Cláudia Filipa Martins Afonso

Mestre em Neurociências
Licenciada em Química com minor em Biologia

Development of *in-cell* NMR methodologies

Dissertação para obtenção do Grau de Mestre em
Química Bioorgânica

Orientador: Professor Doutor Eurico José da Silva Cabrita, Professor Auxiliar com Agregação, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Co-orientador: Doutora Filipa Margarida Barradas de Morais Marcelo, Investigador Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Júri:

Presidente: Professora Doutora Paula Cristina de Sério Branco, Professor Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Arguente: Professora Doutora Maria dos Anjos Lopez de Macedo, Professor Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Vogal: Professor Doutor Eurico José da Silva Cabrita, Professor Auxiliar com Agregação, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2017

Development of *in-cell* NMR methodologies

When I heard the learn’d astronomer,
When the proofs, the figures, were ranged in columns before me,
When I was shown the charts and diagrams, to add, divide, and measure them,
When I sitting heard the astronomer where he lectured with much applause in the lecture-room,
How soon unaccountable I became tired and sick,
Till rising and gliding out I wander’d off by myself,
In the mystical moist night-air, and from time to time,
Look’d up in perfect silence at the stars.

Walt Whitman

Agradecimentos

Gostaria de agradecer à Professora Paula Branco pela ajuda prestada como coordenadora durante o meu percurso no mestrado em Química Bioorgânica. Este agradecimento estende-se, de um modo geral, aos restantes Professores deste mestrado.

Queria igualmente agradecer ao meus orientadores, o Professor Eurico Cabrita e a Dra. Filipa Marcelo, por ter me terem aceite como aluna. Ao Professor Eurico Cabrita ainda pela disponibilidade em falar comigo e pelas correções durante a fase de escrita da dissertação.

Um agradecimento especial ao Micael Silva, pela ajuda nas experiências de *in-cell* e pelo esclarecimento de muitas das dúvidas que lhe coloquei, e ao Doutor Fábio Fernandes, por toda a ajuda durante a aquisição e tratamento de resultados das experiências de FCS.

Gostaria de agradecer aos restantes membros do grupo, sobretudo aos meus colegas de laboratório, o Diogo e a Ana Sofia, pela amizade durante este ano.

Queria agradecer também a ajuda do grupo de cristalografia da FCT-UNL, em particular do Francisco Leisico.

Gostaria também de agradecer aos meus colegas de mestrado, sobretudo à Inês Rosete e ao Mário Almeida, por toda a amizade e ajuda durante estes dois anos. Teria sido mais difícil fazer dois mestrados ao mesmo tempo sem vocês.

E finalmente, um agradecimento muito especial à minha família.

Resumo

O estudo de moléculas num meio biologicamente relevante distingue a espectroscopia de ressonância magnética nuclear (RMN) das restantes técnicas, tais como a cristalografia de raios-X ou a criomicroscopia eletrónica. Devido à sua elevada estabilidade e fraca interação com outros componentes celulares, a proteína GB1 representa a sonda por excelência para estudar os efeitos físico-químicos impostos na estrutura e dinâmica de proteínas pelo aglomerado macromolecular existente nas células, sem comprometer a acquisição de espectros *in-cell* de RMN.

A presente dissertação teve por objectivo investigar as possíveis interações da GB1 com o lisado e ambiente intracelular de *Escherichia coli* de modo a inferir sobre os efeitos físico-químicos impostos por estes dois meios molecularmente densos na estrutura e dinâmica de proteínas. Assim, os parâmetros experimentais críticos para a realização de experiências *in-cell* foram inicialmente otimizados. Subsequentemente, pela monitorização dos desvios químicos de protão e azoto correspondentes aos grupos amida do esqueleto proteico e cadeias laterais de lisinas, assim como de protão e carbono das cadeias laterais de resíduos contendo grupos carbonilo, o comportamento da GB1 foi estudado em lisado e *E. coli*, considerando a proteína pura em água como referência. Além disso, as interações com o ambiente local foram também avaliadas através da determinação do movimento de translação global da proteína através de "diffusion-ordered NMR spectroscopy" (DOSY). Os resultados obtidos sugerem que o meio intracelular é mais viscoso que a correspondente solução composta por 150 mg/ml de lisado e que os resíduos nas regiões "loop" mais flexíveis e expostas ao solvente, ou próximas destas, apresentam uma maior preferência pelos componentes celulares em *E. coli* relativamente ao lisado. Finalmente, foi feita uma comparação dos coeficientes de difusão obtidos por DOSY e espectroscopia de correlação de fluorescência (FCS), a técnica padrão para o estudo da difusão de proteínas.

PALAVRAS-CHAVE

Espectroscopia de resonância magnética nuclear, proteína GB1, aglomeração macromolecular, difusão

Abstract

The characterization of molecules within a biologically relevant environment distinguishes nuclear magnetic resonance (NMR) spectroscopy from other molecular-based biophysical techniques, such as X-ray crystallography and cryo-electron microscopy. Due to its exceptional stability and reduced ability to interact in a specific manner with other cellular components, the GB1 protein represents the quintessential probe to investigate the physiochemical effects imposed by the crowded environment on the structure and dynamics of proteins, without simultaneously compromising the ability to obtain *in-cell* NMR spectra due to binding events.

The general aim of this thesis was to investigate the possible interactions of the GB1 protein with the *Escherichia coli* lysate and intracellular milieu with the purpose of inferring the physiochemical effects imposed by these two crowded environments on the structure and dynamics of proteins. Thus, the experimental parameters critical for performing *in-cell* NMR experiments, including bacterial growth and protein overexpression within *E. coli* cells, were initially optimized. Subsequently, by monitoring proton and nitrogen chemical shifts of backbone amides and lysines side chains, as well as carbon and proton chemical shifts of side chains containing carbonyl groups, the preferential behaviour of GB1 was analysed in lysate and within cells, considering the pure protein in water as the reference state. Furthermore, interactions with the local environment were further examined by determining the overall translational motion of the protein through diffusion-ordered NMR spectroscopy (DOSY). The results obtained suggest that the intracellular environment is much more viscous than its artificially crowded counterpart and that GB1 exhibits a distinct behaviour in *E. coli* than in lysate. Specifically, residues at or near the more flexible and solvent-exposed loop regions of the protein display an increased preference for interaction with cellular components within cells compared to lysate. Finally, a comparison of diffusion coefficients obtained with DOSY and fluorescence correlation spectroscopy (FCS), the standard analytical technique for studying protein diffusion, was made.

KEYWORDS

In-cell NMR spectroscopy, GB1 protein, macromolecular crowding, diffusion.

Contents

List of Figures

List of Tables

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
4D	Four-dimensional
BSA	Bovine serum albumin
CI2	Chymotrypsin inhibitor 2
CK2	Casein kinase 2
CPP	Cell-penetrating peptide
DOL	Degree of labelling
DOSY	Diffusion-ordered spectroscopy
E. coli	*Escherichia coli*
FCS	Fluorescence correlation spectroscopy
GB1	B1 domain of protein G
HIV-1	Human immunodeficiency virus-1
HSQC	Heteronuclear single quantum coherence
IPTG	Isopropyl-β-D-thiogalactopyranoside
LB	Lysogeny broth
MBP	Maltose-binding protein
MRI	Magnetic resonance imaging
MSG	Malate synthase G
MW	Molecular weight
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
OD	Optical density

PDB	Protein Data Bank
PEG	Polyethylene glycol
pI	Isoelectric point
PFTs	Pore-forming toxins
ppm	Parts per million
PVP	Polyvinylpyrrolidone
rpm	Revolutions per minute
SLO	Streptolysin O
SOD1	Superoxide dismutase 1
Tβ4	Thymosin β4 protein
TSP	Trimethylsilylpropanoic acid

1. Introduction

Biological systems are inherently complex and detailed knowledge on the structure of their basic components, *i.e.* biomolecules, is required to characterize their intricate behaviour. Information regarding the three-dimensional structures of target biomolecules is particularly relevant in the design of new therapeutic compounds.[1] Since the first structural elucidation of DNA in 1953 and myoglobin in 1958, the atomic coordinates of more than 132,000 of proteins, nuclei acids and their complexes have been deposited, as of August 2017, in the Protein Data Bank (PDB).[2,3]

Due to the wealth of structural data already collected, the interest of the scientific community is now shifting towards the functional characterization of biomolecules. To fully achieve this, it is essential to validate the structures obtained *in vitro* within a biologically relevant context.[4] However, most biophysical methods employed in the atomic resolution study of biomolecules are constrained to artificial conditions, remote from the native cellular environments in which these molecules exist and exert their proper biological functions. For instance, X-ray crystallography, the first and still the leading and most powerful technique today in the field of structural biology, requires pure and regular crystals to generate diffraction patterns following irradiation with X-ray beams.[5] Furthermore, proteins that contain large mobile or intrinsically disordered regions are usually difficult, if not impossible, to crystallize.[6] By contrast, nuclear magnetic resonance (NMR) spectroscopy represents the most versatile technique in modern structural biology, due to its unique ability to characterize the behaviour of biomolecules in solution.[7]

NMR spectroscopy exploits the intrinsic property of nuclear spin to provide information on the chemical environment surrounding atoms. In brief, nuclei with nonzero spin quantum number *I*, termed “NMR-active nuclei”, exhibit a magnetic moment. When placed in a static magnetic field, the individual magnetic moments of each nucleus align themselves in a discrete number of orientations with respect to the field, rotating continuously about its axis in a motion referred to as Larmor precession. Each orientation or spin state is associated with a specific energy. The physical phenomenon responsible for NMR occurs when nuclei, driven by the absorption of electromagnetic radiation arising from a magnetic field oscillating at the Larmor frequency of the spin, change their state. However, since electrons are charged particles that also possess spin, their precession induces local secondary magnetic fields that alter the strength of the magnetic field experienced by nuclei.[8] Therefore, NMR spectroscopy is sensitive to differences in the chemical environment surrounding nuclei and structural and dynamic information on biomolecules can thus be deduced based on this.[9]

The energy differences between spin states are small, so the corresponding absorption of electromagnetic radiation occurs in the low frequency radio wave region of the spectrum. As a result, NMR spectroscopy does not lead to meaningful tissue or cellular damage, a characteristic which is well suited to its application in the study of biomolecules in living organisms.[10] Yet,

NMR studies have been mostly restricted to dilute and isolated conditions, with buffers being chosen not to mimic the intracellular milieu but instead to optimize parameters such as solubility or signal sensitivity.[10,11] Indeed, a critical distinction between *in vitro* and intracellular environments pertains to the high concentration of macromolecules in the latter, which can extend to 400 g/L in the prokaryotic cytoplasm.[12] These high concentrations of macromolecules lead to excluded volume effects that, in addition to interactions with other molecules and posttranslational modifications, which occur inside living cells and can also influence the structural and dynamic behaviour of biomolecules, are difficult to replicate *in vitro*. Therefore, an increasingly attractive avenue for acquiring information about biomolecules at atomic resolution in their native physiological environment is provided by *in-cell* NMR.[13]

1.1 *In-cell* NMR spectroscopy as a new tool for studying proteins in their native environment

In-cell NMR is a relatively new application of high-resolution NMR spectroscopy which provides atomic resolution of macromolecular structure and dynamics within a highly crowded, viscous and complex, but biologically relevant environment.[14] For this purpose, the resonance frequencies of biomolecules are distinguished from the unlabelled cellular environment and made visible in multi-dimensional correlation experiments following isotopic enrichment with NMR-active nuclides, particularly ^{15}N and ^{13}C.[15] This contrasts with the more well-established technique of *in vivo* NMR, in which characteristic resonance lines are obtained from one-dimensional (1D) spectra of metabolites containing naturally abundant NMR-active nuclides such as ^{1}H and ^{31}P or, alternatively, after infusion with small molecules labelled with NMR-sensitive isotopes, mostly ^{13}C, which later diffuse or become actively transported to the interior of cells in living organisms.[16]

1.1.1 Overview of methodologies

The inherent insensitivity of NMR spectroscopy requires relatively high concentrations of isotopically-enriched biomolecules to be present, in order to distinguish the resonance peaks of interest from the remaining cellular components. To this end, distinct biological systems, delivery methods and labelling techniques can be employed as discussed below, with a special focus on proteins.

Biological systems and delivery methods

The first *in-cell* NMR proof-of-concept was described in 2001 by Serber and colleagues and performed using an *Escherichia coli* (*E. coli*) T7-dependent overexpression system grown in ^{15}N-enriched media. This approach was established for the intracellular observation of bacterial

N-terminal metal-binding domain of mercuric ion reductase (NmerA) and human calmodulin above other cellular components by ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR, which allows for the identification of amide bonds in proteins.[17,18] Since then, recombinant protein expression in *E. coli* has proven to be a suitable and low-cost strategy for *in-cell* NMR studies of proteins that tumble freely in the cytoplasm and do not exceedingly interact with cellular components.

The general protocol consists of a two-step culture, in which transformed bacterial cells are first grown in unlabelled lysogeny broth (LB) medium to an optical density (OD) of 0.6-0.8 and then transferred to isotopically-enriched minimal medium, with the most commonly used strategy employing ammonium chloride (NH_4Cl) and glucose as the sole sources of metabolic precursors. Following induction of gene expression downstream of the T7 promoter, which outperforms endogenous protein synthesis, labelled isotopes become selectively incorporated into the plasmid-encoded polypeptides. This scheme, in which unlabelled LB medium is switched to fresh labelled minimal medium prior to induction, diminishes background resonance peaks and increases spectral quality, because bacterial growth is significantly reduced during overexpression in minimal medium.[18] Accordingly, protein expression, isotopic labelling and NMR measurements occur in the same cell type and sample preparation is relatively straightforward. However, recombinant expression can be difficult to control and protein concentrations can vary between experiments. Moreover, considerable amounts of recombinant proteins in bacteria can lead to the formation of insoluble intracellular aggregates in the shape of inclusion bodies, which is detrimental to solution-state *in-cell* NMR.[19] Nevertheless, if recombinant expression is low and prolonged induction times are necessary to achieve adequate concentrations of intracellular proteins, then diminished cell viability and increased background labelling due to accumulation of metabolites can restrict the applicability of *in-cell* NMR experiments.

While *E. coli* remains the best characterized system for *in-cell* NMR due to its technical simplicity, bacteria may not represent the most adequate framework for studying eukaryotic proteins. Ideally, these proteins should be characterized in a cellular environment which best mirrors the native one. For instance, due to fundamental differences in cell morphology, internal organization and function between prokaryotes such as *E. coli* and eukaryotes, bacteria may lack the required machinery to fold and affect the post-translational modifications of eukaryotic proteins. Moreover, functional partners will likely be absent and the protein will not be targeted to its proper sub-cellular compartment.[20]

In this context, the first successfully conducted eukaryotic *in-cell* NMR experiments were reported in *Xenopus laevis* oocytes by Selenko and colleagues in 2006. Using this model system, the ^{15}N-labelled B1 immunoglobulin-binding domain of *streptococcal* protein G (GB1, 6.2 kDa) was recombinantly expressed and purified from *E. coli* and then delivered to the oocytes by microinjection.[21] This approach provides a higher degree of reproducibility and selectivity than is generally possible by overexpression in bacteria, because a definite amount of isotopically labelled protein is introduced into unlabelled cells following purification from background NMR

signals. Xenopus oocytes have long been employed in biomedical research, in part due to their large size (1 mm diameter) and cell volume (1 µL) allowing convenient manipulation and injection of different materials such as nuclei acids and proteins. Considering the small injection volume per cell (50 nL), an *in-cell* NMR sample with a final volume of 250 µL inside a Shigemi tube containing 200 oocytes requires only 10 µL of labelled protein. The concentration of injected protein required to achieve an adequate signal-to-noise ratio must be determined empirically but is normally in the range of 0.5-3.0 mM.[22]

In other eukaryotic model systems, such as mammalian cells, sample preparation by microinjection is technically unfeasible, because an absurd number of cells would need to be manipulated. Therefore, a distinct delivery mechanism that targets many cells simultaneously while still ensuring high sample deposition and cell viability is required. To this end, three techniques are currently available, namely active transport via cell-penetrating peptide (CPP) tags, passive diffusion through-pore forming toxins and electroporation.

CPPs are short peptides, typically comprising 5 to 30 amino acids, which can translocate across membranes and deliver various cargo molecules, such as proteins, nuclei acids and drugs, with low cellular toxicity.[23,24,25] Although the precise mechanism of entry into cells is still under debate, it is believed that most CPPs have two or more internalization pathways, depending on experimental conditions such as temperature, concentration and weight of the cargo.[26] The first CPP was described independently by two groups in 1988, when it was discovered that the trans-activator of transcription (Tat) protein from the human immunodeficiency virus 1 (HIV-1) could enter cells and translocate to the nucleus.[27,28] In 1998, the peptide sequence of Tat necessary for cellular uptake was identified (^{47}YGRKKRRQRRR^{57}) and the majority of CPPs are indeed cationic peptides rich in arginine (R) and lysine (K) residues.[29] Based on this knowledge, Inomata and colleagues established the first method to observe *in-cell* NMR spectra of proteins delivered by CPPs in 2009.[30] The CPP sequence of Tat was fused to the carboxy terminus human ubiquitin that was mutated at three sites to prevent binding with ubiquitin-interacting proteins. Following isotopic labelling, the fusion protein was able to translocate through the plasma membrane of cultured HeLa cells pre-incubated with pyrenebutyrate, a negatively-charged counterion with high hydrophobicity that facilitates translocation through lipid bilayers.[30,31] Furthermore, the authors also demonstrated that the CPP can be covalently conjugated to the protein through a disulphide bond, which is subject to intracellular cleavage by autonomous reduction.[30]

An alternative approach to CPP-mediated delivery for eukaryotic *in-cell* NMR experiments was also reported in 2009 by means of pore-forming toxins (PFTs). Ogino and colleagues employed the bacterial toxin streptolysin O (SLO) to permeabilize the plasma membrane and introduce the actin-sequestering protein thymosin β4 (Tβ4) into 293F cells.[32] This toxin binds to cholesterol as a monomer and oligomerizes laterally into ring-shaped transmembrane structures containing 50 to 80 subunits to form pores exceeding 30 nm in diameter, through which proteins up to 150 kDa can diffuse in a concentration-dependent manner.[33,34] At high concentrations of SLO, irreversible membrane disruption occurs, leading to cell death by lysis.

At low doses, the plasma membrane can recover by extracellular treatment with Ca^{2+} ions, which permeate the cells and trigger the arachidonic acid cascade that promotes resealing.[34,35] However, incubation of cells with even small amounts of PFTs can induce cellular responses which activate signalling pathways to restore membrane integrity and ion homeostasis. For instance, PFTs stimulate the activity of p38 MAP kinase to re-establish cytosolic ATP and K^+ ion levels while arresting protein synthesis.[36] Therefore, the intracellular environment following PTF-mediated delivery may not represent the true physiological state of the cell.

More recently, a method based on electroporation to introduce proteins into eukaryotic cells for *in-cell* NMR studies was described. Since the first successful intracellular delivery of DNA into mammalian cells in 1982, electroporation constitutes an established technique for introducing nucleic acids, proteins and drugs into cells.[37] It relies on the use of short high-voltage pulses to generate external electric fields that temporarily permeabilize cell membranes. Pore formation at the molecular level is achieved in a timescale of nano- to microseconds by infiltration of water molecules into the bilayer, which leads to reorientation of the adjacent lipids with their hydrophilic head groups pointing towards the water molecules. Once the external electric field is discontinued, pore resealing occurs and it is usually completed only within seconds or minutes after the end of exposure.[38] Using this approach, the group of Selenko demonstrated in 2016 the feasibility of electroporation to observe isotopically labelled α-synuclein in various mammalian cell lines by NMR spectroscopy.[39,40]

The approaches for eukaryotic *in-cell* NMR experiments outlined above depend upon the heterologous expression of proteins in bacterial cells. Instinctively, a protein is best expressed in its natural cellular environment, where the entire machinery required for synthesis, folding, maturation and targeting is present. Conceptually different from *E. coli* expression and protein delivery, a eukaryotic expression system thus represents the ideal choice to support structural and dynamic studies of eukaryote-specific proteins.[41] In this context, Bertrand and colleagues reported the observation of *in-cell* NMR spectra of *S. cerevisiae* and human ubiquitin following overexpression under the control of a methanol-induced AOX1 promoter in the yeast *Pichia pastoris*.[42] Another strategy for eukaryotic *in-cell* NMR using the baculovirus expression system in insect cells was also later described.[43] Furthermore, the research group of Banci exploited the recently developed mammalian cell-based expression systems, which rely on transient transfection with vectors designed for high-level protein production, to record NMR spectra of in living human cells.[44,45] An overview of the possible approaches discussed so far for *in-cell* NMR experiments is illustrated in Figure 1.1.

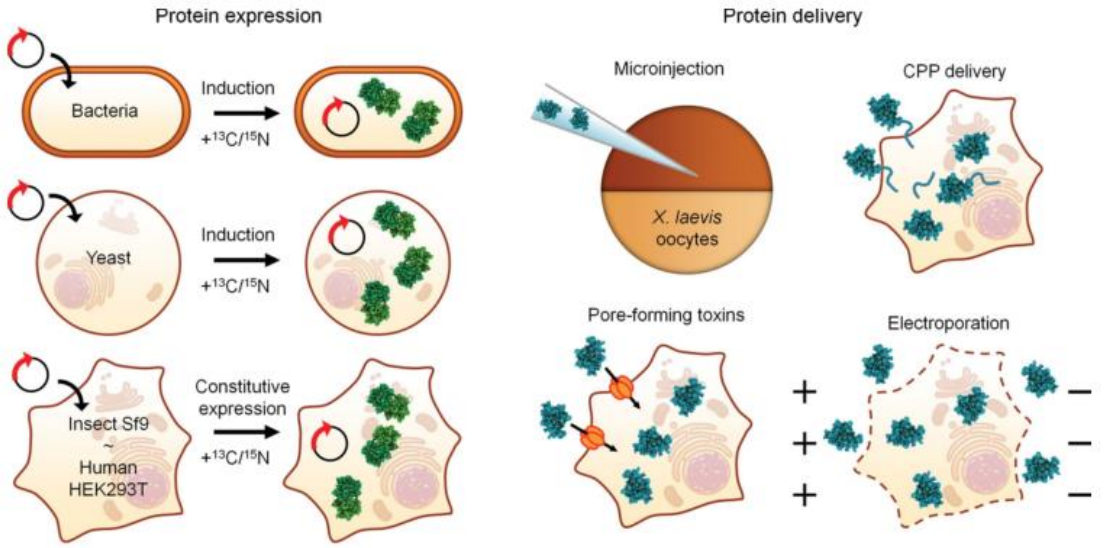

Figure 1.1 – Overview of possible methodologies for *in-cell* NMR experiments. The expression and isotopic labelling of endogenous proteins can be accomplished in distinct biological systems, such as bacteria, yeast, insect and mammalian cells. The introduction of exogenous proteins can be achieved through microinjection into *Xenopus laevis* oocytes or to human cells by cell-penetrating peptides (CPPs), pore-forming toxins or electroporation.[4]

Labelling techniques

Isotopic enrichment of biological macromolecules can be achieved by uniform or selective labelling. The specific labelling strategy will depend on the macromolecule under study, as well as the type of cells that are employed.

In the most frequently used scheme, the target protein is overexpressed from *E. coli* in an isotopically enriched minimal media containing $^{15}NH_4Cl$ and/or ^{13}C-glucose as the only sources of nitrogen and carbon, thus becoming uniformly labelled.[46] In addition to being cost-effective, the production of proteins in prokaryotic expression systems such as *E. coli* occurs without the post-translational modifications that are usually present in polypeptides derived from eukaryotes. Once delivered into eukaryotic cells, the residue-specificity and kinetics of post-translational modifications can be followed by time-resolved NMR spectroscopy.[47] An alternative strategy consists of selective isotopic labelling. This is achieved by supplementing the minimal-expression media with labelled amino acids that are located at the end of biosynthetic pathways, such as lysine, arginine and histidine. Due to less extensive labelling, the resulting spectrum naturally exhibits a lower resolution than if it had been obtained for a uniformly enriched target, but is basically devoid of background signals.[46] The possibility of using auxotrophic *E. coli* host strains, which require an exogenous source of specific amino acids for their growth, allows the selective labelling of the remaining amino acids.[48] However, auxotrophic bacterial strains often present with diminished expression rates, which reduces the quality of *in-cell* NMR spectra.

In theory, uniform ^{13}C [U-^{13}C] enrichment can provide advantages over uniform ^{15}N [U-^{15}N] labelling, largely because the gyromagnetic ratio of ^{13}C is 2.5 times higher than ^{15}N, thus enhancing the intrinsic sensitivity of the former nucleus to NMR signal detection.[49] However, in practice, U-^{15}N labelling is mostly favoured due to lower sample preparation costs and natural abundance. Furthermore, the chemical shifts associated with ^{13}C are confined to a narrower range than ^{15}N and ^{1}H-^{15}N HSQC spectra are easier to interpret than their ^{13}C counterparts. Selective ^{13}C-isotopic labelling of individual amino acids, particularly their methyl and methylene side chain residues, constitutes a more suitable strategy for enhancing sensitivity and minimizing background signals due to unwanted metabolic reactions.[46]

In NMR spectroscopy, the lifetime of the excited state is significantly influenced by the rate of molecular tumbling. As molecules become larger, the overall random tumbling motion begins to slow and the transverse relaxation rate increases. Since the linewidth of NMR signals is inversely proportional to the transverse relaxation rate, the spectra of macromolecules are characterized not only by increased complexity, but also display broader resonances. Moreover, biomolecular NMR experiments frequently rely on coherence transfer through scalar or dipolar couplings between nuclei. Because transfer steps are subjected to transverse relaxation, pulse sequences that use prolonged delays or require multiple transfer steps become less sensitive for larger biomolecules.[50]

The relaxation rate of ^{1}H nuclei in proteins is primarily affected by the substantial number of dipolar interactions with neighbouring protons. For the ^{15}N and ^{13}C heteronuclei, however, the predominant relaxation pathway is the direct dipolar interactions with covalently-bound protons.[50] One popular, albeit expensive avenue for attenuating nuclear relaxation that arises for higher molecular weight proteins involves their overexpression in bacteria grown in deuterated media. Bacteria are sufficiently robust to grow in such media, although at a slower rate and producing lower yields of recombinant proteins.[51] Due to chemical exchange following H_2O-based purification and refolding, protons are re-incorporated at the amide labile sites. Thus, proteins that are deuterated at aliphatic and aromatic positions but otherwise protonated at the amide sites can be prepared and observed in a ^{1}H-^{15}N heteronuclear correlation spectrum. By reducing proton density, perdeuteration decreases the transverse relaxation rates of the remaining protons, which results in a narrowing of their ^{1}H signal linewidths. Furthermore, deuteration of aliphatic sites greatly enhances the lifetime of transverse coherences which, in conjunction with ^{13}C and ^{15}N labelling, extends the size limit of proteins that can be studied by three-dimensional (3D) or four-dimensional (4D) NMR experiments.[50]

However, deuteration removes many of the protons that are necessary for structural and dynamic characterization by nuclear Overhauser effect (NOE)-based distance restraints and two-dimensional (2D) ^{1}H-^{13}C NMR experiments of larger proteins. One ingenious labelling strategy for overcoming this constraint and extending even further the size limit of biomolecules that can be studied by NMR consists on selectively incorporating isotopically labelled methyl groups into proteins. Each methyl group contains three protons that rapidly rotate about its axis, which increases the sensitivity of detection compared to the amide proton due to the three-fold

chemical shift degeneracy and longer transverse relaxation times.[52] Furthermore, protons in methyl groups do not chemically exchange with water as amide protons do, which can reduce and even broaden the resonance frequency intensity beyond detection. Therefore, methyl groups constitute powerful probes of structure and dynamics of high-molecular weight proteins, since they regularly occur in the hydrophobic core and can connect distant regions of primary structure.

