a complex free energy landscape even for small proteins that fold very rapidly.

Results

Relaxation dispersion measurements on gpW

The gpW protein folds in an ultra-fast manner with a time constant of 8 μs at 67º C (pH 6.0). In order to make the folding timescale accessible to RD NMR, which is limited to systems with exchange rates, k_{ex}, in the 100 $s^{-1} \leq k_{ex} \leq$ 2-3000 s^{-1} range (see below), we lowered the temperature to 1º C, exploiting the steep temperature dependence of folding rates that is often observed for fast-folding proteins [32], including gpW. The pH was also reduced to 3.5 in order to destabilize the protein, in the hope that this would lead to increased populations of additional states along the landscape that would become amenable to study via the RD NMR approach. Notably, the structure of the native state of gpW remains unaffected by the low temperature and pH used in the present study. For example, structures of gpW at 21º C/pH 6.5 [31] and 20º C/pH 3.5 (Fig. 1A) are virtually identical (superposition shown in Fig. S1). Further, the pattern of amide ^{1}H and ^{15}N chemical shifts observed in ^{1}H-^{15}N HSQC spectra at 1º (Fig. 1B) are very similar to those obtained at 25º C, confirming that the available native state structures are good models for gpW at 1º C/pH 3.5. However, at the low temperature, resonances show considerable differential line broadening (Fig. 1B) and 3 out of the 61 expected cross-peaks in the ^{1}H-^{15}N HSQC spectrum are broadened beyond detection, signifying the presence of micro-millisecond conformational exchange.

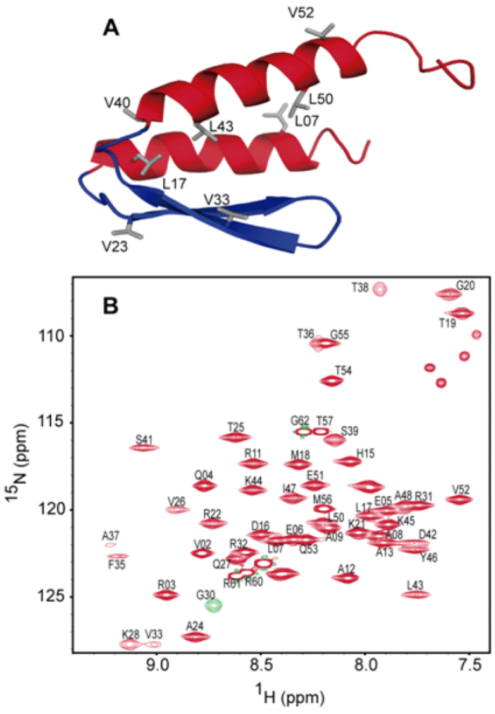


Figure 1. A) Ribbon diagram of the structure of gpW, as determined by solution NMR, 20ºC, pH 3.5 (PDB ID: 2L6R). The α-helical regions and terminal loops are colored red, the β-hairpin is in blue and selected Val and Leu side chains have been highlighted in grey. B) $^{1}H^{N}$-^{15}N HSQC spectrum of gpW, 1ºC, 11.7 T with assignment of resonances as indicated. The cross-peak shown in green is aliased.

In order to study the conformational exchange process in detail a series of RD NMR experiments were recorded, which provide structural information about thermally accessible conformational states that are not directly observable in NMR spectra, so long as they are in equilibrium with the native state. RD NMR measurements were performed using backbone ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, and $^{13}C^{\alpha}$ probes, to investigate changes in secondary structure that accompany the exchange event, and sidechain methyl ^{13}C nuclei to probe differences in hydrophobic packing between the native and invisible conformers. Well-resolved resonances belonging to 22 residues for ^{15}N, 32 for $^{1}H^{N}$, 24 for $^{13}C^{\alpha}$ and 25 for $^{13}C^{O}$ showed exchange contributions $R_{ex} = R_{2,eff}(\nu_{CPMG} = \nu_{min}) - R_{2,eff}(\nu_{CPMG} = \infty) > 5s^{-1}$, where ν_{min} is the minimum CPMG pulsing frequency for a particular experiment (see Materials and Methods). Non-zero values of R_{ex} confirm the presence of one or more excited states in exchange with the native state. Representative RD profiles for a number of spin-probes in the α helices (red and orange) and β hairpin (blue and cyan) are shown in Figure 2.

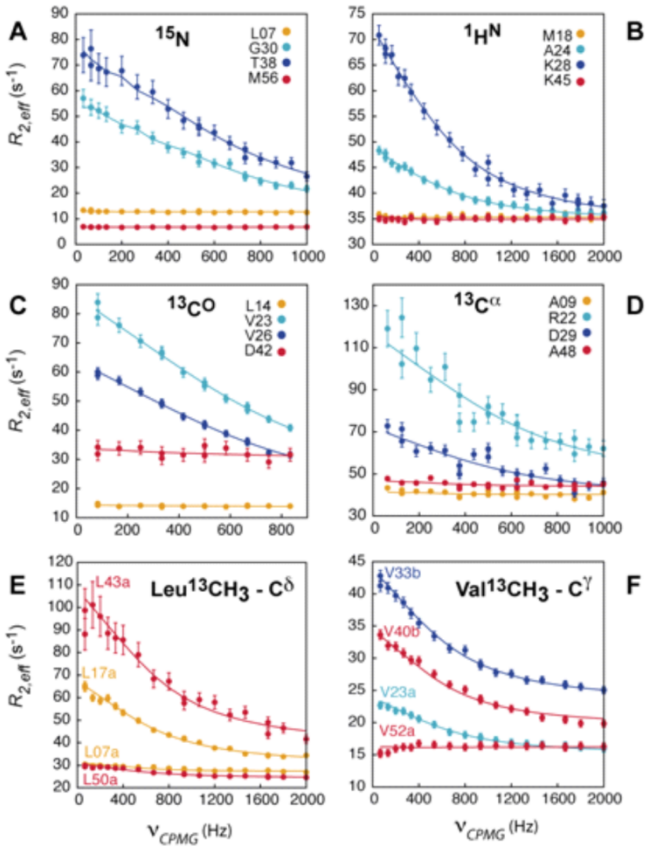


Figure 2. Relaxation dispersion profiles measured using A) ^{15}N, B) ^{1}H^{N}, C) ^{13}C^{O} and D) ^{13}C$^{\alpha}$ probes, recorded at 1° C, 11.7 T, showing the existence of sub-millisecond timescale conformational fluctuations in gpW. Solid lines are best-fits of the data (circles) to a two-state exchange model. Vertical lines associated with each data point denote experimental errors. In general, dispersions (R_{ex} values) are larger for the residues located in the hairpin (cyan and blue) than for the residues in the helices (orange and red). E, F) Methyl-^{13}C dispersion profiles from Leu and Val residues. Stereospecific assignments of the prochiral methyl groups have not been obtained; pairs of methyl groups from a given side-chain are distinguished 'a' and 'b', with 'a' corresponding to the methyl with the most upfield shifted resonance in the ^{13}C dimension.

Interestingly, backbone nuclei of residues in the β sheet show substantial exchange broadening while dispersion profiles of residues in the terminal α helices are relatively flat, consistent with a smaller contribution from exchange. Thus, the exchange process is predominantly associated with the hairpin region of the structure. The -CH_3 containing sidechains that are localized within helices (residues L17, L43, for example, Fig. 2E) and that form a part of the hydrophobic core are broadened considerably because of exchange. Thus, hydrophobic packing may be altered in the invisible state, potentially due to a shift in the register of the helices. The -CH_3 group of V33, which lies at the interface of the β-hairpin and

α-helices, also has a large R_{ex} value (Fig. 2F), either directly because of secondary structural changes in the β-hairpin or as a result of the relative motion between the helices and the hairpin. In addition, one of the methyl groups of residue V40, localized to the C-terminal helix at the helix-sheet interface, also has significant R_{ex} contributions (Fig. 2F).

Obtaining unique populations and extracting structural information

Relaxation dispersion profiles of ^{15}N, $^{1}H^{N}$ and methyl ^{13}C spins with R_{ex}> 5 s^{-1} at both 11.7 and 18.8 T were fit globally to the Bloch-McConnell equations [33] appropriate for a model of two-site exchange, $X \underset{k_{NX}}{\overset{k_{XN}}{\rightleftharpoons}} N$, to extract $k_{ex}= k_{XN} + k_{NX}$ = 4087 ± 42 s^{-1}. The presence of a pronounced minimum in the graph of χ^2_{red} as a function of k_{ex} (Fig. 3A) confirms the robustness of the fitted value of the exchange rate. This value of k_{ex} places the exchange process in the moderately fast regime whereby for all but one or two spins in gpW $k_{ex}>\Delta\omega$, with $\Delta\omega$ being the chemical shift difference, rad/s, between the exchanging sites. For example, α values [34], defining the chemical exchange time-scale, range between mean values of 1.8 to 1.9 (standard deviations of ±0.3) for the different nuclei probed, where a value of α = 2 indicates fast exchange. The moderately fast exchange process challenges the extraction of a robust fractional population of the *X* state, p_X, because in this case only the product $p_X(1-p_X)\Delta\omega^2$ can be obtained reliably from fits of RD data. The problem is illustrated in Fig. 3C that plots the reduced χ^2 from fits of dispersion profiles as a function of p_X. The flat χ^2 *vs* p_X curve indicates that the dispersion profiles can be fit equally well to virtually any value of p_X > 5%, with concomitant adjustment of the $\Delta\varpi$ (ppm) values to maintain the quality of the fit. Since structural information on *X* is obtained from shift differences, a reliable estimate of p_X becomes imperative.

