Structure and regulation of SH2 domain from mouse SH2B1 β

By:

Marym Fahad Albalwi

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Molecular Biology and Biotechnology
Firth Court, Western Bank, Sheffield S10 2TN

July 2021

Abstract

SH2B1 is a member of an adaptor protein family that contains two other proteins (SH2B2 and SH2B3), which regulate signalling pathways initiated by hormones such as insulin and leptin. As a result of alternative splicing of mRNA, four SH2B1 isoforms have been identified (α 765> δ 724> γ 682 > β 670), that differ in their C-terminus downstream of the SH2 domain. The SH2 domain binds to the phosphorylated tyrosine site of receptor tyrosine kinases (RTKs) and causes kinase activation. Additional to the conserved domains of SH2B1 (DD, PH, and SH2), more than 50% of the protein sequence is intrinsically disordered (unstructured region) including the C-terminal tail. The function of the C-terminal diversity is poorly understood. Mouse SH2B1β isoform tail contains a tyrosine residue (Y649), and we hypothesised that pY649 might bind to the tethered SH2 domain and therefore regulate its activity.

The NMR structure of the SH2 domain was calculated using the semi-automated CYANA (version 3.98.5) software. The structure calculation worked, although it required significant manual intervention, including correction of some the automated chemical shift assignments. A new validation method (ANSURR) was applied to check the reliability of the calculated structures, and it is demonstrated that additional hydrogen bond restraints are required to make the rigidity of the structure match experimental data. Phosphorylation using Fer kinase of a construct that includes the SH2 domain and the downstream 45 residues of the intrinsically disordered tail of β, including Y649, caused problems because unexpectedly, Y624 was phosphorylated more rapidly than Y649 and the protein aggregated. We therefore mutated Y624 to Phe. Y624F SH2c was successfully phosphorylated by Fer kinase. ^{15}N HSQC spectra confirmed a binding between pY649 and the pY pocket of the SH2 domain, and identified a new binding pocket in SH2 domain for residues C-terminal of pY. Surprisingly, the affinity between the phosphorylated C-terminus and SH2 was so weak that it could be outcompeted by phosphate in the buffer. The binding affinity between a phospho-tyrosine peptide derived from the C-terminal tail to the SH2 domain was measured using NMR titration. The revealed low affinity was outcompeted with the preferred JAK2 ligand as shown by HSQC spectra. These results provide some new hypotheses about the function of the C-terminal tail of SH2B1β.

Acknowledgements

In the Name of Allah, the Most Merciful, the Most Compassionate. I must acknowledge my limitless thanks to Allah for his blessing by giving me the opportunity and strength to carry out and complete this work.

First and foremost I would like to sincerely thanks my supervisor Professor Mike Williamson for his guidance, help and unlimited support throughout the research and writing process. It has been a great pleasure and honour to have him as my supervisor during my PhD.

During my PhD I met amazing people who helped and supported me from the beginning till the completion of this research. Especial thanks for Nicholas Fowler who offered guidance, help and support throughout solving the NMR structure. Andrea Hounslow deserve a special thanks for her amazing guidance in NMR experiments**,** her aid with computing and while solving the NMR structure. I'm grateful to my advisers Rosie A Staniforth and Stéphane Mensage for useful discussions and advice. Also thanks for Henry Wood, Nicola Baxter, Rob Smith, Svetalana Sedelnikova and Adelina Martín for sharing their knowledge with me. I had a chance to meet an amazing person during this journey who became my closest friend Mahreen U Hassan, we sharing unforgettable memories. To my friend Reem I'm so thankful for your support which helps me to pass difficult days even we are so far apart.

Finally my deepest gratitude goes to my family members. A massive thanks to my dearest mother Sabah, my sisters Marwa, Monira and Amnah, my brothers Hamoud, Ahmad and Omar, and my brother in low Oudah for their motivation, encouragement and unlimited support.

Abbreviation

Amino acid

A	Ala	Alanine
C	Cys	Cysteine
D	Glu	Glutamic acid
E	Asp	Aspartic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
pY	Phosphorylated tyrosine	

NMR spectra

NMR	Nuclear magnetic resonance
1D	One dimensional
2D	Two dimensional
3D	Three dimensional
HSQC	Heteronuclear single-quantum coherence
NOE	Nuclear Overhauser effect
NOESY	NOE spectroscopy

TOCSY	Total correlated spectroscopy
COSY	homonuclear correlation spectroscopy

Software and bioinformatic method

ANSURR	Accuracy of NMR Structure using Random Coil Index and Rigidity
CYANA	Combined assignment and dynamics algorithm for NMR applications
CNS	Crystallography and NMR system
FLYA	Fully automated structure determination of protein in solution
GARANT	General Algorithm for Resonance AssignmeNT
MUSCLE	Multiple Sequence Comparison by log Expectation

Unit

Å	Angstrom
bp	base pair
°C	Degree Celsius
Da	Daltons
kDa	KiloDaltons
g	gram
L	litter
M	Mole
ml	millimole
mM	millimole
MW	Molecular weight marker
ng	Nano gram
ppm	Parts per million
µg	microgram
µl	micromole
µM	micromole

Others

a.a	Amino acids
APS	Ammonium persulfate
ATP	Adenosine-5-triphosphate
Amp	Ampicillin

ESI MS	Electro-Spray Ionization Mass Spectrometry
CSP	Chemical shift perturbation
DTT	Dithiothreitol
D_2O	deuterium oxide
dH2O	Distilled water
DMSO	Dimethyl sulfoxide
DNase 1	Deoxyribonuclease I
E.coli	Escherichia coli
EDTA	Ethylenediamine tetra-acetic acid
ETD MS	Electron Transfer Dissociation Mass Spectrometry
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K_d	Equilibrium dissociation constant
M9	Minimal media
Ni-NTA	Nickel metal affinity (A nickel Nitrilotriacetic Acid Agarose)
nm UV	Nanometer Ultraviolet
OD600	Optical density measure at 600 nm
RMSD	Root-mean-square deviation
SDS page	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
TAE	Tris base, acetate and EDTA
TEV	Tobacco Etch Virus
Tris	Tris (hydroxymethyl) aminomethane
TSP	Sodium 3-trimethylsily-2,2,3,3-($^2H_4$) propionate
TECP	Tris (2-carboxyethyl)phosphine
TEMED	Tetramethylethyldiamine

Protein name

SH2B1	Src homology 2 adaptor protein 1
SH2B1 β	Src homology 2 adaptor protein 1 beta isoform
SH2	Src homology 2
JAK	Janus kinase

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction and study aims

1.1 Signal transduction

In multicellular organisms, cells are able to receive and respond to a wide range of extracellular signals to control their biological activities during their development. The communication system uses extracellular signals (e.g hormones and growth factors), which are normally processed and integrated with other signalling molecules, and carry signals between different cells to provide highly specific and effective communication (Alberts et al, 2015; Schlessinger and Ullrich, 1990).

Eukaryotic cells have a range of signal transmission systems to convert an extracellular signal into an intracellular response. The principal systems are intracellular receptors exemplified by lipophilic hormones which travel via simple diffusion through the plasma membrane; and cell surface receptors which are divided into three classes based on their transduction mechanism: ligand-gated ion channels; G-protein-coupled receptors; and enzyme-coupled receptors (Alberts et al, 2015) .

The fundamental mechanism in enzyme-coupled receptors is that the extracellular signalling molecule (ligand) as a first messenger binds to a specific transmembrane receptor, which acts as signal transducer to convert the extracellular ligand-binding event into an intracellular response via creation of a series of downstream signalling molecules called second messengers. The activation of this system is mainly regulated by the phosphorylation of a set of specific intracellular proteins (Alberts et al, 2015; Blume-Jensen and Hunter, 2001). Important members of the enzyme-coupled receptor family are tyrosine kinase receptors and tyrosine kinase associated receptors, which are exemplified by the insulin receptor and the JAK/STAT receptor, respectively. There are fifty-eight known human receptor tyrosine kinases (RTKs) incorporated into twenty sub-families (Lemmon and Schlessinger, 2010). RTKs regulate critical cellular processes such as proliferation, differentiation, cell survival and metabolism (Blume-Jensen and Hunter, 2001).

The insulin receptor switches from an inactive to active state upon ligand binding. The modular structure of the insulin receptor consists of two extracellular subunits that act as a binding site for the insulin hormone, and two transmembrane domains connected to two intracellular receptor subunits (receptor tyrosine kinase domain) that are linked via disulphide bonds inside the cell. Binding of insulin to the extracellular subunits induces a conformational change of the transmembrane domains, which is followed by rearrangement and activation of the intracellular subunits that stimulate their tyrosine kinase activity. This transmits the signal by cross autophosphorylation of tyrosine residues on one subunit by the kinase on the opposite subunit (Van Obberghen et al., 2001). Phosphorylation of tyrosine residues creates binding sites for a number of cytoplasmic proteins such as an adaptor protein named receptor-bound protein 2 (Grb2) as shown in Figure 1.1. This protein contains an SH2 domain (Src-homology 2), which recognizes the phosphorylated tyrosine pY plus about three downstream residues (pYXNX), and two N and C terminal SH3 domains (Src-homology 3) which recruit and bind to a second adaptor protein Son of Sevenless (Sos) via its proline-rich region. Sos then activates Ras by exchanging GDP to GTP, which initiates a kinase cascade leading to phosphorylation of transcription factors (Holgado-Madruga et al., 1996; Skolnik et al., 1993).

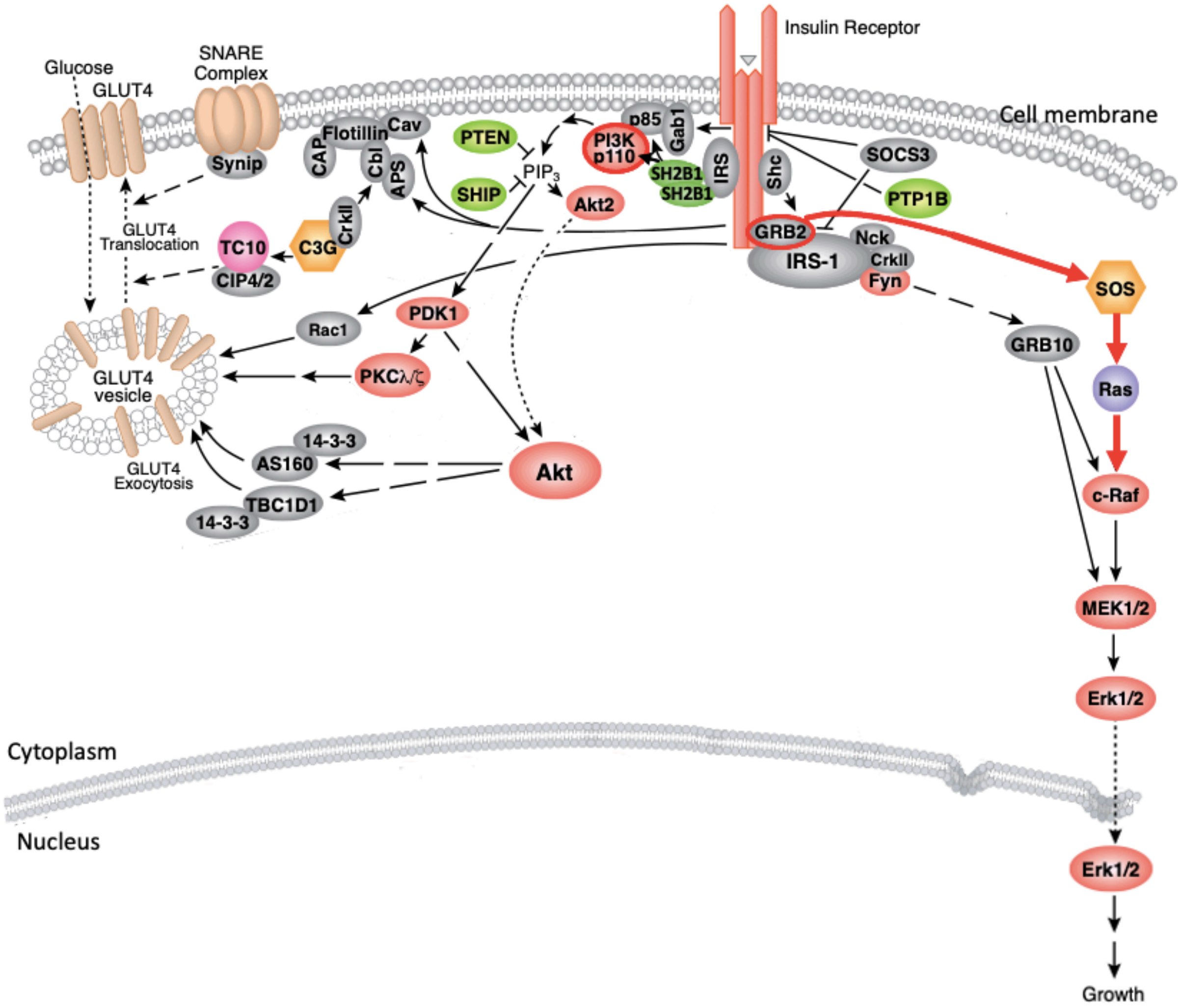


Figure 1.1. An overview diagram of the insulin signalling pathway, taken from Cell Signalling Technology. The diagram shows the downstream signaling proteins; GRB2 which binds to SOS to activate Ras, IRS which recruits PI3K, Grb2 and SH2B1, and small adaptors proteins Nck, Crk and Fyn.

Unlike tyrosine kinase receptors, the Janus kinase/signal transducers and activators of transcription (JAK/STAT) signalling system adopts a simpler mechanism to transfer signals from outside to inside cells, without having a receptor with intrinsic kinase activity. Instead of that the autophosphorylation of the intracellular receptor and cytoplasmic proteins occurs by an associated JAK, Figure 1.2. The receptor dimerises after binding to an extracellular ligand, then JAK starts to phosphorylate a number of tyrosine residues on the receptor to recruit a set of SH2-containing proteins called STATs. The bound STAT proteins are subsequently phosphorylated by JAKs, then dimerise, and dissociate to relocate to the nucleus where they take part in transcriptional activation and gene expression (Schindler et al, 2007).

Both systems, either the protein receptor kinase or associated with protein kinase, have similar recognition mechanisms which rely on phospho-tyrosine sites together with about three to four adjacent amino acids.

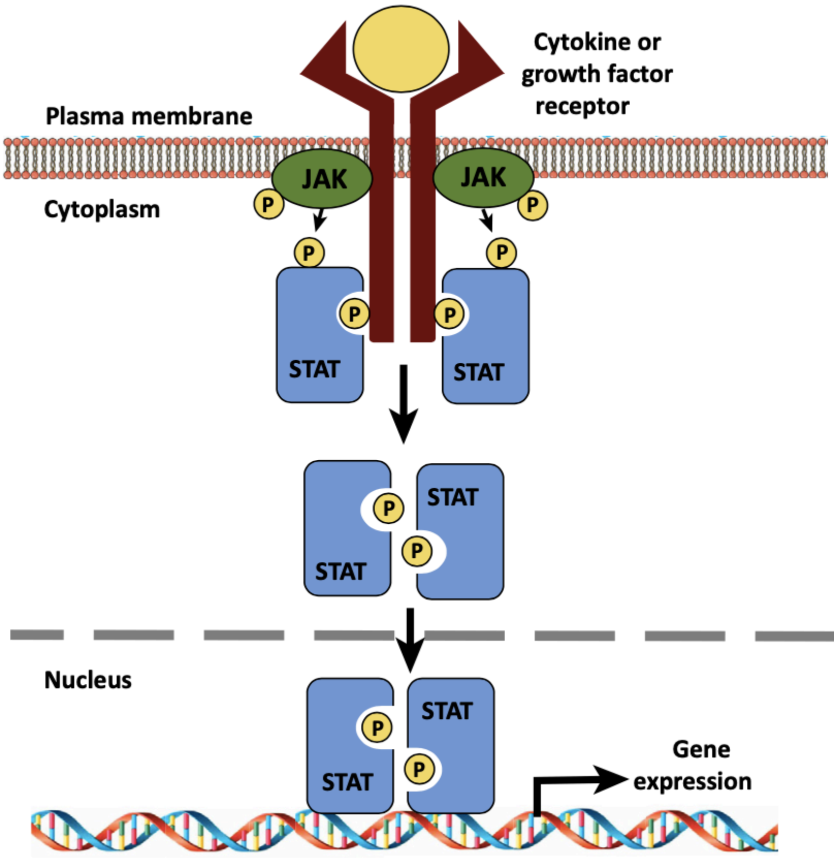


Figure 1.2. Schematic of JAK-STAT signalling pathway. The activated JAKs self-phosphorylate (P) and phosphorylate their associate receptor (P), and recruits STATs which in advance phosphorylate and form a dimer. The diagram is taken from (Dodington, Desai, & Woo, 2018).

1.2 Specificity in intracellular signal transduction

RTK intracellular signalling molecules are mostly proteins and function together to transfer signals that are received by the cell surface receptors into an intracellular response via a series of signalling adaptors (signalling pathway) thus achieving the specific target response.

Theoretically an activated signalling molecule could interact directly with its downstream targets and those targets would be activated only by their upstream signal. However in eukaryotic cells signalling molecules exist in the cytoplasm as a big number of molecules that have relatively broad substrate specificities (e.g SH2 domains) as mentioned in previous examples. This can create problems in specificity, and increase the likelihood of cross-talk between different pathways.

In multicellular organisms in such situations, evolution creates levels of complexity in intracellular communication to achieve the specificity in biological response. One effective strategy is adding extra proteins or polypeptides that interact weakly with several of the existing proteins and modulate the signal, until the required specificity has been obtained (Alberts et al, 2015). Such mechanism appears noticeably in most signalling pathways which have a set of auxiliary proteins that are called adaptor proteins, and another set termed as scaffold proteins (Pawson and Scott, 1997).

Scaffolds are big proteins with no enzymatic activity that are able to form protein complexes by binding to a number of signalling molecules and bringing them together close to their cognate receptor or substrates. Increasing the local concentration (assembly of protein complexes) leads to selective rapid sequential activation thus making the signal transduction more precise and effective (Shaw and Filbert, 2009; Pawson and Scott, 1997).

Insulin receptor substrate 1 (IRS-1) is a scaffold in the insulin signalling pathway, Figure 1.1. It contains a pTyr-binding domain (PTB) that recognizes the phosphorylated motif (NPXY) in the transmembrane region of the receptor (Mardilovich et al., 2009). The C-terminal and the central regions of IRS-1 have more than 20 potential phosphorylation sites on both serine and tyrosine. Phosphorylation of these residues creates binding motifs recognised by a number of signalling proteins in the cytoplasm including PI3K and Grb2. Besides that, IRS-1 contains an N-terminal pleckstrin homology domain (PH) which binds to phospholipid within the plasma membrane.

Adaptor proteins on the other hand comprise two or more small interacting domains that help in the formation of complex groups of interacting signalling proteins. Each module binds to a particular motif in another target protein (e.g peptide sequence) or lipid in the plasma membrane, bringing together protein complexes and thus enhancing specificity (Pawson and Scott, 1997).

Figure 1.1 shows a number of adaptor proteins in the insulin signal transduction pathway such as noncatalytic region of tyrosine kinase (Nck) and Crk proteins: both contain SH2 and SH3 domains, and bind to a range of additional proteins to modulate the signalling pathway, whereas Fyn adaptor protein contains one additional tyrosine kinase domain, and phosphorylation of target proteins by Fyn creates additional sites for recognition by SH2 domains (Sorokin et al., 1998; Latreille et al., 2011; Sun et al., 1996).

1.3 Src Homology 2 B adaptor (SH2B) system

The main class of proteins that contribute to signalling processes inside cells are adaptor proteins. The cell uses a number of these proteins, in order to achieve an appropriate response (Maures et al., 2007). One of these adaptors is Src homology 2 B family proteins (SH2B), which are implicated in variant signalling pathways mediated via JAK tyrosine kinases as well as receptor tyrosine kinases. The SH2B family contains at least three members: SH2B1, SH2B2 and SH2B3, which have been found in different signal transduction pathways, Figure 1.3.

As other classical adaptor proteins it contains multiple binding domains to mediate protein-protein interactions and lacks intrinsic enzymatic activity. The domains within the SH2B protein sequences serve as adaptors that link certain phosphorylated receptor tyrosine kinases and receptor-associated tyrosine kinases, such as the insulin receptor and Janus kinase family on the one hand (Riedel et al., 1997; Maures et al., 2007), and permit specific interactions with cytoplasmic proteins on the other hand.

The SH2B members are present in mammals, such as human, mice and rat, and insects such as *Drosophila melanogaster* (Song et al., 2010). SH2B1 is widely expressed in mammals and conserved in *Drosophila*, but the expression of SH2B2 and SH2B3 are restricted to insulin sensitive and hematopoietic tissues, respectively (Ahmed and Pillay, 2001; Moodie et al, 1999; Rui et al., 1997). The extensive expression of SH2B1 suggests the functionally important role of these proteins.

1.4 Sequence comparison between SH2B family members

The SH2B family members share similar structure of three domains: an N-terminal dimerization domain (DD), a central pleckstrin homology domain (PH), and a C-terminal Src homology 2 (SH2) domain, as illustrated in Figure 1.3. The SH2 domain is a common feature of cytoplasmic proteins that bind to phospho-tyrosine sites, while PH domains typically recognize phosphatidylinositides within the cell membrane. In addition, the N-terminal domain of SH2B members mediates homo-dimerization and hetero-dimerization (Maures et al., 2007).

In between the independently folded domains, there are long unstructured regions that work as flexible linkers. Those linkers have a big diversity in the sequence length and amino acid type in between the family members. However they contain numbers of proline-rich regions, polar residues (such as Ser, Thr, Gln and Asn) and small amino acids (such as Ala and Gly). Also they contain a number of Tyr residues such as Tyr^{439} and Tyr^{494} which are conserved in SH2B1 isoforms, and may contribute to the function of its nearby domains, eg by being potential phosphorylation sites (O'Brien et al., 2003).

SH2B1 and SH2B2 proteins are highly structurally related in the C-terminal SH2 domain with 80% identity but are less similar in both PH domain and N-terminal DD region with 58% and 33% identities, respectively. In contrast, the percentage of similarities between SH2B1 and SH2B3 in the SH2 domain is 72% and in the PH domain is 40% (Nelms et al., 1999).

As a result of alternative splicing of mRNA at the last gene exons, the genes of SH2B1 and SH2B2 generate variant isoforms (Yousaf et al., 2001; Li et al., 2007), whereas no isoform has been discovered yet for SH2B3 protein. The defined isoforms for SH2B1 protein are α, β, γ, and δ (Yousaf et al., 2001). The sequence of these isoforms is identical until Gln631, while the big diversity in between them presents in their carboxyl tail, which is discussed in more detail below. This provides each variant with a unique carboxyl terminus, which may affect the nearby SH2 domain binding specificity or affinity (Rui, 2014; Nelms et al., 1999). In addition, SH2B2 has two isoforms, α and β, which differ in the end of the β isoform which is generated with a truncated C terminal SH2 domain, Figure 1.3.

Although there are strong sequence similarities between them, SH2B2 shows much higher affinity and stimulation to the insulin receptor than SH2B1 (Ahmed and Pillay, 2001), forming different responses to the same specific signal. Moreover the SH2B family members share a similar structural characteristic (SH2 domain) to bind to phosphorylated receptor and kinases, however they show different binding preference: SH2B1 has been reported to interact with phosphorylated Tyr813 in Jak2, whereas SH2B2 binds with the phosphorylated Tyr1158 at the activation loop of insulin receptor (Ahmed et al., 1999; Maures et al., 2007) as a consequence of their oligomeric state. On the other hand SH2B3 has a different function as negative moderator in many signalling pathways such as haematopoiesis and in endothelial cells (Devallière and Charreau, 2011). These observations indicate that SH2B family members (SH2B1, SH2B2, and SH2B3) have functional diversity in cell signalling, not necessarily overlapping or redundant roles.

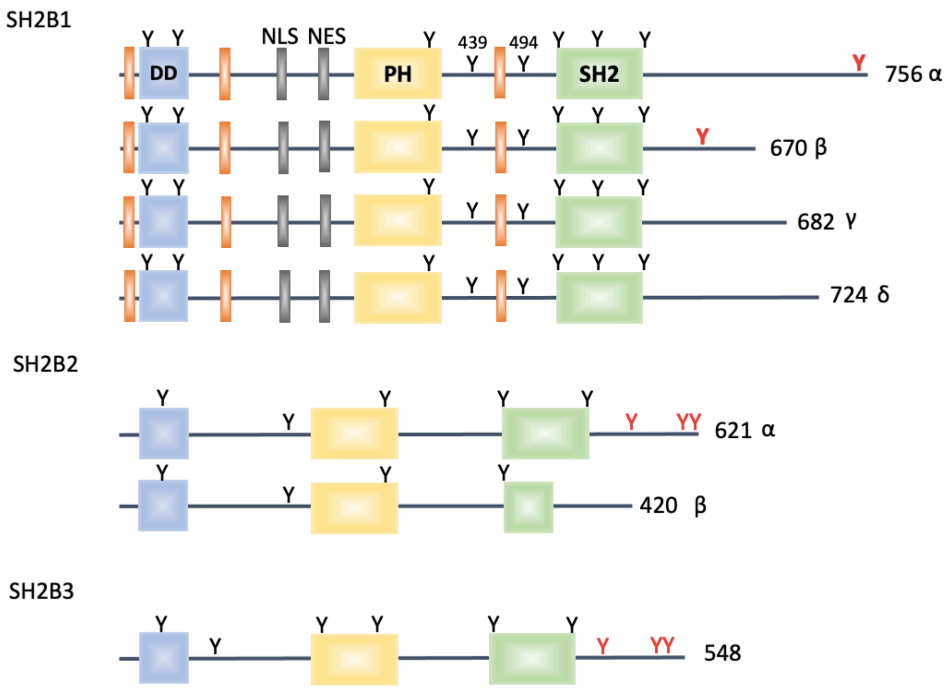


Figure 1.3. A schematic diagram of the SH2B family members SH2B1, SH2B2, and SH2B3 for Mouse. Schematics were drawn on the basis of the protein sequence in uniprot (identifier number SH2B1/Q91ZM2, SH2B2/Q9JID9, SH2B3/O09039). The SH2B1 isoforms (α, β, γ, and δ) and SH2B2 β isoforms are shown. The diagram shows sequences to scale. The boxes indicate the domains within the structure: dimerization domain (DD) in blue, pleckstrin domain (PH) in yellow, scr homology 2 domain (SH2) in green, proline rich sequences in orange, and NLS (nuclear localization sequence) and NES (nuclear export sequence) both in grey. The Y symbols above the sequence represent potential phosphorylation sites. The amino acid length for each protein is shown. The red labelled Tyrosines are the unique residue at the C terminal tail differing between SH2B1 isoforms. The black labelled Tyrosines are the conserved residue in all SH2B members.

1.5 The expression, structure and function of SH2B1 protein

In mammals, SH2B1 cytoplasmic protein is expressed in the central nervous system and peripheral tissues such as brain, liver and skeletal muscle. At the molecular level, neuronal SH2B1 was indicated to control the leptin signalling pathway and indirectly influence insulin sensitivity (Ren et al., 2005; Maures et al., 2007). In support of that, it was proven that the deletion and/or mutation of neuronal SH2B1 genes in mice caused leptin resistance and obesity as well as insulin resistance as a consequence of obesity. Moreover, knockout of peripheral SH2B1 genes leads to insulin resistance and type 2 diabetes regardless of obesity

(Maures et al., 2007, Morris et al., 2008).

To understand more about the function of the protein, the structure features of the protein should be highlighted. As mentioned before, SH2B1 consists of well characterised domains: a DD domain (62 amino acids); a PH domain (110 amino acids); and an SH2 domain (99 amino acids). As well as linkers in between the domains there are a number of tyrosines (eight residues), numerous serines and threonines that are potential phosphorylation sites, plus proline rich sequences (Zhang et al., 2008), Figure 1.3.

In additional to the classical recruitment function of SH2B1 protein as an adaptor to connect a number of proteins to the activated receptor, it has been described as a regulator of activation of receptor tyrosine kinase and non-tyrosine kinase receptors. Although SH2B1 is a cytoplasmic protein (Ren et al., 2005), it has been shown to shuttle between cytoplasm and nucleus (Maures et al., 2007). SH2B1 is a part of the interaction communication pathways, and it is not present in unicellular organisms (Maures et al., 2007).

1.5.1 Src Homology 2 domain (SH2)

The SH2 region is a common domain in cytoplasmic signalling proteins, and it is normally involved in tyrosine kinase signalling pathways. The recruitment of SH2B1 is dependent on mediation by the SH2 domain (527-625 a.a) which is regulated by engaging with tyrosine phosphorylation motifs of the activated target receptor kinase.

From the crystal structure of SH2B1 SH2 domain (5w3r), the fold has been described as a core anti-parallel β-sheet (βB, βC, βD/βD’) followed by two anti-parallel small β-strands (βE and βF), all confined between two α-helices (αA and αB) with the long C-terminus close in space to the N-terminus. In the middle of the carboxyl tail there is a small β-sheet (βG). The N terminal small β-strand (βA) links to αA (Hu and Hubbard, 2006). In addition, the β-strands are linked to each other and to the α-helix via long loops that have a key function in ligand binding at the EF and BG loops as shown in Figure 1.4.

The ability to bind phosphorylated tyrosine residues is common to all SH2 domains as there is 35% identity between all domains (Marengere and Pawson, 1994). However, all SH2 domains do not bind equally (binding affinity and binding residues) to pTyr-containing sequences. In some way, SH2 domains show selectivity for the particular tyrosine phosphorylated residues that they will engage.

The SH2 domain of SH2B1 contains two functional binding regions: a region that recognises pTyr of the target kinase, and a second site that recognises a specific sequence immediately downstream of the pY which is mostly hydrophobic (Bradshaw and Waksman, 2003). The phospho-tyrosine binding cavity of SH2 is built from positive residues: Arg^{555} is the basic conserved residue for recognition of the negatively charge phosphate group (Hu and Hubbard, 2006). Mutation of this critical residue leads to stripping the SH2B1 protein from its ability to bind to phosphotyrosine. This domain has a major role in the adaptor function of the whole SH2B1 protein as discussed below.

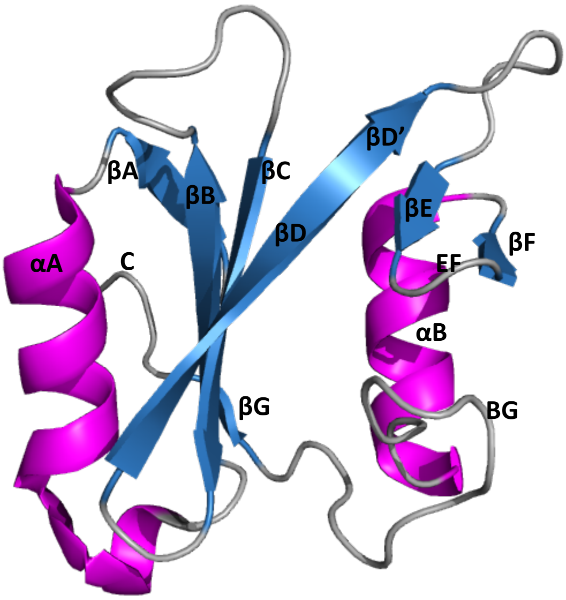


Figure 1.4. Ribbon diagram of SH2B1 SH2 domain based on crystal structure (5w3r) and generated via Pymol program. The α and β helixes are shown in magenta color and the β- strands (βA, βB, βC, βD/βD', βE, βF, and βG) in blue, with the linkage loops in grey, e.g EF and BG.

1.5.1.1 The functional importance of SH2 domain

SH2B1 acts as a linker based on SH2 domain to connect the upstream activated receptor to downstream cytoplasmic signalling, and thus couple a selective group of protein signalling complexes to enhance the catalytic activity of its selectively bound kinase (Maures et al, 2007). As the SH2 of SH2B1 protein has different binding partners, it suggests its adaptor role in different signalling pathways. RTKs are the common substrates for SH2 of SH2B1 (O'Brien et al., 2003; Li et al., 2007).

JAK2 is a tyrosine kinase that mediates leptin signalling by binding to activated cytokine receptors (receptor of long isoform leptin, LERPb). JAK2 recruits to the receptor after activation by associating with its cytokine substrate (growth hormones or leptin), which consequently stimulates a conformational change of JAK2, thus promoting its auto-phosphorylation on multiple tyrosine sites. Neuronal SH2B1 is found to upregulate a leptin signalling pathway through its SH2 which attaches to pTyr^{813} of activated JAK (Morris & Rui, 2009; Li et al., 2007). On the other hand SH2B1 uses its SH2 domain to bind to IRS1 and IRS2, thus recruiting IRSs to the bound JAK2 which in consequence facilitates the phosphorylation of IRSs by JAK2. Phosphorylated residues on IRSs are used as binding sites for other cytoplasmic signalling proteins. Forming JAK2-SH2B1-IRS complexes is able to enhance the downstream PI3-kinase pathway. The binding site of the SH2B SH2 domain for JAK2 and JAK3 of the cytokine receptor-associated kinase has been identified at pTyr^{813}, but the docking point with JAK1 is unknown.

SH2B1 functions as a positive regulator in leptin signalling using multiple mechanisms. First, SH2 of SH2B1 by itself is able to activate Jak2 kinase; the binding between SH2 and JAK2 at pTyr^{813} helps to stimulate the activation of JAK2 as a result of its conformation change. Second, the high affinity between SH2 of SH2B1 and either JAK2 or IRSs prevents JAK2/IRS dephosphorylation as result of blocking the Tyr residue. Third, the strong interaction protects both JAK2 and IRSs from binding with the SH2 domain of inhibitory factors such as suppressor of cytokine signalling (SOCS) or protein tyrosine phosphatase (PTP), respectively (Rui et al., 2000). Finally the leptin-stimulated phosphorylation of IRSs is enhanced by stabilising the formation of a complex of JAK2-IRSs through SH2B1 protein.

A study by Li et al., 2007 indicates that mutation of Arg^{555} to Glu in SH2 domain of SH2B1, or deletion of the SH2 domain in SH2B1, are enough to disrupt the leptin signalling pathway.

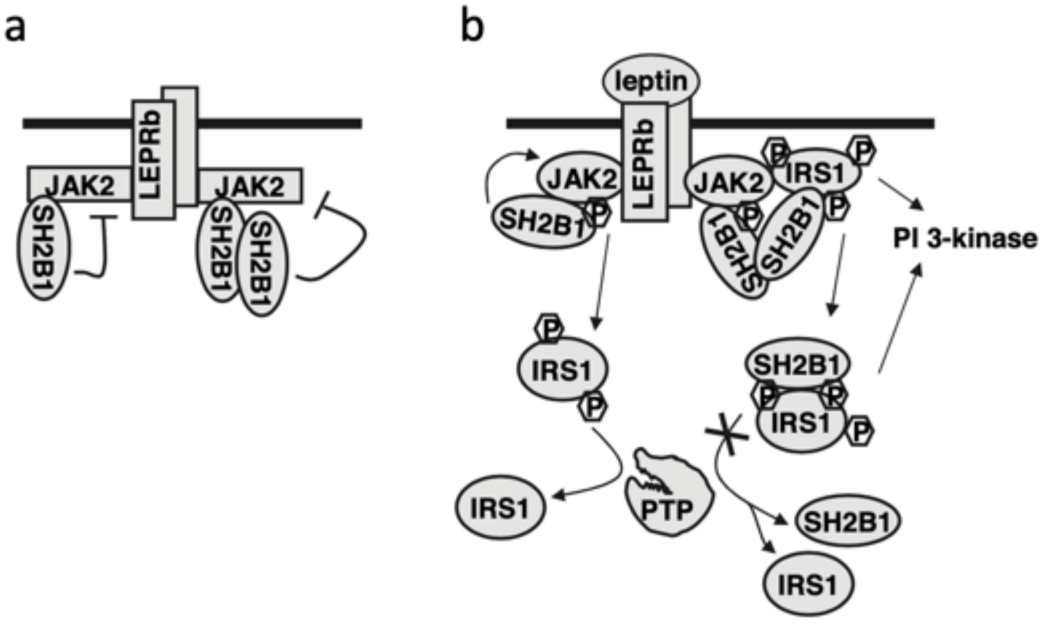


Figure 1.5. Leptin signalling pathway via long isoform leptin of receptor (LEPRb). The diagram is taken from (Li et al., 2007) and shows molecular interaction model of SH2B1 protein through the signalling pathway in the absence and presence of ligand. a). A basal condition in absence of leptin where there is a non-domain interaction with inactive JAK2. b) A stimulation condition in the presence of leptin where there are interactions with active JAK and IRSs thus enhancing downstream signalling. Interaction of SH2B1 with IRSs works to prevent IRSs from dephosphorylation via protein tyrosine phosphatase (PTP). PI3-kinase is phosphatidylinositol 3-kinase signaling pathway, and P is a phosphate group.

SH2B1 protein has a role in regulation of glucose homeostasis, as it is a key player or enhancer in insulin signal transduction, through activation of the insulin receptor (Duan et al., 2004). Binding of insulin to the extracellular α subunit of insulin receptor changes its conformation, and brings the intracellular tyrosine kinase domains (β subunit) close enough to phosphorylate each other, and thus create a direct docking site to the SH2 domain of SH2B1, pTyr^{1158}. The insulin receptor possesses a kinase activity, which is stimulated via binding to SH2B1 protein. Therefore receptor substrate 1 & 2 (IRS) proteins bind by their phosphotyrosine binding domains (PTB) to activated insulin receptor. As a consequence of binding IRS 1 and 2 to the activated IR, they are phosphorylated on a number of their Tyr residues which also become attachment points for the SH2 domains of SH2B1 and other downstream kinases such as IRSs with phosphatidylinositol 3-kinase signalling pathway, and Shc with mitogen activated protein kinase (MAPK) as shown in Figure 1.6 (Fritsche et al., 2008; Morris et al., 2009). Binding SH2B1 to IRS 1 or 2 protects them from dephosphorylation on tyrosines, thereby strengthening the IRS protein mediated pathway.

SH2B1 proteins promote insulin sensitivity via enhancing the insulin receptor catalytic activity which sequentially activates the downstream signalling of insulin receptor.

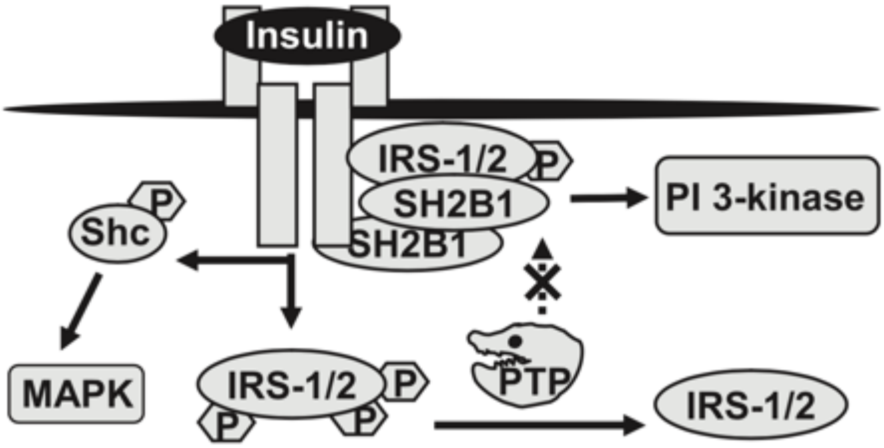


Figure 1.6. A model of the insulin signalling pathway. The diagram is taken from Morris et al., 2008 and displays the activation of IR in response to binding of insulin. The IR starts to recruit SH2B1 via its SH2 domain and IRS 1 and 2 which are phosphorylated and bind to another SH2B1. PTP is protein tyrosine phosphatase, P is a phosphate group, PI3-kinase is phosphatidylinositol 3-kinase, Shc is transforming protein, MAPK is mitogen activated protein kinase.

1.5.2 Pleckstrin homology domain (PH)

Cytoplasmic signalling proteins need to be anchored to the intracellular membrane surface to be close to their interacting or targeting receptors, which are localized in the plasma membrane. Therefore such proteins often contain a domain or membrane anchor that attaches it to the cell membrane, such as the pleckstrin homology domain (PH) (Lemmon, 2007).

At least 61% of studied PH domains are able to interact with membrane phospholipids, and thus target the protein to the plasma membrane (Lenoir et al., 2015). SH2B1 protein is found to contain a PH domain in the middle of the sequence (267-376 a.a), which is predicted to be a lipid binding domain. The method of using a PH domain to localise the SH2B1 protein onto the membrane to be close to the targeting transmembrane receptor is a very reasonable strategy, as it makes the searching and the binding of SH2 domain to the activated receptor fast and robust (Williamson, 2012). Supporting that, it was proven that SH2B1 protein is anchored to the membrane in nerve growth factor (NGF/TrkA) signalling; this membrane association is mostly mediated by the PH domain (Rui et al, 1999b).

Moreover this domain (PH) could have an adaptor function, as the Rui et al., 2000 study indicates that PH with other amino acid regions located in 410 to 555 in SH2B1 protein are essential for the interaction with inactive JAK2. This would have the effect of increasing the protein concentration around the target receptor, which helps to make the binding action via its SH2 domain to the activated receptor rapid and robust.

Furthermore Duan et. al, 2004a suggest that the PH SH2B1 domain could mediate the binding of SH2B1 protein to non-phosphotyrosine regions of IRS 1 and 2, as its deletion impairs the formation of the SH2B1-IRS complex. Both domains SH2 and PH bind independently to IRS proteins, however both domains are required for the full interaction.

A recent study of the function of PH SH2B1 domain (Flores et al., 2019) indicates that in vivo, the PH domain has an essential role in regulation of energy balance and glucose homeostasis, while in vitro the PH domain is able to change the cellular distribution of SH2B1β isoforms, and in PC12 cells the PH domain stimulates NGF neurite outgrowth. However the exact mechanism of PH domain function needs more investigation.

Figure 1.7 illustrates the structure of PH domain of SH2B1 protein as predicted using the Phyre2 structure prediction program (Kelley et al., 2016), and it is close to the NMR structure of the PH domain of SH2B2 protein (Koshiba et al., 1997).

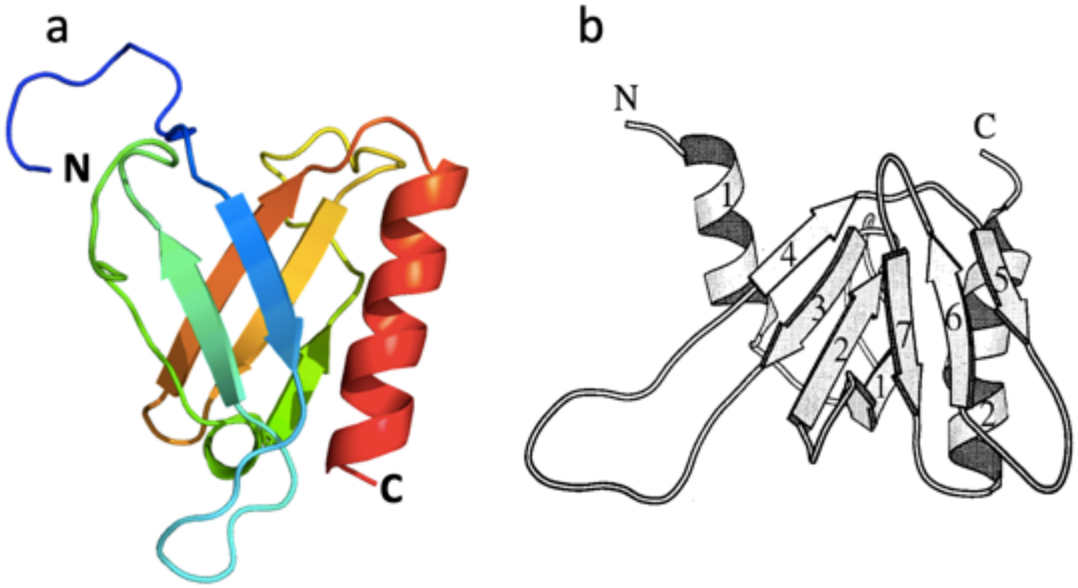


Figure 1.7. The structure of the SH2B1 PH domain of SH2B1. a) The predicted 3D structure of PH domain of SH2B1 using Phyre2 bioinformatics tool. The structure of the SH2B1 PH domain is displayed by PyMol software; the domain consists of two antiparallel β-sheets formed from five β-strands terminating with a C-terminal α-helix. The rainbow colour shows the colour from the N to the C terminus in dark blue to red. b) Ribbon of the NMR structure of PH domain of SH2B2 protein, taken from Koshiba et al., 1997.

1.5.3 Dimerization domain (DD)

Dimerization is a common physical interaction in related proteins which is important for protein-protein interaction and is often seen in transducing signals for regulation (Klemm et al., 1998). SH2B1 is able to dimerise as a homodimer with self-isoforms or a heterodimer with other isoforms such as SH2B2 (51% identity) by mediation of the amino terminal domain (24 to 85 a.a) which has been described as a phenylalanine zipper (Nishi et al., 2005). The topology of the SH2B1 dimerisation domain has been described as a U bisecting region and identified as two α helices (αA and αB) linked by a small β-turn, Figure 1.8.

The functional consequence of SH2B1 dimerization was indicated to promote JAK2 activation leading to enhanced downstream signalling in either the presence or absence of the activating ligand. It was proven that SH2B1 binds weakly at two binding sites (269-410 a.a and 410-555 a.a, which includes the DD sequence and PH domain) to inactive JAK2. Therefore increasing the number of SH2B1 around inactive JAK2 as a result of lower affinity binding between them enables the SH2 domain to bind strongly to JAK2 when it is phosphorylated (Rui et al., 2000).

SH2B1 homo and hetero-dimerization was found to increase the autophosphorylation of JAK2s, which suggests that a SH2B1 dimer promotes the dimerization of bound JAK2, and thus stimulates the transactivation event (Nishi et al., 2005). On the other hand, Maures, et al., 2007 show that a single SH2B1 is effective to enhance the activation of bound JAK2, therefore the N terminal dimerization event of SH2B1 works to stabilise the active state of JAK2. The same idea can be applied to the activation of IR via dimerization of SH2B1, as dimerisation of SH2B1 leads to insulin signalling stimulation by catalysis of IR auto-phosphorylation, decreases the de-phosphorylation of IR, and increases the formation of the complex of IR with IRS 1 and 2.

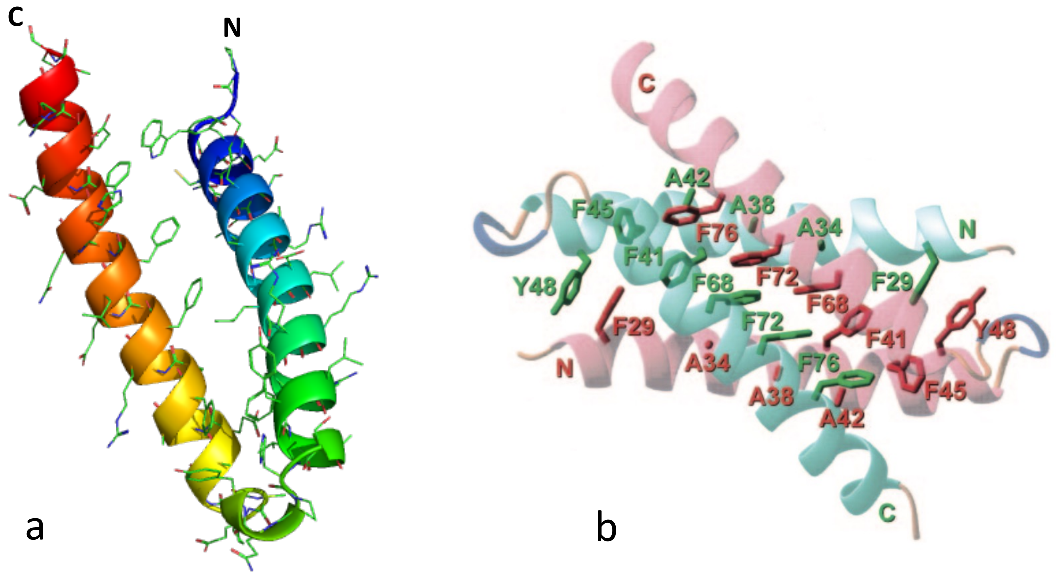


Figure 1.8. The structure of dimerization domain of SH2B1 protein. a) Model of the dimerization domain DD of SH2B1 was generated by Phyre2 bioinformatics modelling tool. The molecule is displayed using Pymol program with rainbow colour; the colour from the N to the C terminus in dark blue to red. b). Predicted structure of dimer state of SH2B1 DD domain via MODELLER program and the crystal structure of SH2B2 DD domain illustrating the amino acid components, green color represents one molecule while pink color represents the other one. Figure taken from Nishi et al., 2005.

1.5.4 Non-domain regions (linkers in between domains)

Residues and specific sequence motifs in interdomain linkers may contribute to the function of their linked functional domains or have independent adaptor function which is associated with their intrinsically disordered nature.

SH2B1 has eight tyrosines and multiple serine and threonine residues in linkers that are potential phosphorylation sites (Figure 1.3). The conserved tyrosine 439 and 494 (YXXL) are able to be phosphorylated by JAK 1 and 2, also the phosphorylation of these residues is stimulated via their targeting the tyrosine kinase receptor, therefore providing a binding site for other SH2 containing proteins (O'Brien et al., 2003; Riedel et al., 1997). Moreover Ser^{96} is phosphorylated by mitogen activated protein kinase (MAPK) which could be another binding site (Rui et al., 1999). The function of the posttranslational modification of these residues is still unknown.

There are a number of proline rich sequences in the linker regions as shown in Figure 1.3. These polypro regions are likely substrates for SH3 domain-containing molecules. As an example the proline sequence in between SH2 and PH domains is predicted to bind to SH3 of Grb2 protein which has a role in the signalling pathway of NGF (Qian et al., 1998).

SH2B1 is present mainly in the cytoplasm, and was shown to shuttle to the nucleus. SH2B1 contains a nuclear localization segment (NLS) (KLKKR^{152}) which is mainly required for its nuclear translocation. Furthermore, it contains a nuclear export signal (NES) (GERWTHRFERLRLSR^{234}). These two regions in between DD and PH domains are conserved in all SH2B1 variants as shown in Figure 1.3. The possible explanation for the nuclear cytoplasmic shuttling of SH2B1 and the presence of the NLS and NES in its sequence has been suggested to be for nucleocytoplasmic function such as transcriptional activation or repression in and/or out of the nucleus (Maures et al., 2009; Chen and Carter-Su, 2004).

1.5.5 Unique Carboxyl sequences of SH2B1 isoforms

The diversity in the C-terminal region (after Gln631) provides each spliced variant of SH2B1 (α, β, γ, and δ) with a unique carboxyl terminus in sequence length and residue type. Alternative splicing events of mRNA play an essential role in formation of proteomic and functional diversity in multicellular organisms (Blencowe, 2006). However the actual functional difference between multiple isoforms of SH2B1 has not yet been investigated.

In brain all four isoforms are found to regulate body weight through energy synthesis and glucose metabolism. On the other hand, one study (Pearce et al., 2014) showed that the β isoform is sufficient to enhance nerve growth factor-induced neurite outgrowth of PC12 cells, whereas the α isoform doesn't have this enhancement ability.

Early studies (Yousaf et al., 2001; Nelms et al., 1999) showed that the SH2B1 variants are expressed widely in cells. In mouse, expression of α and δ isoforms is mostly coupled, in lung, spleen, kidney, thymus and skeletal muscle with a prominence of δ over α in brain and early embryo, while α is more abundant than δ in testis. Moreover, β and γ variants are expressed in pairs in ovary, heart and testis with predominantly β expression, with minimum expression for both variants in other tissues such as lung, brain, thymus, and skeletal muscle (Figure 1.9). Also the distribution of SH2B1 variants in different cells is unalike, as shown by the Doche et al., 2012 study, which shows that the expression of α and δ is mainly in the brain, while the expression of β and γ is ubiquitous.

As a result of these observations it was suggested there is a possible contribution or similarities in function between α and δ, and β and γ. Yousaf et al*.,* 2001 suggested that the expression of SH2B1 isoforms depends on an alternative splicing mechanism that is present in different cells. The transcription of the four isoforms from exon 1 until the first 168 nucleotides of exon 7 is identical, however the 3' end of exon 7 produces different variants as a result of alternative splicing.

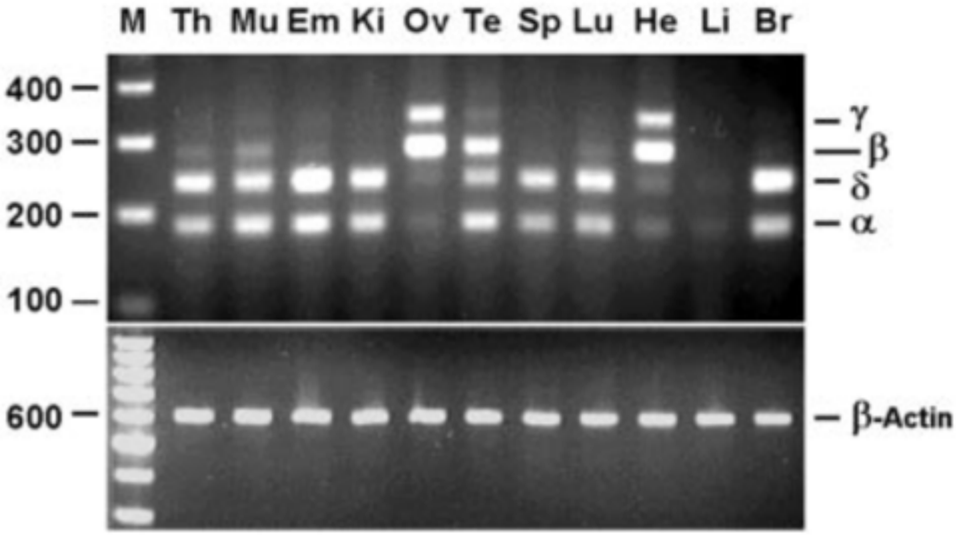


Figure 1.9. Agarose gel of distribution of SH2B1 variants (a, β, γ, and δ) in different mouse tissues. Fragment size of each isoform is; (α 185 bp , β 258 bp , γ 338 bp , δ 238 bp), the gel is taken from Yousaf et al., 2001. The tested mouse tissues are: (from the right) brain, liver, heart, lung, spleen, testis, ovary, kidney, embryo, skeletal muscles, and thymus. The size of the marker is in base pairs.

Moreover a recent study (Joe et al., 2017) proved the contribution of the C terminal tail of α isoform to restrict the ability of the SH2B1 to shuttle through the nucleus, and to inhibit other enhancement abilities such as autophosphorylation of NGF receptor TrkA and phosphorylation of Akt. The difference in function of SH2B1 isoforms is mostly due to their distinct carboxyl termini, thus highlighting the main sequence features can help to understand their functional differences.

In mice and rat, the unique C-terminus of SH2B α and β contains a tyrosine residue, Tyr^{753} and Tyr^{649}, respectively, which may provide sites for signalling molecules which interact uniquely with one isoform, and was suggested to be a site of phosphorylation that may influence the interaction of the adjacent SH2 domain with activated receptor Tyr kinases (Nelms et al., 1999). The phosphorylation of Tyr^{753} at the C terminus of α isoform is able to inhibit the function of the rest of the protein (Joe et al., 2017). The C-terminal sequence RSTSRDP^{639} in the α isoform is similar to the Raf binding sequence for 14-3-3 proteins. Moreover, the sequence PDASSTLLP^{658} in the β isoform at the C-terminus resembles the binding motif of 14-3-3 proteins in the C-terminus of the insulin-like growth factor substrate I (Craparo et al., 1997).

The α variant contains two interesting C-terminal motifs: the sequence NXXY^{754} is similar to the NPXY motif of insulin receptor which is recognized by the phosphoserine binding (PTB) domain of IRS1 proteins (Eck et al., 1996). Also, the SFV^{756} motif is similar to the recognition motif of PDZ domains S/TXV that has been described in a number of regulatory and signalling molecules (Harrison, 1996). Moreover, α, γ and δ isoforms have additional proline-rich sequences located at the C-terminus which is a possible attachment point to other SH3-containing signalling systems. The carboxyl terminus of the δ isoform carries two additional sequences that are similar to nuclear localization sequences (KRRR^{678} and KRGKRKRKR^{700}) which may indicate its role in gene expression, also it contains serine/tryptophan rich sequences (Yousaf et al., 2001).

In addition, it was demonstrated in different studies that the insulin receptor contains a number of docking sites for SH2B1: the γ-isoform interacts with the insulin receptor activation loop β domain at Tyr^{1146} (Nelms et al., 1999), whereas the α-isoform interacts with the insulin receptor activation loop at pTyr^{1158}, pTyr^{1162}, and pTyr^{1163} (Morris et al., 2008). The β-isoform has been reported to bind to activate JAK2 in response to hormones and cytokines (Rui and Carter-Su, 1999).

The existence of unique motifs at the C terminal tail supports the hypothesis of the recruitment function of these tails to specific proteins for signalling assemblies; also the phosphorylation of tyrosine residues in the C terminus of α is suggested to prevent the protein from entering the nucleus and enhancing the NGF signalling pathway (Joe et al., 2017). Although the sequences of SH2B1 isoforms are well-known, the function of the C-terminal diversity is poorly understood.

```
               601                           631               649
SH2B1-α        LEHFRVHPIPLESGGSSDVVLVSYVPSQRQQERSTSRDPAQPSEPPPWTDPPHPGAEEAS
SH2B1-β        LEHFRVHPIPLESGGSSDVVLVSYVPSQRQQGREQAGSHAGVCEGDRCYPDASSTLLPFG
SH2B1-γ        LEHFRVHPIPLESGGSSDVVLVSYVPSQRQQGEQSRSAGEEVPVHPRSEAGSRLGAMQGC
SH2B1-δ        LEHFRVHPIPLESGGSSDVVLVSYVPSQRQQGEQSRSAGEEVPVHPRSENGAPPVTQPSP
               *******************************

SH2B1-α        GAPEVAAATAAAAKERQEKEKAGSGGVQEELVPVAELVPMVELEEAIPGTEAQGGAGSS
SH2B1-β        ASDCVTEHLP--------------------------------------------------
SH2B1-γ        ARATDATPMPPPPSCPSERVTV--------------------------------------
SH2B1-δ        LNPLHGQIPHILGQKRRRGRQKLRQPQPQQPKRGKRKRKRAVEGSRKSWSPWLSWSPWLN

                                               753
SH2B1-α        GDLEVSLMVQLQQLPLGGNGEEGGHPRAINNQYSFV 756 a.a
SH2B1-β        ------------------------------------ 670 a.a
SH2B1-γ        ------------------------------------ 682 a.a
SH2B1-δ        WKRP ------------------------------  724 a.a
```

Figure 1.10. The Carboxyl-terminal amino acid sequence of SH2B1 isoforms α, β, γ, and δ (in mouse) with the length of each variant. The protein sequence until Gln/Q 631 is conserved in all isoforms (grey highlighted). Yellow highlights Tyr (649-753) residues; and the underlined bold sequences are unique to α and β isoforms. The blue highlighted sequences are present only in the δ variant. The motifs in boldface (NXXY and SFV) are identified only in α. Pro-rich sequences are indicated by a waved underline in α and γ and a Ser-Trp in δ is indicated in double underline. The GenBank accession numbers are: mouse α (AF421138), mouse β (AF020526), mouse γ (AF421139), mouse δ (AF380422).

1.6 Roles of intrinsically disordered regions (IDRs) in signalling proteins

The majority of proteins in eukaryotic cells are made up from both structured and unstructured regions (Van Der Lee et al., 2014). Polypeptide segments of proteins that are defined as intrinsically disordered are characterized by their amino acid or sequence composition: less complexity of sequence, low ratio of poly-hydrophobic sequences (I, L, and V) and aromatic residues (F, Y, and W), big number of charged and polar amino acids, and lower content of cysteine and asparagine residues (Babu, 2016; Habchi et al., 2014). The lack of bulky hydrophobic amino acids in the sequence of IDRs, which are order-prompting residues, means that there is no need to prevent the unfavorable interactions with water, thus these sequences lack a stable and defined 2D/3D structure at least *in vitro* under physiological conditions. Regions exhibiting intrinsically disordered properties in proteins are located mainly in between structured domains, and/or at the N and C terminal regions of the protein with sequence length 30 residues or more (Babu, 2016). Alterations or mutations within the disordered segments often lead to developmental diseases in humans (Uversky et al., 2008), as their function is mostly complementary to the structured region of the proteins.

Beside the classical function of IDRs to link different structured parts of the protein, IDR segments are abundant in regulatory proteins that are involved in signal transduction as essential components in signalling machinery. The importance of IDR segments comes from their molecular assembly function, as they contain binding sites or short sequence motifs, which mediate weak interactions with other proteins and thus modulate the formation of signalling protein complexes, thereby increasing specificity and improving the regulation of signalling pathways. Moreover due to their conformational change, they are able to bind to different targeting proteins (Mittag et al., 2010).

Furthermore the intrinsically disordered regions are often regulated via posttranslational modification of accessible sites within IDRs (Bah et al., 2015; Trudeau et al., 2013). An example of regulation controlled by IDRs is autoinhibition. The mechanism is adopted by auxiliary multidomain proteins that are involved in signal transduction to turn them off when not activated, in order to not cause havoc in the cell by stopping unwanted or abnormal activities. The autoinhibition occurs as a result of intramolecular interaction between the binding site (e.g phosphorylated tyrosine) within the IDR with the functional structured part of the protein in order to block activator binding sites (Trudeau et al., 2013).

Such regulatory interactions have been characterised in many signalling proteins, as shown in Figure 1.11. Src kinases contain a C-terminal kinase domain, and N-terminal SH2 and SH3 domains. The inactive kinase (under basal conditions) is repressed via two intramolecular interactions; the SH2 domain interacts with a C-terminal phosphorylated tyrosine, while the SH3 domain interacts with a poly-proline sequence located in between kinase and SH2 domains. Both interactions work to stabilise the inactive conformational state of the kinase by restructuring the kinase active site. The repressed state is disturbed when one or both domains bind their preferred external ligands, thus releasing the kinase to its active state (Lim, 2002).

A similar autoinhibition model appears to be adopted by SH2B1 α isoform via intramolecular interaction with a C terminal tyrosine 753. The phosphorylation of Tyr^{753} is able to repress the nucleocytoplasmic function of the protein by stopping nuclear cycling (Joe et al., 2017). Mutation of this residue Tyr753 restores the protein's ability to cycle through the nucleus, although the exact interactions involved in the inhibitor mechanism are still unclear.

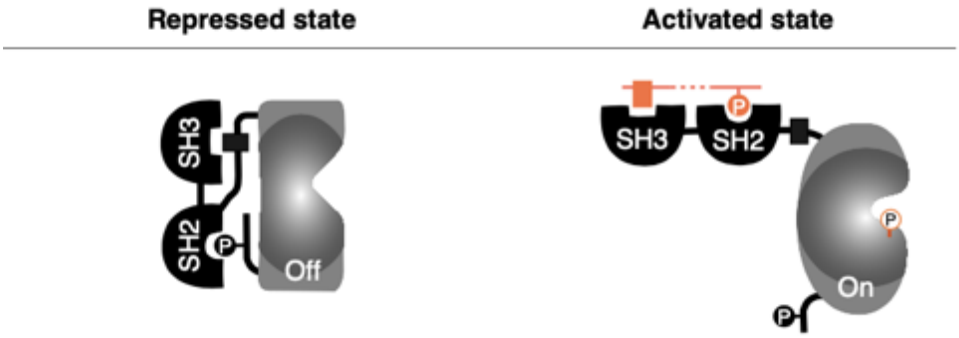


Figure 1.11. Schematic illustration of the autoinhibition mechanism of Src-kinase protein. The schema shows two states of inactive and active Src-kinase; the kinase under basal conditions is repressed by intramolecular interactions involving the SH2 and SH3 domains, whereas intermolecular interactions release the kinase to be in active condition. The schema is taken from Lim, 2002. P in red is the phosphorylated tyrosine, and red square is polyproline sequences.

IDRs are able to adopt different conformation states, which are both flexible and dynamic. This property makes disordered segments useful for regulatory functions. That is because IDRs enable fast association because the interaction is often intramolecular, and IDRs are exposed and therefore have few orientational limitations; and due to the large degree of mobility, IDRs enable fast dissociation, which is clearly an advantage for signal transduction since speed is often more important than strong affinity (Williamson, 2012).

In addition it is commonly observed that IDRs, either for the whole region or for a short distinct segment within IDRs, undergo ordering transitions upon binding to partners. Preformed structural elements work as conformational selection to permit intermolecular contact with specific partners, exposure of a specific phosphorylation site, or to stabilize the active conformation state in weak transient binding (Trudeau et al., 2013; Mittag et al., 2010).

1.7 Study aims

This thesis describes an investigation into the structure and function of the C-terminal SH2 domain of mouse SH2B1, in particular into interactions with the intrinsically disordered C-terminus of the β isoform. As mentioned in section 1.5, SH2B1 isoforms (α, β, γ, and δ) share the same known functional structured domains (N-DD-PH-SH2-C), however there is a big diversity in sequence length and residue type in the carboxyl tail of each isoform (Figure 1.10). Despite the good knowledge about the sequence of SH2B1, there is very little understanding about the function of the intrinsically disordered C-terminus of the different isoforms.

The investigation reveals that SH2B1 β isoform C-terminal tail of mice contains a phosphorylatable amino acid, Y649. This observation raises the question of whether the phosphorylation of this tyrosine has a regulatory role in the activation of the nearby SH2 domain, most obviously a role in autoinhibition. To contribute to understanding the self-regulation mechanism of SH2B1 β isoform we aimed to

1. Optimise the protein expression and the purification to obtain a good concentration of labelled sample of SH2 with the C-terminal tail of β (160 residues) and SH2 protein (118 residues), to aid NMR studies.

2. Phosphorylate the Tyr139 at the C terminal tail of the SH2c in *vitro* by an activated kinase to study the intramolecular interaction upon binding of the phosphorylated tyrosine with SH2 domain using NMR spectroscopy.

3. Measure the binding affinity between the C-terminal phospho-tyrosine ligand and SH2 domain using NMR titration experiments, as sequence analysis for the C-terminal tail of β isoform reveals that it differs from the ideal SH2 high-affinity ligand in JAK2 (O'Brien et al., 2003), and based on that it is expected to bind with only low affinity.

4. Conduct an outcompeting experiment based on titrating the bound SH2-C terminal peptide with JAK2 ligand, because the high effective concentration for the weak binding of the

intramolecular partner does not replace the low concentration of the favourable interaction of the intermolecular partner, thus binding of a C-terminal phosphorylated peptide is predicted to be readily outcompeted by a high-affinity external kinase ligand.

5. The C-terminal long tail of β (45 residues), which is intrinsically disordered, may undergo disorder to order transition upon binding pY to the N terminal SH2 pY pocket for additional support of the weak intramolecular binding by conformational change. This hypothesis was not studied because of unexpected low protein concentration which was not enough to achieve a complete chemical shift assignment.

6. This thesis aims to study the possibility of using automated software (FLYA/CYANA) for a completely automatic NMR structure determination of SH2 protein by a student, and demonstrate the value of the *Accuracy of NMR Structure using Random Coil Index and Rigidity* (ANSURR) method in guiding restraint improvement.

Chapter 2 Materials and Methods

The methods and reagents for various techniques used are given below; any changes employed are described in the relevant sections. All chemicals and reagents used were of analytical grade and purchased from Sigma Aldrich, Novagen, BIO-RAD and Thermo Fisher Scientific. Restriction enzymes and NEBuilder HiFi DNA assembly master mix were obtained from New England Biolabs. All solutions were made using Milli-Q water purified using the Milli-Q system from Millipore. Growth media were made following the lab protocols, using distilled water and sterilized by autoclaving or filtration through 0.2 µM or 0.45 µM filters.

2.1 Culture Media

Table 2.1. Growth medium recipe.

Media type	Reagent per litre
Agar growth	15 g Nutrient Agar
Amp-agar	15 g Nutrient Agar, 1 ml Ampicillin (100 mg/1 ml)
Luria-Bertani (LB)	10 g Bacto Tryptone, 5 g Bacto Yeast Extract, 10 g NaCl.
Amp-LB	10 g Bacto Tryptone, 5 g Bacto Yeast Extract, 10 g NaCl, 1 ml Ampicillin (100 mg/1 ml)
Super Optimal Broth (SOC)	20 g Bacto Tryptone, 5 g Bacto Yeast Extract, 0.5 g NaCl, 0.19 g KCl, adjusted pH to 7, 20 ml of 1M Glucose , 10 ml of 1 M $MgCl_2$, 10 ml of 1 M $MgSO_4$.
M9 Minimal	6 g Na_2HPO_4, 3 g KH_2PO_4, 0.5 g NaCl, adjusted pH of the solution to 7.4, after autoclaving the components in table 2 were added.

Table 2.2. M9 Minimal media recipe

Reagent	Amount per litre
Trace elements (recipe in table 3)	650 µl
1 mg/ml Thiamine	1000 µl
1 M $MgSO_4$	1000 µl
1 M $CaCl_2$	100 µl
1 M Ampicillin	1000 µl
^{15}N source $^{15}NH_4Cl$	1 g dissolved in 4 ml distilled water
^{13}C source ^{13}C-glucose or unlabelled glucose	3 g dissolved in 10 ml distilled water (30% w/v glucose)

Table 2.3. Trace elements recipe.

Reagent	Amount per 100 ml
$CaCl_2.2H_2O$	550 mg
$MnSO_4.H_2O$	140 mg
$CuSO_4.5H_2O$	40 mg
$ZnSO_4,7H_2O$	200 mg
$CoCl_2.6H_2O$	45 mg
$Na2MoO_4.2H_2O$	26 mg
H_3BO_4	40 mg
KI	26 mg

The pH of the solution adjusted to 8.0, and subsequently added EDTA 500 mg. Then pH was re-adjusted to 8.0 and lastly added 375 mg from $FeSO_4.7H_2O$. Finally the final volume was adjusted to 100 mL and autoclaved.

2.2 Bacterial Strains and Expression Vectors

Escherichia coli strains were used: NEB 5-alpha competent cells for high efficiency transformation and BL21 (DE3) for all protein expression (MBP-His-SH2, His-SH2 short protein sequence, His-SH2 long protein sequence, SH2c, and SH2c Y114F). Plasmids used were pET expression vector (6880 bp) which was given and designed by Dr Mesnage's lab (Figure 2.1), and pET-15b (5708 bp) plasmid from GenScript (Figure 2.2).

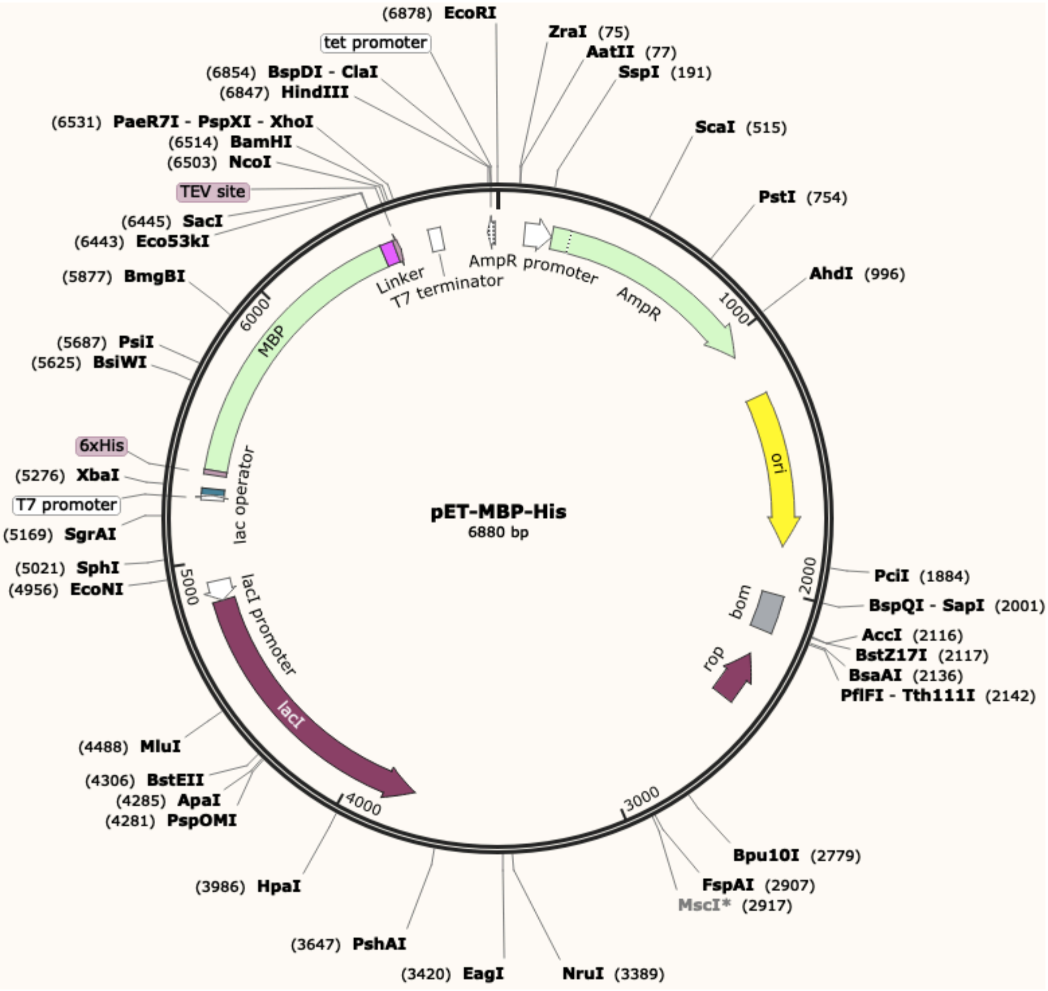


Figure 2.1. pET *E. coli* Vector Map, 6880 bp, showing the location of restriction sites, coding site and origins of replication and translation. The plasmid was designed with N-terminal 6xHis-MBB-TEV tag.