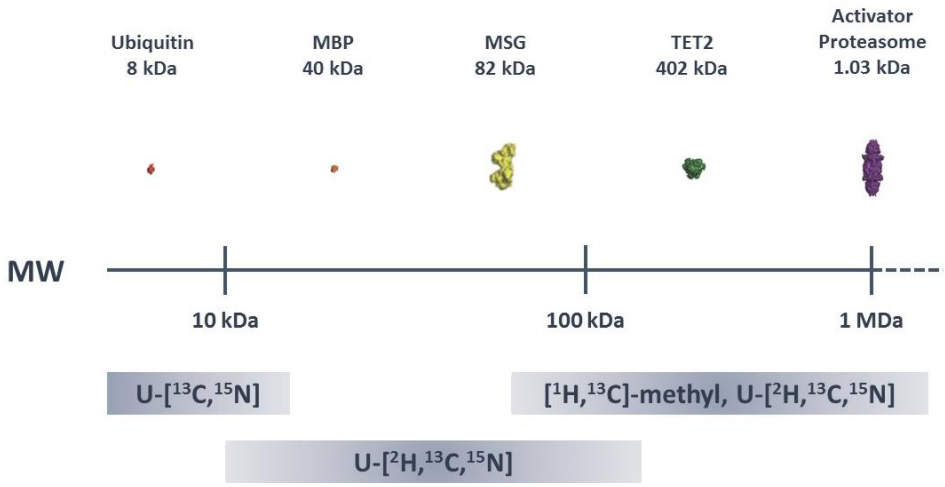

Figure 1.2 – The use of more complex isotopic labelling schemes allows the study of increasingly larger proteins. Structural representations are plotted as a function of molecular weight (MW) for ubiquitin, maltose-binding protein (MBP), malate synthase G (MSG), TET2 and proteasome activator complex. Adapted from *Plevin et al., 2012*.[50]

Excluding post-translational modifications, there are six-methyl containing amino acids in proteins. A variety of strategies for the selective labelling of these groups, while retaining the advantages of perdeuteration have been developed since the late nineties, most of which exploit the metabolic pathways involved in the biogenesis of methyl-containing amino acids. The simplest of these involves the overexpression of proteins in bacteria grown in deuterated media using protonated ^{13}C-pyruvate as the only source of carbon. Pyruvate is a direct precursor to alanine, valine, leucine and isoleucine and its methyl group is incorporated into these amino acids with minimal hydrogen exchange. Using this scheme, it was possible to obtain highly perdeuterated proteins that remain 40-80% protonated at the methyl groups of alanine, valine, leucine and isoleucine at the γ_2-position only.[53] However, due to proton exchange at the methyl group of ^{13}C-pyruvate with D_2O during protein overexpression, a mixture of proteins containing all four possible methyl isotopomers ($^{13}CH_3$, $^{13}CH_2D$, $^{13}CHD_2$ and $^{13}CD_3$) is obtained, which reduces the sensitivity of NMR spectra. Subsequently, a more selective ^{1}H,^{13}C-labelling strategy was developed for the production of deuterated proteins with protonation restricted to isoleucine at the δ_1 position and leucine/valine, respectively, by adding [3,3-$^{2}H_2$,U-^{13}C]-α-ketobutyric and [3,3-$^{2}H_2$,U-^{13}C]-α-ketoisovaleric acids to otherwise D_2O-based minimal medium one hour before induction of expression.[54,55] More recently, it was demonstrated that α-ketoisovaleric acid in which only one methyl group is ^{1}H,^{13}C-labelled allows a more efficient transfer of magnetization

from the methyl probe to the backbone and thus is more suitable for NMR studies of larger proteins.[56]

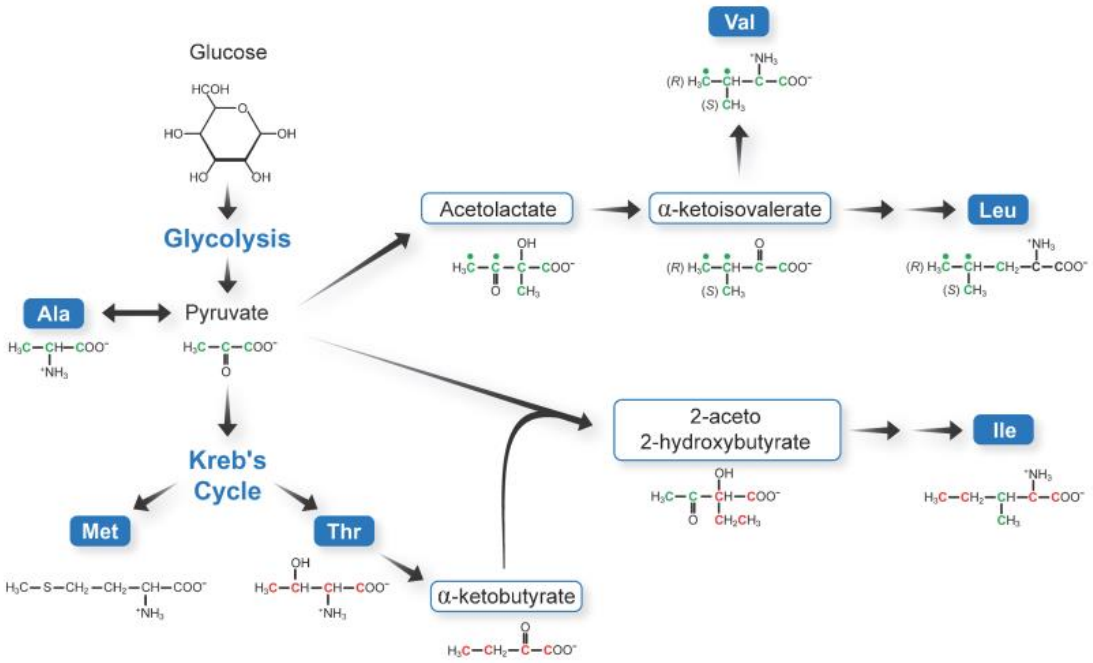

Figure 1.3 – The specific labelling of methyl-group containing amino acids exploits the metabolic pathways involved in their biosynthesis. Pyruvate serves as a precursor to the methyl-containing amino acids alanine, valine, leucine and isoleucine. Alanine is synthesized directly from pyruvate by a transamination reaction. Acetolactate, a precursor to valine and leucine, is obtained from the condensation reaction of two pyruvate molecules. Threonine is converted by a deaminase into α-ketobutyrate, which is then combined with pyruvate to yield isoleucine. The δ_1-methyl group of isoleucine is derived from α-ketobutyrate, while pyruvate originates the methyl group at the γ_2-position.[50]

In addition to ^{15}N and ^{13}C, ^{19}F has also proved to be a valuable probe in the study of proteins. Due to its high gyromagnetic ratio, 83% relative to the proton, and large chemical shift dispersion, which approaches 1000 ppm, this isotope constitutes a very sensitive reporter of structural and conformational changes.[57,58] Moreover, since the ½-spin ^{19}F nucleus is the only stable isotope of fluorine, with an abundance of 100%, and is not present in naturally occurring biomolecules, the resulting NMR spectra is virtually devoid of background signals. Accordingly, 1D ^{19}F NMR experiments generally require lower concentrations and shorter acquisitions times than multi-dimensional NMR methods and, therefore, proteins can be more easily studied near physiological conditions.[58] Also, conveniently enough, the introduction of a single fluorine in exchange for hydrogen at an aromatic side chain site does not lead to meaningful biological or structural changes.[58,59]

A wide variety of methods exist for incorporating fluorine into the protein of interest. Even though chemical synthesis is problematic for any protein of considerable size, several have been labelled using a “semi-synthetic” approach. In this method, a peptide fragment containing a fluorinated amino acid is chemically synthetized and then assembled with the remainder of the protein produced by recombination.[60] Fluorinated amino acids can also be incorporated

following protein expression in growth media containing the desired amino acid analogue. For instance, proteins with fluorinated aromatic amino acids can be easily prepared by either using auxotrophic strains or adding glyphosate, an inhibitor of aromatic amino acid biosynthesis.[61,62]

The use of natural translational machinery to introduce fluorinated analogues of amino acids seldom approaches an efficiency of 95% at a specific location. Moreover, this approach modifies every position of one amino acid simultaneously with varying efficiencies, which can produce protein molecules with heterogeneous structural perturbations. An alternative method for site-specific incorporation of ^{19}F-labelled amino acids into proteins with high translational efficiency and fidelity relies upon the use of an orthogonal tRNA/aminoacyl-tRNA synthetase pair.[63] This technology is based on the degeneracy of three stop codons in mRNA: UAA (ochre), UGA (opal) and UAG (amber).[64] The amber codon is the less commonly used in E. coli (≈7%) and does not terminate the essential genes required for bacterial growth and survival. Moreover, some species do not use the amber codon to signal termination of polypeptide synthesis, but instead to introduce an amino acid.[65] This allowed the development of a genetically engineered translational system in which an exogenous $tRNA_{CUA}$ and its cognate aminoacyl-tRNA synthetase specifically recognise and incorporate an unnatural ^{19}F-labelled amino acid into proteins.[66]

1.1.2 Applications, limitations and future perspectives

Since its inception, *in-cell* NMR spectroscopy has progressively emerged as a promising methodology to investigate the structure and dynamics of biomolecules, particularly proteins, at atomic resolution in their natural environment.

Traditionally, most of the work in this field has focused on prokaryotic hosts such as *E. coli* to achieve recombinant protein expression, because it is easily the most affordable and well-established method. Using the *E. coli* overexpression *in-cell* design, interactions with cofactors and other proteins following sequential expression of independent induction systems, as well as pH determination, conformation and three-dimensional structures have been investigated. For instance, Burz and colleagues developed a protocol for mapping structural changes that occur following protein-protein interactions, which was designated by STINT-NMR. Bacterial cells transformed with two plasmids, each containing different inducible promoters to express the target and interactor proteins at separate time points, are first grown overnight on LB media. The culture is then transferred to [U-^{15}N] minimal medium and expression of the target protein, whose NMR structure is known, is induced with L-arabinose for a specific extent of time. Cells are then washed and placed in unlabelled medium, with the interactor now being produced following induction of expression with isopropyl-β-D-thiogalactopyranoside (IPTG).[67,68] This approach was used to characterize the interaction between ubiquitin (8.5 kDa) and two protein ligands, AUIM (4 kDa) and STAM2 (50 kDa), with results demonstrating chemical shift differences and line broadening between the *in-cell* ^{1}H-^{15}N HSQC spectra of free and bound ubiquitin.[68] A very similar procedure, SMILI-NMR, was developed by the same authors for the screening of

small compounds capable of weakening or enhancing interactions between two components within a biomolecular complex.[69]

In addition, the dependence of the chemical shift and linewidth on binding events has also been used to study protein interactions with cofactors in *E. coli* cells. As an example, Hubbard and colleagues investigated the ion binding preference of the signal transduction protein CheY (15 kDa) under physiological conditions in 2003. By comparing the *in vitro* spectra of the protein in the presence of different ions with the *in-cell* spectrum, the authors demonstrated that CheY preferentially binds to Mg^{2+} ions in the *E. coli* cytoplasm.[70] In the same year, the research group of Dötsch illustrated the relevance of *in-cell* NMR for determining the intracellular pH of *E. coli*. By determining the intensity ratios of $C^{\varepsilon 1}$-H and $C^{\delta 2}$-H cross-peaks in histidine residues, which depend on the protonation and tautomerization state of the imidazole ring, a value of 7.1±0.1 was determined for the cytoplasmic pH in *E. coli*.[71]

Finally, *in-cell* NMR in bacteria has also been employed to study the 3D structure of proteins at concentrations that approach the physiological ones. For instance, the structure of the metal binding protein TTHA1718 at a concentration of 3-4 mM was solved *de novo* in *E. coli* cells by Sakakibara and colleagues in 2009. Because the short *in-cell* sample lifetime does not allow sufficient structural information to be recorded using typical high-dimensionality NMR experiments, the authors relied on nonlinear sampling schemes and distinct labelling strategies, such as [U-^{13}C,^{15}N] and ^{13}C-methyl selective labelling for backbone and side chain assignment, respectively, to reduce measurement times. To improve the signal-to-noise ratio and ensure that only the data of viable cells were acquired, each of the six triple-resonance experiments was repeated several times with fresh samples and the data was later combined.[72] Similarly, the structure of the GB1 protein at a concentration of 250 µM in *E. coli* cells was also recently described in 2016.[73]

Despite being initially established in prokaryotic systems, the full potential of *in-cell* NMR as a method for bridging structure and function of biomolecules in their native environment can only be reached when higher eukaryotic cells become routinely employed in such studies. Much progress has been recently made in this area with the development of different approaches for the expression and delivery of proteins into mammalian cells, including human.

In this context, residue-level information on cellular processes associated with protein modification and maturation have been obtained. For instance, Selenko and colleagues used time-resolved NMR to elucidate the stepwise sequence of phosphorylation events catalysed by casein kinase 2 (CK2) in *Xenopus* oocytes. Experiments were performed using a model substrate, $XT_{111\text{-}132}$GB1, composed of the regulatory protein sequence of the viral SV40 large T antigen fused to the GB1 protein, the latter functioning as a solubility-enhancement tag. The authors observed that CK2 phosphorylation of $XT_{111\text{-}132}$GB1 proceeds via a two-step process, with intermediate release of the mono-phosphorylated substrate.[74] Moreover, the research group of Banci used NMR experiments to follow the maturation events of superoxide dismutase 1 (SOD1) in human cells.[75] SOD1 is a 32 kDa homodimeric enzyme containing an intramolecular disulfide bond and a Cu^{2+}/Zn^{2+} site in each subunit, which catalyses the conversion of superoxide

radicals to hydrogen peroxide and is involved in the onset of the familial form of amyotrophic lateral sclerosis, a fatal neurodegenerative disease.[76] The addition of Zn^{2+} ions to the culture medium induced the dimerization of SOD1, while treatment with Cu^{2+} alone was not sufficient to produce the mature protein, because copper incorporation in SOD1 is tightly depends on the action of a specific chaperone. The role of this chaperone, the CSS protein, was further studied *in-cell*. When Cu^{2+} ions were added to cells co-expressing both SOD1 and CSS, full maturation was observed.[75] Together, these examples provide a brief look on the applicability of *in-cell* NMR to study cellular events.

In addition to proteins, *in-cell* NMR has also been extended to include nuclei acids. The study of these biomolecules inside a cellular environment is, however, challenging due to the underlying inability to overexpress and, therefore, label them. Furthermore, nucleic acids are extremely sensitive to the action of hydrolytic enzymes, such as DNase and RNase, that catalyse their cleavage inside living cells. These difficulties were overcome by Hänsel and colleagues in 2009 to produce the first *in-cell* NMR spectra of synthetically prepared [U-^{13}C,^{15}N]-labelled DNA and RNA sequences, which were chemically modified against enzymatic degradation and then microinjected into *Xenopus* oocytes.[77]

A favourable outcome of *in-cell* experiments depends on overcoming some obstacles. The relative insensitivity of NMR spectroscopy requires a substantial amount of the biomolecule of interest to be soluble inside cells. For most proteins, this is likely problematic because exceeding their physiological concentration, which is commonly on the order of nM to lower µM, can also conceivably affect their behaviour. The steady development of high-field magnets and cryogenic probes will increase the sensitivity of the technique and reduce the required concentration for acquiring NMR experiments with adequate spectral quality. Furthermore, the widespread use of selective labelling schemes based on methyl group detection will enable a further increment in sensitivity.[78]

Nevertheless, the most crucial factor that limits the length and type of experiments that can be recorded is sample viability. In a standard experimental setup, sedimentation and lysis with subsequent leakage of cytoplasmic constituents to the extracellular space naturally occurs within a certain time frame. To stabilize cells and minimize their degradation, some authors have suggested their encapsulation in embedding medium, such as low-melting agarose and Ficoll solutions.[79,80] However, prolonged cell viability invariably requires an exchange of nutrients and oxygen, while simultaneously removing the metabolic waste and maintaining the external pH within a narrow range. For this purpose, bioreactors compatible with conventional NMR instruments have been designed and employed for *in-cell* NMR in bacteria and human cells.[81,82]

Finally, the biggest challenge relies on the fact that a considerable number of globular proteins do not yield *in-cell* NMR spectra. Current evidence dictates these proteins interact with other cellular components and thus tumble so slowly that detection becomes impossible by solution-state NMR due to increased relaxation rates and subsequent broadening of the signals. Disordered proteins, which lack an ordered three-dimensional structure yet display functional roles in biological systems, are naturally more flexible and easier to detect by *in-cell* NMR than

globular domains of the same size. This was demonstrated by the research group of Pielak when a fusion protein constructed from the globular ubiquitin and the disordered α-synuclein was visualized within *E. coli* and subsequently in the cell lysate. The peaks associated with ubiquitin were absent in the *in-cell* spectrum, while those corresponding to α-synuclein were present. However, upon lysing the cells, well-defined peaks of both α-synuclein and ubiquitin were visible. These results indicate that globular and disordered proteins respond differently to macromolecular crowding effects within cells, due to distinct rotational diffusion dynamics.[83] Characterization of these globular proteins in their native cellular environment can be achieved instead by solid-state NMR, which is not inherently limited by the slow tumbling rate of molecules and is thus ideally suited for studying membrane proteins and large complexes or aggregates.

The role of macromolecular crowding on the structure and dynamics of proteins is more carefully considered in the following chapter.

1.2 The significance of macromolecular crowding for protein *in-cell* NMR spectroscopy

As indicated in the previous chapter, *in-cell* NMR can reveal with atomic resolution the structure and dynamics of proteins under their native conditions. Unfortunately, many globular proteins do not yield *in-cell* spectra, even those that are hailed as model systems for *in vitro* NMR studies, such as cytochrome c. The intracellular environment is inherently more complex than the dilute aqueous buffers that are conventionally employed to characterize proteins *in vitro* and the lack of *in-cell* spectra most likely reflects the differing degree to which the physicochemical properties of solutes inside cells influence the behaviour of proteins.

1.2.1 Basic concepts of macromolecular crowding

While a protein of interest may be dilute in the interior of a cell, it is naturally surrounded by a variety of other soluble macromolecules that together are present at high concentrations. For this reason, the intracellular compartment is usually described as highly crowded (Figure 1.4), with the distance between neighbouring macromolecules being of a similar magnitude as their diameter.[84,85] The *E. coli* cytoplasm contains approximately 400 g/L of proteins and RNA, while mammalian cells can have protein and nucleic acid concentrations ranging from 70-300 g/L depending on the cell type.[86] Therefore, prokaryotes exhibit higher concentrations of biological macromolecules than eukaryotes.

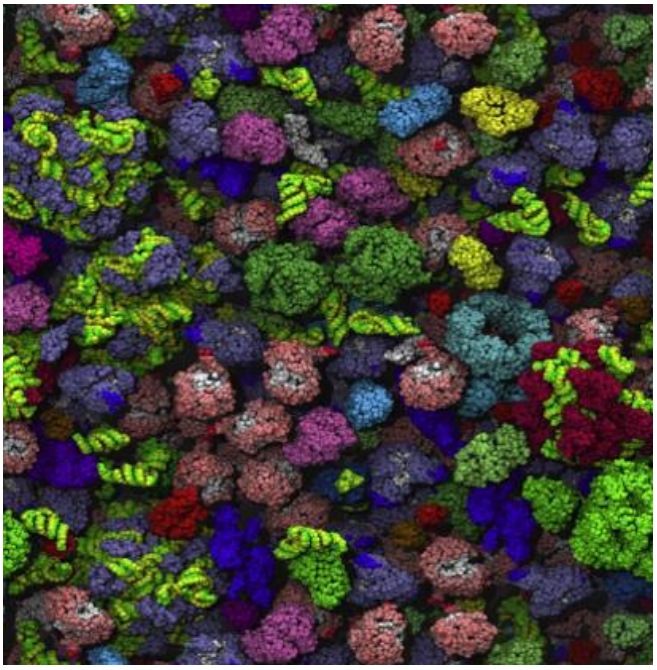

Figure 1.4 **– Schematic representation of the crowded *Escherichia coli* cytoplasm.** This figure was prepared following computer dynamics simulation of the crowded *E. coli* intracellular environment.[85]

Crowding may affect the thermodynamic and kinetic behaviour of proteins by distinct mechanisms, including excluded volume and intermolecular interactions.

The excluded volume effect intuitively arises by realizing that two molecules can never occupy the same space in solution. Because of this steric hindrance, the fraction of volume that a macromolecule can occupy is significantly lower than the total volume of the cell and the randomness of motion and entropy are remarkably reduced. Therefore, the excluded-volume effect increases the free energy and thermodynamic activity of a solute. Accordingly, due to Le Chatelier's principle it should be expected that volume exclusion promotes a more compact state.[87] Indeed, the direct influence of crowding on protein folding has been demonstrated for the small ribosomal protein S16. This study showed that the urea-unfolded ensemble of S16 is more compact, with faster folding rates and slower unfolding kinetics, in the presence of an artificial crowder.[88] Moreover, the effect of volume exclusion on the conformational properties of proteins has been investigated within cells by the group of Pielak using FlgM. This intrinsically disordered protein inhibits the transcription factor σ^{28} which regulates genes essential for flagellar assembly. FlgM is only transiently structured in dilute solutions, but its C-terminal region forms an α-helical structure at biological relevant solute concentrations and within the intracellular compartment.[89] However, disorder in its N-terminal portion remains regardless of macromolecular crowding and binding to its functional partner.[90]

Traditionally, the effects of macromolecular crowding on the behaviour of biomolecules were empirically investigated by using concentrated solutions of uncharged synthetic polymers, such as polyethylene glycol (PEG), polyvinylpyrrolidone (PVP), Ficoll and dextran. These artificial

crowding agents, while capable of interacting weakly with proteins through non-specific van der Waals and hydrophobic forces, mostly mimic the intracellular environment by imposing constraints on the volume that biomolecules can occupy.[91] In this context, the reduction of accessible volume has been shown to stabilize proteins, increase their enzymatic activity and accelerate their aggregation.[92,93]

Although volume exclusion constitutes a validated experimental paradigm for describing the effects of crowding on many aspects of protein behaviour, evidence has demonstrated that other factors should also be considered. The intracellular environment is not solely crowded, but extremely inhomogeneous and compartmentalized into membrane-less dynamic assemblies of components in which diffusive molecular transport with the surrounding cytoplasm occurs.[94] Since macromolecular crowding increases viscosity at distinct length scales within cells, the free diffusion of proteins is not only decreased but displays anomalous characteristics as well. It has been demonstrated using computer simulations that subdiffusion is associated with increased probability of finding a neighbouring target, thus enhancing cellular events that depend on the interaction between many functional partners in comparison with normal diffusion.[95]

Several techniques have been used to reveal the anomalous diffusion of macromolecules within crowded environments, including fluorescence spectroscopy and NMR. As an example, the diffusion of the chymotrypsin inhibitor 2 (CI2, 7.4 kDa) protein has been analysed by NMR-based methods in solutions of glycerol, synthetic uncharged polymers and proteins, such as bovine serum albumin (BSA, 66.5 kDa) and lysozyme (14.3 kDa). In glycerol, the decrease of both translational and rotational diffusion with increasing viscosity followed the Stokes-Einstein law ($D_t=\kappa T/6\pi\eta r$) and the Stokes-Einstein-Debye law ($D_r=\kappa T/8\pi\eta r^3$), respectively (where D_t is the translational diffusion coefficient, D_r is the rotational diffusion coefficient, η is the solution viscosity, κ is the Boltzmann constant, and r is the radius of protein being studied). However, meaningful deviations from the expected behaviour were induced by macromolecular crowding. Synthetic polymers as crowding agents caused a negative deviation from the Stokes laws, which means that increased viscosity reduces diffusion less than expected, and affects translation more than rotation. On the other hand, protein crowders were shown to have the opposite effect, attenuating rotational more than translational diffusion and inducing a positive deviation from the Stokes laws. NMR relaxation experiments revealed that the origin of these observable differences in diffusion behaviour were due to weak and nonspecific interactions between the protein crowders and CI2. Because synthetic polymers and globular proteins induce such dissimilar effects on diffusion, the former might not constitute the best alternative for modelling the effects of intracellular milieu on the behaviour of proteins. Besides, since CI2 is invisible in ^{1}H-^{15}N HSQC spectra from *in-cell* samples, as well as those containing high concentrations of proteins, these results suggest that interactions between proteins explain some of the difficulties in obtaining *in-cell* NMR data from globular proteins.[96]

Unlike excluded volume effects, which represent purely entropic hard-core repulsions, interactions between cytoplasmic components and proteins can be either attractive or repulsive. The repulsive interactions are stabilizing because they promote a more compact state

and thus exacerbate the excluded volume effect. On the contrary, enthalpy-driven attractive interactions are destabilizing since they increase surface exposure and favour unfolding.[97] This was shown when the CI2 protein was resuspended in PVP, BSA and lysozyme. Using NMR-detected amide proton-deuterium exchange, protein stability was then evaluated at residue level. It was observed that while PVP stabilizes CI2, the protein crowders have a destabilizing effect on most residues.[98] Moreover, another study found that *E. coli* lysate, a model crowding agent, destabilizes CI2 in a concentration-dependent manner, which strongly indicates an attractive interaction between the cytosolic components and the protein that is even capable of overcoming the stabilizing entropy-driven repulsions arising from excluded volume effects. In addition, it was noticed that *E. coli* lysates exerted an even stronger destabilizing effect on CI2 than the homogenous protein crowders, which strengthens the biological implications of weak, nonspecific interactions.[99] In conclusion, interactions between proteins can enhance or diminish the excluded volume effects.

Presumably, the chemical origin responsible for the destabilizing effects of physiological crowders arises from a variety of sources, such as complementary charge-charge interactions, hydrogen bonding between amide and carbonyl groups, and hydrophobic forces. To better understand the chemical nature behind protein-protein interactions within the intracellular environment, the research group of Pielak assessed the stability of a mutant variant of CI2 with an isoelectric point (pI) of 6.0 in anionic and cationic protein lysate fractions. Since the net charge of CI2 at neutral pH equals -1.0, the repulsion arising from negative charge-charge interactions with the anionic fraction should favour the compact state and thus increase protein stability. However, using amide proton-deuterium exchange NMR experiments, only destabilization was observed. Furthermore, the anionic fraction is almost as destabilizing as the total protein lysate. The authors concluded that noncomplementary electrostatic interactions are not strong enough to offset the nonspecific attractive interactions between protein surfaces.[100]

1.2.2 The model protein GB1 for probing interactions within biological settings

The biological significance of transient attractive interactions in which proteins participate was first postulated in 1982 by McConkey, who then proposed the term “quinary structure” to designate a fifth level of protein complexity. The existence of quinary structures was put forward following the observation that evolutionarily distant protein homologs are indistinguishable from each other based on isoelectric point and molecular weight.[101] This is unexpected if the external surface of proteins is only required to be hydrophilic and suggests that the intracellular crowded environment is much more organized than previously thought. Because these soft interactions are present in fundamental cellular constituents, such as ribosomes and the cytoskeleton, studies have since attempted to characterize their chemical basis.