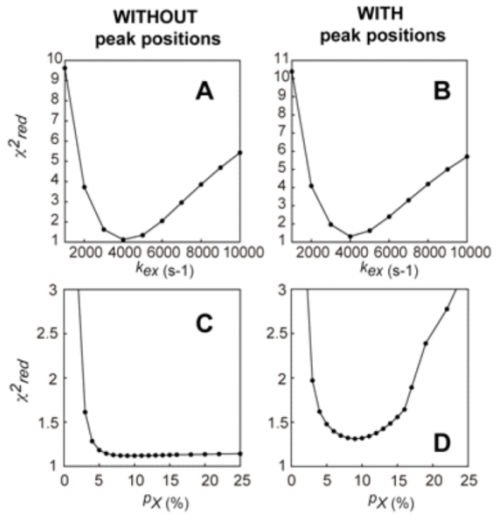


Figure 3. Variation of χ^2_{red} from fits of ^{15}N, $^1H^N$ and methyl ^{13}C dispersion data, 1° C, as a function of exchange parameters, k_{ex} (A,B) or p_X (C,D). Only RD CPMG data acquired at 11.7 and 18.8 T has been included in generating data in panels (A,C) while the static magnetic field dependence of native state peak positions (based on HSQC and HMQC spectra acquired at 11.7 and 18.8 T) has also been included in the analysis in panels (B,D). It is clear that while a unique k_{ex} value, corresponding to a pronounced minimum in the χ^2_{red} *vs* k_{ex} profile (A), can be obtained from fits of RD NMR data exclusively, a unique p_X can only be obtained if RD NMR data are supplemented with peak positions of the native state in HSQC/HMQC spectra recorded at a pair of static magnetic fields.

It has recently been shown that the correlation between p_X and $\Delta\omega$ can be broken for k_{ex} values as large as 4000-6000 s^{-1} by exploiting the fact that peak linewidths and positions depend in different ways on $k_{ex}/\Delta\omega$ and hence on the static magnetic field [35]. Thus, by simultaneously fitting relaxation dispersion profiles and the position of correlations in spectra recorded at a number of static magnetic fields it is possible to extend the exchange time-scale of systems that are amenable to study by RD NMR. Here we have recorded ^1H-^{15}N HSQC and HMQC spectra at 11.7 and 18.8 T and measured shifts in peak positions in HSQC/HMQC data sets and between HSQC data sets. Including these shifts along with ^{15}N, $^1H^N$ and ^{13}C-methyl dispersion profiles in fits breaks the correlation between p_X and $\Delta\varpi$, producing a distinct minimum in the reduced χ^2 *vs* p_X plot corresponding to $p_X = 9.0 \pm 0.3\%$ (Fig. 3D). It is worth noting that in the (moderately) fast exchange limit that is germane here, $\Delta\varpi$ scales as $p_X^{0.5}$. Thus, errors in p_X on the order of 30-40% lead to changes in chemical shift values of no more than 15-20%, that has a negligible effect on the structural insights that can be obtained (see below).

Structural features of the sparsely populated state

Once accurate exchange parameters were determined, ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and methyl ^{13}C chemical shifts of the invisible state were obtained by fitting RD profiles to a simple model of two-state exchange with k_{ex} and p_X fixed. Large ^{15}N chemical shifts (> 4 ppm) were observed for some backbone ^{15}N nuclei suggesting that the protein may be undergoing partial unfolding. To ascertain whether this is the case, differences in ^{15}N chemical shifts between the native and invisible states ($\Delta\varpi_{exp} = \Delta\varpi_{XN} = \varpi_X - \varpi_N$) have been plotted against the corresponding values that would be expected for a global folding-unfolding transition ($\Delta\varpi_{pred} = \Delta\varpi_{UN}$) (Fig. 4A). Intriguingly, the pattern obtained shows a distinct bimodal distribution, with chemical shift differences for residues in the β sheet correlating very well with $\Delta\varpi_{UN}$ values, while the analogous values for the α helices show a much poorer correlation. Other backbone nuclei (Fig. 4 panels B-D), in particular $^{13}C^{\alpha}$ and $^{13}C^{O}$ that are sensitive to secondary structure changes [36], mirror the bimodal distribution pattern observed for ^{15}N, confirming that the β sheet is indeed unfolded in the invisible state. The $^{13}C^{\alpha}$ and $^{13}C^{O}$ chemical shift data are further consistent with the helices remaining intact in *X*, as established by the secondary structure program, TALOS+ [37]. Because signs of $\Delta\varpi_{XN}$ values were not available for many of the ^{13}C probes, we have assumed the most extreme scenario in our analysis of the helical residues whereby in the absence of the sign, ϖ_X has been arbitrarily chosen to be closer to the value expected for a random-coil conformer, Fig. S2. Even with this assumption the helices are preserved in the excited state of gpW. Figure 5 graphs the residue-specific variation in backbone chemical shifts $\Delta\varpi_{RMS}$, showing that significant chemical shift changes (> 0.5 ppm) are localized to the central β sheet region ($<\Delta\varpi_{RMS}> = 0.7 \pm 0.3$), with $\Delta\varpi_{RMS}$ much smaller for the flanking helices($<\Delta\varpi_{RMS}> = 0.2 \pm 0.1$ and 0.2 ± 0.2 for helices 1 and 2, respectively). To summarize, the chemical shift data establish that the β hairpin is unfolded in *X*, but that the α helices remain intact.

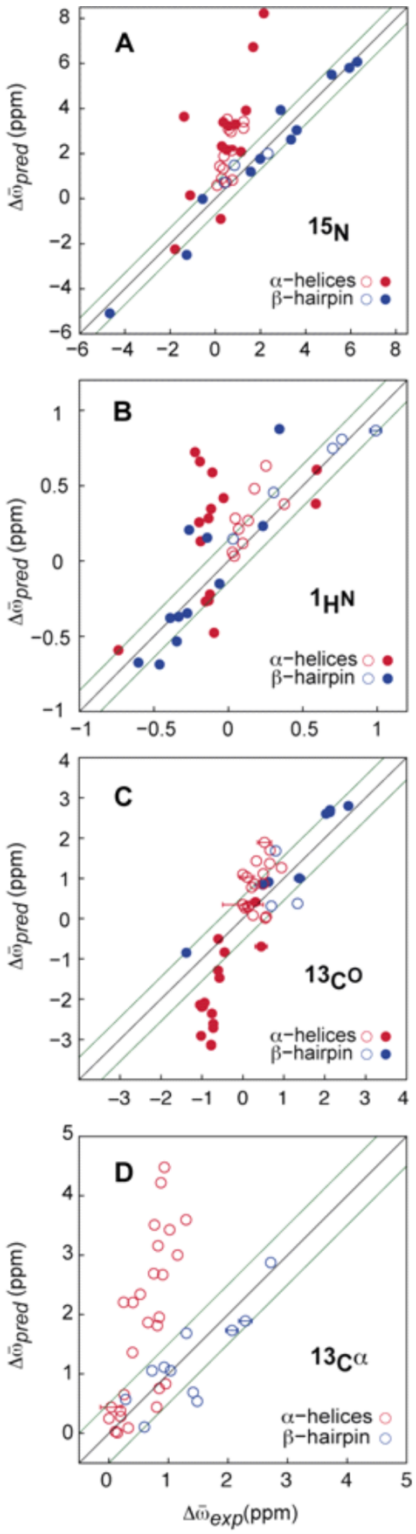


Figure 4. Comparison of experimentally obtained chemical shift differences between native and *X* states of gpW, 1°C, ($\Delta\varpi_{exp} = \varpi_I - \varpi_N$) for A) ^{15}N, B) ^{1}H^{N}, C) ^{13}C^{O} and D) ^{13}C$^{\alpha}$ spins with the corresponding values expected if *X* is in a random-coil conformation ($\Delta\varpi_{pred} = \varpi_U - \varpi_N$). The red (blue) circles denote values derived for residues belonging to helical (hairpin) regions. Values for which the signs of $\Delta\varpi$ have (have not) been obtained experimentally are denoted by filled (open) circles; for the cases where signs are not available, values of $|\Delta\varpi_{exp}|$ and $|\Delta\varpi_{pred}|$ are plotted. Error

bars smaller than the circle diameter have not been explicitly shown. The black line is y=x and the region between the green lines corresponds to the average chemical shift rmsd value for each of the displayed nuclei (from Tamiola *et al*. 2010 [61]).

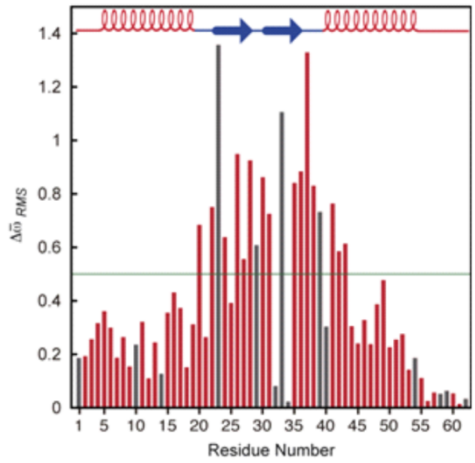


Figure 5. $\Delta\varpi_{RMS}$ (ppm) values plotted as a function of residue number. $\Delta\varpi_{RMS} = \sqrt{\frac{1}{N}\sum_i \left(\frac{\Delta\varpi_i}{\Delta\varpi_{i,STD}}\right)^2}$ where $\Delta\varpi_i$ is the difference between the shifts in the native and *X* states for a particular nucleus in a given residue, $\Delta\varpi_{i,STD}$ corresponds to 1 standard deviation of the distribution of chemical shifts for that nucleus in the context of the amino acid in question, obtained from a database of protein chemical shifts (BMRB, http://www.bmrb.wisc.edu/) and N is the number of nuclei included in the summation (≤4). The nuclei considered are ^{15}N, $^{1}H^{N}$, $^{13}C^{\alpha}$ and $^{13}C^{O}$. Values obtained when only two or fewer $\Delta\varpi$ values are included have been colored in grey. Some of these residues (A10, L14, E34, E49, Q58, R59) were overlapped in the $^{1}H^{N}$-^{15}N HSQC data sets, while others were too broad due to exchange, making it impossible to obtain reliable shift differences (V23, D29, V33, V40).