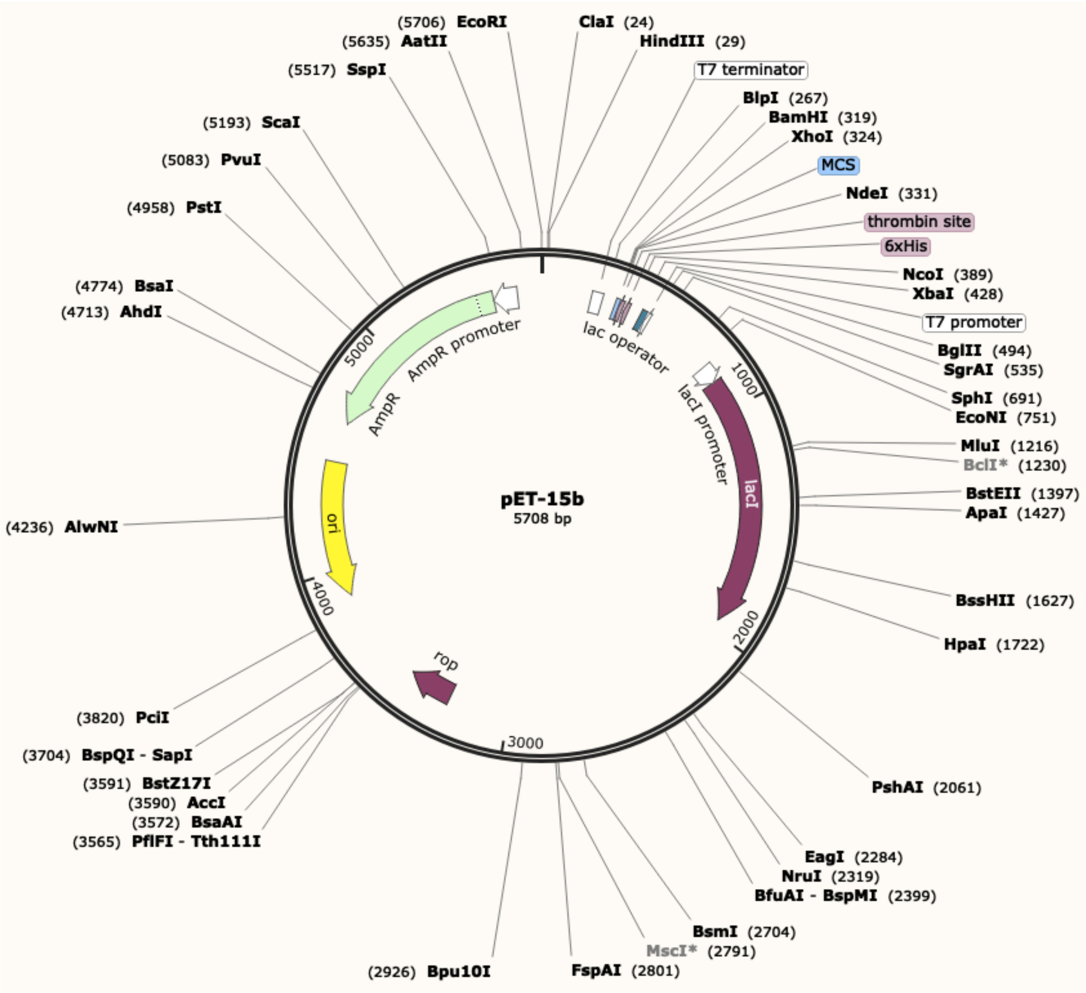


Figure 2.2. pET-15b *E. coli* Vector Map, 5708 bp, illustrating the location of restriction sites, coding regions and origins of replication and translation. The vector includes a N-terminal 6xHis tag.

2.3 Transformation of bacterial strains

E. coli was transformed by adding 5 µl of plasmid (20 to 25 ng/µl) to 50 µl of competent cells. The mixture was incubated on ice for 30 mins before a heat shock in water bath at 42°C for 10 to 30 sec. The culture was placed on ice for 5 mins, then 950 µl SOC medium was added to the mixture followed by an hour incubation shaking at 37°C. 50 µl and 100 µl were taken from the mixture and plated on agar plates containing 50 µg/ml ampicillin. The plates were incubated at 37°C overnight.

2.4 DNA methods

2.4.1 Plasmid amplification, purification and sequencing

NEB 5-alpha *E. coil* competent cells were transformed and plated on Amp-agar (table 2.1) followed by incubation overnight at 37℃. For plasmid amplification, one colony was used to inoculate 10 ml of Amp-LB (table 2.1) which was incubated with overnight shaking at 37℃. QIAprep Spin Miniprep Kit from QIAGEN was used to purify plasmids from 10 ml overnight cultures by following the manufacturer's protocol except that in the final step, the plasmid was eluted with 30 µl of distilled water instead of elution buffer.

20 µl of the resultant plasmid stock (30 ng to 50 ng/µl) was sent to eurofins Genomics for sequencing. The sequencing was done either in both or one directions, using forward T7 promoter and pET reverse primer. The resulting data was visualized and analysed with SnapGene software.

2.4.2 Restriction digestions of DNA

The pET-MBP-His plasmid (6880 bp) was digested with a pair of digestion enzymes: either NcoI and BamHI to insert MBP-His-SH2 gene, or BamHI and XbaI to insert one of the following genes; His-SH2 short protein, His-SH2 long protein, and SH2c.

The restriction digestion was done in a total volume reaction of 50 µl following recommended manufacturer's protocol. Two reaction mixtures were prepared each one containing: 12 µl DNA (20 to 50 ng/µl), 2.5 µl 10× NEBuffer 3.1 in final concentration 1×, 1 µl of NcoI or XbaI restriction enzyme in one reaction and 1 µl of BamHI restriction enzyme in the second reaction (one unit/50 µl reaction), and 9.5 µl distilled water to make total volume 25 µl for each reaction. 5 µl from each reaction was taken as control samples and from the remaining mixture 20 µl from each reaction was mixed to make 40 µl final volume of complete digest sample. The digestion reaction was carried out at 37°C in a water bath for 30 mins to 1 hour.

2.4.3 Agarose gel electrophoresis (analytical method)

Agarose gel electrophoresis was used to fractionate DNA fragments according to length. Linearized plasmids were identified by 1% agarose gel (recipe given in table 2.4). Each 5 µl of samples were mixed with 1 µl of UView^{TM} 6× loading dye. The DNA samples were run on 1% agarose gel alongside 6 µl of DNA marker (4 µl of dH$_2$O, 1 µl of UView^{TM} 6× loading dye, and 1 µl of 1 kb DNA ladder from New England BioLabs). Gel was electrophoresed at 110 V for an hour in in 1× TAE running buffer and visualized under 254 nm UV light.

Table 2.4. Agarose Gel and Buffer recipe.

	Reagent
1% Agarose Gel	1 g agarose powder, 100 ml 1× TAE of 50× TAE stock buffer
50× TAE Buffer	242 g Tris-base, 57.1 ml of 100% acetic acid, 100 ml of 0.5 M EDTA, distilled water was added up to 1 Litre.

2.4.4 Purification of DNA fragments from Agarose gel

Linear DNA fragment bands within the 1% agarose gel were cut with a scalpel under UV light. The gel slices were transferred into a pre-weighed Eppendorf tube. For 100 µg of the gel slice, 100 µL binding buffer was added. The purification was done using GeneJET Gel Extraction Kit according to the manufacturer's instructions. Incubation of the mixture was done at 60°C until the agarose slice was completely melted. The sample-buffer mixture was pipetted onto a silica membrane column and the column was incubated at room temperature for 1 minute. The DNA was spun down to bind it to the membrane. The flow-through was discarded. 500 µl washing buffer was added to the column. The column was spun down to remove the impurities. The DNA was eluted with 30 µl water. The concentration of extracted linear plasmids was measured at 260 nm using a Nano-Drop Lite Microlitre Spectrophotometer.

2.4.5 Design of homologous overlapping gBlock Gene

The SnapGene software was used to design and generate the g.Block overlapping fragments. The target gene sequence of the SH2 domain of SH1B1 β (Homo species*, Q9NRF2*), SH2 of SH1B1 β (Mouse species, *Q91ZM2*), SH2 with the C terminal tail of SH1B1 β (Mouse species, *Q91ZM2)* were taken from http://www.uniprot.org/uniprot. After that Codon Optimization Tool in Integrated DNA Technologies web was used to optimise codon usage of the obtained genes for *E. coli* to express. pET-MBP-His (6880 bp) plasmid was used as a template to create two matched homology overlapping sequences at each end of the target gene.

The same strategies were followed to design three g.Block constructs for SH2 domain as shown in Figures 2.4, 2.6, and 2.7, and one for SH2 domain with the C terminal sequence of SH2B1 β isoform Figure 2.8. The details of each g.Block gene sequences are discussed in Chapter 3.

The first overlap sequence at the 5' end started from the first restriction site with ~33 bp backward; while the second overlap at the 3' end started from the second restriction site with ~33 bp forward that results in a g.Block overlapping fragment, as shown in Figures 2.3 and 2.5, gray coloured sequences. At the end of the g.Block gene and before the second restriction site start, a stop codon (TAA) was added, and a few nucleotides codon for more residues were added before the Trp residue at the start of the SH2 gene; (Gly, Tyr, Pro) in the first SH2 construct, and (Gly, Ser) in the second SH2 construct, as linker. However those extra residues were not added to the third SH2 construct and to the SH2 with C terminal sequence construct. The total length of g.Block genes are shown in Figures 2.4, 2.6, 2.7, and 2.8. All g.Block genes were ordered from Integrated DNA Technologies.

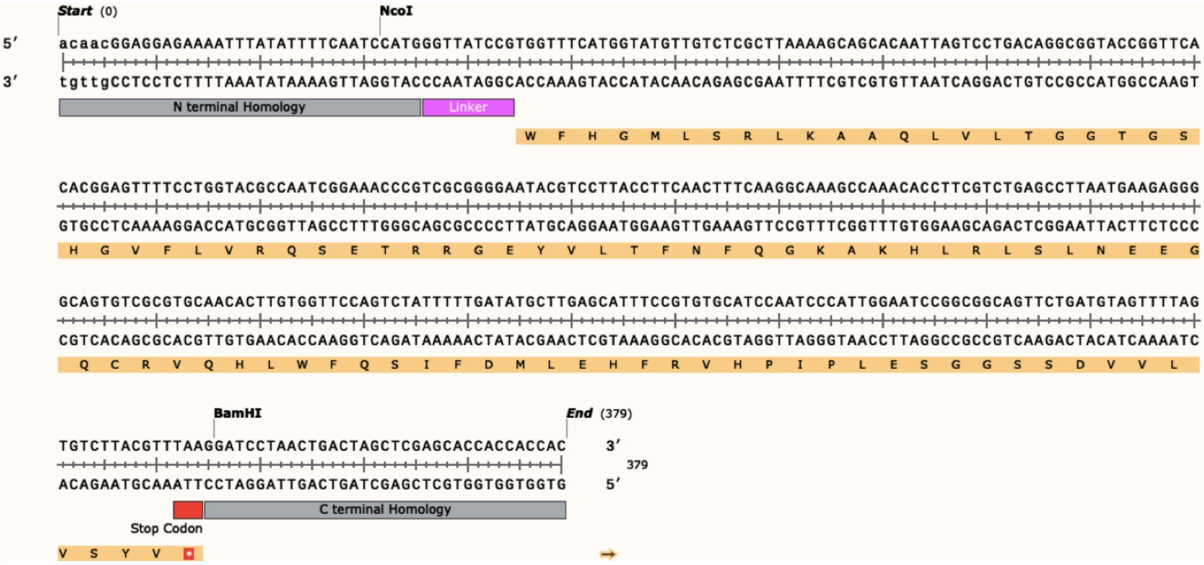

Start (0)
NcoI
5′
acaacGGAGGAGAAAATTTATATTTTCAATCCATGGGTTATCCGTGGTTTCATGGTATGTTGTCTCGCTTAAAAGCAGCACAATTAGTCCTGACAGGCGGTACCGGTTCA
3′
tgttgCCTCCTCTTTTAAATATAAAAGTTAGGTACCCAATAGGCACCAAAGTACCATACAACAGAGCGAATTTTCGTCGTGTTAATCAGGACTGTCCGCCATGGCCAAGT
N terminal Homology
Linker
W F H G M L S R L K A A Q L V L T G G T G S
CACGGAGTTTTCCTGGTACGCCAATCGGAAACCCGTCGCGGGGAATACGTCCTTACCTTCAACTTTCAAGGCAAAGCCAAACACCTTCGTCTGAGCCTTAATGAAGAGGG
GTGCCTCAAAAGGACCATGCGGTTAGCCTTTGGGCAGCGCCCCTTATGCAGGAATGGAAGTTGAAAGTTCCGTTTCGGTTTGTGGAAGCAGACTCGGAATTACTTCTCCC
H G V F L V R Q S E T R R G E Y V L T F N F Q G K A K H L R L S L N E E G
GCAGTGTCGCGTGCAACACTTGTGGTTCCAGTCTATTTTTGATATGCTTGAGCATTTCCGTGTGCATCCAATCCCATTGGAATCCGGCGGCAGTTCTGATGTAGTTTTAG
CGTCACAGCGCACGTTGTGAACACCAAGGTCAGATAAAAACTATACGAACTCGTAAAGGCACACGTAGGTTAGGGTAACCTTAGGCCGCCGTCAAGACTACATCAAAATC
Q C R V Q H L W F Q S I F D M L E H F R V H P I P L E S G G S S D V V L
BamHI
End (379)
TGTCTTACGTTTAAGGATCCTAACTGACTAGCTCGAGCACCACCACCAC
3′
379
ACAGAATGCAAATTCCTAGGATTGACTGATCGAGCTCGTGGTGGTGGTG
5′
C terminal Homology
Stop Codon
V S Y V *

Figure 2.3. The DNA sequence of the first SH2 g.Block designed 379 bp. The Figure shows N terminal and C terminal homology DNA sequences in grey (36 bp backward and 33 bp forward); SH2 protein sequence in orange; a linker DNA sequence in light purple (GYP); a stop codon red; also shows the restriction enzyme sites for NcoI and BamHI.

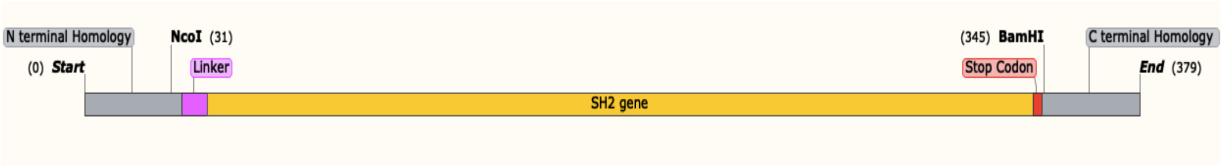


Figure 2.4. Schematic drawing of the first SH2 g.Block gene construct, 379 bp.

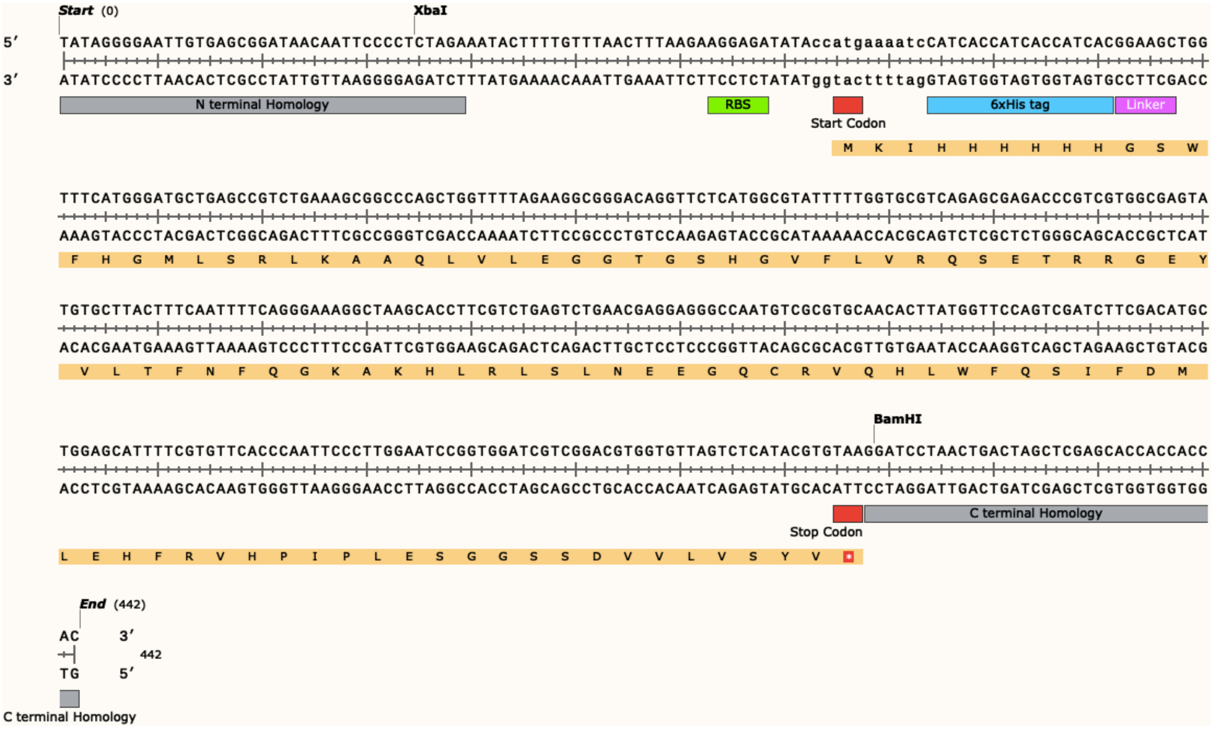

Start (0)
XbaI
5′
TATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATACTTTTGTTTAACTTTAAGAAGGAGATATAccatgaaaatcCATCACCATCACCATCACGGAAGCTGG
3′
ATATCCCCTTAACACTCGCCTATTGTTAAGGGGAGATCTTTATGAAAACAAATTGAAATTCTTCCTCTATATggtactttttagGTAGTGGTAGTGGTAGTGCCTTCGACC
N terminal Homology
RBS
Start Codon
6xHis tag
Linker
M K I H H H H H H G S W
TTTCATGGGATGCTGAGCCGTCTGAAAGCGGCCCAGCTGGTTTTAGAAGGCGGGACAGGTTCTCATGGCGTATTTTTGGTGCGTCAGAGCGAGACCCGTCGTGGCGAGTA
AAAGTACCCTACGACTCGGCAGACTTTCGCCGGGTCGACCAAAATCTTCCGCCCTGTCCAAGAGTACCGCATAAAAACCACGCAGTCTCGCTCTGGGCAGCACCGCTCAT
F H G M L S R L K A A Q L V L E G G T G S H G V F L V R Q S E T R R G E Y
TGTGCTTACTTTCAATTTTCAGGGAAAGGCTAAGCACCTTCGTCTGAGTCTGAACGAGGAGGGCCAATGTCGCGTGCAACACTTATGGTTCCAGTCGATCTTCGACATGC
ACACGAATGAAAGTTAAAAGTCCCTTTCCGATTCGTGGAAGCAGACTCAGACTTGCTCCTCCCGGTTACAGCGCACGTTGTGAATACCAAGGTCAGCTAGAAGCTGTACG
V L T F N F Q G K A K H L R L S L N E E G Q C R V Q H L W F Q S I F D M
BamHI
TGGAGCATTTTCGTGTTCACCCAATTCCCTTGGAATCCGGTGGATCGTCGGACGTGGTGTTAGTCTCATACGTGTAAGGATCCTAACTGACTAGCTCGAGCACCACCACC
ACCTCGTAAAAGCACAAGTGGGTTAAGGGAACCTTAGGCCACCTAGCAGCCTGCACCACAATCAGAGTATGCACATTCCTAGGATTGACTGATCGAGCTCGTGGTGGTGG
Stop Codon
C terminal Homology
L E H F R V H P I P L E S G G S S D V V L V S Y V *
End (442)
AC 3′
442
TG 5′
C terminal Homology

Figure 2.5. The DNA sequence of the second SH2 g.Block designed, 443 bp. The Figures shows N terminal and C terminal homology DNA sequences in grey (33 bp backward and 33 bp forward), Ribosome Binding Sequence in green, Linker-His$_6$-Linker-SH2 protein sequence in orange; a linker DNA sequence in light purple; 6xHistidine tag in light blue, a start and stop codon in red; also shown is the restriction enzyme sites for NcoI and BamHI.

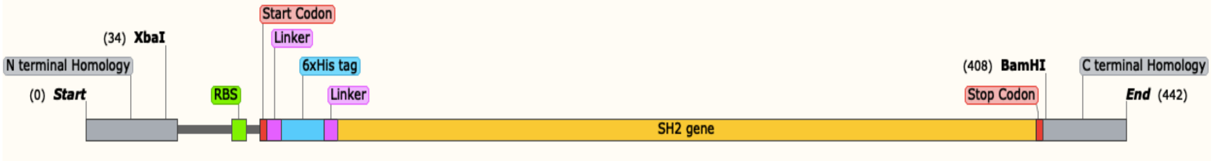


Figure 2.6. An overview map of the second SH2 (short protein) g.Block construct without the MBP tag.

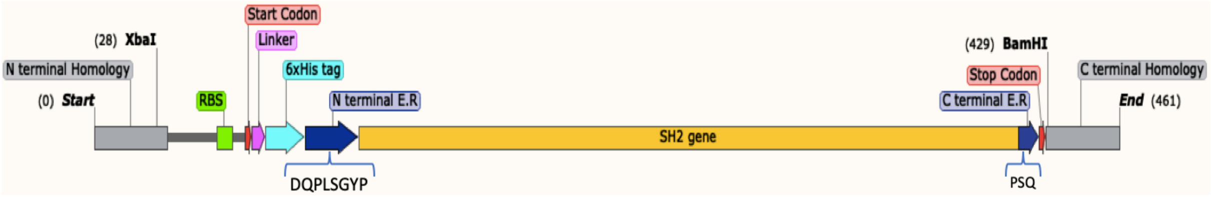


Figure 2.7. An overview map of the third SH2 construct (long protein) 461 pb with the extra DQPLSGYP at N terminus and PSQ at C terminus of the g.Block.

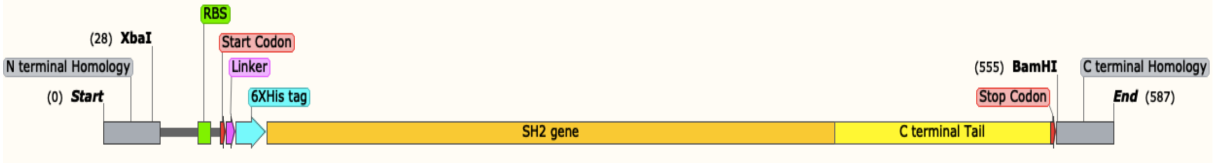


Figure 2.8. Schematic drawing of the SH2 g.Block construct with C terminal tail of SH2B1 β isoform, 586 bp.

2.4.6 Ligation by Gibson assembly

The plasmid was digested with a pair of desired restriction enzymes and then the g.Block gene was designed with overlapping sequences. After that the linear plasmid with the designed gene were ligated using the Gibson assembly method following the lab manual (Gibson et al., 2009).

The conditions of the ligation of g.Block genes with the linearised plasmid were specified by the supplier. The number of pmols of insert and vector for optimal assembly were calculated based on the length and weight of the fragments by using a formula in https://www.neb.com/protocols/2012/12/11/gibson-assembly-protocol-e5510.

Following the manufacturer's protocol, the pellet of g.Block fragments was re-suspended with 50 µl distilled water to a final concentration of 10 ng/µl. After that, the linearised plasmid was ligated with the designed g.Block by Gibson assembly in molar ratio 1:3 of vector to insert. As an example of the assembly calculation, the SH2 g.Block fragment, 379 bp in 10 ng/µl concentration, is 0.0409 pmols. The linear pET-MBP-His vector (50 ng), at a concentration of 7.9 ng/µl for 6869 bp, is 0.011 pmol. A 3× excess of insert is 0.033 pmol which is 0.8 µl ~1 µl. The ligation reaction was carried out in 12 µl final volume: 6 µl vector (7.9 ng/µl), 1 µl insert (10 ng/µl) and 5 µl NEBuilder HiFi DNA assembly master mix. The assembly reaction was incubated at 50°C for one hour. During the incubation the overlapping g.Block of the SH2 gene was ligated with the linearised plasmid via the action of three enzymes: a 5′ exonuclease, a DNA polymerase and a DNA ligase as provided in the kit, Figure 2.9. After the incubation time, 5 µl from the mixture was transformed into 50 µl of NEB 5-alpha competent cells.

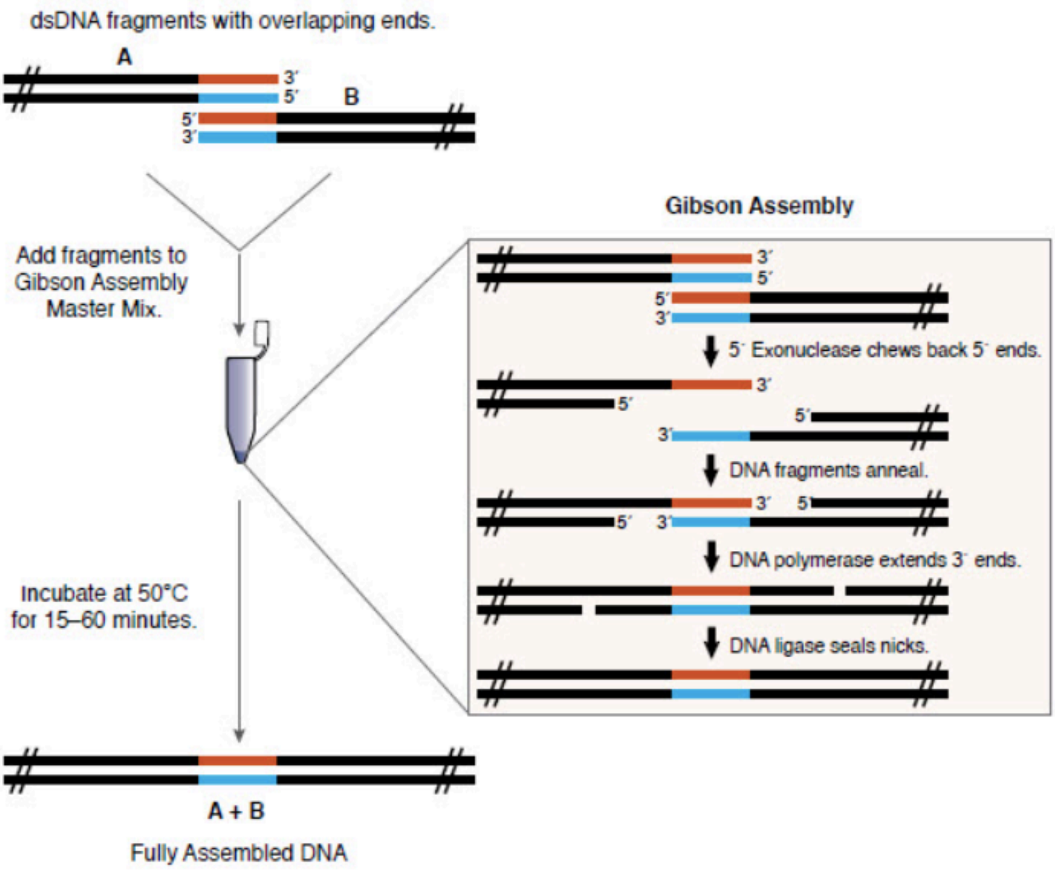

dsDNA fragments with overlapping ends.
A
3′
5′
B
5′
3′
Add fragments to Gibson Assembly Master Mix.
Incubate at 50°C for 15–60 minutes.
Gibson Assembly
3′
5′
5′
3′
5′ Exonuclease chews back 5′ ends.
3′
5′
5′
3′
DNA fragments anneal.
3′
5
5′
3′
DNA polymerase extends 3′ ends.
DNA ligase seals nicks.
A + B
Fully Assembled DNA

Figure 2.9. An overview diagram for Gibson assembly method shows 1. Target DNA gene designed with overlapped sequences at both ends (A) which match with sequences of plasmid at both ends (B), the plasmid digest with desired restriction enzyme before starting the assembly reaction. 2. Gibson assembly reaction into the provided kit which include linear plasmid, target DNA gene, and enzymes for Gibson assembly reaction; exonuclease T5 chews from 5’ to 3’ ends the DNA ends exposing the homologous sequences in blue and red (digestion), Phusion DNA polymerase extends the DNA from the 3 ends (annealing), and Taq ligase start to ligate the nicks (ligation). 3. Final result from Gibson assembly a full assembled DNA. The diagram is taken from www.international.neb.com.

2.5 Protein methods

2.5.1 Protein expression and extraction

BL21 (DE3) *E. coli* cells were transformed with plasmids containing the target gene and were plated onto agar media supplemented with ampicillin, Agar-Amp table 2.1. Following overnight incubation at 37°C, a single colony was picked off and inoculated in 10 ml Amp-LB.

The 10 ml starter culture was used to inoculate 1 L autoclaved medium (LB or M9 media, recipe given in tables 2.2 and 2.1) supplemented with 1 ml ampicillin (100 mg/1ml) followed by incubation at 37°C, 250 rpm until the optical density reached ~0.6 au at 600 nm which was checked by spectrophotometer. Then, protein expression was induced by adding 0.5 mM of Isopropyl β-D-1-thiogalactopyranoside (IPTG) from a 1 M stock solution. After that, the culture was incubated overnight at 25°C and the cells were harvested by 10 mins centrifugation at 18,000 g, at 4°C. Cell paste from overnight culture was suspended in 20 ml lysis buffer (50 mM Tris-HCl, pH 8, 50 mM NaCl with one Mini EDTA-free Protease Inhibitor tablet and DNase 1). The cells were disrupted by sonication via ultrasound on ice for 6× 30 seconds at 16 microns amplitude with 60 sec breaks in between, then the extract was centrifuged at 45,000 g, for 30 mins at 4°C.

2.5.2 Protein purification

2.5.2.1 Immobilized metal affinity chromatography

A nickel Nitrilotriacetic Acid (Ni-NTA) Agarose column was used to purify the overexpressed fusion proteins with His tag. The Ni-NTA column (15 ml resin in 5.0×20 cm column) was equilibrated with washing buffer (40 mM imidazole, 50 mM Tris-HCl, pH 7.5-8, containing 150 mM NaCl) by washing with three times the column volume. Then the supernatant (~20 ml) was applied to the column and washed with the same washing buffer (~100 ml). Wash fractions (3 ml) were collected at a flow rate of 2 ml/min by an ÄKTAprime Plus system. The elution was collected after applying the elution buffer to the column (300 mM imidazole, 50 mM Tris-HCl, pH 7.5-8, containing 150 mM NaCl), then the elution fractions were collected in 3 ml fractions with a flow rate 3 ml/min. The optical density was measured at 280 nm to identify the protein-containing fractions. The washing and elution fractions which showed higher optical density were run on 16% SDS-PAGE.

2.5.2.2 Protease cleavage

Cleavage of overexpressed fusion proteins (His$_6$-MBP-TEV-SH2, the first SH2 construct) by TEV

(Tobacco Etch Virus which was expressed and purified in Dr Mesnage's lab) protease (1.2 mg/ml) was done with different ratios of protease to protein (1:20, 1:50 and 1:100) in buffer (10 mM NaCl, 50 mM Tris pH 8, 5 mM DTT and 5% glycerol). The protease reaction was incubated at room temperature overnight or an hour at 30°C. A 16% SDS gel was used to analyse the cleavage of tagged fusion proteins (His_6-MBP-TEV-SH2) by the protease.

2.5.2.3 Amylose affinity chromatography

An 8 ml Amylose resin packed in a PD-10 column was used to isolate MBP fusions after cleavage of His_6-MBP-TEV-SH2 proteins with TEV protease. The column was equilibrated in washing buffer (0.2 M NaCl, 50 mM Tris pH 8), after applying the sample. After that the column was washed with 15 ml washing buffer, then the elution was performed with elution buffer (0.2 M NaCl, 50 mM Tris pH 8, and 10 mM maltose). The elution fractions were collected in 2.5 ml fractions.

2.5.2.4 Dialysis method

6 M guanidine hydrochloride was used to dissolve the cell pellet from 500 ml growth. Dialysis tubing (Vivaspin^{TM} column 10 kDa MWCO) was cut to the appropriate length (~14 cm). After that, in order to remove the glycerol, the tubing was placed in water for half an hour. The tubing was clipped off at the bottom and the solution was then transferred into it. Then the tubing was clipped off from the top and placed in a beaker containing 1 litre of 50 mM KH_2PO_4, pH 7.4. The beaker was placed in the fridge with a magnetic stirrer. After two hours, the buffer (50 mM KH_2PO_4, pH 7.4) was changed and left stirring for three days with change of buffer after every 24 hr. After three days, the mixture into the tube was poured off into a centrifuge tube and was spun at 18,000 g at 4°C for 20 minutes in a Beckman Avanti J-251 Ultra centrifuge. Then the supernatant and pellet were run on an SDS gel**.**

2.5.2.5 Size exclusion chromatography

Gel filtration was used to separate proteins based on their size (analysis and quantitation of monomer, dimers, trimers, and higher order aggregates for proteins). The gel filtration columns used were: Superdex 200 10/300 or Superdex^{TM} 75 10/300 (GE Healthcare). Before injecting the sample, the columns were washed with the same applied buffer (table 2.5). The gel filtration experiments were performed in different buffer as explained in chapter 3 at flow rate 0.5-1 ml/min using ÄKTA purifier system and ÄKTAprime Plus system. The gel filtration fractions based on the peaks were run in a 16% SDS gel**.**

Table 2.5. Gel Filtration Applied Buffer.

G.F Buffer	Reagent
Buffer 1	50 mM Tris-HCl, pH 8, 150 to 500 mM NaCl
Buffer 2	50 mM Tris-HCL, pH 5.6, 0.5 M NaCl
Buffer 3	50 mM Tris-HCL, pH 7.4, 300 mM NaCl
Buffer 4	50 mM Tris-HCl, pH 8, 300 mM NaCl, 2 mM Dithiothreitol (DTT)

2.6 Protein biochemistry

2.6.1 Protein concentrating and buffer exchange

A vivaspin^{TM} ultrafiltration column (10 kDa MWCO) was used to exchange the buffer and concentrate dilute protein samples. Firstly, the spin column was washed with Milli-Q water by centrifugation. The large volume samples were reduced to the desired concentration or to approximately 1 ml or less in order to exchange the buffer, then the wanted buffer was added and the sample concentrated again to remove unwanted buffer.

2.6.2 Protein quantification and molecular weight

All theoretical extinction coefficients and molecular weights shown in table 2.6 were calculated from the amino acid sequences by ExPASy web. These values were used for the quantification of proteins at A_{280nm} by NanoDrop UV-Vis spectrophotometers according to the Beer-Lambert Law. Also these values were used to optimise the buffer's pH during the purification procedures.

Table 2.6. Physical and Chemical Parameters of Proteins.

Protein	Number of Residues	Molecular Weight Da	Theoretical pI	Extinction Coefficient (M^{-1} cm^{-1})
His_6-MBB-TEV-SH2	498	55272.46	5.96	83310
His_6-SH2	111	12725.36	6.94	13980
His_6-SH2 (with extra N and C terminal residues)	118	13663.55	8.06	15470
SH2 with the C terminal tail of β (SH2c)	160	18092.36	6.69	16960
Mutant SH2c Y115F	160	17935.05	6.54	15470

2.7 Protein analytical methods

2.7.1 SDS PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis)

SDS-PAGE was used to confirm the molecular size of the expressed protein as shown in table 2.6, using 16% resolving and 4% stacking gel acrylamide gels. The gel was prepared with ~4 cm resolving gel and ~2 cm stacking gel. Plates were sealed with tape to prevent the gel from leaking and the assembled plates were fixed in the gel apparatus. 10 ml of 16% resolving gel was prepared (recipe in table 2.7). The poured gel was left for the polymerization. Isopropanol was poured to level the resolving gel. After polymerization of resolving gel, 4 ml of 4% stacking gel was prepared (recipe in table 2.7). The gel was poured into the plates, and a suitably sized comb was inserted. 15 µl of samples were taken and mixed with 5 µl of 4×SDS loading buffer (table 2.8); then the samples were boiled at 95℃ for a minute. 4 µl of pre-stained SDS PAGE marker (250 to 10 kDa from Thermo Fisher) was used as size indicators. 20 µl of samples were loaded into the gel with the help of a micropipette. The gel was electrophoresed for approximately an hour at constant voltage of 180 in 1× running buffer, after which the gel was removed and placed in instant blue stain with shaking for 15 minutes.

Table 2.7. SDS PAGE Layers recipe.

Gel layer	Reagent
16% Polyacrylamide Resolving Gel	2.5 ml of 4× lower buffer (table 8), 4 ml of 40% (w/v) BisAcrylamide, 3.5 ml Milli-Q water, 100 µl of 10% Ammonium Persulphate (APS) (table 2.8), 10 µl Tetramethylethyldiamine (TEMED)
4% Polyacrylamide Stacking Gel	2.5 ml 4× upper buffer (table 8), 1.125 ml of 40% (w/v) BisAcrylamide, 6.3 ml of Milli-Q water, 110 µl of 10% Ammonium Persulphate (APS) (table 2.8), 11 µl Tetramethylethyldiamine (TEMED)

Table 2.8. SDS PAGE Buffers recipes.

SDS PAGE Buffers	Reagent
4× Lower Buffer	1.5 M Tris/HCl, 0.4% (w/v) SDS, pH 8.8
4× Upper Buffer	0.5 M Tris/HCl, 0.4% (w/v) SDS, pH 6.8
4× Loading Buffer	200 mM Tris/HCl, 400 mM DTT , 8% (w/v) SDS, 0.4% (w/v) Bromophenol blue, 40% (v/v) Glycerol, pH 6.8
1× Running Buffer	25 mM Tris/HCl, 0.19 M Glycine, 0.1% (w/v) SDS, pH 8.3
10% (w/v) APS	1 g dissolved in 10 ml distilled water

2.7.2 Mass spectrometry

≥ 20 µl of protein samples were sent to MS facility for analysis by ESI MS (Electro-Spray Ionization Mass Spectrometry) to determine the protein molecular weight.

ETD MS (Electron Transfer Dissociation Mass Spectrometry) was used to analyse the phosphorylation sites on proteins (SH2c and SH2c Y114F) by the trypsin digestion analytical method. ~20 ul of protein samples were run in 16% SDS PAGE gel, followed by staining in InstantBlue dye for 1 h. The relevant bands were cut to small pieces and proteins digested into peptides prior to identification by ETD MS. The process involved destaining the gel pieces, reduction and alkylation followed by trypsin digestion of the proteins and extraction of the peptides from the gel pieces following the manufacture's protocol. The extracted peptides were evaporated overnight in a speedvac to be ready next day to dissolve in suitable buffer, then the analysis by ETD MS was carried out by Adelina Martin in the MS facility.

2.7.3 Bioinformatic analysis

MUSCLE alignment tool (Multiple Sequence Comparison by log Expectation) in https://www.ebi.ac.uk/Tools/msa/muscle/ web was used for multiple sequence alignment of proteins in order to detect the similarity and difference between applied sequences. The protein sequence alignment tool was applied to add extra residues at the N and C termini of SH2 protein by aligning its sequence with other SH2-containing proteins as shown in chapter 3.

2.8 Protein phosphorylation

Group-based Prediction System 3.0 web service was used to predict the phosphorylation site with its cognate protein kinase (Xue et al., 2008). Fer Kinase (from ProQinase) was predicted to be a good enzyme to phosphorylate Tyrosine 139 in the C terminus of SH2 SH2B1 β protein. The phosphorylation of Y139 in the C terminus of SH2c and mutant SH2c Y114F were done through incubation of the kinase with protein in ratio 1:5 for 8 hrs at 30°C with shaking in reaction buffer: 70 mM HEPES-NaOH pH 7.5, 3 mM $MgCl_2$, 3 mM $MnCl_2$, 3 µM Na-orthovanadate, 1.2 mM DTT, 50 µg/ml $PEG_{20000,}$ 1% DMSO, 5 mM ATP. During the incubation time of SH2c Y114F protein, 1 mM DTT was added to the sample every hour. After that the phosphorylated sample was subjected to buffer exchange to be in a suitable buffer for the subsequent NMR experiment. The phosphorylation of Y139 was identified and confirmed by ^{15}N HSQC spectra, ESI and ETD MS.

2.9 NMR (nuclear magnetic resonance) methods

2.9.1 NMR sample

The NMR samples of SH2 protein were prepared in 90% H_2O/10% D_2O, or 99.9% D_2O in 50 mM potassium phosphate pH 6, containing 1 mM TSP (trimethylsilyl propanoic acid), or in 90% H_2O/10% D_2O in 50 mM Tris and 50 mM NaCl pH 6, 1mM TSP. Also the NMR samples of the non-phosphorylated and phosphorylated SH2c and SH2c Y114F proteins were set in two buffers; 90% H_2O/10% D_2O in 100 mM potassium phosphate pH 6-7, with 1 mM TSP or 90% H_2O/10% D_2O in 50 mM Tris and 50 mM NaCl pH 6, 1mM TSP. The protein samples used were uniformly ^{15}N-labelled; ^{13}C, ^{15}N-labelled; and unlabelled.

NMR spectra were obtained for NMR structure determination from ~1 mM double labelled protein in a 5 mm Shigemi tube (300 µl sample). In the case of protein-ligand complexes and other NMR studies, NMR spectra were obtained from a double labelled protein (~1 mM to ~0.06 mM) in a normal 5 mm NMR tube (500 µl sample).

2.9.2 NMR measurement and processing

NMR measurements were recorded on a Bruker spectrometer operating at a proton frequency of 800 MHz at 298K. Spectrometers were controlled by a UNIX workstation running Topspin software for data processing. The data was transferred to LINUX and processed with Felix2007 for spectral analysis.

The chemical shift of ^{1}H was referenced to an internal TSP signal, and ^{15}N and ^{13}C chemical shifts were referenced indirectly to ^{1}H by calculation from their gyromagnetic ratios. Standard 2D and 3D pulse sequences were provided in pulse sequence libraries from the NMR spectrometer manufacturer. 1D ^{1}H and 2D ^{15}N HSQC spectra were collected before every 3D experiment to assess folding, aggregation, stability, and ligand binding or detect unstructured regions of sample.

2.9.3 Backbone assignment (semi-automated)

Backbone assignments of the SH2 protein, and the non-phosphorylated and the phosphorylated SH2c and SH2C Y114F proteins in different buffer were obtained following the same procedures. The backbone resonances of ^{1}HN, ^{15}NH, ^{13}CO, ^{13}Cα and ^{13}Cβ of isotopically labelled protein were assigned using the standard triple resonance spectra listed in table 2.9. All of these spectra collected based on non-uniform sampling to speed up data acquisition. The name of the backbone experiment which listed in table 2.9 shows from where the polarization was transferred, there is no frequency information was obtained for the atoms in brackets. The subscripts show the residue number correlated. For example in HNCA spectrum, HN_i to $C\alpha_{i-1}$ means the HN (i) of a residue correlates to the Cα of the previous residue (i-1), and in square brackets [$C\alpha_{i-1}$] weaker correlations are obtained.

Table 2.9. Standard Triple Resonances NMR spectra.

Spectra	Correlates
^{15}N HSQC	HN to NH
HNCO	HN_i to CO_{i-1}
HN(CA)CO	HN_i to CO_i and [CO_{i-1}]
HNCA	HN_i to $C\alpha_i$ and [$C\alpha_{i-1}$]
HN(CO)CA	HN_i to $C\alpha_{i-1}$
CBCA(CO)NH	HN_i to $C\alpha_{i-1}$ and $C\beta_{i-1}$
HNCACB	HN_i to $C\alpha_i$, $C\beta_i$, [$C\alpha_{i-1}$] and [$C\beta_{i-1}$]

Acquisition parameters are the same for the backbone spectra of; SH2, non-phosphorylated SH2c, and non-phosphorylated and phosphorylated SH2c Y114F double labelled proteins in different buffers which are listed in table 2.10.

The NMR resonance signals observed in the ^{15}N HSQC spectra were picked and numbered systematically using the automatic numbering in Felix2007, to use as a starting point for manual peak picking. In each HN strip, the chemical shift of each HN peak was aligned with the backbone triple resonance spectra to assign the chemical shift of their associated CO, Cα, and Cβ nuclei, to identify their corresponding residues. Peak picking and matching was done manually using Felix2007 after adjusting the referencing using a macro, as explained in Chapter 4.

Table 2.10. Experimental parameters for multidimensional NMR spectra used to assign backbone of SH2, non-phosphorylated and phosphorylated SH2c and SH2c Y114F; and to assign sidechain and structure of SH2 protein. The spectrometer frequency was 800 MHz ^{1}H. Nuc: nuclei recorded, Sw: spectra width, TD: size of fid, Acq: acquisition time.

Experiment	Name spectra on Mag6	D1				D2				D3			
		Nuc	Sw Hz	TD	Acq s	Nuc	Sw Hz	TD	Acq s	Nuc	Sw Hz	TD	Acq s
Backbone													
N^{15} HSQC	hsqcetfpf3gp	$^{1}H_{N}$	9615.3	2048	0.1065	^{15}N	2189.4	256	0.0584				
HNCO	hncogp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	1811.1	100	0.0276
HN(CA)CO	hncacogp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	1811.1	100	0.0276
HNCA	hncagp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	4527.3	100	0.0110
HN(CO)CA	hncocagp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	4527.3	100	0.0110
HNCACB	hncacbgp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	10563.6	160	0.0075
CBCA(CO)NH	hncocacbgp3d.2	$^{1}H_{N}$	9615.3	1024	0.0532	^{15}N	2189.4	40	0.0091	^{13}C	10563.6	160	0.0075
Sidechain													
^{13}CHSQC	hsqcctetgpsisp	$^{1}H_{N}$	9615.3	1024	0.0532	^{13}C	12072.7	256	0.0106				
HBHAcoNH	hbhaconhgpwg3d	$^{1}H_{N}$	9615.3	2048	0.1065	^{15}N	2189.4	40	0.0091	^{1}H	9615.3	128	0.0066
CcoNH	ccconhgp3d.2	$^{1}H_{N}$	8196.7	2048	0.1249	^{15}N	2189.4	40	0.0091	^{13}C	12072.7	128	0.0053
HCcoNH	hccconhgpwg3d2	$^{1}H_{N}$	9615.3	2048	0.1065	^{15}N	2189.4	40	0.0091	^{1}H	9615.3	128	0.0066
HCCHTOCSY	hcchdigp3d2	$^{1}H_{C}$	8196.7	2048	0.1249	^{13}C	12072.7	64	0.0026	^{1}H	8196.7	128	0.0078
HCCHCOSY	hcchcogp3d	$^{1}H_{C}$	8196.7	2048	0.1249	^{13}C	12072.7	64	0.0026	^{1}H	8196.7	128	0.0078
Aromatic													
^{13}C HSQC/TROSY	trosyargpphwg	$^{1}H_{C}$	8196.7	2048	0.1249	^{13}C	5131.3	160	0.0155				
HBCBCGCDHD	hbcbcgcdhdgp	$^{1}H_{\delta}$	9615.3	2048	0.1065	$^{13}C_{\beta}$	3621.7	80	0.0110				
HBCBCGCDCE	hbcbcgcdcehegp	^{1}H	9615.3	2048	0.1065	$^{13}C_{\beta}$	3621.7	80	0.0110				
^{1}H ^{1}H NOESY	noesyphpr	$^{1}H_{N}$	7462.6	2048	0.1372	^{1}H	7501.6	512	0.0341				
Structural (restraint)													
^{15}N NOESY	noesyhsqcfpf3gpsi3d	$^{1}H_{N}$	9615.3	2048	0.1065	^{15}N	2189.4	40	0.0091	^{1}H	9602.1	128	0.0066
^{13}C NOESY	noesyhsqcetgpsi3d	$^{1}H_{C}$	9615.3	2048	0.1065	^{13}C	12072.7	64	0.0026	^{1}H	9602.1	128	0.0066
^{13}C NOESYaro	noesyhsqcetgpsi3d	$^{1}H_{C}$	9615.3	2048	0.1065	^{13}C	5131.3	40	0.0038	^{1}H	9602.1	128	0.0066

2.9.4 Asstools: backbone assignment

The assignment of the spin systems to residues within the protein sequence was completed via the program Asstools (Reed et al., 2003). The program builds up a sequence of matching spin systems using a Monte Carlo simulated annealing method. It works by comparing chemical shifts from a spin system to preceding chemical shifts from all other spin systems, as explained in Chapter 4. Also it matches the carbon chemical shifts (Cα, Cβ, and Co) of spin system to amino acid types. Asstools performs 30 iterative separate runs. In each one, initially spin systems are randomly assigned to a residue in the protein sequence and scores are calculated for the chemical shift matches of self and preceding residue. Also scores are computed for chemical shift similarity of assigned residue type and characteristic random coil amino acid shifts. Each run is done iteratively as assignments are randomly swapped and then scores calculated until a stable score is obtained for three following iterations. Once all runs completed, the output is collated to produce a spin systems list which corresponding to each residue in the protein sequence. Any assignment produced from at least 27 out of the 30 runs is taken as correct. Remaining assignments were then completed by manual inspection of the spectra.

2.9.5 Manual sidechain assignment for SH2 protein

The backbone resonances for ^{1}HN, ^{15}NH, ^{13}CO, ^{13}Cα and ^{13}Cβ of SH2 protein (118 a.a) were assigned using standard triple resonance 3D NMR spectra. The peak lists of these spectra had been already picked manually in the previous manual backbone assignment.

The majority of the aliphatic sidechain peaks were picked and lists prepared manually. HBHAcoNH, HCcoNH and HCcoNH 3D NMR spectra were used to obtain the resonances of the aliphatic carbons and their associated protons, α and β, α, β, and γ protons, and Cα, Cβ, and Cγ, respectively. These spectra were picked manually in Felix2007 after adjusting the referencing of ^{15}N, ^{13}C, ^{1}H with ^{15}N HSQC, and an assigned manual peak list was prepared. HCCH COSY and HCCH TOCSY 3D NMR spectra were used for more information about sidechain protons which are linked to aliphatic carbons.

The remaining resonances, mainly the aromatic sidechains of Trp, Tyr, and Phe, were assigned from ^{1}H-^{1}H through-space connectivity information in the 2D ^{1}H ^{1}H NOESY spectra, supported by 2D ^{1}H ^{1}H TOCSY. In the case of Trp, to assign ε, ζ, η protons which have chemical shifts close together, ^{1}H-^{1}H TOCSY and ^{1}H ^{1}H NOESY spectra were used, and the peaks were picked automatically as before. Because of the similar chemical shifts of aromatic ring protons (δ, ε) of Phe residues which prevented the selective observation of each, 2D HBCBCGCDHD and HBCBCGCDCE spectra were used to detect the aromatic ring protons, to connect through-bond the aliphatic resonances (C_β), which were identified in backbone assignment, to sidechain protons, then these protons were used to assign carbons of aromatic rings in ^{13}C HSQC, peaks picked manually and peak lists prepared. The name of NMR spectra with their acquisition parameters which are used to assign aliphatic and aromatic sidechain resonances are listed in table 2.10.

2.10 CYANA (combined assignment and dynamics algorithm for NMR applications): assignment and structure determination

CYANA automated computational approach (version 3.98.5) with independent programs was used to calculate the 3D NMR structure of SH2 SH2B1 protein (Güntert, 2004). The resonance assignment and structure calculations were based on a set of multidimensional NMR spectra and the amino acid sequence as experimental initial input data.

CYANA calculation was performed on a Linux cluster system. The structure calculation was done with the additional independent programs Felix2007, TALOS-N, and CNS as plug in. All CYANA assignment and calculation was done for residues 1-118 of SH2 as specified in the init.cya macro file. A complete CYANA calculation comprises the following steps.

2.10.1 Peak picking and formatting

After processing 2D and 3D NMR spectra which are listed in table 2.10, referencing and frequency matching all spectra were adjusted manually in Felix2007, then an automated peak picking algorithm of Felix2007 was applied to pick all peaks, thus the peak positions and intensities are identified. The resulting raw peak lists are reformatted to be in XEASY format (with consistent naming of the peak lists according to the experiment type).

2.10.2 Automated complete resonance assignments for SH2 protein

The FLYA routines within CYANA macro, CALC.cya (under the subdirectory demo/flya), were used to obtain the complete resonance assignment, which required as input data a number of multidimensional NMR spectra including the peak lists for backbone, sidechain and NOE experiments as listed in table 2.10 with acquisition parameters, along with the amino acid sequence of the SH2 protein.

The initial chemical shift assignments generate an ensemble of the chemical shift values for each ^{1}H, ^{13}C, and ^{15}N nucleus. This ensemble is obtained by 20 independent runs of the GARANT algorithm which uses different seed values for the random number generator. The GARANT algorithm works based on the knowledge of the amino acid sequence, the magnetization transfer pathways and the experimental given peaks to match between the peaks expected (Bartels et al., 1997). The result of the completed runs is an ensemble of 20 raw chemical shift assignments for the complete protein. In all calculations the defined tolerance values of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N chemical shifts, are used for the matching of peaks with identical assignments. In the ensemble for ^{1}H, ^{13}C, and ^{15}N spin the highly populated chemical shift values are calculated and then selected as the consensus chemical shift value to represent the automated resonances assignment. A number of output files is produced after the assignment calculation as listed in table 2.11.

The resulting resonance assignments require greater manual intervention for errors before proceeding with structure determination due to incomplete data and degeneracy of aliphatic chemical shifts. The manual checking had been done according to previous semi-manual backbone assignments and the assigned manual sidechain peak lists.

Table 2.11. FLYA calculation output files for the SH2 resonance assignments with their description.

Output files	Description
Flya.prot	This file contains a chemical shift for every atom that has been assigned to at least one peak.
Flya.tab	Table with details about the chemical shift assignment of each atom. This file shows the assignment of each atom and whether is strong or weak
Flya.txt	Assignment statistics for each group of atoms
Flya.pdf	Graphical representation of the assignment results in flya.tab represented by a coloured rectangle.
Xxx_exp.peaks	List of expected peaks corresponding to input peak list XXX.peaks
Xxx_as.peaks	Assigned peak list corresponding to input peak list XXX.peaks

2.10.3 Automated NOE assignment and structure calculation

The CYANA structure calculation goes through sequential cycles of automatic NOE assignment, restraint generation, structure calculation, and validation (Güntert, 2004; Schmidt & Güntert, 2013; Herrmann et al., 2002).

The SH2 structure calculation was performed using CYANA macro CALC.cya (under subdirectory demo/auto), based on the amino acid sequence, the chemical shift assignment, unassigned 3D NOESY peak lists (shown in table 2.10), and the dihedral angle restraints obtained from chemical shifts. The CYANA calculation parameters including the number of initial and final structures, number of steps, shift tolerance values of each nuclei in ppm are specify in a macro script, Figure 5.2.

In CYANA the automated NOE assignment and structure calculation are combined in seven sequential cycles of automated NOE assignment, and followed by a final structure calculation. Automatically NOESY cross-peaks are assigned based on the complete chemical shift assignments and unassigned experimental NOE peak lists within the defined tolerance values. ^{1}H-^{1}H distance restraints were derived from ^{13}C and ^{15}N edited HSQC NOESY spectra which is the main source of structural information to determine the NMR structure. In these experiments the ^{13}C and ^{15}N HSQC spectrum is extended into a third dimension which is NOE. Each cross peak arises from an NOE between two protons that possess dipolar coupling. These NOEs are created when protons exchange magnetization in a distance dependant fashion during the mixing time. A signal is only observed if the distance between the protons is less than about 5Å . The ^{15}N NOESY detects NOEs to NH from any other proton in the molecule, whereas the ^{13}C NOESY detects NOEs to CH groups.

The correctness of expected possible NOE assignments is calculated as three probabilities based on the agreement between the given chemical shift values and the experimental peak position, the consistency with an initial 3D structure of the preceding cycle, and network-anchoring, as descried in Chapter 4.

The resulted structure from a given cycle is used in the following cycle as guide the NOE assignments, except for cycle 1, the precision of the structure generally improve in each subsequent cycle. In the first two cycles of the CYANA calculation, constraint combination criterion is used to medium- with long-range distance restraints to decrease the effect of erroneous distance restraints on the calculated structure. The CYANA structure calculation is executed using a standard protocol based on simulated annealing driven by molecular dynamics simulation in torsion angle space. The automated NOESY assignment and structure calculation are followed by a final structure calculation which is generated from an ensemble. A number of output files is produced after each cycle as listed in Table 2.12.

Automated NOESY assignment followed by structure calculation with CYANA does not require manual intervention during cycles, and is based on previously determined chemical shift assignments which permit starting the NOE assignment and structure calculation process directly from the NOESY spectra.

Table 2.11. CYANA structure calculation output files.

Output files	Description
final.noe	NOE assignment details for each peak
final.upl	Final NOE upper distance limits
final.ovw	Target function/violation overview
final.pdf	Final resulting structure
rama.ps	Ramachandran plots

To generate better defined and more rigid structures, explained in chapter 5, further CYANA structure calculations were performed using a list of hydrogen bonds predicted from the previous CYANA calculation as additional input constraints. The obtained hydrogen bonds included in the list were required to be consistent with small values of amide proton temperature coefficient (Baxter & Williamson, 1997).

2.11 Structure constraints

2.11.1 Backbone dihedral angles

After the chemical shift assignments have been completed, dihedral angle restraints for the backbone ϕ and ψ angles are generated using the program TALOS-N (Torsion Angle Likelihood Obtained from Shift and Sequence Similarity) (Shen & Bax, 2013). The idea of this program is to use the given backbone chemical shifts and sequence information to make quantitative predictions for the protein backbone angles ϕ, ψ, and sidechain χ^1. The program searches a database constructed using high resolution crystal structures, then compares their secondary shifts with HN, NH, CO, Hα, Cα, and Cβ chemical shifts of a given protein. Data from each residue (type and chemical shifts) and its two neighbours are compared to those in the database to select the 10 best matches for ϕ and ψ. Matches that indicate consistent values for ϕ and ψ and χ^1 are classified as good. Their average and standard deviation are used as predictions. When matches are inconsistent no predictions are made for the central residues and these residues are subsequently declared ambiguous and were not used. TALOS-N also classifies some residues as mobile based on the small values of their secondary chemical shifts. These were also not used.

2.11.2 Amide proton temperature coefficients (hydrogen bonds)

The correlation between amide proton temperature coefficients and hydrogen bonds was used to predict the donor hydrogen bonds in the loops (Baxter & Williamson, 1997), and also to confirm the existence of hydrogen bonds that were taken from CYANA structure.

A 1 mM SH2 protein sample was prepared in 90% H_2O/10% D_2O in 100 mM potassium phosphate pH 6 in a Shigemi NMR tube. 1 mM TSP in the sample served as the reference for ^1H chemical shifts. ^{15}N HSQC spectra were recorded at 5° intervals from 283 K to 303 K on a Bruker 800 MHz spectrometer. ^{15}N HSQC spectra of the temperature series were processed, then picked within Felix2007 and assigned. Acquisition parameters of HSQC spectra were copied from the HSQC spectrum of SH2 that was used for the backbone assignment. The fitting program performed a least-squares minimization of a linear equation to the chemical shift versus temperature data, and the NH temperature coefficients were obtained from the gradient of the best-fit line, shown in Chapter 4.

2.12 CNS (crystallography and NMR system) recalculation and refinement

Structure re-calculations (100 structures) and refinement in explicit solvent were carried out using CNS. CNS uses a simulated annealing protocol involving a number of stages to produce a set of structures (Brünger et al., 1998).

The standard annealing protocols was used with protein-allhdg parameters. Firstly, a linear string of amino acids of the SH2 protein sequence was taken, along with constraints files using CYANA NOEs assignment, dihedral angles and hydrogen bonds. The atoms movement was simulated by increasing the temperature to around ~5000 K with reducing Van der Waals energy to permit free movement through conformational space, thus not being restrained by local energy. After that gradually the temperature was decreased to ~25 K to 0 K, and then Van der Waals forces were applied slowly, thus movement simulated by the dynamics of torsion angles. That caused reducing the atoms velocity, and hence gradually restrained them into low energy conformations. The total energy of each structure calculated by its violations of an average sum of the experimentally restraints and expected molecular properties, e.g bond lengths and angles, and Van der Waals interactions. A minimised average structure was generated by refinement in explicit solvent. Therefor all 100 calculated structures were refined through an energy minimisation procedure in explicit solvent via the AMBER force field using CNS. During the whole procedure same experimentally restraints were used. Of the ~100 refined structures, a group of 20 lowest energy and which also show a good ANSURR score were taken to represent the protein ensemble.

2.13 ANSURR (accuracy of NMR structures using random coil index and rigidity) validation method

The accuracy of the calculated NMR structures of SH2 were measured by the ANSURR method (Fowler et al., 2020). The ANSURR method evaluates the accuracy of NMR protein structures based on two measurements: the correlation score and the RMSD score. This measuring method combines the matching between the local rigidity from backbone chemical shifts (predicted by Random Coil Index) and the local rigidity from the computed NMR structure (by Floppy Inclusions and Rigid Substructure Topography (FIRST)) into a correlation score which measures the accuracy of the secondary structure, whereas a RMSD score between FIRST and RCI assesses the accuracy of the overall rigidity.

The ANSURR method was applied to follow the accuracy improvement in the final ensemble structure derived from CYANA structure calculations, then the re-calculated and refined structures calculated using the CNS program, and combining with the structure total energy to select the final ensemble.

2.14 Protein binding and dynamics

2.14.1 Protein titration

In order to investigate the binding between SH2 protein and the C terminal tail of β isoform, and obtain the affinity, a series of titrations were carried out using a uniformly ^{15}N labelled SH2 protein and unlabelled (GDRC(pY)PDASST) ligand. The unlabelled phospho-tyrosine peptide was synthesized using the sequence in the region of Tyr139 at the C-terminal end of SH2B1 β isoform (mouse). The concentration of the protein and unlabelled ligand were determined using 1D NMR spectra by comparison to the TSP concentration in the NMR sample which was 0.2 mM, before starting the titration.

The C terminal peptide (from GenScript) was 11 residues long including a phosphate group attached to Tyrosine, and the ligand was dissolved in the same buffer as the protein (50 mM Tris pH 6, 50 mM NaCl, 2 mM DTT) to make a 3 mM stock solution. Ligand titrations was carried out on 0.08 mM protein in 50 mM Tris pH 6, 50 mM NaCl buffer containing 1 mM TSP. Ligand addition was carried out by mixing the free protein sample with a series of ligand concentrations in ratios as listed in Table 2.13.

The series of ^{15}N HSQC spectra were recorded on a Bruker spectrometer operating at a proton frequency of 800 MHz at 298 K, and processed in Felix2007. Acquisition parameters of HSQC spectra were copied from the HSQC spectrum of SH2 protein that was used for the backbone assignment. The extent of binding of ligands to SH2 protein was determined by monitoring the change in chemical shifts of ^{15}N HSQC spectra. The chemical shift changes (CSP) were followed over the HSQC series and then plotted separately against ligand concentration for every assigned residue. The plots of chemical shift changes were further analysed to measure the binding affinity as explained in chapter 6. The weighted average of amide chemical shift differences for each residue was calculated using equation 6.1.

Table 2.12. The titration calculation of SH2 with C terminal ligand, based on y =[v3(c1-c3)]/(c2-c1)], c1 is the required ligand concentration to be added to get desired ligand in mM, c2 is the concentration of ligand stock in mM, v3 is volume of NMR sample in µl, c3 is the previous concentration of ligand present in the NMR sample, y is the final ligand addition to sample.

c1	c2	v3	c3	[v3(c1-c3)]	(c2-c1)	y
0.01	3	500	0	5	2.99	1.7
0.015	3	501.6	0.01	2.508333333	2.985	0.840
0.03	3	502.5	0.015	7.53760469	2.97	2.538
0.04	3	505.04	0.03	5.050448934	2.96	1.706
0.06	3	506.7	0.04	10.13502252	2.94	3.447
0.08	3	510.19	0.06	10.20396825	2.92	3.495
0.1	3	513.69	0.08	10.27385845	2.9	3.543
0.13	3	517.2	0.1	15.51706897	2.87	5.407
0.16	3	522.6	0.13	15.67926829	2.84	5.521
0.2	3	528.16	0.16	21.12652582	2.8	7.545

2.14.2 Outcompeting experiments

JAK2 ligand (FTPDpTyr^{148}ELLTEN) from GenScript was dissolved in the same buffer as the protein-ligand complex (50 mM Tris pH 6, 50 mM NaCl, 2 mM DTT) to make a 0.5 mM stock solution.

The outcompeting NMR experiments were performed by titrating the complex of SH2-C-terminal peptide (0.08 mM protein and 0.2 mM ligand) with a series of concentrations of unlabelled synthetic JAK2 ligand peptide in five titration points up to a final JAK peptide concentration of 0.02 mM, as shown in table 2.14. The outcompeting binding was monitored via chemical shift changes arising upon binding using ^{15}N HSQC spectra, its acquisition parameters were copied from the HSQC spectrum of SH2 that was used for the backbone assignment.

Table 2.13. The outcompeting of SH2-C terminal peptide with JAK ligand, based on y =[v3(c1-c3)]/(c2-c1)], c1 is the required ligand concentration to be added to get desired ligand in mM, c2 is the concentration of ligand stock in mM, v3 is volume of NMR sample in µl, c3 is the previous concentration of ligand present in the NMR sample, y is the final ligand addition to sample.

c1	c2	v3	c3	[v3(c1-c3)]	(c2-c1)	y
0.001	0.5	528.163	0	0.528163	0.499	1.1
0.003	0.5	528.163	0.001	1.056326	0.497	3.2
0.006	0.5	528.163	0.003	1.584489	0.494	6.3
0.01	0.5	528.163	0.006	2.112652	0.49	10.6
0.02	0.5	528.163	0.01	5.28163	0.48	21.1

Chapter 3 Protein construct, production, purification and phosphorylation

This chapter focuses mainly on the production of high quality protein samples for nuclear magnetic resonance (NMR) studies such as 3D structure determination and ligand screening. The production of good quality and quantity protein sample is dependent on the solubility and stability of the protein which is influenced by its native fold. The protein needs to be stable as a monomeric molecule for prolonged NMR measurements, and should ideally be uniformly labelled with the stable isotopes ^{13}C and ^{15}N.

Effective strategies for protein production are based on optimising a number of factors: identification of the best boundaries of the target protein; designing the protein construct; cloning the gene construct into the optimal expression vector using an appropriate assembly method; analytical scale expression of tagged protein using T7 based *Escherichia coli* systems; and solubility screening. Once the tagged protein is expressed in good quantity, the affinity tag enhances the efficiency of the purification procedure and thus results in sufficient purity levels of protein for NMR studies. This chapter describes the testing and optimisation of these steps.

This chapter describes the expression of several different constructs. Initial experiments concentrated on the consensus SH2 domain sequence from SH2B1. This proved to be unstable and not well folded, so a slightly longer version was produced and behaved much better. These constructs are referred to here as SH2. The chapter then describes expression of the SH2 domain together with the rest of the C-terminal sequence from the β isoform, which is referred to as SH2c.

Also in this chapter, optimization of enzymatic phosphorylation of the wild type SH2c protein and the mutant SH2c Y114F by Fer kinase enzyme are discussed in detail. The phosphorylation of SH2c protein was monitored via ESI MS, and the phosphorylated site defined via ETD MS. Also the phosphorylation of mutant protein was checked by ESI MS and the phosphorylated site was confirmed by 2D NMR analytical method based on changes in chemical shift. Two phosphorylateable tyrosine sites by Fer kinase were identified at the C terminal of SH2c protein: 114 and 139 which correspond to 624 and 649, respectively, in the full length protein sequence SH2B1 β isoform.

3.1 Optimising gene construct, expression and purification of SH2-MBP fusion protein

3.1.1 SH2-MBP fusion protein: gene design and expression

A plasmid suitable for expression of the SH2 domain of mouse SH2B1 (residues 527-625*)* was designed and built based on a pET system (6880 bp) (Figure 2.1).The plasmid was modified in Dr Mesnage lab and it is unpublished. This expression plasmid was used because it has an N-terminal histidine tag sequence followed by a maltose binding protein (His$_6$-MBP). The MBP tag should help to enhance the solubility of its fusion protein partner and the His$_6$ sequence is used as an affinity tag in the purification stage (Kapust and Waugh, 1999; Schmitt et al., 1993).

The complete digestion of the pET (6869 bp) vector by the restriction enzyme pair NcoI and BamHI to remove 11 bp and generate a linearised plasmid was done as described in section 2.4.2, and confirmed by a 1% agarose gel (Figure 3.1). The linearized plasmid was purified using a GeneJET Gel Extraction Kit with a good yield of around 10 ng/µl.

The SH2 domain synthetic gene was ligated into the linearised plasmid using the Gibson assembly method (Gibson et al., 2009). It was designed as a g.Block SH2 gene (379 bp) with overlapping ends and was built via SnapGene software as shown in Figure 2.3. This designed g.Block gene was cloned successfully into the linearised protein expression pET plasmid (6869 bp) using the Gibson assembly method. The complete ligation of the plasmid with the target gene to create pET-1 (7183 bp) was verified through Sanger sequencing carried out by Eurofins Genomics (Figure 3.2).

The complete construct of the pET-1 expression system (7183 bp) was transformed into *Escherichia coli* BL21 (DE3) to overexpress the fusion protein (His$_6$-MBP-TEV-SH2, 55kDa).