The structure and dynamics of the B1 immunoglobulin-binding domain of streptococcal protein G (GB1, 6.2 kDa, pI=4.8) have been extensively characterized *in vitro*, particularly by NMR spectroscopy, due to its small size and high solubility. GB1 undergoes a two-state folding

equilibrium and its 56 amino acids are arranged in a four-stranded β-sheet of two symmetrically disposed antiparallel β-hairpins connected by an α-helix [4β+α] (Figure 1.5).[102] Its tightly packed and buried hydrophobic core, with a central part comprising seven solvent inaccessible residues (Leu-5, Leu-7, Ala-26, Phe-30, Ala-34, Phe-52, Val-54), explains its exceptionally high thermal stability, with a reversible melting temperature of 87°C.[102,103]

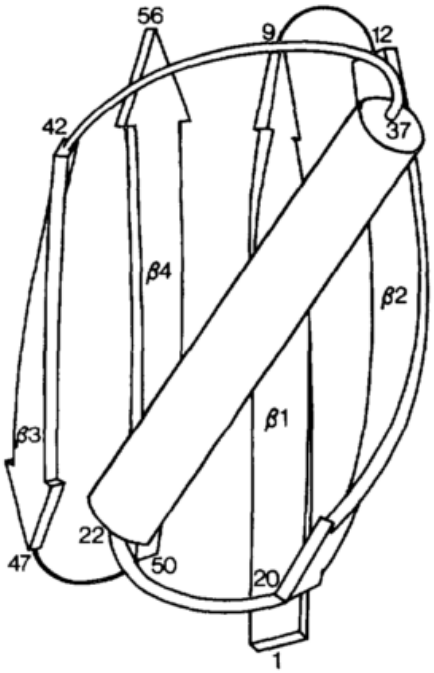

Figure 1.5 – Schematic cartoon diagram representing the folding and secondary elements of GB1. This protein is composed of 56 amino acids that are arranged two symmetrical anti-parallel β sheets connected by an α-helix.[102]

GB1 constitutes the prototypical model protein to investigate the potential of quinary interactions within living cells. Indeed, GB1 is one of the few globular proteins to yield an *in-cell* ^{1}H-^{15}N HSQC spectrum and since it is not intrinsic to either *E. coli* or eukaryotes, the probability of specific interactions with other cytoplasmic components is greatly reduced. So, rather than performing a functional role, this inert protein allows the monitoring of physicochemical properties involved in structure stabilization within biologically relevant environments without compromising the usefulness of NMR experiments due to binding events. For instance, Selenko and colleagues showed that chemical shifts in HSQC spectra of ^{15}N-labelled GB1 in *Xenopus* oocytes closely resembled those of the pure protein in dilute buffer, with some peaks exhibiting a distorted appearance. However, the average intensity of amide cross-peaks was significantly reduced, with a simultaneous increase in proton linewidth. Moreover, the amide groups that participate in hydrogen-bonding within the protein's secondary structure displayed a greater degree of signal reduction or linewidth broadening than those present in loop regions, suggesting that the more dynamic parts of GB1 are less disturbed by the increased intracellular viscosity.[104]

In most microorganisms, including *E. coli*, anionic proteins are significantly more abundant than cationic ones at physiological pH.[105] Because GB1 is negatively charged, the resulting electrostatic repulsions also facilitate the acquisition of high-quality *in-cell* data by overcoming the attractive interactions with other components, thus leading to a more compact state and overall stabilization. The contribution of electrostatic interactions towards protein quinary structure was demonstrated by the group of Pielak. Using NMR-based amide proton-deuterium exchange experiments, a decrease in pH of the *E. coli* cytoplasm from 7.4 to 5.0, which increases the number of proteins with a net positive charge, was correlated with reduced stabilization of the yet negatively charged GB1.[106] Furthermore, by monitoring the intracellular pH through mutagenesis of a lysine with an histidine at position 10 of the GB1 protein, another study evaluated the importance of quinary interactions on the quality of *in-cell* HSQC spectra. The decrease in intracellular pH to 5.0 severely reduces the number of observable cross-peaks due to line broadening. Since the latter is a consequence of faster transverse relaxation rates due to slower molecular tumbling, these results demonstrate the influence of electrostatic protein interactions for *in-cell* NMR experiments.[107]

In line with this rationale, the group of Pielak demonstrated that *E. coli* cytoplasm stabilizes GB1 in comparison with buffer alone at constant pH and temperature. However, when GB1 was resuspended in individual protein crowders, a contrasting behaviour was observed. In the presence of lysozyme, GB1 was destabilized to such an extent that retrieval of quantitative information was not possible. Since lysozyme displays a positive net charge at cytosolic pH, this result can be explained based on the existence of complementary electrostatic interactions with GB1, which lowers the free energy of the denatured state in regard to the folded ensemble. This interpretation does not account for the destabilization of GB1 with BSA, since the latter protein is also anionic at a physiological pH. In this scenario, the nonspecific attractive interactions between both proteins dominate the hard-core repulsions arising from charge-charge effects. Therefore, the choice of crowder markedly affects the type and strength of interactions in which proteins engage.[108] The ability of quinary interactions to modulate protein stability within cells was further supported by a study which described that the mutation of an aspartic acid for a lysine at the surface of the GB1 protein led to its substantial destabilization inside *E. coli*.[109]

1.3 Aims

The ability to characterize the behaviour of biomolecules within a physiologically relevant environment distinguishes NMR from other biophysical techniques. Due to its extraordinary stability and reduced capacity to interact in a specific manner with other cellular components, the GB1 protein represents the quintessential model to probe the physiochemical effects that are imposed by the local crowded environment on the structure and dynamics of proteins, without simultaneously compromising the ability to obtain *in-cell* NMR spectra due to binding events. Thus, in the present thesis, the experimental parameters critical for the success of *in-cell* NMR experiments, such as bacterial growth and protein overexpression, were initially

optimized. Subsequently, by monitoring proton and nitrogen chemical shifts of backbone amides and lysine side chains, as well as carbon and proton chemical shifts of side chains containing carbonyl groups, the preferential behaviour of GB1 was analysed in both *E. coli* lysate and cells, considering the pure protein in water as the reference state. Furthermore, interactions with the local environment were also assessed by determining the overall translational motion of the GB1 protein through diffusion-ordered NMR spectroscopy (DOSY). Finally, a comparison of diffusion coefficients obtained for GB1 with DOSY and fluorescence correlation spectroscopy (FCS), the standard analytical technique for studying protein diffusion, was made.

2. Methods

2.1 Molecular biology

2.1.1 Production of competent *Escherichia coli* cells

Escherichia coli (*E. coli*) BL21 (DE3) competent cells were prepared using an adapted rubidium chloride protocol.[110] In brief, 50 µL of *E. coli* BL21 (DE3) (*NZYTech*) cells were spread onto lysogeny broth (LB) solid agar medium (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl and 15 g/L agar) and incubated at 37°C overnight. A single colony from this plate was then inoculated in 5 mL of LB medium and incubated overnight with agitation at 37°C and 100 rpm. In the following day, 1 mL of starter culture was inoculated into 200 mL of sterile LB medium supplemented with 20 mM $MgSO_4$ and then incubated with agitation at 37°C and 200 rpm. When the optical density (OD) at 600 nm reached between 0.4-0.5, the culture was put on ice for 15 minutes. Cells were then collected by centrifugation (*Centrifuge 5804 R, Eppendorf*) at 4000g for 5 minutes at 4°C and the supernatant discarded. The pellet was resuspended in 60 mL of sterile-filtered TFB1 (30 mM KOAc, 100 mM RbCl, 10 mM $CaCl_2$, 50 mM $MnCl_2$ and 15% glycerol, with the pH adjusted to 5.8 with acetic acid). The cell suspension was incubated for 90 minutes on ice and then centrifuged (*Centrifuge 5804 R, Eppendorf*) at 4000g for 5 minutes at 4°C. The supernatant was discarded and the pellet resuspended in 8 mL of sterile-filtered TFB2 (10 mM MOPS, 75 mM $CaCl_2$, 10 mM RbCl and 15% glycerol, with the pH adjusted to 6.5 with KOH). Finally, this bacterial suspension was divided into 100-µL aliquots, flash frozen in liquid nitrogen and stored at -80°C.

2.1.2 Bacterial transformation

The pET11a expression vector containing the gene encoding the variant T2Q GB1 with resistance to ampicillin was generously provided by Professor Gary Pielak from the University of North Carolina at Chapell Hill.

The T2Q mutation prevents N-terminal deamidation and this variant is thus referred to as only "GB1".[108] The amino acid sequence of this protein is the following:

M**Q**YKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

The expression vector encoding the GB1 protein was transformed into *E. coli* BL21 (DE3) competent cells. For this purpose, 1 µL of the pET11a plasmid was added to 50 µL of E. coli cells and then incubated 15 minutes on ice. Cells were then incubated at 42°C for 40 seconds and transferred to ice. Following 15 minutes, cells were incubated with 500 µL of pre-warmed sterile LB medium (10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl) at 37°C for 1 hour with agitation. The resulting cell suspension was centrifuged at 100g for 3 minutes, the supernatant

removed and 100 µL of the pellet were spread onto LB agar plates containing ampicillin (100 µg/mL), which were left to incubate overnight at 37°C.

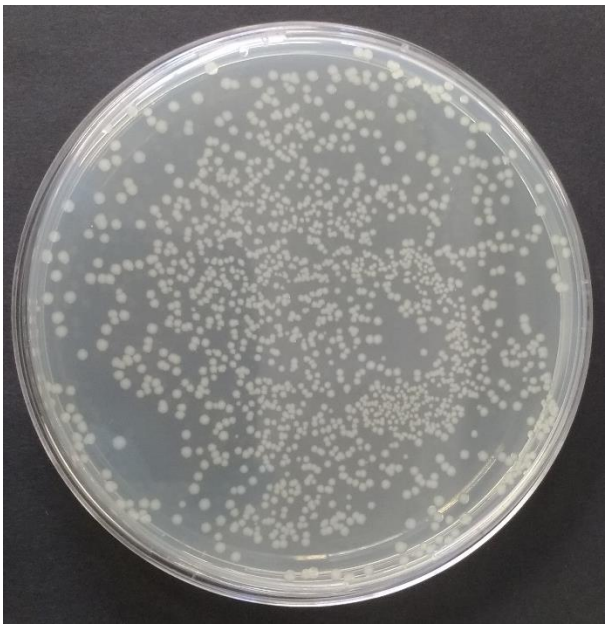

Figure 2.1 – Bacterial transformation with the GB1-encoding plasmid. Representative LB agar plate showing *E. coli* BL21 (DE3) cells transformed with the pET11a expression vector containing the GB1 gene with resistance to ampicillin.

2.2 NMR spectroscopy

2.2.1 *In vitro* NMR experiments

Expression and purification of isotopically labelled GB1

The expression and purification of isotopically labelled GB1 was performed by a student at our group prior to the start of this thesis.[111] For this purpose, a single colony was used to inoculate a 20-mL culture in LB medium containing 100 µg/mL ampicillin, which was then incubated with shaking at 37°C and 200 rpm overnight. In the following day, each 20-mL culture was transferred into 500 mL of M9 minimal medium (50 mM Na_2HPO_4, 20 mM KH_2PO_4, 9 mM NaCl) supplemented with 2 mM $MgSO_4$, 100 µM $CaCl_2$, 100 µM $FeSO_4$, 10 mg/mL thiamine, 4 g/L ^{13}C-glucose, 2.5 g/L $^{15}NH_4Cl$ and 100 µg/mL ampicillin. This culture was left to incubate with shaking at 37 °C and 180 rpm until the OD reached 0.8 at 600 nm. Expression of GB1 was induced with 1 mM of IPTG for 3 hours, after which cells were harvested for 12 minutes at 6000 rpm and 4°C (*Avanti j-26 XPI, Beckman Coulter, JA-10 rotor*) and frozen at -20°C overnight. The cell pellet was resuspended in lysis buffer (10 mM Tris, 1 mM EDTA, pH 7.5) and heated to 80°C for 5 minutes. The lysed cells were cooled on ice and centrifuged for 30 minutes at 15500 rpm and 4°C (*Avanti j-26 XPI, Beckman Coulter, JA-25.50 rotor*). The supernatant was frozen with 30% glycerol and dialyzed against 20 mM Tris at pH 7.5 overnight (*Snake Skin 3.5 K MWCO, Thermo Scientific*). The supernatant was purified by anion exchange chromatography with a HiTrap Q HP (*GE Healthcare*) packed with a diethylaminoethyl cellulose resin on an AKTA fast protein liquid chromatograph (*GE Healthcare*). Buffer A (20 mM Tris, pH 7.5) was used to load the crude

samples onto the column and elute impurities. Buffer B (20 mM Tris, 1 M NaCl, pH 7.5) was used to create a linear gradient of 0-400 mM NaCl. Fractions containing the GB1 protein were concentrated with Amicon Centricon units with 3-kDa membranes (Amicon Ultra-15 3K MWCO, Millipore) at 5000 rpm and 4°C and then further purified by size exclusion chromatography (Superdex 75 10/300 GL, GE Healthcare) with a mobile phase composed of 20 mM potassium phosphate buffer and 50 mM NaCl (pH 6.0). Finally, the pure fractions were dialyzed against water and lyophilized.

Production of Escherichia coli lysate

A single colony from an agar plate was picked to inoculate 20 mL (2x) of LB broth with ampicillin (100 µg/mL). This culture was incubated overnight with agitation at 37°C and 200 rpm. In the following day, each 20 mL of the two starter cultures were placed into 180 mL of LB medium with ampicillin (100 µg/mL) and left to incubate for 9 hours at 37°C and 200 rpm. Cells were then harvested by centrifugation at 3500 rpm (*Avanti j-26 XPI, Beckman Coulter, JA-10 rotor*) for 30 minutes at 4°C and the pellet was then frozen at -20°C overnight.

The cell pellet was resuspended the next day in 30 mL of milli-Q H_2O supplemented with four tablets of protease inhibitor cocktails (*cOmplete ULTRA tablets, Mini, EDTA-free, EASYpack from Roche*). *E. coli* cells were then lysed by sonication (*UP 100H, Hielscher*) on ice for 10 minutes with a duty cycle of 1 minute off and 1 minute on at 80% amplitude. The supernatant was then collected by centrifugation at 8000 rpm (*Avanti j-26 XPI, Beckman Coulter, JA-25.50 rotor*) for 30 minutes at 4°C and lyophilized for 4 days. Approximately 560 mg of dry lysate per 400 mL of culture was obtained, resulting in a yield of 1.4 g/L.

Sample preparation

In vitro protein samples were prepared by dissolving the pure, isotopically labelled GB1 protein at a concentration of 1 mM in milli-Q water and 150 mg/ml of lysate. The pH in both samples was corrected to a final value of 7.3 prior to NMR experiments.

2.2.2 *In-cell* NMR experiments

Sample preparation

The general protocol for obtaining *in-cell* samples was based on the work of Barnes and Pielak.[112] Following transformation of competent *E. coli* cells with the GB1-encoding plasmid, a single colony was picked to inoculate 6 mL of LB medium with ampicillin (100 µg/mL). This culture was incubated overnight with agitation at 37°C and 225 rpm. In the next day, 5 mL of the starter culture was placed in 100 mL of M9 minimal medium (50 mM Na_2HPO_4, 20 mM KH_2PO_4, 9 mM NaCl) supplemented with 2 mM $MgSO_4$, 100 µM $CaCl_2$, 100 µM $FeSO_4$, 10 mg/mL thiamine,

4 g/L ^{13}C-glucose, 1 g/L $^{15}NH_4Cl$ and 100 µg/mL ampicillin. The culture was left to incubate with agitation at 37°C and 220 rpm until the OD reached 0.6-0.8 at 600 nm. GB1 expression was then induced by adding 1 mM of isopropyl β-D-thiogalactopyranoside (IPTG) to the bacterial culture medium, which remained in agitation at 37°C. Overexpression was interrupted at four different time-points (1h, 1h30, 2h and 3h) to optimize the ideal intracellular concentration of GB1 that yields a well-resolved *in-cell* NMR spectrum without the occurrence of protein leakage. Cells were harvested by centrifugation (*Centrifuge 5804 R, Eppendorf*) at 1600g for 20 minutes at 4°C. The supernatant was discarded and the pellet resuspended in 2 mL of M9 medium. Samples for *in-cell* NMR experiments contained resuspended cells in a 90:10 mixture of H_2O:D_2O and were placed in standard 5 mm tubes. Following NMR data acquisition, samples were centrifuged (*Hettich Mikro 120*) for 10 minutes at 2000g at room temperature. The resulting supernatant was removed to assess the occurrence of protein leakage.[112]

The intracellular presence of the GB1 protein at these four-expression time-points was further evaluated by SDS-PAGE. The cell pellet from a 50 mL culture was resuspended in 700 µL of lysis buffer (50 mM Tris and 150 mM NaCl, pH 8.0) containing protease inhibitors (*cOmplete ULTRA tablets, Mini, EDTA-free, EASYpack from Roche*) and sonicated (*UP100H, Hielscher*) for 1 minute with a duty cycle of 4s on and 2s off at 80% amplitude. The lysate was harvested by centrifugation (*Hettich Mikro 120*) at 9400 g for 10 minutes at room temperature.[112] Tricine-SDS-PAGE gels were prepared according to the protocol described by Schägger.[113] Samples containing 20 µL of lysate in loading buffer were incubated at 100°C for 10 minutes and applied to the wells, along with the protein marker (*NZYColour Protein Marker II, NZYTech*). Gels were then run at 80 mV and stained with Coomassie Blue.

Lastly, the viability of transformed *E. coli* cells before and after an NMR experiment lasting 2 hours was assessed by the traditional colony-forming unit assay. In this context, the initial cell suspension was serially diluted 7 times, with 100 µL of the respective 10^{-6} and 10^{-7} fractions from each condition plated onto LB agar media.

2.2.3 Data acquisition and processing

NMR experiments were recorded on a Bruker Avance III 600 MHz spectrometer equipped with a 5-mm inverse detection triple-resonance z-gradient cryogenic probe (TCI). Spectra were processed with TopSpin (version 3.5 from Bruker) and visualized with CCPN (version 2.4.2).[114]

Combined chemical shift perturbations

To evaluate the effects of macromolecular crowding on the behaviour of GB1, proton and nitrogen chemical shifts of amides and lysines, as well as carbon and proton chemical shifts of carbonyl groups, were monitored in water, lysate and within *E. coli* cells at 37°C. Proton chemical shifts in water were referenced against trimethylsilylpropanoic acid (TSP, 0 ppm). The resulting water peak was then used to calibrate the spectra from lysate and *E. coli* samples.

Amide ^{1}H-^{15}N chemical shifts of the protein in water and lysate of isotopically labelled GB1 were acquired with 8 scans in a matrix containing 2048 x 128 data points and a spectral width of 9615.39 Hz x 2311.08 Hz, using the *hsqcetfpf3gpsi2* pulse sequence from the Bruker library. Within *E. coli* cells, ^{1}H-^{15}N SOFAST-HMQC spectra of isotopically labelled GB1 were acquired with 32 scans in a matrix containing 2048 x 128 data points and a spectral width of 9615.39 Hz x 2311.08 Hz, using the *sfhmqcf3gpph* pulse sequence from the Bruker library. GB1 ^{1}H-^{15}N assignments were based on published NMR data denoted by the PDB identifier 2GB1.[115]

Lysine ^{1}H-^{15}N chemical shifts of the protein in water, lysate and *E. coli* were acquired with 8, 32 and 32 scans, respectively, in a matrix containing 2048 x 100 data points and a spectral width of 7692.31 Hz x 425.69 Hz using the H2CN pulse sequence. Instead of providing a direct correlation with the labile and pH-sensitive hydrogens of amine groups, the H2CN sequence is suitable for monitoring proton and nitrogen chemical shifts in neutral and basic conditions through the indirect transfer of magnetization from $^{1}H_{\varepsilon}\rightarrow{}^{13}C_{\varepsilon}\rightarrow{}^{15}N_{\zeta}$ of lysines (Figure 2.2).[116] GB1 ^{1}H-^{15}N assignments of lysines were based on published work.[117]

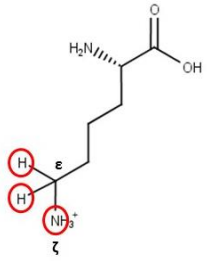

Figure 2.2 – Correlation between nuclei in the H2CN pulse sequence. This sequence allows the observation of lysine side chains at neutral and basic pH, by following the $^{1}H_{\varepsilon}$ and $^{15}N_{\zeta}$ chemical shifts through transfer of magnetization to $^{13}C_{\varepsilon}$.

Carbonyl ^{1}H-^{13}C chemical shifts of the protein in water, lysate and *E. coli* cells were acquired with 16, 8 and 16 scans, respectively, in a matrix containing 1024 x 1 x 96 data points and a spectral width of 8012.82 Hz x 4829.75 Hz x 1811.16 Hz, using the HCCO pulse sequence. The latter correlates the H_{β} of aspartate/asparagine and the H_{γ} of glutamate/glutamine with the carbonyl carbons of these residues (Figure 2.3). GB1 HCCO assignments were based on published work.[118]

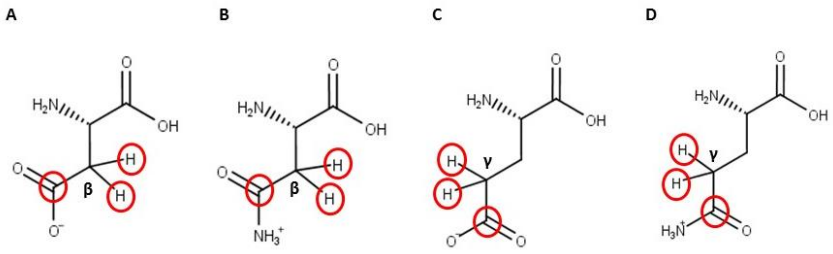

Figure 2.3 – Correlation between nuclei in the HCCO pulse sequence. This sequence allows the observation of (A) aspartate, (B) asparagine, (C) glutamate and (D) glutamine side chains at neutral pH, by following the $^{1}H_{\beta}/^{13}C_{\gamma}$ and $^{1}H_{\gamma}/^{13}C_{\delta}$ chemical shifts of these residues, respectively.

The preferential interactions of GB1 residues with the local environment in both lysate and *E. coli* cells, considering the pure protein in water as the reference state, were assessed by incorporating the chemical shift changes of the nuclei of interest to a sole quantity, the combined chemical shift perturbation ($\Delta\delta_{comb}$), according to the following equation:

$$\Delta\delta_{comb} = \sqrt{(\Delta\delta_H)^2 + (w_i\Delta\delta_{N\ or\ C})^2}$$

where $\Delta\delta_H$ and $\Delta\delta_{N\ or\ C}$ correspond to ^{1}H and ^{15}N or ^{13}C chemical shift differences, respectively, and w_i is a weighing factor which accounts for variations in sensitivity between nuclei. For chemical shifts expressed in parts per million (ppm), a reasonable estimate for determining w_i is given by the formula:

$$w_i = \frac{|\gamma_i|}{|\gamma_H|}$$

where γ_i and γ_H represents the magnetogyric ratios of a nucleus *i* and the proton, respectively. The cut-off value that distinguishes interacting from non-interacting residues was calculated in an iterative procedure as the corrected standard deviation to zero.[119]

DOSY NMR experiments

In addition to chemical shift perturbations, interactions with the local environment were also assessed by analysing the overall translational motion of the GB1 protein at 37°C through diffusion-ordered NMR spectroscopy (DOSY) using the *stebpgp* pulse sequence.[120] The latter scheme uses a stimulated echo and a longitudinal eddy current delay with bipolar gradient pulses for diffusion encoding and decoding. Spectra of GB1 in water, lysate and *E. coli* were recorded with 8, 32 and 16 scans in a matrix of 32 K points and a spectral width of 9615.385 Hz. The gradient strength was linearly incremented from 5% to 95% of its maximum value (53.2 $G.cm^{-1}$) in 32 steps, using a sine shaped smooth square (SMQ10.100) gradient with a 0.9 shape factor. The length of each bipolar gradient pulse, δ/2, was 800 µs in water, 1800 µs in lysate and 2450 µs in *E. coli*. The diffusion time, Δ, between each bipolar gradient pulse was set to 100 ms.

The attenuation in peak intensity by diffusion as a function of gradient amplitude was fitted using the *OriginPro8* software according to the following mono-exponential decay:

$$I = I_0\, e^{-4D(s2\pi\gamma G_i\delta)^2(\Delta+6\tau)}$$

where I_0 is the amplitude at zero gradient strength, γ represents the magnetogyric ratio of the proton (γ_H=4257.7 Hz G^{-1}), g and δ are the strength and duration of the bipolar gradient pulses, respectively, Δ is the diffusion time, τ is the gradient pulse recovery delay and *D* is the diffusion coefficient to be determined.[120]

The diffusion coefficient of water was determined using the *ledbpgp2s* pulse sequence from the Bruker library, which employs a stimulated echo and a longitudinal eddy current delay with bipolar gradient pulses for diffusion encoding and decoding.[121] Spectra of HDO in GB1

protein samples prepared in water and lysate were recorded with 8 scans in a matrix of 32 K points and a spectral width of 12019.23 Hz. The gradient strength was linearly incremented from 5% to 95% and 5% to 35% of its maximum value (53.2 G.cm^{-1}) in water and lysate, respectively, in 32 steps, using a sine shaped smooth square (SMQ10.100) gradient with a 0.9 shape factor. The length of each bipolar gradient pulse, δ/2, was 800 µs in water and 2000 µs in lysate. The diffusion time, Δ, between each bipolar gradient pulse was set to 100 ms.

The attenuation in water peak intensity by diffusion as a function of gradient amplitude fitted using the *OriginPro8* software according to the following mono-exponential decay:

$$I = I_0 \, e^{-D(s2\pi\gamma G_i\delta)^2(\Delta-\frac{\delta}{3})}$$

where I_0 is the amplitude at zero gradient strength, γ represents the magnetogyric ratio of the proton (γ_H=4257.7 Hz G^{-1}), g and δ are the strength and duration of the bipolar gradient pulses, respectively, Δ is the diffusion time and D is the diffusion coefficient to be extracted.[121]

For the comparative study of diffusion coefficients obtained with DOSY and FCS, NMR measurements were performed at 20°C with the fluorescently labelled ^{15}N-GB1 protein at a concentration of 0.37 mM. The *stebpgp* sequence using a stimulated echo and a longitudinal eddy current delay with bipolar gradients for diffusion encoding and decoding was employed.[120] Spectra of the GB1 protein in 0.1 M bicarbonate buffer and 150 mg/ml of *E. coli* lysate were recorded with 64 scans in a matrix of 32 K points and a spectral width of 9615.385 Hz. Gradient strength was linearly incremented in 32 steps from 5% to 95% of its maximum value (53.2 G.cm^{-1}) using a sine shaped smooth square gradient with a 0.9 shape factor (SMQ10.100). The length of each bipolar gradient pulse, δ/2, was 1500 µs and 1800 µs for diffusion measurements of the protein in buffer and 150 mg/ml of lysate, respectively. The diffusion time, Δ, between bipolar gradient pulses was set to 100 ms. The attenuation in peak intensity by diffusion as a function of gradient amplitude was fitted using the *OriginPro8* software according to the same mono-exponential decay presented previously.