Discussion

The RD NMR approach is particularly powerful to characterize low populated protein conformational substates because of the large number of probes that can be used, so that detailed atomic resolution structural information can be obtained [38]. This approach has led to the determination of the structures of on-pathway folding intermediates in the native protein free energy landscape for the slower folding FF [25,26] and Fyn SH3 [27] domains (ms timescale). The FF domain folding intermediate is particularly interesting in that helix 3 is significantly longer than in the native state, leading to the establishment of non-native interactions across

the protein [25,26]. These alternative conformations produce local minima in the free energy surface from which the protein needs to escape before reaching the native structure.

Proteins that fold in a two-state manner have been a major focus of interest in folding studies via mutational analysis [39]. These proteins have, in principle, free energy landscapes devoid of local minima that could be probed by RD NMR. Work over the last 15 years on fast-folding, however, has described a new class of proteins that approach the folding speed limit [40] and thus have free energy surfaces with very low barriers and shallow minima [13]. The shallowness of ultrafast folding free energy surfaces allows for ensembles of partially folded conformations, which would correspond to excited states in nominally two-state folding, to become populated under certain experimental conditions [13]. These low populated partially-folded conformations can be detected by differential scanning calorimetry [41], giving the appearance of complex unfolding behaviour in equilibrium [18] and kinetics [42] experiments, even though the underlying energy landscapes are smooth. The special characteristics of fast-folding proteins have been dramatically demonstrated in recent long timescale atomistic simulations that achieved multiple folding and unfolding events for 12 such proteins [19]. The simulations identify a general theme whereby assembly of native structure for these proteins occurs in a stepwise fashion with key secondary structural elements forming at an early stage of folding that then template the formation of the rest of the secondary structure as well as the establishment of long range native contacts. Fast folding proteins thus seem to accrue structure in steps rather than in a globally cooperative way.

Experimentally, the coarse features of fast-folding free energy surfaces have been derived from fits of data from multiple probes to statistical mechanical models [30,42]. An ultrafast folding protein in which the topography of the free energy surface has been characterized in detail is the villin subdomain [43]. In other cases, such as the engrailed homeodomain (EnHD) [44] and Trp cage [45], similar experimental observations have been interpreted as evidence for the formation of an on-pathway intermediate. However, only in very few examples have these partially folded conformational ensembles been characterized in structural detail via experiment. One example is the 3-helix bundle EnHD intermediate where a mutant was studied that adopts the partly folded conformation at low ionic strength and folds to the native state under conditions of high ionic strength [46]. Motivated by our prior work on the slower folding FF and Fyn SH3 domains and by the fact that rich conformational behaviour has been observed for fast folding proteins by a variety of biophysical techniques [13] we wished to extend our RD NMR studies to explore the free energy landscapes of more rapidly folding molecules. Studies of fast folding proteins by RD NMR are technically challenging, however,

and this is the case for the gpW protein studied here. Folding of gpW has been investigated previously using a variety of different spectroscopic techniques, and a folding relaxation rate of 125,000 s^{-1} has been measured at 67°C [30]. Conformational exchange events on this timescale are well outside the window of RD NMR that is, in general, sensitive to processes that are more than an order of magnitude slower. However, exchange rates can be manipulated by changing temperature and for gpW at 1°C, pH 3.5 exchange was sufficiently slowed so that RD profiles with significant R_{ex} values could be recorded for ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and methyl ^{13}C nuclei, as illustrated in Fig. 2. These were fit well to a two-site exchange process with a rate of interconversion of 4100 s^{-1} (reduced $\chi^2 \sim 1$, Fig. 3), that remains relatively fast on the chemical shift timescale. In this exchange regime it is not possible to obtain accurate chemical shift differences between exchanging sites from analysis of dispersion profiles because of the coupling between p_X and $\Delta\omega$ values. This coupling can be broken, however, through simultaneous fits of both RD data and the chemical shifts of ground state resonances as a function of the static magnetic field, as illustrated in Fig. 3. A value of $p_X = 9\%$ was obtained through a combined analysis of this sort, along with chemical shifts that can be used to obtain structural information about state *X*. Notably, the excited state retains the terminal α helices of the native conformation of gpW, while the β hairpin that connects the helices is unfolded, Figs. 4,5. There is thus a hierarchical ordering of structural stability even in this small protein, with the β hairpin emerging as a subglobal autonomously folding structural unit or foldon.

The partially unfolded state characterized in this study is likely a folding intermediate, though there is no direct kinetic evidence from our NMR experiments because they were recorded under conditions where the population of the fully unfolded state is negligible. Instead, we remeasured the overall folding-unfolding relaxation kinetics of gpW by temperature-jump IR spectroscopy, but this time at pH 3.5, Fig. 6A. We could not reach 1° C directly with T-jump experiments because of the lack of kinetic amplitude under conditions that are so far from the folding midpoint. But, the quantitative analysis of the temperature dependent rate with a simple free energy surface (FES) model (see Supporting Information) produces an extrapolated rate constant of 5000±100 s^{-1}at 1° C, with the uncertainty in the extrapolated rateestimated using a jackknife analysis [47] (±100 s^{-1} is one standard deviation in the rates obtained from the analysis). The obtained value is in reasonable agreement with k_{ex} that we measure by RD NMR at this temperature, 4100±40 s^{-1}, especially considering that the extrapolation is over 20 °C and that $\Delta C_{p,res}$ in the FES model is fixed (see SI). The similarity

in rates suggests that the formation of the *X* state described here corresponds to the rate-limiting step in gpW unfolding. Using a chemical kinetics interpretation, that is most commonly used for RD NMR data, *X* would thus be an on-pathway intermediate placed between the unfolded state and the global transition state, corresponding to a local minimum in the free energy surface, Fig. 6B. The FES analysis of the T-jump data offers an alternative interpretation in which *X* corresponds to the ensemble of partly folded conformations placed at the top of the low-barrier transition observed by IR spectroscopy near the T_m (Fig. 6C). Under highly stabilizing conditions such as 1º C the FES of gpW is predicted to become downhill, but the ensemble ascribed to *X* remains the first excited state that is sampled from *N*, interconverting with nearly single exponential kinetics (inset in Fig 6A) and making up ~10% of the total population (inset to Fig. 6C), consistent with the RD NMR results. Unlike the model of Fig. 6B, *X* does not occupy a local minimum but rather is localized to the saddle point in the FES. In addition, the kinetics are governed essentially by the friction along the landscape, that in turn, is reflected by the temperature dependence of the diffusion on the surface (determined by the intramolecular diffusion coefficient, see Eq. S5 of SI). The RD NMR data does not distinguish between these two possible models, and as described above, is consistent with a two-state interconversion between *N* and *X* states.

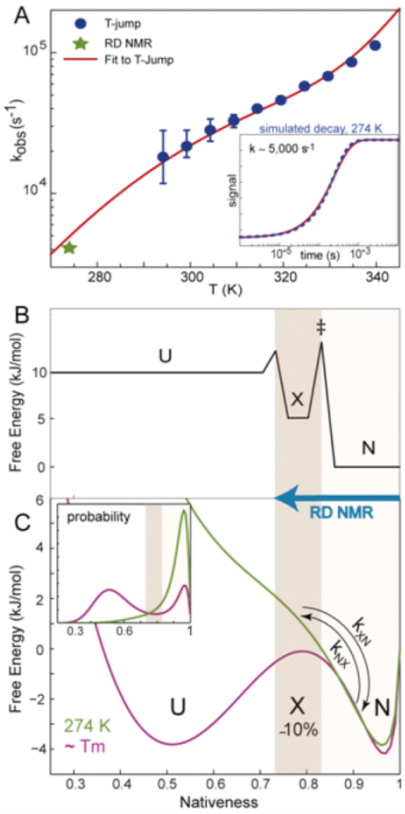


Figure 6. Mechanistic interpretation of the invisible state *X* in light of the comparison between RD NMR results and laser induced temperature-jump experiments. A) Folding-unfolding relaxation kinetics measured by infrared T-jump at pH 3.5 (blue circles) fitted to a 1D FES model (red curve; see Supporting Information), k_{ex} from RD NMR (green star), and the relaxation decay at 1° C as simulated by the FES model (blue in inset) together with a single exponential fit for reference (red in inset). B) Schematic representation of the FES in terms of a chemical kinetics model that separates *U*, *X* and *N* with barriers. C) The FES (main) and probability distribution (inset) obtained for gpW at the denaturation midpoint (magenta) and at 1° C (green) including the region populated by the invisible state *X*. It is noteworthy that the NMR data cannot distinguish between the models in (B) or (C).

Regardless of the preferred interpretation, *X* emerges in both cases as a productive intermediate occurring late in the kinetic folding pathway of gpW to its native conformation. Given that in *X* the α helices are fully formed while the β hairpin is unfolded, the formation of the β hairpin may thus represent the very last step in gpW folding. Interestingly, this conclusion is consistent with the empirical observation that spectroscopic probes sensitive to tertiary interactions between the β hairpin and the α helices report a lower stability than probes sensitive to the α helix structure alone [30]. Moreover, the lower intrinsic stability of the hairpin has also been observed in long-timescale MD simulations of gpW. Finally, we have performed ^{15}N RD measurements on a sample of gpW, 1°C, pH 6.5 to establish whether the intermediate that we have characterized in this work is likely to be important at neutral pH. Notably, extracted $|\Delta\varpi_{XN}|$ values are in good agreement with those obtained at pH 3.5, Fig. S3, providing strong evidence that the excited state remains intact over a broad spectrum of pH values.

In summary, using a combined analysis of RD profiles and the dependence of native state peak positions on the static magnetic field we have been able to probe the free energy landscape of the fast folding α/β gpW domain using solution NMR spectroscopy. We have identified and structurally characterized a thermally accessible sparsely formed, transiently populated state in equilibrium with the native state of gpW. The structure of this 'invisible' state clearly establishes the presence of a hierarchy in the stability of secondary structural elements and the existence of a complex folding process even for this small fast folding, 62-residue protein. gpW is considered to be an incipient downhill folding protein with a small (~ RT) free energy barrier at its T_m (67 ° C) [30]. A previous study has suggested that a combination of strong local interactions and loose hydrophobic packing might contribute to a lowering of the folding free energy barrier [30]. Our NMR characterization of the invisible state

now confirms the involvement of both of these factors in shaping the free energy surface of gpW. Strong local interactions ensure that the N- and C-terminal helices remain structured in the *X* state characterized here, while RD data for Leu and Val residues involved in tertiary interactions strongly suggest that the hydrophobic core is disrupted in the invisible state. The ability to extend RD NMR to faster exchange processes, as described here, increases the range of protein systems that can be studied via this methodology and opens up the exciting possibility of characterizing in greater detail than heretofore possible the free energy landscapes that dictate protein function.