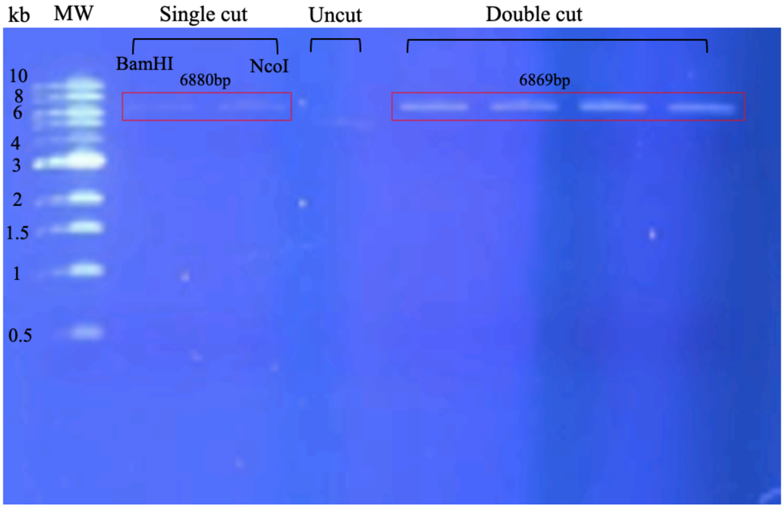
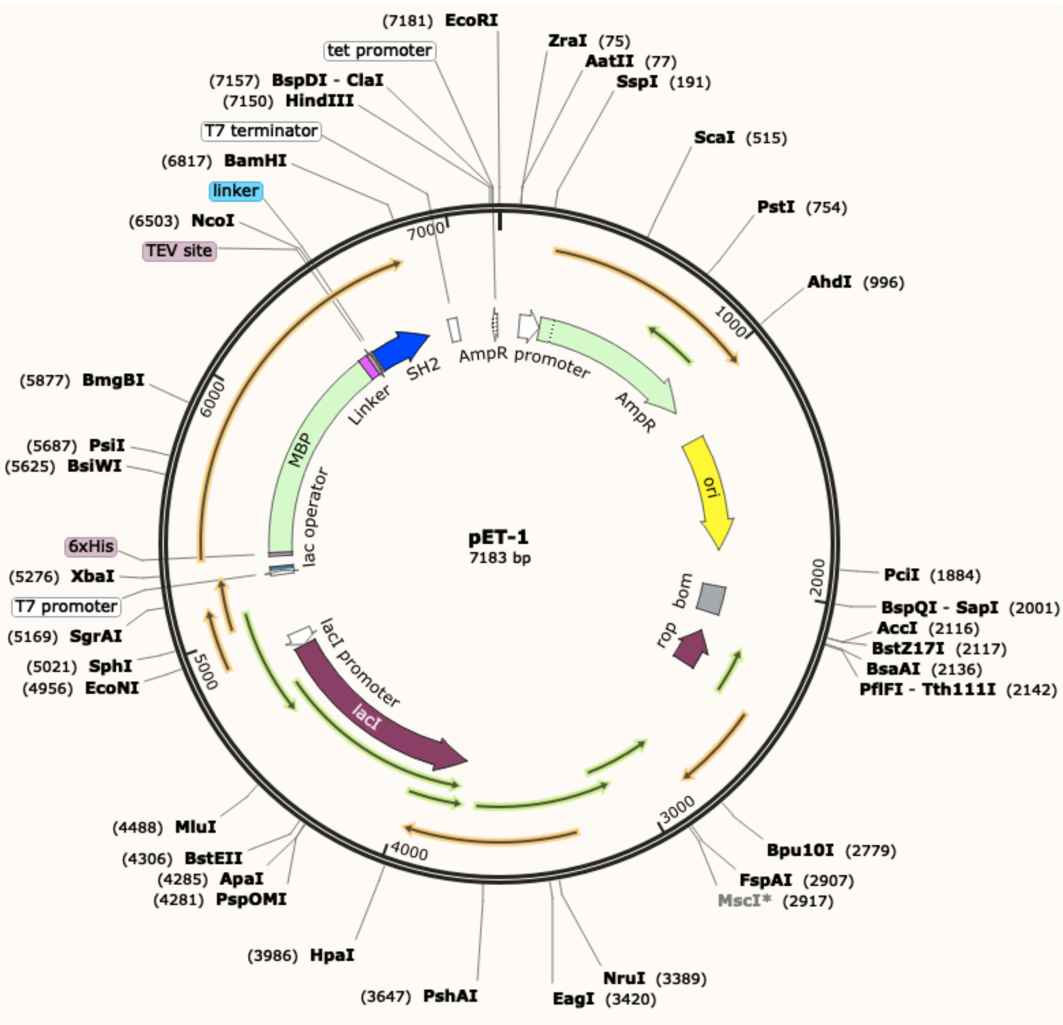


Figure 3.1. Agarose gel of the digestion of pET (6880 bp). Lane 1: Low MW marker DNA shown in kb; lane 2: plasmid cut with BamHI (6880bp); lane 3: plasmid cut with NcoI (6880 bp); lane 4: uncut pET plasmid; lanes 5-8: lanes 5, 6, 7, 8: plasmid cut with both BamHI and NcoI (6869 bp).



```
        10         20         30         40         50         60
GYPWFHGMLS RLKAAQLVLT GGTGSHGVFL VRQSETRRGE YVLTFNFQGK AKHLRLSLNE
        70         80         90        100
EGQCRVQHLW FQSIFDMLEH FRVHPIPLES GGSSDVVLVS YV
```

Figure 3.2. The first gene construct of the SH2 protein. Top. Map of pET-1 plasmid with inserted SH2 g.Block in between NcoI and BamHI (7183 bp). Bottom. The amino acid sequence of SH2 protein including the N-terminal small tag GYP in red (101 a.a).

The expression of His$_6$-MBP-tagged SH2 protein (55 kDa) was tested at 25°C and 37°C with different induction times and IPTG concentrations, as explained in section 2.5.1. Cells were grown in 1L LB media to reach an OD$_{600}$ of 0.6, then induced with 0.5 mM or 1 mM IPTG. Samples of cultures which were incubated at 37°C were collected after 3 hours, and at 25°C were collected after overnight induction.

To analyse samples for the presence of soluble proteins, the cell lysates of bacterial cultures were subjected to centrifugation, then the pellet was resuspended in lysis buffer (200 mM NaCl, 50 mM Tris pH 8, DNase). The resuspended pelleted cells were sonicated for a short period then centrifuged at high speed. The resulting supernatant which contained the soluble cell fraction including the target protein was screened and analysed by 16% SDS PAGE after adding the loading buffer and heating samples at 60°C for a minute.

Comparison of the gel band intensities showed there was protein expression of His$_6$-MBP-TEV-SH2 (55 kDa) at both temperatures with different IPTG concentrations, however there was slightly more soluble protein production at low temperature (25°C overnight with 0.5 mM IPTG) than other conditions (Figure 3.3).

Although good expression of the fusion protein was obtained by induction with 0.5 mM IPTG at 25°C overnight, the His$_6$-MBP (43 kDa) on its own is produced at least to the same level or slightly higher than the fusion protein which is most likely because secondary structure on the mRNA weakens ribosome binding and leads to termination of translation. The quantity of the expressed protein in inclusion bodies was not screened as the concentration of the soluble protein was good.

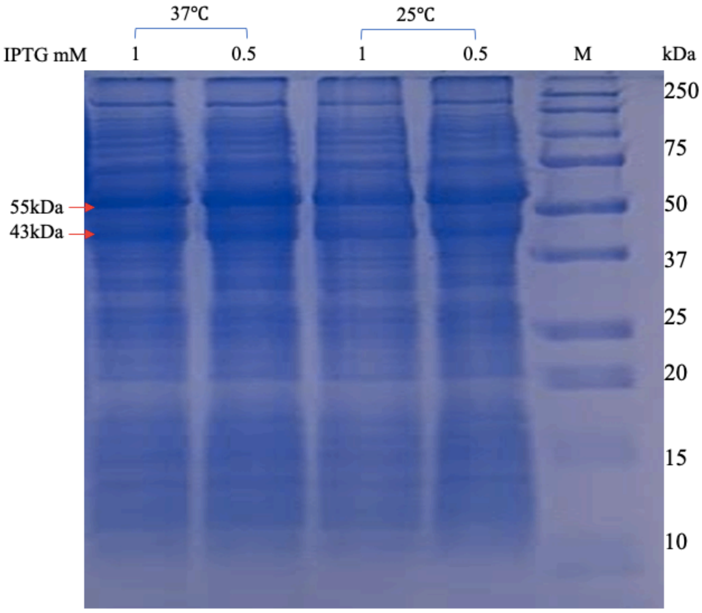


Figure 3.3. SDS PAGE showing the overexpression of fusion protein (His-MBP-TEV-SH2, 55 kDa) in the soluble supernatant at 25°C and 37°C. From right to left, lane 1 M shows the MW markers indicated in kDa. Lanes 2 and 3: expression at 25°C with 0.5 mM and 1mM IPTG overnight; lanes 4, 5: expression at 37°C with 0.5 mM and 1mM IPTG for 3 hrs. 43 kDa is the expected size of His-MBP-TEV tag expression.

3.1.2 Purification of SH2-MBP fusion protein

As the first step of protein purification, an amylose column (8 ml) was used to purify the overexpressed soluble His$_6$-MBP tagged SH2 proteins from the cell extract. The supernatant was applied to the column which was equilibrated in buffer containing 200 mM NaCl, 50 mM Tris pH 8. The column was washed with the same buffer and elution was performed with 200 mM NaCl, 50 mM Tris pH 8 with 10 mM maltose buffer. The collected washing and elution fractions showed higher optical density analysed in SDS-PAGE, Figure 3.4. As shown in the gel there was a quantity of fusion proteins and MBP tag bound to the column (Figure 3.4, lane 3), while some of the fusion proteins present in the unbound material flowed through without binding to the column (Figure 3.4, lanes 1, 2).

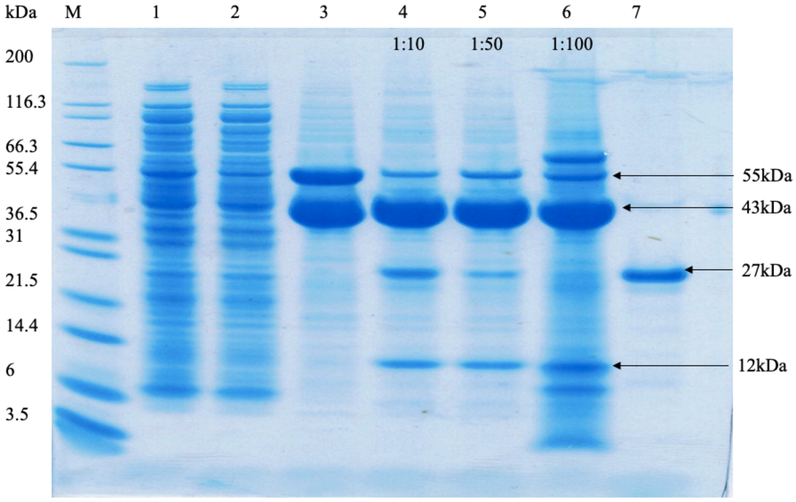


Figure 3.4. SDS-PAGE gel of purification of fusion His-MBP-SH2 protein with amylose column and cleavage reaction of fused proteins with TEV protease. Lanes 1 and 2 unbound fractions from amylose column; lane 3 concentration of elution fractions of amylose column with buffer (200 mM NaCl, 50 mM Tris pH 8 with 10 mM maltose). Cleavage reaction of fusion protein (His-MBP-TEV-SH2, 55kDa) with TEV protease in ratio: lane 4, 1:10, lane 5, 1:50, lane 6, 1:100. Lane 7 TEV enzyme 27 kDa. Fusion protein 55 kDa, His-MBP tag 43 kDa, SH2 protein 12 kDa.

TEV protease is a site-specific endo-protease that cuts between Gln and Ser in the recognition sequence GluAsnLeuTyrPheGlnSer, which was used to cleave the His-MBP tag of the SH2 fusion protein. The cleavage reaction of fusion protein (55 kDa) with TEV protease was tested in different ratios of protease to protein (1:10, 1:50, 1:100) at 30°C for 1h.

In order to measure the efficiency of tag cleavage, samples were extracted before and after cleavage treatment and run in SDS-PAGE. Comparison of gel band intensities revealed similar levels of cleavage (> 85%) of fusion proteins in the samples (1:10) and (1:50) ratio protease to protein in 200 mM NaCl, 50 mM Tris pH 8, 1 mM DTT buffer (Figure 3.4, lanes 4, 5). Moreover, there was less than 15% of fusion proteins that were un-cleavable.

After the cleavage reaction, to separate His$_6$-MBP tag, TEV protease and the remaining fusion protein from SH2 protein, a Ni-NTA column was employed. The TEV protease is His-tagged, so all proteins except the cleaved SH2 should appear in the elution, with only the cleaved SH2 appearing in the run-through and wash. The cleavage sample was applied to a Ni-NTA (10 ml) column which had been equilibrated and washed with 50 mM Tris pH 8, 500 mM NaCl. The elution was performed with the same buffer with the addition of 500 mM imidazole. Both washing and elution fractions were mixed and concentrated to run in SDS-PAGE for analysis (Figure 3.5, lanes 1, 2, 3). It was noticeable in SDS-PAGE that about 50% of the fusion proteins and MBP tag appeared in the unbound fractions with SH2 protein. Moreover a quantity of SH2 protein existed in the elution fractions.

It therefore seems likely that there is an unexpectedly strong interaction between the MBP tag and the SH2 that prevents proper cleavage and separation. The His$_6$-MBP tag (pI 5.3) and fusion proteins (pI 5.8) are defined as acidic according to their isoelectric point values, while the SH2 protein by itself with a pI of 9 is a basic protein. Because of the big difference between the pI values of the His$_6$-MBP tags and fusion proteins compared to the SH2 protein, a strong electrostatic interaction may be occurring in buffer at pH 7-8 with low salt concentration, which may explain the unexpected binding after cleavage of the fusion protein.

To weaken the strong electrostatic interaction between the two molecules (SH2 and His$_6$-MBP tag or SH2 and fusion protein) and thereby separate them by using gel filtration, an experiment was performed with high salt concentration. The sample (1 ml, 1-4 mg/ml) was applied to a Superdex^{TM} 75 10/300 size exclusion column. The GF experiment was performed in 0.5 M NaCl with 50 mM Tris pH 8 buffer at flow rate 1 ml/min using the ÄKTA prime plus system.

As shown on the GF trace in Figure 3.5, there were two peaks eluted. The peak fractions were analysed by SDS-PAGE. The first one, in the void volume, corresponds to aggregate containing fusion protein (55 kDa), His$_6$-MBP tag (43 kDa) and aggregated SH2 proteins (12 kDa). The second peak corresponds to a molecular weight about 40 kDa and also contains the His-MBP tag. Unfortunately, SDS-PAGE gel analysis revealed most of the desired 12 kDa material is present in the first peak in aggregates (Figure 3.5, lanes 5, 6, 7, 8, and 9).

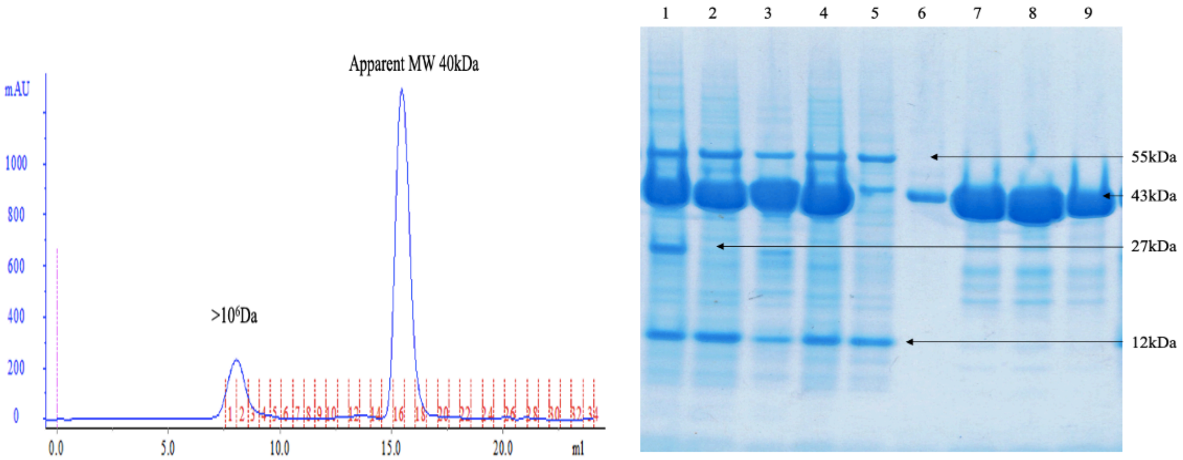


Figure 3.5. SDS gel of Ni purification and gel filtration purification of cleavage fusion protein (His$_6$-MBP-SH2). Right. Lane 1, cleavage of fusion proteins with TEV enzyme; lane 2, Ni column unbound fraction, lane 3, Ni column elution fraction, lane 4, GF loaded sample, lane 5, GF fraction 3, lane 6 GF fraction 16, lane 7 GF fraction 17, lane 8 GF fraction 18. Lane 9 GF fraction 19. 55 kDa fusion protein, 43 kDa MBP tag, 27 kDa TEV enzyme, 12 kDa SH2. Left. Gel filtration trace of cleavage fusion protein purification performed in 0.5 M NaCl with 50 mM Tris pH 8 buffer, the SDS gel is taken from previous SDS gel 3.4.

As the high salt concentration was not enough to separate the two molecules and to reduce the possible electrostatic attraction, an attempt was made to reduce the His$_6$-MBP tag solubility to minimum via performing the gel filtration experiment with buffer at a pH that is close to its pI value (5.03) with a high salt concentration to minimize the electrostatic interaction between the molecules. The GF experiments were performed with buffers 0.5 M NaCl with 50 mM Tris pH 5.6. As a consequence of reducing the pH of the buffer to 5.6, most of the His$_6$-MBP tags and fusion proteins were aggregated and came in the first GF peak with SH2 proteins. On the other hand, the concentration of the salt was no help to separate the SH2 proteins from the MBP tag as shown in an SDS-PAGE gel (Figure 3.6).

Because of the strong interaction between the SH2 protein and cleaved His$_6$-MBP tag or fusion protein after the expression which caused a co-purification, a new expression system was generated for the SH2 gene with N terminal Histidine tag and without the MBP tag.

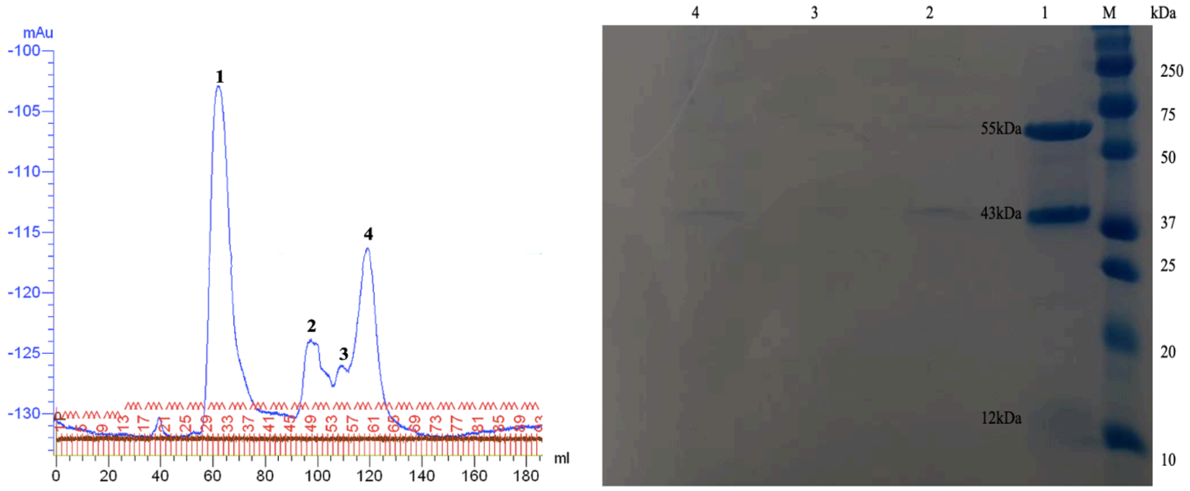


Figure 3.6. The gel filtration purification of cleavage fusion protein (His$_6$-MBP-SH2) at 0.5 M NaCl. Left. The gel filtration trace in buffer 0.5 M NaCl 5.6 pH (injected sample 1.5 ml/ 2.7 mg/ml). Right. The fractions of GF; lane 1, peak1 (55 kDa, 43 kd MBP, 12 kd SH2); lane 2, peak 2; lane 3, peak 3; lane 4, peak 4.

3.2 Optimising gene construct, expression and purification of Histidine tagged SH2 protein

3.2.1 Histidine tagged SH2 protein without MBP tag: gene design and expression

The construct for the MBP fusion had no suitable restriction enzyme digestion sites to allow a simple removal of the MBP and keep the N-terminal Histidine tag. Therefore it has been decided to remove the complete His$_6$-MBP-TEV-SH2 insert, and insert a new SH2 sequence with a His$_6$ tag, again using the Gibson assembly method.

To remove the MBP tag from the pET-1 plasmid (7183 bp), the best selected restriction enzymes were XbaI and BamHI to perform a double digestion in CutSmart buffer. Similar restriction digestion procedures were followed to cut the pET-1 plasmid with those enzymes and remove the His$_6$-MBP-TEV tag (1541 bp), however the digestion also removed the essential DNA sequence of the Ribosome Binding Sequence which was included in the g.Block. The size of the resultant linear plasmid (5642 bp) was confirmed in a 1% agarose gel as shown in Figure 3.7. According to the agarose gel result half of the plasmid was double digested (5742 bp) and the rest was single cut (7183 bp), probably because of inefficient cutting by the BamHI enzyme as shown in the control BamHI cut (lane 3, Figure 3.7). However, the concentration of the double cut purified plasmid (~18 ng/µl) was enough to do the ligation.

A second SH2 gene g.Block (443 bp) was designed via SnapGene software, Figure 2.5, and ordered as a synthetic DNA sequence. Gibson assembly worked efficiently to subclone the linearized plasmid (5642 bp) with a new g.Block SH2 gene. The sequence of the resulting plasmid pET-2 (6016 bp) was confirmed by sequencing by Eurofins Genomics, Figure 3.8.

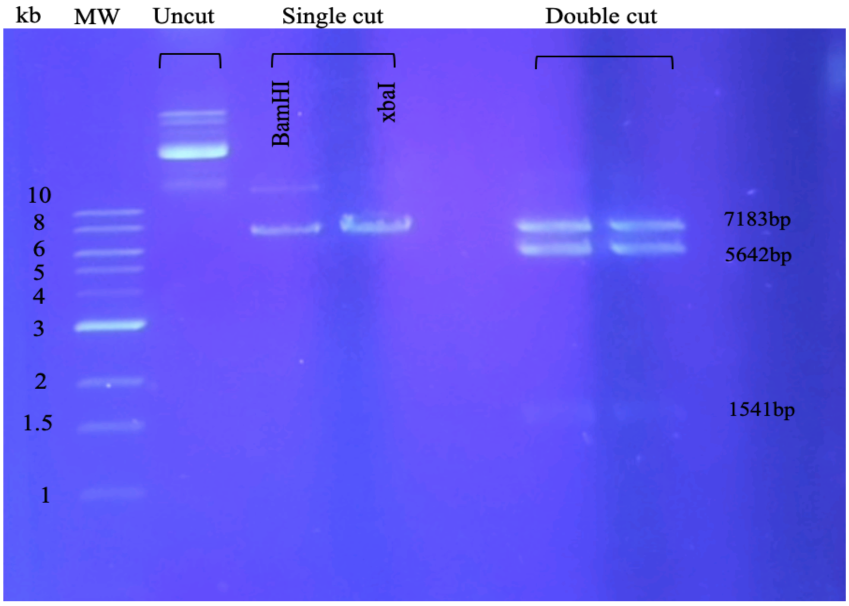
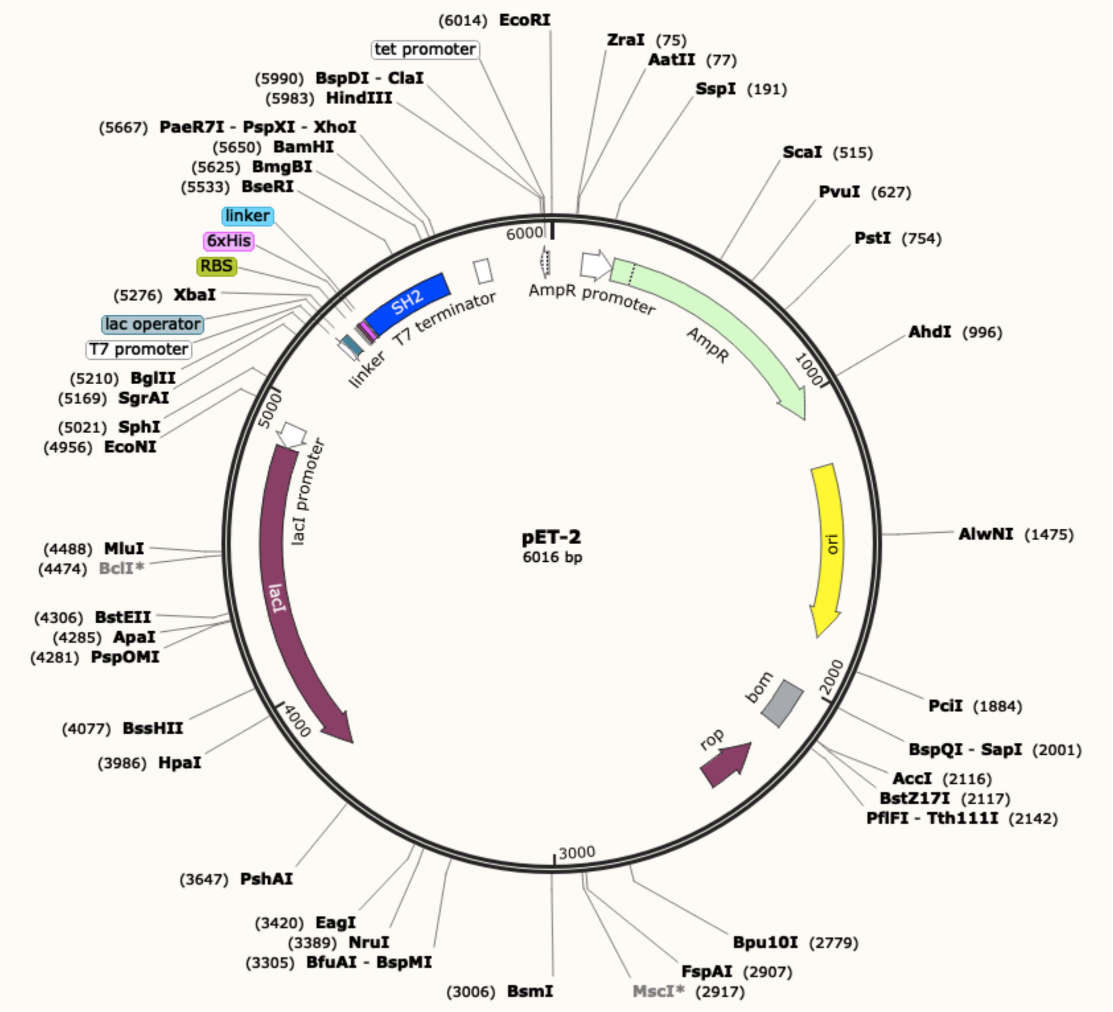


Figure 3.7. 1% agarose gel of the digestion of pET-1 plasmid (7183bp). Lane 1: DNA MW markers. Lane 2: uncut pET-1 plasmid; Lane 3: plasmid cut with BamHI (7183bp); Lane 4: plasmid cut with XbaI (7183 bp); Lanes 6 and 7: plasmid cut with both BamHI and XbaI enzymes. Expected fragment sizes are shown on the right.



```
       10         20         30         40         50         60
KIHHHHHHGS WFHGMLSRLK AAQLVLEGGT GSHGVFLVRQ SETRRGEYVL TFNFQGKAKH

       70         80         90        100
LRLSLNEEGQ CRVQHLWFQS IFDMLEHFRV HPIPLESGGS SDVVLVSYV
```

Figure 3.8. The second construct of SH2 protein. Top. Map of the pET-2 plasmid with the inserted mouse SH2 gene (6016 bp). Bottom. The amino acid sequence of the Histidine tagged SH2 protein consists of the N-terminal sequence of the KI histidine tag sequence followed by a small GS linker in red. The SH2 domain sequence including the tag consists of 109 a.a.

The pET-2 plasmid was transformed into BL21 (DE3) cells for protein expression. A good level of protein expression was demonstrated in the first attempt with the MBP fusion, thus similar conditions were applied to express the His$_6$-SH2 protein (12 kDa), by induction with 0.5 mM IPTG at 25°C overnight. The protein expression was screened first in the collected supernatant. The initial expression of His$_6$-SH2 produced a quantity of soluble protein as homo-oligomers as shown in Figure 3.9, lanes 1, 2 and 3.

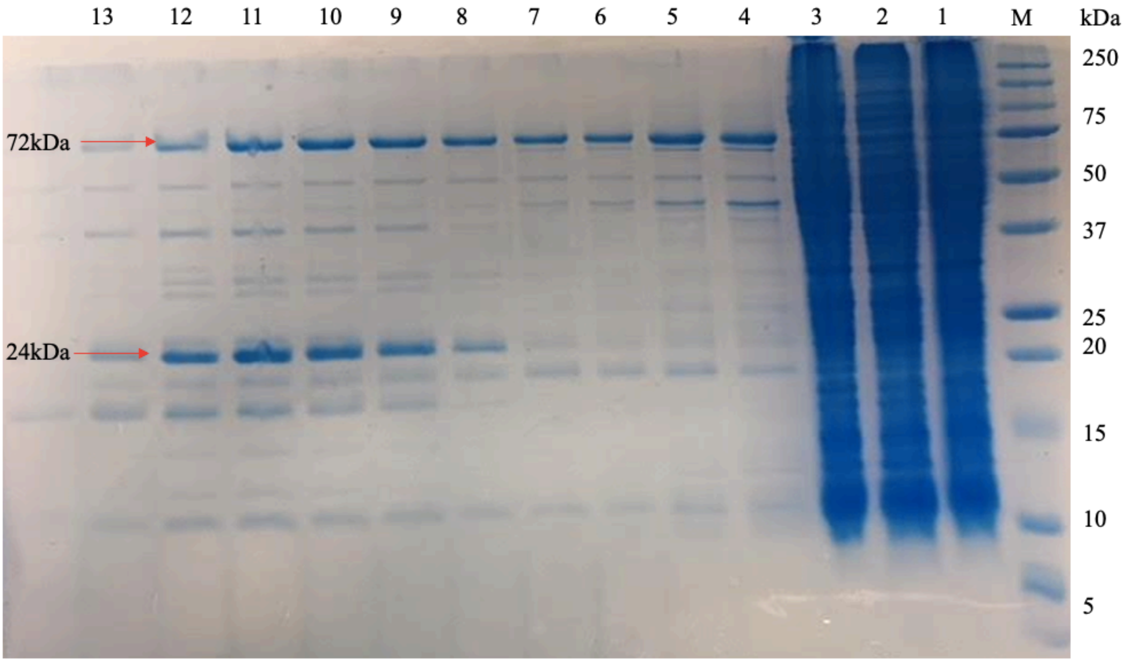


Figure 3.9. SDS-PAGE gel showing the soluble expression of His$_6$-SH2 protein and the purification of His$_6$-SH2 proteins by Ni-NTA column. Lane 1, 2, and 3 are the expression of soluble SH2 following induction with 0.5 IPTG at 25°C overnight. Lanes 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 are elution fractions with buffer (300 mM imidazole, 50 mM Tris pH 8, 300 mM NaCl). Expected sizes are 12 kDa SH2 as a monomer, 24kDa is dimer of SH2, and 72kDa is hexamer of SH2. MW of proteins are displayed in kDa.

In order to check the concentration of the His$_6$-SH2 proteins in the pellet, the obtained pellet from 500 ml growth at 25°C overnight was dialysed for three days as described in section 2.5.2.4. After dialysis for 3 days, the resulting supernatant and the pellet were run on an SDS-PAGE gel. As shown in the gel (Figure 3.10), there is a big concentration of monomer of SH2 proteins in the supernatant (lanes 1 and 2) and the pellet (lanes 3 and 4). A high concentration of expressed protein accumulated into inclusion bodies (in the obtained supernatant and pellet) is normally related to instability of the protein in solution, possibly due to being fully or partly misfolded and thus aggregated.

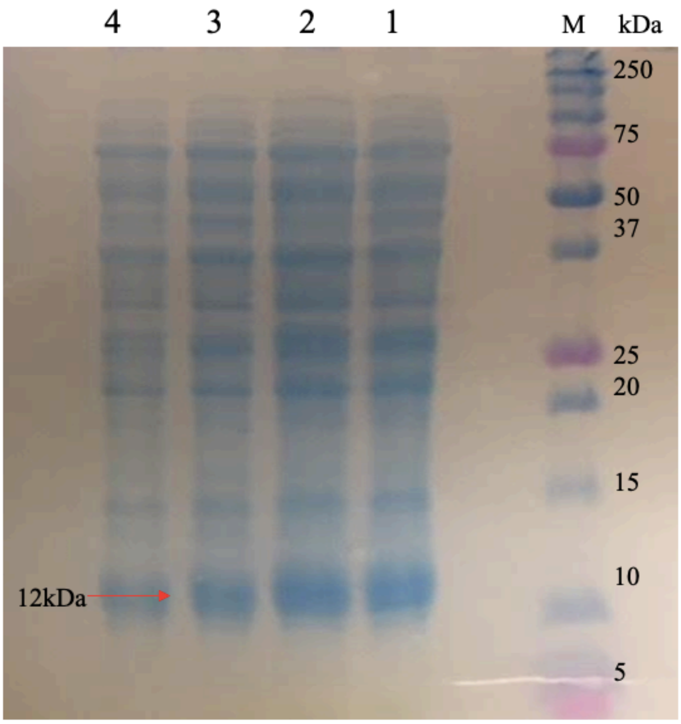


Figure 3.10. SDS-PAGE gel analysis of His$_6$-SH2 pellet dialysis. Lanes 1 and 2 are the supernatant. Lanes 3 and 4 are the pellet: the final dialysis of SH2 in 50 mM KH_2PO_4, pH 7.4 at 4°C; monomer 12 kd. MW of proteins is shown in kDa. The SDS gel shows that cells pellet contained a good concertation of monomer SH2.

3.2.2 Purification of the Histidine tagged SH2 protein without MBP

After expression of the second protein construct in good quantity, a Ni-NTA column was used to purify the overexpressed His$_6$-SH2 protein (12 kDa). ~ 20 ml of the supernatant from 1 Liter culture was applied to the column (10 ml Ni-NTA resin), then it was washed with buffer A (50 mM Tris pH 7.4, 200 mM NaCl, 40 mM imidazole) and the flow-through was collected, followed with eluting bound proteins with buffer B (50 mM Tris pH 7.4, 200 mM NaCl, 300 mM imidazole) and fractions were collected. The elution fractions showing higher optical density in the fraction collection trace were analysed in SDS-PAGE (Figure 3.9, lanes 4-13). On SDS-PAGE gel the eluted fractions were seen to form homo-oligomers (dimer and hexamer) in high concentration.

The elution fractions were concentrated and injected onto a gel filtration column under buffer condition 300 mM NaCl and 50 mM Tris pH 8. Based on the analytic SDS-PAGE gel, the first peak contained homo-hexameric SH2 proteins, the second peak contained homo-dimeric SH2 proteins. However, there was no monomeric SH2 proteins in any peak as revealed in SDS-PAGE (Figure 3.11). It is clear the soluble protein resulting from unfolded monomer associated into stable homo-oligomers.

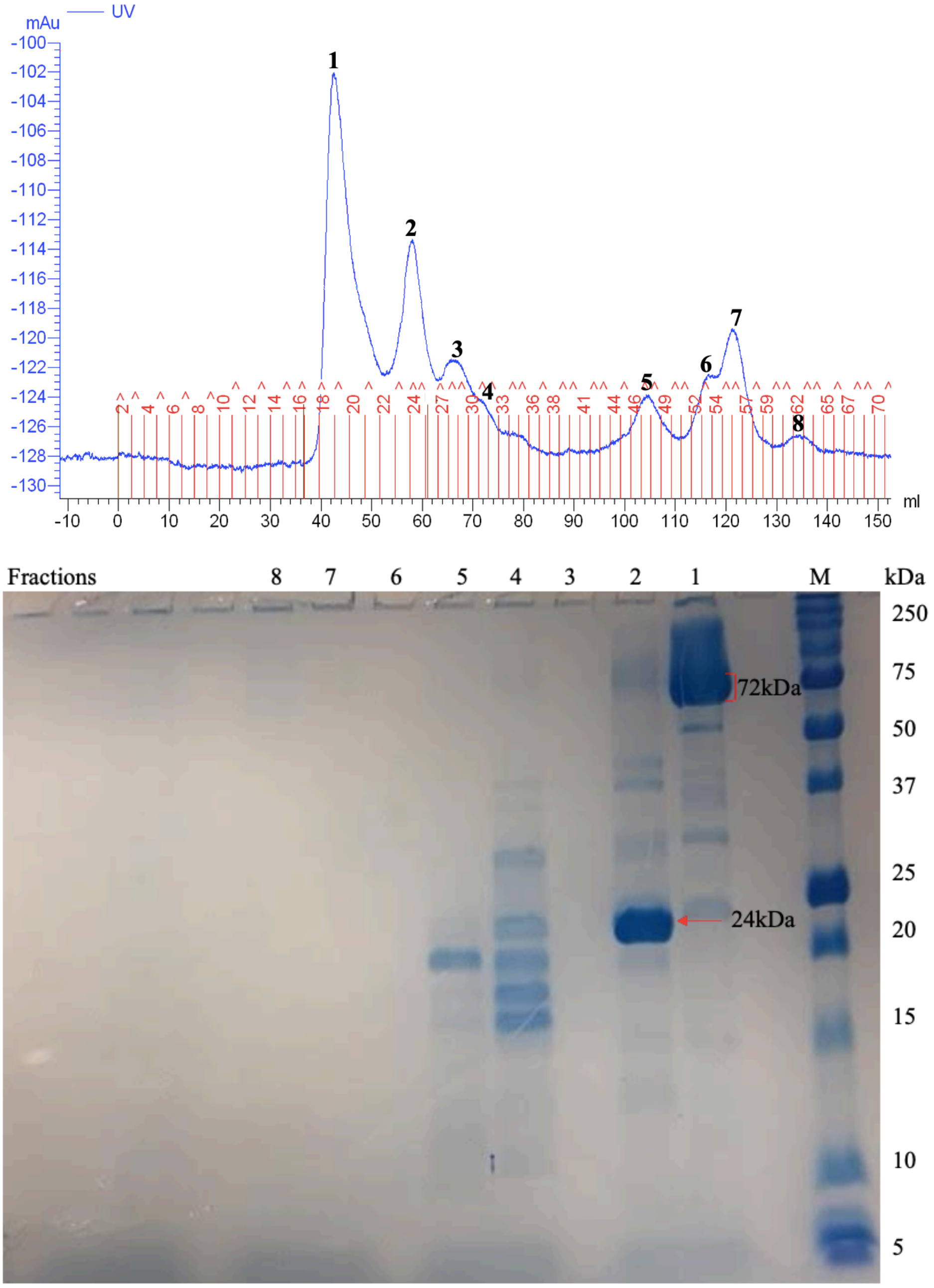


Figure 3.11. The gel filtration purification of the His tagged SH2 protein. Top. The trace of the gel filtration in buffer 300 mM NaCl and 50 mM Tris pH 7.4. Bottom. The GF peaks; lane 1, peak 1 (72 kDa hexamer SH2), lane 2, peak 2 (24 kDa dimer SH2); lane 3, peak 3; lane 4, peak 4; lane 5, peak 5; lane 6, peak 6; lane 7, peak 7; lane 8, peak 8.

According to the physical and the chemical properties of the SH2 proteins, the SH2B1 SH2 domain is expected to be a stable independent folding unit and function as a monomer (Hu et al., 2006). Despite this, the results from the first and second protein expression demonstrate that the overexpressed SH2 protein tends to bind to its cleavage tag (His$_6$-MBP) or forms homo-oligomers causing co-purification as shown in the purification, which implies that the solubility of the SH2 protein is less than expected. Based on the previous observation the SH2 domain boundaries were investigated, and thus whether the start and end of the construct should be altered.

3.2.3 Histidine tagged SH2 protein with extra terminal sequence: gene design and expression

A sequence alignment of the SH2 domain with other adaptor and scaffold containing SH2 proteins showed there are N-terminal and C-terminal matching sequences outside the standard domain boundaries (Figure 3.12). Although the aligned N terminal region preceding the SH2 domain is less well conserved than the C terminal sequences, it seems important for the stable folding of the intact SH2 domain. The defined boundaries of the classical SH2 domain comprise residues 527-625 (in Uniprot) which are somewhat smaller than the full folded domain revealed by Pascal et al., 1994, and Hu and Hubbard, 2006. The paper reporting the crystal structure used a construct with additional residues at both ends with no explanation of the reason for the extension, although it does look like the extensions form an integral part of the protein structure and are thus required for correct folding and stability. Therefore a new g.Block (461 bp) was constructed to contain additional terminal residues to the sequence of the SH2 domain (519-628); eight at the N terminus (DQPLSGYP) and three at the C terminus (PSQ), as shown in Figure 2.13.

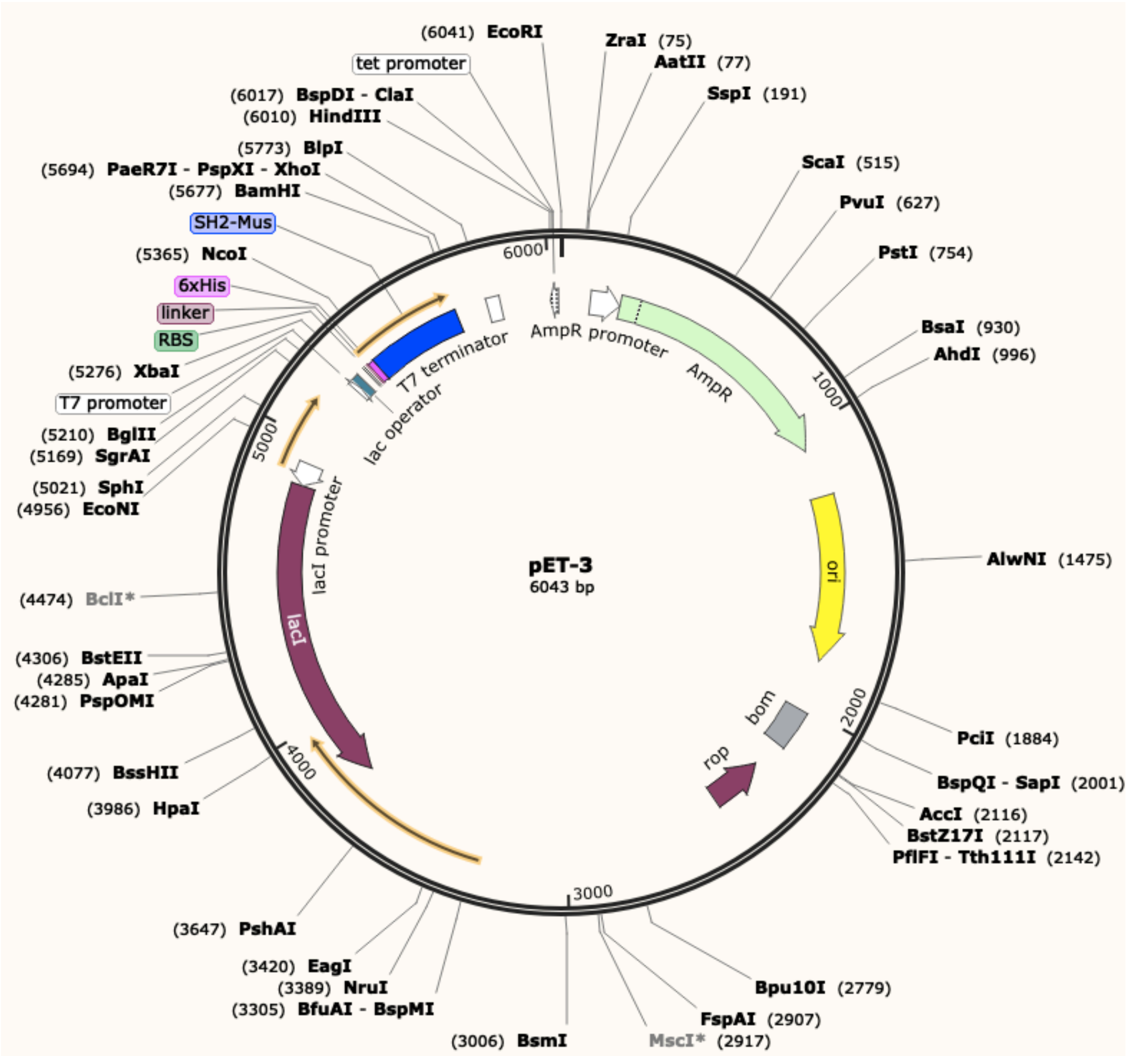
```
Homo.DAPP1|Q9UN19|35-129        TR--------------------WF--TLHRNELKYFKDQ---MSPEPIRILDLTEC----
Homo.STAP1|Q9ULZ2|177-280       PAC-------------------FY--TVSRKEATEMLQK-----NPSLGNMILRPGSDSR
Homo.CLNK|Q7Z7G1|309-419        KRSDRKDVQ----------HNEWYIGEYSRQAVEEAFMK-----ENKDGSFLVRDCSTKS
Homo.SH3BP2|P78314|457-555      SQADTGGDDSDEDYEKVPLPNSVFVNTTESCEVERLFKATSPRGEPQDGLYCIRNSSTKS
Rat.Grab7|Q9QZC5|434-530        TPCSGLSLS-----AAIHRTQPWFHGRISREESQRLIGQ----QGLVDGVFLVRESQRNP
Mus.Shc1|P98083|484-575         PPPQSMSMA-----EQLQ-GEPWFHGKLSRREAEALLQL--------NGDFLVRESTTTP
Mus.sh2b1|Q91ZM2|527-625        GESEGGEGD-----QPLS-GYPWFHGMLSRLKAAQLVLE---GGTGSHGVFLVRQSETRR
                                                       :                            :

Homo.DAPP1|Q9UN19|35-129        ----SAVQFDYSQE--RVNCFCLVFPFRTFYLCAKTGVEADEWI----KILRWK------
Homo.STAP1|Q9ULZ2|177-280       N---YSITIRQEIDIPRIKHYKVMSVGQNYTI----ELEKPVTLPNLFSVIDYFVKETRG
Homo.CLNK|Q7Z7G1|309-419        KEEPYVLAVFY-EN--KVYNVKIRFLERNQQFALGTGLRGDEKFDSVEDIIEHY-KN---
Homo.SH3BP2|P78314|457-555      GK--VLVVWDETSN--KVRNYRI-FEKDSKFY-----LEGEVLFVSVGSMVEHY-HTHV-
Rat.Grab7|Q9QZC5|434-530        QG--FVLSLCHLQ---KVKHYLI-LPSEDEGCLYFSMDDGQTRFTDLLQLVEFH-QLNRG
Mus.Shc1|P98083|484-575         GQ--YVLTGLQ-SG--QPKHLLLVDPE-------GVVRTKDHRFESVSHLISYHMDNH--
Mus.sh2b1|Q91ZM2|527-625        GE--YVLTFNF-QG--KAKHLRL-SLNEEGQC-----RVQHLWFQSIFDMLEHF-RVHP-
                                      :         .     :                    :     ::

Homo.DAPP1|Q9UN19|35-129        -LSQIRKQLNQGEGTIRSRSFIFK------------------------------------
Homo.STAP1|Q9ULZ2|177-280       NLRPFICSTDENTGQEPSMEGRSEKLKKNPHIA---------------------------
Homo.CLNK|Q7Z7G1|309-419        -FPIILIDGKDKTGVHRKQCHLTQPLPLTRHLLPL-------------------------
Homo.SH3BP2|P78314|457-555      -LPSHQS------LLLRHPYGYTGPR----------------------------------
Rat.Grab7|Q9QZC5|434-530        ILP----------CLLRHCCARVAL-----------------------------------
Mus.Shc1|P98083|484-575         -LPIISAGSE---LCLQQPVDRKV------------------------------------
Mus.sh2b1|Q91ZM2|527-625        -IPLESGGSSDVVLVSYVPSQRQQERSTSRDPAQPSEPPPWTDPPHPGAEEASGAPEVAA
                                 :
```

Figure 3.12. ClustalW2 amino acid sequence alignment of SH2 SH2B1 (Mouse, *Q91ZM2*) with scaffold and adaptor proteins. Yellow highlighted residues show the N and C termini of the classical SH2 domain (boundaries), and the red highlighted sequences shows the matching additional sequences between SH2 SH2B1 and other proteins. Each protein sequence is defined by entry name in Uniprot and the N and C terminal boundaries of the SH2 domain. Homo.DAPP1|Q9UN19 is dual adaptor for phosphotyrosine and 3-phosphotyrosine and 3-phosphoinositide (residues 35-129), Homo.STAP1|Q9ULZ2 is signal-transducing adaptor protein 1 (residues 177-280), Homo.CLNK|Q7Z7G1 is cytokine-dependent hematopoietic cell linker (residues 309-419), Homo.SH3BP2|P78314 is SH3 domain-binding protein 2 (residues 457-555), Rat.Grab7|Q9QZC5 is growth factor receptor-bound protein 7 (residues 434-530), Mus.Shc1|P98083 is SHC-transforming protein 1 (residues 484-575), Mus.sh2b1|Q91ZM2 is SH2B adapter protein 1 (residues 527-625).

Restriction digestion of the pET-2 plasmid (6016 bp) was performed as before using BamH1 and Xba1 restriction sites (Figure 3.7), which removed 374 bp including the ribosome binding site (RBS) and His$_6$ sequence which were added to the designed g.Block. The target gene was ligated into a linearized plasmid in between the restriction sites using the Gibson method, then the complete DNA sequence of plasmid pET-3 (6043 bp) was checked by Eurofins Genomics as shown in Figure 3.13.

His$_6$-SH2 protein with additional terminal sequences (14 kDa) was overexpressed from the pET-4 plasmid using *E. coli* BL21 (DE3) cells. Following the optimal conditions as shown in the first attempt for protein expression and extraction, the initial expression yield of soluble His$_6$-SH2 protein at 25°C with 0.5 M IPTG overnight was good as shown in the SDS-PAGE (Figure

3.14, lane 1). ^{15}N ,^{13}C doubled labelled tagged SH2 proteins were overexpressed in 1 Liter M9 minimal media following the protocol of Marley et al., 2001 for NMR study.



```
        10         20         30         40         50         60
KIHHHHHHDQ PLSGYPWFHG MLSRLKAAQL VLEGGTGSHG VFLVRQSETR RGEYVLTFNF

        70         80         90        100        110
QGKAKHLRLS LNEEGQCRVQ HLWFQSIFDM LEHFRVHPIP LESGGSSDVV LVSYVPSQ
```

Figure 3.13. The third construct of the His tagged SH2 protein. Top. Map of pET-3 plasmid with inserted mouse SH2 gene (6043 bp). Bottom. The amino acid sequence of the tagged SH2 protein consists of: the N-terminal histidine tag in red, at the N-terminus, and the C-terminal extra residues underlined. The total His$_6$-SH2 sequence is 118 amino acids.

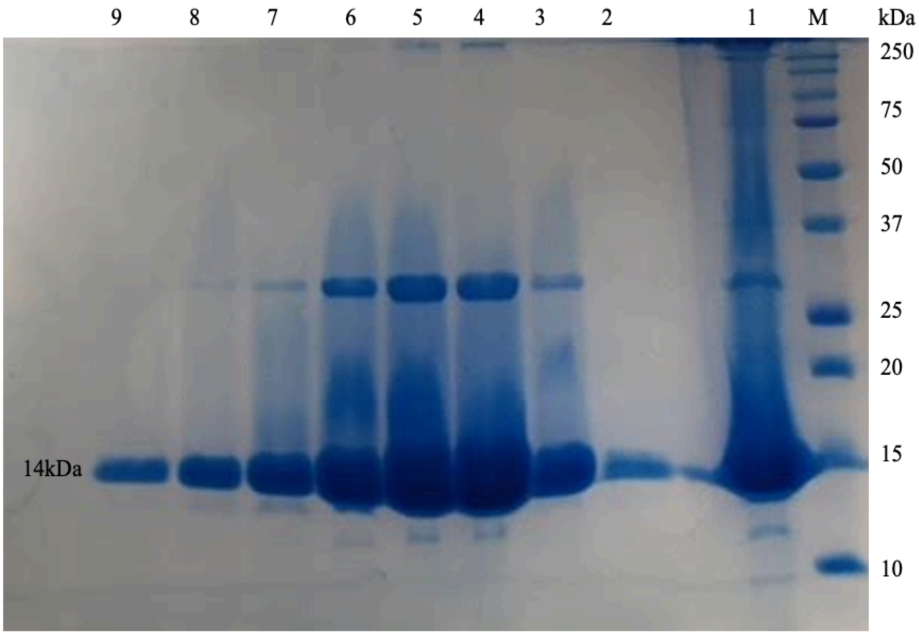


Figure 3.14. 16% SDS-PAGE gel of the supernatant of His$_6$-SH2 protein expression, and purification fractions by Ni column. Lane 1, SH2 protein expression 14 kDa with 0.5 mM IPTG, at 25 °C overnight. The purification of His$_6$-SH2 protein by Ni-NTA column: lane 2, unbound materials; lanes 3, 4, 5, 6, 7, 8, 9 elution fractions with buffer 300 mM imidazole. Monomer His$_6$-SH2 is 14 kDa.

3.2.4 Purification of Histidine tagged SH2 protein with extended terminal sequence

~ 20 ml of supernatant from 1 Liter culture was loaded onto a 20 ml Ni-NTA column to purify the histidine tagged proteins (14 kDa). Good concentrations of proteins were eluted with 50 mM Tris pH 7.5, 200 mM NaCl, and 300 mM imidazole. Purity of protein in the elution fractions was above 80% as estimated using SDS-PAGE (Figure 3.15).

The elution fractions were concentrated to 10 ml and injected onto a Superdex 200 10/300 size exclusion column. The GF experiment was performed in 200 mM NaCl, 50 mM Tris pH 7.5, 2 mM DTT buffer at flow rate 1 ml/min using the ÄKTA prime plus system. As shown on the GF trace a single peak was obtained containing the target protein (14 kDa) in Figure 3.15, and protein purity was above 99% as shown in SDS-PAGE. Because this protein behaved much better than the previous shorter constructs, it was used in all subsequent experiments.

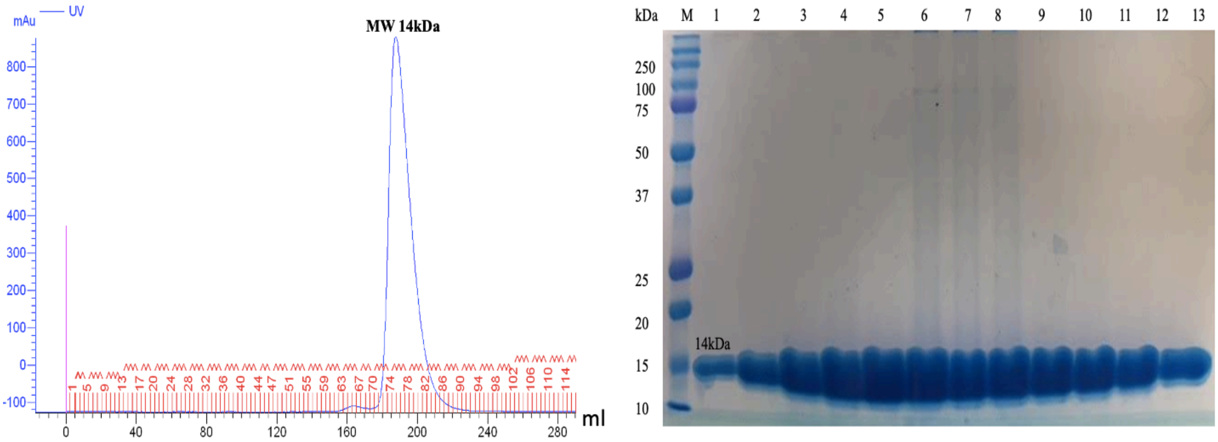


Figure 3.15. The gel filtration purification of the double labelled SH2 protein. Left. The trace of gel filtration performed in 300 mM NaCl, Tris buffer, at pH 7.5. Right. The 16% SDS-PAGE gel of GF fractions: Marker (10-250 kDa), lanes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 peak1 fraction His$_6$- SH2 14 kDa.

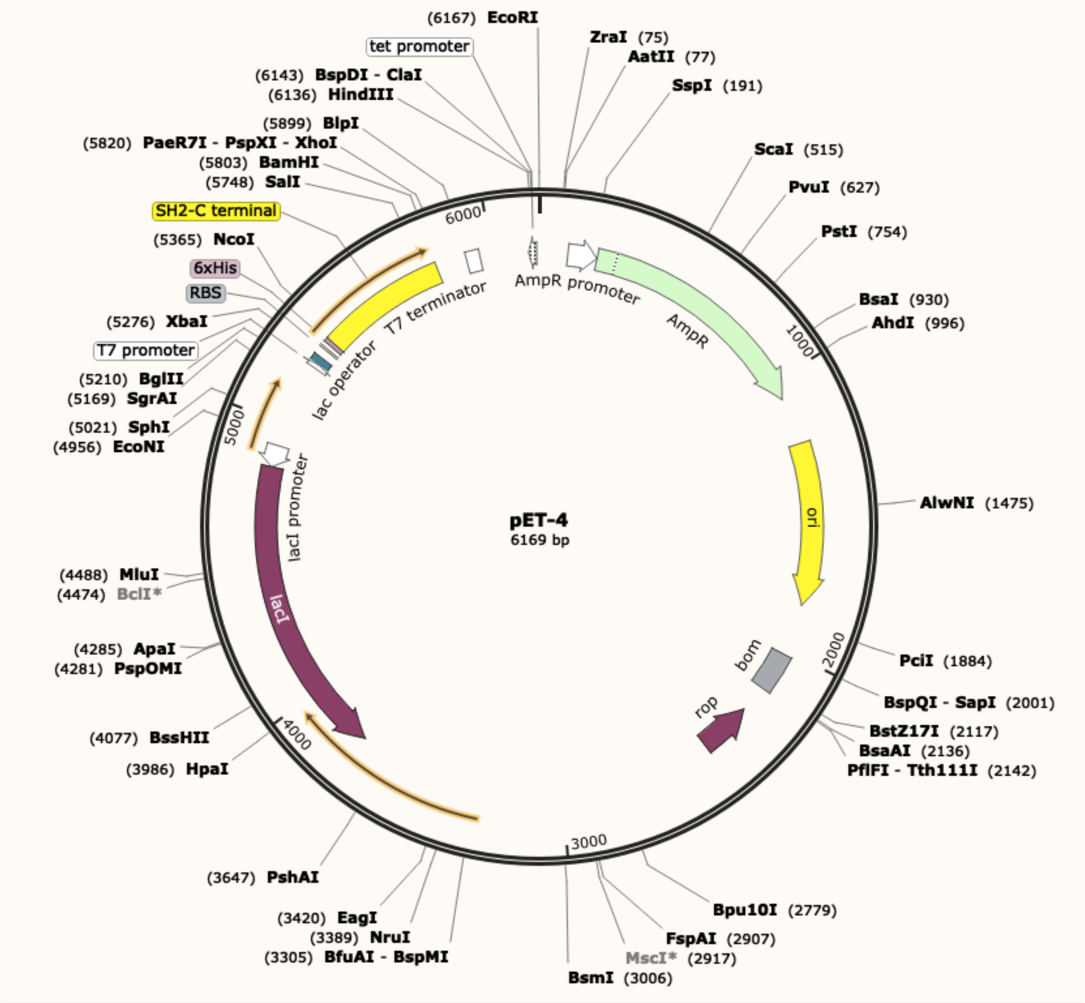
3.3 Optimising DNA construct, expression and purification of SH2 protein with the C terminal tail

3.3.1 SH2 with C terminal tail of SH2B1 β isoform: gene design and expression

A construct of the SH2 domain followed by the C terminus of the mouse SH2B1 β isoform (residues 519-670) was constructed as a DNA gBlock with N and C terminal overlap nucleotide sequences (586 bp) (Figure 2.8).

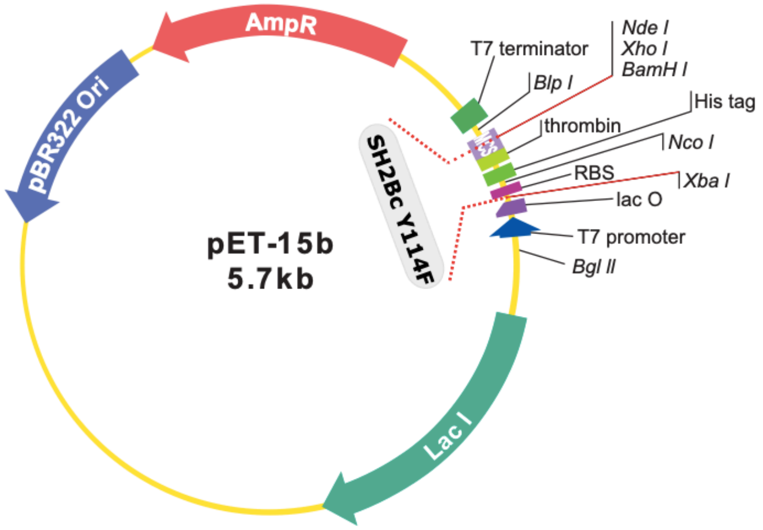
The gene segment encoding the SH2 domain with the C terminal tail of SH2B1 β (SH2c) was inserted into a linearised pET-2 plasmid, created by digestion with XbaI and BamHI, using the Gibson method. The complete DNA sequence of SH2c (pET-4, 6169 bp) was confirmed by Sanger sequencing from Eurofins Genomics (Figure 3.16). Similarly, a gene segment encoding the Y114F mutant SH2c protein was ordered from GenScript to be inserted into pET-15b, Figure 3.17.

His_6 tagged SH2c and SH2c Y114F proteins were overexpressed from plasmids pET-4 and pET-15b respectively using BL21 (DE3) cells, following the same procedure that was used to express SH2 protein described above. A culture of 1 Liter was grown at 25°C overnight after induction with 0.5 mM IPTG. The harvested cells were resuspended in lysis buffer, then cells were lysed using sonication and centrifugated at high speed. Soluble protein was screened by SDS-PAGE (Figure 3.18).



```
        10         20         30         40         50         60
KIHHHHHHDQ PLSGYPWFHG MLSRLKAAQL VLEGGTGSHG VFLVRQSETR RGEYVLTFNF

        70         80         90        100        110        120
QGKAKHLRLS LNEEGQCRVQ HLWFQSIFDM LEHFRVHPIP LESGGSSDVV LVSYVPSQRQ

       130        140        150        160
QGREQAGSHA GVCEGDRCYP DASSTLLPFG ASDCVTEHLP
```

Figure 3.16. Construct of pET-4 expression vector and amino acid sequence of SH2 with C terminus of SH2B1 β isoform. Top. Map of pET-4 plasmid with inserted SH2 with C terminus of mouse SH2B1 β gene (6169 bp). Bottom. The amino acid sequence of SH2c protein consists of: the N-terminal histidine tag in red, the sequence of SH2 domain with the C terminal tail of SH2B1β; the total His-SH2c sequence is 160 a.a. Cysteine residues at the tail are highlighted in light blue. The C-terminal end of SH2B1 β sequence is underlined. Tyrosine 114 and 139 residues are highlighted in yellow.



```
        10         20         30         40         50         60
SSHHHHHHDQ PLSGYPWFHG MLSRLKAAQL VLEGGTGSHG VFLVRQSETR RGEYVLTFNF

        70         80         90        100        110        120
QGKAKHLRLS LNEEGQCRVQ HLWFQSIFDM LEHFRVHPIP LESGGSSDVV LVSFVPSQRQ

       130        140        150        160
QGREQAGSHA GVCEGDRCYP DASSTLLPFG ASDCVTEHLP
```

Figure 3.17. Construct of pET-15b expression vector and amino acid sequence of mutant SH2 with C terminus Y114F. Top. Map of pET-15b plasmid (5.7 kb) where the mutant SH2c Y114F gene was inserted in between BamHI and XbaI sites. Bottom. The amino acid sequence of SH2c Y114F: N terminal tag in red; mutant residue yellow highlighted (Y114F). Total sequence of SH2c Y114F is 160 a.a.

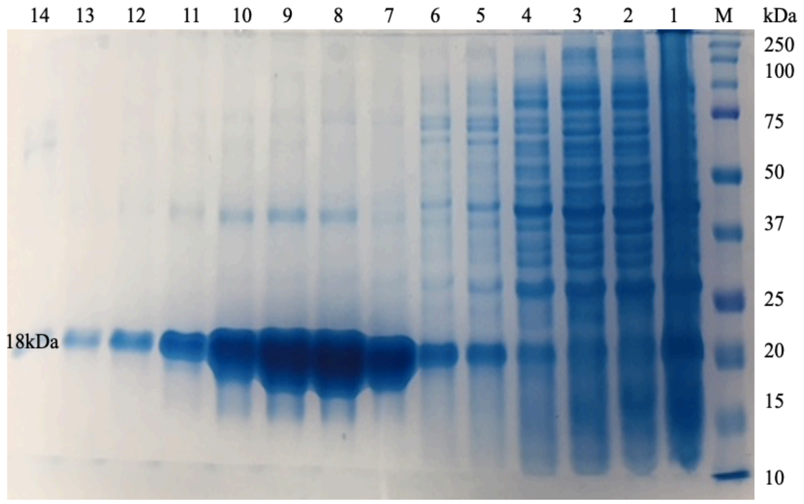


Figure 3.18. 16% SDS-PAGE gel for supernatant of His$_6$ SH2c protein expression and purification of fractions by Ni column. SH2c protein expression (expected 18 kDa) with 0.5 mM IPTG, at 25 °C overnight. The purification of His$_6$-SH2c proteins by Ni-NTA column: lane 1, molecular weight markers; lane 2, unbound materials; lanes 3, 4, 5, 6, washing fractions with 40 mM imidazole, and lanes 7, 8, 9, 10, 11, 12, 13, 14 elution fractions with buffer containing 300 mM imidazole.

3.3.2 Purification of SH2 with C terminus of SH2B1 β (SH2c) and mutant SH2c Y114F

The overexpressed SH2c and mutant SH2c Y114F were purified in two steps as before. First a 20 ml Ni-NTA column was used to purify the histidine tagged proteins (18 kDa) from 1 Liter culture. High concentrations of proteins were eluted with 50 mM Tris pH 8, 200 mM NaCl, and 300 mM imidazole. As shown in Figure 3.19 the purity of protein was estimated to be above 80%.

After that for the second purification round the elution fractions were concentrated to 10 ml and injected onto a Superde 200 10/300 gel filtration size column. The GF experiment was performed in 200 mM NaCl, 50 mM Tris pH 8, 2 mM DTT buffer at flow rate 1 ml/min using the ÄKTA prime plus system. As shown in Figure 3.19 on the GF trace a single peak was obtained containing the target protein (18 kDa), and SDS-PAGE indicates the purity of the protein was above 99%.

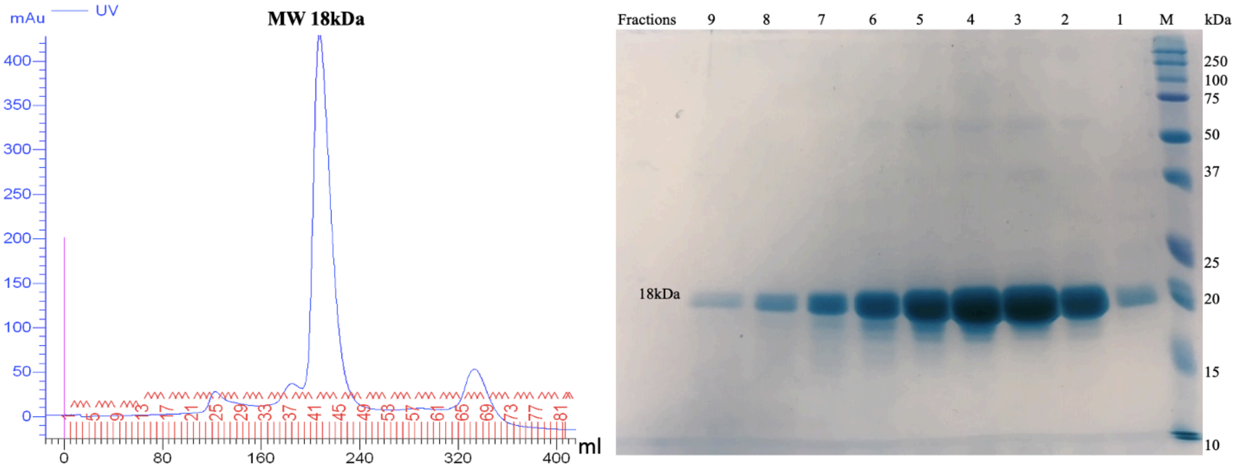


Figure 3.19. The gel filtration purification of the double labelled wild type and mutant SH2c protein. Left: The trace of gel filtration performed in 300 mM NaCl, Tris pH 8, 2 mM DTT. Right. The 16% SDS-PAGE gel of GF fractions: Marker (10-250 kDa), lanes 3, 4, 5, 6, 7, 8, 9 peak1 fractions histidine tagged SH2c 18 kDa (monomer).

3.4 Phosphorylation with activated Kinase

3.4.1 Phosphorylation of the wild type SH2c and mutant SH2c Y114F

In order to phosphorylate Tyr139 in the C terminal tail of SH2c protein, Group-based Prediction System 3.0 database was used to identify phosphorylation sites and define the corresponding kinase for a particular phosphorylation event (Xue et al., 2008). A number of kinases were predicted to phosphorylate Tyr139 at the C terminus of the protein, however Fer kinase showed a higher prediction score value (7.2) than other kinases meaning more probability that the residue is phosphorylated. Moreover although there were a number of phosphorylation sites at the C terminal tail of SH2c, Fer kinase was predicted to be specifically able to phosphorylate Tyr139 and no other potential sites. Fer kinase (ProQinase) was therefore used to carry out the phosphorylation in the reaction buffer which was recommended from the supplier, however the incubation time was optimised to obtain sufficient quantities of phosphorylated proteins.

The phosphorylation of Tyr139 in the C terminus of SH2c was done through incubating the Fer kinase with protein in ratio 4:100 at 30°C with shaking in reaction buffer; 70 mM HEPES-NaOH pH 7.5, 3 mM $MgCl_2$, 3 mM $MnCl_2$, 3 μM Na-orthovanadate, 1.2 mM DTT, 50 μg/ml $PEG_{20000,}$ 1% DMSO, 5 mM ATP, withdrawing samples at different time intervals (0, 2, 4, 6, and 8 hrs) to check the phosphorylation by mass spectrometry. During the incubation time, there was a clear precipitation after 6 hrs which rapidly increased after 8 hrs.

The mutant SH2c Y114F protein was phosphorylated with Fer kinase for 5 hrs using the same conditions as the wild type protein, adding 2 mM DTT every hour during the incubation time. A sample was taken every hour (0, 1, 2, 3, 4, and 5) to follow the phosphorylation by MS.

3.4.2 Monitoring the phosphorylation by Mass spectrometry

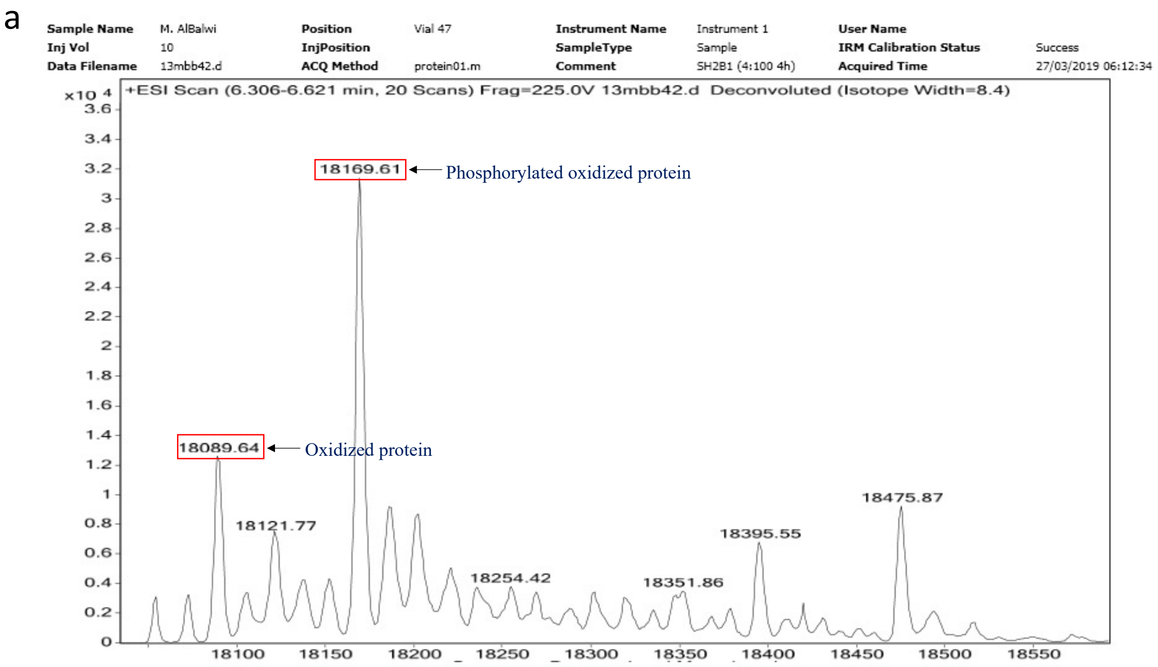
The protein phosphorylation was followed during the incubation time by analysing the samples through ESI MS (Electro-Spray Ionization Mass Spectrometry) for a quantitative analysis of the extent of phosphorylation and to achieve a complete or maximum protein phosphorylation. Moreover Electron Transfer Dissociation Mass Spectrometry (ETD MS) was used for identification of distinct phosphorylation sites after subjecting the protein to trypsin digestion and investigation of each peptide fragment. All samples analysed via ETD MS were extracted from 16% SDS gels that were corresponding to protein bands.