The diffusion coefficient of water was determined using the *ledbpgp2s* pulse sequence from the Bruker library, which employs a stimulated echo and a longitudinal eddy current delay with bipolar gradient pulses for diffusion encoding and decoding.[121] Spectra of HDO in GB1 protein samples labelled with the fluorescent dye in buffer and lysate were recorded with 8 scans in a matrix of 32 K points and a spectral width of 12019.23 Hz. The gradient strength was linearly incremented from 5% to 40% of its maximum value (53.2 G.cm^{-1}) in 32 steps using a sine shaped smooth square (SMQ10.100) gradient with a 0.9 shape factor. The length of each bipolar gradient pulse, δ/2, was 1500 µs in buffer and lysate. The diffusion time, Δ, between each bipolar gradient pulse was set to 100 ms. The attenuation in water peak intensity by diffusion as a function of gradient amplitude was fitted using the *OriginPro8* software according to the same mono-exponential decay presented previously.

2.3 Fluorescence correlation spectroscopy

2.3.1 Sample preparation

The ^{15}N-labelled GB1 protein (5.24 mg; 0.83 µmol) was dissolved at 10 mg/ml in 0.1M bicarbonate buffer (98.86 mM $NaHCO_3$ and 1.14 mM Na_2CO_3, pH 8.3). The amine-reactive Alexa Fluor 488 carboxylic acid (1.0 mg; 1.55 µmol; *Molecular Probes, Invitrogen*) was dissolved in 100 µl of DMSO and slowly added to the solution containing the GB1 protein. The reaction mixture was left to incubate in the dark and at room temperature for 2 hours with continuous stirring. Subsequently, the conjugation was quenched with 100 µl of 1.5 M hydroxylamine (pH 8.5) and the reaction was left to incubate for 1 hour at room temperature. The fluorescently labelled ^{15}N-GB1 protein was purified with Amicon Centricon units with 3-kDa membranes (*Amicon Ultra-15 3K MWCO, Millipore*) at 5000 rpm and overnight dialysis in 0.1M bicarbonate buffer.

2.3.2 Data acquisition

FCS experiments

Fluorescence correlation spectroscopy (FCS) is a widely used technique for the analysis of physical parameters that are capable of producing intensity fluctuations when deviations from thermal equilibrium of small fluorescent ensembles occur.[122] One such parameter is diffusion. In this case, the intensity fluctuations arising from fluorescent molecules diffusing through the submicrometric detection volume of a confocal microscope are measured as a function of time and fitted to the following autocorrelation function $G(\tau)$:

$$G(\tau) = \frac{\langle \delta F(t) F(t+\tau) \rangle}{\langle F(t) \rangle^2}$$

which measures the self-similarity of the fluctuating signals diffusing through the small confocal volume after a lag time τ. Its value reflects the probability of the signal recorded at various times belonging to the same molecular ensemble. For instance, if the signal is detected after an extended lag period τ, the fluctuations at $t + \tau$ are likely caused by different particles than those at time t and thus the autocorrelation $G(\tau)$ approaches zero. In contrast, if the lag period τ is very narrow, virtually the same particles are responsible for generating the fluctuations at time t and $t + \tau$ and therefore the autocorrelation $G(\tau)$ is closer to 1.[123]

To determine diffusion coefficients with FCS, the dimensions of the confocal volume were calibrated first using a reference dye with a known diffusion coefficient. For this purpose, rhodamine 110, with a diffusion coefficient of 440 µm^{2}s^{-1} determined at 22.5°C, was chosen.[124] Considering the following relation between the diffusion time τ_D, which corresponds to the average residence time in the detection volume, and the diffusion coefficient D:

$$\tau_D = \frac{w_0^2}{4D}$$

the lateral size of the focus, w_0, can be extracted. Subsequently, this same equation can be used to calculate the diffusion coefficient of the ^{15}N-GB1 protein by substituting τ_D with the value retrieved from its autocorrelation curve.

Fluorescence measurements were performed at 20^{o}C using a Leica TCS SP5 (*Leica Microsystems CMS GmbH, Mannheim, Germany*) inverted confocal microscope (DMI6000). A 63x apochromatic water immersion objective with a numerical aperture of 1.2 (*Zeiss, Jena Germany*) was used in each experiment. Detection was carried out using avalanche photodiodes (APDs) in combination with a 500-550 band-pass filter. Excitation of the Alexa 488 labelled ^{15}N-GB1 (28.4 nM in milliQ water and 150 mg/ml of lysate) was performed with the 488-nm line of the argon laser. The dimensions of the confocal volume were determined by calibration with Rhodamine 110 (10 nM) in milliQ water using the same optical setup as the samples.

3. Results

3.1 *In-cell* NMR protocol optimization

As indicated previously in the methods chapter, the production of GB1 in *E. coli* was induced during four distinct time-periods (1h, 1h30, 2h and 3h) in minimal medium to assess the influence of overexpression on the quality of the resulting ^{1}H-^{15}N SOFAST-HMQC spectra (Figure 3.1).

It was observed that 1h after induction (Figure 3.1.A), most resonances belonging to GB1 protein are already visible, albeit weakly, with negligible background signals arising from ^{15}N incorporation into other cellular components or metabolic degradation. However, inspection of the remaining spectra (Figures 3.1.B, 3.1.C and 3.1.D) reveals that with increasing induction time, the protein cross-peaks become sharper and more well-defined.

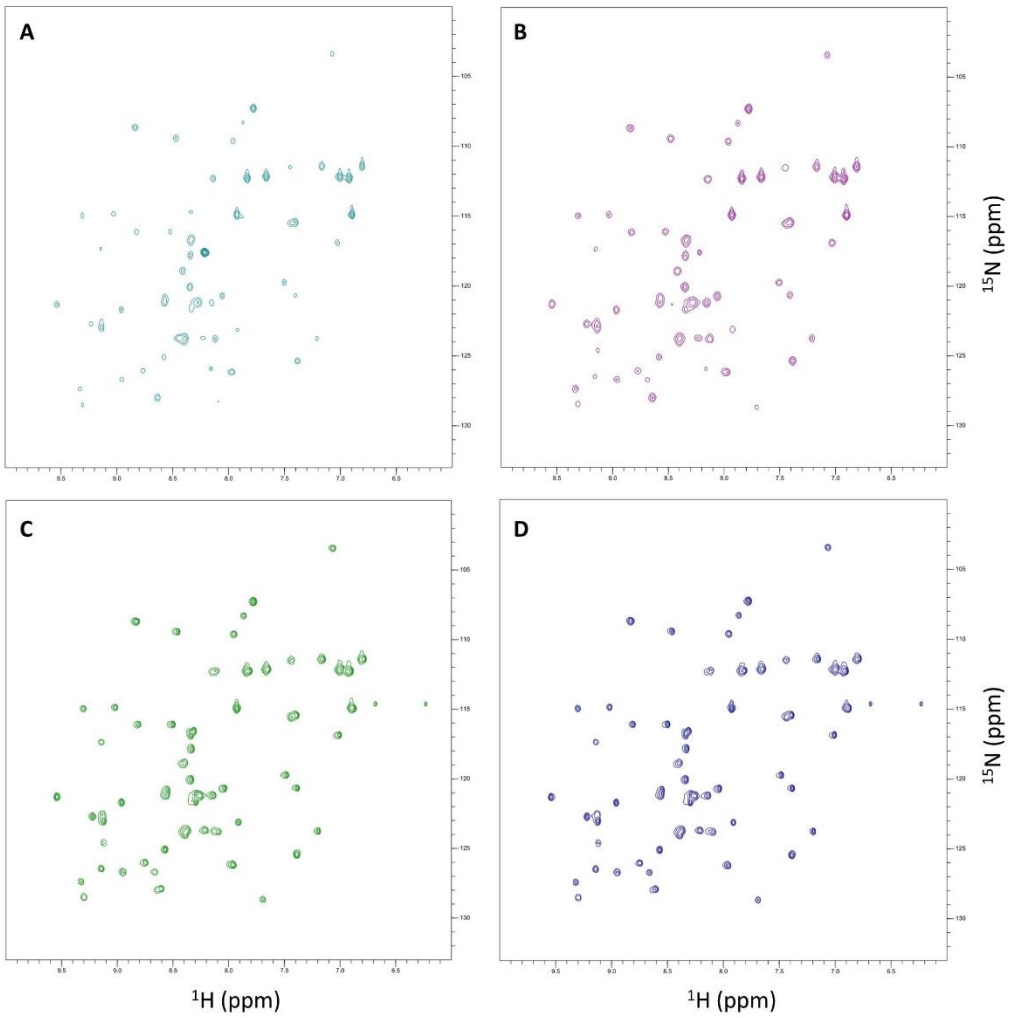

Figure 3.1 – SOFAST ^{1}H-^{15}N HMQC spectra from GB1-expressing *E. coli* acquired at four distinct time-points after induction of protein expression in ^{15}N-labelled minimal medium. (A) HMQC spectra of the *in-cell* slurry recorded after 1h, (B) 1h30, (C) 2h, and (D) 3h of induction.

The presence of the GB1 protein within *E. coli* cells was further confirmed by Tricine-SDS-PAGE of samples obtained following induction of protein expression. As indicated in Figure 3.2, this protein is also observed by electrophoresis after incubation of the culture medium with IPTG for 1h, 1h30, 2h and 3h (corresponding to lanes 2, 3, 4 and 5, respectively). In contrast, GB1 appears to be absent from samples in which protein expression was not induced with IPTG for 1h, 1h30, 2h and 3h (corresponding to lanes 7, 8, 9 and 10, respectively).

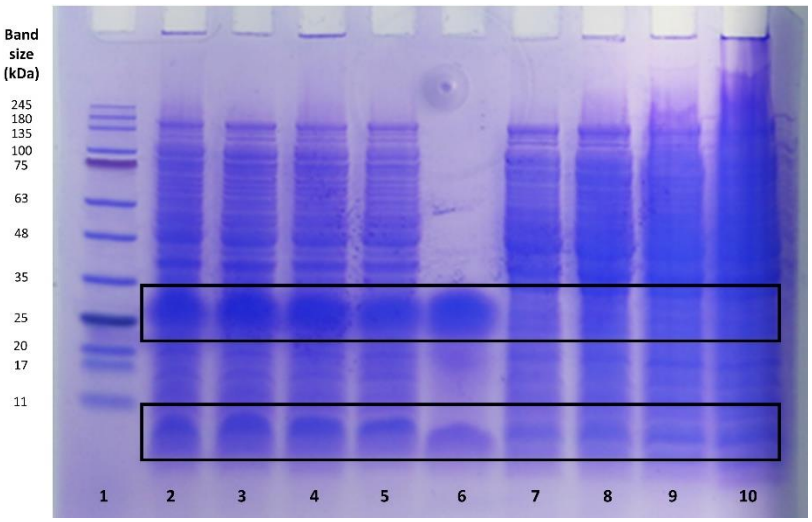

Figure 3.2 – Tricine-SDS-PAGE of samples from *in-cell* NMR experiments. Representative gel showing the presence (lanes 2, 3, 4 and 5) or absence (lanes 7, 8, 9 and 10) of the GB1 protein. The protein marker was loaded onto the first lane. Meanwhile, lanes 2, 3, 4 and 5 correspond to samples in which GB1 expression was induced for 1h, 1h30, 2h and 3h, respectively, following addition of IPTG to the culture medium. A pure GB1 standard at a concentration of 1 mM was applied to the sixth lane. Lastly, lanes 7, 8, 9 and 10 correspond to control samples in which GB1 expression was not induced with IPTG for 1h, 1h30, 2h and 3h, respectively.

To determine the contribution of expression levels on protein leakage, the *in-cell* samples were centrifuged and the resulting supernatant analysed by NMR. As illustrated in Figure 3.3.A, the GB1 cross-peaks are nearly absent in the supernatant of *in-cell* samples collected after 1h of induction. However, the three-remaining spectra contain progressively more GB1-specific peaks as the time frame allowing overexpression of this protein increases from 1h to 1h30 (Figure 3.3.B), 2h (Figure 3.3.C) and, finally, 3h (Figure 3.3.D). Indeed, following 2h and 3h of protein induction with IPTG, the number of GB1 ^{1}H-^{15}N peaks observed with *in-cell* and their respective supernatant samples is the same (Figure 3.4). This strongly suggests that the aforementioned peaks within the *in-cell* spectra at these two time-points do not correspond to the presence of the GB1 protein within cells, but instead in the extracellular environment. Moreover, a dramatic increase in the average volume per number of GB1 cross-peaks is observed in the supernatant of *in-cell* samples as protein induction develops from 1h30 to 3h (Figure 3.5). Thus, protein leakage from the cells to the surrounding media appears to occur when the overexpression level exceeds a certain limit.

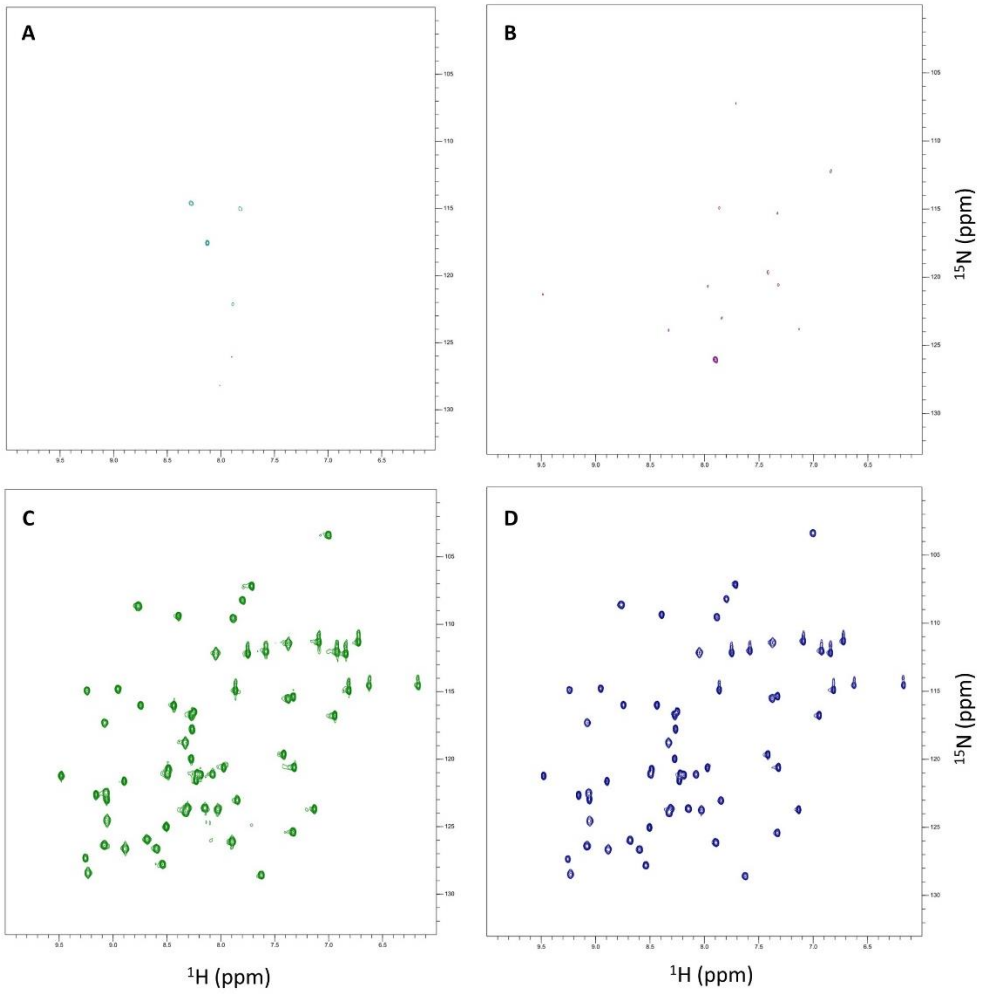

Figure 3.3 – SOFAST ^{1}H-^{15}N HMQC spectra from the supernatant of GB1-expressing *E. coli* acquired at four distinct time-points after induction of protein expression in ^{15}N-labelled minimal medium. (A) ^{1}H-^{15}N HMQC spectra of the supernatant from *in-cell* NMR samples recorded after 1h, (B) 1h30, (C) 2h, and (D) 3h of protein induction with IPTG.

Figure 3.4 – Number of GB1-specific ^{1}H-^{15}N cross-peaks of *in-cell* NMR samples and their respective supernatants acquired at four distinct expression time-periods. The number of ^{1}H-^{15}N cross-peaks observed in SOFAST ^{1}H-^{15}N HMQC spectra in the supernatant of *in-cell* NMR samples greatly increases as GB1 overexpression progresses from 1h30 to 3h post-induction with IPTG.

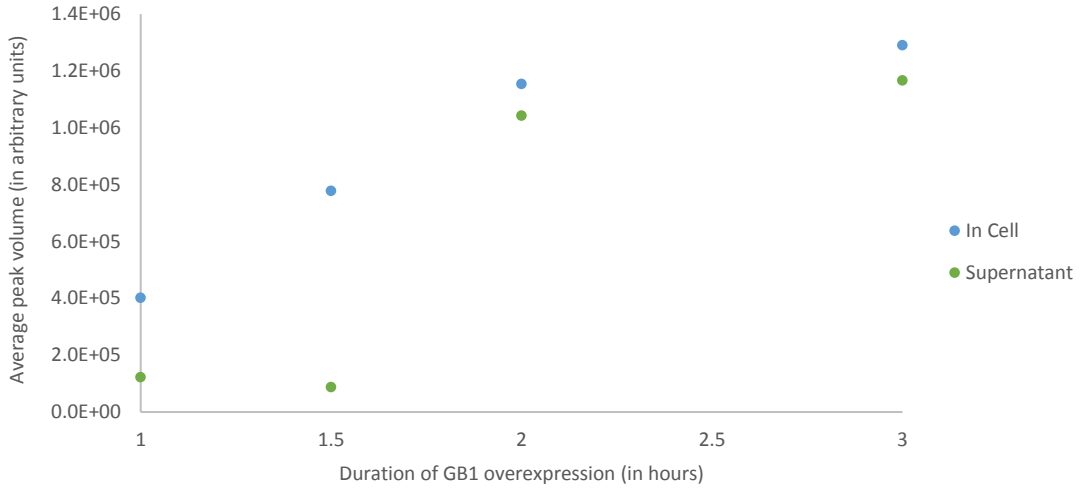

Figure 3.5 – Average peak volume of *in-cell* NMR samples and their respective supernatants acquired at four distinct expression time-periods. The volume per number of GB1-specific ^{15}N-^{1}H peaks observed in SOFAST-HMQC spectra in the supernatant of *in-cell* NMR samples significantly increases as protein induction with IPTG progresses from 1h30 to 3h.

Protein leakage can seriously undermine the usefulness of *in-cell* NMR experiments. Due to extensive differences in viscosity between the intra- and extracellular environments, even a small amount of protein present on the outside will give rise to cross-peaks that are considerably more intense than those originated within cells. In cases in which leakage does occur, the signals associated with a protein on the outside will dominate the resulting NMR spectrum, since cross-peaks of the same species within the intracellular milieu exhibit much lower intensities due to line broadening caused by cellular viscosity. Therefore, as a compromise between the spectral quality of specific GB1 ^{1}H-^{15}N cross-peaks and leakage with increasing expression levels, an intermediate induction time of 1h15 was selected to compare protein conformation and dynamics between three distinct conditions: water, lysate and *E. coli* cells. The presence of a still substantial number of viable bacterial cells following a 2-hour NMR experiment was confirmed by performing colony-forming unit assays, as illustrated in Figure 3.6.

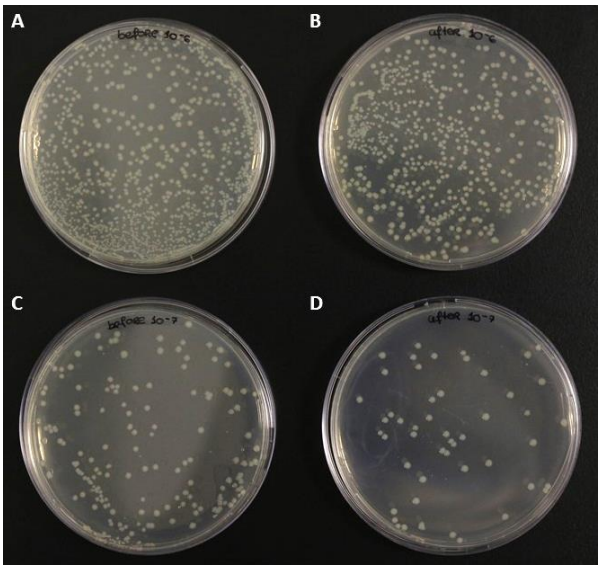

Figure 3.6 – Colony-forming unit assays before and after *in-cell* NMR experiments. Viability of *E. coli* following a 2-hour *in-cell* NMR experiment was assessed by colony-forming unit assays. For this purpose, the bacterial suspension was serially diluted seven times, with 100 µL of the 10^{-6} and 10^{-7} fractions corresponding to before and after *in-cell* NMR experiments plated onto solid media. (A) before and (B) after *in-cell* sample at a dilution of 10^{-6}. (C) before and (D) after *in-cell* sample at a dilution of 10^{-7}.

3.2 Study of GB1 backbone amide interactions in lysate and *E. coli*

To evaluate the effect of a crowded intracellular environment on the conformation of GB1, the combined ^{1}H-^{15}N chemical shift perturbations ($\Delta\delta_{comb}$) of backbone amide groups in this protein were compared in lysate and *E. coli* cells. In this context, as previously mentioned in section 2.2.3, specific ^{1}H-^{15}N chemical shift changes were mapped according to the following equation:

$$\Delta\delta_{comb} = \sqrt{(\Delta\delta_H)^2 + (w_i\Delta\delta_N)^2}$$

in which w_i represents the ratio between the magnetogyric ratios of nitrogen and proton and $\Delta\delta$ corresponds to chemical shift variations for each GB1 residue between two conditions.

Considering the pure protein in water as the reference state, a perturbation of every one of the 55-backbone amide groups that give rise to cross-peaks in ^{1}H-^{15}N spectra was observed (Figure 3.7) in the *E. coli* lysate. Specifically, the most affected residues by interactions with lysate molecules are K10 and T11 present in the more dynamic loop region that connects the β_1 and β_2 sheets. Additionally, residue L12 located at the beginning of the β_2 sheet exhibits the highest chemical shift perturbation in GB1. In contrast, the least perturbed residues generally correspond to those in which the backbone amide groups are involved in hydrogen bonding within the GB1's α-helix secondary structure. A more intuitive display of chemical shift perturbations as a function of secondary structural elements in GB1 is presented in Figure 3.8. As observed, residues belonging to the loop that connects β_1 and β_2 sheets exhibit the highest median of combined chemical shift perturbations (0.1025). On the contrary, the secondary structural element of GB1 with the lowest median of combined chemical shift changes is the α-

helix (0.06618). The spectra and respective assignments for the GB1 protein in water and lysate can be found in Appendix section 7.1.

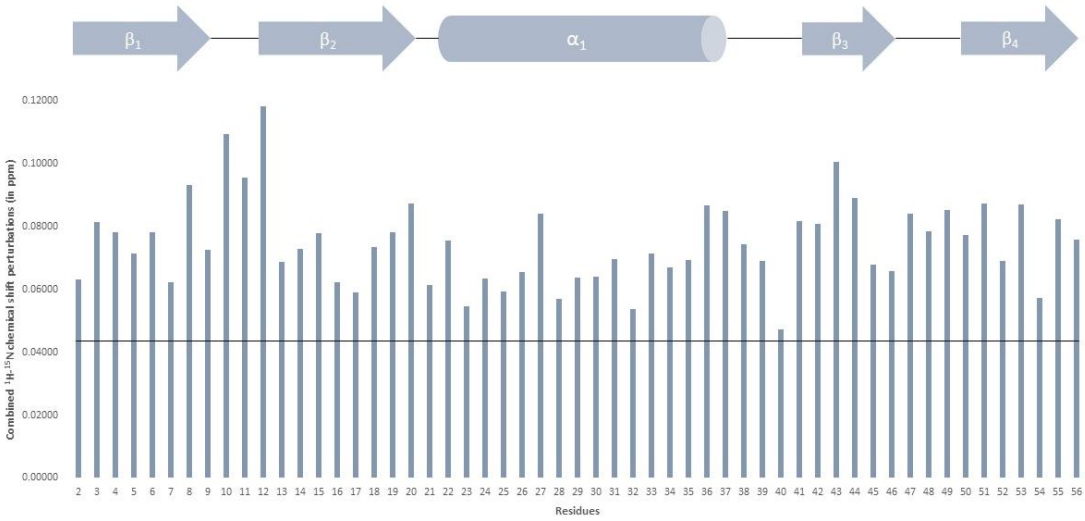

Figure 3.7 – Combined ^{1}H-^{15}N chemical shift perturbations of the GB1 protein in *E. coli* lysate. Residue-specific chemical shift changes of GB1 in lysate when considering the pure protein in water as the reference state. The black line corresponds to the chemical shift cut-off value (0.041), above which residues are classed as being affected by the differing lysate environment. Secondary structural elements of GB1 are shown at the top of the chart. Residues K10 in the first loop region and L12 located at the start of the β_2 sheet display the highest chemical shift perturbations when placed in lysate.

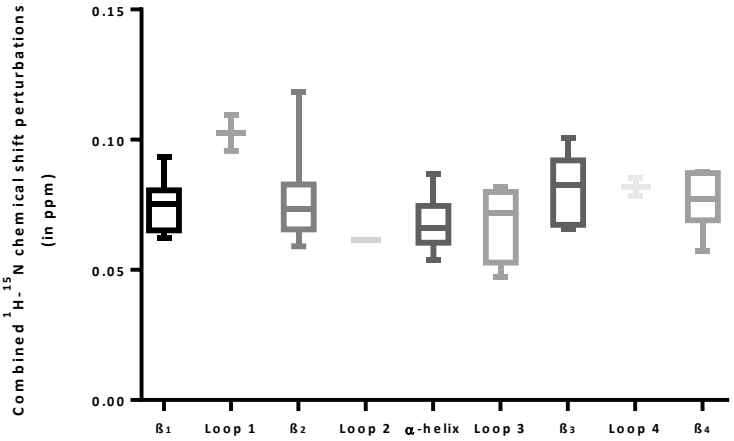

Figure 3.8 – Box-plot depicting the combined ^{1}H-^{15}N chemical shift perturbations in *E. coli* lysate as a function of secondary structural elements in GB1. The horizontal band inside the box represents the median of data in each group. As illustrated, the residues comprising the first loop, which connects β_1 and β_2 sheets, present with the highest median of combined chemical shift changes. Meanwhile, residues belonging to the α-helix exhibit the lowest median value of combined chemical shift perturbations.

Analysing the chemical shift perturbations of the GB1 protein within *E. coli* cells, again considering the pure protein in water as the reference state, preferential interactions with residues K10 and T11 in the first loop region and residues D40 and G41 in the third loop region were observed (Figure 3.9). In addition, residues N8, L12 and W43 in β_1, β_2 and β_3 sheets and residue N37 in the α-helix, which are close to the mentioned loop regions, exhibit a significant degree of perturbation within *E. coli* cells, with residue T55 located at the end of the β_4 sheet

displaying the highest combined chemical shift change. Residues Y33 and E42 tend to overlap with the surrounding ^{1}H-^{15}N cross-peaks and are difficult to assign with reasonable precision. However, the ^{1}H-^{15}N cross-peak corresponding to residue T25 seems to disappear when the protein is located within *E. coli* cells. The absence of residue F52 is related with the use of a band-selective ^{1}H excitation pulse (PC9) in the SOFAST-HMQC sequence, centred at 7.6 instead of 8.6 ppm, which resulted in a shorter pulse length than necessary to excite the proton of this amide.[126] Again, a box-plot representation of chemical shift changes as a function of secondary structural elements in GB1 is presented in Figure 3.10. As indicated, residues comprising the first loop display the highest median (0.0738) of combined ^{1}H-^{15}N chemical shift perturbations. Again, like the previous case of the GB1 protein in lysate, the secondary structural element of GB1 with the lowest median of combined chemical shift changes is the α-helix (0.0247). The spectra and respective ^{1}H-^{15}N backbone assignments for the GB1 protein in *E. coli* cells can be found in Appendix section 7.1.