Experimental Section

Protein Expression and Purification

A 62-residue gpW construct was chosen for the current study that does not include the C-terminal 6 residues that induce aggregation [30]. The T2V mutation, present in the construct used in previous structural studies of this protein [30,31], was retained. *E. coli* BL21(DE3) cells expressing gpW with a C-terminal histidine-tag were grown at 37ºC in minimal media. Protein expression was induced with IPTG (isopropyl-β-D-thiogalacto-pyranoside) when the cell culture reached an OD_{600} value of 0.7-0.8. After 4 hours of growth at 37ºC, the cells were harvested and centrifuged at 8000 rpm for 30 minutes. The pellet was resuspended in 20 mM Tris, 150 mM NaCl, 8 M urea, pH 8 until homogeneous. The lysate was ultracentrifuged at 18000 rpm for 45 minutes and the supernatant gently sonicated and loaded on a Ni affinity column followed by elution with an imidazole gradient. Fractions containing gpW were pooled, the imidazole subsequently dialyzed out, and the His-tag cleaved by incubation with Ulp1 protease for 1 hour at 37ºC. A second run through the Ni column was used to remove the cleaved His-tag. The protein was then concentrated with a 3kDa MWCO Centricon centrifugal concentrator for a last reverse phase (RP) HPLC step. The RP-HPLC uses a 0-95% water-acetonitrile gradient with 0.1% trifluoroacetic acid to remove salts and other impurities. Fractions containing pure gpW were pooled and lyophilized. The purity of the protein was checked by SDS-polyacrylamide gel electrophoresis and its identity verified by ESI mass spectrometry.

Sample Preparation for NMR Spectroscopy

[U-^{15}N] and selectively ($^{13}CH_3$) methyl labeled gpW, expressed in minimal media containing $^{15}NH_4Cl$ and [1-^{13}C]-glucose as the sole nitrogen and carbon sources respectively, [48] was used to record ^{15}N, $^{1}H^{N}$ and $^{13}CH_3$ RD experiments. [U-^{15}N,^{13}C] gpW was expressed in minimal

media containing $^{15}NH_4Cl$ and $^{13}C_6$-glucose, while [U-^{15}N] and selectively C^{α} labeled gpW was produced using minimal media with $^{15}NH_4Cl$ and [2-^{13}C]-glucose [48]. These samples were used for $^{13}C^{O}$ and $^{13}C^{\alpha}$ dispersion measurements, respectively. All NMR samples comprised approximately 1 mM protein in 20 mM glycine buffer (pH 3.5), 1 mM sodium azide, 10% D_2O/90% H_2O.

NMR Spectroscopy and Data Analysis

Backbone and sidechain chemical shift assignments were obtained using standard triple resonance experiments [49] recorded at either 11.7 or 14.1 T, 1 ºC.RD data were recorded at field strengths of 11.7 and 18.8 T, 1º C, with the temperature measured using a thermocouple inserted into an NMR tube. ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and methyl ^{13}C RD profiles were recorded with pulse schemes described previously [50-55]. For the measurement of $^{1}H^{N}$ dispersions a modified sequence was used whereby the amide magnetization at the start of the constant-time CPMG element is anti-phase with respect to ^{15}N. In addition, a (backbone amide specific) REBURP pulse of 2.00 ms (11.7 T) or 1.25 ms (18.8 T) duration was applied on the ^{15}N channel during the first INEPT transfer period in experiments recording ^{15}N and $^{1}H^{N}$ profiles to eliminate intense side-chain correlations arising from Arg residues. A pre-scan delay of 2.5 s was used in all cases, with constant-time CPMG relaxation delays of 30, 18, 24, 16 and 30 ms for measurement of ^{15}N, $^{1}H^{N}$, $^{13}C^{O}$, $^{13}C^{\alpha}$ and methyl ^{13}C RD profiles, respectively. Dispersion data were collected at 17 CPMG fields ranging from 33.3 to 1000 Hz for ^{15}N (11.7 and 18.8T), 21 CPMG fields from 55.6 to 2000 Hz for $^{1}H^{N}$ (11.7 and 18.8 T), 10 (12) CPMG fields from 83.3 to 833.3 (1000) Hz for $^{13}C^{O}$ at 11.7 T (18.8 T), 10 (13) CPMG fields from 62.5 to 1000 Hz for $^{13}C^{\alpha}$, 11.7 T (18.8 T), and 17 CPMG fields ranging from 66.7 to 2000 Hz for $^{13}CH_3$ (11.7 and 18.8 T). Errors in $R_{2,eff}$ were estimated by acquiring at least three duplicate points at selected ν_{CPMG} values.

Additional experiments were measured to determine the signs of the chemical shift differences, $\Delta\varpi$, obtained from fits of CPMG RD data. For ^{15}N $\Delta\varpi$ values, signs were obtained by comparing peak positions in $^{1}H^{N}$-^{15}N HSQC spectra acquired at 11.7 and 18.8 T and in $^{1}H^{N}$-^{15}N HSQC and HMQC spectra [56], also measured at 11.7 and at 18.8 T. Signs of $^{1}H^{N}$ shift differences were obtained from an analysis of $^{1}H^{N}/^{15}$N zero- and double-quantum coherence RD profiles [57] and also from a comparison of the direct dimension chemical shifts in HSQC spectra recorded at 11.7 and 18.8 T [58]. Signs for $^{13}C^{O}\Delta\varpi$ values were obtained from a comparison of resonance positions in $^{1}H^{N}$-$^{13}C^{O}$ data sets recorded at 11.7 and 18.8 T where the $^{13}C^{O}$ chemical shift arises either from evolution of single quantum or multiple quantum

($^{1}H^{N}$-$^{13}C^{O}$ or ^{15}N-$^{13}C^{O}$) coherence [59]. In total, 31, 26 and 25 signs were determined for 53, 52 and 47 measured ^{15}N, $^{1}H^{N}$ and $^{13}C^{O}\Delta\varpi$ values. It was not possible to measure signs for $^{13}C^{\alpha}$ shift differences because of the limited quality of $^{1}H^{\alpha}$ – $^{13}C^{\alpha}$ correlation maps.

NMR data were analyzed using the NMRPipe suite of programs [60],with peak intensities quantified with FuDA (http://pound.med.utoronto.ca/software.html), as previously described [61]. Spectra were visualized with NMRDraw [60] and Sparky [62]. Exchange parameters (p_i, k_{ex}, $\Delta\varpi$) were obtained by simultaneously fitting the dispersion profiles and HS/MQC data to a global two-site exchange model using the program CHEMEX (in-house written software, available upon request) that numerically propagates the Bloch-McConnell equations [33] in the simulation of the experimental data.

Infrared Temperature-Jump Kinetics

gpW samples for infrared kinetic measurements were prepared at 4 mg/ml protein concentration in 20 mM deuterated glycine buffer, after achieving complete deuteration of the exchangeable amide protons by performing multiple cycles of lyophilization and dilution in D_2O. The samples were adjusted to pD = 3.5 after correction for the isotope artifact on the readout of the glass electrode of the pH-meter. Infrared laser-induced temperature jump measurements were performed using a custom built apparatus [63]. Briefly, the fundamental wavelength of a Nd:YAG laser (Continuum Surelite I) operating at a repetition rate of 4 Hz is shifted to 1907 nm (a frequency at which D_2O exhibits strong absorption) by passing through a 1m path Raman cell (Light Age Inc.) filled with a high pressure mixture of H_2 and Ar. The output of the Raman cell is < 10 ns pulses of about 30 mJ which induce a local temperature jump of about 10-12 degrees when focused onto the sample. The response of the sample to this fast perturbation is monitored using a quantum cascade laser tunable in the range 1605-1690 cm^{-1} (Daylight solutions) to probe the amide I band of the protein under study. For gpW we tuned the laser to 1632 cm^{-1} to match the absorption maximum of the α-helix amide I band. The light transmitted from the sample was then detected using a fast MCT detector (Kolmar Technologies) coupled to an oscilloscope (Tektronix DPO4032). Samples were held in a cell formed by two MgF_2 windows separated by a 50μm teflon spacer and thermostated at the proper base temperature using two Peltier thermoelectric coolers (TE Technology Inc.) in a custom-built sample holder. The transmission of D_2O was used as an internal thermometer to measure the amplitude of the temperature jump.

Acknowledgements

C.S.M acknowledges support from the Spanish Ministry of Economy and Competitiveness in the form of a pre-doctoral fellowship (BES-2009-029117). A.S is the recipient of a postdoctoral fellowship from the Canadian Institutes of Health Research (CIHR). This work was funded through CIHR and NSERC research grants to L.E.K and grants CSD2009-00088 and BIO2011-28092 (Spanish Ministry of Economy and Competitiveness) and ERC-2012-ADG-323059 to V.M. L.E.K. holds a Canada Research Chair in Biochemistry.

Supporting Information

A description of the FES model is provided, along with one figure showing the superposition of NMR derived structures of gpW at pH 3.5 and 6.5, a second figure establishing that the α-helices are intact in the excited state of gpW and a third figure providing evidence that state *X* characterized at pH 3.5 is populated at neutral pH as well. This information is available free of charge via the internet at http://pubs.acs.org.

References

(1) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789-796.

(2) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598-1603.

(3) Lazaridis, T.; Karplus, M. *Biophys. Chem.* **2003**, *100*, 367-395.

(4) Bai, Y.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192-197.

(5) Chamberlain, A. K.; Handel, T. M.; Marqusee, S. *Nat. Struct. Mol. Biol.* **1996**, *3*, 782-787.