3.4.2.1 Phosphorylation of SH2c

ESI MS spectrum data of all analysed samples which were incubated at 30°C for 2 and 4 hrs showed two observed peaks (18090 Da and 18170 Da), Figure 3.20 (a), which mostly corresponded to wild type SH2c proteins (18092 Da). As known from the protein sequence there are three cysteine residues in the C terminal tail (133, 138 and 154) which can potentially be oxidised during the phosphorylation. Therefore the first peak at 18090 Da is expected to be oxidised protein as it has decreased in mass by 2 Da due to a disulphide bond formation. Although cysteine residues were reduced by adding 1.2 mM of DTT in the phosphorylation buffer, the presence of Mn^{+2} ions in the same buffer greatly decreased the stability of DTT (Charrier & Anastasio., 2012; Getz et al., 1999). Reducing the concentration of DTT in the reaction buffer caused cysteine residues to be oxidised rapidly and form a disulphide bond within the C terminal tail of SH2c protein.

In addition the second peak in the MS trace as shown in Figure 3.20 (a) corresponds to the phosphorylated oxidised SH2c protein (expected mass 18170 Da). The molecular weight of the protein increased by 80 Da due to the phosphorylation of one residue and at the same time lost 2 Da because of forming a disulphide bond between a pair of cysteine residues.

The phosphorylation of SH2c protein was followed in samples which were incubated at 30°C for 0, 2, and 4 hrs by ESI MS. The last two samples (6 and 8 hrs) could not be analysed as all of the protein precipitated. The phosphorylation reaction took several hours to go to completion. It is expected that the oxidisation of protein increased during the incubation time, which led to formation of a cross disulphide bond between two oxidised cysteines at the tail which may affect the stability of the protein and cause precipitation.



Sample Name	M. AlBalwi	Position	Vial 47	Instrument Name	Instrument 1	User Name	
Inj Vol	10	InjPosition		SampleType	Sample	IRM Calibration Status	Success
Data Filename	13mbb42.d	ACQ Method	protein01.m	Comment	SH2B1 (4:100 4h)	Acquired Time	27/03/2019 06:12:34

b

	Incubation time	% of sequence coverage	Peptide counts unique	Number of pY	pY site	Peptide sequence include phospho-tyrosine	Mass of pY peptide Da
Sample 1	0h	65.2	8	0	0	na	na
Sample 2	2h	65.2	7	1	114	VHPIPLESGGSSDVVLVS(Y)VPSQR	2601.284
Sample 3	4h	56.5	6	1	114	VHPIPLESGGSSDVVLVS(Y)VPSQR	2601.284
Sample 4	6h	56.5	7	1	114	VHPIPLESGGSSDVVLVS(Y)VPSQR	2601.284
Sample 5	8h	55.3	6	2	114	VHPIPLESGGSSDVVLVS(Y)VPSQR	2601.284
					139	C(Y)PDASSTLLPFGASDCVTEHLP	2616.091

c

```
        10         20         30         40         50         60
KIHHHHHHDQ PLSGYPWFHG MLSRLKAAQL VLEGGTGSHG VFLVRQSETR RGEYVLTFNF
        70         80         90        100        110        120
QGKAKHLRLS LNEEGQCRVQ HLWFQSIFDM LEHFRVHPIP LESGGSSDVV LVSYVPSQRQ
       130        140        150        160
QGREQAGSHA GVCEGDRCYP DASSTLLPFG ASDCVTEHLP
```

Figure 3.20. MS analysis of SH2c protein sample phosphorylated at 30°C with Fer kinase. a) ESI MS trace of intact phosphorylated SH2c protein after 4 hrs of incubation time at 30°C. b) Summary table of ETD MS results of the trypsin digested SH2c that was phosphorylated at different incubation times; 0, 2, 4, 6, and 8 hrs. c) The amino acid sequence of SH2c protein showing the possible cleavage sites (K and R) by Trypsin enzyme.

To provide molecular details about which residue was phosphorylated, all samples incubated at 30°C for 0, 2, 4, 6, and 8 hrs were checked by ETD MS after extracted from the SDS gel. Figure 3.20 (c) shows there are thirteen possible cleavage sites in the sequence of the SH2c protein, and thus around thirteen peptides expected to be identified. The result of ETD MS shows that between 6 and 8 peptides were identified in each sample, which means in total the achieved sequence coverage for SH2c protein was varying between 55 to 65% in samples, Figure 3.20 (b).

According to the ETD MS analysis, in all samples the peptides containing Tyr114 (V96-R120) and Tyr139 (C138-P160) were detected. Tyr114 in the unstructured part of SH2 domain was phosphorylated more rapidly than Tyr139 as identified in samples 2, 4 and 6 however Tyr139 was phosphorylated after 8 hrs, Figure 3.20 (b). This was a surprise, because Fer kinase was chosen because it is expected to phosphorylate Tyr139 much faster than Tyr114.

The formation of a disulphide bond within the C terminus of SH2c protein seems to happen faster than the phosphorylation. In the protein sequence, Tyr139 comes after Cys138, thus it is possible that the oxidation of of Cys138 early, forming a disulphide bond with another Cysteine residue (Cys 133 or 154), might hide Tyr139 and make it not easily accessible to be phosphorylated by Fer kinase.

It is surprising that Tyr114 at the end of the SH2 domain was the preferentially phosphorylated residue. There was considerable precipitation of the protein after phosphorylation of Tyr114. It is likely that pTyr114 may create a binding site for another SH2 domain, that would start early and increase during the incubation time leading it to form large molecular assemblies and cause protein aggregation.

3.4.2.2 Phosphorylation of SH2c Y114F

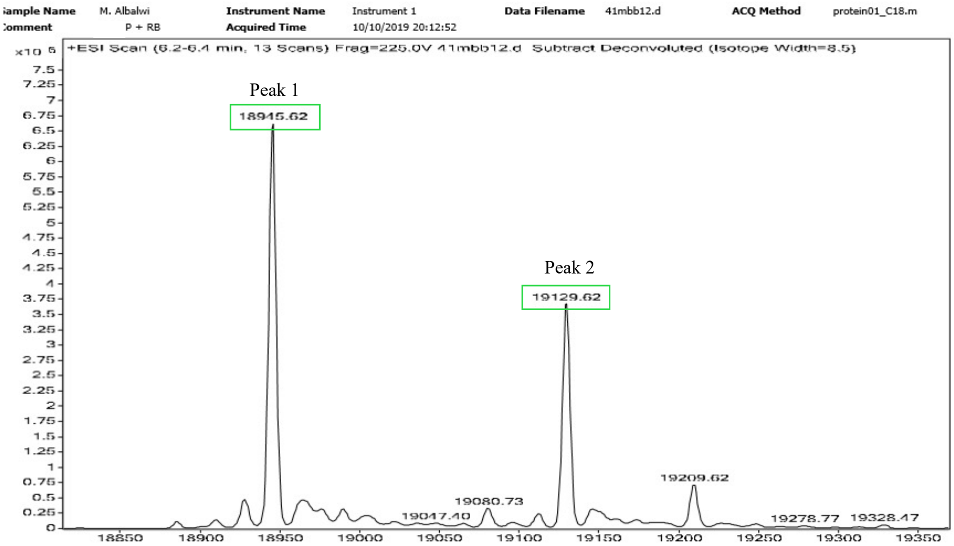
The problems with the unexpected phosphorylation of Tyr114, and the precipitation of the phosphorylated protein, led us to create the Y114F mutant which cannot be phosphorylated at residue 114. The phosphorylation of mutant SH2c Y114F protein at 30°C for 5 hrs (0, 1, 2, 3, 4, and 5) was monitored by ESI MS spectrometry.

In the MS spectrum of the 0 hr sample there were unexpectedly two peaks with different molecular weight (18946 Da and 19130 Da), Figure 3.21. The expected mass of the mutant protein is 17935 Da. The extra mass does not correspond to anything recognizable.

The two peaks were detected before starting the incubation and therefore presumably relate to the un-phosphorylated mutant protein. During the incubation time for 1, 2, 3, 4, and 5 hrs, phosphorylation of an amino acid was detectable in both peaks, as both signals increased their mass by 80 Da (18946 to 19026 Da, 19130 Da to19210 Da).

Moreover the level of protein phosphorylation compared with the non-phosphorylated protein in MS traces can be estimated from the relative intensities of the peaks. According to Figure 3.21, the level of phosphorylated protein increased gradually, while there was a significant decrease in the corresponding peaks for un-phosphorylated proteins. It can therefore be concluded that the phosphorylation was successful, even there is no explanation for the observed extra masses.

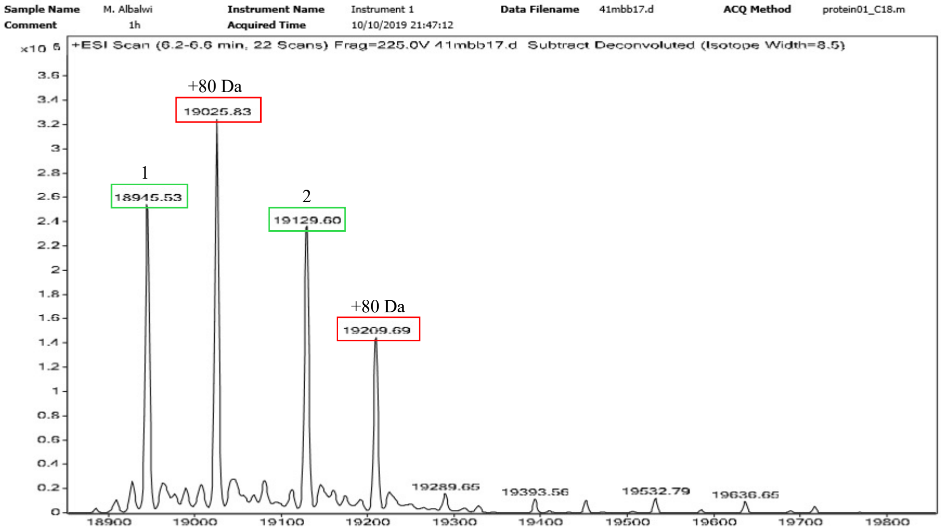
In summary, the samples analysed by MS showed that there was an increase in phosphorylation with time clearly visible once oxidation of the disulphide bonds had been prevented, by adding more DTT during the incubation time of the phosphorylation reaction. However the MW of observed peaks were different than the actual MW of SH2c Y114F protein (17935 Da) by 1011 and 1195 Da.



ample Name M. Albalwi Instrument Name Instrument 1 Data Filename 41mbb12.d ACQ Method protein01_C18.m

omment P + RB Acquired Time 10/10/2019 20:12:52

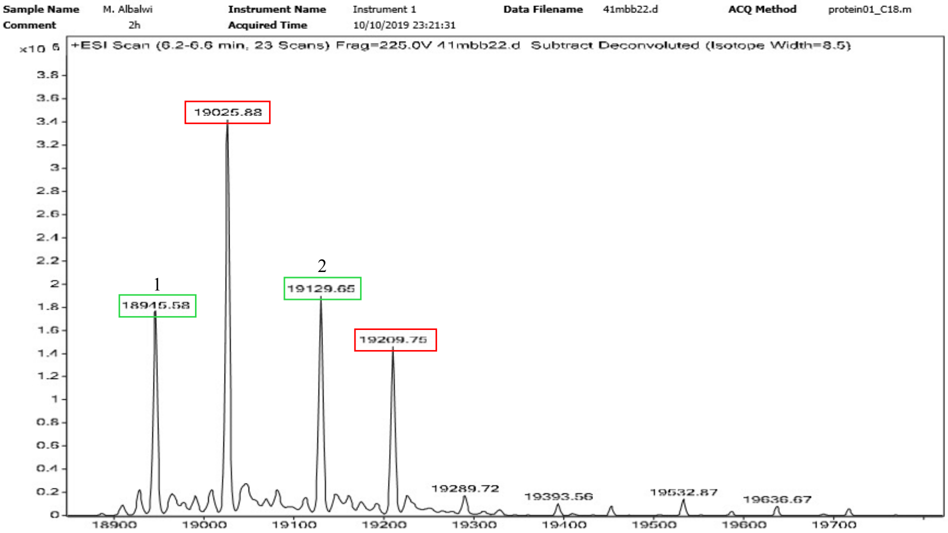
+ESI Scan (6.2-6.4 min, 13 Scans) Frag=225.0V 41mbb12.d Subtract Deconvoluted (Isotope Width=8.5)



Sample Name M. Albalwi Instrument Name Instrument 1 Data Filename 41mbb17.d ACQ Method protein01_C18.m

Comment 1h Acquired Time 10/10/2019 21:47:12

+ESI Scan (6.2-6.6 min, 22 Scans) Frag=225.0V 41mbb17.d Subtract Deconvoluted (Isotope Width=8.5)



Sample Name M. Albalwi Instrument Name Instrument 1 Data Filename 41mbb22.d ACQ Method protein01_C18.m

Comment 2h Acquired Time 10/10/2019 23:21:31

+ESI Scan (6.2-6.6 min, 23 Scans) Frag=225.0V 41mbb22.d Subtract Deconvoluted (Isotope Width=8.5)

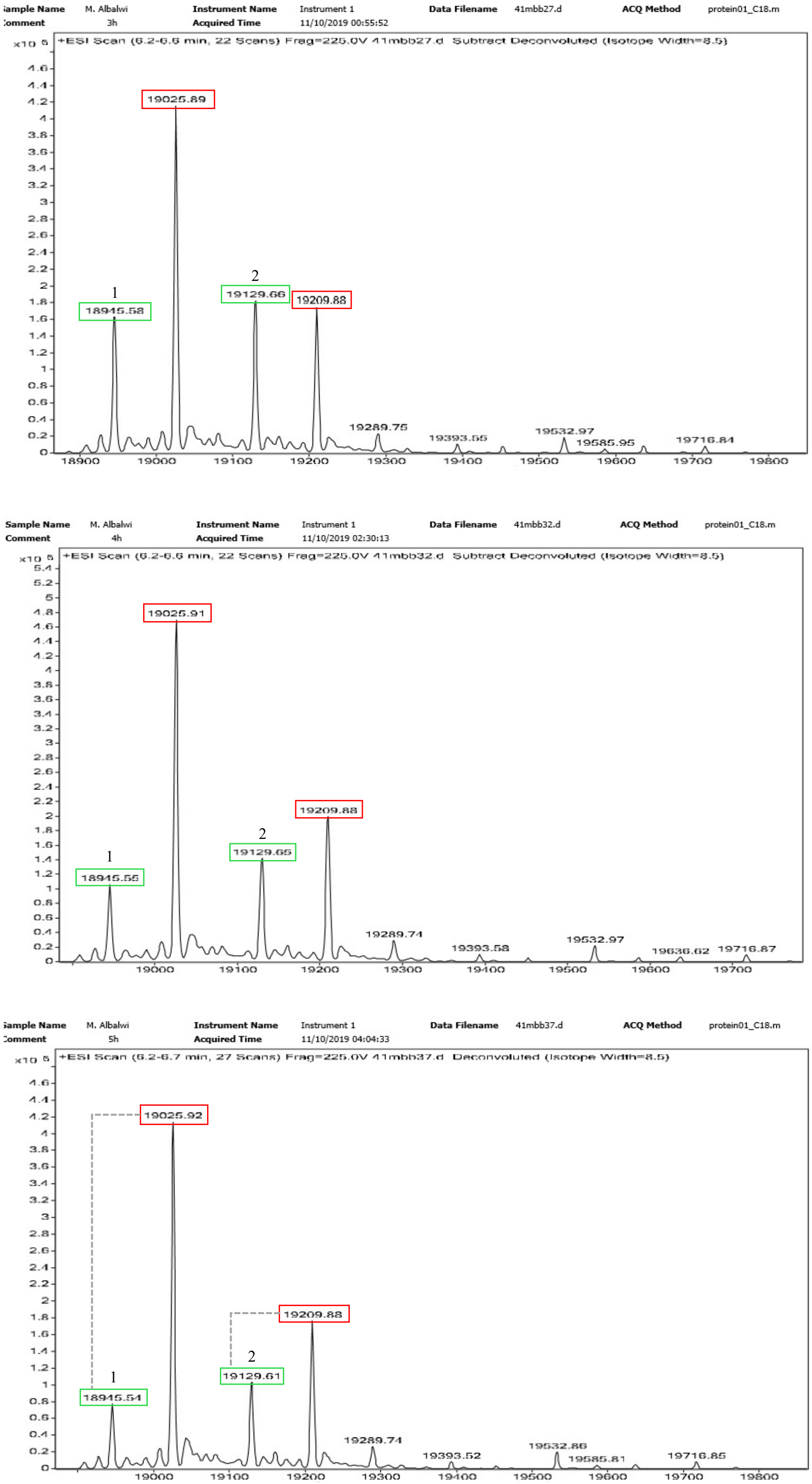


Figure 3.21. ESI MS spectrum analysis of phosphorylated SH2c Y114F protein samples at 30°C for 0, 1, 2, 3, 4, and 5 hrs. In the MS trace the un-phosphorylated proteins are denoted by green boxes, while the phosphorylated proteins are surrounded by red boxes. Peaks 1 and 2 refer to the detectable peaks for the un-phosphorylated proteins. MS traces from the top to the bottom represent the result coming from the same protein sample at different incubation times 0, 1, 2, 3, 4, and 5 hrs, respectively.

3.4.3 Checking the phosphorylation by ^{15}N HSQC NMR

3.4.3.1 Phosphorylation of SH2c Y114F

2D ^{15}N HSQC spectra were used for identification of distinct phosphorylation sites in the mutant protein. The double labelled SH2c Y114F protein was phosphorylated using a ratio of 4:100 kinase to protein for 6 hrs by Fer kinase as before. A high concentration of protein was used for the phosphorylation (1.5 mM) to overcome the loss of protein during the phosphorylation, and thus to end with good quantity of phosphorylated protein for NMR. Most of the protein was lost during the phosphorylation and buffer exchange, and 0.06 mM of protein was obtained. The loss may be due to dimerization of protein as a result of forming cross disulfide bonds, or after the phosphorylation of the tyrosine at the tail.

^{15}N HSQC spectra of phosphorylated and un-phosphorylated SH2c Y114F proteins were assigned in 100 mM potassium phosphate pH 6, 2 mM DTT, 10% D_2O, 2 mM TSP. The weighted average chemical shift change (^{15}N and ^{1}H) of the phosphorylated and un-phosphorylated protein was calculated using equation 6.1 to identify the phosphorylation of Tyr139 by changes in its chemical shift.

According to Figure 6.3 (a) in chapter 6 the weighted average shift change in the chemical shift of Tyr139 backbone NH was approximately $\Delta\delta = 0.1$ ppm and slightly less to neighbouring residues (137, 141, 142, and 143), upon phosphorylation of Y139 by Fer kinase. As explained in Section 6.1.2 the observed chemical shift changes verified the phosphorylation of Tyr139, but not the binding to the phospho-tyrosine pocket in the SH2 domain. Huang et al., 2020, and Bienkiewicz et al., 1999 indicate that the phosphorylation of Tyr induces a small downfield shift for the backbone amide. The shift change is small because of the distal position of the hydroxyl group (OHη) of Tyr which binds covalently to the phosphate (PO_4^{3-}) group, thus does not have a big impact on the backbone amide shift.

In summary, the SH2 and the SH2c protein were produced and purified, and the SH2c was phosphorylated using Fer kinase. Unexpectedly, phosphorylation was faster at Tyr114 than at

Tyr139, even though Fer kinase was selected because it should phosphorylate faster at Tyr139. For NMR analysis, it is preferred to use a protein that was free from problems of phosphorylation at Tyr114, and so the SH2c Y114F mutant was prepared. Phosphorylation of SH2c and SH2c Y114F both produced large amounts of precipitation. In part this is thought to be due to oxidation of the cysteines and formation of intermolecular disulphide bridges. This oxidation could be slowed by addition of extra DTT, but was nor prevented. It may also be due to the phosphorylation of Tyr114 producing a recognition site for a second SH2 domain, leading to formation of dimers and further aggregates.

Chapter 4 Chemical shift assignment of SH2 protein

This chapter discusses two methods of NMR signal assignment: a manual interactive method and an automatic computational method (CYANA). Chemical shift assignment is the process of assigning the signals observed in a number of NMR spectra to specific nuclei in the protein sequence. Chemical shift assignment is useful for at least three important things. It provides structural information in the form of dihedral angle restraints, for example using the program TALOS-N. It is the essential first step in protein NMR structure determination, because it provides the basis for the assignment of NOEs to generate distance restraints. It is also essential for interpretation of chemical shift perturbation mapping. For all of these, it is important that the assignment should be as complete and as accurate as possible. One of the questions addressed in this chapter is therefore which method provides a more complete and accurate assignment. For the automated assignment, the program CYANA was used (Güntert, 2004; Herrmann et al., 2002). This program was written by Peter Güntert, mostly while working in the group of Prof Wüthrich, and is one of the most popular methods for automated assignment. It feeds into the FLYA program, which is the leading program for automated resonance assignment.

This chapter describes how the chemical shift assignments were used to produce a dihedral angle restraint file. It also describes the generation of another set of structure restraints, namely hydrogen bonds. As discussed in the next chapter, hydrogen bonds are important restraints for improving the rigidity of NMR structures. There are two main ways of generating hydrogen bond lists from NMR data. The most secure method uses measurements of the *J* coupling across hydrogen bonds, known as $^{3h}J_{NC'}$ (Cornilescu et al., 1999). The coupling constant is typically of the order of 1 Hz, which means that for most proteins, transverse relaxation is sufficiently efficient that signals from this coupling are too small to be seen. The other way of generating hydrogen bond lists is to look at predicted regular secondary structure, and add restraints by hand to ensure typical hydrogen bonding patterns, which need to be supported by NOEs. This is generally seen as a justifiable procedure, even if not fully evidenced by experimental data. In this chapter, the use of amide proton temperature coefficients as a way of producing hydrogen bond restraints was explored. The defined donor hydrogen bond residue data were used as supplemental information for subsequent structure calculation.

All spectral data for SH2 assignment and structure calculation were collected from a 1 mM double labelled SH2 protein sample in 100 mM potassium phosphate buffer in 10% D_2O or 100% D_2O at pH 6. The NMR experiments were performed on a Bruker 800 MHz spectrometer at 298 K. The spectral data was transferred to LINUX computers and processed with Felix2007 for analysis.

4.1 Manual resonance assignment

4.1.1 Sequential assignment of backbone resonances

The backbone resonances of ^{1}HN, ^{15}N, ^{13}C′, ^{13}Cα, Hα and ^{13}Cβ of SH2 protein were assigned using the ^{15}N HSQC spectrum and 3D standard triple resonance spectra (HNCO, HNCACO, HNCA, CBCACONH, and HNCACB with HBHACONH). The sets of nuclei correlated in these spectra, and acquisition parameters of each backbone spectrum are shown in table 2.10.

The ^{1}H, ^{15}N HSQC spectrum is an NMR experiment that shows ^{1}H-^{15}N correlations for directly bonded nuclei. It should therefore contain one peak for each backbone amide (except for the N terminus which is an amine, and prolines which have no amide proton), plus peaks for other sidechain NH that exchange slowly enough, ie tryptophan, glutamine, asparagine, and arginine Hε. The auto-peak picking function in Felix2007 detected 106 out of 118 backbone amide signals of His_6 tagged SH2 protein (excluding sidechain amide signals). There are 5 prolines in SH2, so one would expect 112 signals, with some of the N-terminal His-tag signals overlapped.

Two main methods were used for the signal assignment. The first was to set up the Felix program so that it displays a set of backbone spectra based on the same amide H and N frequencies, as shown in Figure 4.1. This makes it easy to identify likely sequential pairs manually, with the cursors linked between windows. Different spin systems can be loaded manually for comparison. The group of backbone nuclei associated with a certain amide is called a spin system, therefore the amide number in ^{15}N HSQC is the spin system number, table 4.1. The second was the use of the Asstools program (Reed et al., 2003), which is based on a Monte-Carlo simulated annealing method as described in detail in section 2.9.4. Following Asstools run, sequential matches were checked manually. At that point, spin systems can be deleted if they are clearly not backbone resonances or are a minor species; assignments can be removed or changed if in error (eg an incorrect pCα/Cα assignment); and sequence-specific assignments can be fixed within the program. Asstools is then run again until a self-consistent set of assignments is produced. Figure 4.1 shows a typical example of a set of triple resonance spectra. The final assignment list is shown in Table 4.1.

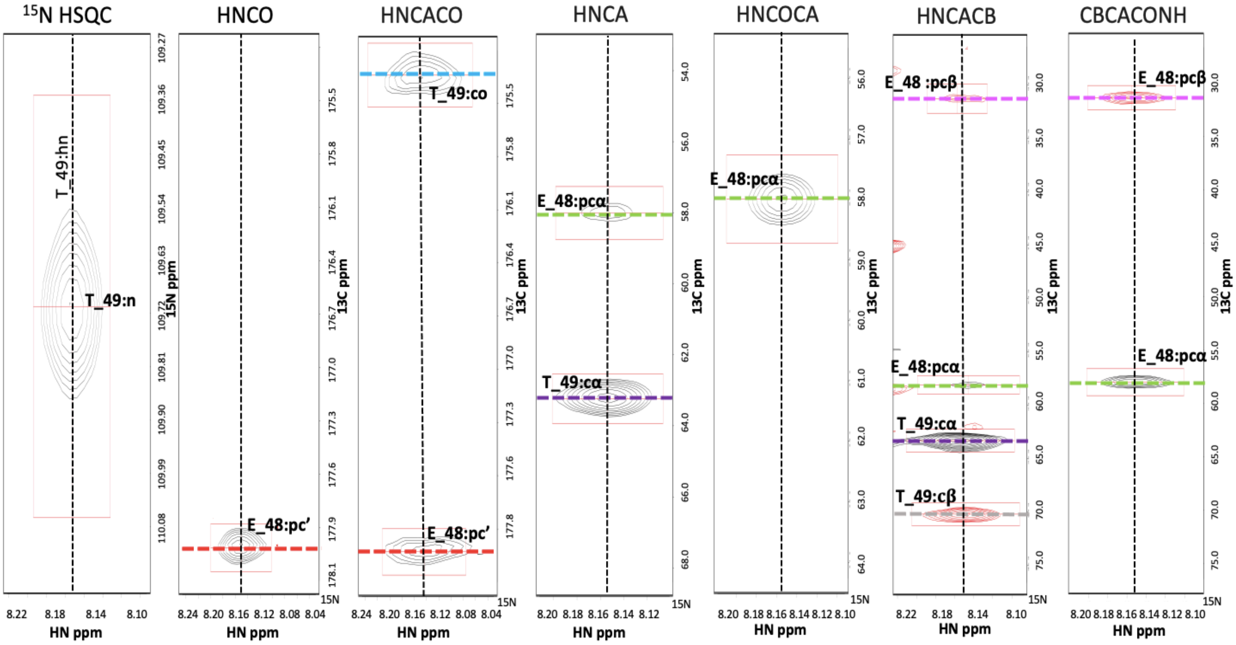


Figure 4.1. Representative slices of triple resonance backbone spectra as displayed in Felix2007 showing the backbone spin system of residue T49 (HN, N, C′, Cα, Cβ) and pC′, pCα, pCβ of preceding residue E48. NH peak of T49 in ^{15}N HSQC spectra was aligned with others triple resonances spectra to assign the chemical shift of its associated Co, Cα, Cβ nuclei. Each slice is labelled with spectrum name on the top, and signals in different spectra which are assigned to the same nucleus are labelled with the same dashed line colour. 15N: ^{15}N dimension (ppm), HN: ^{1}H dimension (ppm), 13C: ^{13}C dimension (ppm).

In the ^{15}N HSQC, 103 of the 106 amide signals were assigned straightforwardly to their backbone resonances. The remaining three spin systems were assigned to Trp17, Gln61 and His81. The amide signals for these three were visible with low intensity in ^{15}N HSQC spectra but their corresponding backbone resonances were not identified in the triple resonance spectra. However, the corresponding backbone resonances of Trp17, Gln61 and His81 residues were identified in the spin system of the following residue and confirmed later on in the sidechain assignment, as shown in Figure 4.2. The amide proton of Trp17 has an unusually high field chemical shift (5.57 ppm, Table 4.1). NMR spectra are typically processed with a convolution difference window function, which removes the water signal very effectively (Waltho and Cavanagh, 1993). It also reduces the intensity of signals that are close in chemical shift to the water resonance, and it is likely that the Trp17 signal has been affected in this way, explaining its low intensity.

With these three added in, the final SH2 polypeptide backbone resonance assignment was complete except for the amide resonance of Gly20 and the N terminal linker (Lys^1-Ile^2). It is common for the signals from residue 2 in a protein to be weak or missing. One of the His_6 tag residues was missing, which could be due to overlap with another His-tag signal or possibly to intermediate chemical exchange or fast exchange with solvent (Englander et al., 1992).

All amide sidechain peaks in ^{15}N HSQC were excluded from the backbone assignment. The sidechain signals of Asn (Nδ-Hδ2) and Gln (Nε-Hε2) are characteristic because they form pairs with different proton chemical shifts and the same nitrogen chemical shifts (cross-peak), and were not identified in the triple resonance spectra. Also, Nε-Hε groups of Arg were defined because the Nε chemical shift is outside the spectral region and is therefore aliased, and identified by changing the spectral width (SW) in ^{15}N HSQC spectra. Nε-Hε sidechain peaks of Trp17 and Trp83 appeared in the bottom left corner of ^{15}N HSQC spectra around 10 ppm, as shown in Figure 4.2.

The final sequential assignment of SH2 backbone resonances showed that in total 95% of backbone resonances for non-proline residues 9 to 118 were assigned (96% of ^{1}H, 96% of ^{15}N, 96% of Cα, 94% of Cβ, 92 % of C'), the missing signals being due to Trp17, Gly20, Gln61, and His81 (Table 4.1).

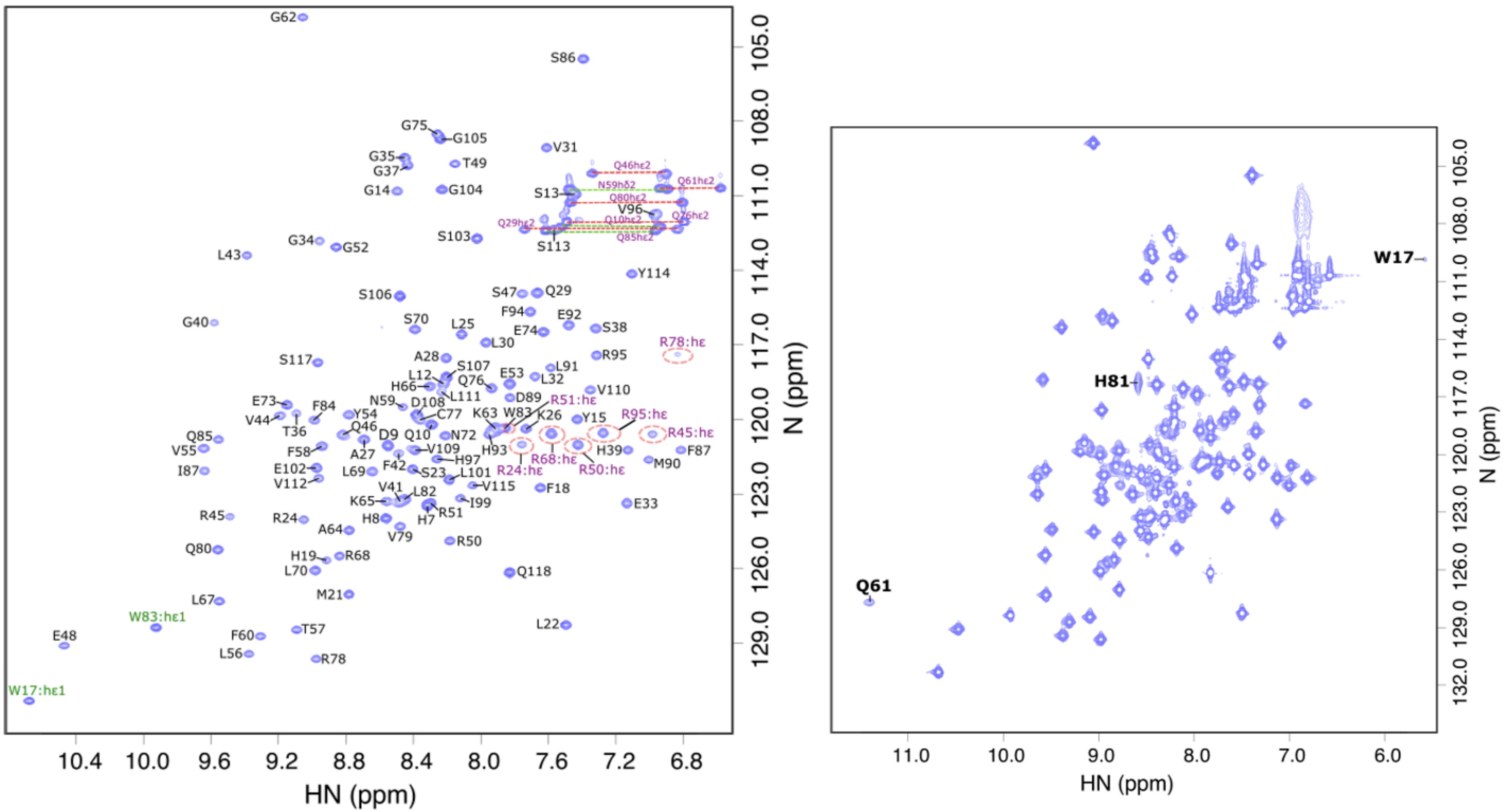
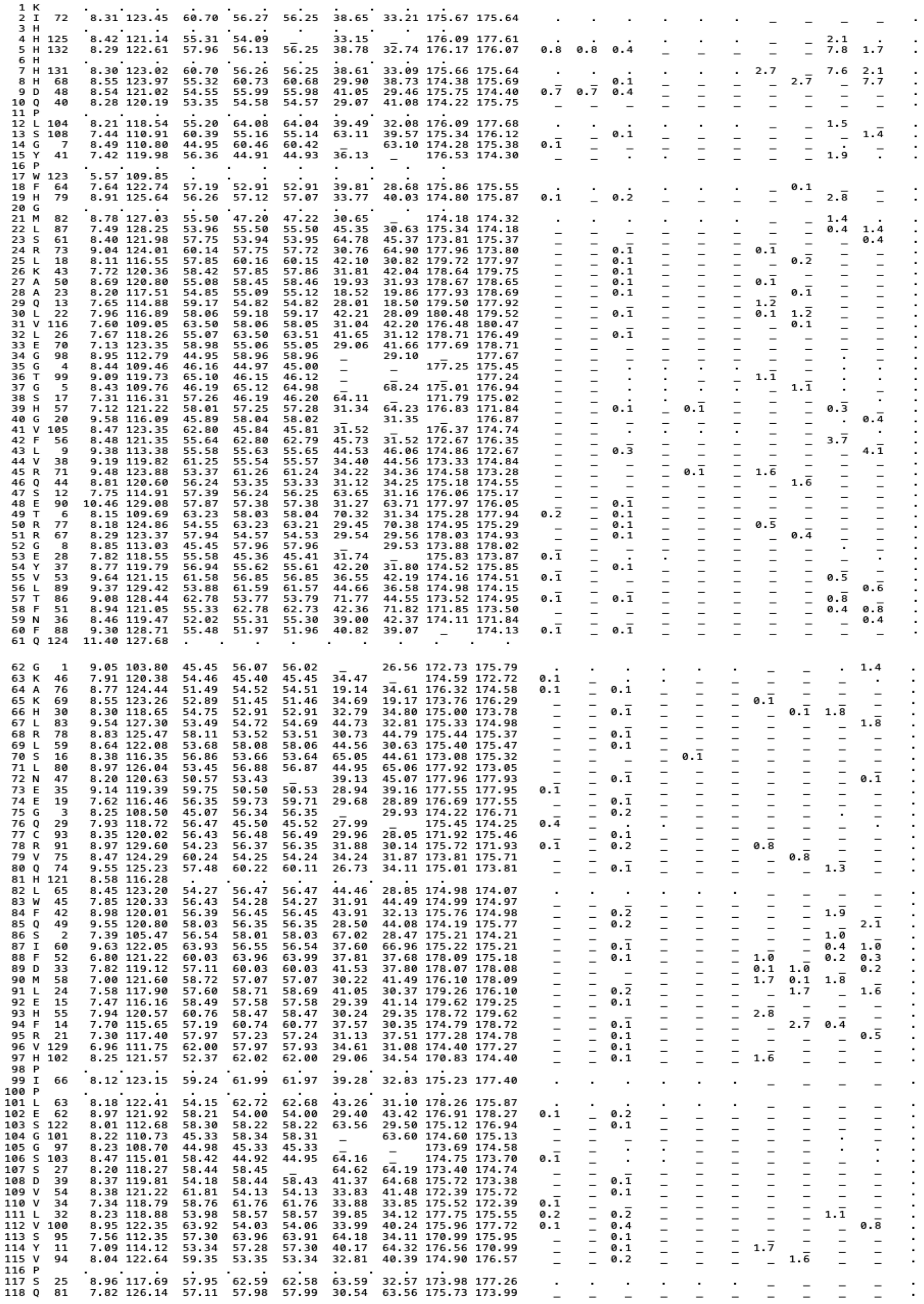


Figure 4.2. ^{1}H,^{15}N HSQC spectrum of SH2 (double labelled sample) recorded at pH 6 and 298 K. Right, ^{1}H ^{15}N HSQC shows the low intensity peaks (W17, Q61, H81). Left, ^{1}H ^{15}N HSQC shows the backbone resonances of SH2 indicated by sequence number and residue name. Red circle peaks and green labelled peaks illustrate the sidechain NHε of Arg and NHε of Trp, respectively, which were assigned by NOESY spectra. The top right corner contains the sidechain amide (Nε-Hε2 of Gln, Nδ-Hδ2 of Asn) peaks (cross-peaks) which are labelled based on NOESY assignment, and other nearby amide backbone peaks. N: ^{15}N dimension (ppm), H: ^{1}H dimension (ppm).

Table 4.1. Output spin systems of SH2 protein from the Asstools backbone assignment program. The first column shows the residue number, the second column is the residue name, the third column is the spin system number, the following lines are spin system nuclei (HN, N, cα, pcα, pcα, cβ, pcβ, c′, pc′), respectively, where HN, N, cα, cβ, c′ are nuclei of the same residue and pcα, pcβ, pc′ nuclei of the preceding residue.



```
  1 K        .      .      .      .      .      .      .      .      .      .      .
  2 I  72   8.31 123.45  60.70  56.27  56.25  38.65  33.21 175.67 175.64    .    .    .    .    .    .    _    _    _    _    .
  3 H        .      .      .      .      .      .      .      .      .      .      .
  4 H 125   8.42 121.14  55.31  54.09      _  33.15      _ 176.09 177.61    .    .    .    .    .    .    _    _  2.1    .    .
  5 H 132   8.29 122.61  57.96  56.13  56.25  38.78  32.74 176.17 176.07  0.8  0.8  0.4    _    _    _    _    _  7.8  1.7    .
  6 H        .      .      .      .      .      .      .      .      .      .      .
  7 H 131   8.30 123.02  60.70  56.26  56.25  38.61  33.09 175.66 175.64    .    .    .    .    .    .  2.7    _  7.6  2.1    .
  8 H  68   8.55 123.97  55.32  60.73  60.68  29.90  38.73 174.38 175.69    _    _  0.1    _    _    _    _  2.7    _  7.7    .
  9 D  48   8.54 121.02  54.55  55.99  55.98  41.05  29.46 175.75 174.40  0.7  0.7  0.4    _    _    _    _    _    _    _    .
 10 Q  40   8.28 120.19  53.35  54.58  54.57  29.07  41.08 174.22 175.75    _    _    _    _    _    _    _    _    _    _    .
 11 P        .      .      .      .      .      .      .      .      .      .      .
 12 L 104   8.21 118.54  55.20  64.08  64.04  39.49  32.08 176.09 177.68    .    .    .    .    .    .    _    _  1.5    _    .
 13 S 108   7.44 110.91  60.39  55.16  55.14  63.11  39.57 175.34 176.12    _    _  0.1    _    _    _    _    _    _  1.4    .
 14 G   7   8.49 110.80  44.95  60.46  60.42      _  63.10 174.28 175.38  0.1    _    _    _    _    _    _    _    .    _    .
 15 Y  41   7.42 119.98  56.36  44.91  44.93  36.13      _ 176.53 174.30    _    _    .    .    _    _    _    _  1.9    .    .
 16 P        .      .      .      .      .      .      .      .      .      .      .
 17 W 123   5.57 109.85       .      .      .      .      .      .      .      .
 18 F  64   7.64 122.74  57.19  52.91  52.91  39.81  28.68 175.86 175.55    .    .    .    .    .    .    _  0.1    _    _    .
 19 H  79   8.91 125.64  56.26  57.12  57.07  33.77  40.03 174.80 175.87  0.1    _  0.2    _    _    _    _    _  2.8    _    .
 20 G        .      .      .      .      .      .      .      .      .      .      .
 21 M  82   8.78 127.03  55.50  47.20  47.22  30.65      _ 174.18 174.32    .    .    .    .    .    .    _    _  1.4    .    .
 22 L  87   7.49 128.25  53.96  55.50  55.50  45.35  30.63 175.34 174.18    _    _    _    _    _    _    _    _  0.4  1.4    .
 23 S  61   8.40 121.98  57.75  53.94  53.95  64.78  45.37 173.81 175.37    _    _    _    _    _    _    _    _    _  0.4    .
 24 R  73   9.04 124.01  60.14  57.75  57.72  30.76  64.90 177.96 173.80    _    _  0.1    _    _    _  0.1    _    _    _    .
 25 L  18   8.11 116.55  57.85  60.16  60.15  42.10  30.82 179.72 177.97    _    _  0.1    _    _    _    _  0.2    _    _    .
 26 K  43   7.72 120.36  58.42  57.85  57.86  31.81  42.04 178.64 179.75    _    _  0.1    _    _    _    _    _    _    _    .
 27 A  50   8.69 120.80  55.08  58.45  58.46  19.93  31.93 178.67 178.65    _    _  0.1    _    _    _  0.1    _    _    _    .
 28 A  23   8.20 117.51  54.85  55.09  55.12  18.52  19.86 177.93 178.69    _    _  0.1    _    _    _    _  0.1    _    _    .
 29 Q  13   7.65 114.88  59.17  54.82  54.82  28.01  18.50 179.50 177.92    _    _    _    _    _    _  1.2    _    _    _    .
 30 L  22   7.96 116.89  58.06  59.18  59.17  42.21  28.09 180.48 179.52    _    _  0.1    _    _    _  0.1  1.2    _    _    .
 31 V 116   7.60 109.05  63.50  58.06  58.05  31.04  42.20 176.48 180.47    _    _    _    _    _    _    _  0.1    _    _    .
 32 L  26   7.67 118.26  55.07  63.50  63.51  41.65  31.12 178.71 176.49    _    _  0.1    _    _    _    _    _    _    _    .
 33 E  70   7.13 123.35  58.98  55.06  55.05  29.06  41.66 177.69 178.71    _    _    _    _    _    _    _    _    _    _    .
 34 G  98   8.95 112.79  44.95  58.96  58.96      _  29.10      _ 177.67    _    _    _    _    _    _    _    _    .    _    .
 35 G   4   8.44 109.46  46.16  44.97  45.00      _      _ 177.25 175.45    _    _    .    .    .    .    _    _    .    .    .
 36 T  99   9.09 119.73  65.10  46.15  46.12      _      _      _ 177.24    _    _    .    .    _    _  1.1    _    .    .    .
 37 G   5   8.43 109.76  46.19  65.12  64.98      _  68.24 175.01 176.94    _    _    .    .    .    .    _  1.1    .    _    .
 38 S  17   7.31 116.31  57.26  46.19  46.20  64.11      _ 171.79 175.02    _    _    .    .    _    _    _    _    _    .    .
 39 H  57   7.12 121.22  58.01  57.25  57.28  31.34  64.23 176.83 171.84    _    _  0.1    _  0.1    _    _    _  0.3    _    .
 40 G  20   9.58 116.09  45.89  58.04  58.02      _  31.35      _ 176.87    _    _    _    _    _    _    _    _    .  0.4    .
 41 V 105   8.47 123.35  62.80  45.84  45.81  31.52      _ 176.37 174.74    _    _    .    .    .    .    _    _    _    .    .
 42 F  56   8.48 121.35  55.64  62.80  62.79  45.73  31.52 172.67 176.35    _    _    _    _    _    _    _    _  3.7    _    .
 43 L   9   9.38 113.38  55.58  55.63  55.65  44.53  46.06 174.86 172.67    _    _  0.3    _    _    _    _    _    _  4.1    .
 44 V  38   9.19 119.82  61.25  55.54  55.57  34.40  44.56 173.33 174.84    _    _    _    _    _    _    _    _    _    _    .
 45 R  71   9.48 123.88  53.37  61.26  61.24  34.22  34.36 174.58 173.28    _    _    _    _  0.1    _  1.6    _    _    _    .
 46 Q  44   8.81 120.60  56.24  53.35  53.33  31.12  34.25 175.18 174.55    _    _    _    _    _    _    _  1.6    _    _    .
 47 S  12   7.75 114.91  57.39  56.24  56.25  63.65  31.16 176.06 175.17    _    _    _    _    _    _    _    _    _    _    .
 48 E  90  10.46 129.08  57.87  57.38  57.38  31.27  63.71 177.97 176.05    _    _  0.1    _    _    _    _    _    _    _    .
 49 T   6   8.15 109.69  63.23  58.03  58.04  70.32  31.34 175.28 177.94  0.2    _  0.1    _    _    _    _    _    _    _    .
 50 R  77   8.18 124.86  54.55  63.23  63.21  29.45  70.38 174.95 175.29    _    _  0.1    _    _    _  0.5    _    _    _    .
 51 R  67   8.29 123.37  57.94  54.57  54.53  29.54  29.56 178.03 174.93    _    _  0.1    _    _    _    _  0.4    _    _    .
 52 G   8   8.85 113.03  45.45  57.96  57.96      _  29.53 173.88 178.02    _    _    _    _    _    _    _    _    .    _    .
 53 E  28   7.82 118.55  55.58  45.36  45.41  31.74      _ 175.83 173.87  0.1    _    .    .    _    _    _    _    _    .    .
 54 Y  37   8.77 119.79  56.94  55.62  55.61  42.20  31.80 174.52 175.85    _    _  0.1    _    _    _    _    _    _    _    .
 55 V  53   9.64 121.15  61.58  56.85  56.85  36.55  42.19 174.16 174.51  0.1    _    _    _    _    _    _    _  0.5    _    .
 56 L  89   9.37 129.42  53.88  61.59  61.57  44.66  36.58 174.98 174.15    _    _    _    _    _    _    _    _    _  0.6    .
 57 T  86   9.08 128.44  62.78  53.77  53.79  71.77  44.55 173.52 174.95  0.1    _  0.1    _    _    _    _    _  0.8    _    .
 58 F  51   8.94 121.05  55.33  62.78  62.73  42.36  71.82 171.85 173.50    _    _    _    _    _    _    _    _  0.4  0.8    .
 59 N  36   8.46 119.47  52.02  55.31  55.30  39.00  42.37 174.11 171.84    _    _    _    _    _    _    _    _    _  0.4    .
 60 F  88   9.30 128.71  55.48  51.97  51.96  40.82  39.07      _ 174.13  0.1    _  0.1    _    _    _    _    _    _    _    .
 61 Q 124  11.40 127.68       .      .      .      .      .      .      .      .      .      .

 62 G   1   9.05 103.80  45.45  56.07  56.02      _  26.56 172.73 175.79    .    .    .    .    .    .    _    _    .  1.4    .
 63 K  46   7.91 120.38  54.46  45.40  45.45  34.47      _ 174.59 172.72  0.1    _    .    .    _    _    _    _    _    .    .
 64 A  76   8.77 124.44  51.49  54.52  54.51  19.14  34.61 176.32 174.58  0.1    _  0.1    _    _    _    _    _    _    _    .
 65 K  69   8.55 123.26  52.89  51.45  51.46  34.69  19.17 173.76 176.29    _    _    _    _    _    _  0.1    _    _    _    .
 66 H  30   8.30 118.65  54.75  52.91  52.91  32.79  34.80 175.00 173.78    _    _  0.1    _    _    _    _  0.1  1.8    _    .
 67 L  83   9.54 127.30  53.49  54.72  54.69  44.73  32.81 175.33 174.98    _    _    _    _    _    _    _    _    _  1.8    .
 68 R  78   8.83 125.47  58.11  53.52  53.51  30.73  44.79 175.44 175.37    _    _  0.1    _    _    _    _    _    _    _    .
 69 L  59   8.64 122.08  53.68  58.08  58.06  44.56  30.63 175.40 175.47    _    _  0.1    _    _    _    _    _    _    _    .
 70 S  16   8.38 116.35  56.86  53.66  53.64  65.05  44.61 173.08 175.32    _    _    _    _  0.1    _    _    _    _    _    .
 71 L  80   8.97 126.04  53.45  56.88  56.87  44.95  65.06 177.92 173.05    _    _    _    _    _    _    _    _    _    _    .
 72 N  47   8.20 120.63  50.57  53.43      _  39.13  45.07 177.96 177.93    _    _  0.1    _    _    _    _    _    _  0.1    .
 73 E  35   9.14 119.39  59.75  50.50  50.53  28.94  39.16 177.55 177.95  0.1    _    _    _    _    _    _    _    _    _    .
 74 E  19   7.62 116.46  56.35  59.73  59.71  29.68  28.89 176.69 177.55    _    _  0.1    _    _    _    _    _    _    _    .
 75 G   3   8.25 108.50  45.07  56.34  56.35      _  29.93 174.22 176.71    _    _  0.2    _    _    _    _    _    .    _    .
 76 Q  29   7.93 118.72  56.47  45.50  45.52  27.99      _ 175.45 174.25  0.4    _    .    .    _    _    _    _    _    .    .
 77 C  93   8.35 120.02  56.43  56.48  56.49  29.96  28.05 171.92 175.46    _    _  0.1    _    _    _    _    _    _    _    .
 78 R  91   8.97 129.60  54.23  56.37  56.35  31.88  30.14 175.72 171.93  0.1    _  0.2    _    _    _  0.8    _    _    _    .
 79 V  75   8.47 124.29  60.24  54.25  54.24  34.24  31.87 173.81 175.71    _    _    _    _    _    _    _  0.8    _    _    .
 80 Q  74   9.55 125.23  57.48  60.22  60.11  26.73  34.11 175.01 173.81    _    _  0.1    _    _    _    _    _  1.3    _    .
 81 H 121   8.58 116.28       .      .      .      .      .      .      .      .
 82 L  65   8.45 123.20  54.27  56.47  56.47  44.46  28.85 174.98 174.07    .    .    .    .    .    .    _    _    _    _    .
 83 W  45   7.85 120.33  56.43  54.28  54.27  31.91  44.49 174.99 174.97    _    _    _    _    _    _    _    _    _    _    .
 84 F  42   8.98 120.01  56.39  56.45  56.45  43.91  32.13 175.76 174.98    _    _  0.2    _    _    _    _    _  1.9    _    .
 85 Q  49   9.55 120.80  58.03  56.35  56.35  28.50  44.08 174.19 175.77    _    _  0.2    _    _    _    _    _    _  2.1    .
 86 S   2   7.39 105.47  56.54  58.01  58.03  67.02  28.47 175.21 174.21    _    _    _    _    _    _    _    _  1.0    _    .
 87 I  60   9.63 122.05  63.93  56.55  56.54  37.60  66.96 175.22 175.21    _    _  0.1    _    _    _    _    _  0.4  1.0    .
 88 F  52   6.80 121.22  60.03  63.96  63.99  37.81  37.68 178.09 175.18    _    _  0.1    _    _    _  1.0    _  0.2  0.3    .
 89 D  33   7.82 119.12  57.11  60.03  60.03  41.53  37.80 178.07 178.08    _    _    _    _    _    _  0.1  1.0    _  0.2    .
 90 M  58   7.00 121.60  58.72  57.07  57.07  30.22  41.49 176.10 178.09    _    _    _    _    _    _  1.7  0.1  1.8    _    .
 91 L  24   7.58 117.90  57.60  58.71  58.69  41.05  30.37 179.26 176.10    _    _  0.2    _    _    _    _  1.7    _  1.6    .
 92 E  15   7.47 116.16  58.49  57.58  57.58  29.39  41.14 179.62 179.25    _    _  0.1    _    _    _    _    _    _    _    .
 93 H  55   7.94 120.57  60.76  58.47  58.47  30.24  29.35 178.72 179.62    _    _    _    _    _    _  2.8    _    _    _    .
 94 F  14   7.70 115.65  57.19  60.74  60.77  37.57  30.35 174.79 178.72    _    _  0.1    _    _    _    _  2.7  0.4    _    .
 95 R  21   7.30 117.40  57.97  57.23  57.24  31.13  37.51 177.28 174.78    _    _  0.1    _    _    _    _    _    _  0.5    .
 96 V 129   6.96 111.75  62.00  57.97  57.93  34.61  31.08 174.40 177.27    _    _  0.1    _    _    _    _    _    _    _    .
 97 H 102   8.25 121.57  52.37  62.02  62.00  29.06  34.54 170.83 174.40    _    _  0.1    _    _    _  1.6    _    _    _    .
 98 P        .      .      .      .      .      .      .      .      .      .      .
 99 I  66   8.12 123.15  59.24  61.99  61.97  39.28  32.83 175.23 177.40    .    .    .    .    .    .    _    _    _    _    .
100 P        .      .      .      .      .      .      .      .      .      .      .
101 L  63   8.18 122.41  54.15  62.72  62.68  43.26  31.10 178.26 175.87    .    .    .    .    .    .    _    _    _    _    .
102 E  62   8.97 121.92  58.21  54.00  54.00  29.40  43.42 176.91 178.27  0.1    _  0.2    _    _    _    _    _    _    _    .
103 S 122   8.01 112.68  58.30  58.22  58.22  63.56  29.50 175.12 176.94    _    _  0.1    _    _    _    _    _    _    _    .
104 G 101   8.22 110.73  45.33  58.34  58.31      _  63.60 174.60 175.13    _    _    _    _    _    _    _    _    .    _    .
105 G  97   8.23 108.70  44.98  45.33  45.33      _      _ 173.69 174.58    _    _    .    .    _    _    _    _    .    .    .
106 S 103   8.47 115.01  58.42  44.92  44.95  64.16      _ 174.75 173.70  0.1    _    .    .    _    _    _    _    _    .    .
107 S  27   8.20 118.27  58.44  58.45      _  64.62  64.19 173.40 174.74    _    _    _    _    _    _    _    _    _    _    .
108 D  39   8.37 119.81  54.18  58.44  58.43  41.37  64.68 175.72 173.38    _    _  0.1    _    _    _    _    _    _    _    .
109 V  54   8.38 121.22  61.81  54.13  54.13  33.83  41.48 172.39 175.72    _    _  0.1    _    _    _    _    _    _    _    .
110 V  34   7.34 118.79  58.76  61.76  61.76  33.88  33.85 175.52 172.39  0.1    _    _    _    _    _    _    _    _    _    .
111 L  32   8.23 118.88  53.98  58.57  58.57  39.85  34.12 177.75 175.55  0.2    _  0.2    _    _    _    _    _  1.1    _    .
112 V 100   8.95 122.35  63.92  54.03  54.06  33.99  40.24 175.96 177.72  0.1    _  0.4    _    _    _    _    _    _  0.8    .
113 S  95   7.56 112.35  57.30  63.96  63.91  64.18  34.11 170.99 175.95    _    _  0.1    _    _    _    _    _    _    _    .
114 Y  11   7.09 114.12  53.34  57.28  57.30  40.17  64.32 176.56 170.99    _    _  0.1    _    _    _  1.7    _    _    _    .
115 V  94   8.04 122.64  59.35  53.35  53.34  32.81  40.39 174.90 176.57    _    _  0.2    _    _    _    _  1.6    _    _    .
116 P        .      .      .      .      .      .      .      .      .      .      .
117 S  25   8.96 117.69  57.95  62.59  62.58  63.59  32.57 173.98 177.26    .    .    .    .    .    .    _    _    _    _    .
118 Q  81   7.82 126.14  57.11  57.98  57.99  30.54  63.56 175.73 173.99    _    _    _    _    _    _    _    _    _    _    .
```

4.1.2 Sidechain assignment

The majority of methyl and methylene aliphatic resonances were assigned manually using HBHAcoNH, CcoNH, and HCcoNH spectra. The assignment process typically started from the known backbone chemical shift assignments of Cα and Cβ and used these to correlate and identify the aliphatic carbon sidechain atoms (Cγ and Cδ) in CcoNH spectra (which correlate the sidechain carbons to the N and H of the following residue), then detected their associated protons (Hα, Hβ, Hγ, and Hδ) in HBHAcoNH and HCcoNH spectra, as shown in Figure 4.3.

In total ~80% of the sidechain assignment (663 nuclei) was completed manually for SH2 protein residues 9 to 118 using a number of 2D and 3D NMR spectra. The nuclei correlated in each experiment and acquisition parameters of the NMR spectra used in the sidechain resonance assignment are shown in Table 2.12. The complete sidechain assignment relies on a combination of aliphatic and aromatic sidechain resonances.

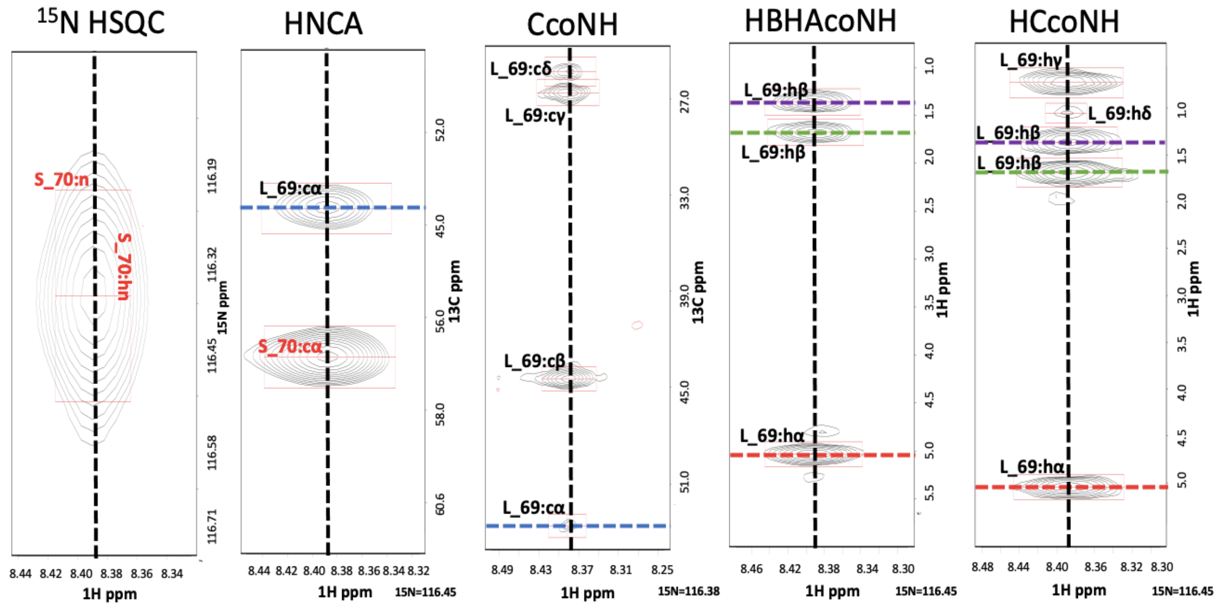


Figure 4.3. An example of slices from N HSQC, HNCA, CcoNH, HBHAcoNH, and HCcoNH NMR spectra used for aliphatic sidechain assignment. The slices for residue S70 were used to obtain the corresponding aliphatic resonances of the preceding residue L69. Resonances are labelled and signals in different spectra which were assigned to the same nucleus are linked with the same dash line colour. 15N: ^{15}N dimension (ppm), 1H: ^{1}H dimension (ppm), 13C: ^{13}C dimension (ppm).

A number of nuclei beyond CHγ including the terminal methyl of Met and methylene of Lys residues were not observed according to the previous method likely because of poor signal dispersion and frequency degeneracies. Moreover Trp17, Gln61 and His81 residues were sequentially correlated to aliphatic resonances using HBHAcoNH, CcoNH, HCcoNH spectra as they were assigned from the preceding residue. However because of their low intensity they were not useful to assign the sidechain of their preceding residues (Pro16, Phe60, and Gln80). Likewise the aliphatic sidechain of Gln10, Tyr15, His19, His97, Ile99, and Val115 residues were not obtained as the following residue is Pro or the missing Gly20, including the last residue Q118.

Because more than 80% of aliphatic sidechain proton resonances had been assigned from the previous experiments, the assignment of the sidechain amide resonances of Gln, Asn and Arg residues was straightforward. ^{15}N NOESY spectra were used to assign the the signal groups of Nδ-Hδ2 of Asn, Nε-Hε2 of Gln and Nε-Hε of Arg in ^{15}N HSQC spectra using the advantage of knowing the correlated intra residue protons, Figure 4.4. Using these assignments, all amide sidechain groups were identified except for the amide groups of Asn72 and Gln118 which did not appear in ^{15}N HSQC. In addition, NHε2 of Gln10 could not be correlated to any other intraresidue protons, and thus was not assigned.

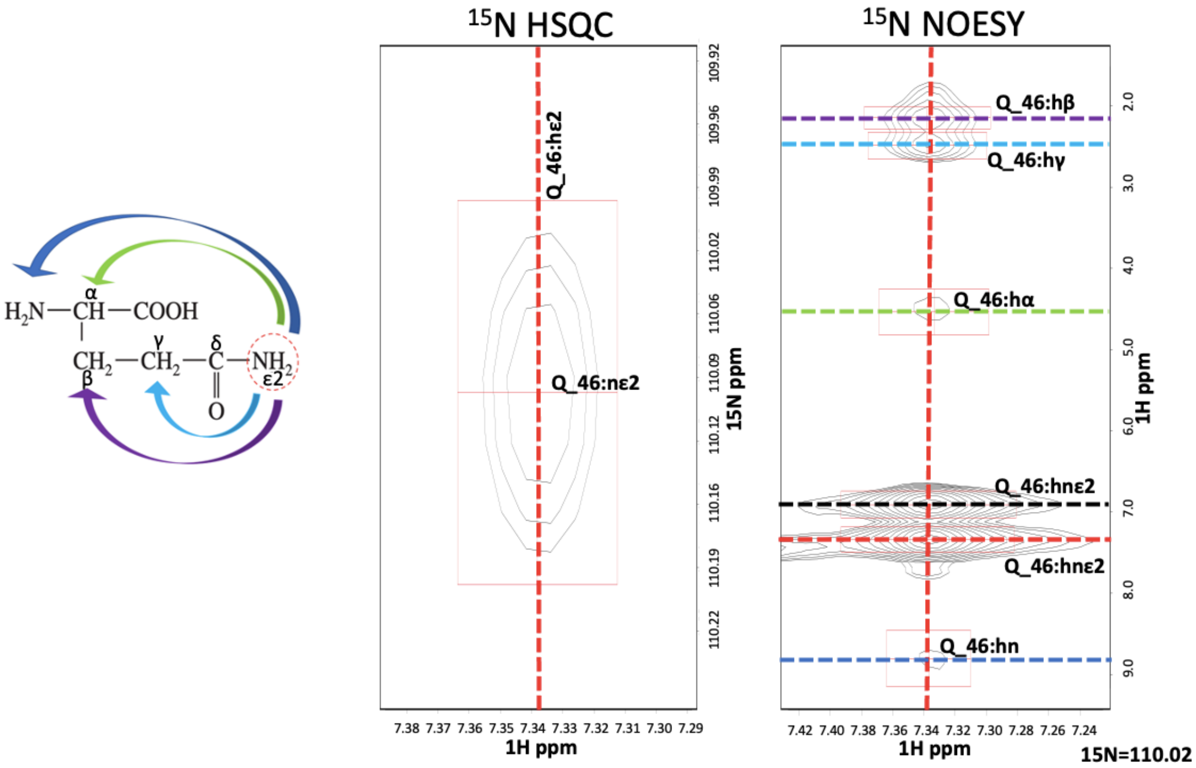


Figure 4.4. An example of slices from ^{15}N HSQC and ^{15}N NOESY NMR spectra used to assign NH sidechain resonances. The slices for residue Q46 are shown connecting through bond the NHε2 aliphatic resonance to intraresidue nuclei in ^{15}N NOESY spectra. Signals are labelled in each spectra with residue name and nuclei type, the same nuclei are linked with the same dashed line colour. 15N: ^{15}N dimension (ppm), and 1H: ^{1}H dimension (ppm).

Aromatic residues compose 16% of the SH2 protein sequence following the His-tag (18 residues), most of them in the folded part as shown in the secondary structure prediction from TALOS-N, Figure 4.11. The aromatic resonances were identified and assigned using ^{13}C HSQC, HBCBCGCDHD, and ^{13}C NOESY aromatic spectra.

The Cβ resonances, which were identified previously in the backbone assignment, were linked through-bond to detect the sidechain δ aromatic ring protons in 2D HBCBCGCDHD spectra (Yamazaki et al., 1993). After that, these assigned δ protons were analyzed to assign their associated δ carbons from the aromatic rings in ^{13}C HSQC aromatic spectra as shown in Figure 4.5. Based on this method all CHδ of aromatic rings were identified except for Phe 42, despite overlapping of Phe and Tyr signals. Moreover Cδ1 of Trp and Cδ2 of His residues in ^{13}C HSQC spectra were readily identified as negative peaks because they are not attached to carbon atoms in the ring. This was used to differentiate these Cδ and assign them through the link to the defined δ1 proton.

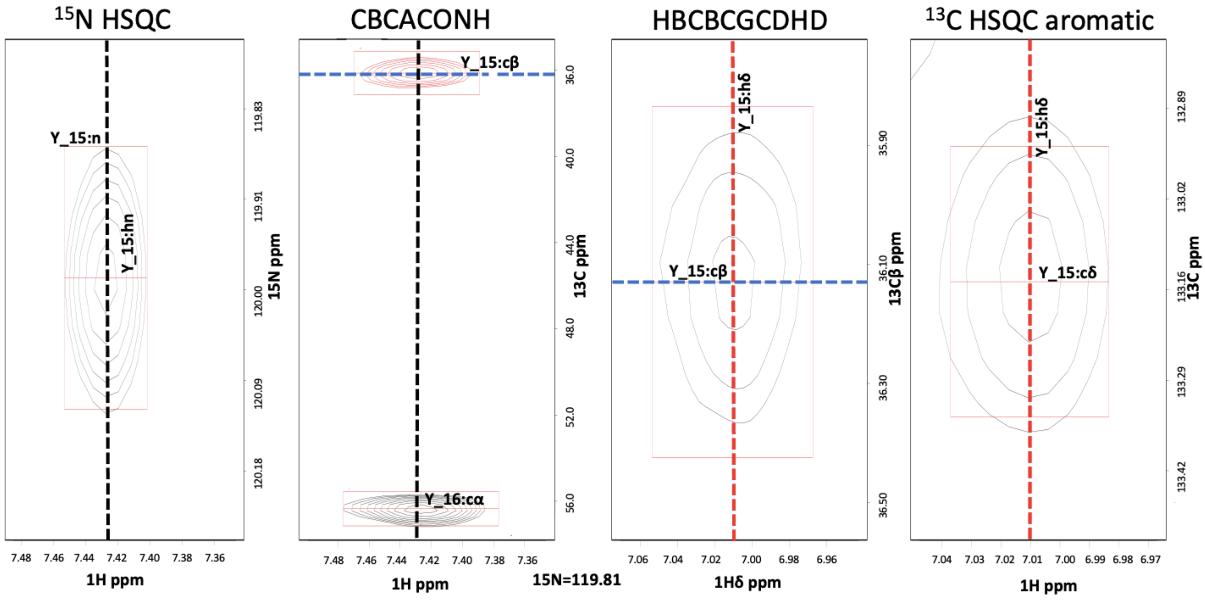
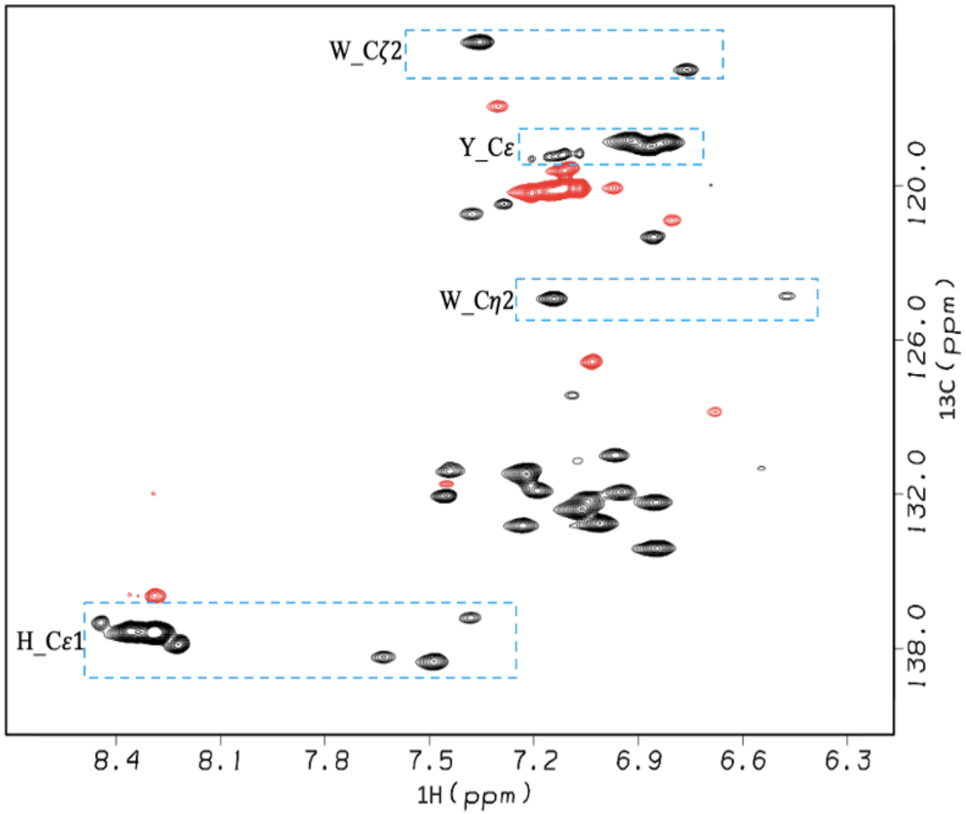
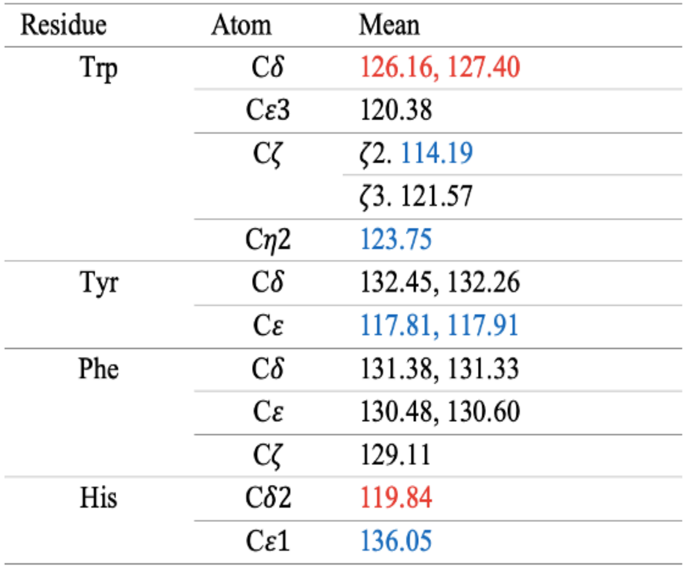


Figure 4.5. An example of slices from ^{15}N HSQC, CBCACONH, HBCBCGCDHD, and ^{13}C HSQC aromatic NMR spectra used for CHδ aromatic sidechain assignment. The slices for residue Y15 are shown connecting through bond the Cβ aliphatic resonance to Hδ using HBCBCGCDHD, then using ^{13}C HSQC aromatic spectra to obtain its corresponding Cδ resonance. Resonances are labelled in different spectra with residue name and nuclei type; the same nuclei are linked with the same dashed line colour. 15N: ^{15}N dimension (ppm), 1H: ^{1}H dimension (ppm), 13C: ^{13}C dimension (ppm).