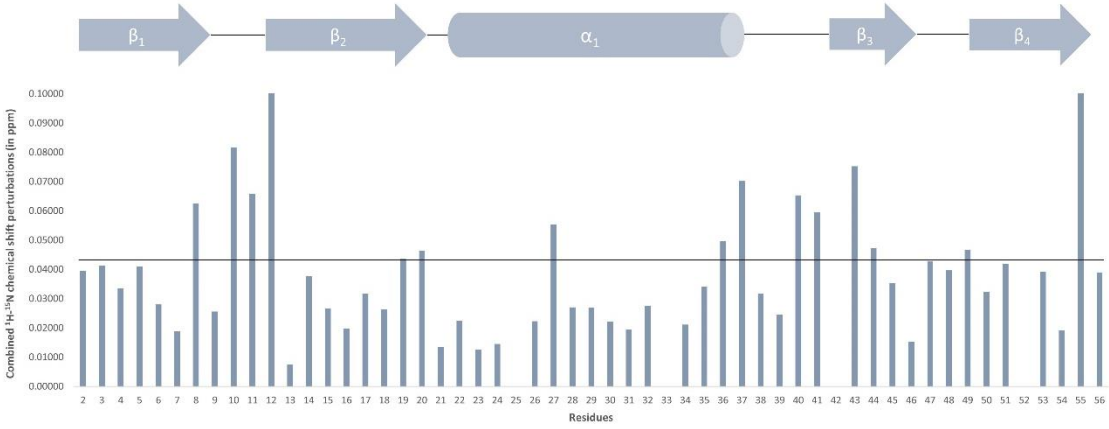

Figure 3.9 – Combined ^{1}H-^{15}N chemical shift perturbations of the GB1 protein in *E. coli* cells. Residue-specific chemical shift changes of GB1 in *E. coli* cells considering the pure protein in water as the reference state. The black line corresponds to the chemical shift cut-off value (0.043), above which residues are classed as being affected by the differing *E. coli* intracellular environment. Preferential interactions with residues N8, K10, T11, L12, E19, A20, E27, D36, N37, D40, G41, W43, T49 and T55 occur when GB1 is located within the *E. coli* cytoplasm. Secondary structural elements of the GB1 protein are shown at the top of the chart.

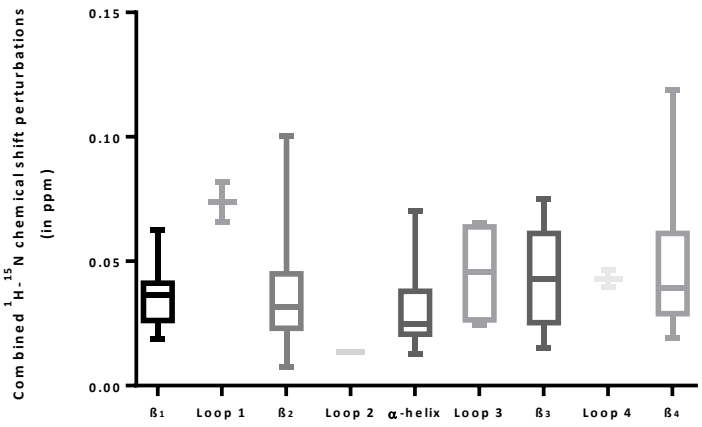

Figure 3.10 – Box-plot depicting the combined ^{1}H-^{15}N chemical shift perturbations in *E. coli* cells as a function of secondary structural elements in GB1. The horizontal band inside each box represents the median of data in each group. As illustrated, the residues comprising the first loop exhibit the highest median of combined chemical shift changes. Contrary to the when the protein is placed in lysate, residues comprising the second loop display the lowest median value of combined chemical shift perturbations when GB1 is overexpressed in *E. coli*.

The effect of a crowded environment on the linewidth and intensity of GB1 signals was also evaluated. As illustrated in Figure 3.11, a dramatic increase of 107.6% on the average value of proton linewidths retrieved from 2D ^{1}H-^{15}N *in-cell* experiments is observed for each backbone amide cross-peak of the GB1 protein, considering the pure protein in water as the reference state. Since GB1 displays marginal interaction with other cellular components, this reduction in rotational diffusion is likely due to increased intracellular viscosity. Furthermore, the most affected residues in this case appear to be those located within the α-helix secondary structural elements of GB1. However, only a slight increase of 13.4% in proton linewidths is evident when GB1 is placed in lysate. Thus, it is possible that the viscosity of the intracellular environment is much higher than that of the *E. coli* lysate. A more intuitive display of the differences in proton linewidths as a function of secondary structural elements in GB1 is presented in Figures 3.12 and 3.13. As observed, the residues comprising the α-helix of GB1 display the highest median of fractional change in proton linewidths in both lysate and *E. coli* cells (in lysate: 0.1295; in *E. coli*: 0.5530).

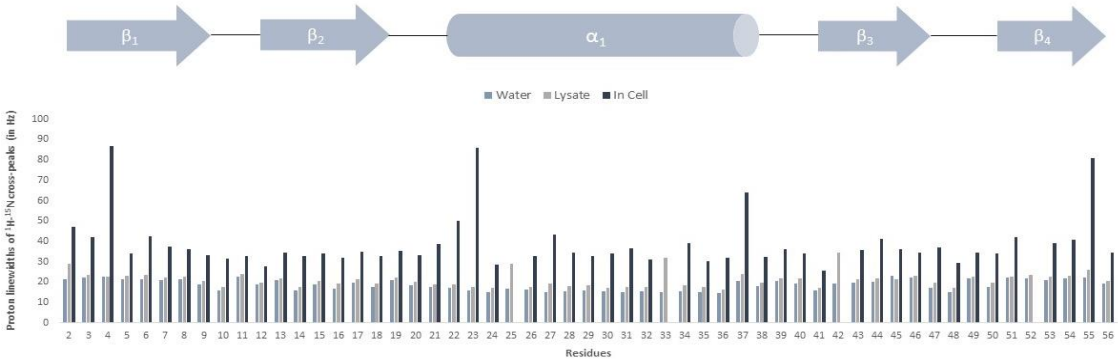

Figure 3.11 – Proton linewidths of ^{1}H-^{15}H cross-peaks in the spectra of the GB1 protein in water, lysate and *E. coli* cells. Proton linewidths (in Hz) as extracted from 2D ^{1}H-^{15}N experiments of GB1 backbone amides in water, lysate and *E. coli* cells. Due to increased intracellular viscosity, a significant increase of 107.6% in linewidth is observed when the GB1 is overexpressed and located within *E. coli* cells, as compared to its reference state in water. A slight increase of 13.4% in proton linewidth is also present when the protein is placed in lysate. Secondary structural elements of the GB1 protein are shown at the top of the chart.

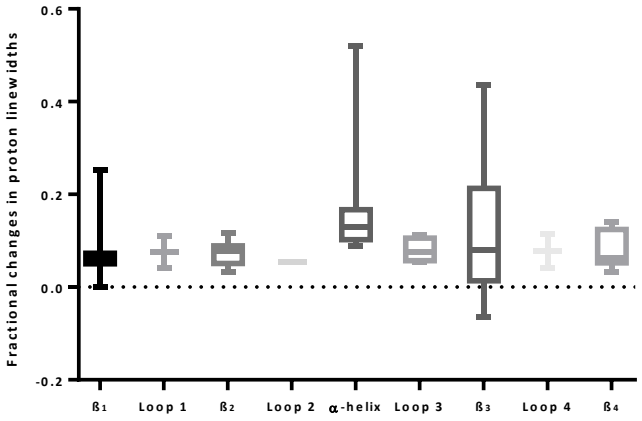

Figure 3.12 – Fractional changes in proton linewidths in lysate as a function of secondary structural elements in GB1. Proton linewidths were extracted from 2D ^{1}H-^{15}N experiments of GB1 backbone amides in water and lysate. Fractional changes were calculated for each residue by dividing the linewidth of the pure protein in water with the corresponding value in lysate, and then subtracting it from 1. The horizontal band inside each box represents the median of data in each group. As illustrated, when GB1 is placed in lysate, the residues comprising its α-helix exhibit the greatest degree of line broadening.

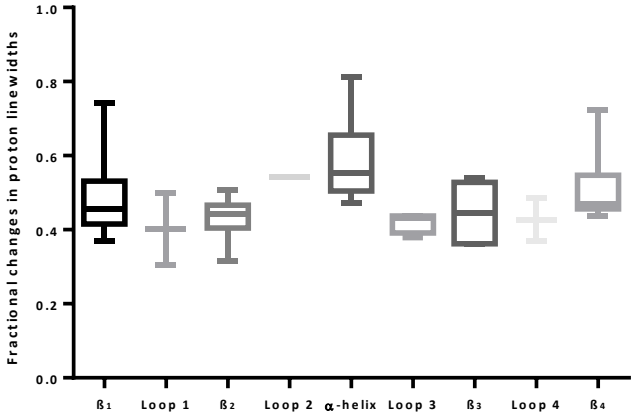

Figure 3.13 – Fractional changes in proton linewidths in *E. coli* cells as a function of secondary structural elements in GB1. Proton linewidths were extracted from 2D ^{1}H-^{15}N experiments of GB1 backbone amides in water and *E. coli* cells. Fractional changes were calculated for each residue by dividing the linewidth of the pure protein in water with the corresponding value in cells, and then subtracting it from 1. The horizontal band inside each box represents the median of data in each group. As illustrated, the residues within the α-helix display the greatest degree of line broadening when GB1 is overexpressed in *E. coli*.

In line with the evidence that the *E. coli* cytoplasm exhibits a much higher viscosity than its lysate, a significant reduction in the intensity of ^{1}H-^{15}N cross-peaks of backbone amides is seen following protein overexpression in these cells, with an intermediate decrease being observed when GB1 is present in lysate (Figure 3.14).

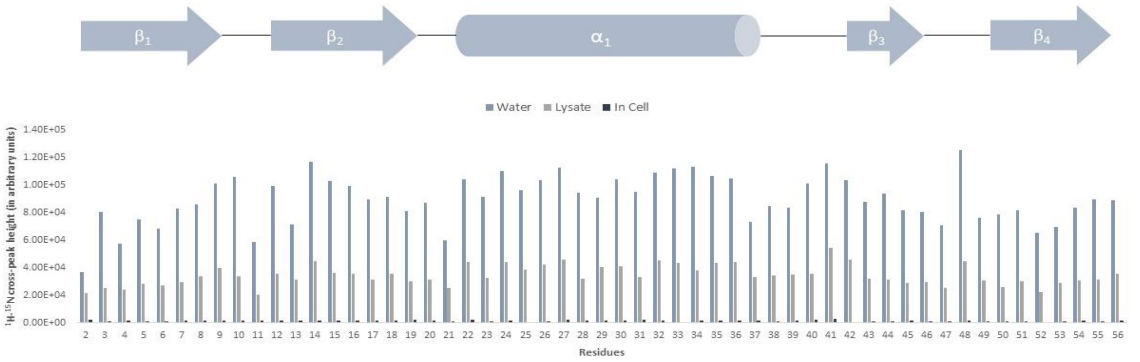

Figure 3.14 – ^{1}H-^{15}N cross-peak heights in the spectra of the GB1 protein in water, lysate and *E. coli* cells. Absolute cross-peak heights (in arbitrary units) as retrieved from 2D ^{1}H-^{15}N experiments of GB1 residues in water, lysate and *E. coli* cells. Due to increased intracellular viscosity, a dramatic reduction in peak intensity of GB1 residues is readily observed when the protein is located within *E. coli* cells. A still significant decrease in peak intensity can be seen when the protein is placed in lysate, as compared to its reference state in water. Secondary structural elements of the GB1 protein are shown at the top of the chart.

In addition to chemical shift perturbations, interactions with the local environment were also investigated through diffusion-ordered spectroscopy (DOSY). The basic rationale behind DOSY experiments relies on the application of a defocusing gradient to encode the spatial position of a molecule, which can subsequently be decoded by a second refocusing gradient of equal duration and magnitude. However, since molecules diffuse in solution in a size-dependant manner, complete refocusing does not occur and the detected signals are attenuated by an amount which reflects their diffusion coefficients. Furthermore, using the Stokes-Einstein equation ($D_t=\kappa T/6\pi\eta r$), it is possible to quantify variations in the hydrodynamic radius by relating the ratio of diffusion coefficients of the water and protein and water (D_{HDO}/D_{GB1}) with the inverse ratio of the corresponding radii (r_{GB1}/r_{HDO}). Assuming a static water radius, changes in the hydrodynamic radii value can then be assigned to the influence of molecular interactions between the protein and the environment. The determined diffusion coefficients and respective hydrodynamic radii ratios for the GB1 protein in water and lysate at 37°C are presented in Table 3.1.

Table 3.1 – Diffusion coefficients and hydrodynamic radius ratio of the GB1 protein in water and lysate. Diffusion coefficients and hydrodynamic radius ratio of GB1 (1 mM) in water and lysate at 37°C and pH 7.3 were extracted from DOSY experiments.

	Diffusion Coefficients *(D) ($\mu m^2 s^{-1}$)*		**Hydrodynamic radii ratio** (D_{HDO}/D_{GB1})
Condition	GB1	HDO	
Water	322.2 ±4.5	3944.6	12.2
150 mg/ml lysate	38.0 ±1.7	2326.3	61.2

As expected, due to the increase in solution viscosity, the diffusion coefficient of GB1 in the presence of 150 mg/ml of lysate was decreased by a factor of roughly 10 times, considering its value in water. Furthermore, when accounting for the effect of viscosity in protein diffusion, the apparent hydrodynamic radius of GB1 increases 5-fold. Unfortunately, it was not possible to retrieve a diffusion coefficient for GB1 within *E. coli* cells, because even applying the maximum gradient pulse duration (δ/2=2450 μs) allowed by the probe, only marginal signal attenuation was observed.

3.3 Study of GB1 side chain interactions in lysate and *E. coli*

To evaluate the effect of a crowded intracellular environment on the conformation of GB1, the combined ^{1}H-^{15}N and ^{1}H-^{13}C chemical shift perturbations of lysines and carbonyl side chains in this protein were also compared in lysate and *E. coli* cells using the exact same approach as before. Lysine $^{1}H_{\varepsilon}$-$^{15}N_{\zeta}$ and $^{1}H_{(\beta \text{ or } \gamma)}$-$^{13}C_{(\text{side chain carbonyl})}$ chemical shifts were obtained from H2CN and HCCO experiments, respectively, as described in section 2.2.3.

3.3.1 ^{1}H-^{15}N combined chemical shift perturbations of lysine side chains

Again, considering the pure protein in water as the reference state, an overall interaction between the *E. coli* lysate and each one of the six-lysine side chains was observed (Figure 3.15). Specifically, residue K31 located within the α-helix of GB1 simultaneously presents with the highest and lowest combined ^{1}H-^{15}N chemical shift changes. Since both cross-peaks share the same $^{15}N_{\zeta}$-nuclei, this variation is accounted for by differences in the chemical shift values of both H_{ε} protons. The spectra and respective assignments for the lysine side chains of GB1 in water and lysate can be found in Appendix section 7.2.

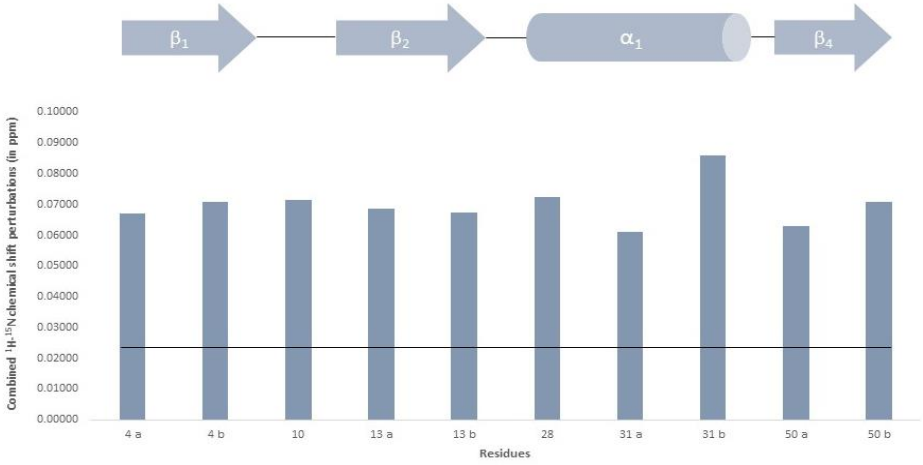

Figure 3.15 – Combined ^{1}H-^{15}N chemical shift perturbations of lysine side chains present in the GB1 protein in *E. coli* lysate. Chemical shift changes of lysine side chains when GB1 is placed in lysate, considering the pure protein in water as the reference state. The black line corresponds to the chemical shift cut-off value (0.020), above which residues are classed as being affected by the distinct lysate environment. Secondary structural elements of GB1 are shown at the top of the chart. Residue K31 exhibits the highest and lowest combined ^{1}H-^{15}N chemical shift variations simultaneously.

In the ^{1}H-^{15}N H2CN spectrum of the GB1 protein within *E. coli* cells, it was observed that 8 of the existent 10 peaks associated with lysine side chains in water are missing. Analysing the chemical shift perturbations of the 2 remaining ^{1}H-^{15}N cross-peaks, corresponding to residues K10 and K28, an extremely small cut-off value (0.00082) was determined. In this context, while both lysines appear to interact with the surrounding *E. coli* cytoplasm, residue K28 exhibits the highest chemical shift perturbation of the two (Figure 3.16). The spectra for the lysine side chains of GB1 in *E. coli* cells and respective assignments can be found in Appendix section 7.2.

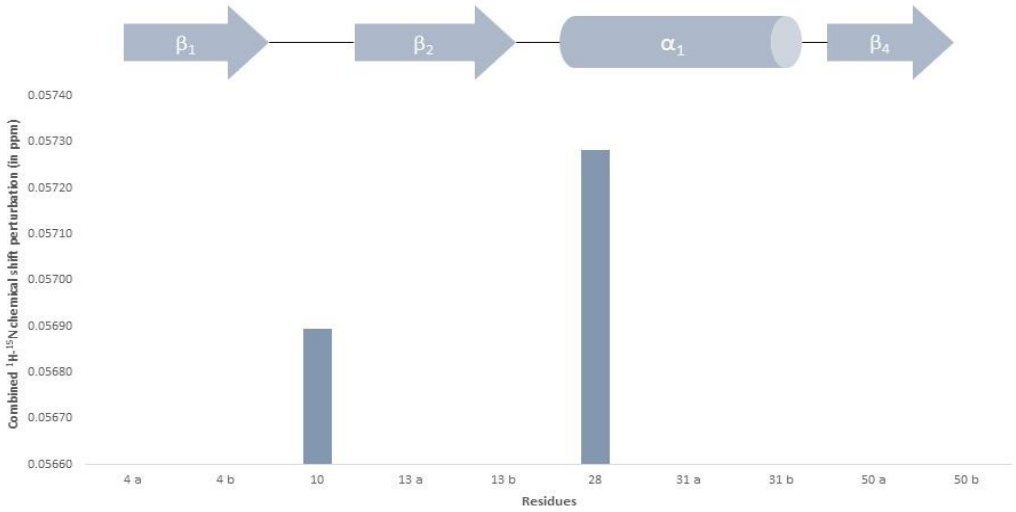

Figure 3.16 – Combined ^{1}H-^{15}N chemical shift perturbations of lysine side chains present in the GB1 protein within *E. coli* cells. Chemical shift variations in lysine side chains when GB1 is located within *E. coli* cells, considering the pure protein in water as the reference state. Secondary structural elements of GB1 are shown at the top of the chart. Since the determined chemical shift cut-off value was so small (0.00082), no line was drawn. Therefore, both K10 and K28 residues interact with the *E. coli* cytoplasm.

3.3.2 ^{1}H-^{13}C combined chemical shift perturbations of carbonyl side chains

Examining the ^{1}H-^{13}C combined chemical shift changes of the existent cross-peaks in the HCCO spectra of GB1, a general interaction between the carbonyl side chains and lysate contents was observed (Figure 3.17). More specifically, the side chain of residue E27 located within the α-helix of GB1 exhibits the highest combined ^{1}H-^{13}C chemical shift perturbation. The spectra and respective assignments for the carbonyl side chains of GB1 in water and lysate are presented in Appendix section 7.3.

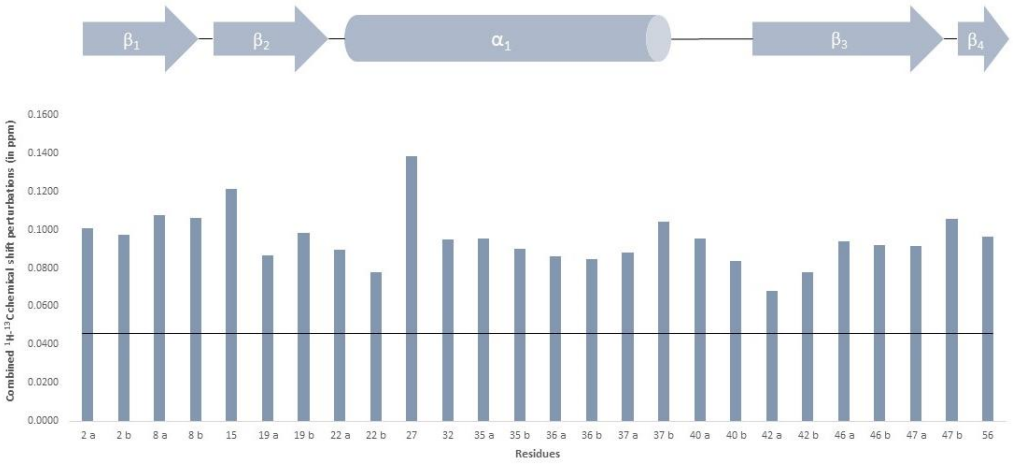

Figure 3.17 – Combined ^{1}H-^{13}C chemical shift perturbations of carbonyl side chains present in the GB1 protein in *E. coli* lysate. Chemical shift changes of carbonyl side chains when GB1 is placed in lysate, considering the pure protein in water as the reference state. The black line corresponds to the chemical shift cut-off value (0.042), above which residues are classed as being affected by the differing lysate environment. Secondary structural elements of GB1 are shown at the top of the chart.

In the ^{1}H-^{13}C HCCO spectrum of the GB1 protein within *E. coli* cells, it was observed that 7 of the possible 26 cross-peaks associated with carbonyl side chains are missing (Figure 3.18). Looking at the chemical shift perturbations of the 19 remaining ^{1}H-^{13}C peaks, only the one corresponding to the side chain of residue E42 does not appear to be affected by the *E. coli* cytoplasm. Moreover, while in lysate the most affected carbonyl side chain is that of residue E27, in the *E. coli* cytoplasm the highest chemical shift perturbation is associated with residue N35. The spectra for the carbonyl side chains and respective assignments can be found in Appendix section 7.3.

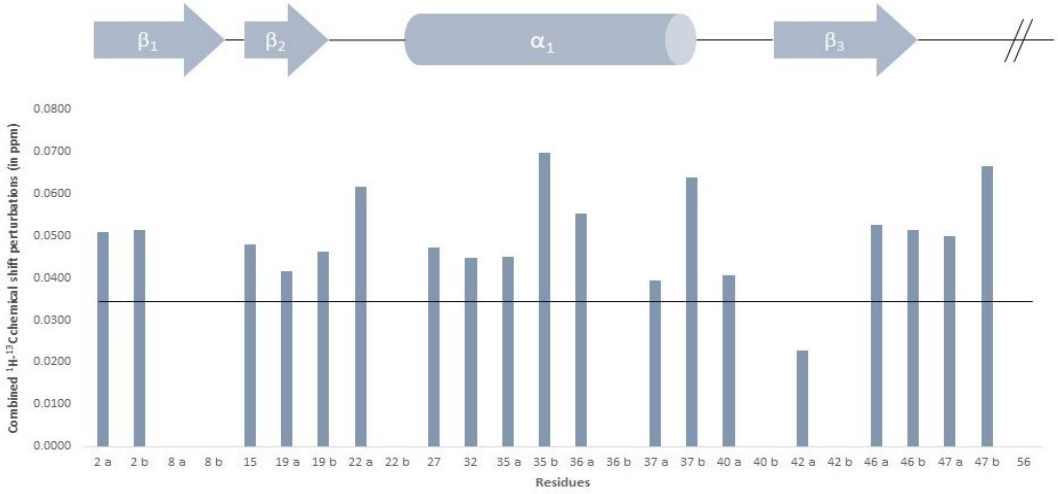

Figure 3.18 – Combined ^{1}H-^{13}C chemical shift perturbations of carbonyl side chains present in the GB1 protein within *E. coli* cells. Chemical shift changes of carbonyl side chains when GB1 is located within *E. coli* cells, considering the pure protein in water as the reference state. The black line corresponds to the chemical shift cut-off value (0.033), above which residues are classed as being affected by the distinct *E. coli* intracellular environment. Secondary structural elements of GB1 are shown at the top of the chart.