(6) Hu, W.; Walters, B. T.; Kan, Z.-Y.; Mayne, L.; Rosen, L. E.; Marqusee, S.; Englander, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 7684-7689.

(7) Bai, Y. *Chem. Rev.* **2006**, *106*, 1757-1768.

(8) Maity, H.; Maity, M.; Krishna, M. M.; Mayne, L.; Englander, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4741-4746.

(9) Hoang, L.; Bédard, S.; Krishna, M. M.; Lin, Y.; Englander, S. W. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12173-12178.

(10) Chamberlain, A. K.; Marqusee, S. *Adv. Protein Chem.* **2000**, *53*, 283-328.

(11) Bartlett, A. I.; Radford, S. E. *Nat. Struct. Mol. Biol.* **2009**, *16*, 582-588.

(12) Jackson, S. E. *Fold. Des.* **1998**, *3*, R81-R91.

(13) Muñoz, V. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 395-412.

(14) Eaton, W. A.; Muñoz, V.; Thompson, P. A.; Henry, E. R.; Hofrichter, J. *Acc. Chem. Res.* **1998**, *31*, 745-753.

(15) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193-197.

(16) Naganathan, A. N.; Muñoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 480-481.

(17) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A. *Science* **2012**, *335*, 981-984.

(18) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Munoz, V. *Science* **2002**, *298*, 2191-2195.

(19) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517-520.

(20) Palmer, A. G.; Kroenke, C. D.; Loria, J. P. *Methods Enzymol.* **2001**, *339*, 204-238.

(21) Mittermaier, A.; Kay, L. E. *Science* **2006**, *312*, 224-228.

(22) Palmer III, A. G.; Massi, F. *Chem. Rev.* **2006**, *106*, 1700-1719.

(23) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. *Structure* **2010**, *18*, 923-933.

(24) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 4685-4690.

(25) Korzhnev, D. M.; Religa, T. L.; Banachewicz, W.; Fersht, A. R.; Kay, L. E. *Science* **2010**, *329*, 1312-1316.

(26) Korzhnev, D. M.; Vernon, R. M.; Religa, T. L.; Hansen, A. L.; Baker, D.; Fersht, A. R.; Kay, L. E. *J. Am. Chem. Soc.* **2011**, *133*, 10974-10982.

(27) Neudecker, P.; Robustelli, P.; Cavalli, A.; Walsh, P.; Lundström, P.; Zarrine-Afsar, A.; Sharpe, S.; Vendruscolo, M.; Kay, L. E. *Science* **2012**, *336*, 362-366.

(28) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. *Nature* **2011**, *477*, 111-114.

(29) Vallurupalli, P.; Hansen, D. F.; Kay, L. E. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 11766-11771.

(30) Fung, A.; Li, P.; Godoy-Ruiz, R.; Sanchez-Ruiz, J. M.; Muñoz, V. *J. Am. Chem. Soc.* **2008**, *130*, 7489-7495.

(31) Sborgi, L.; Verma, A.; Muñoz, V.; de Alba, E. *PloS one* **2011**, *6*, e26409.

(32) Naganathan, A. N.; Doshi, U.; Muñoz, V. *J. Am. Chem. Soc.* **2007**, *129*, 5673-5682.

(33) McConnell, H. M. *J. Chem. Phys.* **1958**, *28*, 430-431.

(34) Millet, O.; Loria, J. P.; Kroenke, C. D.; Pons, M.; Palmer III, A. G. *J. Am. Chem. Soc.* **2000**, *122*, 2867-2877.

(35) Vallurupalli, P.; Bouvignies, G.; Kay, L. E. *J. Phys. Chem. B* **2011**, *115*, 14891-14900.

(36) Wishart, D. S.; Bigam, C. G.; Holm, A.; Hodges, R. S.; Sykes, B. D. *J. Biomol. NMR* **1995**, *5*, 67-81.

(37) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J. Biomol. NMR* **2009**, *44*, 213-223.

(38) Baldwin, A. J.; Kay, L. E. *Nat. Chem. Biol.* **2009**, *5*, 808-814.

(39) Naganathan, A. N.; Muñoz, V. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 8611-8616.

(40) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76-88.

(41) Naganathan, A. N.; Sanchez-Ruiz, J. M.; Muñoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 17970-17971.

(42) Li, P.; Oliva, F. Y.; Naganathan, A. N.; Muñoz, V. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 103-108.

(43) Kubelka, J.; Henry, E. R.; Cellmer, T.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 18655-18662.

(44) Religa, T. L.; Johnson, C. M.; Vu, D. M.; Brewer, S. H.; Dyer, R. B.; Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 9272-9277.

(45) Neuweiler, H.; Doose, S.; Sauer, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 16650-16655.

(46) Religa, T.; Markson, J.; Mayor, U.; Freund, S.; Fersht, A. *Nature* **2005**, *437*, 1053-1056.

(47) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical recipes in C*; Cambridge University Press: Cambridge, 1988.

(48) Lundström, P.; Teilum, K.; Carstensen, T.; Bezsonova, I.; Wiesner, S.; Hansen, D. F.; Religa, T. L.; Akke, M.; Kay, L. E. *J. Biomol. NMR* **2007**, *38*, 199-212.

(49) Sattler, M.; Schleucher, J.; Griesinger, C. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93-158.

(50) Mulder, F. A. A.; Skrynnikov, N. R.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *J. Am. Chem. Soc.* **2001**, *123*, 967-975.

(51) Hansen, D. F.; Vallurupalli, P.; Kay, L. E. *J. Phys. Chem. B* **2008**, *112*, 5898-5904.

(52) Ishima, R.; Torchia, D. A. *J. Biomol. NMR* **2003**, *25*, 243-248.

(53) Vallurupalli, P.; Hansen, D. F.; Lundström, P.; Kay, L. E. *J. Biomol. NMR* **2009**, *45*, 45-55.

(54) Lundström, P.; Hansen, D. F.; Kay, L. E. *J. Biomol. NMR* **2008**, *42*, 35-47.

(55) Lundström, P.; Vallurupalli, P.; Religa, T. L.; Dahlquist, F. W.; Kay, L. E. *J. Biomol. NMR* **2007**, *38*, 79-88.

(56) Skrynnikov, N. R.; Dahlquist, F. W.; Kay, L. E. *J. Am. Chem. Soc.* **2002**, *124*, 12352-12360.

(57) Orekhov, V. Y.; Korzhnev, D. M.; Kay, L. E. *J. Am. Chem. Soc.* **2004**, *126*, 1886-1891.

(58) Bouvignies, G.; Korzhnev, D. M.; Neudecker, P.; Hansen, D. F.; Cordes, M. H.; Kay, L. E. *J. Biomol. NMR* **2010**, *47*, 135-141.

(59) Hansen, D. F.; Vallurupalli, P.; Lundström, P.; Neudecker, P.; Kay, L. E. *J. Am. Chem. Soc.* **2008**, *130*, 2667-2675.

(60) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.

(61) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586-590.

(62) Kneller, D.; Kuntz, I. *J. Cell. Biochem.* **1993**, *53*, 254.

(63) Desai, T. M.; Cerminara, M.; Sadqi, M.; Muñoz, V. *J. Biol. Chem.* **2010**, *285*, 34549-34556.

(64) Tamiola, K.; Acar, B.; Mulder, F. A. *J. Am. Chem. Soc.* **2010**, *132*, 18000-18003.

Supporting Information

Probing the free energy landscape of the fast folding gpW protein by relaxation dispersion NMR

Celia Sanchez-Medina, Ashok Sekhar, Pramodh Vallurupalli, Michele Cerminara, Victor Muñoz and Lewis E. Kay

Supporting Methods

Free Energy Surface Model and Analysis of T-jump Kinetics

The free energy surface model that we use for the analysis of the infrared temperature-jump data has been described in detail before[1,2]. In general, it is a mean-field model that generates a one-dimensional free energy surface (FES) according to a single order parameter termed nativeness (n) that is assumed to operate as a good reaction coordinate for folding. The order parameter nativeness is defined as the probability of finding any given residue in the native dihedral angles (the specific sets of dihedral angles that are found in the native 3D structure). Therefore it is a local order parameter that goes from 0 for the fully unfolded state (all residues in non-native angles) to 1 for the native structure (all residues in native angles). Using this definition it is straightforward to define a relation that determines the changes in conformational entropy of the protein as a function of nativeness[2]:

$$\Delta S_{conf}(n) = N\left(-R\left[n\ln(n)+(1-n)\ln(1-n)\right]+(1-n)\Delta S_{conf,res}\right) \qquad [S1]$$

where n is nativeness, $\Delta S_{conf,res}$ is the average cost in conformational entropy of fixing one residue with native angles and N is the number of residues in the protein. The function that determines the changes in stabilization enthalpy (energy from interactions different from solvation free energy) as a function of nativeness is taken to be a Markov chain, which is equivalent to assume that native interactions are broken at a constant rate as the protein becomes increasingly disordered. Particularly, the stabilization enthalpy functional is defined as

$$\Delta H(n) = N\Delta H_{res}\left[\left(1-x^{(1-n)}\right)/(1-x)\right] \qquad [S2]$$

where ΔH_{res} is the total stabilization energy per residue, and the parameter x determines the rate of change of the stabilization energy as a function of n. Thus, the decay in stabilization enthalpy as the protein moves away from the fully native state ($n=1$) is exponential, and the term within square brackets in Eq. S2 is the fraction of the enthalpy present at any given value of nativeness. The protein specific parameter x determines how sharply the stabilization enthalpy decays; x is determined by the structural class and the specifics of the protein native structure and sequence[2]. Here we have directly used Eq. S2 with ΔH_{res} and x as adjustable

parameters to fit the experimental laser-induced temperature-jump data (rates and amplitudes).