In addition, the chemical shift statistics in BMRB databases (http://www.bmrb.wisc.edu) show that the average chemical shift of a number of aromatic sidechain carbons have a bigger variance than others, which was used to distinguish between them in ^{13}C HSQC aromatic spectra. The knowledge of the chemical shift average of aromatic sidechain carbons was a useful working basis to define which group of signals belong to which aromatic carbons of particular residue types in ^{13}C HSQC spectra, as shown in Figure 4.6. For example, the average chemical shifts of Cζ2 and Cη2 of Trp, Cε1 of His and Cε of Tyr residues are different from other aromatic carbons which helped to differentiate their peaks in ^{13}C HSQC spectra. After assignment to residue type, these signals were assigned to a particular residue, partly on the basis of intramolecular NOEs in ^{15}N NOESY or ^{13}C NOESY spectra to known protons.

In ^{15}N HSQC HNε1 of Trp signals were assigned by employing the NOEs between these peaks and the defined chemical shifts of Hβ and Hδ in ^{15}N NOESY. Also for confirmation these amide protons (Hε1 of Trp in ^{15}N HSQC spectra) correlated to Hδ1 and Hζ2 protons in 2D ^{1}H ^{1}H NOESY spectra (Figure 4.7). This assignment method could not be applied to assign Hδ1 and Hε2 of His because of overlapped signal.





Residue	Atom	Mean
Trp	Cδ	126.16, 127.40
	Cε3	120.38
	Cζ	ζ2. 114.19
		ζ3. 121.57
	Cη2	123.75
Tyr	Cδ	132.45, 132.26
	Cε	117.81, 117.91
Phe	Cδ	131.38, 131.33
	Cε	130.48, 130.60
	Cζ	129.11
His	Cδ2	119.84
	Cε1	136.05

Figure 4.6. Left; ^{13}C HSQC spectrum for aromatic reisudes of SH2. Red colour shows the nuclei expected to have negative intensity in constant time ^{13}C spectra because of the number of carbon atoms attached, while black colour peaks indicate the positive carbons. The blue dash box are surrounded groups of carbon peaks that have distinguishable chemical shifts, Trp:Cζ2, Tyr:Cε, Trp:Cη2, His:Cε1. 1H: ^{1}H dimension (ppm), 13C: ^{13}C dimension (ppm). Right; Typical range of values for chemical shifts from different aromatic residues shown in dashed boxes, taken from biological magnetic resonances bank (BMRB) databases.

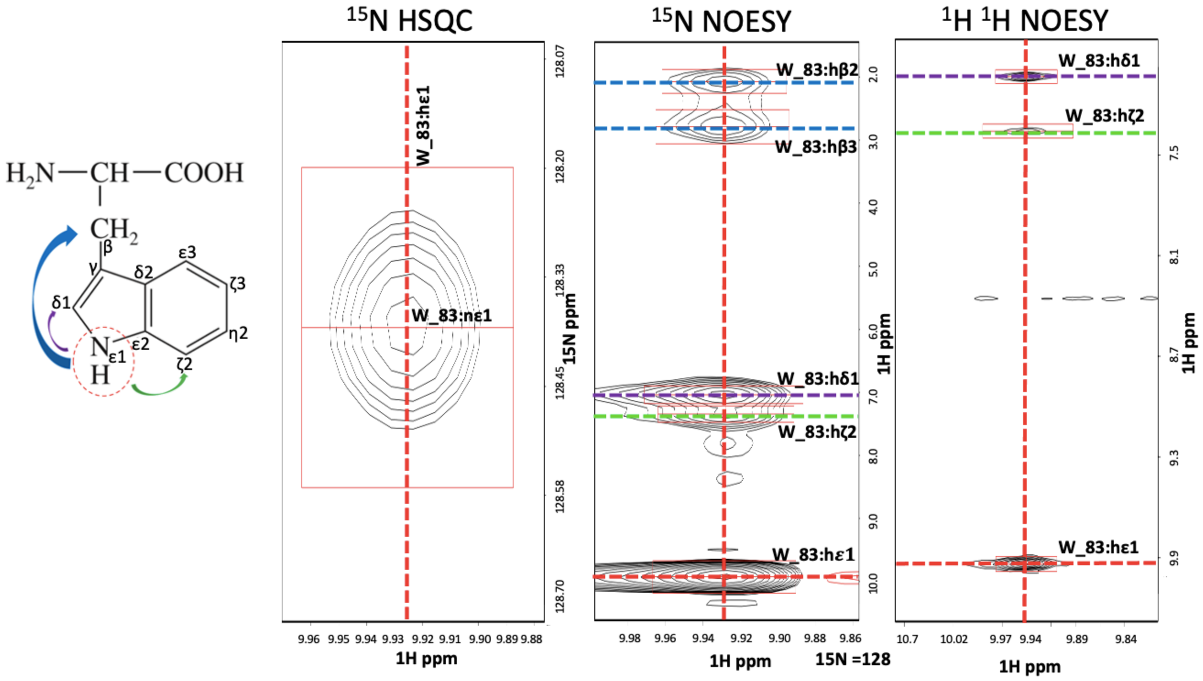


Figure 4.7. Representative slices of amide sidechain assignment spectra. The slices for residue W83 are shown, indicating NOEs connecting through space between amide sidechain proton and nearby aromatic ring protons. Resonances are labelled in different spectra with residue name and nuclei type; the same nuclei are linked using the same dashed line colour. 15N: ^{15}N dimension (ppm), 1H: ^{1}H dimension (ppm).

In a similar way, the overlapped peaks of Cε3 and Cζ3 of Trp, and Cε and Cζ of Phe residues, which typically have very similar carbon chemical shift values, could be distinguished on the basis of strong NOEs to identified protons in ^{13}C NOESY.

In proteins, the two Hδ protons of Tyr and Phe are in different environments and can therefore appear as two signals. The same is true for the Hε pair, and for the attached carbons. The actual appearance in NMR spectra depends on the rate of 180° flips around the Cβ/ Cγ/ Cζ axis. When the flip rate is slow compared to the chemical shift difference, two sets of peaks are seen, while when the flip rate is fast, they average to give a single observed signal (Wüthrich et al., 1975). In SH2, all the pairs of Phe and Tyr aromatic protons are in fast exchange and have a single averaged shift. The manual resonances assignment of the backbone and sidechain nuclei are listed in appendix A.

4.2 Automated resonance assignment

One of the aims of this work was to compare the shift assignments done manually and by automated methods. The automatic resonance assignment calculation was done by the GARANT algorithm (Güntert, 2009; Bartels et al., 1997) within the CYANA program. The required input files consist of the protein sequence and a set of experimental NMR peak lists from multidimensional spectra as listed in Table 2.12.

The peak picking was performed manually for ^{15}N HSQC, ^{13}C HSQC, ^{13}C HSQC aromatic, HNCO, HNCACO, HNCA, CBCACONH, HNCACB, HBHAcoNH, CcoNH, HCcoNH and HBCBCGCDHD spectra, whereas an automatic peak-picking algorithm in Felix2007 was applied over all the diagonal cross peak spectra (HCCH-TOCSY, HCCH-COSY, ^{1}H ^{1}H NOESY, ^{15}N NOESY, and ^{13}C NOESY and ^{13}C NOESY aromatic).

After the peak lists were prepared, the automated resonance assignment calculation for SH2 was done using a standard automated chemical shift assignment script CALC.cya within CYANA version 3.98.5. A number of nuclei were excluded from the calculation as they are difficult to detect as they have fast exchange with solvent, such as Cζ2 of Arg, NHζ of Lys, Hγ of Thr and Hγ of Ser.

The automated chemical shift assignment was performed in two steps, following the recommended procedure (Güntert, 2004): an ensemble of chemical shift assignments was computed followed by combining the resultant raw chemical shift assignments into a single consensus resonance list. The ensemble of chemical shift assignments was obtained from 20 runs of the GARANT algorithm, in which the iteration size for one generation was 100. Each independent run started from the same experimental peak lists but using a different random seed value, and optimised the match between observed peaks and expected peaks based on the knowledge of the amino acid sequence and the magnetization transfer pathways in the spectra used (Bartels et al., 1997). The matching was done with the recommended tolerance values of 0.03 and 0.4 ppm for the ^{1}H dimension and for the ^{13}C and ^{15}N dimensions, respectively.

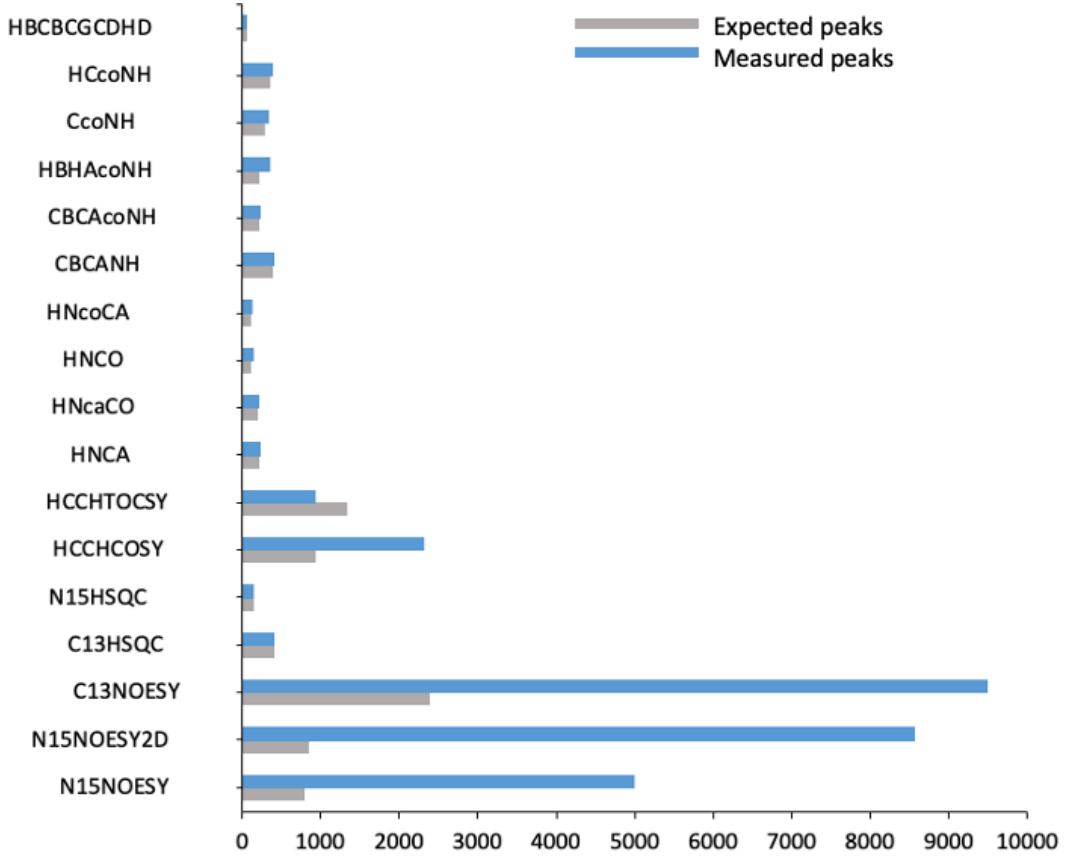
A scoring function estimates the match between measured and expected peaks, to distinguish between correct and incorrect resonance assignments. The score incorporates the essential features of a correct resonance assignment such as the presence of expected peaks in the spectra used, the positional alignment of peaks that originate from the same atoms, and the statistical agreement of the assigned resonance frequencies with a chemical shift database assembled from the known resonance assignments of many proteins (Güntert, 2003). A good assignment should have a score $\geq$ 80%.

4.2.1 Peak picking

The identification of the NMR signals in multidimensional spectra is referred to as peak picking and is the first step to analysis of NMR spectra. The extent to which the peak picking, whether manual or automatic, yielded the expected number of peaks was measured by comparing the number of the experimentally observed peaks in each peak list with the theoretical expected peak number, which was predicted from matching the knowledge of the primary structure and the magnetisation transfer pathway in the spectra used, as explained in López-Méndez and Güntert, 2006.

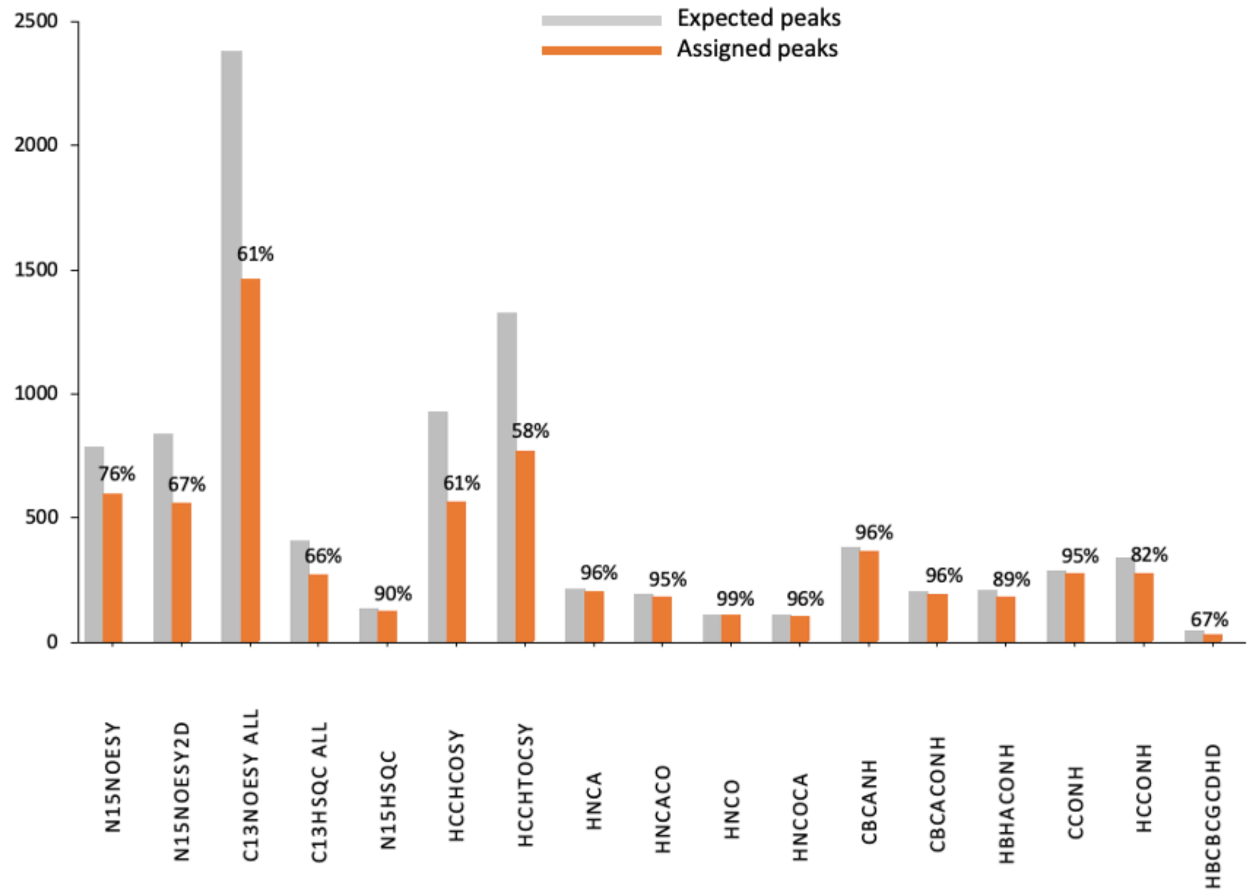
Any peak picking, whether manual or automated, has to be able to distinguish noise and artifacts from peaks, and should have some way of dealing with overlapping peaks. It is difficult to make these distinctions automatically. Therefore a manual inspection was done to remove noise from the peak lists and make some judgement about strongly overlapping peaks. Nevertheless the total number of measured peaks in the peak lists was much higher than the number of expected peaks for most of the automated picked spectra, as shown in Figure 4.8. The discrepancy is particularly marked for NOESY spectra. This is because for many of the triple resonance spectra the number of peaks per residue is limited, and they are all expected to have a similar intensity. However in NOESY spectra, there is a much larger dynamic range of peak intensity, and it is therefore necessary to set the threshold much lower in order to catch all the genuine peaks, which results in a large increase in the number of noise peaks detected. For the HCCH-TOCSY spectrum although the automated peak picker found fewer observed peaks than predicted as shown in Figure 4.8, the overlooked peaks happened mostly because of abundant cross-peak overlap especially near the diagonal, and incomplete magnetization transfer in the spectra for the longer aliphatic sidechains (Bax et al., 1990; Ikura et al., 1991). On the other hand, in the manually picked spectra of the backbone and the non-diagonal aliphatic sidechains, the number of measured peaks was almost the same or slightly higher than the number of expected peaks (Figure 4.8). This is because the backbone and the simple aliphatic sidechain spectra have less signal overlap and are more sensitive (Bax et al., 1993).

The effectiveness of peak picking was reflected in the percentage of predicted peaks that could be assigned. It was observed that more peaks were missing in the automatically picked spectra due to the imperfections of the automated peak picking algorithm and the difficulties of analysing spectra with a large number of overlapped signals. However a high percentage of assigned peaks were obtained for most of the spectra which were picked using conventional manual methods (Figure 4.9).



Spectra	Expected peaks	Measured peaks
N15NOESY	786	4982
N15NOESY2D	835	8555
C13NOESY all	2379	9487
C13HSQC all	409	399
N15HSQC	136	133
HCCHCOSY	927	2306
HCCHTOCSY	1324	928
HNCA	212	230
HNcaCO	189	203
HNCO	106	139
HNcoCA	108	125
CBCANH	379	403
CBCAcoNH	201	225
HBHAcoNH	205	348
CcoNH	286	326
HCcoNH	341	389
HBCBCGCDHD	46	55
ALL	8869	28305

Figure 4.8. Number of observed peaks and expected peaks that could be assigned in automated resonance assignment of SH2. The number of peaks expected is based on the amino acid sequence and the ideal magnetization transfer pathway (López-Méndez and Güntert, 2006), and in the NOESY spectra the peaks expected are all ^{1}H-^{1}H distances shorter than 4.5 Å in the CYANA reference structure (random.pdb). Measured peaks are blue bars and expected peaks are grey bars. The aliphatic and aromatic peak lists of ^{13}C HSQC and ^{13}C NOESY were combined into one list.



Spectra	Expected peaks	Assigned peaks
N15NOESY	786	596
N15NOESY2D	835	558
C13NOESY all	2379	1461
C13HSQC all	409	268
N15HSQC	136	122
HCCHCOSY	927	564
HCCHTOCSY	1324	766
HNCA	212	204
HNcaCO	189	180
HNCO	106	105
HNcoCA	108	104
CBCANH	379	363
CBCAcoNH	201	193
HBHAcoNH	205	182
CcoNH	286	273
HCcoNH	341	278
HBCBCGCDHD	46	31
ALL	8869	6248

Figure 4.9. Number of expected peaks and assigned peaks of each NMR spectrum that were used for automated resonance assignment of SH2. The number of expected peaks is based on the amino acid sequence and the ideal magnetization transfer pathway (López-Méndez and Guntert, 2006), and in the NOESY spectra the peaks expected are all ^{1}H-^{1}H distances shorter than 4.5 Å in the CYANA reference structure (random.pdb). The percentage on the top of the bar represents the predicted peaks that could be assigned. Expected peaks in grey bars and assigned peaks in orange bars.

4.2.2 Automated resonance assignment by CYANA

A complete automated resonance assignment was obtained for SH2 (1-118 a.a, 1441 atoms). Overall ~ 76% of assigned atoms were scored highly confident (strong) as they were assigned within the tolerance values. The majority of the strong assigned nuclei came from backbone and aliphatic sidechain resonances (Figure 4.10). The automated chemical shift assignment is listed in appendix A.

The accuracy of the automated consensus chemical shifts was assessed by comparison with the resonance assignment obtained by conventional methods, excluding resonances that were not assigned manually. It has been suggested that ≥ 90% of completeness in resonance assignment is required before proceeding with NOESY assignment and structure calculation, in order to have confidence in the structure calculation (Güntert, 2003). A consensus chemical shift assignment was considered to agree with the reference assignment if the two corresponding chemical shift values differed by less than 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N.

99% of the automated backbone chemical shift assignments ($^{1}H_{N}$, ^{15}N, $^{13}C_{o}$, $^{13}C_{\alpha}$ and $^{13}C_{\beta}$) of SH2 (9-118 a.a) were considered to agree with the reference assignment. The only resonances that were not assigned correctly are the amide chemical shifts of Trp17 and Gln61, which are two of the four resonances that had missing signals (but assigned H and N) in the manual assignment. In addition the amide resonances of Gly20 and backbone resonances of Pro16 could not be confirmed whether they were assigned correctly as they were missing in the reference assignment.

For aliphatic methyl and methylene sidechain groups, the agreement between automated assignments and reference assignments was 97%. The differences to the reference assignment were caused by the exchange of resonance assignments within a residue; this permutation is less likely to affect the automated derived structure than a completely wrong assignment or an exchange of assignments between different residues.

The automated method succeeded in assigning most of the backbone and aliphatic sidechain resonances correctly because their assignment was based on number of pair related spectra which gave a redundancy among different spectra. The abundance of multiple peaks for a given atom helped to lower the degree of ambiguity by allowing cross checking during the automatic calculation even when a few expected peaks were missing.

On the other hand, most of the NOE-based chemical shift assignments did not match with the reference assignment. This includes the amide sidechains of Arg, Asn and Gln, and the methyl groups of aromatic ring: only 53% of them were assigned correctly (agreed with manual assignment) as their assignment basically relies on NOEs.

The automated assignment of those nuclei depends on the quality of the information content of the NOE input data, which was incomplete according to the number of assigned peaks, resulting from failure of the automated peak picking algorithm to deal with complex spectra. In addition, the assignment of missing true peaks could be replaced with wrong assignments coming from noise and artefacts in the NOE peak lists. CYANA continues to check and refine its assignments during the course of structure calculations, and it is possible that some of these incorrect assignments would have been corrected. On the other hand, the accuracy of the automated structure calculation relies heavily on the accuracy of the first iteration of the structure calculation, which in turn depends on the accuracy of the initial set of assignments (Güntert and Buchner, 2015). It is therefore important that the assignments should be as correct as possible before the start of the structure calculation. SH2 is a small protein with sharp signals and relatively little overlap. It also has only a small number of signals missing, presumably because of exchange processes. One would therefore expect that automated methods would handle SH2 better than most other proteins. It can be concluded that automated assignments should always be checked manually before proceeding to structure calculation. A comparison of automated resonance assignment and reference assignment is presented in appendix A.

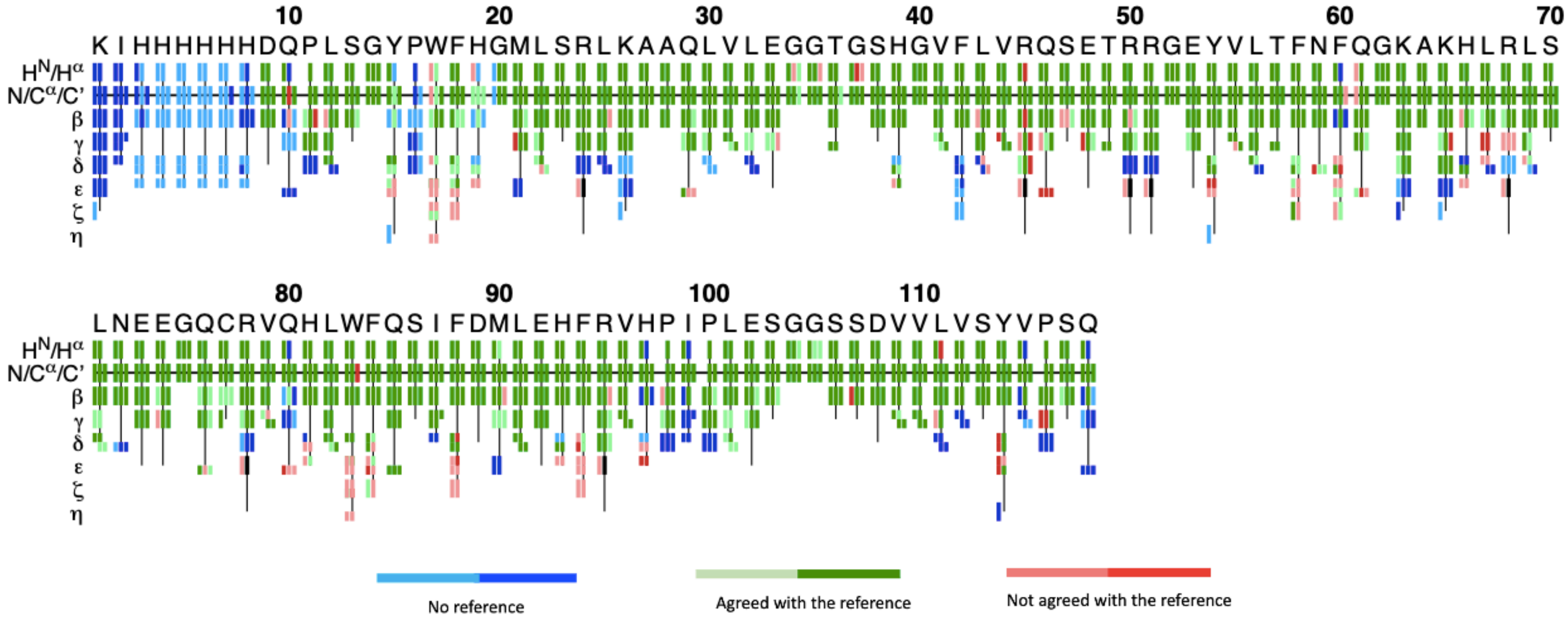


Figure 4.10. An overview graphic of CYANA assignment for SH2 residues 1-118, using manual assignment as a reference. On the top are the residue and residue name. In general, light colours indicate less confident assignments and dark colours illustrate strong assignments; blue color indicates no reference assignment available, green color shows assignment by CYANA agreed with reference assignment within defined tolerance value, red color shows the assignments that do not agree with reference assignment, the black color represents the assignments that were found in reference assignment but not assigned by CYANA. The row labelled displays for each residue: HN on the left and Hα in the center, and N, Cα, C' shows for each residue, from left to right, the N, Cα, and C' assignments. The next rows show the side-chain assignments: in the center the heavy atom assignment and hydrogen atoms to the left and right. In branched side-chains, the row is divided into upper and lower.

4.3 Additional structural restraints

4.3.1 Dihedral angles

Most of the structural restraints used for the structure calculation come from NOEs, and are discussed in Chapter 5. Here, the additional restraints on dihedral angles and hydrogen bonds are discussed.

Restraints for the backbone ϕ, ψ and the sidechain χ^1 angles of SH2 were generated by TALOS-N (Shen and Bax., 2013). The program uses the given backbone chemical shifts (HN, NH, CO, Hα, Cα, and Cβ) with sequence information to make quantitative predictions for the protein backbone angles ϕ and ψ, also χ^1 sidechain torsion angles.

The backbone dihedral angles ϕ and ψ were predicted for SH2. 93 out of 118 residues were classified as strong or generous as their matches show consistent values for ϕ and ψ. These reliable restraints were applied as dihedral restraints for the structure calculation and refinement. TALOS generates a standard deviation for the predicted angle. In the structure calculation, this value was used to produce upper and lower bounds for the angle. However, residues for which their matches were inconsistent, and were therefore classified by TALOS as Warn or Dynamic, were eliminated from the list.

The sidechain dihedral angles χ^1 were obtained for 54 residues and were employed as additional restraints in the structure calculation.

TALOS-N gives the secondary structure classification of the amino acids of a protein from the given backbone chemical shifts. The secondary structure prediction for SH2 showed a central β sheet consisting of three long β strands (β 2, 3, and 4) followed by two short strands (β 5 and 6), the β sheet being located between N and C terminal α helices (α 1 and 2). In addition to that there are two small β strands at the N and the C terminus of the protein (β 1 and 7). Also there are a number of long loops connecting the structured regions of the protein (Figure 4.11).

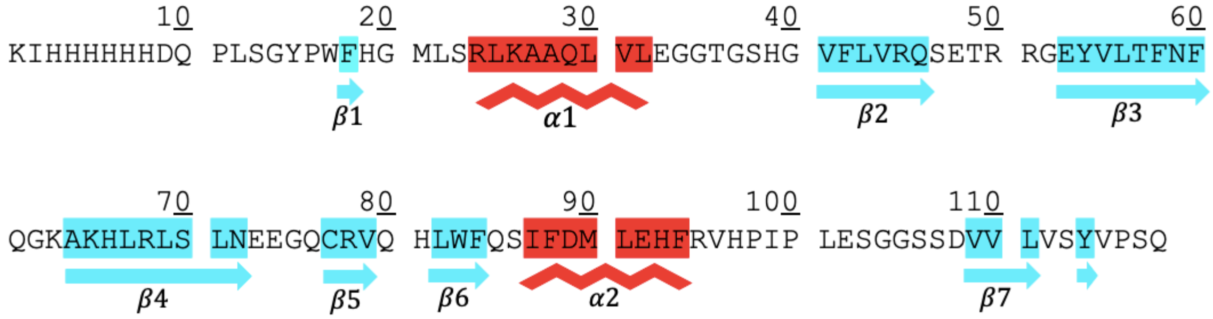

10 20 30 40 50 60
KIHHHHHHDQ PLSGYPWFHG MLSRLKAAQL VLEGGTGSHG VFLVRQSETR RGEYVLTFNF
β1
α1
β2
β3
70 80 90 100 110
QGKAKHLRLS LNEEGQCRVQ HLWFQSIFDM LEHFRVHPIP LESGGSSDVV LVSYVPSQ
β4
β5
β6
α2
β7

Figure 4.11. Secondary structure of SH2 as predicted by TALOS-N. The secondary structure resulting per-residue likelihood estimates are displayed; aqua for beta-strand and red for alpha-helix. Below the sequence are the secondary structure elements based on TALOS-N prediction. Above the sequence is the numbering of the amino acid sequence of SH2 protein.

4.3.2 Chemical shift temperature coefficients

Chapter 5 describes structure calculations of SH2. During the course of the calculations, the results were analysed using a new program, ANSURR (Fowler et al., 2020) which compares the rigidity of the resulting structure to the experimental rigidity as measured by a modified Random Coil Index calculation (Berjanskii and Wishart, 2005). This comparison indicated that the structures generated were too floppy and lacked restraints. This is a very typical problem with NMR structure calculations, and does not indicate any specific problems with the SH2 structure calculation, but presented an opportunity to investigate whether the accuracy of the structure could be improved by addition of extra restraints, in particular hydrogen bonds. This section discusses how hydrogen bond restraints could potentially be generated from the measurement of amide proton temperature coefficients.

Studies carried out by the Williamson lab (Baxter & Williamson, 1997; Tomlinson & Williamson, 2012) have shown that amide temperature coefficients (that is, the change in chemical shifts of amide protons as a result of change in temperature) are a useful guide to hydrogen bonding. They defined a cutoff of -4.5 ppb/K, and concluded that amide protons with a temperature coefficient more positive than -4.5 ppb/K have a high probability of being involved in intramolecular hydrogen bonding. This measure defines the amide group involved in the hydrogen bond, but does not identify the group that it forms a hydrogen bond to. Chapter 5 investigates how such restraints can be applied: here the identification of amide protons that match the criterion is discussed.

Intramolecular hydrogen bonds between the amide backbone and carbonyl groups were investigated for non-proline residues of SH2 by chemical shift temperature coefficients. The temperature coefficients of amide protons were calculated from a series of ^{15}N HSQC spectra at 288, 293, 298, and 303K, at pH6. Spectra were referenced to TSP in 1D spectra at each temperature. For measuring the NH chemical shift at each temperature the HSQC cross peaks were picked in Felix2007. After that, a straight line was fitted to the chemical shift versus temperature data by least squares minimisation, and the NH chemical shift temperature dependence was calculated as the gradient of that line. The fitted data plot is displayed in Figure 4.12.

103 out of 105 backbone amide protons of SH2 showed a linear temperature dependence. The exceptions were Gly20 and Trp17, which could not be observed. Most of the amide temperature coefficients are negative (ie, the chemical shift value decreases as the temperature is raised), which is related to the predominately down field shift of hydrogen bonded amide protons (Baxter and Williamson, 1997). Positive temperature coefficient values were observed for Gln61 and Leu69 residues. The NH coefficient experimental data of SH2 is presented in Table 4.2.

Amide protons were defined as donors if they have chemical shift coefficient above -4.5 ppb/K as discussed in Baxter and Williamson, 1997. Sixty-five amide protons have temperature gradients greater than -4.5 ppb/K and thus were considered to be potential hydrogen bond donors, Table 4.2. Many of those residues are part of regular secondary structures according to TALOS-N, Figure 4.11. Residues Leu25, Lys26, Ala27, Ile87, and Asp89 are expected to be located at the N-terminal end of an α helix, as shown in Figure 4.11 and 4.13. They have temperature coefficients that are more negative than -4.5 ppb/K, implying that they are not involved in hydrogen bonds. The first four amides in an α-helix are not hydrogen bonded to carbonyls in the helix, although they may be hydrogen bonded to sidechain oxygens forming an “N-cap” motif (Richardson and Richardson, 1988). This result therefore usefully defines the N-terminal ends of the helices. Residues Gln46, Ala64, His66, Arg68, Leu71 and Trp83 are within β strands but have coefficients more negative than -4.5 ppb/K so are less likely to be involved in hydrogen bonding. The geometry of β strands implies that for the outer strands of a sheet, every other residue points out from the sheet and therefore is not involved in hydrogen bonding. In most cases, these results therefore confirm the topology of the sheet and which direction the outer strands are pointing, which is explained in detail in chapter 5. The most interesting result was that the temperature coefficient data predicted several hydrogen bonded donor residues in loops and in the C-terminal unstructured region of the protein (Figure 4.13).

Table 4.2. Amide chemical shift temperature dependences for SH2 protein. Yes indicates that the residue displayed coefficient > -4.5 ppb/K, thus is predicted to be involved in a backbone hydrogen bond as a donor.

Residue	Δδ/ΔT ppb/K	Residue	Δδ/ΔT ppb/K	Residue	Δδ/ΔT ppb/K
D9	-7.31	**R50**	-6.05	**I87**	-5.11
Q10	-7.35	**R51**	-6.59	**F88**	-2.60 yes
L12	-3.18 yes	**G52**	-5.79	**D89**	-4.91
S13	-5.81	**E53**	-4.31 yes	**M90**	-2.87 yes
G14	-4.83	**Y54**	-3.66 yes	**L91**	-3.60 yes
Y15	-2.27 yes	**V55**	-3.44 yes	**E92**	-1.66 yes
F18	-1.32 yes	**L56**	-2.00 yes	**H93**	-2.09 yes
H19	-3.86 yes	**T57**	-1.49 yes	**F94**	-1.29 yes
M21	-6.81	**F58**	-4.15 yes	**R95**	-1.32 yes
L22	-1.80 yes	**N59**	-3.42 yes	**V96**	-2.65 yes
S23	-6.32	**F60**	-2.49 yes	**H97**	-3.49 yes
R24	-2.85 yes	**Q61**	1.78 yes	**I99**	-4.53
L25	-5.78	**G62**	-0.38 yes	**L101**	-8.18
K26	-6.58	**K63**	-6.59	**E102**	-6.98
A27	-4.96	**A64**	-6.97	**S103**	-3.44 yes
A28	-2.83 yes	**K65**	-3.24 yes	**G104**	-2.41 yes
Q29	-0.69 yes	**H66**	-7.29	**G105**	-5.55
L30	-2.39 yes	**L67**	-2.10 yes	**S106**	-8.14
V31	-3.53 yes	**R68**	-5.31	**S107**	-3.83 yes
L32	-0.92 yes	**L69**	1.82 yes	**D108**	-5.42
E33	-0.52 yes	**S70**	-2.54 yes	**V109**	-3.82 yes
G34	-9.09	**L71**	-8.10	**V110**	-2.85 yes
G35	-0.44 yes	**N72**	-2.37 yes	**L111**	-7.94
T36	-7.53	**E73**	-2.09 yes	**V112**	-6.37
G37	-5.26	**E74**	-1.95 yes	**S113**	-7.61
S38	-2.39 yes	**G75**	-3.46 yes	**Y114**	-3.52 yes
H39	-2.98 yes	**Q76**	-3.44 yes	**V115**	-3.61 yes
G40	-4.47	**C77**	-4.27 yes	**S117**	-3.27 yes
V41	-2.97 yes	**R78**	-3.04 yes	**Q118**	-7.12
F42	-3.31 yes	**V79**	-3.22 yes		
L43	-2.53 yes	**Q80**	-8.93		
V44	-2.11 yes	**H81**	-5.66		
R45	-0.49 yes	**L82**	-2.79 yes		
Q46	-4.82	**W83**	-6.56		
S47	-0.76 yes	**F84**	-2.85 yes		
E48	-10.39	**Q85**	-4.32 yes		
T49	-1.58 yes	**S86**	-0.96 yes		

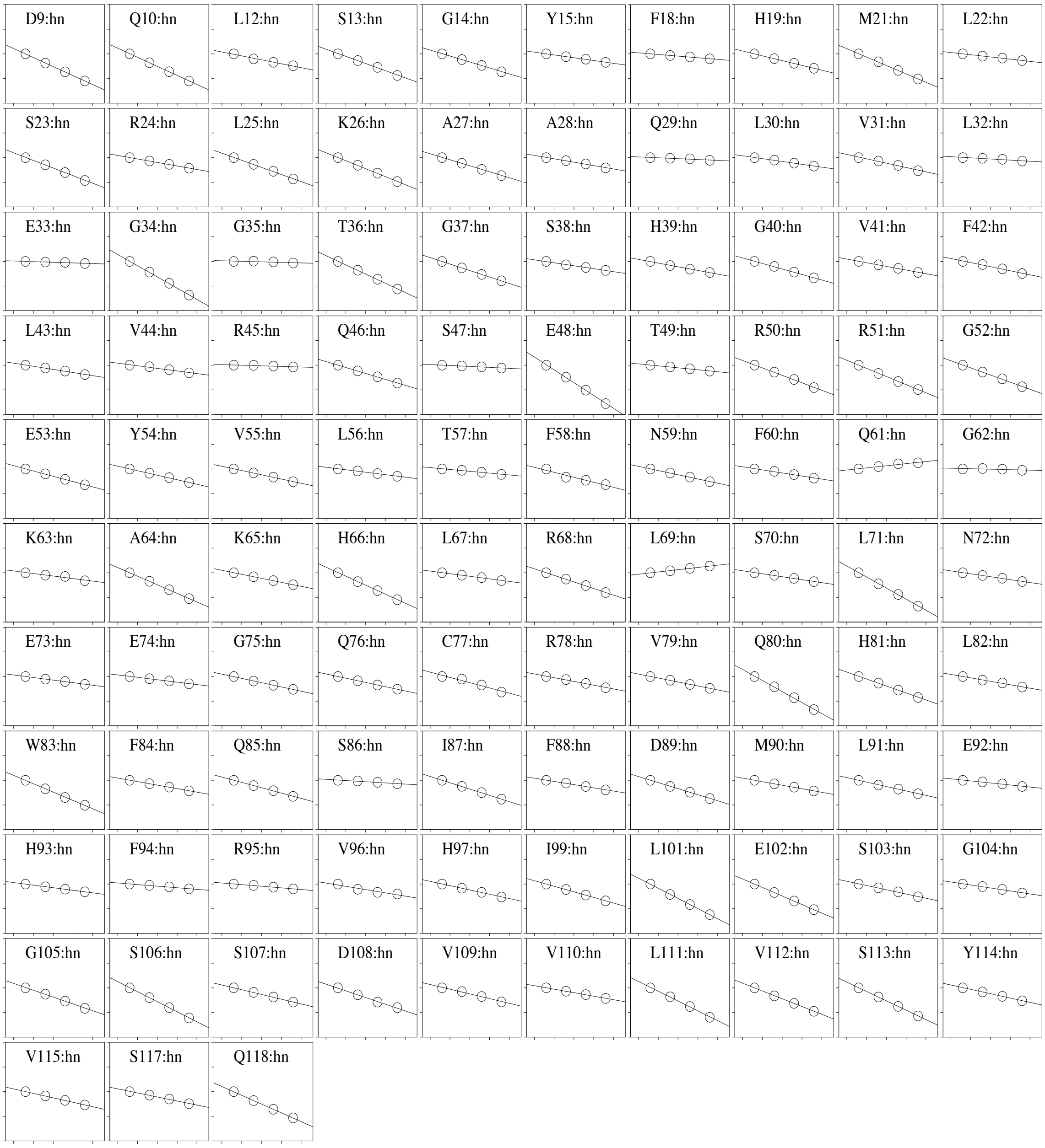


Figure 4.12. Dependence of chemical shift on temperature for backbone amide protons of SH2. The best fit lines are identified from which the temperature coefficient was obtained. On each graph, the x-axis displays the temperature range from 283 K to 303 K with 5° intervals, and the y-axis spans 0.8 ppm with a tick mark interval of 0.2 ppm.

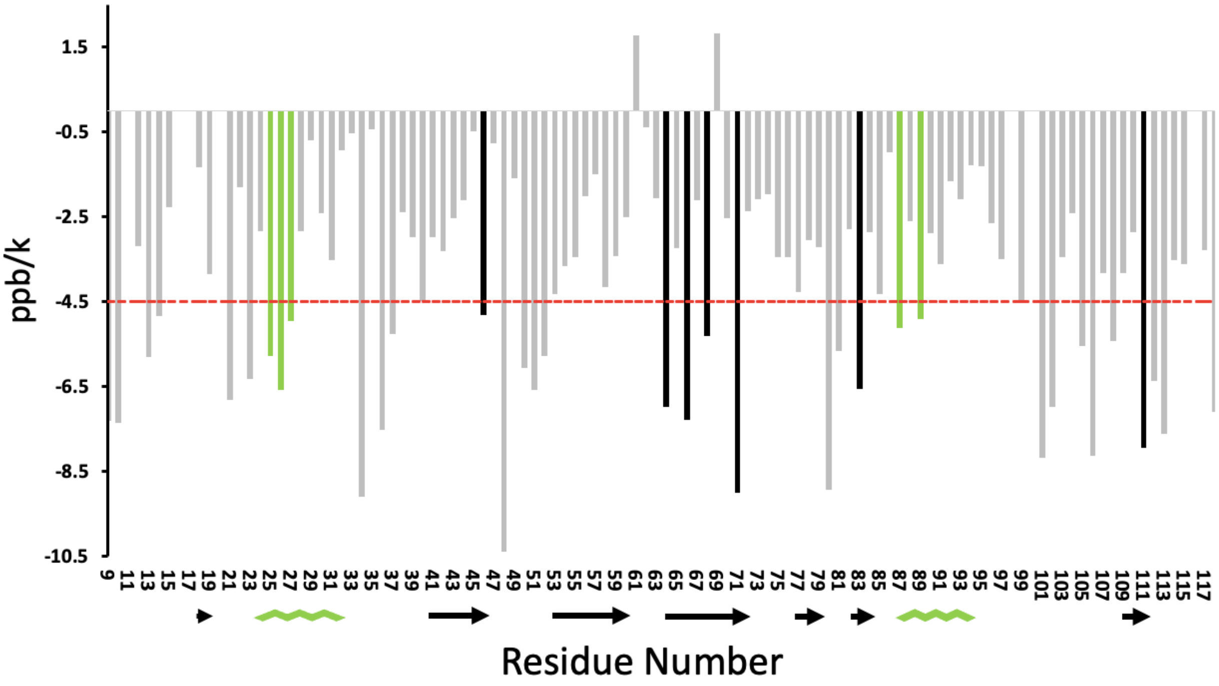


Figure 4.13. NH temperature coefficients (ppb/K) of SH2. The amide temperature coefficient data are plotted against residue number. The red line represents the -4.5 ppb/K cut-off value which was used as an indicator for whether the amide proton is involved in intramolecular hydrogen bonding. The temperature coefficient values above the cut-off (-4.5 ppb/K) are likely hydrogen-bond donors, and below are not hydrogen bonded. Below the residue number at the bottom is the secondary structure prediction based on TALOS-N. The green bars are residues in α helix and black bars are residues in β strands which have coefficients < -4.5 ppb/K.

Chapter 5 NMR structure determination of SH2

This chapter describes in detail the NMR structure determination of the SH2 domain. Here different procedures to calculate the NMR protein structure using an automated method are proposed, and new validation criteria for structure accuracy are presented. The structure calculation was carried out in four stages: automated NOESY assignment and structure calculation using the CYANA program; structure re-calculation and refinement by the Crystallography and NMR system (CNS) program; the selection of the ensemble structure; and the structure validation. The NOE statement sets up the database for distance restraints such as interproton distances. Figure 5.1 illustrates the general flow-chart of the NMR structure determination of SH2. CYANA is the most popular method for automated structure calculation (using the FLYA routines within CYANA). Up till now, at Sheffield, structure calculations have always been carried out by a laborious manual assignment of NOESY spectra followed by iterative calculations, checking structure restraints at each stage. This project had the aim of calculating the structure of the SH2 domain, and doing it using the CYANA automated method, as a way of testing whether this automated method works, specifically when used by someone who has had no previous experience of structure calculation. This was the first time that CYANA had been used in the lab.

A set of structure calculation attempts were executed using the CYANA (3.98.5) program based on automated NOE assignment. The methodology of CYANA structure calculation is described in general terms. The output from CYANA is a set of unrefined structures, plus the final set of restraints derived from the calculations. The obtained output restraints (NOEs and dihedral restraints) from the CYANA structure calculation were then combined with the temperature coefficient data described in the previous chapter, to generate an extra H-bond restraint list. This H-bond list was refined by four subsequent rounds of calculations in CYANA (calculation rounds 2, 3, 4 and 5 in Figure 5.1) based on analysis of the preceding calculation, including the use of the ANSURR program described below, to improve the resulting structure. The success of the CYANA structure calculations is documented here.

The NMR structure of SH2 was then recalculated and refined in explicit solvent using the

Crystallography and NMR system (CNS) program using the final set of restraints from the CYANA calculations. After that, two set of 20 structures were selected on the basis of the ANSURR analysis score (a new method, discussed here) and another set based on the total energy. For the final selection, both a good ANSURR score and a low total energy was required, to select the final ensemble structures and to represent the NMR structure for SH2.

Finally the geometric validation of the final SH2 structures was assessed according to conventional measurements of structure accuracy such as Ramachandran distribution and the RMSD among the ensemble structures. This chapter concludes with a discussion of the success and problems of these calculations.

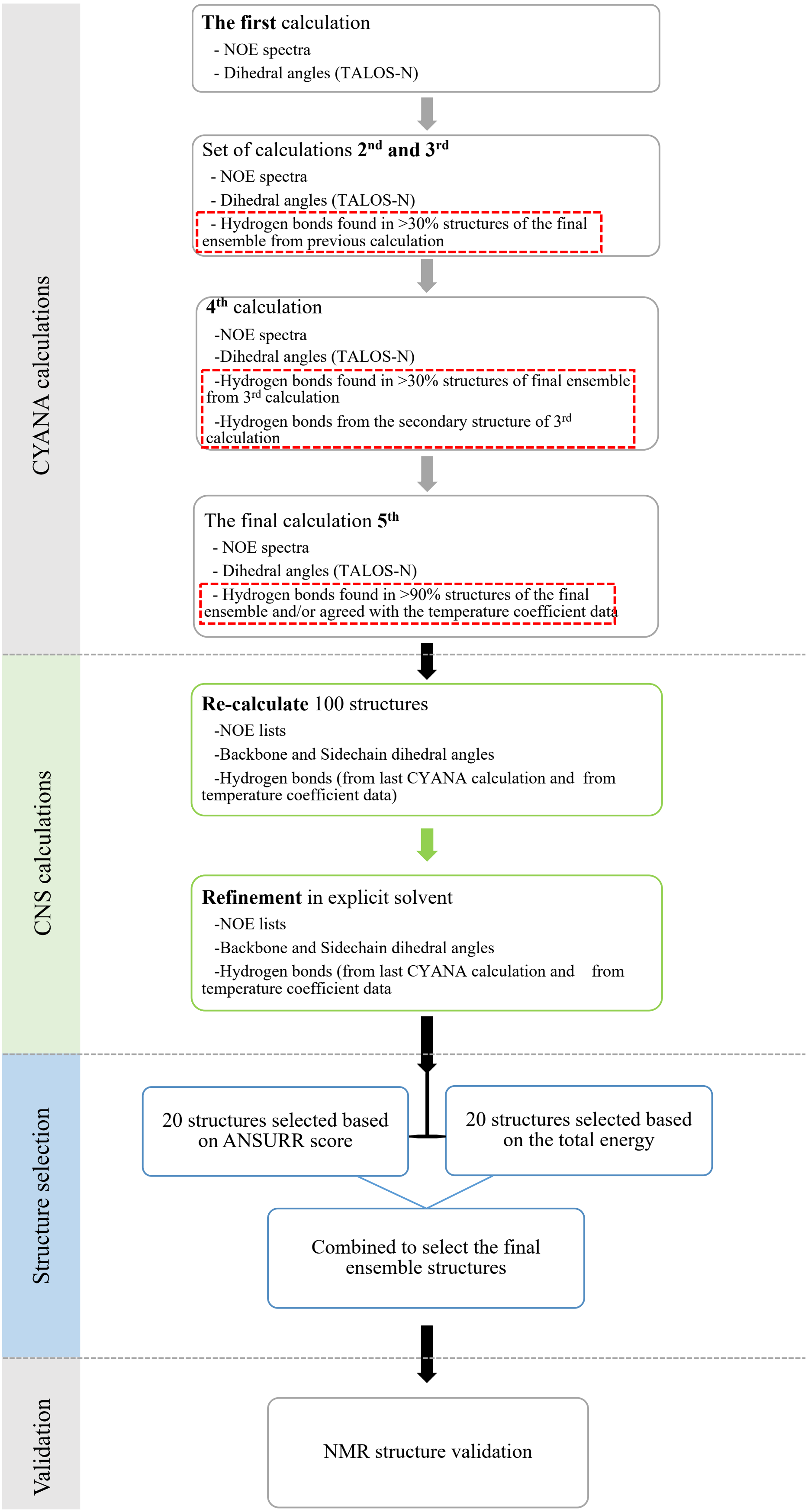

CYANA calculations
The first calculation
- NOE spectra
- Dihedral angles (TALOS-N)
Set of calculations 2nd and 3rd
- NOE spectra
- Dihedral angles (TALOS-N)
- Hydrogen bonds found in >30% structures of the final ensemble from previous calculation
4th calculation
-NOE spectra
-Dihedral angles (TALOS-N)
-Hydrogen bonds found in >30% structures of final ensemble from 3rd calculation
-Hydrogen bonds from the secondary structure of 3rd calculation
The final calculation 5th
- NOE spectra
- Dihedral angles (TALOS-N)
- Hydrogen bonds found in >90% structures of the final ensemble and/or agreed with the temperature coefficient data
CNS calculations
Re-calculate 100 structures
-NOE lists
-Backbone and Sidechain dihedral angles
-Hydrogen bonds (from last CYANA calculation and from temperature coefficient data)
Refinement in explicit solvent
-NOE lists
-Backbone and Sidechain dihedral angles
-Hydrogen bonds (from last CYANA calculation and from temperature coefficient data
Structure selection
20 structures selected based on ANSURR score
20 structures selected based on the total energy
Combined to select the final ensemble structures
Validation
NMR structure validation

Figure 5.1. A general scheme of the NMR structure calculation for SH2 protein. The flow-chart is divided into four stages.

5.1 Overview of CYANA automated NOESY assignment and structure calculation

The overall CYANA structure calculation consists of seven connected cycles of automated NOESY assignment and structure calculation using torsion angle dynamics driven simulated annealing procedures (Güntert, 2004). These iterative cycles are followed by a final cycle of structure calculation using only unambiguous distance restraints, with all remaining unassigned NOE peaks set aside and not used in the calculation. Initially each cycle starts from 100 conformers with random values of torsion angles, and each conformer is calculated with 10000 steps of torsion angle dynamics. After that, the 20 conformers are selected which have the lowest target function value, to represent the final structure as a bundle (Güntert et al., 2015). From each CYANA cycle, a number of output files is generated which contain the resulting NOE assignments and the calculated structures (cycleNo.noe, cycleNo.upl, cycleNo.ovw, cycleNo.pdb). The content of these files is explained in Section 2.10.3. The information is transferred onto the subsequent cycles through the bundle of structures resulting from the preceding cycle, which is used as a guide to the next NOE assignments. Other than that, the same input data are used in every cycle of structure calculation as specified in the CYANA documentation (Güntert, 2004).

The algorithm for automated NOESY assignment is executed on the basis of a probabilistic framework. There are a number of filtering criteria applied within the structure calculation cycles (Güntert et al., 2015). In the beginning of each cycle, consistency checking criteria are applied to measure the agreement between peak positions as listed in NOESY peak lists and their given corresponding chemical shift values within tolerance ranges as specified in CYANA macros. As a consequence of the agreement, one or more initial assignments is created for each individual NOESY cross peak. A Gaussian probability value (P_{shifts}) is used to express the fitting or matching quality between the input data of the peak position and the chemical shift values. All NOESY peaks with one unique initial assignment are treated as unambiguous distance restraints (Güntert, 2004). In contrast, NOESY peaks with more than one assignment give rise to ambiguous upper distance restraints, and because most NOE peaks cannot be assigned unambiguously just on the basis of chemical shift information, more filtering criteria are applied.

Assessment criteria evaluate each initial assignment possibility according to its presence in other symmetry related positions, and based on assignment possibilities of other cross peaks between the same or adjacent residues (network-anchoring) without relying on the 3D structure, and compatible with the short range covalent distance restraints. The probability from network anchoring ($P_{network}$) has a big impact in the first cycle of CYANA structure calculation as no structure-based criteria can be applied yet. Overall the network anchoring functions initially to lower the ambiguity of NOE cross peak assignments.

After that, each initial assignment possibility for ambiguous NOESY cross peaks is assessed based on the agreement probability with the 3D ensemble structures from the preceding cycle ($P_{structure}$). This probability function calculates the number of conformers in the structure bundle where the distance is shorter than the upper bound, with an acceptable value of violation which is set high in the early cycles and decreases to almost zero in the last one. The first CYANA cycle is excluded here as no preliminary structure is generated yet, thus $P_{structure} = 1$.

Only assignment possibilities for which the combined probability of the three previous weighting factors is higher than a defined threshold are acceptable, and are retained to use in subsequent steps to generate distance constraints, as described in Equation 5.1. Any NOESY assignments that fail this test are eliminated from the assignment list as false NOESY cross peaks. In addition, upper distance bounds are calibrated according to the volume of NOE cross peaks which are extracted from NOESY peak lists.

$$P_{total} = P_{shifts} \times P_{network} \times P_{structure} \geq P_{min} \quad \text{(Equation 5.1)}$$

In the first cycle for structure determination before applying the 3D structure bundle based filtering for the NOE assignments, there is a big chance for false distance restraints to be selected from the noise and artifact peaks in the input peak lists. Therefore a constraint combination criterion is implemented in the first two cycles of CYANA to eliminate structural distortion that arises from erroneous NOE assignments (Herrmann et al., 2002a). The constraint combination consists of making random sets of two or more distance restraints, and requiring that at least one member of each set be satisfied. The aim is to reduce the risk that incorrect restraints will distort the structure calculation, while still allowing correct restraints to operate. The restraints combination thus helps to lower the impact of erroneous distance restraints on the resulting structure (Güntert, 2003).

After that, the resultant distance restraints are used to calculate an ensemble of structures, which are then added to the input data for the following structure calculation cycle, calculated using simulated annealing with torsion angle dynamics. In the final cycle of structure calculation the NOE assignments are taken from cycle 7 and then an additional filtering is applied to ensure that every NOE has one assignment to a single pair of hydrogen atoms, otherwise it is discarded from the input for the structure calculation.

These steps are described in detail in papers from Güntert et al. It is worth adding that the CYANA program itself comes with little documentation, meaning that it is not straightforward to work out what the program is doing, whether the input is correct, and what the output files mean.

5.2 Initial structure calculation attempt by CYANA

5.2.1 Data requirements

The automated NOE assignment followed by structure calculation of SH2 protein was executed using a standard CYANA macro (version 3.98.5) CALC.cya in the demo folder under the auto directory. A first calculation was done using minimum basic conformational constraint data (distances and dihedral angles) to check the reliability of the NOESY cross peak assignment from unrefined NOE peak lists, as the NMR structure calculation primarily depends on distance information (Braun, 1987).

The required input data listed in the CYANA macro for the structure calculation are illustrated in Figure 5.2, and consist of the protein amino acid sequence, a complete chemical shift assignment list, experimental NOE peak lists from multidimensional spectra, and dihedral angles as conformational constraints from the TALOS-N program (external source).

```
peaks            := C13NOESY.peaks,N15NOESY..peaks,C13NOESYaro.peaks     # NOESY peak lists in XEASY format
prot             := Chemical_shifts.prot                                  # names of chemical shift list(s)
Restraints       := Diherdal_angles.aco                                   #  additional (non-NOE) restraints
tolerance        := 0.04, 0.03, 0.45                                      # shift tolerances: H, H', C/N', C/N
#calibration_constant:=6.7E5,8.2E5,8.0E4                                  # calibration constants, automatic if empty
structures       := 100,20                                                # number of initial, final structures
steps            := 10000                                                 # number of torsion angle dynamics steps
randomseed       := 434726                                                # random number generator seed

noeassign peaks=$peaks prot=$prot autoaco
```

Figure 5.2. CYANA macro file (*CALC.cya*) for automated NOE assignment and structure calculation. The macro script specifies the defined parameters and the input data for the combined calculation run. The first three lines show the input data lists; the 3D NOESY spectra lists of C13 NOESY for aliphatic residues, N15 NOESY, and C13 NOESY for aromatic residues, chemical shifts list (*Chemical_shifts.prot*), and torsion angle restraints (*Dihedral_angles.aco*). The following lines (4th, 6th, 7th,and 8th) detail the defined parameters; the tolerance values for the chemical shift matching is 0.04 ppm for 1H dimension, 0.03 ppm for 15N or 13C bound 1H dimension, 0.45 ppm for both dimensions 15N or 13C. The calculation started with 100 conformers generated from random torsion angle values, and then 20 conformers which have lowest target function value represent the final structure. Each conformer is generated after 10000 torsion angle dynamics steps. The random number generator for random torsion angle values and initial velocities for torsion angle dynamics is set with a seed value of 434726. The *noeassign* command is used on the above specified peak and chemical shift lists, and *autoco* is used for the automatic generation of torsion angle restraints to favour regions of the Ramachandran plot.

Identification of NOE signals in the multidimensional spectra was done automatically as described in section 4.2.1 using the automated peak picking algorithm in Felix2007, however this time before peak picking, the contouring level was decreased to pick more low intensity peaks. That was because it was clear from the CYANA automated resonance assignment (section 4.2.1) that lots of peaks were missing from NOE spectra mainly because of strong peak overlap and missing low intensity peaks. This observation confirmed the fact that the automated peak picking algorithm does not function well with complex spectra with a large dynamic range such as NOESY (Güntert, 2009). The experimental NOESY peak lists for SH2 structure calculation: ^{15}N NOESY, ^{13}C NOESY for aliphatic atoms, and ^{13}C NOESY for aromatic atoms, were sorted in XEASY format which contains peak positions and volumes.

Because the main structural restraints for the CYANA structure calculation are upper distance restraints derived from automated NOE assignment, this requires the knowledge of proton chemical shifts from which the NOEs are generated (Vögeli et al., 2016). Therefore as input requirement data the chemical shift values of SH2 were assigned automatically by CYANA (3.98.5) using refined and unrefined peak lists, as explained in detail in Chapter 4, section 4.2.2. The automated resonance assignments were then manually corrected according to the resonance assignments made manually using interactive methods before they were used as input data for the automated NOESY assignment and structure calculation. The majority of the automated chemical shift assignment is correct which helps to reduce the ambiguity of cross peak assignment, although it did include a large number of noise and artefact signals in the NOESY peak lists. Tests made at this stage showed that it is highly beneficial to impose manual corrections on the automated NOE assignments, as this significantly reduced the number of incorrect assignments made.

In addition, dihedral angle restraints (ϕ and ψ) were generated according to the given backbone chemical shifts in the SH2 protein using the TALOS-N program, described in detail in Section 4.3.1. The TALOS-N predicted angles (listed in pred.tab) were converted into torsion angle restraints using a CYANA macro (TalosAngleRestraints.cya) to use as input restraints complementary to the NOE distance restraints for the NMR structure calculation. The macro reads the 93 pairs of dihedral angles (ϕ and ψ) that are classified as strong or generous. Error values are given a default value of double the standard deviation listed by TALOS-N (Cartwright, 2015). The predicted dihedral angles of proline residues were excluded from the CYANA torsion angles list, even for those classified as strong, namely 98, 100 and 116.

5.2.2 Results of the first structure calculation

Seven iteration cycles of automated NOESY assignment followed with structure calculation were carried out using the input data sets as described in the above section. The information for the first CYANA structure calculation for SH2 protein is summarised in Table 5.1, including the NOE assignment details of each individual cycle.

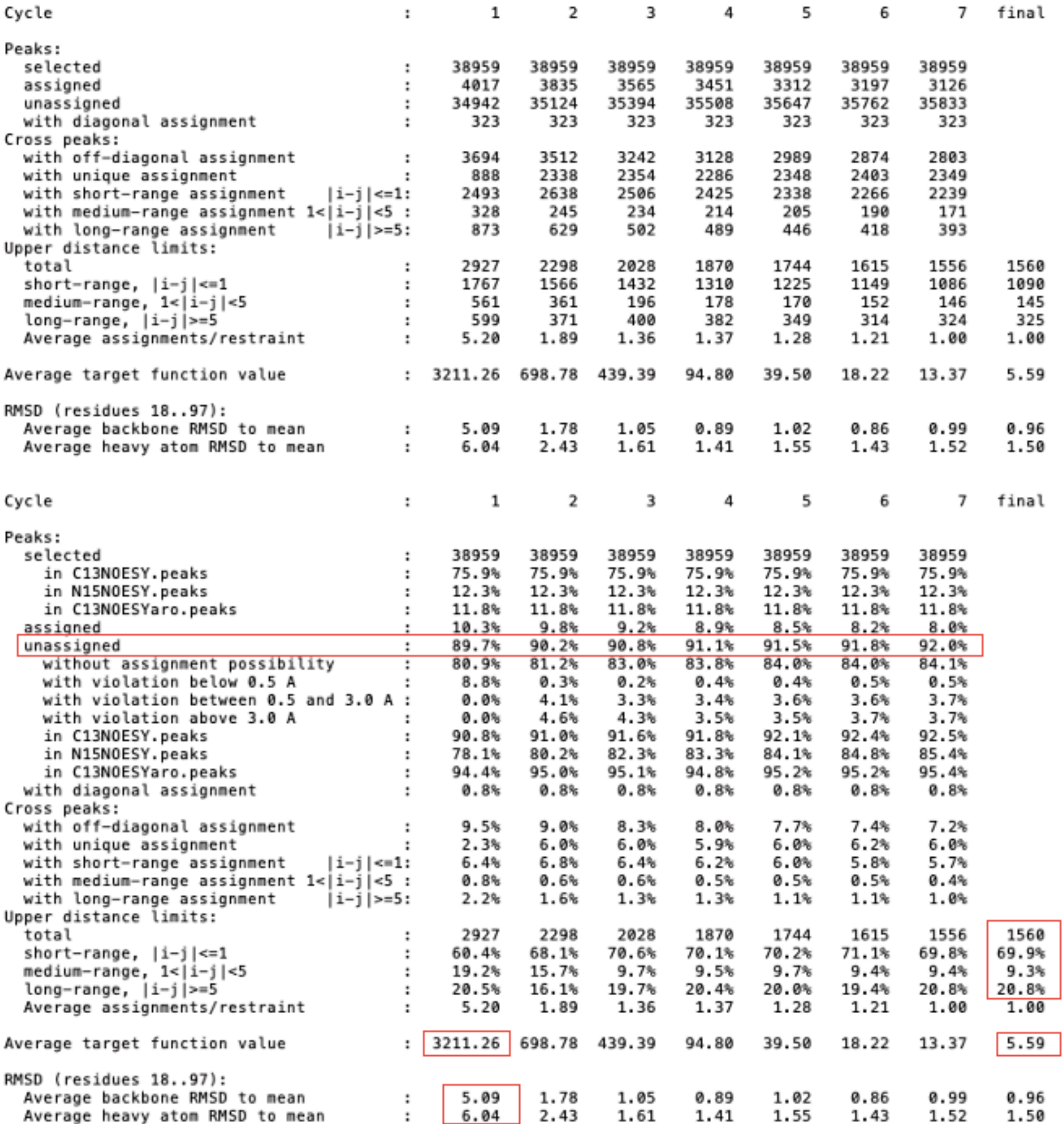
The automated calculation was applied using raw NOESY peak lists. From all selected NOESY peaks (38959), from first towards last cycles, between 10.3% and 8% were assigned and converted to upper distance restraints, and then used for the structure calculation.

As a consequence of applying additional criteria for NOE assignment, which depend on the 3D structures calculated from the preceding cycle, there was a big decrease in average number of possible assignments per cross peak from about 5 in cycle 1 to 1.9 in cycle 2 and continued to cycle 6 with only just over one. Moreover applying more filtering criteria has an impact on the total number of unambiguously assigned NOESY peaks, which reduced from 2927 in cycle 1 to 1556 in cycle 7. In the final cycle, 1560 unambiguous NOEs were defined, comprising 20.8% of upper long range distances with 69.9% and 9.3% of short and medium ranges distances, respectively, as shown in Table 5.1.

The average target function value of CYANA quantifies the agreement between the given constraints and the calculated structure (Güntert, 2003; Vögeli et al., 2016). In this structure calculation for SH2 protein the presence of a high number of artifact peaks in the input peak lists is revealed in a big value of target function for the three early cycles of the calculation. although there was a notable decrease in the target function for the 20 lowest energy conformers starting from cycle 4 towards the final cycle with 5.59 Å^2 (Table 5.1). The precision of the SH2 structure determination in the form of the root mean square deviation (RMSD) average values of backbone and heavy atoms within the final bundle conformers started to improve gradually from cycle 1 to subsequent cycles as the criteria for accepting assignments and NOE constraints were implemented in advanced cycles of the calculation. Although cycle 2 showed good convergence within the bundle of structures, a slightly better defined bundle of structures appeared in the folded part of the protein (18-97 a.a) in cycle 3 and continued to achieve convergence, with the last cycle having a RMSD 0.96 Å and 1.50 Å for the backbone and sidechain respectively. The progress of the NMR structure determination of SH2 protein during the CYANA cycles is shown in Figure 5.3.

A reasonable Ramachandran distribution resulted from the final bundle of structures for SH2; 75.8% of residues in the most favoured, 22.4% in additional allowed regions, 0.8% in generously allowed regions and 0.9% in disallowed regions. By comparison to good structures, this is however still poor.

Table 5.1. Summary table of structure calculation for SH2 protein, based on automated NOESY assignment using CYANA 3.98.5. The first half of the table was created with the cyanatable command and shows results of each cycle obtained with raw NOESY peak lists for SH2 protein, with absolute values. The second half of the table was generated with cyanatable -lp command and shows the obtained result of each cycle with percentage values. Both tables present the same data comprising one column for each CYANA cycle (1 to 7) with the final cycle of structure calculation. The important numbers that assess the outcome of the structure determination of SH2 are surrounded in red.



```
Cycle                                     :       1       2       3       4       5       6       7   final

Peaks:
  selected                                :   38959   38959   38959   38959   38959   38959   38959
  assigned                                :    4017    3835    3565    3451    3312    3197    3126
  unassigned                              :   34942   35124   35394   35508   35647   35762   35833
  with diagonal assignment                :     323     323     323     323     323     323     323
Cross peaks:
  with off-diagonal assignment            :    3694    3512    3242    3128    2989    2874    2803
  with unique assignment                  :     888    2338    2354    2286    2348    2403    2349
  with short-range assignment    |i-j|<=1:    2493    2638    2506    2425    2338    2266    2239
  with medium-range assignment 1<|i-j|<5 :     328     245     234     214     205     190     171
  with long-range assignment     |i-j|>=5:     873     629     502     489     446     418     393
Upper distance limits:
  total                                   :    2927    2298    2028    1870    1744    1615    1556    1560
  short-range, |i-j|<=1                   :    1767    1566    1432    1310    1225    1149    1086    1090
  medium-range, 1<|i-j|<5                 :     561     361     196     178     170     152     146     145
  long-range, |i-j|>=5                    :     599     371     400     382     349     314     324     325
  Average assignments/restraint           :    5.20    1.89    1.36    1.37    1.28    1.21    1.00    1.00

Average target function value             : 3211.26  698.78  439.39   94.80   39.50   18.22   13.37    5.59

RMSD (residues 18..97):
  Average backbone RMSD to mean           :    5.09    1.78    1.05    0.89    1.02    0.86    0.99    0.96
  Average heavy atom RMSD to mean         :    6.04    2.43    1.61    1.41    1.55    1.43    1.52    1.50


Cycle                                     :       1       2       3       4       5       6       7   final

Peaks:
  selected                                :   38959   38959   38959   38959   38959   38959   38959
    in C13NOESY.peaks                     :   75.9%   75.9%   75.9%   75.9%   75.9%   75.9%   75.9%
    in N15NOESY.peaks                     :   12.3%   12.3%   12.3%   12.3%   12.3%   12.3%   12.3%
    in C13NOESYaro.peaks                  :   11.8%   11.8%   11.8%   11.8%   11.8%   11.8%   11.8%
  assigned                                :   10.3%    9.8%    9.2%    8.9%    8.5%    8.2%    8.0%
  unassigned                              :   89.7%   90.2%   90.8%   91.1%   91.5%   91.8%   92.0%
    without assignment possibility        :   80.9%   81.2%   83.0%   83.8%   84.0%   84.0%   84.1%
    with violation below 0.5 A            :    8.8%    0.3%    0.2%    0.4%    0.4%    0.5%    0.5%
    with violation between 0.5 and 3.0 A :    0.0%    4.1%    3.3%    3.4%    3.6%    3.6%    3.7%
    with violation above 3.0 A            :    0.0%    4.6%    4.3%    3.5%    3.5%    3.7%    3.7%
    in C13NOESY.peaks                     :   90.8%   91.0%   91.6%   91.8%   92.1%   92.4%   92.5%
    in N15NOESY.peaks                     :   78.1%   80.2%   82.3%   83.3%   84.1%   84.8%   85.4%
    in C13NOESYaro.peaks                  :   94.4%   95.0%   95.1%   94.8%   95.2%   95.2%   95.4%
  with diagonal assignment                :    0.8%    0.8%    0.8%    0.8%    0.8%    0.8%    0.8%
Cross peaks:
  with off-diagonal assignment            :    9.5%    9.0%    8.3%    8.0%    7.7%    7.4%    7.2%
  with unique assignment                  :    2.3%    6.0%    6.0%    5.9%    6.0%    6.2%    6.0%
  with short-range assignment    |i-j|<=1:    6.4%    6.8%    6.4%    6.2%    6.0%    5.8%    5.7%
  with medium-range assignment 1<|i-j|<5 :    0.8%    0.6%    0.6%    0.5%    0.5%    0.5%    0.4%
  with long-range assignment     |i-j|>=5:    2.2%    1.6%    1.3%    1.3%    1.1%    1.1%    1.0%
Upper distance limits:
  total                                   :    2927    2298    2028    1870    1744    1615    1556    1560
  short-range, |i-j|<=1                   :   60.4%   68.1%   70.6%   70.1%   70.2%   71.1%   69.8%   69.9%
  medium-range, 1<|i-j|<5                 :   19.2%   15.7%    9.7%    9.5%    9.7%    9.4%    9.4%    9.3%
  long-range, |i-j|>=5                    :   20.5%   16.1%   19.7%   20.4%   20.0%   19.4%   20.8%   20.8%
  Average assignments/restraint           :    5.20    1.89    1.36    1.37    1.28    1.21    1.00    1.00

Average target function value             : 3211.26  698.78  439.39   94.80   39.50   18.22   13.37    5.59

RMSD (residues 18..97):
  Average backbone RMSD to mean           :    5.09    1.78    1.05    0.89    1.02    0.86    0.99    0.96
  Average heavy atom RMSD to mean         :    6.04    2.43    1.61    1.41    1.55    1.43    1.52    1.50
```

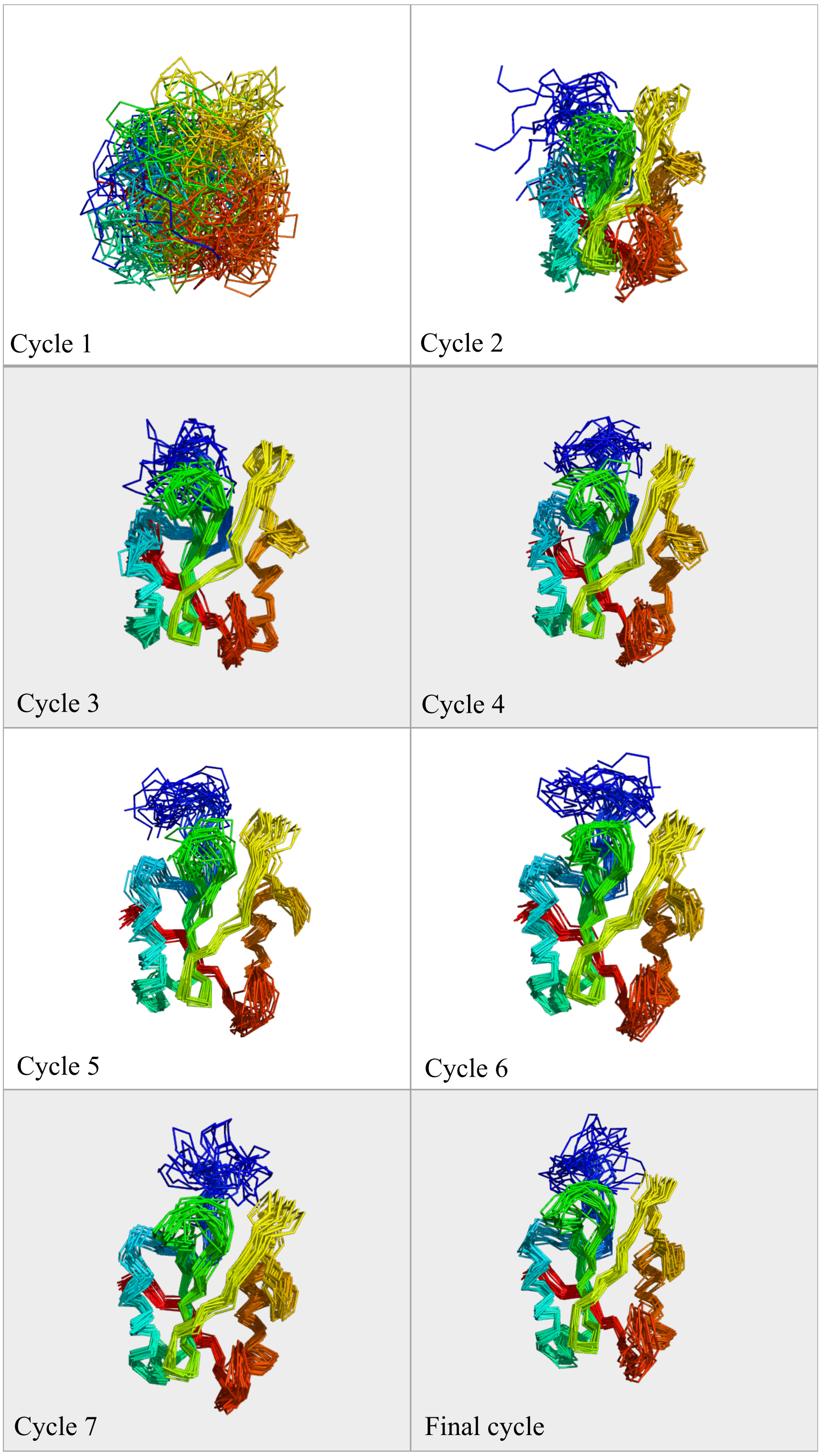


Figure 5.3. The progress of the NMR structure determination of SH2 protein obtained by CYANA program in seven iterative cycles of automated NOESY assignment followed by structure calculation. The ensemble structure shows the 20 conformers that have the lowest residual target function values for each cycle (1 to 7) including the final bundle structure (overlay 1-118 a.a). Structures overlayed are displayed using Pymol program with rainbow colour; the colour from the N to the C terminus in dark blue to red for clarity of the individual helices and strands. Number on the left-bottom of each structure refers to CYANA cycle number from 1-7.