3.4 Comparative study between diffusion coefficients obtained by NMR and FCS experiments

3.4.1 Conjugate reaction between the GB1 protein and the Alexa Fluor 488 succinimidyl ester

Lastly, the diffusion coefficients obtained by two different spectroscopic techniques, NMR and FCS, were compared. For this purpose, the ^{15}N-GB1 protein was labelled with the fluorescent probe Alexa Fluor 488 succinimidyl ester. The latter reacts through a nucleophilic addition-elimination mechanism at their carbonyl carbon atoms with the amine groups of the GB1 protein, including its amine terminus and the ε-amino groups of its six lysines, to form stable amide bonds (Figures 3.19 and 3.20). Since the ε-amine group of lysines has a pK_a ranging between 10.7 to 11.4, the conjugation reaction was carried out in bicarbonate buffer with a slightly basic pH of 8.5 to maintain the amine group in the non-protonated form.[117]

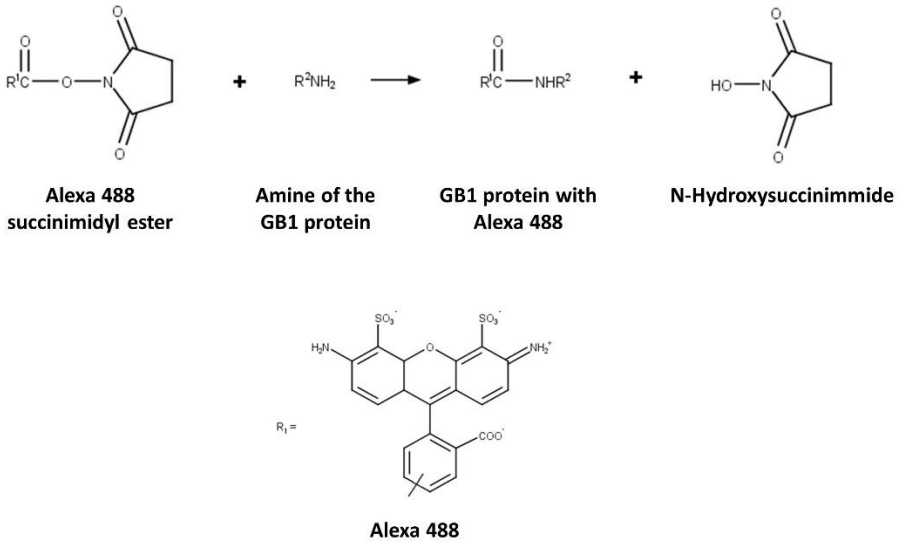

Figure 3.19 – Reaction scheme between the Alexa Fluor 488 succinimidyl ester and the primary amines of the GB1 protein. The substitution reaction of a succinimidyl ester with a primary amine leads to the formation of a stable amide bond and elimination of the N-hydroxysuccinimide leaving group.[125]

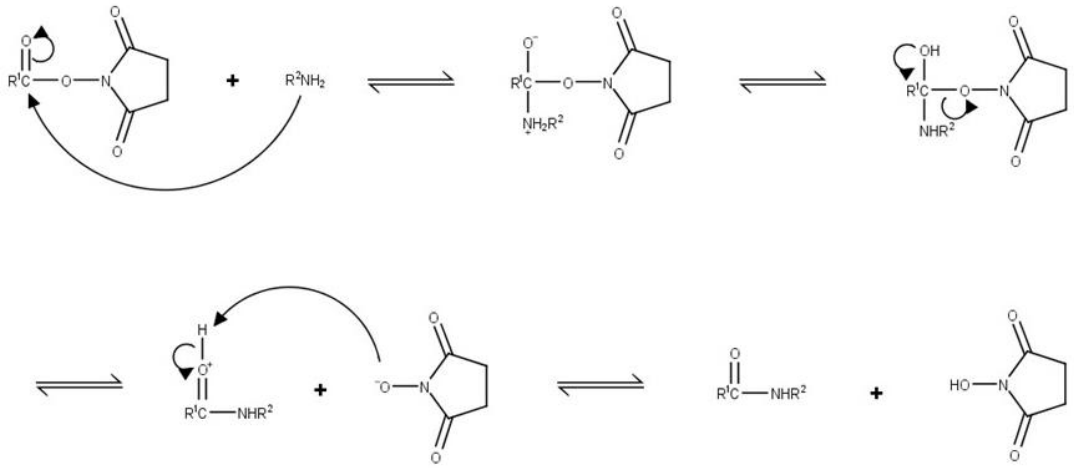

Figure 3.20 – Mechanism for the reaction between the Alexa 488 succinimidyl ester and the primary amines of the GB1 protein. The reaction proceeds via the creation of a tetrahedral intermediate following nucleophilic addition of the amine group. The elimination of the leaving N-hydroxysuccinimide group restores the carbon-oxygen bond in the carbonyl and leads to the formation of a stable amide bond.[125]

The efficiency of the labelling reaction was evaluated by determining the average number of dye molecules coupled to the GB1 protein. For this purpose, the absorbance spectra of the free Alexa 488 and the protein-dye conjugate were recorded (Figure 3.21). The degree of labelling (DOL) is calculated according to the following formula:

$$DOL = \frac{A_{max}/\varepsilon_{dye}}{A_{GB1}/\varepsilon_{GB1}}$$

where A_{max} is the absorbance maximum value in the spectrum of the protein-dye conjugate, ε_{dye} is the extinction coefficient of the Alexa 488 dye at its absorbance maximum and equal to 71000 $M^{-1}cm^{-1}$, A_{GB1} is the absorbance of GB1 at 280 nm and ε_{GB1} is the extinction coefficient of the protein at 280 nm and equal to 9970 $M^{-1}cm^{-1}$. Because the Alexa 488 fluorophore also absorbs electromagnetic radiation at 280 nm, it is necessary to correct the contribution of the dye to the absorbance of the GB1 protein:

$$A_{GB1} = (A_{280} - A_{max}) \times \frac{A_{280\ free\ dye}}{A_{\max\ free\ dye}}$$

where A_{280} is the absorbance registered at 280 nm in the spectrum of the protein-dye conjugate, while $A_{280\ free\ dye}$ is the absorbance at 280 nm and $A_{max\ free\ dye}$ is the absorbance maximum in the spectrum of the free Alexa 488. Therefore, a DOL of 1.3 was found for the conjugate reaction, which corresponds to an average labelling with the Alexa dye at more than one position in each GB1 molecule.

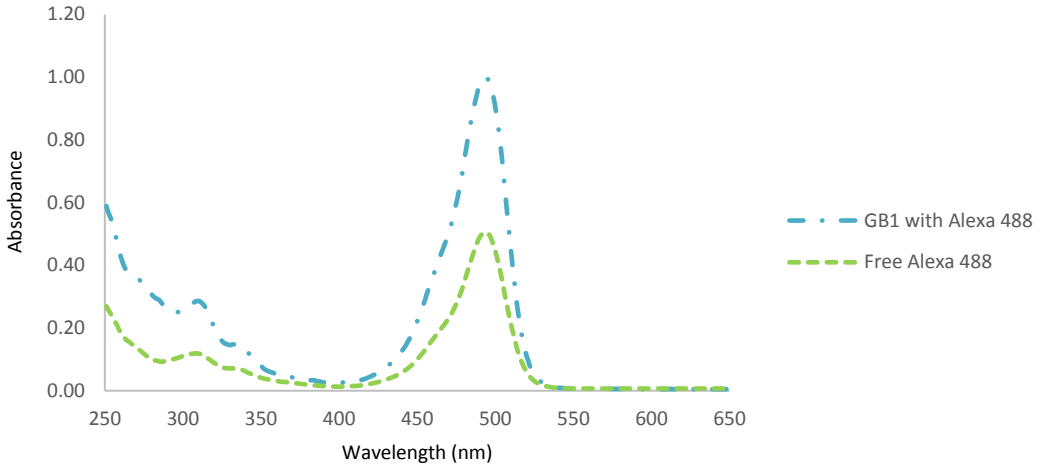

Figure 3.21 – Absorption spectra of the free Alexa 488 dye and the fluorescently labelled ^{15}N-GB1 protein between 250 and 650 nm. The concentration of the labelled protein can be determined by the Lambert-Beer law, using the absorbance value measured at 280 nm. However, since the Alexa 488 dye also absorbs electromagnetic radiation at this wavelength, it becomes necessary to introduce a correction.

3.4.2 Diffusion studies of the fluorescently labelled ^{15}N-GB1 protein with FCS experiments

As previously mentioned in the methods section 2.3.2, the measurement of diffusion coefficients with FCS relies on the temporal analysis of fluctuations in fluorescence intensity in the narrow detection volume of a confocal microscope, which can subsequently be fitted to an autocorrelation function *G(τ)*. The latter measures the degree of self-similarity between the oscillating signals that diffuse within the confocal volume. Therefore, if a signal is detected following an extended lag time τ, then the fluctuations at time *t + τ* arise from different particles than those at time *t* and *G(τ)* tends to zero. However, as τ becomes smaller, the same particles are responsible for producing oscillations at time *t* and *t + τ* and *G(τ)* approaches 1.

The determination of the diffusion coefficient with FCS requires a previous calibration of the confocal volume with a reference dye. In this context, the fluorophore rhodamine 110 with a diffusion coefficient of 440 $\mu m^2 s^{-1}$ was chosen. Considering the mathematical relation between the average residence time of molecules within the detection volume, τ_D, which can be extracted from its autocorrelation function (Figure 3.22), and the diffusion coefficient *D,*

$$\tau_D = \frac{w_0^2}{4D}$$

a value of 191 nm was obtained for the lateral size of the focus, w_0. Thus, the same equation was then used to determine the diffusion coefficient of the ^{15}N-GB1 protein by substituting τ_D with the value retrieved from its autocorrelation curve (Figure 3.23). Accordingly, a diffusion coefficient of 121.1 $\mu m^2 s^{-1}$ was found for this molecular ensemble in water at 20°C.

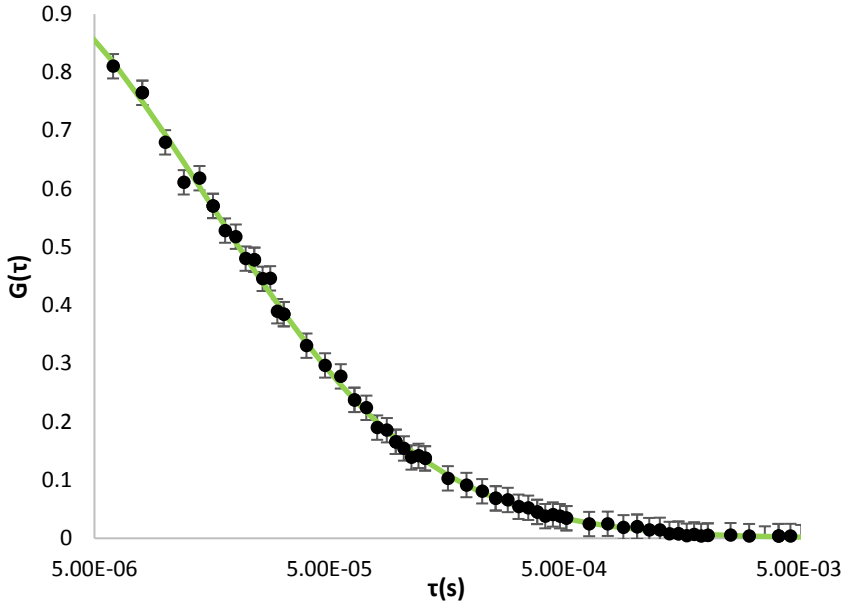

Figure 3.22 – Autocorrelation curve for the reference dye rhodamine 110 in water at 20°C. The average diffusion time of rhodamine 110 molecules within the confocal detection volume was extracted from this autocorrelation curve as equal to 20.7 µs.

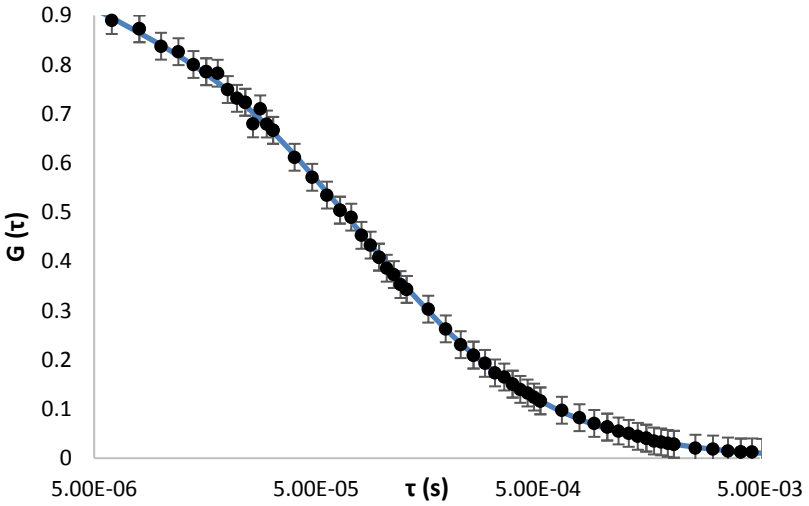

Figure 3.23 – Autocorrelation curve for the ^{15}N-GB1 protein conjugated with the Alexa 488 dye in water at 20°C. The average diffusion time of ^{15}N-GB1 coupled to Alexa 488 was retrieved from this autocorrelation curve as equal to 75.3 µs.

Next, the diffusion of the ^{15}N-GB1 protein was evaluated in 150 mg/ml of *E. coli* lysate. Again, by extracting the value of τ_D from its autocorrelation curve (Figure 3.24), the diffusion coefficient obtained for the protein was 14.5 $\mu m^2 s^{-1}$. A summary of the values registered in FCS experiments is provided in Table 3.2. As anticipated, the diffusion coefficient of the ^{15}N-GB1 protein fluorescently conjugated with Alexa 488 is decreased approximately 8 times when in the presence of 150 mg/ml of lysate. This reduction is reflected in the average residence time of the

protein within the detection volume, which is significantly increased when surrounded by the crowder.

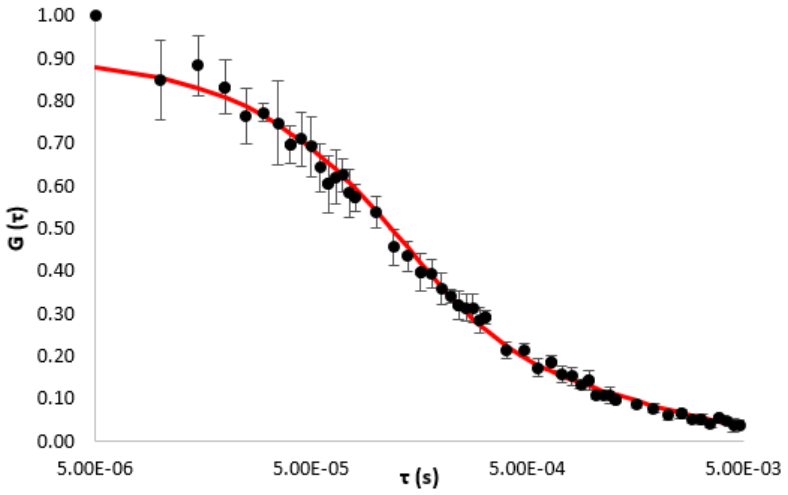

Figure 3.24 – Autocorrelation curve for the ^{15}N-GB1 protein conjugated with the Alexa 488 dye in 150 mg/ml of lysate at 20ºC. The average diffusion time of ^{15}N-GB1 coupled to Alexa 488 in lysate was extracted from this autocorrelation curve as equal to 747.2 µs.

Table 3.2 – Values obtained in FCS experiments for the ^{15}N-GB1 protein coupled to the Alexa 488 dye in water and lysate. Diffusion coefficients (in $\mu m^2 s^{-1}$) and average diffusion time (in µs) of 28.4 nM GB1 in water, with and without 150 mg/ml of lysate, as extracted from FCS experiments at 20ºC.

	Diffusion Coefficients *(D) ($\mu m^2 s^{-1}$)*	***Average Diffusion Time*** *(µs)*
^{15}N-GB1 with Alexa 488 (without lysate)	121.1	75.3
^{15}N-GB1 with Alexa 488 (150 mg/ml lysate)	14.5	747.2

3.4.3 Diffusion studies of the fluorescently labelled ^{15}N-GB1 protein with DOSY experiments

Finally, the diffusion coefficients of the fluorescently labelled ^{15}N-GB1 in the absence and presence of *E. coli* lysate were evaluated by DOSY experiments. The determined diffusion coefficients and respective hydrodynamic radii ratios for the GB1 protein in water and lysate with DOSY experiments at 20ºC are presented in Table 3.3. In comparison with FCS, an almost 2-fold smaller diffusion coefficient for GB1 in the absence of lysate was determined with DOSY. However, in the presence of lysate, the diffusion coefficients obtained by both techniques are in fact quite similar.

Table 3.3 – Values obtained in DOSY experiments for the ^{15}N-GB1 protein coupled to the Alexa 488 dye in bicarbonate buffer and lysate. Diffusion coefficients and hydrodynamic radius ratio of 0.37 mM GB1 in 90% 0.1M bicarbonate buffer and 10% D_2O, with and without 150 mg/ml of lysate, as extracted from DOSY experiments at 20°C.

	Diffusion Coefficients *(D) (μm² s⁻¹)*		**Hydrodynamic radii ratio** (D_{HDO}/D_{GB1})
Condition	GB1	HDO	
^{15}N-GB1 with Alexa 488 (without lysate)	55.6 ±2.5	1938.9	34.9
^{15}N-GB1 with Alexa 488 (with 150 mg/ml lysate)	20.4 ±3.2	1462.6	71.7

4. Discussion

The ability to characterize biomolecules, particularly proteins, with atomic resolution within a physiologically relevant environment differentiates NMR spectroscopy from other techniques employed in the study of biological macromolecules. Recent advances in labelling strategies and recombinant expression systems have made possible the acquisition of high-quality protein spectra within cells. Due to its exceptional stability and reduced capacity to interact in a specific manner with other cellular components, the GB1 protein represents the quintessential probe to investigate the physiochemical effects imposed by the crowded environment on the structure and dynamics of proteins. In the present thesis, the experimental conditions that limit the success of *in-cell* NMR experiments in *E. coli*, such as bacterial growth and overexpression, were optimized at first. Subsequently, the existent interactions between GB1 and its environment were investigated in lysate and *E. coli* cells, considering the pure protein in water as the reference state, by monitoring proton and nitrogen chemical shifts of backbone amides and lysine side chains, as well as carbon and proton chemical shifts of carbonyl-containing side chains. Interactions were further assessed by determining the translational motion of the protein through diffusion-ordered NMR spectroscopy (DOSY). Finally, a comparison of diffusion coefficients obtained for GB1 with DOSY and fluorescence correlation spectroscopy (FCS), the standard analytical technique for studying protein diffusion, was made.

The degree of overexpression constitutes a critical factor in the ability to observe GB1 exclusively within cells, since leakage of this cytoplasmic protein was positively correlated with induction time. Leakage compromises the usefulness of *in-cell* NMR experiments, because the extensive differences in viscosity between the intra- and extracellular environments originates cross-peaks with distinct linewidths. Thus, even when a small amount of leakage occurs, the signals associated with the protein on the outside will overshadow those of the same species within cells, since the cross-peaks of the latter display significantly lower intensities due to line broadening. In this context, as a compromise between leakage and spectral quality, both afforded by increased expression levels, an intermediate duration of 1h15 for protein induction following addition of IPTG to the culture medium was found to be optimal for *in-cell* NMR experiments. Furthermore, the switch from enriched growth medium to isotopically labelled minimal medium a couple of hours prior to the induction of expression not only minimizes the costs associated with *in-cell* NMR experiments, but also increases their sensitivity by reducing the number of unspecific cross-peaks. Another important aspect that needs to be considered is *E. coli* growth. In accordance to what is described in the literature, it was observed that the ideal bacterial pellet for *in-cell* NMR experiments comprises approximately 30% of the sample's final volume. In fact, when the *E. coli* slurry exceeded this value, shimming of the magnetic field was impossible to accomplish. Lastly, the placement of *E. coli* in a NMR tube to perform experiments at 37^{o}C allows the acquisition of GB1 spectra within these cells for at least two hours, since a considerable number of colony-forming bacteria is still evident at this time-point.

To investigate if the GB1 protein interacts with the crowded cellular environment, ^{1}H-^{15}N spectra of backbone amide bonds were acquired in lysate and *E. coli* cells, considering the pure protein in water as the reference state. The results obtained from these experiments suggest that the viscosity displayed by an artificially crowded solution, such as 150 mg/ml of lysate, and the intracellular milieu are markedly different. Indeed, a significant increase in proton linewidths from 2D ^{1}H-^{15}N experiments was observed following GB1 expression within cells, in comparison to the protein in lysate. Moreover, the effect of viscosity in signal line broadening in lysate and cells, considering the pure protein in water as the reference state, was more pronounced in the hydrogen-bonded amide groups of the α-helix. This is in accordance with the published literature, since Selenko also found that residues involved in GB1's secondary structure display a greater degree of line broadening than those in the loop regions.[104] Intuitively, it is quite reasonable that amide groups of the more mobile and dynamic elements of a protein are less affected, since the lower conformational restrictions imposed by its structure are better able to accommodate increases in viscosity. Further evidence supporting the existence of meaningful differences in viscosity between the *E. coli* lysate and intracellular environment was given by DOSY experiments. Indeed, it was impossible to obtain a decay of GB1 signals with increasing gradient strength and thus retrieve a diffusion coefficient for the protein within cells. However, the determination of hydrodynamic radii ratios of GB1 in water and lysate demonstrated that in the latter, despite the increase in solution viscosity, the protein behaves as a 5-fold heavier molecular ensemble.

Viscosity does not appear to be directly responsible for the observed differences in combined chemical shift perturbations of GB1 in lysate and *E. coli* cells, since higher values of this quantity were observed with the less viscous lysate samples. In addition, contrary to the case of line broadening, residues exhibiting the lowest median for combined chemical shift changes of amide groups are present within the α-helix, while those with the highest values of this quantity are located, particularly, in or at the border of the first loop region connecting the β_1 and β_2 sheets (residues K10 and L12, Figure 4.1). These observations can be understood based on transient interactions between the more dynamic and solvent-exposed loop region 1 with the soluble *E. coli* cytoplasmic contents simultaneously present in both lysate and cells. However, results also suggest that the GB1 protein interacts differently with the *E. coli* lysate and the intracellular environment. For instance, while an interaction with each one of the 55-backbone amide groups was observed in lysate, only certain residues were shown to interact with the *E. coli* cytoplasm. This preferential interaction with the intracellular milieu occurs not only with the residues in or at the border of the first loop region (N8, K10, T11 and L12, Figure 4.1) but also with those in or near the third loop region connecting the α-helix with the β_3 sheets (N37, D40, G41, W43, Figure 4.1). Thus, the GB1 protein seems to exhibit a different behaviour at the dynamic and solvent-exposed loop regions within cells in comparison to its artificially crowded counterpart.

The structure determination of the GB1 protein inside living *E. coli* cells was recently described using NMR methods.[73] By calculating the combined ^{1}H-^{15}N chemical shift differences

from the published assignments available at the PDB database for the protein in M9 medium (with the identification code 2N9K) and within cells (with the identification code 2N9L), an interaction between specific GB1 residues and the *E. coli* cytoplasm was observed. In fact, the interacting residues described in this thesis, except for those at positions 40 and 55, were also found when comparing the ^{1}H-^{15}N backbone assignments mentioned above. However, an increased number of statistically significant chemical shift changes were observed in the case of the published assignments, due to a smaller cut-off value. Furthermore, unlike the results of this thesis, the occurrence of interactions with the intracellular *E. coli* components does not exhibit a structural preference for the solvent-exposed secondary elements of the GB1 protein, since residues with high chemical shift perturbations appear to be randomly scattered. These differences could be due to the insertion of an additional glycine residue at position 2 and the absence of the T2Q mutation in the GB1 variant employed. In addition, the reference state for the comparison between the *in vitro* and *in-cell* ^{1}H-^{15}N backbone chemical shifts available at the PDB database was not water but M9 minimal medium instead.

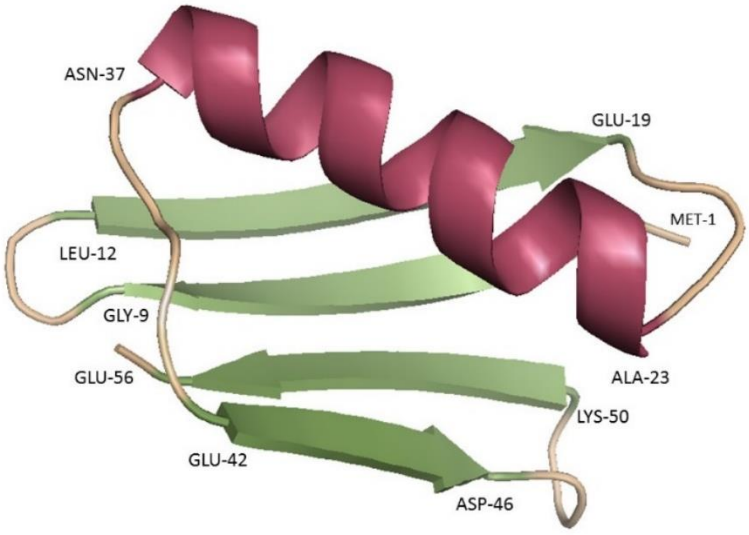

Figure 4.1 – Schematic cartoon diagram depicting the positions of residues within the GB1 protein according to its secondary structural elements. This representation was visualized by loading the PDB file with the identification code "2GB1" with PyMol Viewer (version 1.7.4) for educational purposes. Residues N8, K10, T11 and L12 are at or near loop region 1 (pale yellow), which connects the β_1 and β_2 sheets (in green). Meanwhile, residues N37, D40, G41 and W43 are located at or near the third loop region (pale yellow), which connects the α-helix (in red) and the β_3 sheet (in green) together.

The contribution of electrostatic interactions to the stability of GB1 within cells was previously demonstrated by the group of Pielak, following the observation that surface point mutations are more destabilizing in the *E. coli* cytoplasm than in buffer. Moreover, this effect was considerably more pronounced when the point mutation altered the net charge of the protein by +2, promoting the occurrence of attractive interactions between the GB1 surface and the negatively charged cytosolic contents of *E. coli*.[109] Electrostatic interactions between some of these oppositely charged residues, designated by salt bridges, have been demonstrated in

crystal structures of GB1.[117,127] Therefore, the existence of charge-charge interactions involving the GB1 protein were investigated in this thesis by following proton and nitrogen chemical shifts of lysine side chains, as well as carbon and proton chemical shifts of carbonyl-containing side chains of aspartate and glutamate residues. While a general interaction between the *E. coli* lysate and charged side chains was observed, the results from these experiments point to the presence of an intra-helical interaction between E27 (*i*) and K31 (*i*+4), because the highest chemical shift perturbations in ^{1}H-^{13}C HCCO and ^{1}H-^{15}N H2CN experiments were registered for these residues (Figure 4.2). However, no evidence of this salt bridge was found within cells, since residue K31 does not originate a cross-peak in the *in-cell* ^{1}H-^{15}N H2CN spectrum of GB1. The absence of most peaks in ^{1}H-^{15}N H2CN experiments within *E. coli* cells is likely due to viscosity inducing resonance broadening beyond detection, because the more dynamic and solvent-accessible lysine side chains K10 and K28 (with the exception of K4) are seen. The latter side chains are incapable of forming intramolecular donor-acceptor interactions due to being so thoroughly extended into the solvent.[127] However, an intermolecular salt bridge between K10 and D40 has been demonstrated in crystal structures of GB1 and since small but significant chemical shift changes were observed for these residues in ^{1}H-^{13}C HCCO and ^{1}H-^{15}N H2CN experiments, its existence within the intracellular environment cannot be excluded. Nonetheless, the more perturbed carbonyl-containing side chains within *E. coli* cells were those of residues N35, N37 and D47 and their combined chemical shift changes presumably reflect interactions with cellular components. In summary, the side chains of residues K10, K28, N35, N37 and D47 constitute the interacting probes of the GB1 protein with the intracellular environment in *E. coli* cells. Particularly, the involvement of residue K28 in electrostatic interactions with the cytosol was previously demonstrated by the group of Pielak. Using NMR-based amide proton-deuterium exchange experiments, the decrease in cytosolic pH resulted in a significant increase of attractive charge-charge interactions at this residue site.[106]

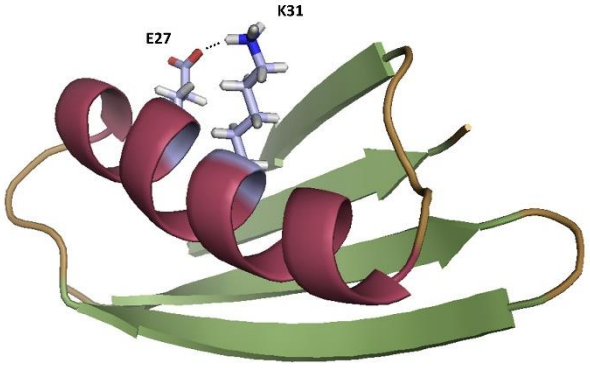

Figure 4.2 – Schematic cartoon diagram depicting the residues and respective side chains involved in the intra-helical salt bridge E27-K31. This representation was visualized by loading the PDB file with the identification code "2GB1" with PyMol Viewer (version 1.7.4). The salt-bridge between the negatively charged side chain of E27 (oxygen in red) and the positively charged K31 (hydrogen in white) within the α-helix has been described in crystal structures of the GB1 protein.[117]

As mentioned in the introduction chapter, the net negative charge of GB1 at physiological pH within the cytoplasm facilitates the acquisition of *in-cell* NMR spectra due to electrostatic repulsions with the abundant anionic proteins of *E. coli*. Therefore, the reduced number of GB1 residues that interact with its environment, both in lysate and within cells, is not surprising. However, a degree of weak and transient interactions, which do not compromise the ability to record high-quality spectra, is still evident. In this context, the acquisition of *in-cell* data for the ubiquitin variant E16W, using the same experimental procedure as that employed for GB1, did not yield a spectrum of the protein. Since the E16W mutant of ubiquitin exhibits an isoelectric of 8.2 and is positively charged at the physiological pH of *E. coli*, this result can be rationalized on the presence of attractive electrostatic interactions with other cytoplasmic components.