The change in heat capacity as a function of the order parameter is treated in a similar way, and defined by the following expression

$$\Delta C_p(n) = \Delta C_{p,res}\left[\left(1 - x_c^{(1-n)}\right)/(1 - x_c\right] \qquad \text{[S3]}$$

where $\Delta C_{p,res}$ is the change in heat capacity per residue and x_c the characteristic rate of change of the heat capacity as a function of n. Combining Eqs. S1-S3 we obtain the free energy surface as a function of the order parameter

$$\Delta G(n) = \Delta H(n) - T\Delta S(n) + \Delta C_p(n)\left[(T - T_m) + T\ln(T/385)\right] \qquad \text{[S4]}$$

where T_m is the denaturation midpoint temperature (temperature at which the protein is half unfolded) for the protein and 385 K is the convergence temperature for the folding entropy[1]. Eq. S4 can be used directly to calculate the FES of a protein as a function of temperature from which the probability distribution can be obtained.

The relaxation kinetics are calculated as simple diffusion on the FES assuming a diffusion coefficient (D) that is constant as function of the nativeness parameter n, but depends on the solvent viscosity and temperature according to the following expression

$$D = \frac{k_0}{\eta(T)}\exp(-NE_a/RT) \qquad \text{[S5]}$$

where η is the viscosity of water as a function of temperature, k_o is the pre-factor and E_a is the activation energy per residue. Note that in this model, the temperature dependence of the relaxation rateis determined by D(equation S5) and from any changes in the FES shape induced by temperature as defined in equation S4.

For all the calculations described in this work we use a fixed value of $\Delta S_{conf,res}$ =16.5 J/(mol.K), which we generated by calculating the average value obtained from the Robertson and Murphy analysis of a database of calorimetric data [3]. For additional simplicity we fixed $\Delta C_{p,res}$ to the average value of 53.5 (52.5 J/(mol.K) obtained empirically by Robertson and Murphy[3]) and x_c to 0.0136, that has been used before by us[1]. We thus fit the infrared laser induced temperature-jump kinetic data (relaxation rate and amplitude as a function of temperature) to a model of four parameters (ΔH_{res}, x, k_0 and E_a), using N=61 for gpW. The overall parameters obtained from the fit to the gpW data at pH 3.5 are: ΔH_{res} = 4.23 kJ/mol, x=0.94, k_0=1.37x10^{12} s^{-1} and E_a =0.877 kJ/mol.

References

(1) Naganathan, A. N.; Doshi, U.; Muñoz, V. *J*. *Am*. *Chem*. *Soc*. **2007**, *129*, 5673-5682.

(2) De Sancho, D.; Muñoz, V. *Phys*. *Chem*. *Chem*. *Phys*. **2011**, *13*, 17030-17043.

(3) Robertson, A. D.; Murphy, K. P. *Chem*. *Rev*. **1997**, *97*, 1251-1268.

(4) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J*. *Biomol*. *NMR* **2009**, *44*, 213-223.

(5) Tamiola, K.; Acar, B.; Mulder, F. A. *J*. *Am*. *Chem*. *Soc*. **2010**, *132*, 18000-18003.

(6) Skrynnikov, N. R.; Dahlquist, F. W.; Kay, L. E. *J*. *Am*. *Chem*. *Soc*. **2002**, *124*, 12352-12360.

(7) Orekhov, V. Y.; Korzhnev, D. M.; Kay, L. E. *J*. *Am*. *Chem*. *Soc*. **2004**, *126*, 1886-1891.

(8) Bouvignies, G.; Korzhnev, D. M.; Neudecker, P.; Hansen, D. F.; Cordes, M. H.; Kay, L. E. *J*. *Biomol*. *NMR* **2010**, *47*, 135-141.

(9) Lundström, P.; Teilum, K.; Carstensen, T.; Bezsonova, I.; Wiesner, S.; Hansen, D. F.; Religa, T. L.; Akke, M.; Kay, L. E. *J*. *Biomol*. *NMR* **2007**, *38*, 199-212.

(10) Vallurupalli, P.; Bouvignies, G.; Kay, L. E. *J*. *Phys*. *Chem*. *B* **2011**, *115*, 14891-14900.

(11) Hansen, D. F.; Vallurupalli, P.; Lundström, P.; Neudecker, P.; Kay, L. E. *J*. *Am*. *Chem*. *Soc*. **2008**, *130*, 2667-2675.

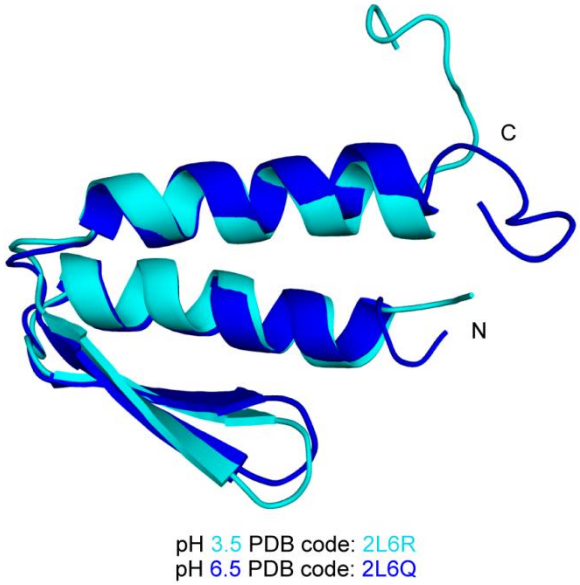


Figure S1. Superposition of NMR derived structures of gpW at pH 3.5 (blue, PDB: 2L6R), and at pH 6.5 (cyan, PDB: 2L6Q). Structures were superimposed using backbone heavy atoms (excluding CO).

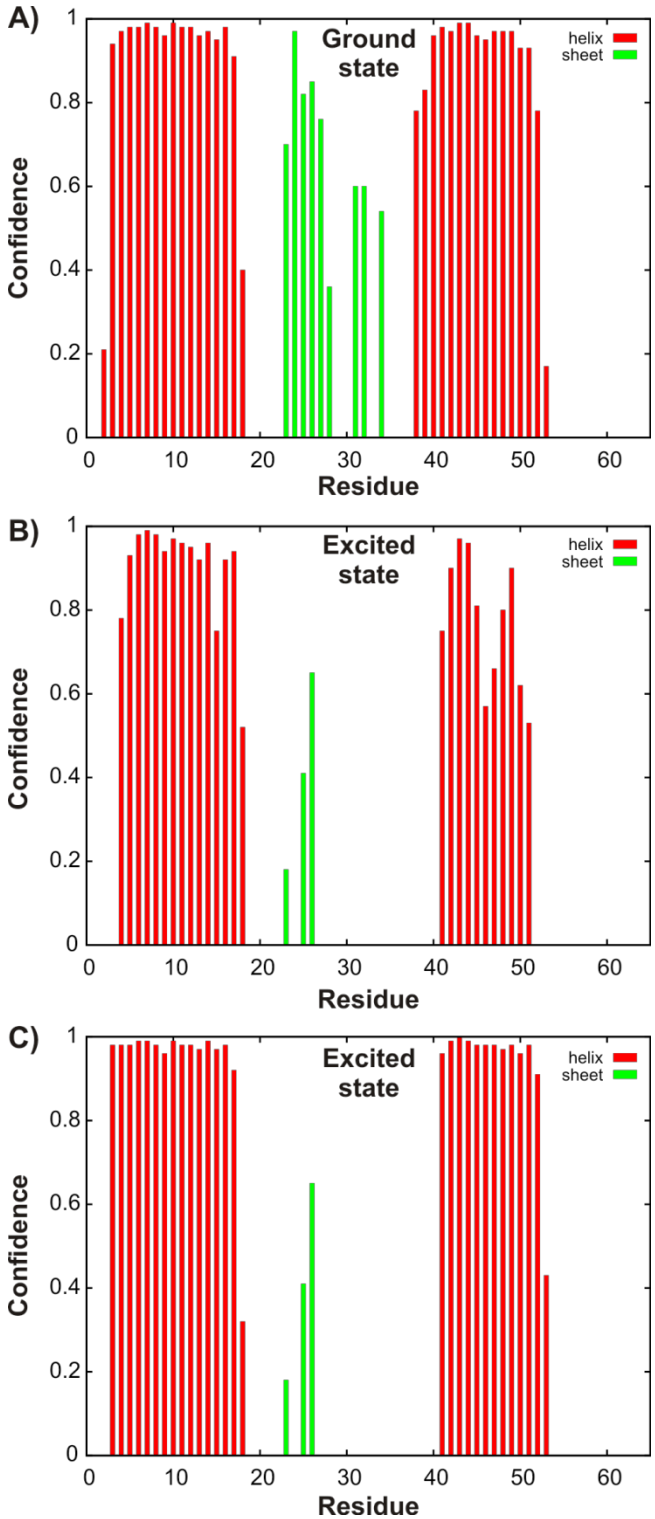


Figure S2. Confidence of secondary structural elements of state *X*, as predicted by TALOS+ [4]. A) Predicted secondary structure of the native state. B) Predicted secondary structure for the excited state. For residues localized to helices in the native gpW conformation and where signs of $\Delta\varpi_{XN} = \varpi_X - \varpi_N$ could not be determined experimentally, they were chosen so as to position ϖ_X closer to the random coil chemical shift. C) As in B but with the signs of $\Delta\varpi_{XN}$ chosen such that ϖ_X is placed further from the random coil chemical shift. Where the signs

were not available for nuclei in the beta hairpin in the native conformation, they were chosen so as to move ϖ_X closer to the random coil chemical shift (in both B and C).