5.2.3 Reliability of the first structure calculation

The conventional criteria that are normally used to measure the accuracy of the calculated NMR structure such as low value of the backbone root mean square deviation within the final ensemble structure (RMSD), small number of restraint violations, and small value of the total energy of structure are less informative with the automated structure calculation method that discards ambiguous NOESY restraints (Schmidt et al., 2013; Güntert, 2003). Other assessment criteria for CYANA results that are established and discussed in Herrmann et al. 2002b allow checking the reliability of the calculated structure on the basic of automated NOESY assignment.

The first general criterion is to check the percentage of the unassigned cross peaks, which should not exceed 20% among all input NOEs peak lists through CYANA cycles. As shown in Table 5.1, between 89% and 92% of NOE peaks were discarded across CYANA cycles in the first structure calculation. The percentage of unassigned cross peaks indicates that the peak list contains a big number of artifact peaks. This is expected because no manual inspection was applied to remove the noise peaks from the peak list. In retrospect, it would have been better to manually remove more of the obvious noise and artefact peaks before starting CYANA, and probably to raise the contour threshold slightly to eliminate some of the noise peaks.

The second important criterion is the average residual target function value of the first and the last CYANA cycles which should be below the acceptable value for NMR structure < 250 $\mathring{A}^2$ and < 10 $\mathring{A}^2$, respectively. In this structure calculation the target function value of cycle 1 is high: about 3211 $\mathring{A}^2$, which mostly reflects a severe inconsistency in between the experimental constraints data and the first calculated structure. On the other hand the target function value of the final cycle is 5.59 $\mathring{A}^2$, well in the range below the limiting value, which reveals that the final calculated structure fulfils the conformational constraints input data.

Furthermore it has been shown that the ability to obtain a correctly folded structure in the initial cycle of the structure calculation has a big impact on the accuracy of the final calculated structure using automated NOESY assignment (Güntert, 2003). The importance of achieving convergence in the first cycle of the structure calculation comes from the fact that each subsequent cycle depends on the obtained 3D structure from the preceding cycle, excepting cycle 1. Therefore inconsistent NOE peaks with the obtained 3D structure of the previous cycle will remain unassigned also in the following cycle, and thus the correctly folded structure should appear in the first cycle as a consequence of assigning all essential NOE peaks.

The CYANA calculation result of SH2 protein shows that the quality of NOE assignment based on the given input data was not sufficient to achieve convergence in cycle 1 because the average RMSD of the backbone atoms (of the well-structured part) between the individual conformers to their mean coordinates is 5.59 Å. The big RMSD value for cycle 1 is above the acceptable range (< 3 Å) and indicates insufficient or incomplete input data to build a correct structure in the initial cycle. Again this is likely to be due to a large number of noise and artefact peaks in the initial peak list.

Although the final ensemble structure looks well-defined with a small RMSD value for backbone atoms within the bundle of conformers (~1 Å), there is a large distribution of conformations in loop regions and the N and the C terminal unstructured part of the protein, Figure 5.2. That was expected as there are not enough distance restraint data to characterise these ill-defined regions (Rosato et al., 2013).

Another criterion to evaluate the quality of the automated CYANA calculation is the RMSD drift between the mean atom coordinates of the first and the last cycles for the backbone atoms of the structured region of the SH2 protein. This is 3.3 Å which is slightly above the acceptable value of < 3 Å (Herrmann et al., 2002a).

In conclusion, the ensemble produced by this first set of CYANA calculations looks reasonable (but not good). It would clearly have been preferable in hindsight to clean up the NOE peak list before presenting it to CYANA.

5.3 Set of structure calculations by CYANA

5.3.1 Input data

The initial attempt to calculate the NMR structure of SH2 using constraints obtained from automated structure calculation (and manual chemical shift assignments) did not succeed to achieve convergence in the first cycle, and thus resulted in unsatisfactory final ensemble structures. In the CYANA macro for structure calculation (CALC.cya) there is an option to add an extra constraints list obtained from an external source. Using additional constraints such as hydrogen bonds with the other input data used in the first CYANA calculation (section 5.3) can be sufficient to achieve good convergence in the first cycle of the structure calculation. As discussed above, the resultant structure from cycle 1 is the key for subsequent cycles driven by 3D structure-based assignment, thus finding the correct fold in this cycle is important for the structure calculation to result in a correct structure (Güntert, 2004).

To create a list of hydrogen bond restraints, a direct identification approach was followed in the 2nd and 3rd CYANA structure calculations. This method was based on reported hydrogen bonds generated from the CYANA structure calculation. CYANA lists H-bonds in the final.ovw file that are found in more than 30% of the final calculated ensemble structures. Rounds 2 and 3 of CYANA calculations therefore added these backbone H-bonds iteratively to the input restraints list.

Furthermore an indirect identification method was introduced in the 4th CYANA structure calculation based on inferring residues which are likely to form backbone H-bonds in regular secondary structure, using the final 3D structural models from the preceding 3rd CYANA calculation. A number of residues located on the β strands of the protein were predicted to form backbone hydrogen bonds with nearby residues on adjacent strands. The new hydrogen bonds were added to the H-bond list as extra restraints for the following structure calculation. Figure 5.4 shows all the expected hydrogen bonds in the NMR structure of the SH2 protein, shown as a secondary structure diagram which was derived on the basis of the final 3D bundle of structures (final.pdb) of the 3rd CYANA structure calculation.

Finally, a temperature coefficient method was used to confirm the final H-bond list, to determine whether obtained hydrogen bonds from CYANA calculations agreed with the temperature coefficient data or not. Only the residues that have a temperature coefficient value more positive than -4.5 ppb/K were included in the final CYANA calculation as the extra constraints list. Also the reported hydrogen bonds from CYANA calculation were added to the input restraints list if they were found in more than 90% of final ensemble structures in cases of absent temperature coefficient support evidence. The identification procedure to create extra restraints from H-bonds as a part of NMR structure determination for SH2 protein is shown in Figure 5.1.

Before adding the obtained hydrogen bond list from direct and indirect methods to the CYANA structure calculation, CYANA macro Hbonds.cya was used to generate a pair of upper (2 Å to 3 Å range) and lower (1.80 Å to 2.70 Å range) distance constraints for the given atoms. After that, hydrogen bond upper and lower distance restraints lists were added to the next CYANA structure calculation for SH2 protein as additional restraints with the other input as used in the first CYANA calculation: amino acid sequence of SH2 protein, chemical shift values obtained from manual resonance assignment, multidimensional NOE peak list, dihedral angles list, and upper and lower hydrogen bonds list.

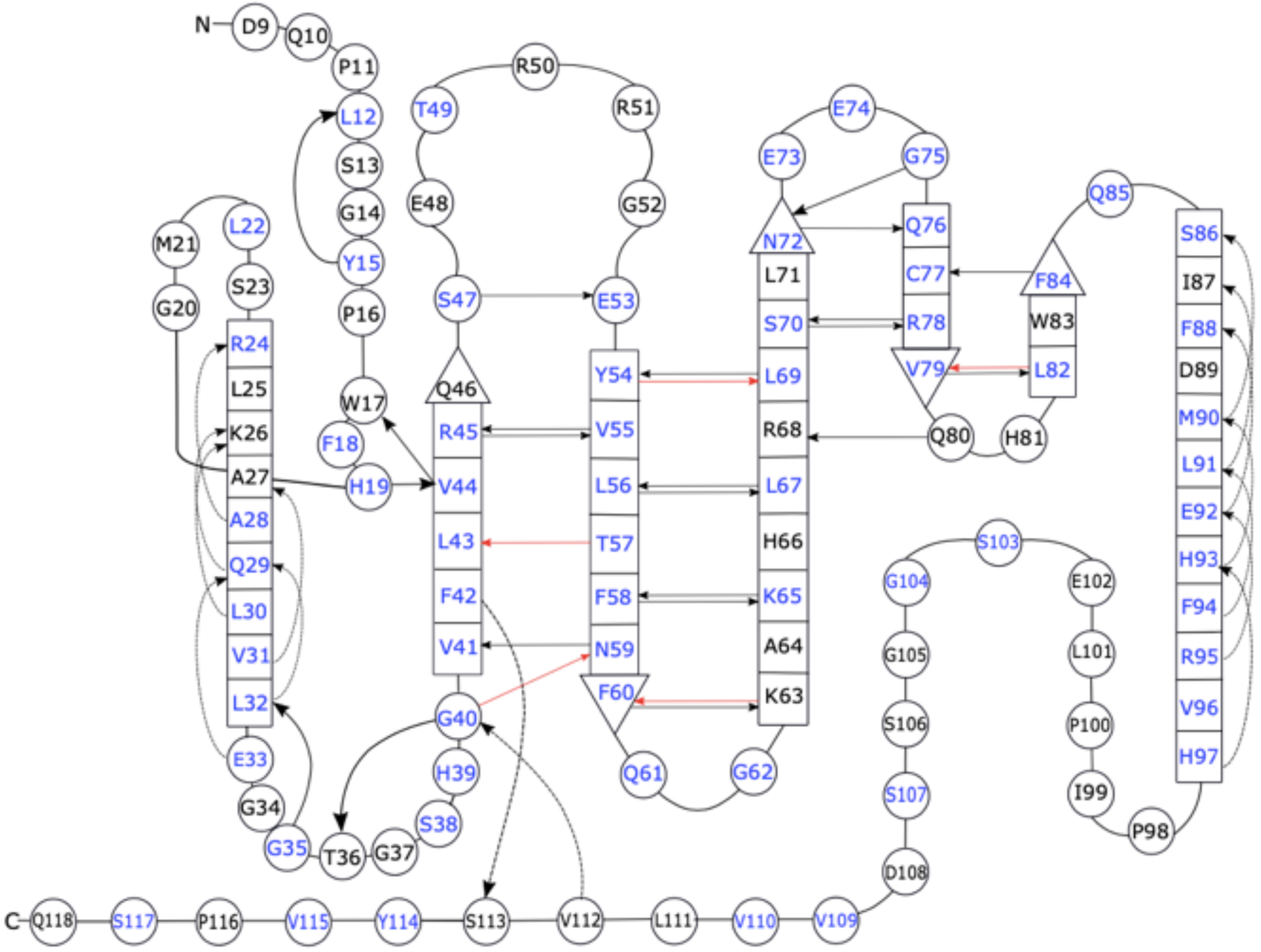


Figure 5.4. A secondary structure bead diagram of SH2 protein. The 2D schema is drawn based on the 3D final structure from the 3rd CYANA structure calculation. Thin arrows display the hydrogen bond between the NH and the CO of the backbone groups, red arrows the H-bonds removed from the 5th CYANA calculation, big bars and arrow represent α-helices and β-strands, respectively.

5.3.2 Results of structure calculations

From the first CYANA structure calculation which was completed as explained above without input H-bond restraints, twenty-one residues were reported to form hydrogen bonds in more than 30% of structures of the final ensemble. All residues were confirmed by temperature coefficient data to be likely to be involved in backbone hydrogen bonds. The obtained H-bonds were used in the following calculation (2nd CYANA) as an additional input list. A similar procedure was followed to find out more hydrogen bonds after the second CYANA structure calculation, and an extra new seven residues forming backbone hydrogen bonds were listed in the output final.ovw file. These new predicted H-bonds were used in the next structure calculation (3rd CYANA).

In the fourth CYANA structure calculation another indirect identification method was introduced to find out more residues participating in intramolecular hydrogen bonding. From the final 3D bundle structures of the preceding structure calculation (3rd CYANA), residues in the β strand polypeptide chains running alongside each other are expected to be linked by hydrogen bonds between their backbone NH and CO groups. Figure 5.5 shows an example of the NH group of Arg 45 pointed to the CO group of Val 55. Although these residues run alongside each other, there were not enough distance restraints to form hydrogen bonds between them during the CYANA calculation. A further fifteen H-bonds were predicted to make in total 43 pairs of upper and lower H-bond distance restraints, which were added to the input data of the 4th CYANA structure calculation.

Adding hydrogen bond restraints to the structure calculation forces the CYANA program to significantly restrict the conformational freedom of the resulting structure using the input H-bonds. Furthermore those H-bond restraints were assigned on the basis of indirect experimental data where there is a chance for incorrect assignment. Therefore the last structure calculation was carried out with more caution to avoid introducing bias in the final CYANA structure. To end the structure calculation, the final 5th CYANA calculation was completed as before but with only using the highly confident hydrogen bonds that were found in more than 90% of the final ensemble structures of the preceding calculation.

The final H-bond restraints list obtained from CYANA calculations is in agreement with the temperature coefficient data (> -4.5 ppb/K). However in addition, V112 and Q80 were reported to form backbone hydrogen bonds with G40 and R68, respectively, because they were found in more than 90% of the final structures in the ensemble, which means it is highly likely that they are genuinely present. These residues were included in the H-bond restraints list although they do not have supporting experimental evidence from the temperature coefficient data. The original analysis of temperature coefficients in proteins (Baxter and Williamson, 1997) found exceptions to the -4.5 ppb/K rule, so this is acceptable. In total the final list of H-bond restraints which were used in the final 5th CYANA structure calculation contained 38 pairs of upper and lower H-bond distance restraints. Out of sixty-eight residues suggested to contribute to backbone hydrogen bond interaction as a donor by the temperature coefficient method, 53% of them were identified from the CYANA structure calculations and used in the calculations. The information of backbone hydrogen bond donors from temperature coefficient data and their defined acceptor atoms from CYANA calculations is listed in Table 5.2.

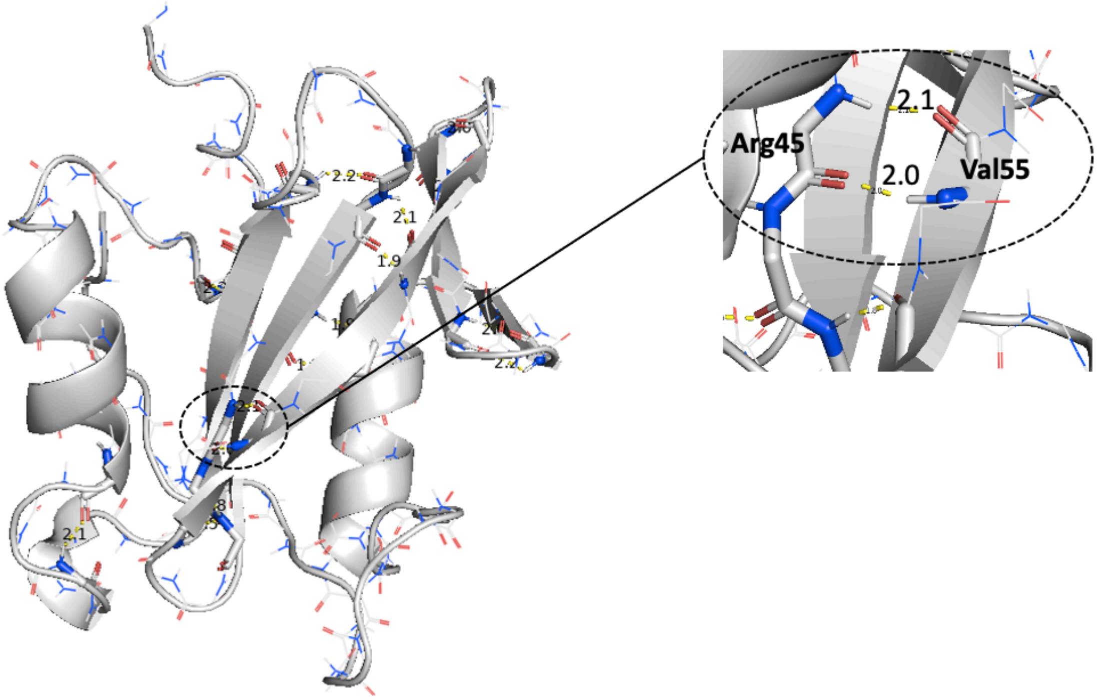


Figure 5.5. 3D structures of SH2 protein resulting from the first CYANA structure calculation displayed using Pymol shows backbone atoms (N, Cα, C_o) of residues in SH2 protein. The circled region (black dashes) focusses on an example of hydrogen bond interaction between backbone amide NH and carboxyl CO of Arg 45 and Val 55 atoms with distance measurements predicted from the secondary structure. N and C letters denote the N and C termini of the protein.

Table 5.2. Secondary structural data for the hydrogen bonds in the SH2 protein structure. Yes/No indicate whether temperature cofficient data (T/C) suggests the backbone amide proton is a hydrogen bond donor or not. Hydrogen bonds restraints used in CYANA calculations (CYANA1, CYANA2, CYANA3, and CYANA4) are listed in final.ovw, or predicted from the secondary structure from a previous calculation as in CYANA4. In the CYANA calculation 5, the H-bond restraints are the only ones that used from the final.ovw of the 4th calculation.

Residue	H-bond						Residue	H-bond					
	T/C	Cyana round						T/C	Cyana round				
		1	2	3	4	5			1	2	3	4	5
D9	No						T49	Yes					
Q10	No						R50	No					
L12	Yes						R51	No					
S13	No				Q10		G52	No					
G14	No						E53	Yes					
Y15	Yes	L12	L12	L12	L12	L12	Y54	Yes			L69	L69	
F18	Yes						V55	Yes	R45	R45	R45	R45	R45
H19	Yes			V44	V44	V44	L56	Yes			L67	L67	L67
M21	No						T57	Yes		L43	L43	L43	
L22	Yes						F58	Yes			K65	K65	K65
S23	No						N59	Yes	V41	V41	V41	V41	V41
R24	Yes						F60	Yes		K63	K63	K63	K63
L25	No						Q61	Yes					
K26	No						G62	Yes					
A27	No						K63	No	F60	F60	F60	F60	
A28	Yes	R24	R24	R24	R24	R24	A64	No					
Q29	Yes	K26	K26	K26	K26	K26	K65	Yes			F58	F58	F58
L30	Yes	K26	K26	K26	K26	K26	H66	No					
V31	Yes	A27	A27	A27	A27	A27	L67	Yes			L56	L56	L56
L32	Yes		Q29	Q29	Q29	Q29	R68	No					
E33	Yes		Q29	Q29	Q29	Q29	L69	Yes			Y54	Y54	Y54
G34	No						S70	Yes		R78	R78	R78	R78
G35	Yes			L32	L32	L32	L71	No					
T36	No						N72	Yes			Q76	Q76	Q76
G37	No						E73	Yes					
S38	Yes	G35	G35	G35	G35	G35	E74	Yes					
H39	Yes	T36	T36	T36	T36	T36	G75	Yes			N72	N72	N72
G40	Yes			N59	N59		Q76	Yes					
V41	Yes						C77	Yes					
F42	Yes	S113	S113	S113	S113	S113	R78	Yes	S70	S70	S70	S70	S70
L43	Yes		T57	T57			V79	Yes			L82	L82	L82
V44	Yes			W17	W17	W17	Q80	No				R68	R68
R45	Yes		V55	V55	V55	V55	H81	No					
Q46	No						L82	Yes			V79	V79	
S47	Yes			E53	E53	E53	W83	No					
E48	No						F84	Yes	C77	C77	C77	C77	C77

Residue	H-bond					
	T/C	Cyana round				
		1	2	3	4	5
Q85	Yes					
S86	Yes					
I87	No					
F88	Yes					
D89	No					
M90	Yes	S86	S86	S86	S86	S86
L91	Yes	I87	I87	I87	I87	I87
E92	Yes	F88	F88	F88	F88	F88
H93	Yes	M90	M90	M90	D89	D89
F94	Yes	L91	L91	L91	M90	M90
R95	Yes	E92	E92	E92	E92	E92
V96	Yes					
H97	Yes	H93	H93	H93	H93	H93
I99	No				V109	
L101	No					
E102	No					
S103	Yes					
G104	Yes					
G105	No					
S106	No					
S107	Yes					
D108	No					
V109	Yes					
V110	Yes					
L111	No					
V112	No	G40	G40	G40	G40	G40
S113	No					
Y114	Yes					
V115	Yes					
S117	Yes					
Q118	No					

5.3.3 Evaluation of CYANA structure calculations

As discussed in Schmidt et al., 2013 and Güntert, 2003, a number of assessment criteria are established to assess the success of CYANA structure calculations: the total number of discarded NOE peaks, the average target function of the first and last cycle, the RMSD radius value of the first cycle, and the RMSD drift value.

Through CYANA structure calculations (CYANA 2nd, 3rd, 4th, and 5th) the majority of input NOE peaks were unassigned: the percentage of discarded peaks (between 90% to ~92%) is much higher than the acceptable number (< 20%) of total NOE peaks (Güntert, 2004). This indicates that the unrefined NOE peak lists contained many more artifact peaks than real peaks, which can be avoided if a manual intervention is applied to remove noise and artifact peaks. In addition, rejecting a big number of cross peaks could indicate imperfections of the experimental NMR input data in term of error in picked or inaccurately positioned peaks. This will cause inconsistency between the input cross peaks and the given chemical shift values within the defined tolerance ranges.

Figure 5.6 (a) shows in all CYANA calculations the number of assigned NOE peaks in cycle 1 was almost the same, then started to reduce in subsequent cycles because of applying more criteria for accepting cross peaks. On the other hand as a consequence of adding more input restraints data in the 2nd , 3rd, 4th and 5th calculations the total number of assigned NOE peaks in the final cycle increased between 7% to 9% comparing with the first CYANA calculation which was completed without H-bond restraints. Assigning more NOEs in the final cycle led to a slight increase in the percentage of the long range upper distance restraints in the final structure, Figure 5.6 (b). That means that adding more input restraint data to the CYANA structure calculation supports the CYANA program in assigning more NOE peaks, and was therefore a useful step.

No substantial change occurred to the overall target function value of the first cycle in all CYANA structure calculations, which was higher than the reasonable starting value > 250 Å^2. The presence of larger number of noise and artifact peaks in the input lists is very likely to be the origin of the big value of the target function in the early cycles. However, the overall target function value of the final cycle for all structure calculations was within the acceptable average range < 10 Å^2, which reveals that the resultant final structure of all calculations meets the conformational constraint data.

Achieving convergence in the first cycle is important as discussed earlier in Section 5.3.3, as it has a big impact on the accuracy of the final calculated structure using automated NOESY assignment. In all calculations there was a slight decrease in the RMSD value from the mean coordinates of the backbone in cycle 1, however the decrease was not enough to achieve the optimal convergence, as the RMSD is still above 3 Å, Figure 5.6 (c). In all calculations the RMSD radius reduced towards the last cycle to end with a small average value within the final structure ensemble (between 0.96 Å and 0.36 Å), however this value cannot serve as indicator for the structure quality as it reflects the precision not the accuracy. Figure 5.5 (d) indicates that the RMSD drift value started to decrease with using more restraints in the input data, to achieve 2.13 Å in the last CYANA calculation as it should be (<3 Å). Using additional hydrogen bond constraints as input helped to improve the overall convergence in the last cycle but is still not enough to achieve comparable convergence in the early cycles.

The Ramachandran statistics of the last structure calculation for SH2 protein comparing with the first calculation showed a good improvement in terms of increasing the percentage of residues in most favoured regions from 75.8% to 80.9% and decreasing the percentage of residue in generously allowed and disallowed regions from 1.7% to 0.1% as illustrated in Figure 5.7. Fowler et al., 2020 concluded that Ramachandran quality is a good measure of the accuracy of NMR structures, suggesting that including of hydrogen bond restraints has produced a significant improvement in the resulting structures.

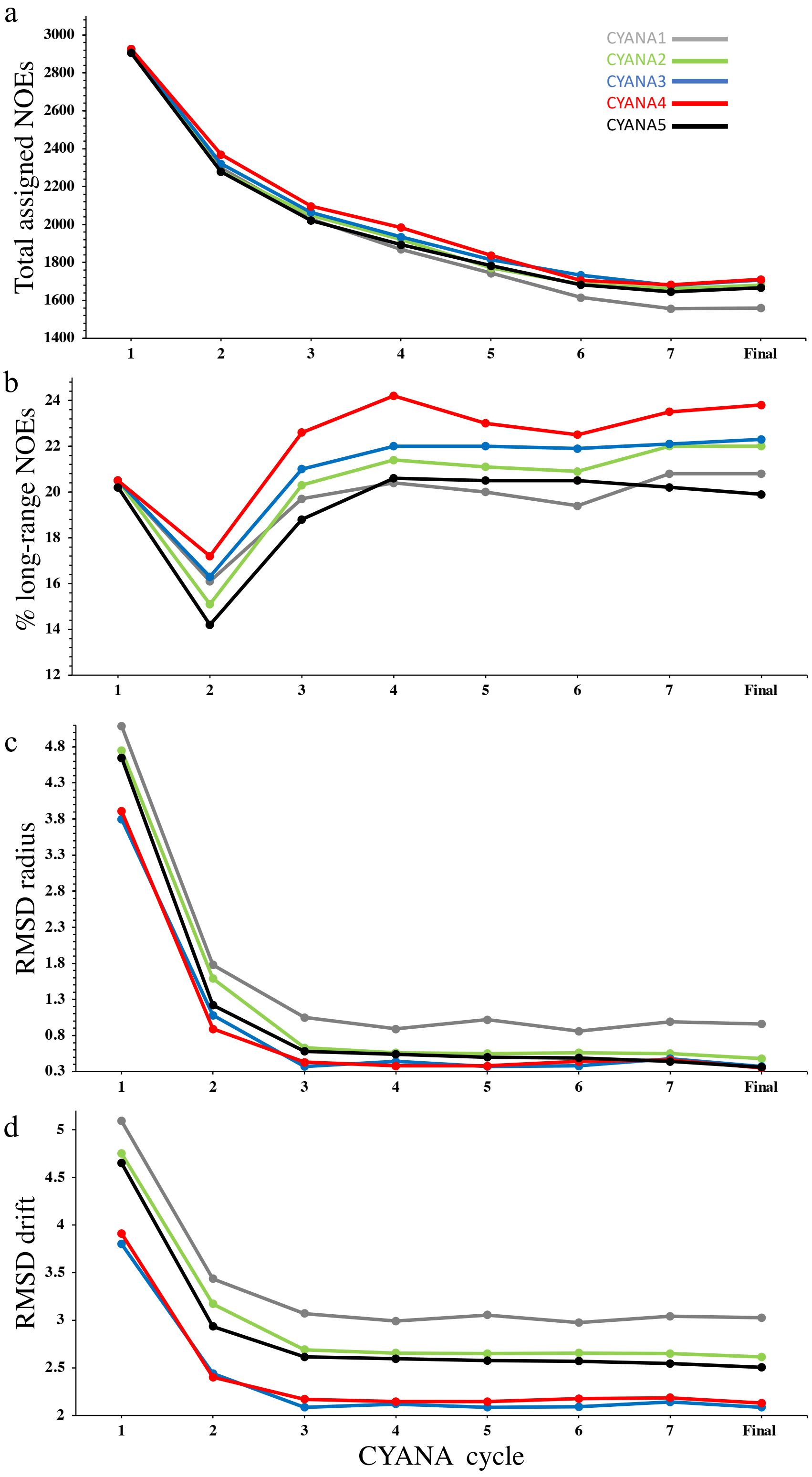
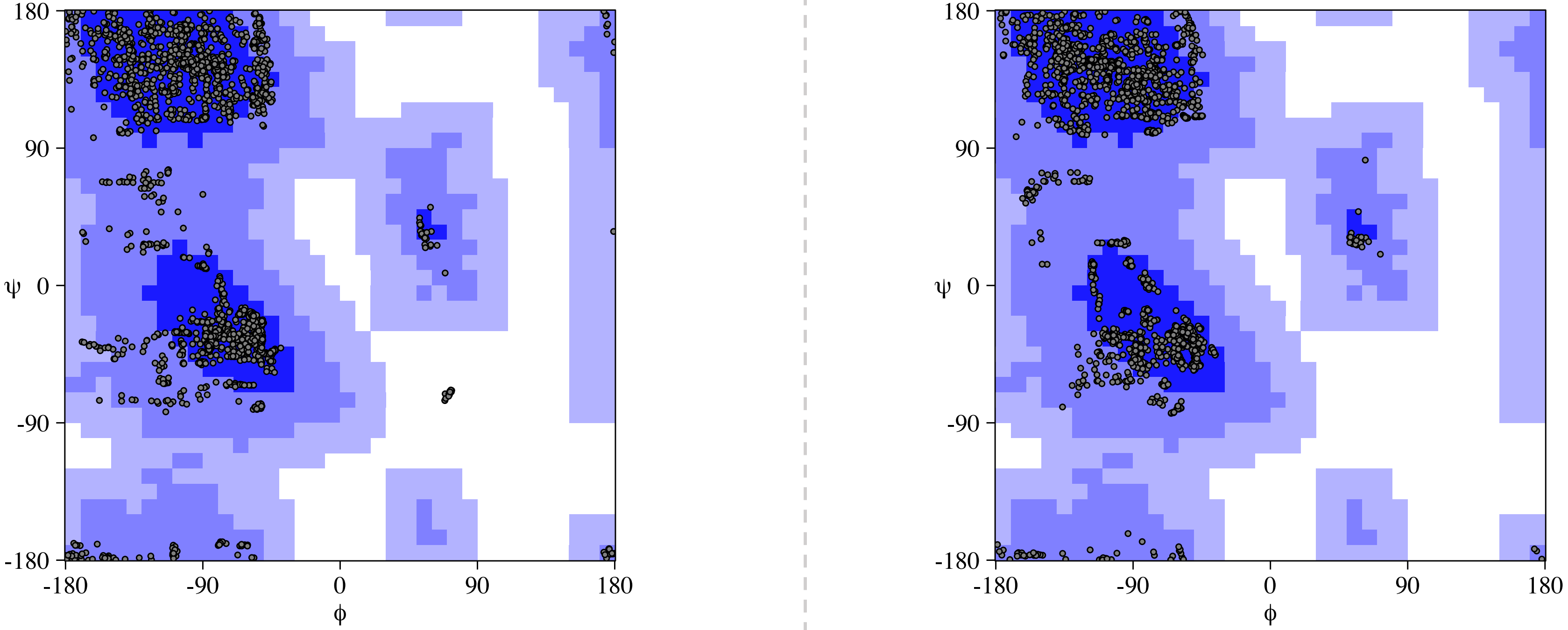


Figure 5.6. Evolution of characteristic parameters for NMR structures in seven cycles with the final cycle of CYANA structure calculation for SH2 protein. (a) The percentage of the total number of assigned NOE peaks. (b) the percentage of the upper long distance restraints from the total assigned NOE. (c) the backbone average RMSD drift to mean (calculated as the RMSD between the mean coordinates of the bundle of conformers after first cycle and each following cycle). (d) the RMSD drift value between each cycle and the final cycle.



75.8% in most favored regions
22.4% in additionally allowed regions
0.8% in generously allowed regions
0.9% in disallowed regions

80.9% in most favored regions
18.9% in additionally allowed regions
0.1% in generously allowed regions
0.0% in disallowed regions

Figure 5.7. Ramachandran plot shows distribution of phi (ϕ) and psi (ψ) dihedral angles for the residues in the final ensemble structures of SH2 protein. On the left is the 1st CYANA 1 structure calculation, and on the right is the final 5th CYANA structure calculation.

5.4 Structure determination and refinement by CNS

5.4.1 Input data

Standard simulated annealing calculations (such as those carried out by CYANA) aim to fold the protein into the correct fold, and get as close as possible to the correct structure. The force fields used in these calculations are aimed at rapid calculation, and at using the experimental data to guide the calculation to produce structures as quickly as possible. They are not designed to produce energetically good protein structures. Thus for example, CYANA uses torsion angle dynamics and does not permit changes to bond lengths or bond angles. It does not include any terms for coulombic interactions or hydrogen bonds, and there is no van der Waals attraction term, simply a van der Waals repulsion to prevent atoms from overlapping. All of this means that one should not expect the structures produced directly from CYANA to look like geometrically good proteins, and indeed they do not. The energies calculated by CYANA are essentially restraint violations (plus van der Waals overlaps etc) which means the free energies are always positive. Clearly this does not represent "real" proteins. It is well established (Ikeya et al., 2016) that in order to produce good structures, it is important to use more realistic force fields. In particular, recent work (Fowler et al., 2020) emphasises that NMR structures are a joint optimisation of experimental restraints and geometrical factors; and that because NMR produces relatively a small number of fairly imprecise restraints, the knowledge-based geometrical terms are very important to produce accurate structures. Hence it is important to optimise the structures from CYANA in more realistic force fields. In particular it is important to optimise the structures using explicit solvent in the simulations, to get hydrogen bonds more realistic. This requires the addition of a large number of water molecules to the system and therefore makes the calculations much more lengthy. It also produces completely different energetics, because the energies tend to be dominated not by the restraint violation terms but by factors such as bond energy and coulombic terms (Linge et al., 1999). The energies are more realistic, and in particular are usually negative, as they should be for a folded protein.

The structure calculation for SH2 and the following refinement procedure in solvent were performed using crystallography and NMR system program (CNS) based on a simulated annealing and restrained molecular dynamics protocol (Brünger et al., 1998). Using CNS program to re-generate SH2 structures expands the option of adding more restraints such as sidechain dihedral angles which are restricted when using CYANA software.

Structures were re-calculated and refined in subsequent rounds using a collection of restraints data: a list of unambiguous NOE distances, backbone and sidechain dihedral angles, and hydrogen bonds. The full set of unambiguous NOESY restraints derived from the last 5th CYANA structure calculation were used. In total 1667 distance restraints comprises: 68.7 % of short-range, 11.4 % of medium-range, and 19.9 % of long-range, Figure 5.8. These NOE upper distance lists were converted to be in XPLOR format to use in CNS calculation, by cyana2cns.cya macro.

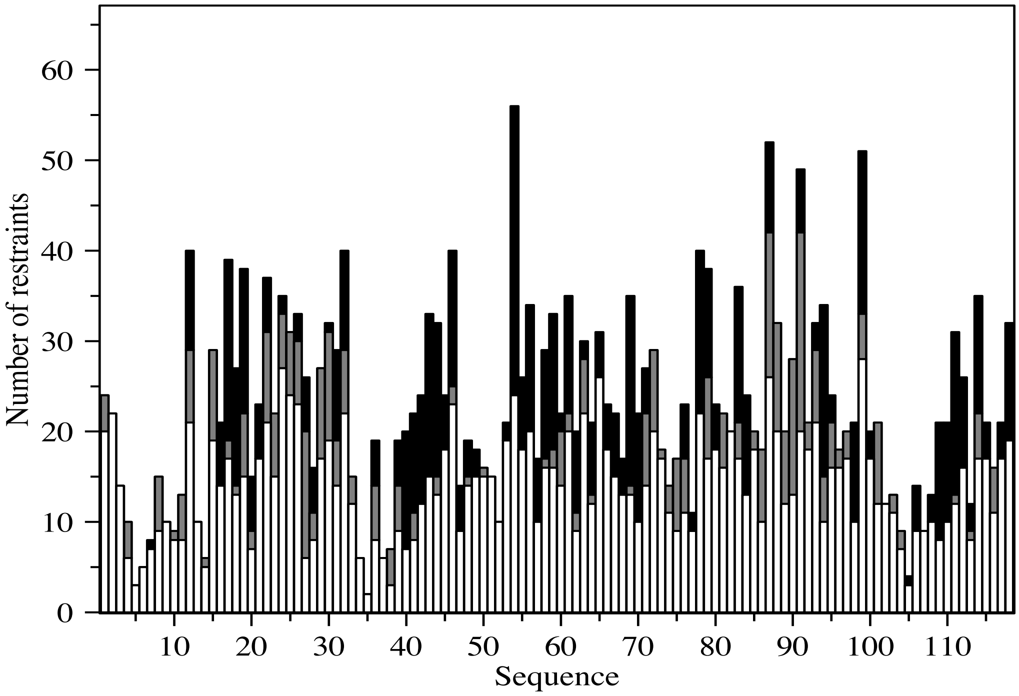


Figure 5.8. Distribution of NOE distance restraints for each SH2 residue from the last 5th CYANA structure calculation. From the bottom to the top; short-range (intra-residue, white), medium-range (dark grey), and long-range (black).

In addition, Section 4.3.1 discussed in detail the dihedral angle prediction for SH2 from the TALOS-N program. A script within TALOS-N software, talos2xplor.com, was used to read unambiguous ϕ and ψ angles (classified by TALOS as strong and generous) which are listed in pred.tab and convert them to torsion angle restraints. 93 pairs of angle restraints were converted to XPLOR format for the structure calculation; the values of ϕ and ψ were set as their actual predicted values from TALOS-N and their widths were set to either 20° or double of the predicted standard deviation (DPHI and DPSI), whichever is the largest value (Cartwright, 2015). In addition TALOS-N reported χ^1 sidechain torsion angles which are listed in predChi1.tab. From these predicted angles, fifty-four unambiguous angles were added as further information to the previous torsion angles list.

The hydrogen bond restraints list consists of 38 hydrogen bonds between backbone NH-CO groups that were used as input in the last 5th CYANA calculation. Residues T49, Y54, E73, E74 and Q76 are suggested by the temperature coefficient method to be involved in backbone hydrogen bonded interactions as donors. Those residues are not reported from previous CYANA structure calculations or the structural model, therefore their acceptor partners are still not defined. Because they are located on loops they should be treated with some caution due to the inherent ambiguity in the assignment of the hydrogen bond acceptor. In this case one could consider generating ambiguous restraints for these amides, for example a short distance to any oxygen atom in the protein, or (more simply) to any oxygen atom within the same loop. Thus they can be given multiple possibilities in the assignment of nearby hydrogen bond acceptors.

The final torsion angle restraints set contained 147 angles comprised of 93 pairs of backbone angles and 54 sidechain angles. Moreover the full hydrogen bond restraints set consists of 43 bonds: the five bonds discussed above from loops, written as ambiguous restraints to acceptor oxygens within the same loop, additional to the 38 H-bonds located in the regular secondary structure regions.

5.4.2 Re-calculation and refinement results

One hundred NMR structures of SH2 were generated using CNS. Those structures have various overall energy values ranked in order of increasing total energy shown in Figure 5.9 (a); from 775 kcal/mol to more than 10000 kcal/mol. The higher value of the total energy is mainly caused by distance and dihedral restraint violations. This is a fairly typical profile for an NMR structure calculation, in that most of the structures have fairly similar energy, with only the final four or so having significantly higher energy, indicating a problem in the calculation.

The subsequent widely used procedure is to select the lowest total energy structures from the calculated structures and refine them in explicit solvent against a full force field to minimize the energy (Linge et al., 2003). There is no general consensus on how to select the most appropriate structures for refinement (Spronk et al., 2004). The general assumption is that the structures with the lowest energy from simulated annealing have the fewest restraint violations and so are likely to be good starting points for refinement.

Therefore all 100 structures from the CNS simulated annealing calculation were taken and refined in explicit solvent. As shown in Figure 5.9 (b), there is a strong correlation between the total energy of the structures before and after the refinement. Significantly, the two highest energies before refinement (not shown in Figure 5.9) are the only two structures to refine to a structure with an overall positive energy, again indicating a problem in the initial structure calculation. However as shown in Figure 5.9 (a), not all of the lowest energy structures prior to refinement are the lowest energies after refinement. For example structure No.51 does not have a particularly low total energy before the refinement, but after refining in solvent it is the second lowest total energy structure. The important conclusion that it is more reliable to refine all calculated structures.

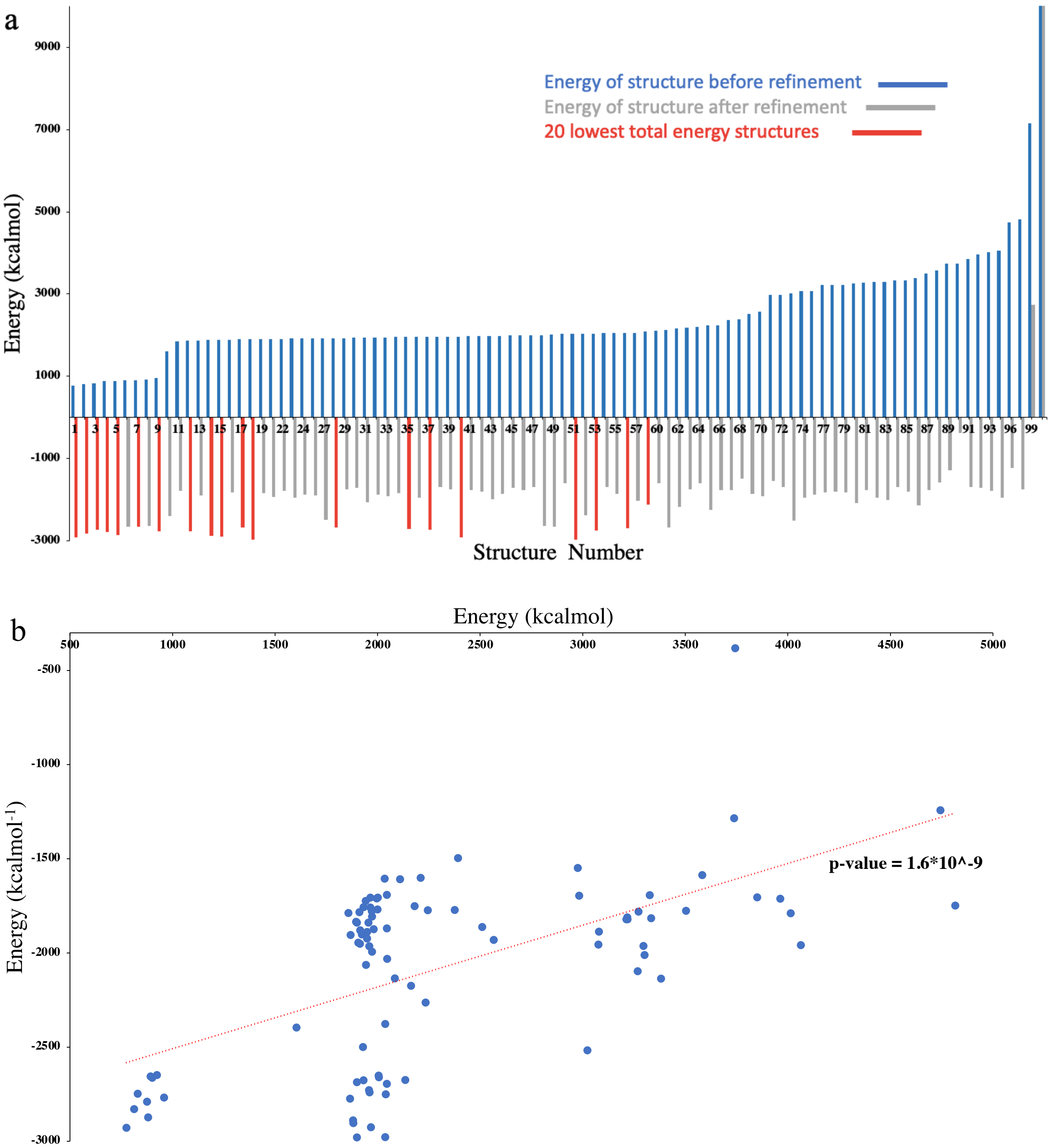


Figure 5.9. The total energies of 100 structures calculated using crystallography and NMR system program (CNS). (a) shows the total energy for each structure before and after the refinement. Blue bars indicate the total energy of structure before the refinement, ranked in order of increasing overall energy level. Grey bars represent the total energy of the structures after refining in explicit solvent. Red bars are the 20 lowest total energy structures after the refinement. (b) represents the correlation between the total energy of the structures before and after the refinement (structures with positive energy are un-refined and structures with negative energy are refined in water); the last two models unrefined and refined (which have total energy >10000 kcalmol) are excluded from the plot and the calculation. The trendline of the best fit (in red dash line), the Pearson correlation coefficient (r value), and the p-value are shown. A significant correlation is indicated by $p < 0.05$.

5.4.3 Analysis of the structures

The accuracy of the NMR structures of SH2 were measured by the *Accuracy of NMR structures using random coil index and rigidity* (ANSURR) method, recently developed in Sheffield, which is the first measuring method comparing the input backbone chemical shifts data with the calculated NMR structure. As explained in Fowler et al., 2020, the ANSURR method evaluates the accuracy of NMR protein structures based on two measurements: the correlation score and the RMSD score. This measuring method combines the matching between the local rigidity from backbone chemical shifts (predicted by Random Coil Index) and the local rigidity from the computed NMR structure (by Floppy Inclusions and Rigid Substructure Topography (FIRST)) into a correlation score which measures the secondary structure, whereas a RMSD score between FIRST and RCI assesses the overall rigidity. Initially this method was used to follow the accuracy improvement in the final ensemble structure derived from CYANA structure calculations, and then the re-calculated and refined structures calculated using the CNS program.

ANSURR score was calculated for the final 20 conformers for each CYANA structure calculation (1st, 2nd, 3rd, 4th, and 5th). As shown in Figure 5.10, there is a continuing significant improvement in ANSURR score with adding more restraints in CYANA structure calculations including the last 5th calculation where fewer hydrogen bonds restraints were used. The RMSD and correlation scores are ranked centile scores comparing to all NMR structures in the PDB, meaning that a structure as good as the median PDB structure has a score of 50%, and the maximum possible is 100. The ANSURR score is the sum of the RMSD and correlation scores, and therefore has a maximum possible value of 200.

There is an increase in the overall rigidity of the ensemble structure (RMSD score) while adding more restraints, and also with removing a number of un-confident restraints to avoid introducing bias in the final resulting structure. Moreover the improvement in the correlation score carried into the last CYANA calculation even though the final ensemble structures were computed with fewer restraints. As shown in Figure 5.4, five out of six of the H-bonds removed in the 5th CYANA structure calculation are formed between residues located on the middle or at the end of β strands. These H-bonds seem to force the secondary structure to be too organised, and removing them returns flexibility to the secondary structure as measured by RCI which explains the significant increase in the correlation score of the last calculation, as shown in Figure 5.11. That means that in this case, adding more restraints enhanced the overall rigidity of the structure but it did not make it more accurate.

In addition the ANSURR scores were calculated for all 100 re-generated structures and 94 solvent-refined structures, and are shown in Figure 5.10. There is an overall improvement in ANSURR score for structures after refining in solvent which derives from increases in both the correlation and RMSD scores. There is no big change in the correlation score for structures before and after the refinement, because in general refined structures in explicit solvent did not display any major changes in protein fold, and thus no major change in the secondary structure. On the other hand there is a large increase in the RMSD score for structures after refining in explicit solvent due to the big improvements in hydrogen bonding which works to rigidify the whole protein structure (Linge et al., 2003).

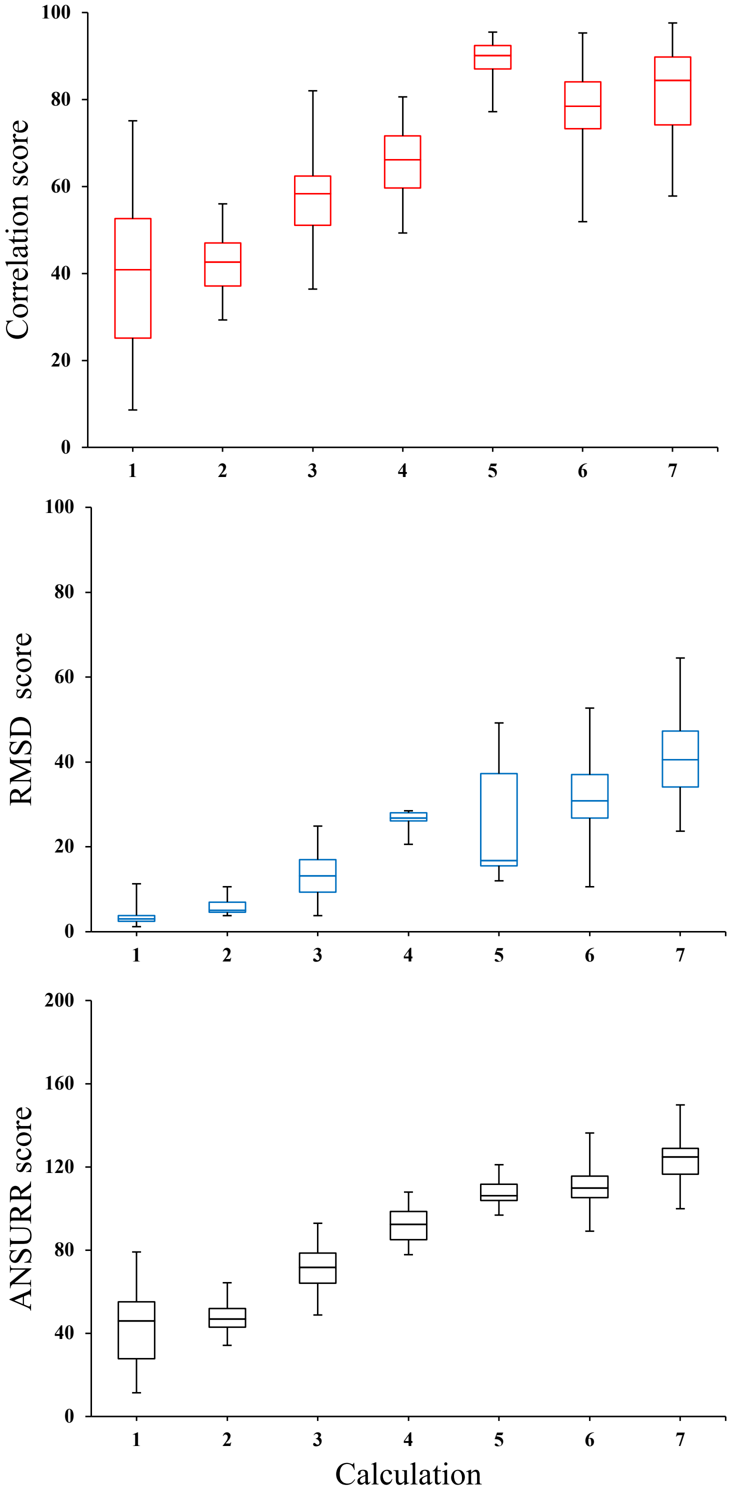


Figure 5.10. ANSURR analysis of structure calculations by CYANA and CNS. The sample size of each CYANA calculation consists of the final 20 structures of SH2; CYANA 1, 2, 3, 4, and 5.The last two points of the boxplot consist of 100 structures of the unrefined structures (6) and 94 refined structures (7) from CNS (four of the structures did not refine and two are clearly wrong). The middle line in each box shows the median. The ANSURR score is the sum of RMSD and correlation scores and therefore runs from 0 to 200.

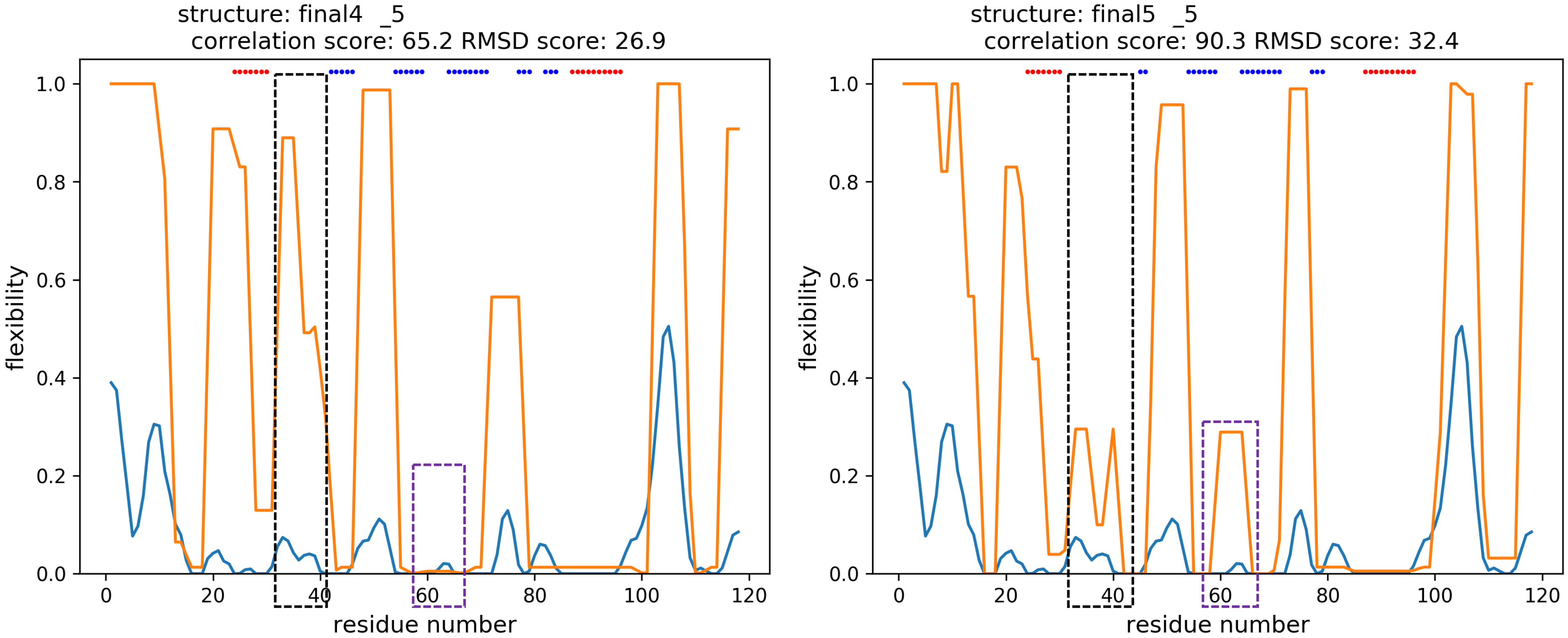


Figure 5.11. An example for ANSURR analysis of structure 5 from the 4th and the 5th CYANA calculations. In all the plots the blue line indicates the predicated flexibility calculated by RCI and the orange line represents the predicated flexibility computed by FIRST. In the top of each plot are the structure's number and name, the measured RMSD and correlation scores by ANSURR. The secondary structure shows in each plot on the top as red bars (for α-helix) or blue bars (for β-strand). The black and the purple dash boxes surrounded same residues area (H39-T36, V41-N59, and F60-K63) as an example where there is a change or difference in the predicted flexibility (by RCI) from removing those H-bonds in the 5th CYANA calculation.

5.4.4 Ensemble structure selection and validation

The initial step in the structure validation is to select a set of models from a large number of calculated structures to represent the NMR structure. Typically the structure selection is applied to unrefined structures on the basis of their total energy or the size of the experimental violated restraints (Spronk et al., 2004). After that the selected set of structures are carried through a sequential refining procedure to represent the final NMR structure.

As was shown earlier there is a correlation between the total energy of structure before and after the refinement, however all calculated structures were refined in explicit solvent to achieve more accuracy in the final structure selection. In order to select the final set of ensemble structures from the 94 solvent-refined structures to represent the final SH2 NMR structure, two validation criteria were considered: ANSURR score and total energy. Structures 99 and 100 were excluded from the structure selection as they have large total energy.

As explained previously ANSURR analysis is an indicator of structure accuracy, hence the ANSURR score was measured for all solvent-refined structures. ANSURR scores are rank percentile scores measured against all NMR structures in the PDB, and are the sum of RMSD and correlation scores. Figure 5.12 shows that in general all refined structures have ANSURR score above the average. A score of 200 is the maximum possible, and a score of 50 for either RMSD or correlation is achieved by the median PDB structure. All solvent-refined structures displayed correlation score higher than the median, whereas relatively few structures (14 out of 94) exhibited RMSD score above the median. Fowler et al., 2020 showed that β-sheet structures tend to have worse RMSD scores than helical structures. However this still means that the SH2 structure has a poorer RMSD structure (ie, it is too floppy) than the median NMR structure.

Structures were sorted on the basis of their ANSURR score and the 20 best refined structures, which have highest ANSURR scores above 130, are displayed in Figure 5.13. It is not necessary for those structures to have the lowest total energy.

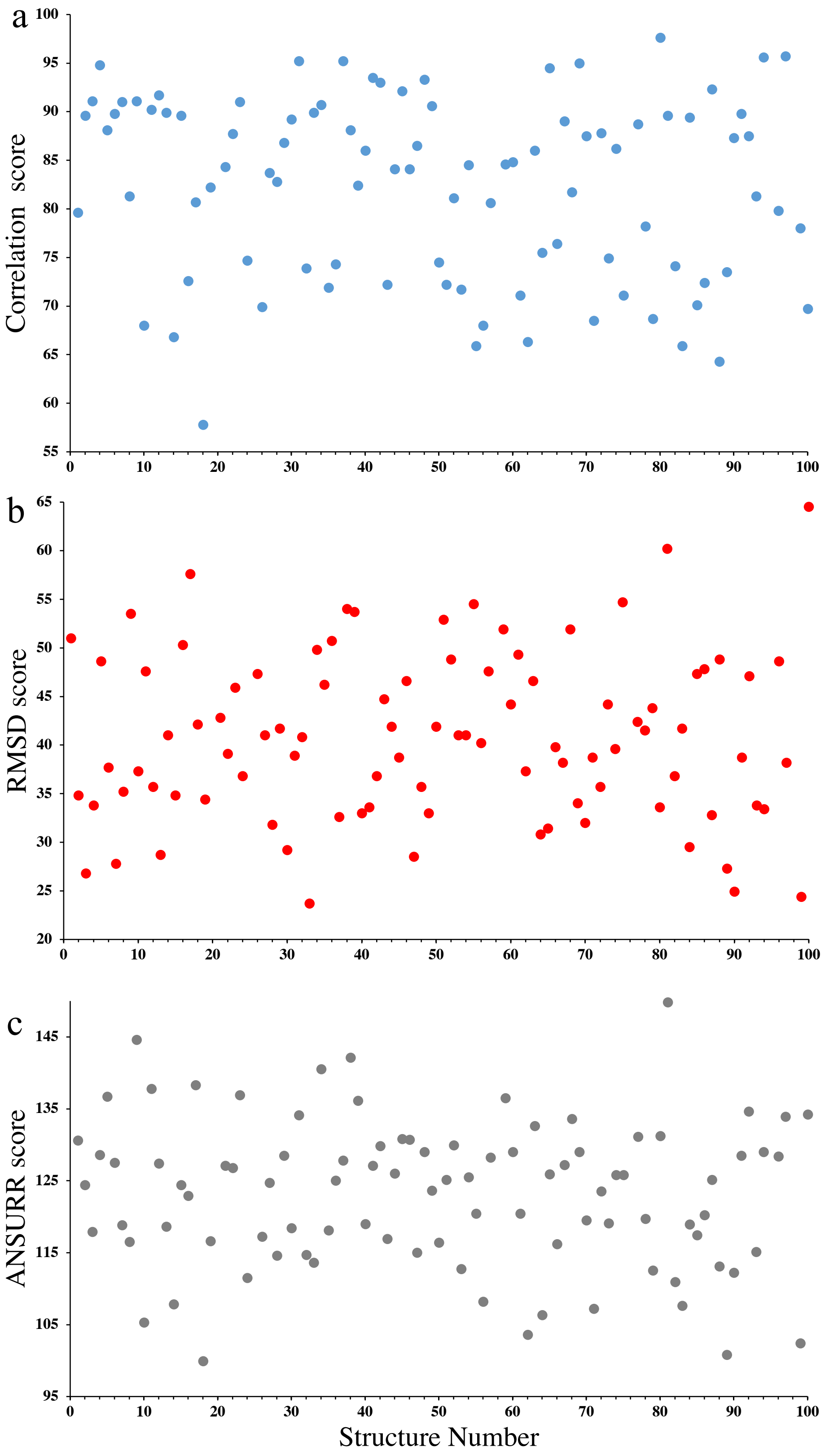


Figure 5.12. ANSURR analysis for solvent-refined structures. Each score is displayed against the number of the refined structure. (a) Correlation score, (b) RMSD score, and (c) summed ANSURR score.

It is shown in Fowler et al., 2020 that the total energy of a structure is a good measurement to estimate its accuracy. Furthermore the total energy of a structure reflects the size of the restraint energy as it contributes to the experimental restraints, thus a structure with low energy tends to be more accurate because it means having a better fit to the experimental input data (Spronk et al., 2004). However if no manual intervention is applied to check or remove violated restraints after each structure calculation, the total energy of refined structures is expected to be higher than the acceptable average.

Because there is no obvious correlation between the total energy of the solvent-refined structures and their ANSURR score, another 20 ensemble refined structures were selected based on their total energy to compare with the first ensemble models which were selected on the basis of their ANSURR score, as shown in Figure 5.13. The two sets of ensembles were taken to further quality assessment.

There is a confirmed strong relationship between the structure accuracy and the geometrical quality, such that the structures with the best backbone geometry also tend to be the most accurate (Fowler et al., 2020). Thus Ramachandran analysis was applied for both ensemble models whether based on total energy or ANSURR score, to act as an independent validation of whether either selection criterion has genuinely improved structure accuracy. The 20 ensemble models with the highest ANSURR score have better overall Ramachandran distribution than the ensemble models with the lowest total energy, calculated both as means of good percentage of residues in favoured region, and as low percentage of residues in generously allowed and disallowed regions, as illustrated in Figure 5.14.

Lastly the final ensemble NMR structures for SH2 protein were chosen based on a combination between the two validation criteria: a small number of structures were selected as they have both low total energy and good ANSURR scores, shown in Figure 5.13.

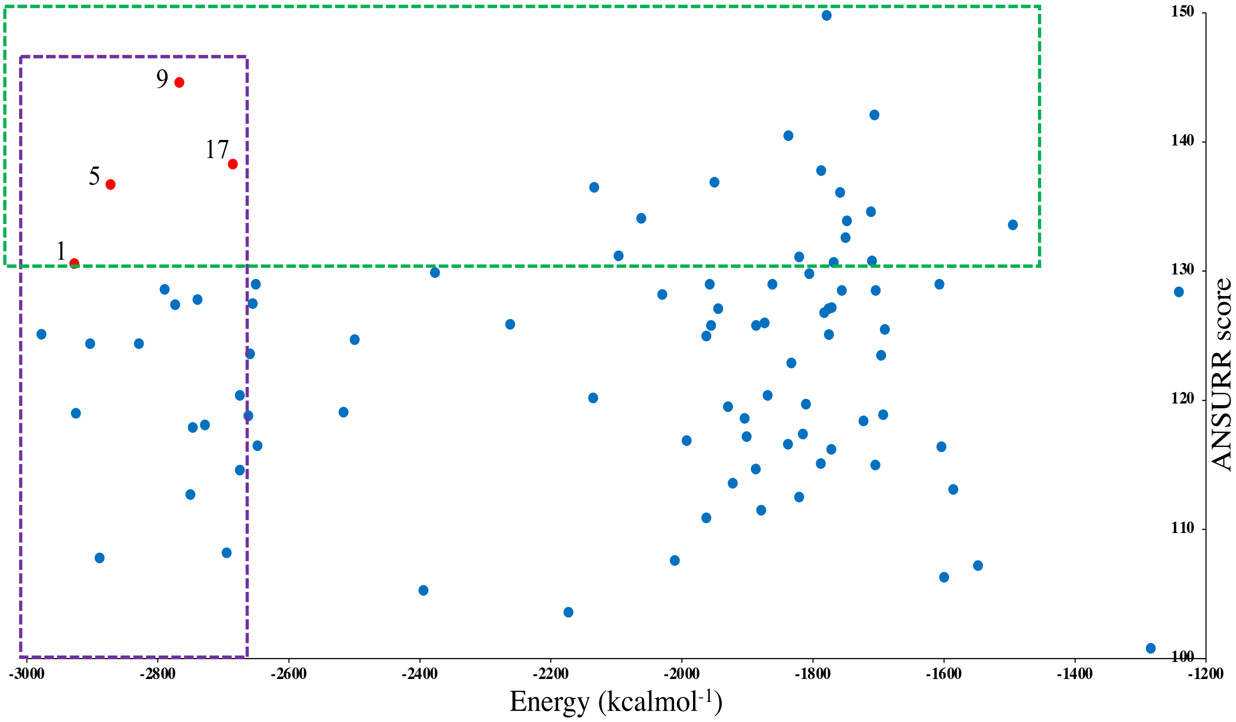
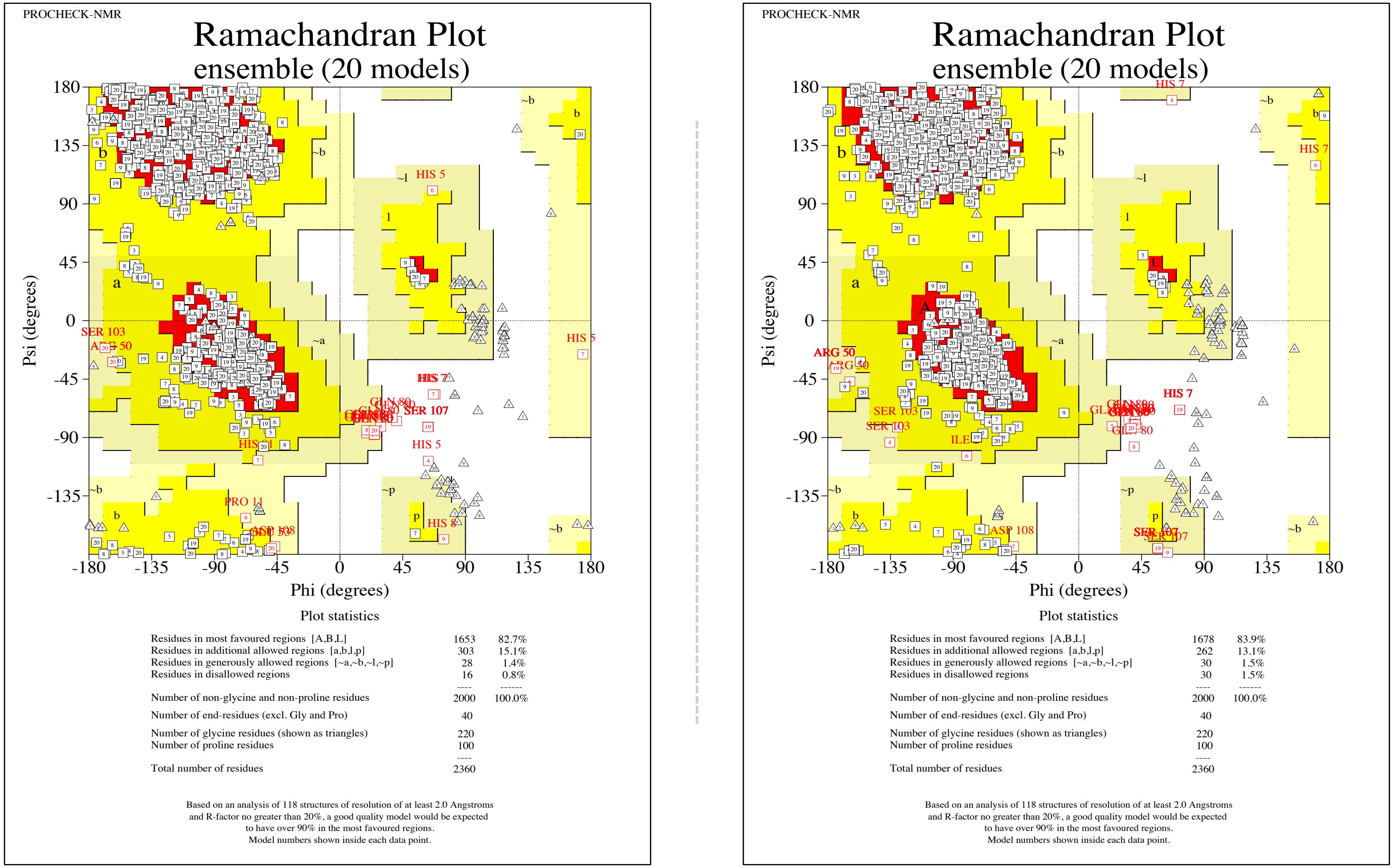


Figure 5.13. Performance of ANSURR score against total energy of refined structure. Red circles show the structures with both a good ANSURR score and low total energy. The 20 lowest energy structures are surrounded with a purple dash square, and the 20 structures with highest ANSURR scores are surrounded by a green dash square.



Plot statistics

```
Residues in most favoured regions  [A,B,L]                1653     82.7%
Residues in additional allowed regions  [a,b,l,p]          303     15.1%
Residues in generously allowed regions  [~a,~b,~l,~p]       28      1.4%
Residues in disallowed regions                              16      0.8%
                                                          ----    ------
Number of non-glycine and non-proline residues            2000    100.0%

Number of end-residues (excl. Gly and Pro)                  40

Number of glycine residues (shown as triangles)            220
Number of proline residues                                 100
                                                          ----
Total number of residues                                  2360
```

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.
Model numbers shown inside each data point.