Lastly, the comparison of diffusion coefficients determined with DOSY and FCS revealed a significant difference in the values obtained in dilute conditions. Due to the submicrometric detection volume of a confocal microscope in FCS experiments, the timescale of translational diffusion events is on the order of microseconds, while the value for diffusion time, Δ, in DOSY was kept constant at 100 milliseconds. When the timescale of diffusion is very narrow, the probability of a protein in finding obstacles on its path is smaller and, thus, the value obtained is closer to its self-diffusion coefficient. On the contrary, as the timescale of diffusion increases, so does the probability of a protein in encountering other species on its way, leading to a reduction in its diffusion coefficient value.

Following this logic, it should be expected that differences in coefficients obtained with DOSY and FCS are more pronounced in crowded rather than dilute conditions, since the former contains a much higher concentration of molecules. However, the exact opposite occurred. In this scenario, this likely reflects the distinct optimal conditions for performing DOSY and FCS experiments. Indeed, the concentration necessary for obtaining well-fitted autocorrelation curves of proteins in FCS is on the nanomolar order of magnitude. However, the minimum concentration required for DOSY experiments is on the micromolar range. When GB1 was placed in the artificially crowded solution of 150 mg/ml of lysate, the diffusion coefficients obtained with FCS and DOSY were quite similar, which indicates that differences obtained in dilute conditions were due to the impossibility of employing the same sample preparation protocol in both techniques. Nonetheless, individual proteins often exist within cells at a nanomolar concentration and diffusion coefficients obtained with FCS are presumably more relevant from a biological standpoint than those retrieved with DOSY.

5. Conclusions and Future Perspectives

The general aim of this dissertation was to investigate the possible interactions of the GB1 protein with the *E. coli* lysate and intracellular milieu, using nuclear magnetic resonance (NMR) spectroscopy, with the purpose of inferring the physiochemical effects imposed by these two crowded environments on the structure and dynamics of proteins. The results presented here indicate that the intracellular environment is much more viscous than its artificially crowded counterpart comprising 150 mg/ml of lysate. However, the increase in line broadening due to reduced molecular tumbling rates in both cases is more evident in the hydrogen-bonded amide groups of residues in the α-helix. The results obtained also suggest that GB1 displays a distinct behaviour in cells with respect to lysate. More specifically, residues at or near the more flexible and solvent-exposed loop regions of the GB1 protein exhibit an increased preference for cellular components within cells compared to lysate. In addition, the intramolecular salt bridge between the side chains of residues E27 and K31 that is present in lysate was found to be absent within cells. In the latter case, the side chains of residues K10, K28, N35, N37 and D47 constitute the interacting probes with the intracellular environment.

GB1 is a small globular protein devoid of a functional role and extrinsic to most cells. Thus, it would be interesting to test if the tendencies outlined above represent a general feature of proteins, including structurally disordered ones, within cells. In addition, the *E. coli* cytoplasm contains a high natural abundance of anionic proteins, which facilitates the acquisition of *in-cell* spectra of negatively charged proteins, such as GB1, at its physiological pH. For instance, the decrease in intracellular pH of *E. coli*, which increases the number of proteins with a net positive charge, has been shown to negatively interfere with the number of observable cross-peaks due to increased electrostatic interactions. Therefore, another future avenue to consider could be examining the response rate of the GB1 side chain probes to changes in local environment, such as pH or ionic strength, within lysate and cells. Furthermore, the results presented in this dissertation should be correlated with those obtained from NMR approaches designed to measure protein dynamics, such as amide proton-deuterium exchange experiments in dilute and crowded conditions.

6. Bibliography

1. Blundell, T.L., Sibanda, B.L., Montalvão, R.W., Brewerton, S., Chelliah, V., Worth, C.L., Harmer, N.J., Davies, O. and Burke, D., 2006. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 361(1467),* pp.413-423.

2. Shi, Y., 2014. A glimpse of structural biology through X-ray crystallography. *Cell*, *159*(5), pp.995-1014.

3. T*he Protein Data Bank (PDB)*. Available from https://www.rcsb.org/pdb/. [August 17, 2017]

4. Luchinat, E. and Banci, L., 2017. In-cell NMR: a topical review. *IUCrJ*, *4*(2), pp.108-118.

5. Garman, E.F., 2014. Developments in x-ray crystallographic structure determination of biological macromolecules. *Science*, *343*(6175), pp.1102-1108.

6. Campbell, I.D., 2002. The march of structural biology. *Nature reviews Molecular Cell Biology*, *3*(5), pp.377-381.

7. Feng, W., Pan, L. and Zhang, M., 2011. Combination of NMR spectroscopy and X-ray crystallography offers unique advantages for elucidation of the structural basis of protein complex assembly. *Science China Life Sciences*, *54*(2), pp.101-111.

8. Claridge, T.D., 2016. *High-resolution NMR techniques in organic chemistry* (Vol. 27). Elsevier.

9. Sugiki, T., Kobayashi, N. and Fujiwara, T., 2017. Modern technologies of solution Nuclear Magnetic Resonance Spectroscopy for three-dimensional structure determination of proteins Open Avenues for Life Scientists. *Computational and Structural Biotechnology Journal*, *15*, pp.328-339.

10. Reckel, S., Hänsel, R., Löhr, F. and Dötsch, V., 2007. In-cell NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *51*(2), pp.91-101.

11. Li, C. and Liu, M., 2013. Protein dynamics in living cells studied by in-cell NMR spectroscopy. *FEBS letters*, *587*(8), pp.1008-1011.

12. Zimmerman, S.B. and Trach, S.O., 1991. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*. *Journal of Molecular Biology*, *222*(3), pp.599-620.

13. Serber, Z., Corsini, L., Durst, F. and Dötsch, V., 2005. In-cell NMR Spectroscopy. *Methods in Enzymology*, *394*, pp.17-41.

14. Selenko, P. and Wagner, G., 2006. NMR mapping of protein interactions in living cells. *Nature Methods*, *3*(2), pp.80-81.

15. Kumar, T.K.S., Thurman, R. and Jayanthi, S., 2013. In-cell NMR spectroscopy-*In vivo* monitoring of the structure, dynamics, folding, and interactions of proteins at atomic resolution. *Journal of Analytical & Bioanalytical Techniques*, *4*(1), p.e112.

16. van der Graaf, M., 2010. In vivo magnetic resonance spectroscopy: basic methodology and clinical applications. *European Biophysics Journal*, *39*(4), pp.527-540.

17. Serber, Z., Keatinge-Clay, A.T., Ledwidge, R., Kelly, A.E., Miller, S.M. and Dötsch, V., 2001. High-resolution macromolecular NMR spectroscopy inside living cells. *Journal of the American Chemical Society*, *123*(10), pp.2446-2447.

18. Serber, Z., Ledwidge, R., Miller, S.M. and Dötsch, V., 2001. Evaluation of parameters critical to observing proteins inside living *Escherichia coli* by in-cell NMR spectroscopy. *Journal of the American Chemical Society*, *123*(37), pp.8895-8901.

19. Singh, A., Upadhyay, V., Upadhyay, A.K., Singh, S.M. and Panda, A.K., 2015. Protein recovery from inclusion bodies of *Escherichia coli* using mild solubilization process. *Microbial Cell Factories*, *14*(1), p.41.

20. Theriot, J.A., 2013. Why are bacteria different from eukaryotes?. *BMC Biology*, *11*(1), p.119.

21. Serber, Z., Selenko, P., Hänsel, R., Reckel, S., Löhr, F., Ferrell, J.E., Wagner, G. and Dötsch, V., 2006. Investigating macromolecules inside cultured and injected cells by in-cell NMR spectroscopy. *Nature Protocols*, *1*(6), pp.2701-2709.

22. Selenko, P. and Wagner, G., 2007. Looking into live cells with in-cell NMR spectroscopy. *Journal of Structural Biology*, *158*(2), pp.244-253.

23. Heitz, F., Morris, M.C. and Divita, G., 2009. Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *British Journal of Pharmacology*, *157*(2), pp.195-206.

24. Guo, Z., Peng, H., Kang, J. and Sun, D., 2016. Cell-penetrating peptides: Possible transduction mechanisms and therapeutic applications. *Biomedical reports*, *4*(5), pp.528-534.

25. Copolovici, D.M., Langel, K., Eriste, E. and Langel, U., 2014. Cell-penetrating peptides: design, synthesis, and applications. *ACS Nano*, *8*(3), pp.1972-1994.

26. Tünnemann, G., Martin, R.M., Haupt, S., Patsch, C., Edenhofer, F. and Cardoso, M.C., 2006. Cargo-dependent mode of uptake and bioavailability of TAT-containing proteins and peptides in living cells. *The FASEB Journal*, *20*(11), pp.1775-1784.

27. Frankel, A.D. and Pabo, C.O., 1988. Cellular uptake of the Tat protein from human immunodeficiency virus. *Cell*, *55*(6), pp.1189-1193.

28. Green, M. and Loewenstein, P.M., 1988. Autonomous functional domains of chemically synthesized human immunodeficiency virus Tat *trans*-activator protein. *Cell*, *55*(6), pp.1179-1188.

29. Nagahara, H., Vocero-Akbani, A.M., Snyder, E.L., Ho, A., Latham, D.G., Lissy, N.A., Becker-Hapak, M., Ezhevsky, S.A. and Dowdy, S.F., 1998. Transduction of full-length TAT fusion proteins into mammalian cells: TAT-p27Kip1 induces cell migration. *Nature Medicine*, *4*(12), pp.1449-1452.

30. Inomata, K., Ohno, A., Tochio, H., Isogai, S., Tenno, T., Nakase, I., Takeuchi, T., Futaki, S., Ito, Y., Hiroaki, H. and Shirakawa, M., 2009. High-resolution multi-dimensional NMR spectroscopy of proteins in human cells. *Nature*, *458*(7234), pp.106-109.

31. Takeuchi, T., Kosuge, M., Tadokoro, A., Sugiura, Y., Nishi, M., Kawata, M., Sakai, N., Matile, S. and Futaki, S., 2006. Direct and rapid cytosolic delivery using cell-penetrating peptides mediated by pyrenebutyrate. *ACS Chemical Biology*, *1*(5), pp.299-303.

32. Ogino, S., Kubo, S., Umemoto, R., Huang, S., Nishida, N. and Shimada, I., 2009. Observation of NMR signals from proteins introduced into living mammalian cells by reversible membrane permeabilization using a pore-forming toxin, streptolysin O. *Journal of the American Chemical Society*, *131*(31), pp.10834-10835.

33. Palmer, M., Harris, R., Freytag, C., Kehoe, M., Tranum-Jensen, J. and Bhakdi, S., 1998. Assembly mechanism of the oligomeric streptolysin O pore: the early membrane lesion is lined by a free edge of the lipid membrane and is extended gradually during oligomerization. *The EMBO journal*, *17*(6), pp.1598-1605.

34. Walev, I., Bhakdi, S.C., Hofmann, F., Djonder, N., Valeva, A., Aktories, K. and Bhakdi, S., 2001. Delivery of proteins into living cells by reversible membrane permeabilization with streptolysin-O. *Proceedings of the National Academy of Sciences*, *98*(6), pp.3185-3190.

35. Bhakdi, S. and Tranum-Jensen, J., 1986. Membrane damage by pore-forming bacterial cytolysins. *Microbial Pathogenesis*, *1*(1), pp.5-14.

36. Gonzalez, M.R., Bischofberger, M., Frêche, B., Ho, S., Parton, R.G. and van der Goot, F.G., 2011. Pore-forming toxins induce multiple cellular responses promoting survival. *Cellular Microbiology*, *13*(7), pp.1026-1043.

37. Neumann, E., Schaefer-Ridder, M., Wang, Y. and Hofschneider, P.H., 1982. Gene transfer into mouse lyoma cells by electroporation in high electric fields. *The EMBO Journal*, *1*(7), pp.841-5.

38. Yarmush, M.L., Golberg, A., Serša, G., Kotnik, T. and Miklavčič, D., 2014. Electroporation-based technologies for medicine: principles, applications, and challenges. *Annual Review of Biomedical Engineering*, *16*, pp.295-320.

39. Binolfi, A., Limatola, A., Verzini, S., Kosten, J., Theillet, F.X., Rose, H.M., Bekei, B., Stuiver, M., Van Rossum, M. and Selenko, P., 2016. Intracellular repair of oxidation-damaged α-synuclein fails to target C-terminal modification sites. *Nature Communications*, *7*.

40. Theillet, F.X., Binolfi, A., Bekei, B., Martorana, A., Rose, H.M., Stuiver, M., Verzini, S., Lorenz, D., van Rossum, M., Goldfarb, D. and Selenko, P., 2016. Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature*, *530*(7588), pp.45-50.

41. Aricescu, A.R., Lu, W. and Jones, E.Y., 2006. A time-and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallographica Section D: Biological Crystallography*, *62*(10), pp.1243-1250.

42. Bertrand, K., Reverdatto, S., Burz, D.S., Zitomer, R. and Shekhtman, A., 2012. Structure of proteins in eukaryotic compartments. *Journal of the American Chemical Society*, *134*(30), pp.12798-12806.

43. Hamatsu, J., O'Donovan, D., Tanaka, T., Shirai, T., Hourai, Y., Mikawa, T., Ikeya, T., Mishima, M., Boucher, W., Smith, B.O. and Laue, E.D., 2013. High-resolution heteronuclear multidimensional NMR of proteins in living insect cells using a baculovirus protein expression system. *Journal of the American Chemical Society*, *135*(5), pp.1688-1691.

44. Banci, L., Barbieri, L., Luchinat, E. and Secci, E., 2013. Visualization of redox-controlled protein fold in living cells. *Chemistry & Biology*, *20*(6), pp.747-752.

45. Barbieri, L., Luchinat, E. and Banci, L., 2016. Characterization of proteins by in-cell NMR spectroscopy in cultured mammalian cells. *Nature Protocols*, *11*(6), pp.1101-1111.

46. Maldonado, A.Y., Burz, D.S. and Shekhtman, A., 2011. In-cell NMR Spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *59*(3), pp.197-212.

47. Theillet, F.X., Liokatis, S., Jost, J.O., Bekei, B., Rose, H.M., Binolfi, A., Schwarzer, D. and Selenko, P., 2012. Site-specific mapping and time-resolved monitoring of lysine methylation by high-resolution NMR spectroscopy. *Journal of the American Chemical Society*, *134*(18), pp.7616-7619.

48. Waugh, D.S., 1996. Genetic tools for selective labeling of proteins with α-^{15}N-amino acids. *Journal of Biomolecular NMR*, *8*(2), pp.184-192.

49. Pielak, G.J., Li, C., Miklos, A.C., Schlesinger, A.P., Slade, K.M., Wang, G.F. and Zigoneanu, I.G., 2009. Protein nuclear magnetic resonance under physiological conditions. *Biochemistry*, *48*(2), pp.226-234.

50. Plevin, M.J. and Boisbouvier, J., 2012. Isotope-labelling of methyl groups for NMR studies of large proteins. *Recent Developments in Biomolecular NMR, Royal Society of Chemistry*, pp.1-24.

51. Pfuhl, M., and Driscoll, P. C., 2000. Protein nuclear magnetic resonance spectroscopy in the new millennium. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *358*(1766), pp.513-545.

52. Tugarinov, V. and Kay, L.E., 2005. Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. *ChemBioChem*, *6*(9), pp.1567-1577.

53. Rosen, M.K., Gardner, K.H., Willis, R.C., Parris, W.E., Pawson, T. and Kay, L.E., 1996. Selective methyl group protonation of perdeuterated proteins. *Journal of Molecular Biology*, *263*(5), pp.627-636.

54. Gardner, K.H. and Kay, L.E., 1997. Production and incorporation of ^{15}N, ^{13}C, ^{2}H (^{1}H-δ1 methyl) isoleucine into proteins for multidimensional NMR studies. *Journal of the American Chemical Society*, *119*(32), pp.7599-7600.

55. Goto, N.K., Gardner, K.H., Mueller, G.A., Willis, R.C. and Kay, L.E., 1999. A robust and cost-effective method for the production of Val, Leu, Ile (δ1) methyl-protonated ^{15}N-, ^{13}C-, ^{2}H-labeled proteins. *Journal of Biomolecular NMR*, *13*(4), pp.369-374.

56. Tugarinov, V. and Kay, L.E., 2003. Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *Journal of the American Chemical Society*, *125*(45), pp.13868-13878.

57. Gakh, Y.G., Gakh, A.A. and Gronenborn, A.M., 2000. Fluorine as an NMR probe for structural studies of chemical and biological systems. *Magnetic Resonance in Chemistry*, *38*(7), pp.551-558.

58. Danielson, M.A. and Falke, J.J., 1996. Use of ^{19}F NMR to probe protein structure and conformational changes. *Annual Review of Biophysics and Biomolecular Structure*, *25*(1), pp.163-195.

59. Gregory, D.H. and Gerig, J.T., 1991. Structural effects of fluorine substitution in proteins. *Journal of Computational Chemistry*, *12*(2), pp.180-185.

60. Chaiken, I.M., Freedman, M.H., Lyerla, J.R. and Cohen, J.S., 1973. Preparation and studies of ^{19}F-labeled and enriched ^{13}C-labeled semisynthetic ribonuclease-S' analogues. *Journal of Biological Chemistry*, *248*(3), pp.884-891.

61. Post, J.F., Cottam, P.F., Simplaceanu, V. and Ho, C., 1984. Fluorine-19 nuclear magnetic resonance study of 5-fluorotryptophan-labeled histidine-binding protein J of Salmonella typhimurium. *Journal of Molecular Biology*, *179*(4), pp.729-743.

62. Kim, H.W., Perez, J.A., Ferguson, S.J. and Campbell, I.D., 1990. The specific incorporation of labelled aromatic amino acids into proteins through growth of bacteria in the

presence of glyphosate: application to fluorotryptophan labelling to the H^+-ATPase of *Escherichia coli* and NMR studies. *FEBS Letters*, *272*(1-2), pp.34-36.

63. Jackson, J.C., Hammill, J.T. and Mehl, R.A., 2007. Site-specific incorporation of a ^{19}F-amino acid into proteins as an NMR probe for characterizing protein structure and reactivity. *Journal of the American Chemical Society*, *129*(5), pp.1160-1166.

64. Wang, L. and Schultz, P.G., 2005. Expanding the genetic code. *Angewandte Chemie International Edition*, *44*(1), pp.34-66.

65. Wals, K. and Ovaa, H., 2014. Unnatural amino acid incorporation in *E. coli*: current and future applications in the design of therapeutic proteins. *Frontiers in Chemistry*, *2*(15).

66. Hammill, J.T., Miyake-Stoner, S., Hazen, J.L., Jackson, J.C. and Mehl, R.A., 2007. Preparation of site-specifically labeled fluorinated proteins for ^{19}F-NMR structural characterization. *Nature Protocols*, *2*(10), pp.2601-2607.

67. Burz, D.S., Dutta, K., Cowburn, D. and Shekhtman, A., 2006. In-cell NMR for protein-protein interactions (STINT-NMR). *Nature Protocols*, *1*(1), pp.146-152.

68. Burz, D.S., Dutta, K., Cowburn, D. and Shekhtman, A., 2006. Mapping structural interactions using in-cell NMR spectroscopy (STINT-NMR). *Nature Methods*, *3*(2), pp.91-93.

69. Xie, J., Thapa, R., Reverdatto, S., Burz, D.S. and Shekhtman, A., 2009. Screening of small molecule interactor library by using in-cell NMR spectroscopy (SMILI-NMR). *Journal of medicinal chemistry*, *52*(11), pp.3516-3522.

70. Hubbard, J.A., MacLachlan, L.K., King, G.W., Jones, J.J. and Fosberry, A.P., 2003. Nuclear magnetic resonance spectroscopy reveals the functional state of the signalling protein CheY *in vivo* in *Escherichia coli*. *Molecular Microbiology*, *49*(5), pp.1191-1200.

71. Shimba, N., Serber, Z., Ledwidge, R., Miller, S.M., Craik, C.S. and Dötsch, V., 2003. Quantitative identification of the protonation state of histidines in vitro and in vivo. *Biochemistry*, *42*(30), pp.9227-9234.

72. Sakakibara, D., Sasaki, A., Ikeya, T., Hamatsu, J., Hanashima, T., Mishima, M., Yoshimasu, M., Hayashi, N., Mikawa, T., Wälchli, M. and Smith, B.O., 2009. Protein structure determination in living cells by in-cell NMR spectroscopy. *Nature*, *458*(7234), pp.102-105.

73. Ikeya, T., Hanashima, T., Hosoya, S., Shimazaki, M., Ikeda, S., Mishima, M., Güntert, P. and Ito, Y., 2016. Improved in-cell structure determination of proteins at near-physiological concentration. *Scientific Reports*, *6*, p.38312.

74. Selenko, P., Frueh, D.P., Elsaesser, S.J., Haas, W., Gygi, S.P. and Wagner, G., 2008. *In situ* observation of protein phosphorylation by high-resolution NMR spectroscopy. *Nature Structural & Molecular Biology*, *15*(3), pp.321-329.

75. Banci, L., Barbieri, L., Bertini, I., Luchinat, E., Secci, E., Zhao, Y. and Aricescu, A.R., 2013. Atomic-resolution monitoring of protein maturation in live human cells by NMR. *Nature Chemical Biology*, *9*(5), pp.297-299.

76. Rakhit, R. and Chakrabartty, A., 2006. Structure, folding, and misfolding of Cu,Zn superoxide dismutase in amyotrophic lateral sclerosis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, *1762*(11), pp.1025-1037.

77. Hänsel, R., Foldynová-Trantírková, S., Löhr, F., Buck, J., Bongartz, E., Bamberg, E., Schwalbe, H., Dötsch, V. and Trantírek, L., 2009. Evaluation of parameters critical for observing nucleic acids inside living *Xenopus laevis* oocytes by in-cell NMR spectroscopy. *Journal of the American Chemical Society*, *131*(43), pp.15761-15768.

78. Rahman, S., Byun, Y., Hassan, M.I., Kim, J. and Kumar, V., 2017. Towards understanding cellular structure biology: In-cell NMR. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1865*(5), pp.547-557.

79. Reckel, S., Löhr, F. and Dötsch, V., 2005. In-Cell NMR Spectroscopy. *ChemBioChem*, *6*(9), pp.1601-1606.

80. Bodart, J.F., Wieruszeski, J.M., Amniai, L., Leroy, A., Landrieu, I., Rousseau-Lescuyer, A., Vilain, J.P. and Lippens, G., 2008. NMR observation of Tau in *Xenopus* oocytes. *Journal of Magnetic Resonance*, *192*(2), pp.252-257.

81. Sharaf, N.G., Barnes, C.O., Charlton, L.M., Young, G.B. and Pielak, G.J., 2010. A bioreactor for in-cell protein NMR. *Journal of Magnetic Resonance*, *202*(2), pp.140-146.

82. Kubo, S., Nishida, N., Udagawa, Y., Takarada, O., Ogino, S. and Shimada, I., 2013. A gel-encapsulated bioreactor system for NMR studies of protein-protein interactions in living mammalian cells. *Angewandte Chemie International Edition*, *52*(4), pp.1208-1211.

83. Barnes, C.O., Monteith, W.B. and Pielak, G.J., 2011. Internal and global protein motion assessed with a fusion construct and in-cell NMR spectroscopy. *ChemBioChem*, *12*(3), pp.390-391.

84. Minton, A.P. and Wilf, J., 1981. Effect of macromolecular crowding upon the structure and function of an enzyme: glyceraldehyde-3-phosphate dehydrogenase. *Biochemistry*, *20*(17), pp.4821-4826.

85. McGuffee, S.R. and Elcock, A.H., 2010. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Computational Biology*, *6*(3), p.e1000694.

86. Theillet, F.X., Binolfi, A., Frembgen-Kesner, T., Hingorani, K., Sarkar, M., Kyne, C., Li, C., Crowley, P.B., Gierasch, L., Pielak, G.J. and Elcock, A.H., 2014. Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). *Chemical Reviews*, *114*(13), pp.6661-6714.

87. Kuznetsova, I.M., Zaslavsky, B.Y., Breydo, L., Turoverov, K.K. and Uversky, V.N., 2015. Beyond the excluded volume effects: mechanistic complexity of the crowded milieu. *Molecules*, *20*(1), pp.1377-1409.

88. Mikaelsson, T., Ådén, J., Johansson, L.B.Å. and Wittung-Stafshede, P., 2013. Direct observation of protein unfolded state compaction in the presence of macromolecular crowding. *Biophysical Journal*, *104*(3), pp.694-704.

89. Dedmon, M.M., Patel, C.N., Young G.B. and Pielak, G.J., 2002. FlgM gains structure in living cells. *Proceedings of the National Academy of Sciences*, *99*(20), pp.12681-12684.

90. Smith, A.E., Zhou, L.Z. and Pielak, G.J., 2015. Hydrogen exchange of disordered proteins in *Escherichia coli*. *Protein Science*, *24*(5), pp.706-713.

91. Kuznetsova, I.M., Turoverov, K.K. and Uversky, V.N., 2014. What macromolecular crowding can do to a protein. *International Journal of Molecular Sciences*, *15*(12), pp.23090-23140.

92. Dhar, A., Samiotakis, A., Ebbinghaus, S., Nienhaus, L., Homouz, D., Gruebele, M. and Cheung, M.S., 2010. Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. *Proceedings of the National Academy of Sciences*, *107*(41), pp.17586-17591.

93. Hatters, D.M., Minton, A.P. and Howlett, G.J., 2002. Macromolecular crowding accelerates amyloid formation by human apolipoprotein C-II. *Journal of Biological Chemistry*, *277*(10), pp.7824-7830.

94. Brangwynne, C.P., 2013. Phase transitions and size scaling of membrane-less organelles. *The Journal of Cell Biology*, *203*(6), pp.875-881.

95. Guigas, G. and Weiss, M., 2008. Sampling the cell with anomalous diffusion-the discovery of slowness. *Biophysical Journal*, *94*(1), pp.90-94.

96. Wang, Y., Li, C. and Pielak, G.J., 2010. Effects of proteins on protein diffusion. *Journal of the American Chemical Society*, *132*(27), pp.9392-9397.

97. Wang, Y., Sarkar, M., Smith, A.E., Krois, A.S. and Pielak, G.J., 2012. Macromolecular crowding and protein stability. *Journal of the American Chemical Society*, *134*(40), pp.16614-16618.

98. Miklos, A.C., Sarkar, M., Wang, Y. and Pielak, G.J., 2011. Protein crowding tunes protein stability. *Journal of the American Chemical Society*, *133*(18), pp.7116-7120.

99. Sarkar, M., Smith, A.E. and Pielak, G.J., 2013. Impact of reconstituted cytosol on protein stability. *Proceedings of the National Academy of Sciences*, *110*(48), pp.19342-19347.

100. Sarkar, M., Lu, J. and Pielak, G.J., 2014. Protein crowder charge and protein stability. *Biochemistry*, *53*(10), pp.1601-1606.

101. McConkey, E.H., 1982. Molecular evolution, intracellular organization, and the quinary structure of proteins. *Proceedings of the National Academy of Sciences*, *79*(10), pp.3236-3240.