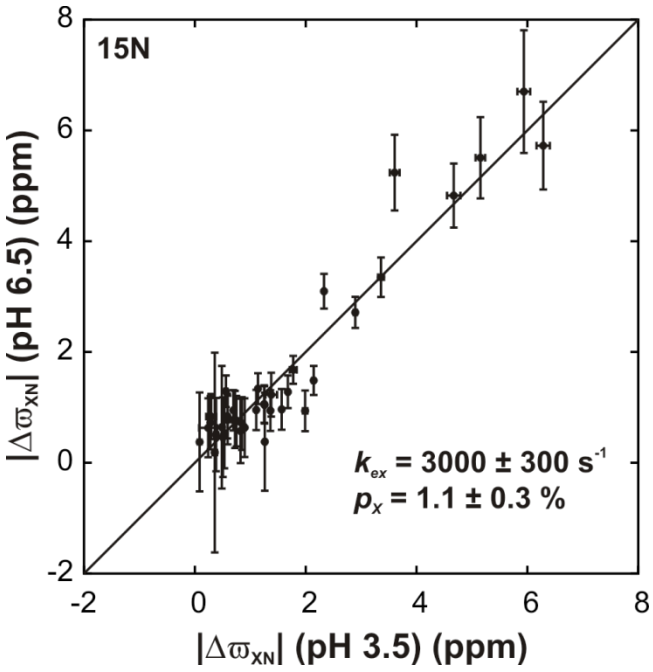


Figure S3. The excited state observed for gpW at pH 3.5 is also present at pH 6.5. Correlation between $|\Delta\varpi_{XN}|$ values for backbone ^{15}N nuclei, pH 3.5,*vs*those obtained at pH 6.5. Chemical shift differences at pH 6.5 were extracted from fits of ^{15}N RD profiles recorded at 11.7 T on a sample of ^{15}N-labeled gpW dissolved in a buffer comprised of 25 mM sodium phosphate, 50 mM NaCl, 0.5 mM EDTA, 0.5 mM NaN_3/10% D_2O, pH 6.5. Dispersion profiles were fit to a two-state model to extract p_X, k_{ex} and $\Delta\varpi_{XN}$ values. Although these values may not be accurately determined at pH 6.5 because of the availability of data at only a single static magnetic field, the clear correlation obtained between $|\Delta\varpi_{XN}|$ at pH 3.5 and 6.5 establishes that very similar excited states are present at the two pH values.

Table S1

^{15}N and $^{1}H^{N}$ chemical shifts for the native state, ϖ_N, and differences in chemical shifts between excited and native states $\Delta\varpi_{XN} = \varpi_X - \varpi_N$ of the gpW protein, 1°C.

Res	ϖ_N (ppm) $^{15}N^a$	ϖ_N (ppm) $^{1}H^{N\,a}$	ϖ_{RC} (ppm) $^{15}N^b$	ϖ_{RC} (ppm) $^{1}H^{N\,b}$	$\Delta\varpi_{XN}$ (ppm) $^{15}N^c$	$\Delta\varpi_{XN}$ (ppm) $^{1}H^{Nc}$
M1	-	-	-	-	-	-
V2	122.49	8.78	121.75	8.20	0.49 ± 0.02	0.000 ± 0.032
R3	124.88	8.95	125.71	8.47	0.75 ± 0.02	-0.096 ± 0.002
Q4	118.61	8.76	122.03	8.49	-1.26 ± 0.01	0.000 ± 0.289
E5	120.04	7.89	122.37	8.52	+0.29 ± 0.09	0.253 ± 0.005
E6	121.70	8.34	121.60	8.39	0.00 ± 1.00	0.000 ± 0.095
L7	121.77	8.42	123.10	8.21	-0.35 ± 0.05	0.069 ± 0.006
A8	121.51	7.93	124.74	8.21	+0.59 ± 0.07	-0.133 ± 0.003
A9	121.03	8.17	123.26	8.19	0.82 ± 0.01	0.000 ± 0.114
A10	d	d	123.20	8.21	d	d
R11	117.33	8.53	120.64	8.26	+0.90 ± 0.02	-0.153 ± 0.003
A12	123.92	8.08	125.81	8.35	-0.38 ± 0.04	0.000 ± 0.097
A13	121.88	7.90	122.87	8.16	0.00 ± 0.61	-0.196 ± 0.003
L14	d	d	121.01	8.04	d	d
H15	117.22	8.06	119.39	8.41	0.51 ± 0.02	-0.116 ± 0.003
D16	121.42	8.49	121.57	8.27	-1.11 ± 0.02	-0.124 ± 0.002
L17	120.33	7.97	122.41	8.10	+1.14 ± 0.02	-0.186 ± 0.002
M18	117.39	8.31	120.91	8.36	0.53 ± 0.03	0.000 ± 0.190
T19	108.70	7.53	115.43	8.25	+1.68 ± 0.02	-0.224 ± 0.002
G20	107.59	7.58	111.53	8.46	+2.90 ± 0.03	+0.343 ± 0.003
K21	121.30	8.03	121.29	8.18	-0.56 ± 0.03	-0.142 ± 0.002
R22	120.78	8.73	122.55	8.35	+1.99 ± 0.04	-0.391 ± 0.003
V23	e	e	122.31	8.23	e	e
A24	127.29	8.81	128.49	8.46	+1.57 ± 0.02	-0.276 ± 0.002
T25	115.83	8.61	113.34	8.08	-1.26 ± 0.02	-0.347± 0.002
V26	119.96	8.90	123.00	8.23	+3.61 ± 0.09	-0.604 ± 0.010
Q27	122.79	8.61	124.80	8.46	2.33 ± 0.01	0.032 ± 0.008
K28	127.76	9.13	122.68	8.44	-4.67 ± 0.12	-0.462 ± 0.005
D29	e	e	121.52	8.39	e	e
G30	103.89	8.72	109.40	8.35	+5.15 ± 0.09	-0.334 ± 0.003
R31	119.73	7.74	121.22	8.19	0.85 ± 0.02	0.304 ± 0.003
R32	122.43	8.57	123.15	8.43	-0.41 ± 0.02	0.015 ± 0.014
V33	127.72	9.02	122.21	8.27	5.27 ± 0.21	0.701 ± 0.023
E34	d	d	124.79	8.43	d	d
F35	122.66	9.19	121.44	8.32	1.94 ± 0.20	0.990 ± 0.039
T36	110.45	8.22	116.26	8.07	+5.94 ± 0.12	-0.059 ± 0.007
A37	121.97	9.22	126.66	8.41	4.67 ± 0.27	0.763 ± 0.031
T38	107.25	7.92	113.33	8.13	+6.29 ± 0.12	-0.264 ± 0.003
S39	115.94	8.14	118.57	8.37	+3.36 ± 0.05	+0.233 ± 0.003
V40	122.95	e	122.65	8.27	1.63 ± 0.39	e
S41	116.43	9.06	120.07	8.47	-1.38 ± 0.10	-0.739 ± 0.007
D42	121.91	7.75	122.11	8.36	0.00 ± 0.88	+0.593 ± 0.007
L43	124.86	7.74	122.62	8.12	-1.77 ± 0.06	+0.588 ± 0.011

K44	118.86	8.53	122.26	8.27	+0.36 ± 0.04	-0.132 ± 0.004
K45	120.84	7.88	122.28	8.30	-0.21 ± 0.07	-0.032 ± 0.010
Y46	122.27	7.75	121.37	8.13	+0.24 ± 0.16	0.375 ± 0.003
I47	119.32	8.34	123.24	7.86	+1.36 ± 0.02	0.176 ± 0.005
A48	119.78	7.80	128.01	8.39	+2.15 ± 0.03	-0.108 ± 0.003
E49	d	d	119.73	8.31	d	d
L50	120.74	8.20	123.00	8.25	-0.76 ± 0.02	0.000 ± 0.112
E51	118.58	8.23	121.56	8.36	-0.69 ± 0.02	0.000 ± 0.170
V52	119.40	7.54	121.54	8.20	0.72 ± 0.02	-0.191 ± 0.002
Q53	121.72	8.28	124.81	8.56	0.59 ± 0.02	0.047 ± 0.007
T54	112.59	8.15	115.72	8.27	+1.25 ± 0.02	0.097 ± 0.003
G55	110.42	8.18	111.32	8.45	0.28 ± 0.05	0.133 ± 0.003
M56	119.93	8.19	120.08	8.25	0.01 ± 0.71	0.028 ± 0.005
T57	115.50	8.21	115.34	8.24	0.00 ± 0.43	0.040 ± 0.008
Q58	d	d	122.84	8.43	d	d
R59	d	d	122.92	8.42	d	d
R60	123.61	8.56	123.13	8.44	0.05 ± 0.09	0.000 ± 0.021
R61	123.80	8.61	123.20	8.48	0.09 ± 0.05	0.000 ± 0.010
G62	115.51	8.29	-	-	0.18 ± 0.03	0.000 ± 0.032

[a] ^{15}N and $^{1}H^{N}$ chemical shifts in the native state (ϖ_N) were measured from an $^{1}H^{N}$-^{15}N HSQC spectrum recorded at 1 ºC.

[b] Random coil (RC) values, ϖ_{RC}, were predicted using the Neighbor Corrected Intrinsically Disordered Protein Library[5].

[c] Values of $\Delta\varpi$ have been measured by RD NMR experiments, as described in the text. Where possible, signs of $\Delta\varpi_{XN}$ (^{15}N) have been obtained by a comparison of peak positions in (i) sets of HSQC and HMQC experiments recorded at 11.7 T and 18.8T and/or in (ii) pairs of HSQC experiments measured at 11.7, 18.8T, as described previously [6].Signs could not be obtained for those shift values that do not have + or -. Signs of $^{1}H^{N}$ $\Delta\varpi_{XN}$ values were determined on the basis of (i) $^{1}H^{N}/^{15}N$ zero-quantum and double-quantum coherence relaxation dispersion experiments[7] or (ii) from a comparison of peak positions in the directly detected dimension of HSQC experiments acquired at 11.7 T and 18.8 T [8].

[d] Overlapped peaks.

[e] Peaks do not appear in spectra.

Table S2

$^{13}C^{O}$ and $^{13}C^{\alpha}$ chemical shifts for the native state, ϖ_N, of the gpW protein and differences in shifts between excited and native states $\Delta\varpi_{XN} = \varpi_X - \varpi_N$, 1°C.