Plot statistics

```
Residues in most favoured regions  [A,B,L]                1678     83.9%
Residues in additional allowed regions  [a,b,l,p]          262     13.1%
Residues in generously allowed regions  [~a,~b,~l,~p]       30      1.5%
Residues in disallowed regions                              30      1.5%
                                                          ----    ------
Number of non-glycine and non-proline residues            2000    100.0%

Number of end-residues (excl. Gly and Pro)                  40

Number of glycine residues (shown as triangles)            220
Number of proline residues                                 100
                                                          ----
Total number of residues                                  2360
```

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.
Model numbers shown inside each data point.

Figure 5.14. Ramachandran distribution plot of ϕ and ψ angles for the best 20 selected ensemble structural models; left the plot of the 20 ensemble structures that have the best ANSURR score, right the plot of the 20 ensemble structures that have lowest total energy. Residues in red letters are in the generously allowed and disallowed regions.

5.4.5 Validation of the NMR ensemble structure

The final ensemble NMR structures for SH2 protein consist of four models; 1, 5, 9, and 17, Figure 5.13. Those models converged to a similar folded structure.

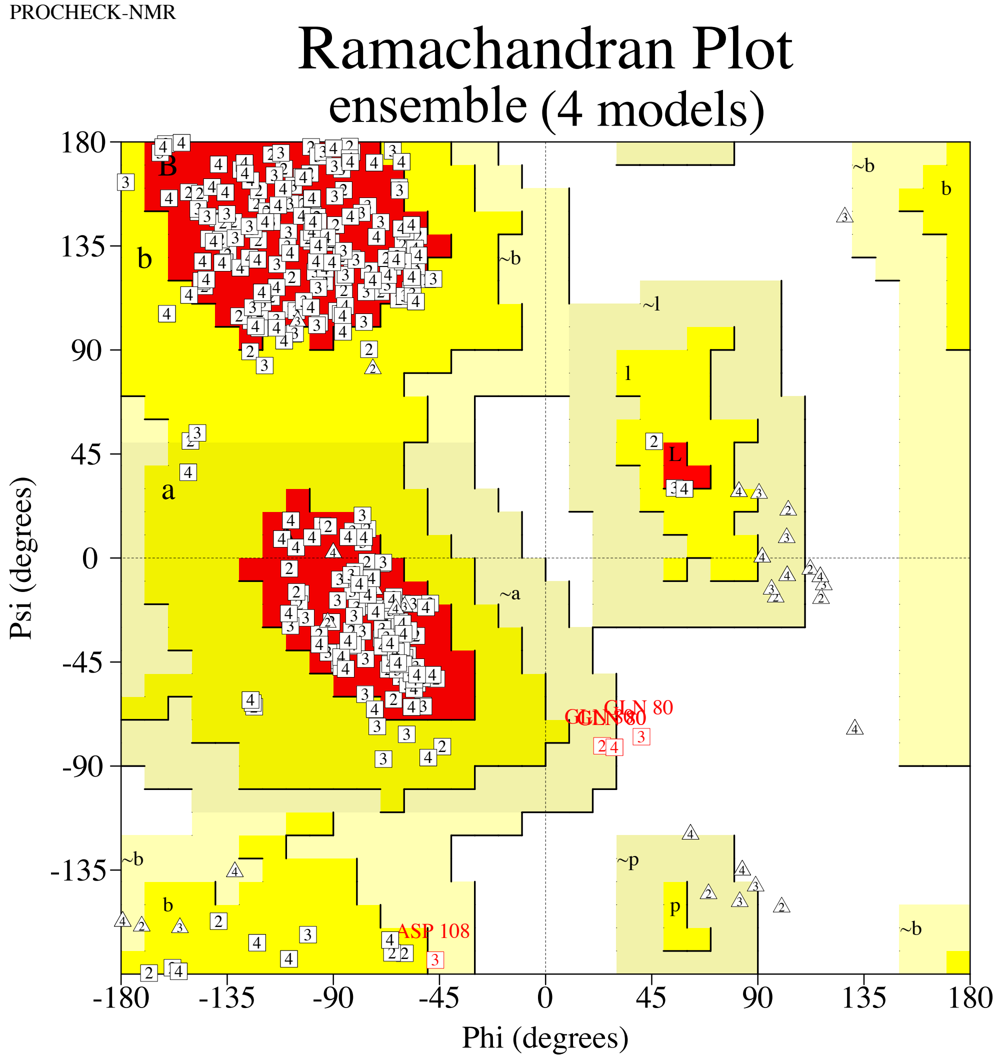
In general the structures have similar number of total energy and violated restraints. Typically, computed models from an NMR structure calculation are not in complete agreement with the input experimental data due to a range of different reasons such as the presence of different conformations of the same protein in solution (Nabuurs et al., 2004). However in this study where the NMR structure of SH2 protein was determined using an automated NOE assignment and structure calculation with no manual inspection applied to check for inconsistent and incompatible experimental restraints, the disagreement is higher than the average case. Therefore the NOE and dihedral angle energies are the highest among the other parameters. The total energy average of the ensemble with a breakdown of the energies are presented in table 5.3.

The structure accuracy is related strongly to its geometrical quality, thus the validation of the final ensemble structures was checked by Ramachandran analysis. PROCHECK-NMR (Laskowski et al., 1996) was used to generate the Ramachandran plot for each residue in the ensemble structures of SH2 as shown in Figure 5.15 where the average breakdown of residues shows the proportion mentioned on the bottom of the plot. The Ramachandran distribution diagram shows the majority of residue backbone angles are located in favourable regions, while one residues does not. Although about 86% of all residues are in the favourable allowed region, the average quality of the ensemble structures is still below the ideal percentage (> 90%) which represent a good structure. However, it is worth noting that the Ramachandran distribution for this ensemble is better than that for either the good ANSURR ensemble or the low energy ensemble, which in turn are better than the entire ensemble. This provides a good indication that this strategy is a promising way to select good structures from the ensemble.

The positional uncertainty in the molecular coordinates in the final ensemble is represented as the coordinate RMSD value. It is known that NMR protein structures calculated in solution due to their internal dynamics are expected to adopt variant conformations around the correct average (Spronk et al., 2004). Although the precision of the ensemble structure is not necessarily a good method to measure the structure quality (Zhao & Jardetzky, 1994), however a good quality of NMR structure is expected to have low RMSD value. Fowler et al., 2020 found a good correlation between ANSURR score and RMSD precision. In this study where all distance restraints were derived from automated NOE assignment (no manual intervention), the precision of the ensemble is an important measure of structure quality.

The RMSD average of the final ensemble models from the mean for the backbone and sidechain atoms of the structured region of the protein (18-97 a.a) are 0.817 Å and 1.314 Å, respectively. The overall precision of the ensemble is on average, the precision breakdown (backbone and sidechain) of each residue in the sequence of SH2 protein shown in Figure 5.16, a and c. It is clear in Figure 5.16 the big variation in RMSD of the backbone and sidechain atoms appeared at the N and the C termini, additional to slight variation at the big loop connecting βI strand to βII strand, which adopted different folds due to their flexibility in the protein structure and due to the lack of NOEs and hydrogen bond restraints. Although the other regions of the protein have good precision. The NMR structural statistics of the average ensemble models are presented in Table 5.3.

PROCHECK-NMR



Plot statistics

Residues in most favoured regions [A,B,L]	344	86.0%
Residues in additional allowed regions [a,b,l,p]	51	12.8%
Residues in generously allowed regions [~a,~b,~l,~p]	4	1.0%
Residues in disallowed regions	1	0.2%
	----	------
Number of non-glycine and non-proline residues	400	100.0%
Number of end-residues (excl. Gly and Pro)	8	
Number of glycine residues (shown as triangles)	44	
Number of proline residues	20	

Total number of residues	472	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.
Model numbers shown inside each data point.

Figure 5.15. Ramachandra distribution plot of ϕ and ψ angles for the SH2 final ensemble structures. Residues in red letters are in the generously allowed and disallowed regions.

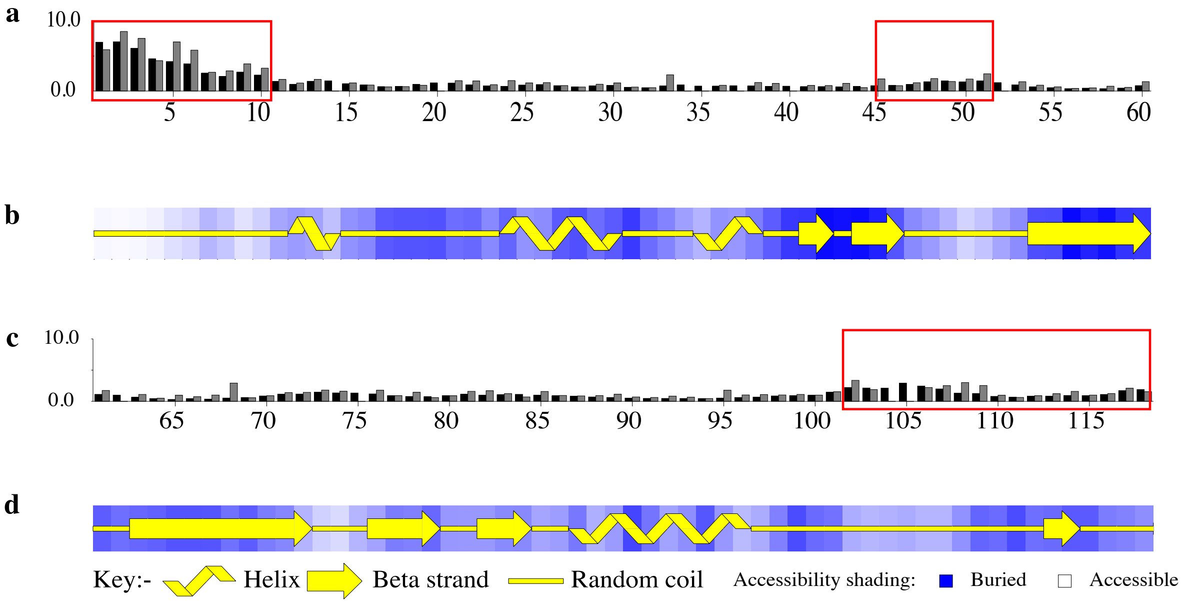


Figure 5.16. The geometrical properties of the final NMR ensemble structure of the SH2 protein. (a) and (c) show histograms of the RMSD deviations from mean coordinates of the backbone-chain in black bars and the side-chain in grey bars, the numbers on the bottom are protein sequence. (b) and (d) illustrate a schematic diagram for the predicted secondary structure and the average accessibility along the protein sequence which is represented for each residue as shading. Red boxes surround the most variable parts of the protein according to RMSD. Graphs were generated by PROCHECK-NMR program.

Table 5.3. NMR statistics for the final ensemble structures for SH2 protein.

NMR distance and dihedral constraints	
Distance constraints	
Total NOE	1667
Sequential (|i – j| = 1)	1146
Medium-range (|i – j| < 5)	190
Long-range (|i – j| > 5)	331
Hydrogen bonds	43
Total dihedral angle restraints	
• ϕ	93
• ψ	93
• X^1	54
Structure statistics	
• Distance constraints violation (>0.0)	230
• Dihedral angle constraints violation (>0.0001)	57
Average rms deviation (Å)	
• Heavy atoms	1.314
• Backbone	0.817
Total average energy (kcal mol^{-1})	-2813
• NOE	327
• Bonds	62.5
• Angles	296.4
• Improper	178
• Dihe	640
• Vdw	-357
• Elec	-4059

5.5 Comparison between NMR and crystal structures

Overall the final NMR structural models of SH2 are well packed into an average well defined structure, Figure 5.18, a. The average secondary structure of the ensemble consists of a central β sheet; three long antiparallel strands (I V41-Q46, II Y54-F60, and III K63-N72), followed by two antiparallel small strands (IV G75-V79 and V L82-F84), all surrounded by two α helices (I 24R-32L and II 87I-V96), Figure 5.15. Moreover the fold of the protein agrees well with the TALOS-N secondary structure prediction as shown in chapter 4, Figure 4.11.

A comparison between the resultant NMR structure of (mouse) SH2 protein and the existing crystal structure of (human) SH2 domain 5w3r.pdb (Hu et al., 2006) allows us to conclude that the structures have mostly a very similar fold. The extra residues at the C and N terminal ends of SH2 protein look to form an integral part of the protein structure as they are close to each other in space and are thus required for correct folding and stability. Also there are a number of residues different between the protein sequences of those two structures including the N terminal tags, which are displayed in the sequence alignment in Figure 5.18. When superimposed by the regular regions of the secondary structure (S23-H97 to S533-H607), the RMSD between the backbone atoms of the NMR and the crystal structures is 1.569 Å, as shown in Figure 5.18, b.

The obvious difference between the two structures appears as low precision in long loops (between β strands β1-β2 (Q46-Y54), β3-β4 (N72-G75), and β4-β5 (V79-L82)) and the beginning of the unstructured C terminal tail (I99-V109) as they are poorly defined in the NMR structure. Also there is a slightly variation at the end part of the first α helix and β strand which in the crystal structure are one residue longer than in the NMR structure, whereas the end of the third β strand in the NMR structure is slightly longer than the crystal structure. Analysis of the ANSURR calculation for this ensemble suggests that the NMR structures in general are too floppy than the expected RCI measure, as shown in the flexibility plots produced by ANSURR for each structure in the final ensemble in Figure 5.17. In other words, it is likely that the NMR structure for this loop is incorrect and is missing one or more restraints (possibly a hydrogen bond), which could markedly improve the accuracy of this loop.

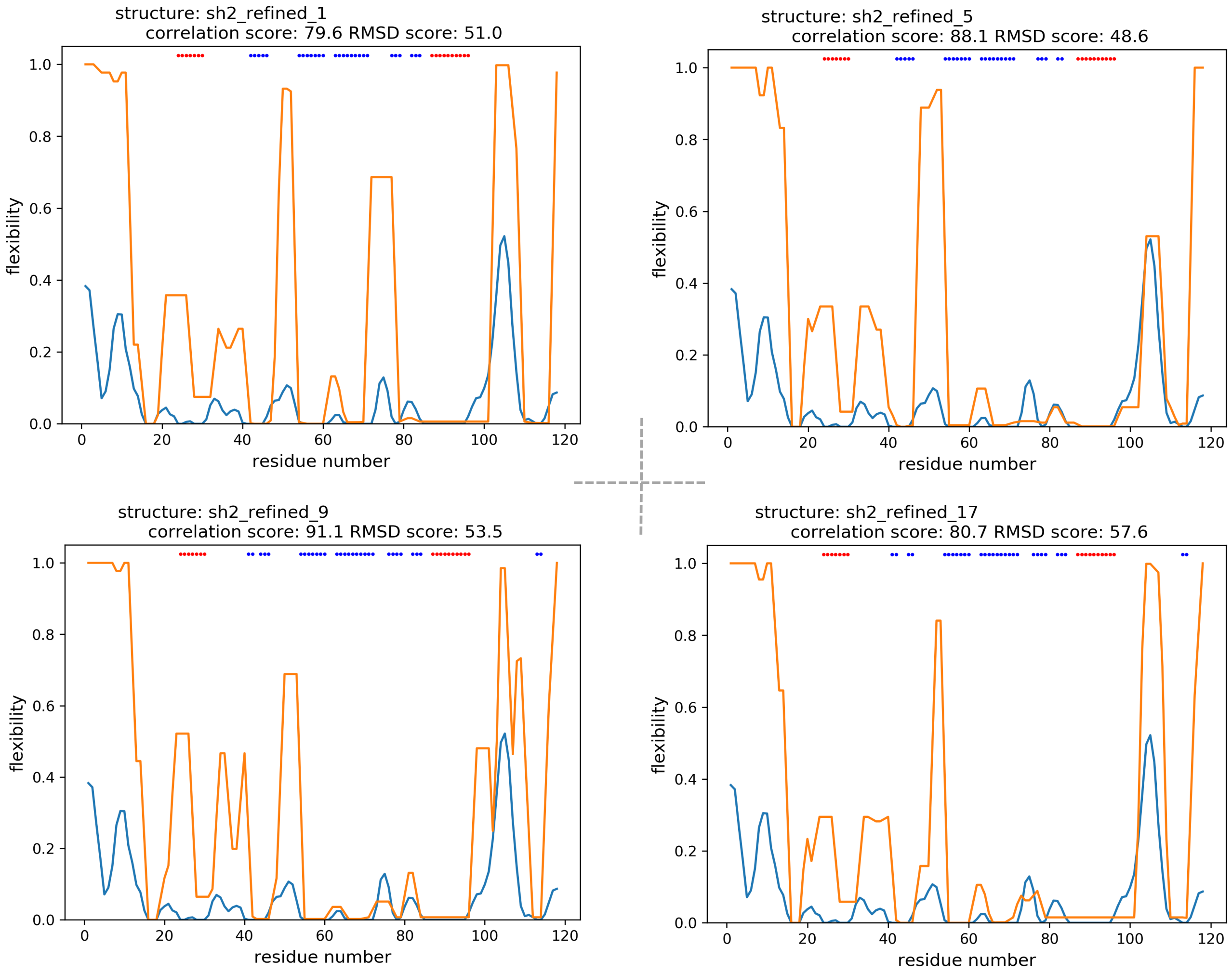
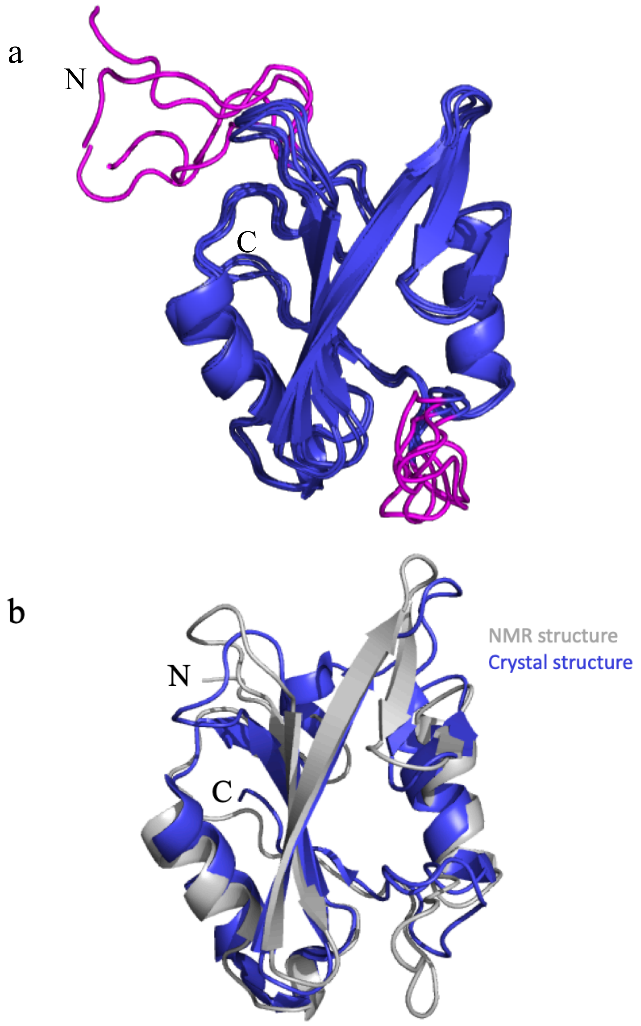


Figure 5.17. ANSURR analysis of each refined structure (1, 5, 9, and 17) in the NMR final ensemble for the SH2 domain. In all the plots the blue line indicates the predicated flexibility calculated by RCI and the orange line represents the predicated flexibility computed by FIRST. In the top of each plot are the measured RMSD and correlation scores by ANSURR, and the structure's number. The secondary structure shows in each plot on the top as red bars (for α-helix) or blue bars (for β-strand).



```
                       10        20        30        40        50        60
SH2B1|mouse|   KIHHHHHHDQPLSGYPWFHGMLSRLKAAQLVLEGGTGSHGVFLVRQSETRRGEYVLTFNF
SH2B1|homo |   ----GSHMDQPLSGYPWFHGMLSRLKAAQLVLTGGTGSHGVFLVRQSETRRGEYVLTFNF
                      520       530       540       550       560       570

                       70        80        90       100       110
SH2B1|mouse|   QGKAKHLRLSLNEEGQCRVQHLWFQSIFDMLEHFRVHPIPLESGGSSDVVLVSYVPSQ  →(118)
SH2B1|homo |   QGKAKHLRLSLNAAGQCRVQHLHFQSIFDMLEHFRVHPIPLESGGSSDVVLVSYVPSS  →(628)
                      580       590       600       610       620
```

Figure 5.18. A comparison of the final NMR structure of SH2 with the crystal structure. (a) Superimposition of the final SH2 ensemble structures consisting of four refined structures (1, 5, 9 and 17). The magenta colour represents the low precision regions of the protein. (b) Superimposition of the average of the ensemble NMR structure with the crystal structure (5w3r.pdb) after removing the N terminal tag from both proteins. Dark blue is the crystal structure, and grey is the NMR structure. (N) N-terminal and (C) C-terminal, both conformers are overlaid using the backbone atoms (N, Cα, and Co). The structure superimposition was created by PyMol. The bottom is the ClustalW sequence alignment of SH2 protein of two species; mouse is the SH2 protein of the NMR structure (118 a.a) and homo (human) is the SH2 protein of the crystal structure (628 a.a). Residue numbers are shown on the top or bottom along the protein sequence. The grey highlighted residues are the N terminal tags and the different residues between the two sequences. The yellow highlighted residue is the mutated residue in the protein sequence of the crystal structure.

5.6 Discussion

The NMR structure determination framework presented in Figure 5.1 was used to produce an NMR structure ensemble that has quite good RMSD and a good Ramachandran distribution. ANSURR of the final ensemble shows that both RMSD and correlation scores are above the median, and thus that the structure is good compared to the average NMR structure. However, the final NMR structure of SH2 is slightly different from the reference crystal structure (5w3r) especially in one loop, and more floppy than it should be as proven from ANSURR score. The difference between those two structures is the most commonly pinpointed variation between NMR and crystal structures (Fowler et al., 2020), which are the loops and the unstructured regions.

The first procedure in the NMR structure determination was the CYANA structure calculations. The importance of the CYANA calculation is mainly the NOE assignments produced by CYANA, which were used in the next calculation procedures.

As discussed earlier in this chapter the key of success in the CYANA structure calculation is obtaining the correct fold structure in the first cycle, because the obtained structure in cycle 1 depends on the NOE distances derived from the automated NOE assignment. The initial NOE assignment relies basically on the agreement between the input data as the first filtering criterion imposed before the introduction of other features such as network anchoring and constraints combination. For a successful NMR structure calculation based on automated NOE assignment with no manual intervention, arguably it is the quality of the input data which is critical. Therefore two input data should be checked carefully for consistency: the chemical shift assignments and the NOE peak lists.

In this study the experimental spectra which were used for assignment and the NOE spectra which were used in the structure calculation were checked for chemical shift referencing. However there was no further checking for the consistency between the input data, such as the obtained chemical shift values and the peak positions in the peak lists, to be sure they are matching within the tolerance defined values. Also the presence of a large number of noise peaks in the peak list reduced their quality and interfered with the real peaks. In hindsight, it would have saved a lot of time to remove many weak signals from the NOE peak list.

For the chemical shifts, a complete and accurate list of at least 90% completeness is required for reliable automated NOE assignment by CYANA. As mentioned in Güntert, 2004 the CYANA algorithm can overcome a big number of incomplete chemical shifts (whether wrong or missing). It is however important to note that some peaks are more important than others. Missing assignments are not important for chemical shifts which contribute to only a small number of NOEs, but there are a small number of essential chemical shifts that can cause a significant distortion in the resulted structure, such as the chemical shifts of protons in aromatic rings which are involved in lots of NOEs. In our study the assignment of the aromatic rings was incomplete with a number of other residues which make in total < 80% of the resonance assignment completed.

It is not realistic to achieve a good NMR structure with only one CYANA run, unless a complete chemical shift assignment and very high quality of peak lists are used, which is hard to reach from real experimental data. The CYANA structure calculation should be repeated until the CYANA performance meets the recommended validation criteria by improving the quality of the input data (chemical shifts and peak lists).

Moreover increasing the weight of the relative conformational constraints in each CYANA calculation may restrict the structure to fit the input data, however as shown in this study, that will not help to correct the NOE assignment in the first cycle as those restraints data will be used just for structure calculation not during the automated NOE assignment.

Furthermore available software for automated NOE assignment and structure calculation such as CYANA can minimize the manual intervention, but not cancel it. In this study no manual inspection was applied to check the resulted automated NOE assignment from the CYANA calculation nor remove the violated restraints, which were used in the next calculation and refinement procedures. The assignments were checked manually before using CYANA, and a number of errors in the CYANA assignment were corrected by forcing it to use the manual assignments.

Although the resulting NMR structure for the SH2 protein is clearly still too floppy (from the ANSURR analysis), it is not essentially wrong but still needs further improvement. In particular, it would be ideal to find further restraints for the loop connecting β-strands 1 and 2.

Chapter 6 NMR study of the interaction between the SH2 domain and the C terminal tail

This chapter focuses on the interactions between the intrinsically disordered C-terminal tail of SH2B1 β isoform and the SH2 domain that precedes it, to examine whether this interaction is likely to have a biological function. Because the tail contains a tyrosine (Tyr139) that is suggested to be phosphorylated in *vivo* by KinasePhos web tool (Huang et al., 2005), the study concentrates on the effect of phosphorylation of Tyr139, mainly using NMR.

The first approach was to detect interactions between the SH2 domain and the tethered pY139 on the basis of the induced chemical shift changes observed in ^{15}N HSQC spectra recorded from the labelled protein under non-phosphorylated and phosphorylated states. Also the changes in the chemical shift were used to determine the bound ligand location.

The second approach was NMR titration experiments of SH2 protein with a synthetic C-terminal peptide [GDRCpTyrPDASST]. The experiments involved repeated ^{15}N HSQC measurements with serial additions of ligand stock to the protein sample. The valuable information extracted from the titration experiments were CSPs which were used to identify the residues involved in the binding interaction with the C terminal titrated ligand, and to provide information about the affinity of the interaction.

The determined 3D NMR structure of SH2B domain which was described in chapter 5 was used to map the interacting residues based on CSP data upon binding to the tethered pTyr139 site and to the free C-terminal ligand. These results were used to characterise the location of those residues and understand the recognition mechanism of SH2 domain in both cases.

6.1 Characterisation of protein-tethered pTyr139 ligand interaction

The tagged SH2c Y114F mutant protein (160 a.a) was expressed as a ^{15}N and ^{13}C labelled protein as described in Chapter 3. The phosphorylation of Tyr139 at the C-terminus of the SH2c was done using Fer Kinase enzyme as it was predicted by Group-based Prediction System 3.0 database to be able to phosphorylate this site (Xue et al., 2008) as explained in Section 3.3.

After that the backbone resonance assignments of the double labelled unphosphorylated SH2c and phosphorylated SH2c proteins including the downstream 45 residues of intrinsically disordered tail were assigned using the ^{15}N HSQC spectrum and the standard triple resonance NMR spectra HNCO, HNCACO, HNCA, CBCACONH, and HNCACB. The sequential backbone assignment method which was explained previously in Chapter 4 to assign SH2 protein was followed to assign ^{1}HN, ^{15}N, ^{13}C′, ^{13}Cα, and ^{13}Cβ backbone nuclei of the unphosphorylated and phosphorylated SH2c proteins. The processing of the spectra, chemical shift referencing and peak picking were performed in Felix2007. The sets of nuclei correlated in these spectra, and acquisition parameters of each backbone spectrum of the non-phosphorylated and phosphorylated proteins, are shown in table 2.10.

All observed NH signals in the ^{15}N HSQC spectra were picked and systematically numbered in Felix2007. After that each ^{1}H-^{15}N signal was analysed and correlated to its local group of backbone nuclei (Cα, Cβ, C′, pCα, pCβ and pC′) which is referred to as a spin system in the triple resonance spectra. Then each obtained spin system was matched up with its neighbouring spin system and with the residue type in the protein sequence using simulated annealing in the Asstools program (Reed et al., 2003). In the ^{15}N HSQC spectrum, the observed cross-peaks of amide side-chains were identified as they are not correlated to nuclei in the triple resonance spectra, and so were excluded from the backbone assignment. In total for SH2c protein, 144 out of 160 residues are expected to be identified and assigned, excluding the N-terminal tag residues and eight proline residues.

The backbone resonances of the two proteins were obtained in two different buffers: 100 mM potassium phosphate, 2 mM DTT and 10% D_2O at pH 7 (preferred buffer for NMR study as it is non-protonated buffer; Kozak et al., 2016) and 50 mM Tris, 50 mM NaCl, 2 mM DTT and 10% D_2O at pH 7.

The backbone amide chemical shifts of non-phosphorylated SH2c and the phosphorylated SH2c (pTyr139) protein were assigned and compared, to confirm the likely phosphorylation site of Tyr139 at the C-terminus of the SH2c protein, and study the expected interaction between the SH2 domain to the tethered phospho-tyrosine 139 site. The study was based on analysis of the ^{1}H and ^{15}N chemical shift perturbations of the SH2c proteins in ^{15}N HSQC spectra. CSP or induced chemical shift changes occur as result of interaction or conformational change (Zuiderweg, 2002; Williamson, 2013).

The bound ligand location is defined based on chemical shift changes. The weighted averages of the chemical shifts of the backbone nuclei (^{1}H, ^{15}N) were calculated using the following equation which is described in Williamson, 2013:

$$\Delta\delta_{HN} = \sqrt{(\Delta\delta H)^2 + (0.14 \text{ x } \Delta\delta N)^2} \qquad \text{(Equation 6.1)}$$

in which ΔδH is the change in chemical shift of the amide proton, ΔδN is the change in the amide nitrogen chemical shift, and 0.14 is an average weighting factor for the ^{15}N shift comparing to ^{1}H shift. The Δδ chemical shifts of ^{1}H and ^{15}N for each observed residue in the ^{15}N HSQC spectra are shown in Appendix B.

6.1.1 Backbone resonance assignments of the labelled SH2c Y114F in phosphate buffer

The backbone assignments of the non-phosphorylated SH2c protein (~1 mM) were obtained in 100 mM potassium phosphate, 2 mM DTT and 10% D_2O at pH 7. In the ^{15}N HSQC spectrum 142 amide signals were observed and assigned to their specific residue in the protein sequence by identifying its corresponding backbone nuclei using the triple resonance experiments.

The amide signals of residues Trp17 and Gly20 were missing. The absence of Trp17 is most likely due to its unusually high field HN chemical shift, as discussed in Chapter 3. The absence of Gly20 may be due to intermediate chemical exchange or fast exchange with solvent (Englander and Mayne, 1992). In addition the amide peaks for Gln61 and His81 appeared in low intensity in the ^{15}N HSQC spectrum and their corresponding backbone resonances were not assigned in the triple resonance spectra. However the corresponding backbone resonances of Trp17, Gly20, Gln61 and His81 were assigned from the spin system of the following residue. In total the backbone assignment was complete for SH2c protein in phosphate buffer at pH 7 except for amide resonances of residues: N-terminal tag, Trp17 and Gly20, Figure 6.1. The backbone amide signals of the downstream 45 residues of the intrinsically disordered tail (from V115 to L159) appeared in a cluster around 8.3 ppm in the ^{15}N HSQC spectrum that is compatible with the expected distribution of NMR chemical shifts for random coil (Gibbs et al., 2017; Cornilescu et al., 1999).

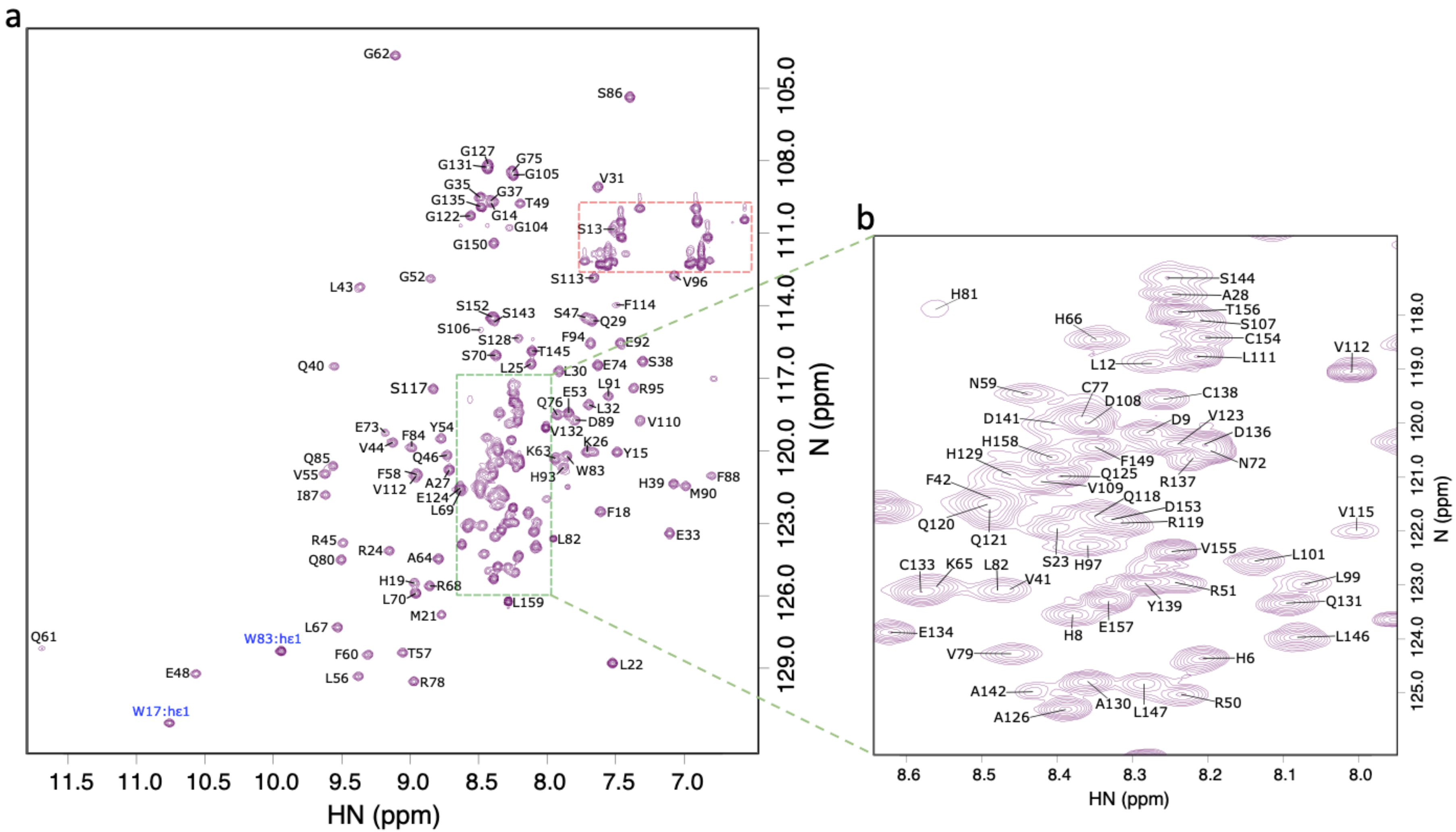


Figure 6.1. ^{1}H-^{15}N HSQC spectrum of the labelled SH2c protein in 100 mM phosphate buffer at pH7. a) The full ^{1}H-^{15}N HSQC spectrum and, b) An expansion for the overlapped crowded region which is surrounded by the green dashed box. The assignments of the backbone amide resonances are labelled by residue type and sequence number in black, the side chain amide peaks of Asn and Glu residues are surrounded by a red dashed box, and the sidechain resonances of Trp residues are shown in blue.

The phosphorylation of Tyr139 at the C-terminus of the labelled SH2c was performed as explained in Section 3.3, and about 60 µM of phosphorylated protein in 100 mM potassium phosphate buffer at pH 7 was obtained. The low concentration of the protein sample was due to precipitation during the phosphorylation experiment, and losses during buffer exchange in order to remove unwanted components that were used during the phosphorylation reaction which may affect the quality of the NMR spectra.

Figure 6.2 shows an overlay of the ^{15}N HSQC spectra of non-phosphorylated and phosphorylated SH2c proteins which were recorded in identical conditions in phosphate buffer. The majority of the amide signals were in similar positions. Therefore for those peaks where no shift changes were observed, their assignments were copied from the previous assignment of the non-phosphorylated protein. On the other hand there were a number of peaks slightly shifted and a few new peaks. For those peaks, triple resonance spectra were used to assign their backbone resonances. Because of the low concentration of the phosphorylated protein sample, there were six unassigned residues: Q10, W17, G20, T36, E102 and S103. In total, 138 observed backbone amide peaks in the ^{15}N HSQC spectrum were assigned to their correlated residue in the protein sequence.

In both backbone assignments of the non-phosphorylated and phosphorylated proteins, due to the low quality of CBCACONH and HNCACB spectra most of the Cβ resonances were not assigned, however the HNCO and HNCA spectra were more sensitive hence ~96% of HN, ^{1}N, Cα, and C′ were assigned.

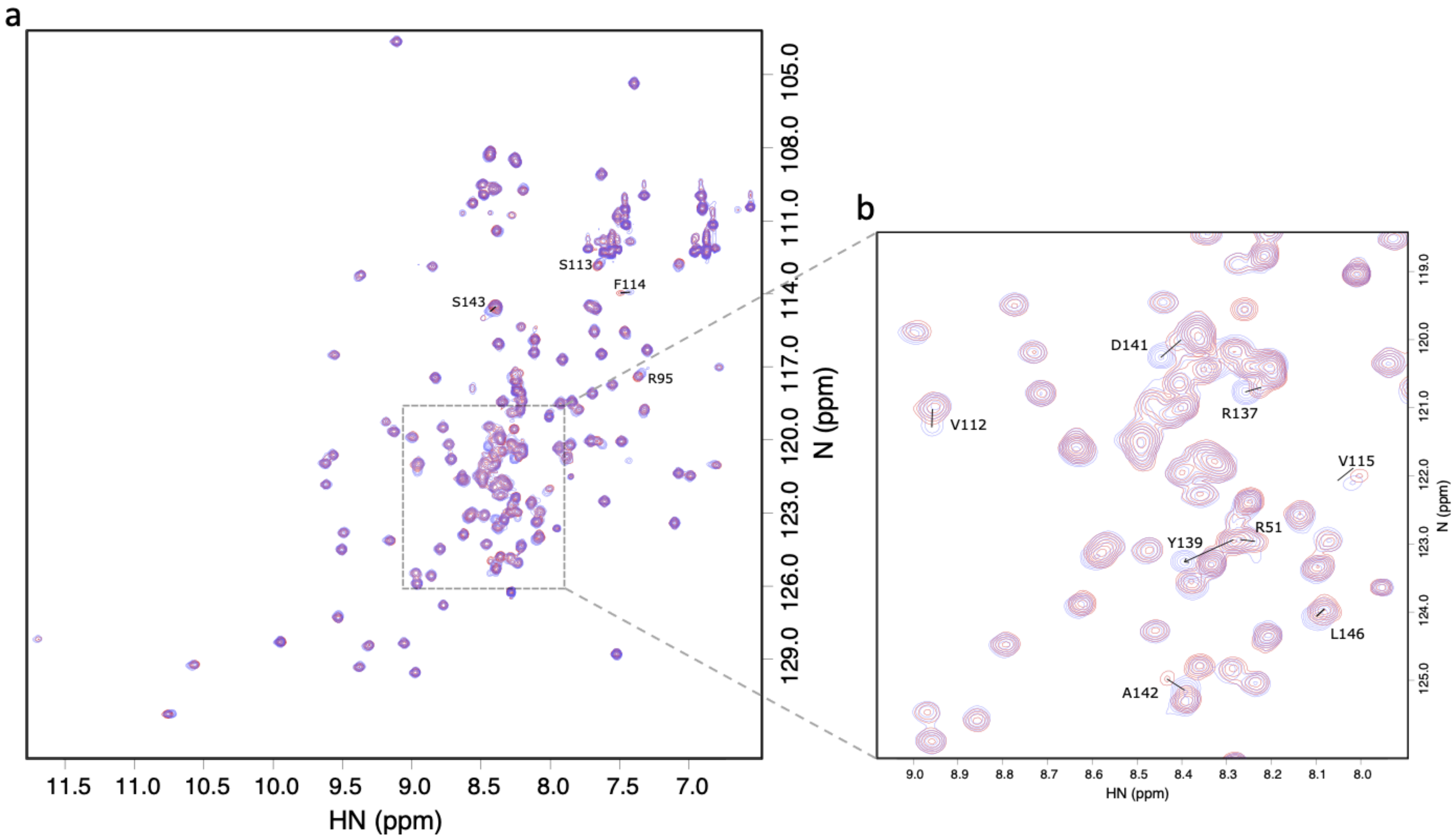


Figure 6.2. Overlaid ^{1}H-^{15}N HSQC spectra of the non-phosphorylated and phosphorylated SH2c proteins in phosphate buffer at pH7. a) The full ^{1}H-^{15}N HSQC spectrum, and b) *A*n expansion for the overlapped region which is surrounded by the grey dashed box. The assignments of the backbone amide resonances of the shifted and new residues are labelled by residue type and sequence number in black. The red spectrum is the non-phosphorylated protein, and the blue spectrum is the phosphorylated protein.

6.1.2 Analysis of the interaction of the SH2 with the tethered ligand (pY139) in phosphate buffer

The Δδ chemical shift changes of backbone amide resonances of the SH2c proteins that were obtained under unphosphorylated and phosphorylated states were calculated using equation 6.1. Figure 6.3 shows there are a number of residues exhibiting Δδ chemical shift changes above the estimated threshold (0.015 ppm). Δδ chemical shifts are listed in appendix B.

The largest shift change occurred for Tyr139 around Δδ ~ 0.1 ppm. The limited chemical shift perturbation is consistent with the general observation for the phosphorylation of Tyr which leads to a small downfield shift for the backbone amide, because of the distal position of the hydroxyl group (OHη) of Tyr which is attached covalently to the phosphate (PO_4^{3-}) group, and thus does not have a big impact on the backbone amide shift (Theillet et al., 2012). The phosphorylation of Tyr does not cause a large perturbation of the backbone amide resonances, whereas it does induce a big downfield chemical shift for the aromatic CHε resonances (Δδ 0.3 to 3 ppm) (Huang et al., 2020; Bienkiewicz et al., 1999).

Moreover there were perturbations for the nearby residues: R137, D141, A142, and S143 (Figure 6.3), because the phosphorylation does not just induce chemical shift changes of the modified residue itself but also the neighbouring residues due to changes in the local environment of Tyr139 (Huang et al., 2020). These observations verified the result of the Mass spectrometry analysis for the phosphorylation of Tyr139, as explained in Chapter 3.

As a consequence of the Tyr139 phosphorylation there was expected to be an interaction between pTyr139 and the positive phosphate-binding pocket residues in the SH2 domain, however there were no observed chemical shift changes in the component residues of the binding pocket (Hu et al., 2003; Waksman et al, 1992), although there were small shifts (Δδ ~ 0.03) for residues Arg50 and Arg95, shown in Figure 6.3. Arg50 is one of the residues that is situated within the active loop binding site (which connects strand β1 to β2), and is known not to have direct interactions with the phospho-tyrosine. It is hydrogen bonded to the sidechain of S47 which interacts directly with the phosphate group of the phospho-tyrosine

(McKercher et al., 2017). However there is no observed phospho-tyrosine interaction for residue S47 as shown in Figure 6.1.

The lack of interaction between pY139 and the SH2 binding pocket was a surprise. It is more likely that because the sequence of the C-terminal tail of SH2c (pY^{139}XXA^{142}) differs from the ideal SH2 high-affinity ligand (pYXXL/I, JAK2), it is therefore expected to bind with only low affinity (O'Brien et al., 2003; Hu et al., 2003; McKercher et al., 2017). The presence of phosphate ions in the protein sample at high concentration (100 mM potassium phosphate) may be saturating the positive phosphate-binding pocket in the SH2 domain, thus preventing the pTyr139 from fully binding. The missing interactions with the phosphate binding site may be why it compensated for the loss of interaction by binding to Arg50.

In addition there were observed shift changes for residue F114 (Δδ 0.07 ppm) and its nearby primary sequence residues V112, S113 and V115 (Figure 6.3). TALOS-N secondary structure prediction result for SH2 protein shows that residue F114 part of a small β strand at the C-terminus of the SH2 domain, Figure 6.3. One of the possible consequences of pTyr139 competing with the phosphate ions to bind to the positive binding pocket in the SH2 domain is increasing the mobility of the tail which in turn may have caused a conformational change for the fulcrum point of the C terminal tail.

The chemical shift change results reveal that there was a successful phosphorylation of the Tyr139 at the C-terminus of SH2c, but there was no full binding between the tethered phospho-tyrosine 139 site and the phosphate-binding pocket in the SH2 domain. Instead it binds weakly in a different orientation.

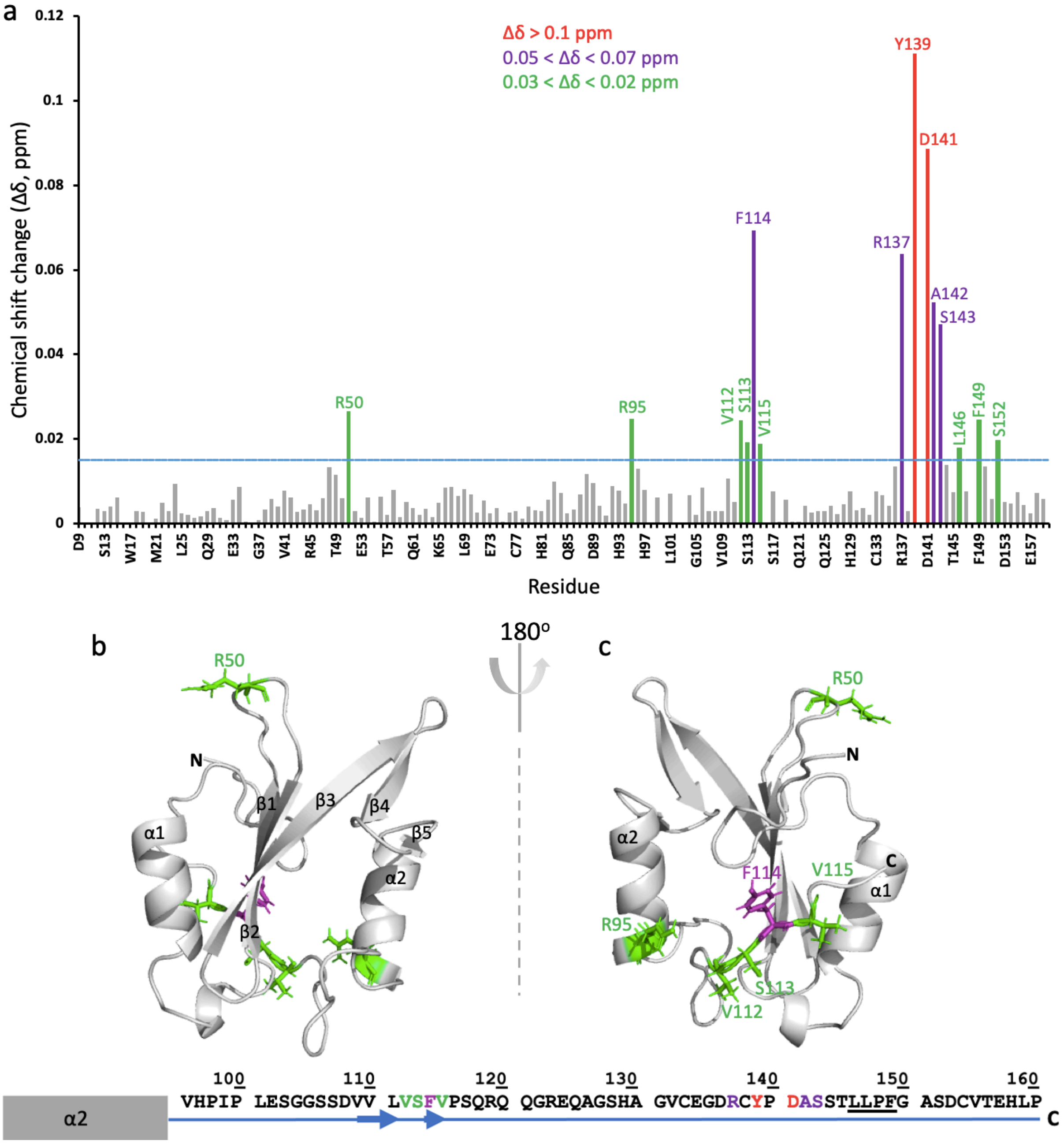


Figure 6.3. The chemical shift changes of the SH2c in phosphate buffer at pH 7. a) The changes in the chemical shift of ^{1}H and ^{15}N of the non-phosphorylated and phosphorylated SH2c proteins were compared, the chemical shift perturbations are plotted versus residue type and number. Bars coloured in accordance to their classification of chemical shift change as: largest shift (Δδ > 0.1 ppm, red), medium shift (0.05 < Δδ < 0.07 ppm, purple) and small shift (0.03 < Δδ < 0.02 ppm, green). The estimated threshold is the blue dashed line (0.015 ppm). b) Mapping the NMR chemical shift changes of ^{1}H and ^{15}N onto the 3D structure of SH2 domain, b) the front and c) the back faces of the SH2 domain (180° rotated view). A scheme of the C-terminal tail of the SH2c protein (V96-P160 a.a) corresponding to the residue type and number are shown below the 3D structure. Residues Y^{139}PDA are the expected C-terminal ligand that binds to SH2 domain. The underlined residues are the hydrophobic patch residues (LLPF^{149}). The green and purple labelled residues (VSF^{114}V) are the shifted residues which are part of SH2 domain. TALOS-N predicts short β-sheet stretches for residues 109-111 and 114-115.

6.1.3 Backbone resonance assignment of SH2c Y114F in Tris buffer

To remove all the phosphate ions that may be saturating the phosphate-binding pocket in the SH2 domain, thus possibly preventing the binding to the phospho-tyrosine site, the protein sample's buffer was exchanged to 50 mM Tris, 50 mM NaCl, 2 mM DTT and 10% D_2O at pH 7. After that the backbone sequential assignments of SH2c were obtained based on the labelled non-phosphorylated protein using previous assignments made with triple resonance spectra.

Excluding the proline and the N-terminal tag residues, 132 out of a possible 144 were assigned in the ^{15}N HSQC spectrum to their sequential backbone resonances in the triple resonance spectra (Figure 6.4). A number of residues remain unassigned in the ^{15}N HSQC spectrum: D9, W17, G20, M21, S23, T36, E48, R50, Q61, H81, S103, and S106. That is mostly because of the low quality of the triple resonance spectra in Tris buffer; as commented above for the phosphate spectra, the HNCO and HNCA spectra were more sensitive than the CBCACONH and HNCACB spectra. Therefore 92% and 87% were assigned for the chemical shift of the Cα and the C′, respectively, whereas only 53% of the Cβ resonances were assigned.

To compare the obtained ^{1}H-^{15}N assignments for the non-phosphorylated protein with the backbone assignments of the phosphorylated SH2c protein in Tris buffer, the same labelled phosphorylated SH2c protein sample which was used previously to assign its backbone resonances in phosphate buffer was used, and the sample's buffer was exchanged to 50 mM Tris, 50 mM NaCl, 2 mM DTT and 10% D_2O at pH 7. During the buffer exchange ~20 µM of the protein concentration was lost, and the remaining 40 µM double labelled protein was used to assign the backbone resonances using 3D triple resonance experiments.

Out of 144 expected residues for the phosphorylated protein, 111 residues were assigned in the ^{15}N HSQC spectrum (Figure 6.5). Due to the low concentration of the protein, 30 residues were unassigned: D9, Q11, L12, S13, G14, W17, G20, M21, S23, R24, T36, F42, S47, E48, R50, Q61, C77, F84, D89, S103, G104, S106, S107, V112, S113, F114, V115, H129, E134, G135, R137, A142, and S144.

In total the backbone resonances of the SH2c pTyr139 were assigned in Tris buffer: 77% of ^{1}HN and ^{15}N, 69 % of Cα, 58 % of C′, whereas the most of Cβ resonances were not assigned.

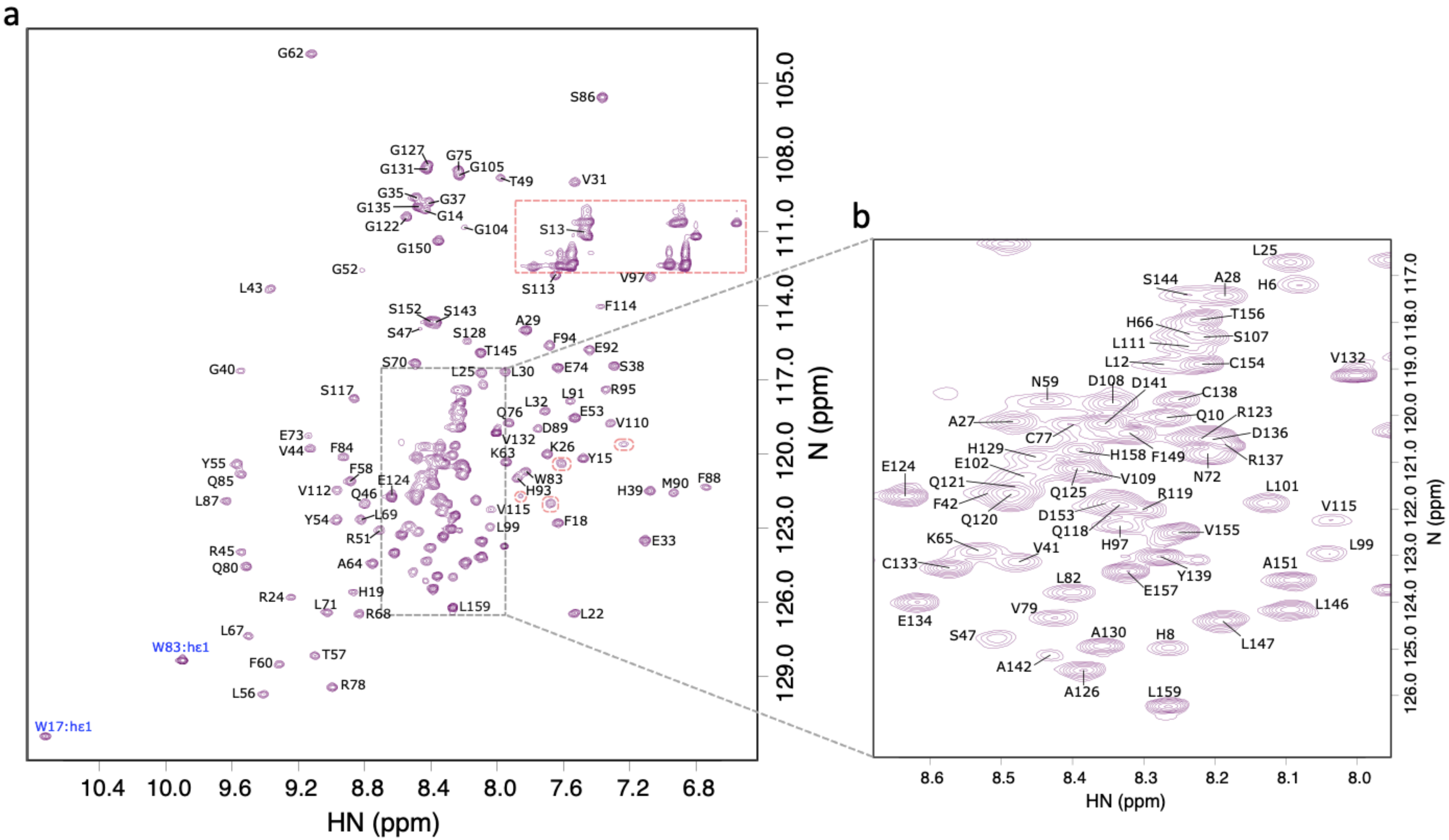


Figure 6.4. ^{1}H-^{15}N HSQC assignments of the non-phosphorylated SH2c protein in Tris buffer at pH7. a) The full ^{1}H-^{15}N HSQC spectrum, and b) An expansion for the overlapped region which is surrounded by the grey dashed box. The backbone amide resonance assignments are labelled in black. The sidechain resonance assignments of Trp residues are labelled in blue. The sidechain amide peaks of Asn and Glu residues are surrounded by a red dashed box. The red dash ovals are surrounded the sidechain amide peaks of Arg residues.

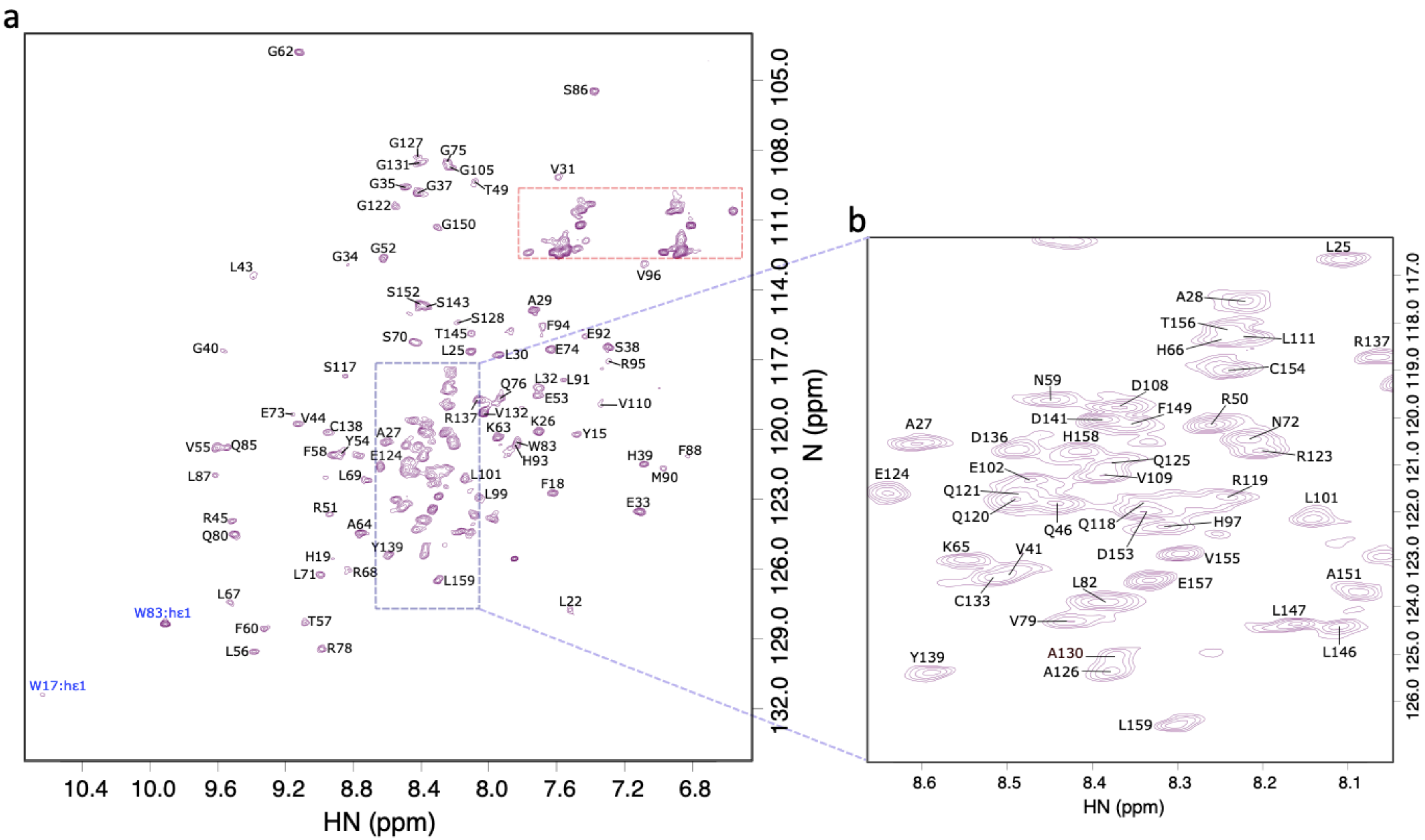


Figure 6.5. ^{1}H-^{15}N HSQC assignments of the phosphorylated SH2c protein in Tris buffer at pH7. a) The full ^{1}H-^{15}N HSQC spectrum, and b) *A*n expansion for the overlapped region which is surrounded by the purple dashed box. The resonance assignments are labelled in black, whereas the sidechain amide peak assignments of the Trp residues are labelled in blue. The sidechain peaks of Asn and Glu residues are surrounded by a red dashed box.

6.1.4 Analysis of the binding interaction of the SH2 with the tethered ligand (pY139) in Tris buffer

Although there were a number of missing residues in the assignment of the phosphorylated SH2c, due either to the low protein concentration or to the interaction with pY139, the reminding 77% assigned residues were enough to compare with the assignment of the non-phosphorylated SH2c Y114F protein.

The Δδ chemical shift of ^{1}H-^{15}N between the non-phosphorylated and phosphorylated states of SH2c proteins were calculated using equation 6.1. Residues that have Δδ shift above the calculated mean plus standard deviation (threshold 0.09) were analysed, and coloured in accordance to their classification of chemical shifted range (Figure 6.6.a). Note the different y axis scale compared to Figure 6.3, implying much stronger interactions than were seen in phosphate buffer, and confirming our hypothesis that phosphate in the buffer can compete with the pY-binding site interaction. Δδ chemical shifts are listed in appendix B.

The largest chemical shift changes ($\Delta\delta > 0.3$) were observed for Tyr139 and its neighbours D136 and C138 at the C-terminal tail of SH2c protein. The strong downfield shift for the backbone amide of the Tyr139 is not just due to the phosphorylation of Tyr itself (because in phosphate buffer, the changes around Y139 are much smaller) but more likely to arise from the binding to the phosphate-binding pocket in the SH2 domain, which is also verified by the chemical shift changes in neighbouring residues.

Figure 6.6.a shows that residues Q46, T49, R51, G52, E53, and Y54 displayed chemical shift changes $\Delta\delta > 0.1$ ppm. These residues are located in the loop that connects strands β1 and β2. Mapping those residues onto the 3D structure of SH2 domain reveals that they are clustered within the phosphate-binding pocket and probably have a significant role in the interaction with the phosphate group of tethered pTyr139. Moreover residues L22, A27, Q29, and G34 were shifted mostly because of the indirect interaction or conformational change upon the binding to the pTyr139.

In addition, the amide signals of residues L69 and F88 exhibited Δδ CSP above the estimated threshold, as was the NHε1 sidechain of W17 (Δδ 0.09 ppm). According to the data mapped onto the 3D structure in Figure 6.6.b and c those residues are situated on the back of the SH2 domain. These residues do not interact with the pY ligand in classical SH2 interactions. A typical SH2 ligand has the sequence motif pYxφφ where φ is a hydrophobic residue (JAK2 has the sequence pTyr^{148}ELLT), and the hydrophobic residue at Y+3 interacts with a hydrophobic pocket adjacent to the pY pocket, but separated by a peptide strand that crosses the binding site (Fig. 6.3.b). The SH2c sequence is pTyrPDASSLLPF, and so does not have large hydrophobic residues at Y+3. It does however have hydrophobic residues at Y+6, Y+7 and Y+9, and it is possible for the peptide chain to reach around the back of SH2 and make contact with L69, F88 and W17 sidechains.

In summary, Δδ chemical shift mapping data suggests that SH2 domain binds weakly to the pTyr139 using a similar binding interface to other known SH2 domains, however it reveals a unique interface that contains hydrophobic interactions involving residues after pTyr139, at positions Y+6 to Y+9, as shown in Figure 6.6.

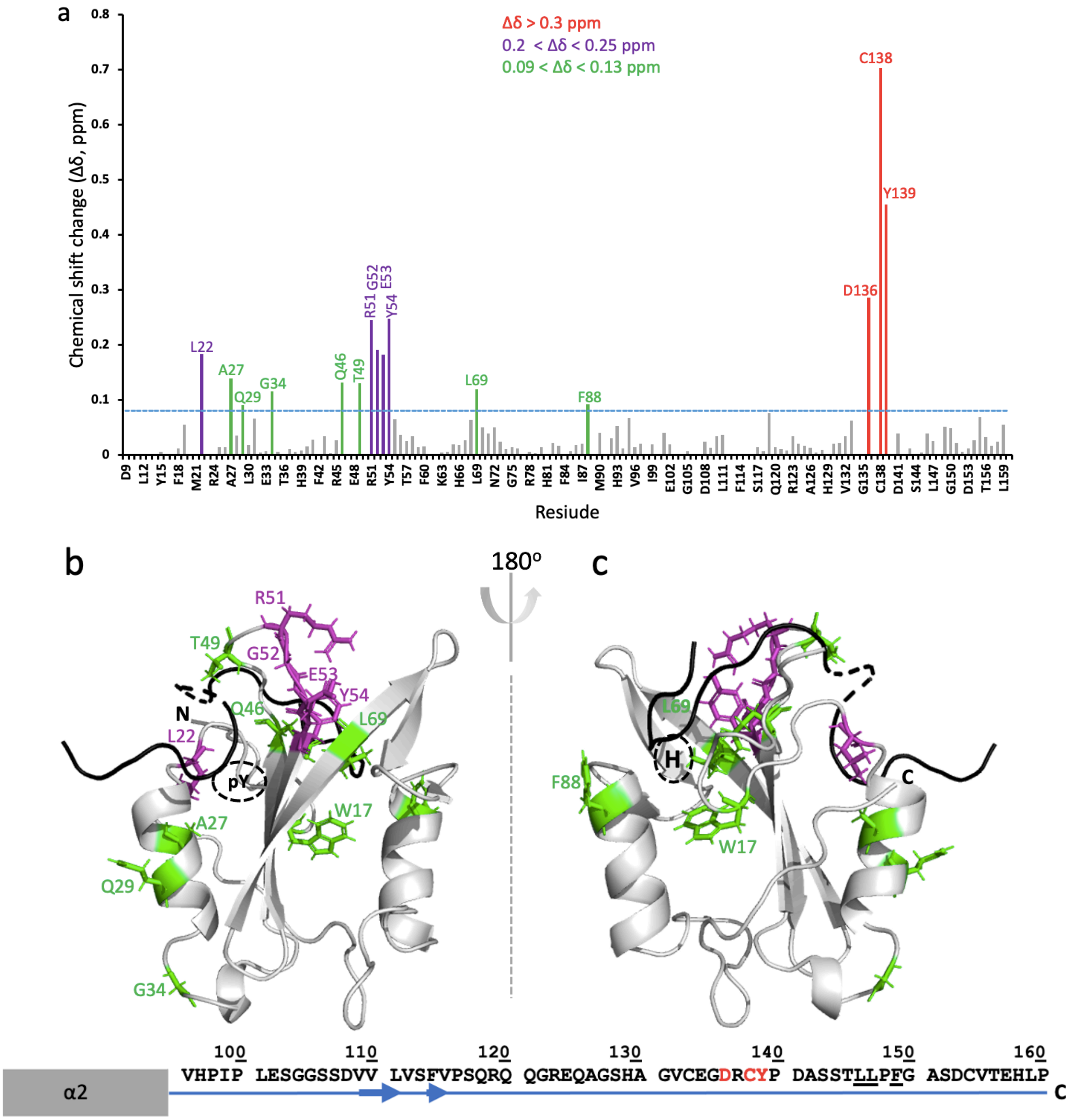


Figure 6.6. The chemical shift changes of the SH2c Y114F upon binding to tethered pY139 in Tris buffer at pH 7. a) The changes in the chemical shift of ^{1}H and ^{15}N of the non-phosphorylated and phosphorylated SH2c protein in Tris buffer at pH7 were compared. The chemical shift perturbations are plotted versus residue type and number. Residues were coloured in accordance to their classification of chemical shifted range to: small (0.09 < Δδ < 0.13 ppm, green), medium (0.2 > Δδ > 0.25 ppm, purple) and large (Δδ > 0.3 ppm, red). b) Mapping the chemical shift changes upon binding to pY139 C terminal onto the 3D NMR structure of the SH2 domain. The front face of the protein shows the phosphate binding pocket (Q46, T49, R51, G52, E53, and Y54), c) The back face of the protein (180° rotated view) shows the hydrophobic interface (W17:NHε1, L69, and F88). A scheme of the C-terminal tail of the SH2c protein (V96-P160 a.a) corresponding to the residue type and number are shown below the 3D structure. The SH2 domain is in grey, and the C terminal tail (N-CEGDRCY^{139}PDASSTLLPFGAS-C) is in black. pY refers to the phospho-tyrosine binding pocket and H refers to the hydrophobic binding pocket. The green and purple parts are residues that showed small or medium CSP, respectively.

6.2 Characterisation of protein interaction with a C-terminal peptide GDRCpTyrPDASST

The synthetic C-terminal phospho-ligand was titrated into the double labelled SH2 protein, to confirm the binding model of the SH2 domain to the tethered phospho-tyrosine 139 site at the C-terminal tail, investigate the structural changes upon the binding to the ligand, and determine the equilibrium dissociation constant of the binding interaction.

6.2.1 Backbone resonance assignments of the double labelled SH2 protein in Tris buffer

The amide backbone resonances of the SH2 protein were assigned in 50 mM Tris, 50 mM NaCl, 2 mM DTT and 10% D_2O at pH 6. The amide resonance assignments were based on the previous SH2 assignments in phosphate buffer at pH 6, because from the overlaid ^{15}N HSQC spectra almost all the peaks in the ^{15}N HSQC spectrum displayed identical chemical shifts, however for the slightly shifted peaks, HNCA and HNCO spectra were used to assign and confirm them.

Excluding the N-terminus and proline residues, 100 out of 105 possible amide resonances in the ^{15}N HSQC spectrum were assigned to their residues in the protein sequence (Figure 6.7.a). The unassigned residues were W17, G20, E48, Q61, and L111. Residues E48 and Q61 already displayed very low intensity peaks when they were assigned in the phosphate buffer, and residues W17 and G20 were missing.

All sidechain amide groups of Arg residues, except R51, were assigned using HCcoNH and ^{15}N NOESY spectra.

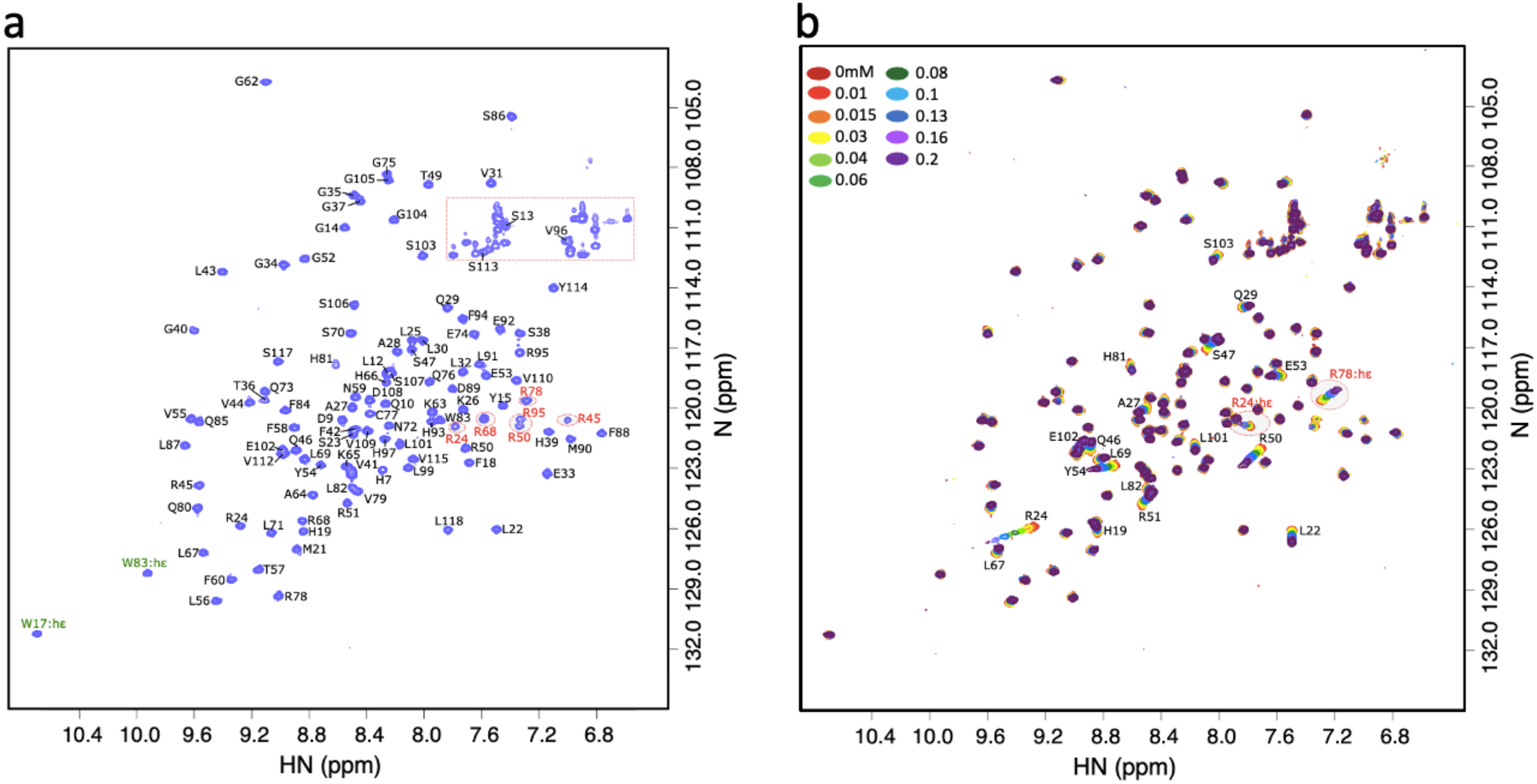


Figure 6.7. ^{1}H-^{15}N HSQC assignments of the labelled SH2 protein in Tris buffer at pH6. a) The full ^{1}H-^{15}N HSQC spectrum. The resonance assignments are labelled in black, the sidechain amide peak assignments of the Trp residues are labelled in green, and Arg sidechain peaks are in red. The sidechain peaks of Asn and Glu residues are surrounded by a red dashed box. b) Overlay of NMR titration experiments of ^{15}N HSQC spectra of 0.08 mM SH2 domain, showing the ^{1}H-^{15}N perturbations upon addition of the C terminal ligand.