102. Gronenborn, A.M. and Clore, G.M., 1993. Structural studies of immunoglobulin-binding domains of streptococcal protein G. *Immunomethods*, *2*(1), pp.3-8.

103. Frank, M.K., Dyda, F., Dobrodumov, A. and Gronenborn, A.M., 2002. Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nature Structural & Molecular Biology*, *9*(11), pp.877-885.

104. Selenko, P., Serber, Z., Gadea, B., Ruderman, J. and Wagner, G., 2006. Quantitative NMR analysis of the protein G B1 domain in *Xenopus laevis* egg extracts and intact oocytes. *Proceedings of the National Academy of Sciences*, *103*(32), pp.11904-11909.

105. Spitzer, J. and Poolman, B., 2009. The role of biomacromolecular crowding, ionic strength, and physicochemical gradients in the complexities of life's emergence. *Microbiology and Molecular Biology Reviews*, *73*(2), pp.371-388.

106. Cohen, R.D. and Pielak, G.J., 2016. Electrostatic contributions to protein quinary structure. *Journal of the American Chemical Society*, *138*(40), pp.13139-13142.

107. Cohen, R.D., Guseman, A.J. and Pielak, G.J., 2015. Intracellular pH modulates quinary structure. *Protein Science*, *24*(11), pp.1748-1755.

108. Monteith, W.B. and Pielak, G.J., 2014. Residue level quantification of protein stability in living cells. *Proceedings of the National Academy of Sciences*, *111*(31), pp.11335-11340.

109. Monteith, W.B., Cohen, R.D., Smith, A.E., Guzman-Cisneros, E. and Pielak, G.J., 2015. Quinary structure modulates protein stability in cells. *Proceedings of the National Academy of Sciences*, *112*(6), pp.1739-1742.

110. Green, R. and Rogers, E.J., 2013. Chemical transformation of *E. coli*. *Methods in Enzymology*, *529*, pp.329-336

111. Silva, M., 2015. *Design of bio-inspired ionic liquids for protein stabilisation*. Master's thesis in Bioorganic Chemistry, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

112. Barnes, C.O. and Pielak, G.J., 2011. In-cell protein NMR and protein leakage. *Proteins: Structure, Function, and Bioinformatics*, *79*(2), pp.347-351.

113. Schägger, H., 2006. Tricine-SDS-PAGE. *Nature Protocols*, *1*(1), pp.16-22.

114. Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E.D., 2005. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins: Structure, Function, and Bioinformatics*, *59*(4), pp.687-696.

115. Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. and Clore, G.M., 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science*, *253*(5020), pp.657-661.

116. André, I., Linse, S. and Mulder, F.A., 2007. Residue-specific pK_a determination of lysine and arginine side chains by indirect ^{15}N and ^{13}C NMR spectroscopy: application to *apo* calmodulin. *Journal of the American Chemical Society*, *129*(51), pp.15805-15813.

117. Tomlinson, J.H., Ullah, S., Hansen, P.E. and Williamson, M.P., 2009. Characterization of salt bridges to lysines in the protein G B1 domain. *Journal of the American Chemical Society*, *131*(13), pp.4674-4684.

118. Wallerstein, J., Weininger, U., Khan, M.A.I., Linse, S. and Akke, M., 2015. Site-specific protonation kinetics of acidic side chains in proteins determined by pH-dependent carboxyl ^{13}C NMR relaxation. *Journal of the American Chemical Society*, *137*(8), pp.3093-3101.

119. Schumann, F.H., Riepl, H., Maurer, T., Gronwald, W., Neidig, K.P. and Kalbitzer, H.R., 2007. Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. *Journal of Biomolecular NMR*, *39*(4), pp.275-289.

120. Ferrage, F., Zoonens, M., Warschawski, D.E., Popot, J.L. and Bodenhausen, G., 2003. Slow diffusion of macromolecular assemblies by a new pulsed field gradient NMR method. *Journal of the American Chemical Society*, *125*(9), pp.2541-2545.

121. Wu, D.H., Chen, A.D. and Johnson, C.S., 1995. An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. *Journal of Magnetic Resonance, Series A*, *115*(2), pp.260-264.

122. Haustein, E. and Schwille, P., 2004. Fluorescence correlation spectroscopy: an introduction to its concepts and applications. *Biophysics Textbook Online*.

123. Ries, J. and Schwille, P., 2012. Fluorescence correlation spectroscopy. *BioEssays*, *34*(5), pp.361-368.

124. Gendron, P.O., Avaltroni, F. and Wilkinson, K.J., 2008. Diffusion coefficients of several rhodamine derivatives as determined by pulsed field gradient-nuclear magnetic resonance and fluorescence correlation spectroscopy. *Journal of Fluorescence*, *18*(6), pp.1093-101.

125. Solomons, T.W.G., Fryhle, C.B. and Snyder, S.A., 2014. *Organic Chemistry*. Hoboken.

126. Schanda, P., Kupče, Ē. and Brutscher, B., 2005. SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. *Journal of Biomolecular NMR*, *33*(4), pp.199-211.

127. Shi, X. and Rienstra, C.M., 2016. Site-specific internal motions in GB1 protein microcrystals revealed by 3D ^{2}H-^{13}C-^{13}C solid-state NMR spectroscopy. *Journal of the American Chemical Society*, 138(12), pp.4105-4119.

7. Appendix

7.1 ^{1}H-^{15}N backbone assignment of GB1 in water, lysate and *E. coli* cells

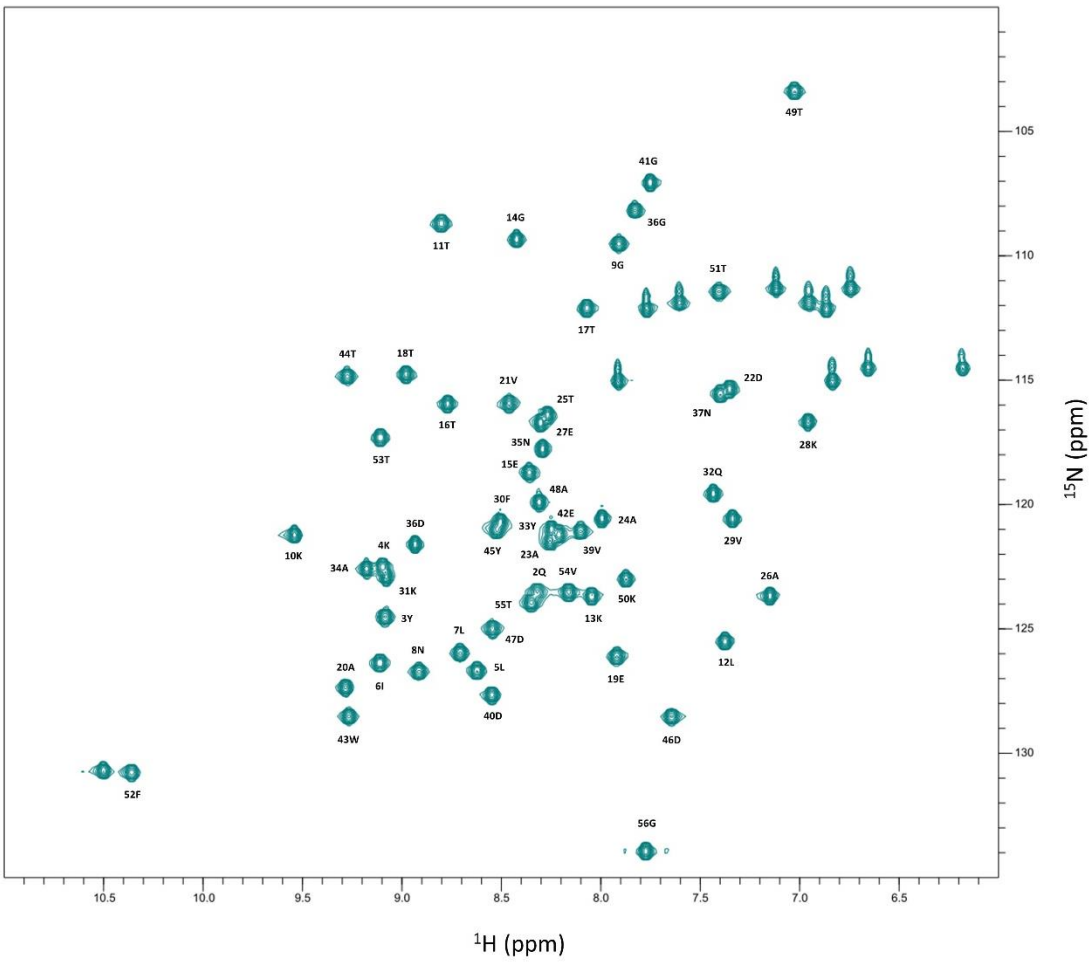

Figure 7.1 – ^{1}H-^{15}N HSQC spectrum and respective backbone amide assignment of the GB1 protein in water.

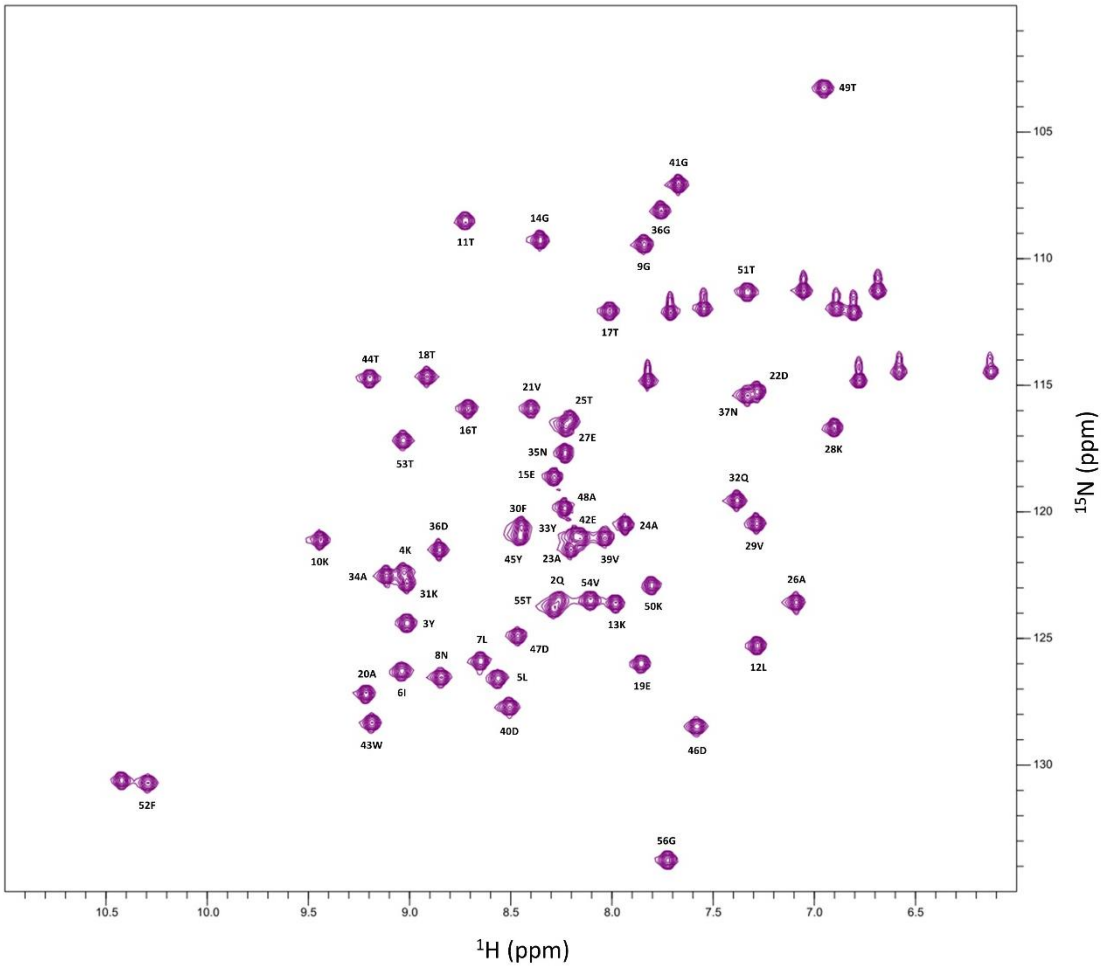

Figure 7.2 - ^{1}H-^{15}N HSQC spectrum and respective backbone amide assignment of the GB1 protein in *E. coli* lysate.

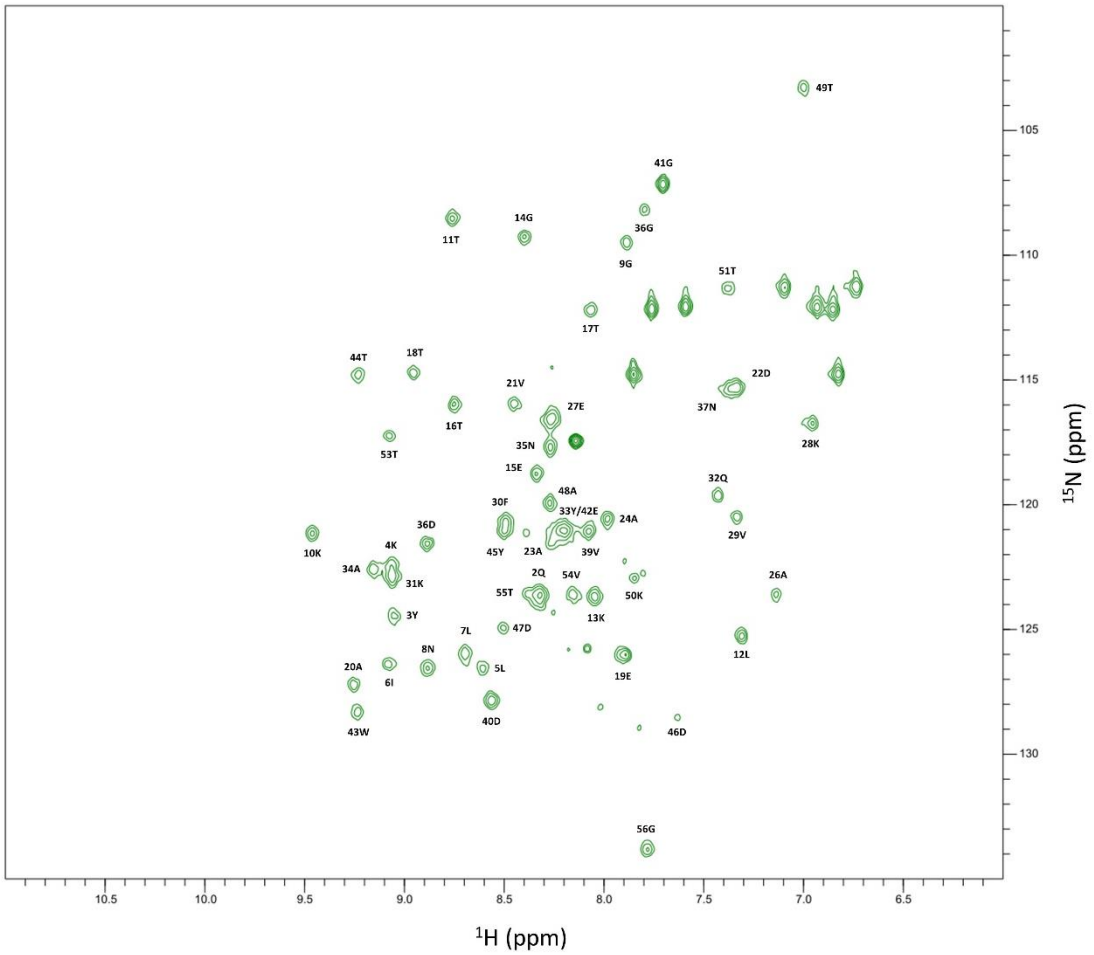

Figure 7.3 – ^{1}H-^{15}N SOFAST-HMQC spectrum and respective backbone amide assignment of the GB1 protein within *E. coli* cells.

7.2 ^{1}H-^{15}N lysine side chain assignment of GB1 in water, lysate and *E. coli* cells

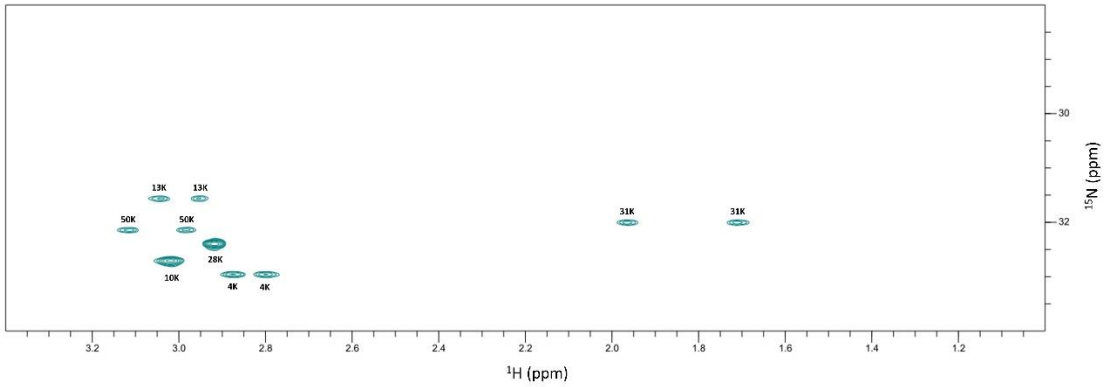

Figure 7.4 – ^{1}H-^{15}N H2CN spectrum and respective lysine side chain assignment of the GB1 protein in water.

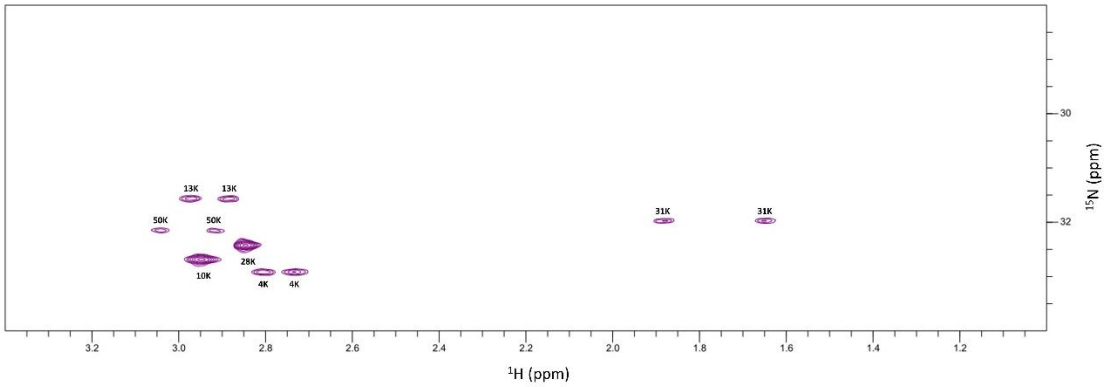

Figure 7.5 - ^{1}H-^{15}N H2CN spectrum and respective lysine side chain assignment of the GB1 protein in *E. coli* lysate.

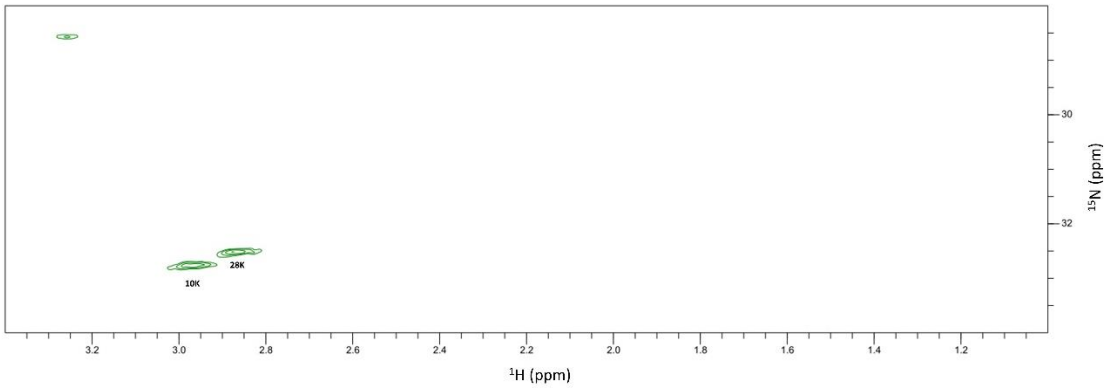

Figure 7.6 - ^{1}H-^{15}N H2CN spectrum and respective lysine side chain assignment of the GB1 protein within *E. coli* cells.

7.3 ^{1}H-^{13}C carbonyl side chain assignment of GB1 in water, lysate and *E. coli* cells

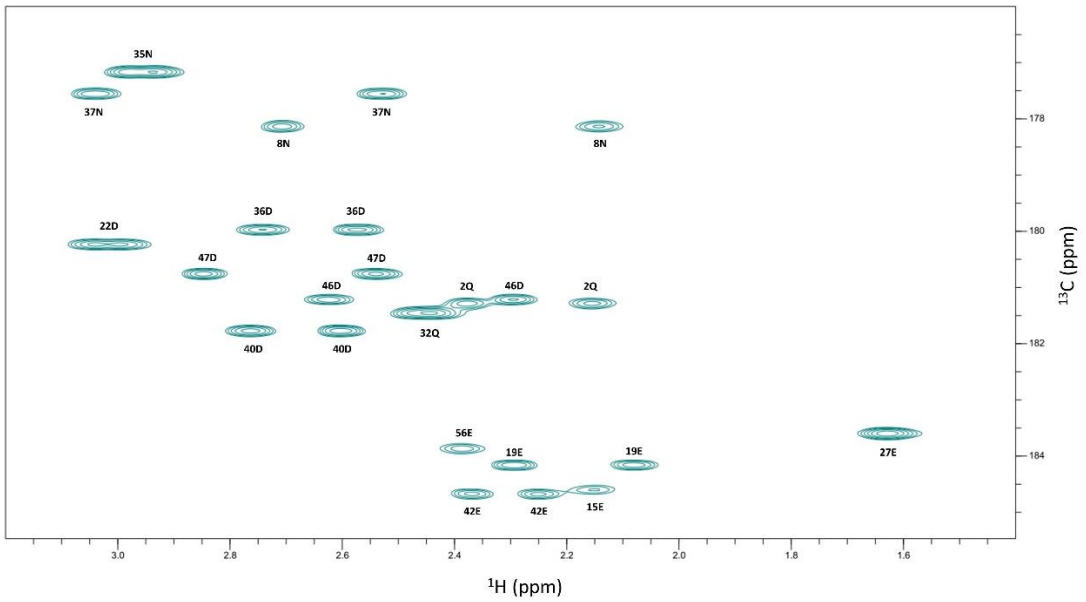

Figure 7.7 – ^{1}H-^{13}C HCCO spectrum and respective carbonyl side chain assignment of the GB1 protein in water.

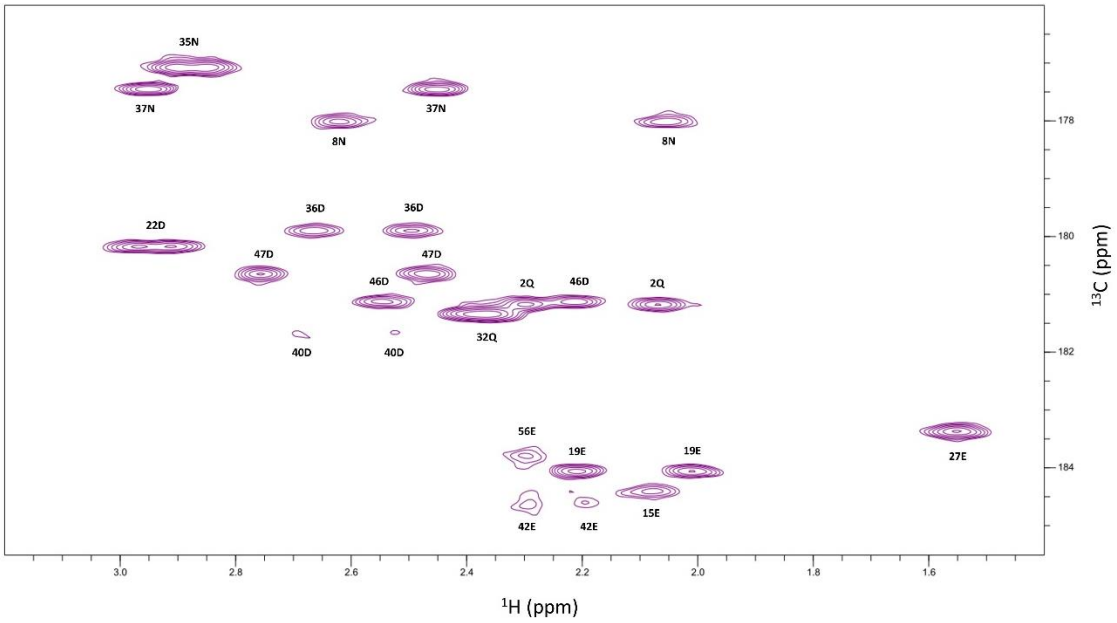

Figure 7.8 - ^{1}H-^{13}C HCCO spectrum and respective carbonyl side chain assignment of the GB1 protein in *E. coli* lysate.

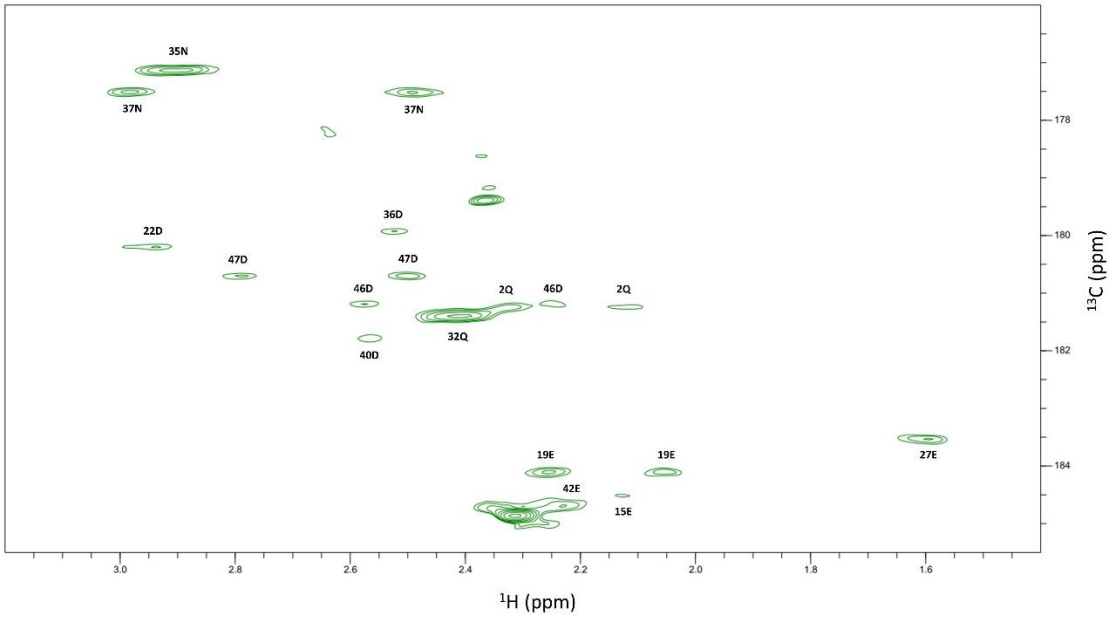

Figure 7.9 – ^{1}H-^{13}C HCCO spectrum and respective carbonyl side chain assignment of the GB1 protein within *E. coli* cells.