Res	ϖ_N (ppm) $^{13}C^{O}$ [a]	ϖ_N (ppm) $^{13}C^{\alpha}$ [b]	ϖ_{RC} (ppm) $^{13}C^{O}$ [c]	ϖ_{RC} (ppm) $^{13}C^{\alpha}$ [c]	$\Delta\varpi_{XN}$ (ppm) $^{13}C^{O}$ [d]	$\Delta\varpi_{XN}$ (ppm) $^{13}C^{\alpha}$ [e]
M1	172.27	f	-	-	-0.39 ± 0.01	f
V2	176.53	62.81	176.03	62.12	-0.60 ± 0.01	g
R3	177.36	58.76	176.07	56.07	-0.61 ± 0.02	0.75 ± 0.05
Q4	178.12	59.31	175.76	55.80	-0.76 ± 0.02	0.77 ± 0.03
E5	178.56	58.57	176.47	56.70	-0.93 ± 0.01	0.66 ± 0.03
E6	178.68	h	176.48	56.64	-1.00 ± 0.01	h
L7	177.15	57.82	177.19	55.24	0.57 ± 0.01	g
A8	180.62	f	177.49	52.49	-0.77 ± 0.01	f
A9	f	54.78	177.67	52.57	f	0.25 ± 0.04
A10	178.73	f	177.86	52.50	0.50 ± 0.03	f
R11	179.22	60.28	176.07	56.06	-0.77 ± 0.01	0.88 ± 0.05
A12	178.79	f	177.36	52.44	0.33 ± 0.02	f
A13	f	54.92	177.87	52.58	f	0.53 ± 0.02
L14	177.35	58.00	177.43	55.17	0.25 ± 0.03	g
H15	177.29	59.38	174.38	55.95	-1.02 ± 0.02	1.02 ± 0.03
D16	177.40	57.22	176.13	54.22	0.95 ± 0.03	1.15 ± 0.04
L17	180.01	57.29	177.83	55.49	-0.97 ± 0.02	g
M18	177.34	f	176.51	55.55	-0.45 ± 0.03	f
T19	174.59	61.20	175.00	61.95	+0.31 ± 0.12	0.85 ± 0.02
G20	174.83	h	173.98	45.24	-1.38 ± 0.01	h
K21	176.29	56.26	176.60	56.15	0.70 ± 0.05	0.60 ± 0.03
R22	f	57.96	176.04	56.07		2.29 ± 0.11
V23	172.91	59.81	175.71	62.19	+2.58 ± 0.03	g
A24	175.25	51.44	177.89	52.55	+2.12 ± 0.02	0.93 ± 0.02
T25	173.69	h	174.53	61.87	+0.49 ± 0.18	h
V26	173.39	59.74	176.09	62.39	+2.13 ± 0.02	g
Q27	174.79	f	175.79	55.77	+1.38 ± 0.12	f
K28	f	55.75	176.22	56.44	h	1.42 ± 0.05
D29	175.19	55.07	176.87	54.53	0.80 ± 0.09	1.48 ± 0.04
G30	173.15	h	174.06	45.27	+0.63 ± 0.01	h
R31	173.60	54.10	176.20	55.83	+2.03 ± 0.02	2.07 ± 0.10
R32	h	f	175.91	55.90	h	f
V33	f	61.58	176.02	62.30	f	h
E34	175.00	f	176.04	56.48	0.04 ± 4.08 [i]	f
F35	f	57.41	176.01	57.98	f	0.28 ± 0.04
T36	h	j	173.63	61.58	h	j
A37	h	55.50	177.96	52.62	h	2.72 ± 0.07
T38	175.08	63.62	174.71	61.94	1.34 ± 0.03	1.30 ± 0.02
S39	h	j	174.44	58.28	h	j
V40	176.85	66.56	176.16	62.35	+0.45 ± 0.15	g
S41	176.90	61.92	174.18	58.33	-0.72 ± 0.08	1.30 ± 0.04
D42	178.14	56.97	176.25	54.30	0.53 ± 0.17	0.90 ± 0.03
L43	177.59	57.49	177.58	55.41	0.55 ± 0.02	g

K44	179.20	60.83	176.59	56.35	-0.72 ± 0.01	0.93 ± 0.03
K45	177.73	59.62	176.26	56.46	-0.57 ± 0.02	0.83 ± 0.02
Y46	176.46	58.66	175.61	57.83	0.29 ± 0.05	0.95 ± 0.03
I47	176.48	66.19	175.45	60.82	0.10 ± 0.18	g
A48	f	54.64	177.87	52.44	f	0.41 ± 0.03
E49	178.72	58.58	176.58	56.63	-1.05 ± 0.01	0.84 ± 0.04
L50	179.20	57.07	177.50	55.36	0.66 ± 0.01	g
E51	177.46	58.33	176.35	56.52	0.50 ± 0.02	0.81 ± 0.05
V52	177.56	64.64	176.19	62.42	0.66 ± 0.01	g
Q53	176.76	57.15	175.98	55.79	0.22 ± 0.02	0.40 ± 0.04
T54	f	62.70	174.97	62.05	f	0.26 ± 0.04
G55	173.97	45.57	174.23	45.32	0.06 ± 0.05	0.00 ± 0.06
M56	176.28	55.91	176.63	55.47	0.00 ± 0.50	0.04 ± 0.18
T57	f	62.24	174.45	61.86	f	0.19 ± 0.03
Q58	f	55.70	175.75	55.73	f	0.11 ± 0.07
R59	175.73	f	176.04	55.95	0.13 ± 0.02	f
R60	175.67	56.14	175.95	55.86	0.12 ± 0.02	0.20 ± 0.02
R61	175.47	f	176.57	56.12	0.00 ± 0.05	f
G62	-	h	-	-	-	h

[a] $^{13}C^{O}$chemical shifts were measured from 2D ($^{13}C^{O}$,$^{1}H^{N}$) HNCO-based spectra recorded on a U-^{15}N/^{13}C gpW sample, 1^{o}C.

[b]$^{13}C^{\alpha}$ shifts were measured on a U-^{15}N and ^{13}C-selectively labeled sample (using 2-^{13}C-glucose as the ^{13}C source[9]).

[c]ϖ_{RC} values were predicted using the Neighbor Corrected Intrinsically Disordered Protein Library[5].

[d]$\Delta\varpi_{XN}$ values were obtained by fitting $^{13}C^{O}$RD CPMG profiles using fixed values for k_{ex} (4087 ± 42 s^{-1}) and p_X (9.0 ± 0.3 %), that, in turn, were generated by simultaneous fits of ^{15}N, $^{1}H^{N}$ and ^{13}C-methyl RD curves and the static magnetic field dependence of ^{15}N chemical shifts as described in the text [10]. Signs of $\Delta\varpi_{XN}$ values were obtained from peak positions in $^{1}H^{N}$-$^{13}C^{O}$correlation maps where the $^{13}C^{O}$ chemical shift was recorded either as single quantum or multiple quantum (^{15}N-$^{13}C^{O}$or$^{1}H^{N}$-$^{13}C^{O}$) coherence[11]. Experiments were recorded at 11.7 and 18.8T.

[e]$^{13}C^{\alpha}$ shift differences were obtained by fitting RD CPMG profiles using fixed values for k_{ex} (4087 ± 42 s^{-1}) and p_X (9.0 ± 0.3 %), as described above. Dispersion profiles derived from I or V were not included because the $^{13}C^{\alpha}$ spins are not isolated ($^{13}C^{\beta}$ labeling) while L is not labeled at the C-α position using 2-^{13}C-glucose as the ^{13}C source[9]. Signs of $\Delta\varpi_{XN}$ values could not be obtained because of the limited quality of $^{1}H^{\alpha}$ – $^{13}C^{\alpha}$ correlation maps.

[f] Overlapped peaks.

[g] I,L,V residues that cannot be quantified.

[h] Very weak peak or does not appear in the spectra.

[i] Not used in analysis

[j] Peak is overlapped with the water signal

Table S3

^{13}C-methyl chemical shifts for the native state of the gpW protein, ϖ_N ,and chemical shift differences between the excited and native states $\Delta\varpi_{XN} = \varpi_X - \varpi_N$, 1°C.

Residue Number[a]	ϖ_N (ppm) ^{13}C[b]	$\Delta\varpi_{XN}$ (ppm) ^{13}C[c]
M1ε	16.887	0.07 ± 0.06
V2γa	20.517	d
V2γb	e	e
L7δa	24.054	0.56 ± 0.01
L7δb	e	e
A8β	e	e
A9β	18.165	0.16 ± 0.03
A10β	18.615	0.30 ± 0.02
A12β	e	e
A13β	e	e
L14δa	23.940	0.69 ± 0.01
L14δb	e	e
L17δa	22.152	1.78 ± 0.02
L17δb	26.283	1.14 ± 0.02
M18ε	16.342	0.32 ± 0.01
V23γa	20.483	0.85 ± 0.00
V23γb	20.607	0.00 ± 0.04
A24β	23.099	f
V26γa	21.157	0.00 ± 0.16
V26γb	22.059	1.09 ± 0.01
V33γa	21.461	0.18 ± 0.05
V33γb	21.775	1.31 ± 0.01
A37β	18.496	0.56 ± 0.01
V40γa	e	e
V40γb	23.664	1.11 ± 0.01
L43δa	21.276	2.52 ± 0.09
L43δb	26.096	1.94 ± 0.08
A48β	18.117	0.19 ± 0.03
L50δa	22.168	0.71 ± 0.01
L50δb	25.059	0.00 ± 0.14
V52γa	21.241	0.00 ± 0.06
V52γb	21.672	0.40 ± 0.03
M56ε	16.818	0.09 ± 0.05

[a] The letters a and b next to the atom labels refer to upfield (a) or downfield (b) peaks as they appear in the ^{13}C dimension.

[b] Native state $^{13}CH_3$ chemical shifts were obtained from a ^{1}H-^{13}C correlation map recorded at 1ºC.

[c] Chemical shift differences for the methyl group were obtained by fitting RD CPMG profiles acquired at 11.7 and 18.8 T, 1ºC, as described in the text.

[d] The peak shows coupling.

[e] Overlapped peaks

[f] Peak is either very weak, does not appear in the spectrum or it is impossible to get a reliable $\Delta\varpi$ value.

[g] Only some atoms of Leu, Val, Ala and Met sidechains have been used in the analysis, in accordance with the labeling scheme expected from growth with 1-^{13}C-glucose as the carbon source [9].