6.2.2 Analysis of the binding interaction of the SH2 with C-terminal ligand

Unlabelled phospho-tyrosine peptide (11 residues, GDRCpTyrPDASST) was synthesized using the sequence in the region of Tyr139 at the C-terminal end of SH2B1 β isoform (mouse).

The ^{1}H and ^{15}N chemical shift perturbation of the labelled SH2 upon interaction with the C terminal phospho-peptide was monitored through ^{15}N HSQC spectra (Fig 6.7.b). Spectra of SH2 protein (0.08 mM) were recorded in the absence and presence of the ligand, and it was titrated with unlabelled phospho-peptide (3 mM) in molar ratio increases up to a final protein to peptide ratio of 1:2.5. In total eleven titration points were obtained for each amide peak, and the ligand was added to a final concentration of 0.2 mM, to achieve close to saturation of the protein binding site, table 2.13.

As observed in the overlaid ^{15}N HSQC spectra shown in Figure 6.7.b, upon addition of the phospho-peptide, most of the chemical shift perturbations of the SH2 signals were in the fast exchange regime. The CSP of each peak in ^{15}N HSQC spectra was recorded. A number of peaks shifted gradually upon binding to the ligand or as a consequence of the conformational change induced by the binding. This gradual change allowed for tracking the chemical shift changes for each amide resonance from free to ligand-bound states of the SH2 protein, thus allowing a transfer of assignments without the need for further experiments to assign new peaks.

The weighted average of amide chemical shift differences for each residue was calculated using equation 6.1, and the Δδ chemical shifts of each observed residue in ^{15}N HSQC spectra are shown in appendix B. Residues that have Δδ CSP value above the calculated mean plus standard deviation (threshold 0.04) were analysed, and were coloured in accordance with their classification of chemical shifted range, Figure 6.8.a.

As shown in Figure 6.8.a the largest chemical shift perturbation was observed for residue R24 (Δδ ~ 0.3 ppm) which is located at the beginning of helix α1, also there were shifts for the amide side-chain (HNε) of this residue during the titrations (Δδ 0.08 ppm) (Figure 6.7.b). These changes imply an interaction of the HN and Nε of R24 with the phospho-tyrosine C-terminal peptide. As shown in Waksman et al., 1992, Waksman et al., 1993, and Hu et al., 2006, residue R24 forms a hydrogen bond between the non-terminal nitrogen (Nε) and the phosphate oxygen of the pTyr, and another hydrogen bond between the terminal nitrogen (N) and the aromatic ring of the pTyr of JAK2 peptide. Thus the CSP for R24 is entirely consistent with the binding of pTyr in the normal SH2 binding pocket.

In addition there were medium ^{1}H-^{15}N CSP values detected for the corresponding residues that are situated at the active binding loop that connects strands β1 to β2: Q46, S47, R50, R51, E53 and Y54. In the absence of ligand, this loop is known to be in an open conformational state, whereas in the ligand-bound state its conformation changes to cover and bind to the phosphate group. Some residues in this loop such as S47 are known to have a hydrogen bond with the phosphate group of the pTyr as revealed in the crystal structure of the complex of SH2 domain with JAK2 ligand (McKercher et al., 2018). Moreover the chemical shift of the sidechain of R78 (HNε) was perturbed during the titrations (Δδ 0.1 ppm) (Figure 6.7.b), which suggests that a new sidechain NHε interaction occurred for this residue with the phospho-tyrosine peptide. As shown in Figure 6.8.b these residues are located in the active binding loop site with R24 and are clustered within the binding pocket and are key components of the phosphate-binding pocket of SH2 domain.

H19, L22, A27, and Q29 displayed a small to medium chemical shift change. These residues are all close to R24. As a consequence of the dual interactions between R24 and the phospho-tyrosine ligand, their chemical shifts can be explained from the conformational changes induced by complex formation.

On the other hand slight perturbation was observed for residue L67 which is likely due to interactions with the amino acids C-terminal to the pTyr139 of the peptide; whereas the shifts for L69 are mostly related to the conformational change. As shown in Figure 6.8.a, residues L101, E102 and S103 with L82 were among the residues that undergo changes in ^{1}H-^{15}N chemical shift upon binding. These residues are the components of the second binding pocket of SH2 which is a hydrophobic patch that includes charged residues (Figure 6.8.b). These residues, which create the hydrophobic patch, orientate the molecules for complex formation by interactions with residue Y+3 (Ala^{+3}) to pTyr in the peptide.

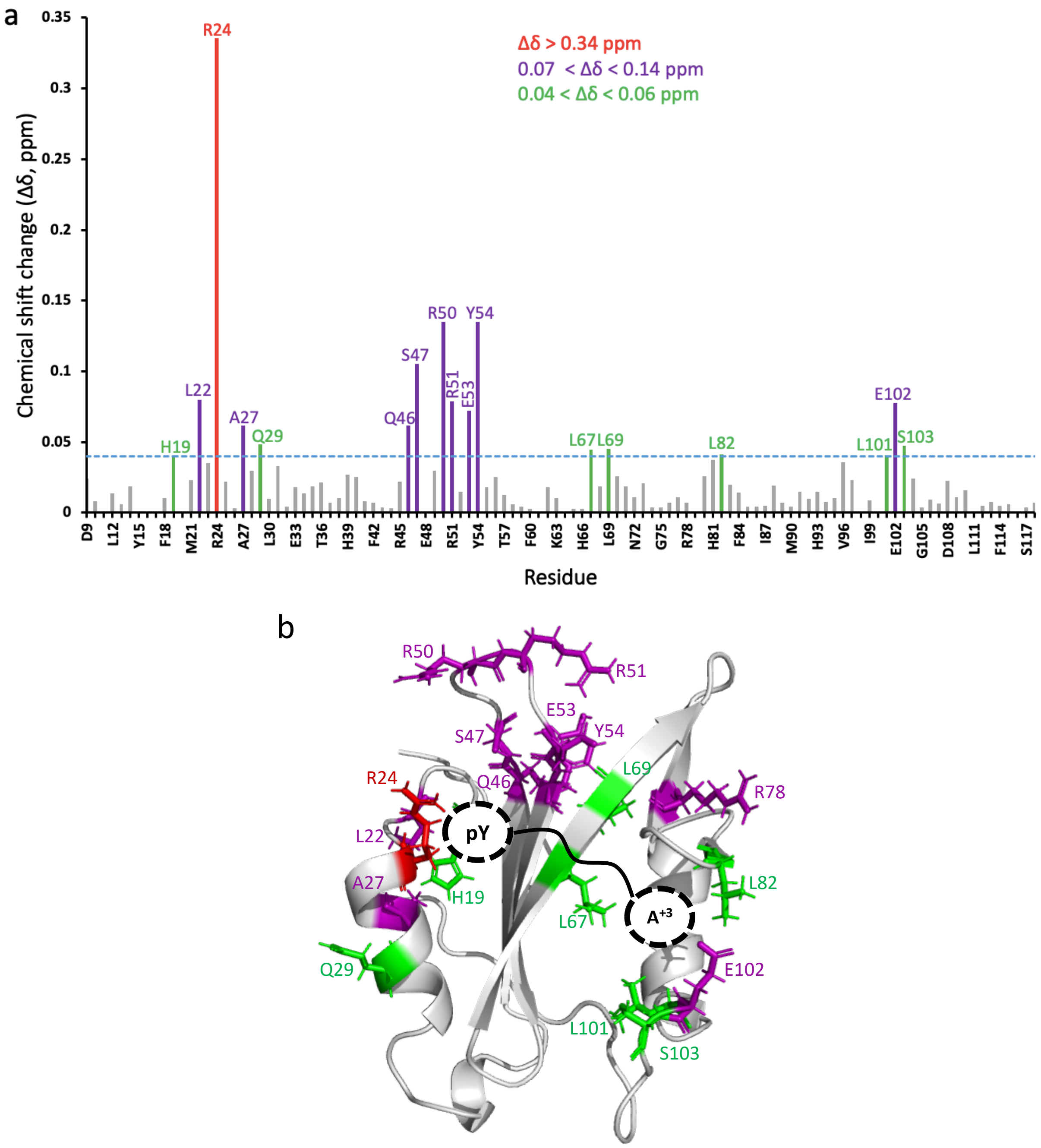


Figure 6.8. The chemical shift changes of SH2 upon addition of C-terminal phospho-tyrosine ligand. a) The perturbations in the chemical shift of ^{1}H and ^{15}N of the SH2 protein upon binding to the C-terminal ligand are plotted versus residue type and number. Residues were coloured in accordance with their classification of chemical shifted range to: small (0.04 < Δδ < 0.06 ppm, green), medium (0.07 > Δδ > 0.14 ppm, purple) and large (Δδ > 0.34 ppm, red). b) Mapping the NMR chemical shift perturbations onto the 3D NMR structure of SH2 domain, including the amide sidechain of Arg24 and Arg78. The front face of the SH2 domain shows the phosphate binding pocket (pY) and the hydrophobic binding pocket (A^{+3}).

6.2.3 Estimation of the dissociation constant (Kd) by NMR titration

The dissociation constant (K_d) for the binding of the SH2 domain to the C-terminal ligand was derived from the chemical shift perturbations ($\Delta\delta$) of SH2 during the NMR titration experiments. To measure the K_d the changes in chemical shift of SH2 domain with increasing concentration of the phospho-peptide were defined by the following equation (6.2) where P and L are the free protein and the free ligand concentrations, respectively, and PL is the concentration of protein-ligand complex.

$$P + L \underset{K_d}{\rightleftharpoons} PL \qquad \text{(Equation 6.2)}$$

Observations of the shifted residues in Figure 6.7 show the binding interactions of SH2 domain with the ligand were in the fast limit on the chemical shift timescale, which is the regime expected for weak interactions (Zuiderweg, 2002; Williamson, 2013). In this situation the ratio of free to ligand-bound protein at any titrated point, n, is proportional to the chemical shift change between the free protein (δ_i) and the complete ligand-bound state (δ_f), as shown in the following equation (6.3)

$$\Delta\delta_n/\Delta\delta_{max} = [PL]/[P]_n \qquad \text{(Equation 6.3)}$$

in which $\Delta\delta_n$ is the chemical shift difference between free and bound protein at titration point n, $\Delta\delta_{max}$ is the maximum chemical shift change which corresponds to (δ_f - δ_i), and $[PL]/[P]_n$ is the concentration ratio of the ligand-bound protein to the total protein.

The equilibrium dissociation constant is measured by connecting the known values of the protein concentration $[P]_n$ and the ligand concentration $[L]_n$ at titrated point (n) to each other by equation (6.4)

$$\Delta\delta_n = \Delta\delta_{max}/2\ [(1 + K_d/[P]_n + [L]_n/[P]_n)- \{(1+ K_d/[P]_n + [L]_n/[P]_n)^2 - 4[L]_n/[P]_n\}^{1/2}] \qquad \text{(Equation 6.4)}$$

Equation (6.4) was used to fit $\Delta\delta_n$ values against the known ligand and protein concentration by non-linear least squares to measure the K_d and $\Delta\delta_{max}$ values of the binding. The fitting saturation curve was applied using the Solver module in Excel program, (Harris, 1998).

6.2.3.1 Kd of the C-terminal ligand binding

The Δδ chemical shift changes of ^{1}H and ^{15}N (ppm) of 100 amino acid residues of SH2 were plotted against the total concentration of the ligand (mM) to obtain saturation curves (Figure 6.9). A non-linear least square algorithm with the given total protein concentration was used to determine the K_d and $\Delta\delta_{max}$ values.

The best fit K_d value was obtained by averaging the individual K_d values of a set of eight selected residues: H19, S23, L25, A28, Q29, V31, L67, and S103. Those residues were chosen as they displayed the best saturation curves (Figure 6.10). The NMR titration data for the binding of SH2 domain to the C-terminal phospho-peptide reveals an average dissociation constant (K_d) of ~0.3 mM, as expected from the observed fast chemical shift changes. This is a weak binding for an SH2/pY binding interaction, but is expected because the sequence of the peptide does not correspond well to the consensus SH2 ligand sequence, in particular because it does not possess a large hydrophobic residue at pY+3.

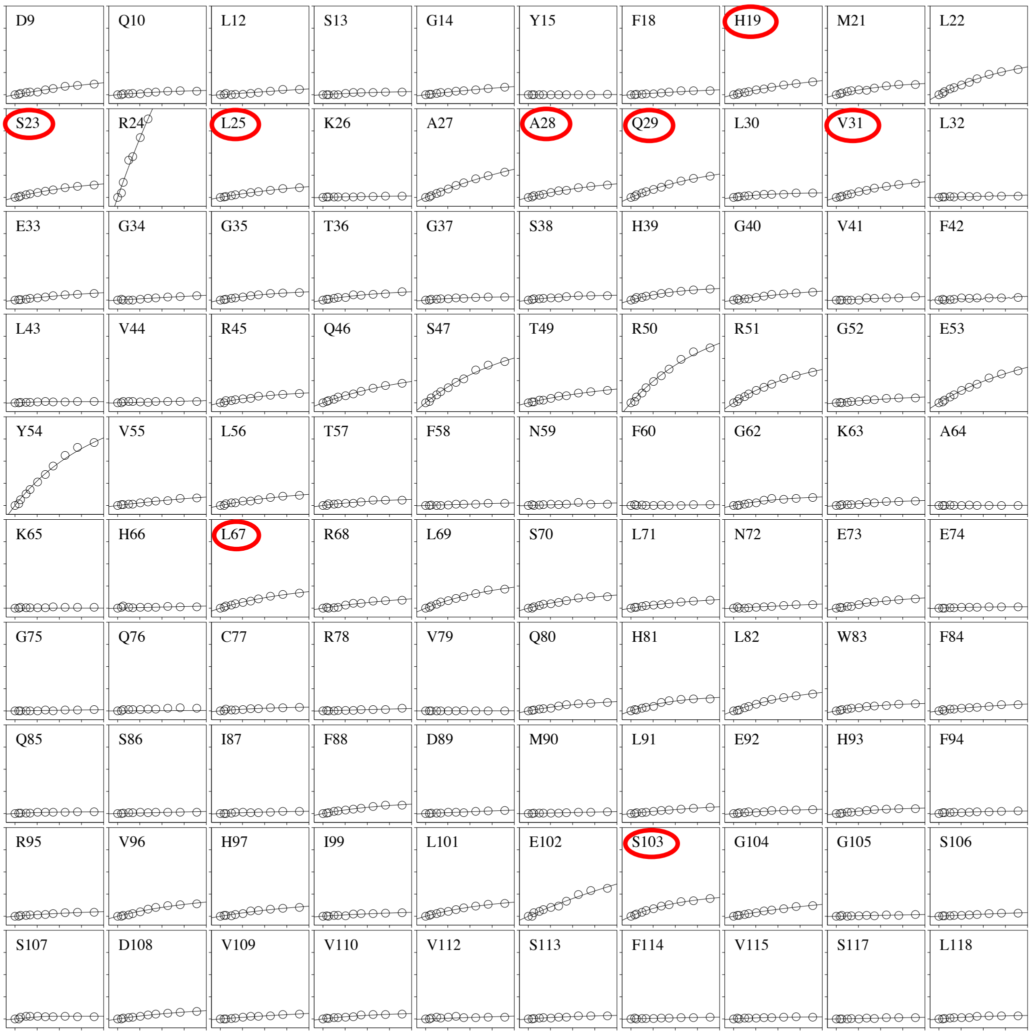


Figure 6.9. ^{1}H and ^{15}N chemical shift perturbations of individual SH2 residues upon titration with C-terminal peptide. Δδ ^{1}H-^{15}N CSP were observed for 11 titration points. The y-axis shows the chemical shift of Δδ ^{1}H-^{15}N nuclei (0.0-0.04 ppm), and the x-axis is the C terminal ligand concentration (0-0.2 mM). Eight residues were selected to measure the average binding affinity that displayed the best fitted saturation curves.

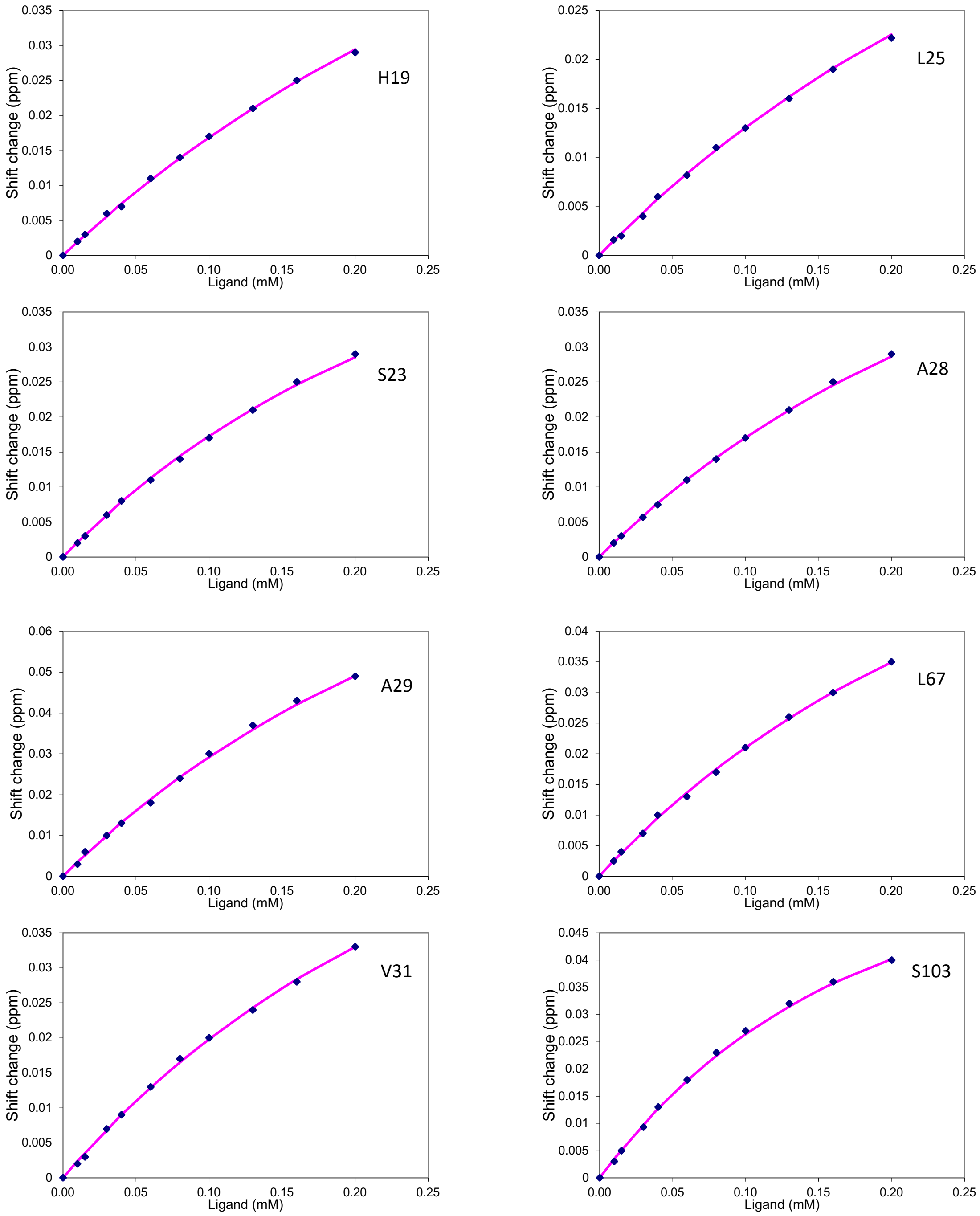


Figure 6.10. Binding affinity of selected residues of SH2 domain with addition of C-terminal ligand. The dissociation constant K_d for each residue was calculated by fitting the ^{1}H-^{15}N chemical shift perturbations of residues: H19, L25, S23, A28, A29, L67, V31, and S103. (■ observed CSP, ---- calc).

Table 6.1. The dissociation constant K_d values of a set of eight selected residues. Estimated error for the average K_d value is ±0.1, which is the standard deviation of the resulting ensemble of the K_d values.

Residue	K_d (mM)
H19	0.443
S23	0.248
L25	0.395
A28	0.294
Q29	0.297
V31	0.270
L67	0.268
S103	0.118
Average K_d	0.291

6.2.4 Competing binding between the C-terminal ligand and JAK2 ligand

A recent study (Joe et al., 2017) shows the phosphorylation of tyrosine 753 at the C terminus of SH2B1 α isoform regulates the function of the whole protein, although the regulation mechanism is still unclear. According to that it has been hypothesised that the binding ability of the SH2 domain of SH2B1 β isoform is repressed via binding to the tethered C terminal phospho-tyrosine 139, but that the repressed state of SH2 is disrupted by outcompeting with its high affinity substrate, Janus kinase 2 (JAK2) peptide. This hypothesis is built on the fact that the binding affinity of SH2-JAK2 (K_d 0.320 μM) (McKercher et al., 2018) is ~ 1000 times stronger than the estimated binding affinity for C-terminal peptide (K_d 300 μM).

To test the hypothesis of competitive binding between the C-terminal ligand (GDRCpTyr^{139}PDASST) and the JAK2 ligand (FTPDpTyr^{148}ELLTEN) to the SH2 domain, an outcompeting binding experiment was performed. The complex of SH2-C-terminal peptide (0.08 mM protein and 0.2 mM ligand) was titrated with unlabelled synthetic JAK2 ligand peptide (0.5 mM stock) in five titration points up to the final JAK peptide concentration of

0.02 mM, Table 2.14. The concentration of the C terminal peptide was ten times more than the JAK2 peptide. The outcompeting binding was monitored via chemical shift changes arising upon binding using ^{15}N HSQC spectra (Figure 6.11). There were gradual shifts and disappearances for a number of peaks upon binding to the JAK2 ligand or as a consequence of the conformational change. This does not provide a quantitative estimate of the affinity of the JAK2 ligand, but it does show that it is capable of displacing the C-terminal peptide and that the affinity is at least ten times stronger than the affinity for C-terminal peptide.

As proven in different studies, SH2 domain has a two-pronged mechanism of phospho-tyrosine peptide recognition. First, the N-terminal phosphate binding pocket in which both ligands (C-terminal and JAK2) are expected to compete for similar residues or to bind to similar residues within the active binding site. Second, the C-terminal binding pocket which is known to bind to the third residue after pTyr, however due to the diversity in the sequence C-terminal to pTyr of both peptides, they are expected to bind differently to residues that are situated within the binding site.

The crystal structure of the complex of SH2 domain with JAK2 ligand reveals that residues R24, R45 and R68 create the positively charged recognition site of the phosphate-binding pocket, as their sidechains are hydrogen bonded to the phospho-tyrosine (McKercher et al., 2018). In addition to that, the phosphate group is stabilised via a limited hydrogen bonding network involving residues S47, E48, and T49, which are located within the active binding loop that connects strand β1 to β2, as shown in Figure 6.12. Thus, recognition of the phosphate group involves the same interactions for both ligands.

On the other hand, previous observations of CSP upon binding SH2 to C-terminal ligand reveal there were slightly different mechanisms to recognise and bind to the phospho-tyrosine. The CSP data of R24 with other residues on the active binding loop site (Q46, S47, R50, R51, E53, Y54) suggest these are residues involved in the interaction with the phospho-tyrosine. Furthermore the positively changed sidechain interaction of R45 and R68 with the phosphate group which are found in the SH2-JAK2 complex are replaced by the sidechains of R24 and R78 in the SH2-C terminal complex.

The differences in the phospho-tyrosine binding interface of SH2 for the two peptides was used to prove competitive binding between the C-terminal and JAK2 peptides. Residues R45 and T49 are involved in binding with the phosphate group in JAK2 but not in C-terminal. Also those residues can be identified easily in the ^{15}N HSQC spectrum as they appear in a well separated region of the spectrum, thus allowing for tracking their chemical shift changes upon outcompeting binding. As shown in Figure 6.11, the amide signals of R45 and T49 displayed down-field and up-field shifts, respectively, upon binding to the JAK2 ligand. The shift changes of these signals demonstrates that the JAK2 ligand successfully outcompetes C-terminal. In other words, JAK2 binds more strongly than C-terminal, and binding of JAK2 is likely to be able to displace the bound C-terminus *in vivo*, especially in the presence of phosphate in solution. As shown the high concentration of phosphate ions in the NMR sample saturated the pY binding pocket in the non-phosphorylated SH2 domain; and for the phosphorylated protein, even though the phosphorylation of Tyr139 happened, the phosphate ions were able to outcompete the pY139 binding, which is expected to have weak affinity as discussed before.

It is expected that the conserved residues such as R45 within the active binding pocket in the SH2 domain participate in the phosphate binding, suggesting that the recognition of pTyr will be conserved among SH2 domains. However the observed results indicate that the binding of the phosphate group via the phosphate-binding pocket is controlled by the diversity of the ligand sequence. Therefore the sequence of the phospho-tyrosine peptide is not just critical to account for differing affinities of SH2 domain toward substrates, but also controls the binding orientation of pTyr into the binding cavity in SH2 domain. R45 can be considered as the center and base of the binding pocket, due to the dual hydrogen bonding to the oxygens of the phosphate group.

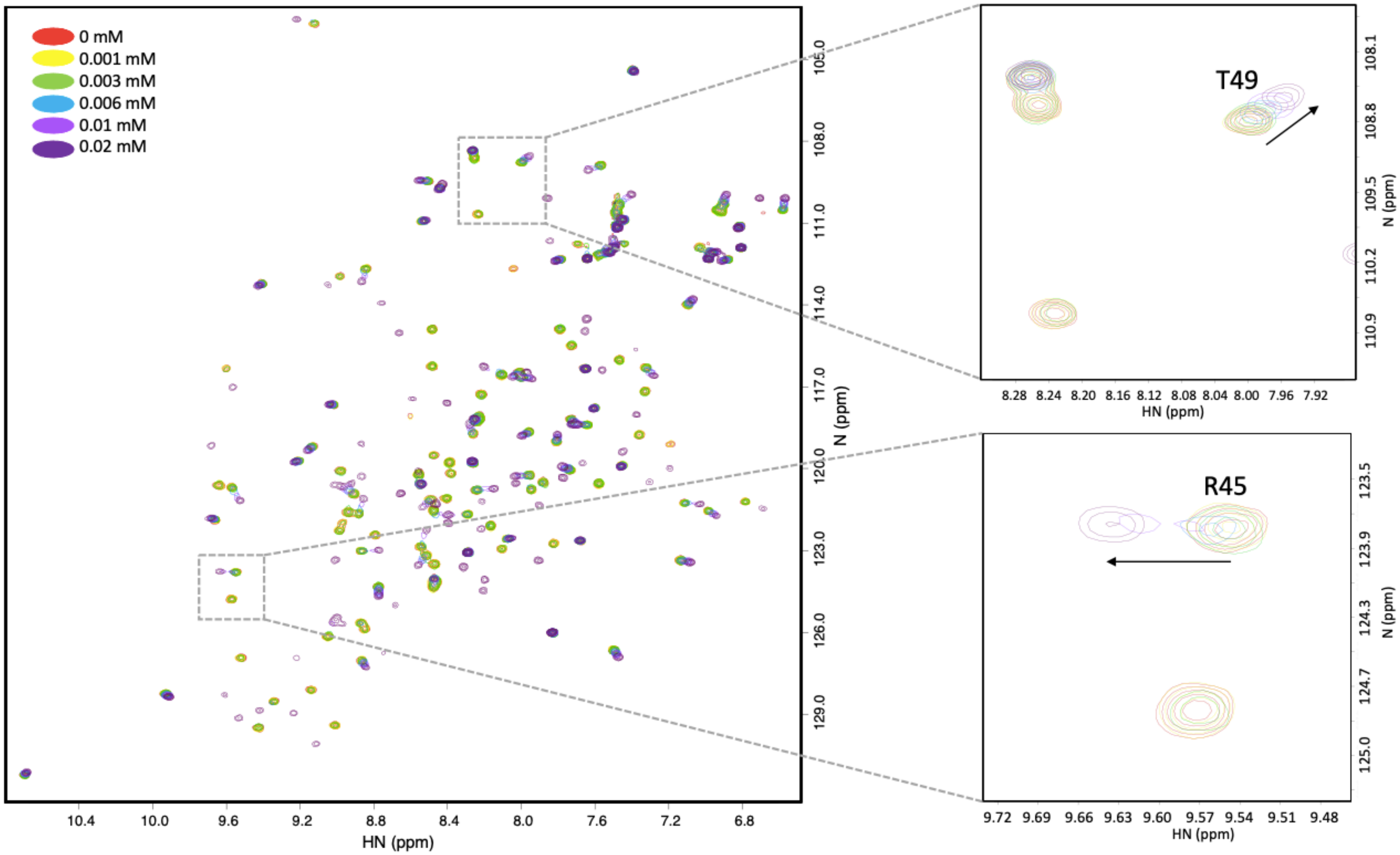


Figure 6.11. Overlaid ^{15}N HSQC spectra of competition experiments of SH2-C terminal complex upon addition of the JAK2 ligand. The ^{1}H-^{15}N chemical shift perturbations show expansions of R45 and T49 upon binding to JAK2 ligand, surrounded by grey dashed boxes.

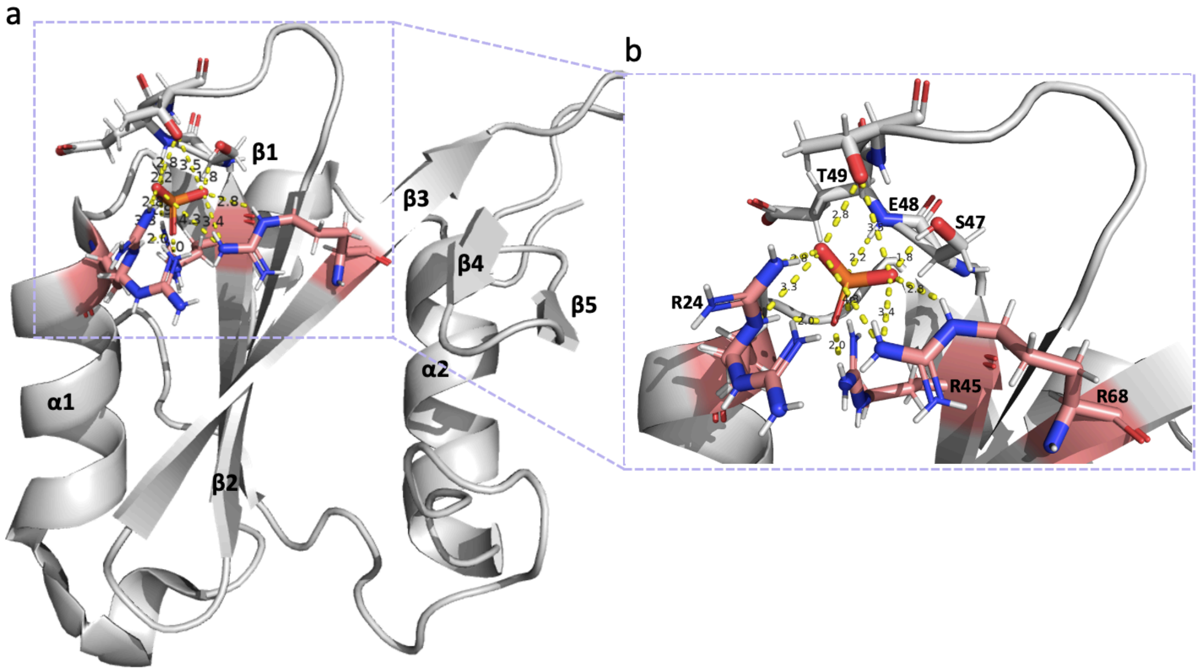


Figure 6.12. Crystal structure of the complex of SH2 domain and the phosphate group in JAK2 peptide (5w3r) as illustrated in McKercher et al., 2018. a) The SH2 domain components are labelled in black. b) An expansion of the phosphate binding pocket showing residues (R24, S47, E48, T49, R45, and R68, coloured by elements) that are involved in interaction with the phosphate group of Tyr in JAK2 peptide, with measured distance in yellow dash lines. The positively charged sidechains of Arg residues are coloured in pink by elements. The red and orange coloured molecule is the phosphate group of Tyr.

6.3 Discussion

In this study it has been demonstrated that SH2 domain binds to the phospho-tyrosine of the tethered C-terminal tail and phospho-tyrosine in the C-terminal peptide via the phosphate-binding pocket, with some differences in residues involved or cooperating in phosphate binding.

The initial binding interaction studies were carried out in phosphate buffer. A surprising observation in these studies was that there appeared to be very little interaction between the pY and the SH2 phosphate binding pocket. The amino acid sequence of the C-terminus does not match the consensus binding sequence for SH2 domains, but the pY is tethered to the SH2 domain, so it was expected that there would be significant interaction, because of the intramolecular nature of the binding. However, very little binding was observed in 100 mM phosphate buffer. When the buffer was changed, using Tris instead of phosphate, a significant interaction was observed, matching the expected interaction. It therefore appears that 100 mM phosphate is able to compete successfully with the binding of the C-terminal tail. The physiological concentration of phosphate in eukaryotic cells is approximately 2.1 mM, (Xu et al., 2019). It is concluded that the binding of the C-terminal tail of SH2B1 β isoform to the SH2 domain is finely tuned so that in physiological phosphate the tail binds to the SH2 domain but can be easily displaced by consensus ligands such as JAK.

The NMR titration results indicate that SH2 domain follows a similar mechanism to those seen for other SH2 domains to bind to the phospho-tyrosine peptide via a two-pronged model. SH2 domain recognises and binds the phospho-peptide via two binding pockets: the N-terminal phosphate-binding pocket and the C-terminal hydrophobic pocket. According to CSP data residues R24, Q46, S47, R50, R51, E53, and Y54 are clustered within the N terminal binding site and involved in the interactions with the phosphate group of Tyr in the peptide, whereas residues L82, L101, E102, and S103 are situated within the C terminal binding site and create hydrophobic network interactions with Ala^{+3} which is a small hydrophobic residue. As shown in Figure 6.8.b the C-terminal peptide crosses the front of the central β sheet of the SH2 domain.

On the other hand the observations from binding the SH2 domain to tethered pTyr139 reveal a different mechanism of binding to the phospho-tyrosine peptide. As shown in Figure 6.6, more residues that are situated within the active binding site participate in the binding interaction with the phosphate group or affected by conformational change such as T49 and G52. Residues R24, S47 and R50 were not assigned in the complex, and thus it is not clear if they bind to the phosphate group of the tethered pTy139. Other residues (Q46, R51, E53, Y54), which displayed CSP values above the threshold, are clustered within the phosphate binding site in the SH2 domain.

Although the binding of SH2 domain to the phosphate group of the tethered pTyr139 site is similar, in that the phosphate binds to the same residues, there is a different hydrophobic binding interface which consists of residues L69, F88 and sidechain NHε1 of W17. Moreover there were no shifts for residues L82, L101, and E102 (S103 is unassigned), which are the components of the hydrophobic pocket as shown in the titration result for the SH2 domain with the C-terminal peptide.

By contrast, the binding of the synthetic C-terminal peptide GDRCpTyr^{139}PDASST does seem to affect the ‘consensus’ binding residues, namely the phosphate binding site and the hydrophobic patch that includes residues L101, E102 and S103. In other words, the synthetic peptide binds in the consensus position, but the full-length tethered C-terminal tail does not. An obvious difference between these two ligands is that the full-length tail contains several hydrophobic amino acids that come immediately after the C-terminal end of the synthetic ligand. Therefore it has been proposed that residues L69, F88 and NHε1 of W17 interact with one or more of the extended hydrophobic residues (Y^{139}PDA^{142}SST$\underline{\text{LL}}$P$\underline{\text{F}}$) at the C-terminal tail (Figure 6.3), which are not found in more standard SH2 ligands. A number of studies have identified that the only residues that contact SH2 domains are the pTyr with the following three residues in the peptide (Bradshaw and Waksman, 2003). However, the normal β C-terminus only has alanine at Y+3, which is not a large enough residue to interact properly with the hydrophobic pocket. The existence of the different binding mechanism of SH2 to the extended amino acids after the tethered pTyr139 indicates an important role for the diversity of the tail residue type. It has been proposed that the extended hydrophobic tail of the β isoform (L^{+7}, L^{+8}, F^{+10}) forms a different recognition motif, which causes the C-terminal tail to fold round the ‘side’ of SH2 and contact a hydrophobic region on the ‘back’ face (Figure 6.6).

Chapter 7 General discussion and future work

The SH2B1 protein is found in several signal transduction pathways including those downstream of leptin and insulin, which highlights its functional importance as well as its importance as a therapeutic target involved in specific diseases such as obesity and type 2 diabetes.

This thesis focuses on studying the structure and the intramolecular regulation model of the SH2 domain of SH2B1 β. The structural study used the SH2 domain as a model system to investigate the possibility of solving the solution protein structure using the automatic structure calculation program CYANA, with the new promising validation method ANSURR.

Although SH2B1 isoforms share the same common functional and structural domains, there is less known about the function of the intrinsically disordered C-termini, which is where the different isoforms differ. Characterisation of the biological role of the C-terminal tail will provide greater understanding of the functional diversity of the isoforms. Our investigation shows that the C-terminus of the β isoform contains a phosphorylatable tyrosine which may carry a biological function when it is phosphorylated, because it binds to the nearby SH2 domain and therefore inactivates it, and the binding is outcompeted by a preferred phospho-tyrosine ligand, which displaces the C-terminal tail and creates a functional binding site.

7.1 Expression of monomeric SH2 protein

Initial experiments described in Chapter 3 attempted to express the SH2 protein using the consensus boundaries of an SH2 domain, as defined in Uniprot. The protein was expressed with and without an MBP tag, however our results proved that this version of the protein is unstable and not well folded, and thus tends to form homo-oligomers. Addition of a few residues at the C and N termini of the consensus SH2 sequence, based on sequence alignment with other SH2-containing proteins, led to the expression of a monomeric and well behaved protein. The extra residues at the both ends look to form an integral part of the protein structure and are thus required for correct folding and stability. Our structure, presented in Chapter 5, confirms this, as does the crystal structure.

7.2 Determination of the NMR structure of SH2 using fully automatic software

Once the protein had been assigned (Chapter 4), the NMR structure of SH2 protein was determined following an established framework as shown in chapter 5. Using available software for automated structure calculation can reduce the human intervention but does not eliminate it. The automated structure calculation relies basically on the quality of the input data, specifically both the complete chemical shifts list and the NOE peak lists.

Although the CYANA algorithm can handle incomplete chemical shift assignments, it still requires approximately 90% complete chemical shift assignment for reliable automated NOE assignment. That is because an almost complete and accurate chemical shift assignment is critical in NOE assignment as any assignment errors or gaps cause distortion in the resulting structure. In this case a manual intervention is needed.

The automated chemical shift assignment for SH2 protein was obtained firstly using FLYA, then it was checked by manual assignment, in order to create a list with a higher percentage of correct assignments, as discussed in Chapter 4. The automated backbone assignment was nearly correct, but the sidechain assignment had a number of errors. These results highlight the importance of manually checking all assignments before proceeding to the structure calculation.

Beside a complete chemical shift assignment, a good NOE peak list with a low number of noise peaks is required as well for an accurate NOE assignment. Here the second limitation is NOE peak lists as they contained a big number of unreal peaks, such as noise and spectral artifacts, which may interfere with the real peaks or possibly replace missing assignments during NOE assignment by CYANA. In this case it was hard to remove noise peaks from the experimental NOE data, which led to CYANA having problems to achieve a good NMR structure. Therefore a new framework was established based on repeating CYANA calculations, and then using the CYANA output restraints with additional hydrogen bonds derived from temperature coefficients to add more H-bond restraints to subsequent calculations. The additional constraints in each CYANA calculation improved the resulting structure.

A new validation method for structure accuracy called ANSURR was used to check the calculated structure. This method which was created by Nicholas Fowler in our lab, is based on making a relationship between the calculated structure and the experimental data. The resultant structure was checked in every stage throughout the calculations: CYANA runs, CNS calculation, and water refinement. ANSURR scores showed improvement in the resultant structure at each stage. At the end ANSURR scores were combined with the total energy of the structure to select the final ensemble of structures that represent the NMR structure of the SH2 domain. This may be a useful general method for producing a better NMR ensemble.

7.3 Phosphorylation of SH2c and mutant SH2c Y114F by activated kinase

The phosphorylation by Fer kinase targeted tyrosine 139 at the C-terminal tail of SH2c protein. There were two limitations while doing the phosphorylation: an unexpectedly rapid phosphorylation of tyrosine 114 at the end of the SH2 domain, and oxidation of cysteines residues in the C-terminal tail which leads to the formation of disulphide bonds and caused protein precipitation, as described in chapter 3. The oxidation of cysteines results from reducing the DTT concentration in the reaction buffer due to the presence of Mn^{+2} ions.

To overcome those limitations, tyrosine 114 was mutated to phenylalanine, and to compensate for the loss of DTT, more DTT was added every hour during the incubation time. Addition of more DTT during the incubation was still not enough to overcome the rapid oxidation of cysteines, therefore the protein still formed disulphide bonds and precipitated. Despite the loss of protein, phosphorylated protein was obtained at low concentration but in adequate amounts for NMR studies.

7.4 The self-regulation mechanism of SH2B1 β protein

It has been hypothesised that the SH2 protein might be inhibited by weak interaction with its tethered phospho-tyrosine site 139 at the C-terminus of the β isoform, and that the autoinhibited state might be disrupted by an outcompeting interaction with good ligands such as JAK2. To investigate the hypothesis two procedures are needed; demonstrate an intramolecular interaction between the SH2 and pTyr139, and measure the binding affinity to compare with the preferred ligand.

As explained in chapter 6, our results indicate that SH2 domain binds to pTyr139 at the C-terminal tail via the phosphate-binding pocket. The intramolecular interaction was outcompeted by phosphate ions in the buffer, which suggests a very weak binding affinity. The C-terminal tail of the β isoform does not contain the typical hydrophobic residue 3 residues after the pY, but it does contain hydrophobic residues 7, 8 and 10 residues after the pY, which weakens the binding. On the other hand there was a remarkable result in Chapter 6 showing that there is a new binding interface for SH2 consisting of L69, F88 and NHε1 of W17, which may interact with the extended hydrophobic residues after pY139 (N----GDRCpTyr^{139}PDASSTLLPFGASD----C).

Furthermore titration of SH2 protein with a short C-terminal phospho-tyrosine peptide without the extended region (GDRCpTyr^{139}PDASST) reveals the expected SH2 binding behaviour to the phospho-tyrosine ligand with slightly different cooperating residues in both pockets: binding to the phosphate group via the N-terminal phosphate binding pocket, and binding to a hydrophobic residue at the C-terminal hydrophobic pocket. The estimated binding affinity between the SH2 and C-terminal peptide was shown to be weak (~300 µM) compared to the binding affinity to the preferred ligand JAK2 (~0.320 µM). Binding to the full-length C-terminal peptide will be stronger, meaning that the weak binding of the C-terminal tail prevents binding of the SH2 domain to unphosphorylated JAK2.

7.5 Future work

7.5.1 Further investigations

7.5.1.1 Prevention of precipitation of the SH2c Y114F protein during phosphorylation

To obtain a high quantity of phosphorylated protein sample and do further investigations, the protein precipitation during the phosphorylation needs to be stopped.

As explained in Chapter 3 the three cysteine residues at the C-terminus were reduced by adding 1.2 mM DTT into the phosphorylation buffer during the phosphorylation. Even with adding more DTT during the incubation time, still the presence of Mn^{+2} ions in the kinase buffer greatly decreased the stability of DTT. Reducing the concentration of DTT in the reaction buffer caused cysteine residues to be oxidised and form a cross disulphide bond with another cysteine residue, sequentially reducing the stability of the protein and causing precipitation.

It is possible that this problem could be removed by using TCEP instead of DTT, which has an advantage over DTT that it is not affected by Mn^{2+} ions in the buffer which might consequently lead to stopping the protein precipitation. Another option would be to mutate cysteine residues to Alanine to avoid problems arising from DTT reduction and oxidation.

Obtaining a good quantity of protein sample is required to obtain a good NMR spectrum which is needed for further studies, such as to make a complete chemical shift assignment of SH2c while binding to pY139 and using the assignment to find out if the C-terminal sequence undergoes a disorder to order transition upon binding SH2 to pY139; and to define which residues interact with the C-terminal ligand using NOEs.

7.5.1.2 Measurement of the exact Kd

Investigation of the binding of SH2 domain to the phospho-tyrosine of tethered C-terminal peptide, described in Chapter 6, reveals that the phosphate group of the tethered pTyr139 site binds to the phospho-binding pocket in SH2 domain similarly to other SH2 domains, however the extended hydrophobic tail of the β isoform (L^{+7}, L^{+8}, F^{+10}) forms a different recognition motif, which causes the C-terminal tail to fold round the 'side' of SH2 and contact a hydrophobic region on the 'back' face, which is a new pocket.

The binding affinity between the phospho-tyrosine C terminal ligand and SH2 domain is expected to be weak as it is outcompeted by phosphate ions in the buffer. However the exact Kd should be measured and compared with the affinity of the preferred ligand JAK2, by using a long pY C-terminal ligand including the bulky hydrophobic sequence (L^{+7}, L^{+8}, F^{+10}).

7.5.2 New proposals

7.5.2.1 Cysteine residues at C-terminal tail may have a role in dimerization via forming cross disulphide bonds

It has been hypothesised that the phosphorylated tyrosine 139 might interact with its own SH2 domain, as an internal regulatory mechanism. There was a remarkable observation described in Chapter 6, which is that the phosphorylation of Tyr139 is strongly dependent on the redox state of the solution. There are three cysteines close together in the C-terminal end of the β isoform (C133, C138 and C154), and our results suggest that the tyrosine only gets phosphorylated when the cysteines are reduced. This could be a novel regulatory mechanism for SH2B1 function (β isoform), which needs to be investigated.

7.5.2.2 Possible function for the phosphorylated Tyr114 at the end of SH2

In Chapter 3 our results show surprisingly that Tyr114 at the end of the SH2 domain was phosphorylated rapidly, even though it does not look like a consensus sequence for Fer kinase. There was considerable precipitation of the protein after phosphorylation of Tyr114. It has been speculated that pTyr114 may create a binding site for another SH2 domain, that would start early and increase during the incubation time leading it to form large molecular assemblies and cause protein aggregation. In vivo, phosphorylation of Tyr114 could possibly produce protein dimerization, which is known to be important for example for binding and activating JAK2.

These comments suggest that Tyr114 may be phosphorylated *in vivo* (Huang et al., 2005), thus it may have a key function. This observation needs to be studied further.

References

Ahmed, Z. and Pillay, T. S. (2001). Functional Effects of APS and SH2-B on Insulin Receptor Signalling. *Biochemical Society Transactions*, 29, 529–534.

Babu, M. M. (2016). The Contribution of Intrinsically Disordered Regions to Protein Function, Cellular Complexity, and Human Disease. *Biochemical Society Transactions*, 44(5), 1185–1200.

Bah, A., Vernon, R. M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., Sonenberg, N., Kay, L. E. and Forman-Kay, J. D. (2015). Folding of an Intrinsically Disordered Protein by Phosphorylation as a Regulatory Switch. *Nature*, 519, 106–109.

Baltensperger, K., Karoor, V., Paul, H., Ruoho, A., Czech, M. P. and Malbon, C. C. (1999). The β -Adrenergic Receptor Is a Substrate for the Insulin Receptor Tyrosine Kinase. *Journal of Biological Chemistry*, 271, 1061–1064.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997). GARANT - A General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra. *Journal of Computational Chemistry*, 18(1), 139–149.

Bax, A., Clore, G. M. and Gronenborn, A. M. (1990). ^{1}H- ^{1}H Correlation via Isotropic Mixing of ^{13}C Magnetization, a New Three-Dimensional Approach for Assigning ^{1}H and ^{13}C Spectra of ^{13}C-Enriched Proteins. *Journal of Magnetic Resonance*, 88(2), 425–431.

Bax, A. and Grzesiek, S. (1993). Methodological Advances in Protein NMR. *Accounts of Chemical Research*, 26(4), 131–138.

Baxter, N. J. and Williamson, M. P. (1997). Temperature Dependence of ^{1}H Chemical Shifts in Proteins. *Journal of Biomolecular NMR*, 9(4), 359–369.

Berjanskii, M. V. and Wishart, D. S. (2005). A Simple Method to Predict Protein Flexibility Using Secondary Chemical Shifts. *Journal of the American Chemical Society*, 127(43), 14970–14971.

Bienkiewicz, E. A. and Lumb, K. J. (1999). Random-Coil Chemical Shifts of Phosphorylated Amino Acids. *Journal of Biomolecular NMR*, 15(3), 203–206.

Blencowe, B. J. (2006). Alternative Splicing: New Insights from Global Analyses. *Cell*, 126(1), 37–47.

Blume-Jensen, P. and Hunter, T. (2001). Oncogenic Kinase Signalling. *Nature*, 411(6835), 355–365.

Bradshaw, J. and Waksman, G. (2003). Molecular Recognition by SH2 Domains. *Advances in Protein Chemistry*, 61, 161–210.

Braun, W. (1987). Distance Geometry and Related Methods for Protein Structure Determination from NMR Data. *Quarterly Reviews of Biophysics*, 19(3–4), 115–157.

Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, P. W. (2015). *Molecular Biology of the Cell*. Garland Science, Taylor and Francis Group.

Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grossekunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998). Crystallography and NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5), 905–921.

Cartwright, H. (2015). Artificial Neural Networks. *New York: Springer Protocols*.

Charrier, J. G. and Anastasio, C. (2012). On Dithiothreitol (DTT) as a Measure of Oxidative Potential for Ambient Particles: Evidence for the Importance of Soluble Transition Metals. *Atmospheric Chemistry and Physics Discussions*, 12(5), 11317–11350.

Chen, L. and Carter-Su, C. (2004). Adapter Protein SH2-Bβ Undergoes Nucleocytoplasmic Shuttling: Implications for Nerve Growth Factor Induction of Neuronal Differentiation. *Molecular and Cellular Biology*, 24(9), 3633–3647.

Cornilescu, G., Delaglio, F. and Bax, A. (1999). Protein Backbone Angle Restraints from Searching a Database for Chemical Shift and Sequence Homology. *Journal of Biomolecular NMR*, 13(3), 289–302.

Cornilescu, G., Ramirez, B. E., Frank, M. K., Clore, G. M., Gronenborn, A. M. and Bax, A. (1999). Correlation between ${}^{3h}J_{NC'}$ and Hydrogen Bond Length in Proteins. *Journal of the American Chemical Society*, 121(26), 6275–6279.

Craparo, A., Freund, R. and Gustafson, T. A. (1997). 14-3-3 (ε) Interacts with the Insulin-like Growth Factor I Receptor and Insulin Receptor Substrate I in a Phosphoserine-dependent Manner. *Journal of Biological Chemistry*, 272(17), 11663–11670.

Devalliére, J. and Charreau, B. (2011). The Adaptor Lnk (SH2B3): An Emerging Regulator in Vascular Cells and a Link between Immune and Inflammatory Signaling. *Biochemical Pharmacology*, 82(10), 1391–1402.

Doche, M. E., Bochukova, E. G., Su, H. W., Pearce, L. R., Keogh, J. M., Henning, E., Cline, J. M., Dale, A., Cheetham, T., Barroso, I., Argetsinger, L. S., O'Rahilly, S., Rui, L., Carter-Su, C.

and Farooqi, I. S. (2012). Human SH2B1 Mutations Are Associated with Maladaptive Behaviors and Obesity. *Journal of Clinical Investigation*, 122(12), 4732–4736.

Dodington, D. W., Desai, H. R. & Woo, M. (2018). JAK/STAT – Emerging Players in Metabolism. *Trends in Endocrinology and Metabolism*, 29(1), 55–65.

Duan, C., Li, M. and Rui, L. (2004). SH2-B Promotes Insulin Receptor Substrate 1 (IRS1)- and IRS2-Mediated Activation of the Phosphatidylinositol 3-Kinase Pathway in Response to Leptin. *Journal of Biological Chemistry,* 23(1), 1–7.

Duan, C., Yang, H., White, M. F. and Rui, L. (2004). Disruption of the SH2 - B Gene Causes Age-Dependent Insulin Resistance and Glucose Intolerance . *Molecular and Cellular Biology*, 24(17), 7435–7443.

Eck, M. J., Dhe-Paganon, S., Trüb, T., Nolle, R. T. and Shoelson, S. E. (1996). Structure of the IRS-1 PTB Domain Bound to the Juxtamembrane Region of the Insulin Receptor. *Cell*, 85(5), 695–705.

Englander, S. W. and Mayne, L. (1992). Using Hydrogen-Exchange Labeling and Two - Dimensional NMR. *Annual Reviews*, 21, 243–265.

Flores, A., Argetsinger, L. S., Stadler, L. K. J., Malaga, A. E., Vander, P. B., DeSantis, L. C., Joe, R. M., Cline, J. M., Keogh, J. M., Henning, E., Barroso, I., de Oliveira, E. M., Chandrashekar, G., Clutter, E. S., Hu, Y., Stuckey, J., Sadaf Farooqi, I., Myers, M. G. and Carter-Su, C. (2019). Crucial Role of the SH2B1 PH Domain for the Control of Energy Balance. *Diabetes*, 68(11), 2049–2062.

Fowler, N. J., Sljoka, A. and Williamson, M., P. (2020). A Method for Validating the Accuracy of NMR Protein Structures. *Nature Communications*, 11(6321).

Fritsche, L., Weigert, C., Häring, H.-U. and Lehmann, R. (2008). How Insulin Receptor Substrate Proteins Regulate the Metabolic Capacity of the Liver--Implications for Health and Disease. *Current medicinal chemistry*, 15(13), 1316–1329.

Getz, E. B., Xiao, M., Chakrabarty, T., Cooke, R. and Selvin, P. R. (1999). A Comparison between the Sulfhydryl Reductants Tris (2-Carboxyethyl) Phosphine and Dithiothreitol for Use in Protein Biochemistry 1. *Analytical Biochemistry*, 80, 73–80.

Gibbs, E. B., Cook, E. C. and Showalter, S. A. (2017). Application of NMR to Studies of Intrinsically Disordered Proteins. *Archives of Biochemistry and Biophysics*, 628, 57–70.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A. and Smith, H. O. (2009). Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases. *Nature*

Methods, 6(5), 343–345.

Güntert, P. (2003). Automated NMR Protein Structure Calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43(3–4), 105–125.

Güntert, P. (2004). Automated NMR Structure Calculation with CYANA. *Methods in molecular biology*, 278, 353–378.

Güntert, P. (2009). Automated Structure Determination from NMR Spectra. *European Biophysics Journal*, 38(2), 129–143.

Güntert, P. and Buchner, L. (2015). Combined Automated NOE Assignment and Structure Calculation with CYANA. *Journal of Biomolecular NMR*, 62(4), 453–471.

Habchi, J., Tompa, P., Longhi, S. and Uversky, V. N. (2014). Introducing Protein Intrinsic Disorder. *Chemical Reviews*, 114, 6561–6588.

Harris, D. C. (1998). Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver. *Journal of Chemical Education*, 75(1), 119-121.

Harrison, S. C. (1996). Peptide-Surface Association: The Case of PDZ and PTB Domains. *Cell*, 86(3), 341–343.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002a). Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology*, 319(1), 209–227.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002b). Protein NMR Structure Determination with Automated NOE-Identification in the NOESY Spectra Using the New Software ATNOS. *Journal of Biomolecular NMR*, 24(3), 171–189.

Holgado-Madruga, M., Emlet, D. R., Moscatello, D. K., Godwin, A. K. and Wong, A. J. (1996). A Grb2-Associated Docking Protein in EGF- and Insulin-Receptor Signalling. *Nature*, 379(2), 560–564.

Hu, J. and Hubbard, S. R. (2006). Structural Basis for Phosphotyrosine Recognition by the Src Homology-2 Domains of the Adapter Proteins SH2-B and APS. *Journal of Molecular Biology*, 361(1), 69–79.

Hu, J., Liu, J., Ghirlando, R., Saltiel, A. R. and Hubbard, S. R. (2003). Structural Basis for Recruitment of the Adaptor Protein APS to the Activated Insulin Receptor. *Molecular Cell*, 12(6), 1379–1389.

Huang, B., Liu, Y., Yao, H. and Zhao, Y. (2020). NMR-Based Investigation into Protein Phosphorylation. *International Journal of Biological Macromolecules*, 145, 53–63.

Huang, H. Da, Lee, T. Y., Tzeng, S. W. and Horng, J. T. (2005). KinasePhos: A Web Tool for Identifying Protein Kinase-Specific Phosphorylation Sites. *Nucleic Acids Research*, 33(2), 226–229.

Ikeya, T., Ikeda, S., Kigawa, T., Ito, Y. and Güntert, P. (2016). Protein NMR Structure Refinement Based on Bayesian Inference. *Journal of Physics*, 699(1).

Ikura, M., Kay, L. E. and Bax, A. (1991). Improved Three-Dimensional ^{1}H-^{13}C-^{1}H Correlation Spectroscopy of a ^{13}C-Labeled Protein Using Constant-Time Evolution. *Journal of Biomolecular NMR*, 1(3), 299–304.

Joe, R. M., Flores, A., Doche, M. E., Cline, J. M., Clutter, E. S., Vander, P. B., Riedel, H., Argetsinger, L. S. and Carter-Su, C. (2017). Phosphorylation of the Unique C-Terminal Tail of the Alpha Isoform of the Scaffold Protein SH2B1 Controls the Ability of SH2B1α To Enhance Nerve Growth Factor Function. *Molecular and Cellular Biology*, 38(6).

Kapust, R. B. and Waugh, D. S. (1999). *Escherichia Coli* Maltose-Binding Protein Is Uncommonly Effective at Promoting the Solubility of Polypeptides to Which It Is Fused . *Protein Science*, 8(8), 1668–1674.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. (2016). The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nature Protocols*, 10(6), 845–858.

Klemm, J. D., Schreiber, S. L. and Crabtree, G. R. (1998). Dimerization as a Regulatory Mechanism in Signal Transduction. *Annual Review of Immunology*, 16, 569–592.

Koshiba, S., Kigawa, T., Kim, J. H., Shirouzu, M., Bowtell, D. and Yokoyama, S. (1997). The Solution Structure of the Pleckstrin Homology Domain of Mouse Son-of-Sevenless 1 (MSos1). *Journal of Molecular Biology*, 269(4), 579–591.

Kozak, S., Lercher, L., Karanth, M. N., Meijers, R., Carlomagno, T. and Boivin, S. (2016). Optimization of Protein Samples for NMR Using Thermal Shift Assays. *Journal of Biomolecular NMR*, 64(4), 281–289.

Latreille, M., Laberge, M. K., Bourret, G., Yamani, L. and Larose, L. (2011). Deletion of Nck1 Attenuates Hepatic ER Stress Signaling and Improves Glucose Tolerance and Insulin Signaling in Liver of Obese Mice. *American Journal of Physiology - Endocrinology and Metabolism*, 300(3), 423–434.

Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C.

J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E. and Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13), 6589–6631.

Lemmon, M. A. and Schlessinger, J. (2010). Cell Signaling by Receptor-Tyrosine Kinases. *Cell*, 141(7), 1117–1134.

Lemmon, M. A. (2007). Pleckstrin Homology (PH) Domains and Phosphoinositides. *Cell Biology of Inositol Lipids and Phosphates*, 93(74), 81–93.

Lenoir, M., Kufareva, I., Abagyan, R. and Overduin, M. (2015). Membrane and Protein Interactions of the Pleckstrin Homology Domain Superfamily. *Membranes*, 5(4), 646–663.

Li, M., Li, Z., Morris, D. L. and Rui, L. (2007). Identification of SH2B2β as an Inhibitor for SH2B1- and SH2B2α- Promoted Janus Kinase-2 Activation and Insulin Signaling. *Endocrinology*, 28(10), 1615-1621.

Li, Z., Zhou, Y., Carter-Su, C., Myers, M. G. and Rui, L. (2007). SH2B1 Enhances Leptin Signaling by Both Janus Kinase 2 Tyr^{813} Phosphorylation-Dependent and -Independent Mechanisms. *Molecular Endocrinology*, 21(9), 2270–2281.

Lim, W. A. (2002). The Modular Logic of Signaling Proteins: Building Allosteric Switches from Simple Binding Domains. *Current Opinion in Structural Biology*, 12(1), 61–68.

Linge, J. P. and Nilges, M. (1999). Influence of Non-Bonded Parameters on the Quality of NMR Structures: A New Force Field for NMR Structure Calculation. *Journal of Biomolecular NMR*, 13(1), 51–59.

Linge, J. P., Williams, M. A., Spronk, C. A. E. M., Bonvin, A. M. J. J. and Nilges, M. (2003). Refinement of Protein Structures in Explicit Solvent. *Proteins: Structure, Function and Genetics*, 50(3), 496–506.

Liu, B. A., Jablonowski, K., Raina, M., Arcé, M., Pawson, T. and Nash, P. D. (2006). The Human and Mouse Complement of SH2 Domain Proteins-Establishing the Boundaries of Phosphotyrosine Signaling. *Molecular Cell*, 22(6), 851–868.

López-Méndez, B. and Güntert, P. (2006). Automated Protein Structure Determination from NMR Spectra. *Journal of the American Chemical Society*, (2), 13112–13122.

Mardilovich, K., Pankratz, S. L. and Shaw, L. M. (2009). Expression and Function of the Insulin Receptor Substrate Proteins in Cancer. *Cell Communication and Signaling*, 7, 1–15.

Marengere, L. E. and Pawson, T. (1994). Structure and Function of SH2 Domains. *Journal of*

Cell Science, 18, 97–104.

Maures, T. J., Chen, L. and Carter-Su, C. (2009). Nucleocytoplasmic Shuttling of the Adapter Protein SH2B1β (SH2-Bβ) Is Required for Nerve Growth Factor (NGF)-Dependent Neurite Outgrowth and Enhancement of Expression of a Subset of NGF-Responsive Genes. *Molecular Endocrinology*, 23(7), 1077–1091.

Maures, T. J., Kurzer, J. H. and Carter-Su, C. (2007). SH2B1 (SH2-B) and JAK2: A Multifunctional Adaptor Protein and Kinase Made for Each Other. *Trends in Endocrinology and Metabolism*, 18(1), 38–45.

McKercher, M. A., Guan, X., Tan, Z. and Wuttke, D. S. (2018). Diversity in Peptide Recognition by the SH2 Domain of SH2B1. *Proteins: Structure, Function and Bioinformatics*, 86(2), 164–176.

Mittag, T., Kay, L. E. and Forman-Kay, J. D. (2010). Protein Dynamics and Conformational Disorder in Molecular Recognition. *Journal of Molecular Recognition*, 23(2), 105–116.

Moodie, S. A., Alleman-Sposeto, J. and Gustafson, T. A. (1999). Identification of the APS Protein as a Novel Insulin Receptor Substrate. *Journal of Biological Chemistry*, 274(16), 11186–11193.

Morris, D., Cho, K. W., Zhou, Y. and Rui, L. (2008). SH2B1 Directly Enhances Insulin Action by Both Stimulating the Insulin Receptor and Inhibiting Tyrosine Dephosphorylation of IRS Proteins. *Diabetes*, 57(9), A383–A383.

Morris, D. L., Cho, K. W., Zhou, Y. and Rui, L. (2009). SH2B1 Enhances Insulin Sensitivity by Both Stimulating the Insulin Receptor and Inhibiting Tyrosine Dephosphorylation of Insulin Receptor Substrate Proteins. *Diabetes*, 58(9), 2039–2047.

Morris, D. and Rui, L. (2009). Recent Advances in Understanding Leptin Signaling and Leptin Resistance. *American Journal of Physiology - Endocrinology and Metabolism*, 297(28), 1247–1259.

Nabuurs, S. B., Spronk, C. A. E. M., Vriend, G. and Vuister, G. W. (2004). Concepts and Tools for NMR Restraint Analysis and Validation. *Concepts in Magnetic Resonance Part A: Bridging Education and Research*, 22(2), 90–105.

Nelms, K., O'Neill, T. J., Li, S., Hubbard, S. R., Gustafson, T. A. and Paul, W. E. (1999). Alternative Splicing, Gene Localization, and Binding of SH2-B to the Insulin Receptor Kinase Domain. *Mammalian Genome*, 10(12), 1160–1167.

Nishi, M., Werner, E. D., Oh, B.-C., Frantz, J. D., Dhe-Paganon, S., Hansen, L., Lee, J. and

Shoelson, S. E. (2005). Kinase Activation through Dimerization by Human SH2-B. *Molecular and Cellular Biology*, 25(7), 2607–2621.

O'Brien, K. B., Argetsinger, L. S., Diakonova, M. and Carter-Su, C. (2003). YXXL Motifs in SH2-Bβ Are Phosphorylated by JAK2, JAK1, and Platelet-Derived Growth Factor Receptor and Are Required for Membrane Ruffling. *Journal of Biological Chemistry*, 278(14), 11970–11978.

Van Obberghen, E., Baron, V., Delahaye, L., Emanuelli, B., Filippa, N., Giorgetti-Peraldi, S., Lebrun, P., Mothe-Satney, I., Peraldi, P., Rocchi, S., Sawka-Verhelle, D., Tartare-Deckert, S. and Giudicelli, J. (2001). Surfing the Insulin Signaling Web. *European Journal of Clinical Investigation*, 31(11), 966–977.

Pascal, S. M., Singer, A. U., Gish, G., Yamazaki, T., Shoelson, S. E., Pawson, T., Kay, L. E. and Forman-Kay, J. D. (1994). Nuclear Magnetic Resonance Structure of an SH2 Domain of Phospholipase C-γ1 Complexed with a High Affinity Binding Peptide. *Cell*, 77(3), 461–472.

Pawson, T. and Scott, J. D. (1997). Signaling through Scaffold, Anchoring, and Adaptor Proteins. *Science*, 278(5346), 2075–2080.

Pearce, L. R., Joe, R., Doche, M. E., Su, H. W., Keogh, J. M., Henning, E., Argetsinger, L. S., Bochukova, E. G., Cline, J. M., Garg, S., Saeed, S., Shoelson, S., O'Rahilly, S., Barroso, I., Rui, L., Farooqi, I. S. and Carter-Su, C. (2014). Functional Characterization of Obesity-Associated Variants Involving the α and β Isoforms of Human SH2B1. *Endocrinology*, 155(9), 3219–3226.

Qian, X., Riccio, A., Zhang, Y. and Ginty, D. D. (1998). Identification and Characterization of Novel Substrates of Trk Receptors in Developing Neurons. *Neuron*, 21(5), 1017–1029.

Reed, M. A. C., Hounslow, A. M., Sze, K. H., Barsukov, I. G., Hosszu, L. L. P., Clarke, A. R., Craven, C. J. and Waltho, J. P. (2003). Effects of Domain Dissection on the Folding and Stability of the 43 KDa Protein PGK Probed by NMR. *Journal of Molecular Biology*, 330(5), 1189–1201.

Ren, D., Li, M., Duan, C. and Rui, L. (2005). Identification of SH2-B as a Key Regulator of Leptin Sensitivity, Energy Balance, and Body Weight in Mice. *Cell Metabolism*, 2(2), 95–104.

Richardson, J. and Richardson, D. (1988). Amino Acid Preferences for Specific Locations at the Ends of α Helices. *Science*, 242(4886), 1624.

Riedel, H., Wang, J., Hansen, H. and Yousaf, N. (1997). PSM, an Insulin-Dependent, Pro-Rich, PH, SH2 Domain Containing Partner of the Insulin Receptor. *Journal of Biochemistry*, 122(6), 1105–1113.

Rosato, A., Tejero, R. and Montelione, G. T. (2013). Quality Assessment of Protein NMR Structures. *Current Opinion in Structural Biology*, 23(5), 715–724.

Rui, L. (2014). SH2B1 Regulation of Energy Balance, Body Weight, and Glucose Metabolism. *World Journal of Diabetes*, 5(4), 511–526.

Rui, L. and Carter-Su, C. (1999). Identification of SH2-Bβ as a Potent Cytoplasmic Activator of the Tyrosine Kinase Janus Kinase 2. *Proceedings of the National Academy of Sciences of the United States of America*, 96(13), 7172–7177.

Rui, L., Günter, D. R., Herrington, J. and Carter-Su, C. (2000). Differential Binding to and Regulation of JAK2 by the SH2 Domain and N-Terminal Region of SH2-Bbeta. *Molecular and Cellular Biology.*, 20(9), 3168–3177.

Rui, L., Herrington, J. and Carter-Su, C. (1999). SH2-B, a Membrane-Associated Adapter, Is Phosphorylated on Multiple Serines/Threonines in Response to Nerve Growth Factor by Kinases within the MEK/ERK Cascade. *Journal of Biological Chemistry*, 274(37), 26485–26492.

Rui, L., Mathews, L. S., Hotta, K., Gustafson, T. A. and Carter-Su, C. (1997). Identification of SH2-Bbeta as a Substrate of the Tyrosine Kinase JAK2 Involved in Growth Hormone Signaling. *Molecular and Cellular Biology*, 17(11), 6633–6644.

Schindler, C., Levy, D. E. and Decker, T. (2007). JAK-STAT Signaling: From Interferons to Cytokines. *Journal of Biological Chemistry*, 282(28), 20059–20063.

Schlessinger, J. and Ullrich, A. (1990). Signal Transduction by Receptors with Tyrosine Kinase Activity. *Cell*, 61, 203–212.

Schmidt, E. and Güntert, P. (2013). Reliability of Exclusively NOESY-Based Automated Resonance Assignment and Structure Determination of Proteins. *Journal of Biomolecular NMR*, 57(2), 193–204.

Schmitt, J., Hess, H. and Stunnenberg, H. G. (1993). Affinity Purification of Histidine-Tagged Proteins. *Molecular Biology Reports*, 18(3), 223–230.

Shaw, A. S. and Filbert, E. L. (2009). Scaffold Proteins and Immune-Cell Signalling. *Nature Reviews Immunology*, 9(1), 47–56.

Shen, Y. and Bax, A. (2013). Protein Backbone and Sidechain Torsion Angles Predicted from NMR Chemical Shifts Using Artificial Neural Networks. *Journal of Biomolecular NMR*, 56(3), 227–241.

Skolnik, E. Y., Batzer, A., Li, N., Lee, C. H., Lowenstein, E., Mohammadi, M., Margolis, B. and Schlessinger, J. (1993). The Function of GRB2 in Linking the Insulin Receptor to Ras Signaling Pathways. *Science*, 260(5116), 1953–1955.

Song, W., Ren, D., Li, W., Jiang, L., Cho, K. W., Huang, P., Song, Y., Liu, Y. and Rui, L. (2010). SH2B Regulation of Growth, Metabolism and Longevity in Both Insects and Mammals. *Cell Metabolism*, 11(5), 427–437.

Sorokin, A., Reed, E., Nnkemere, N., Dulin, N. O. and Schlessinger, J. (1998). Crk Protein Binds to PDGF Receptor and Insulin Receptor Substrate-1 with Different Modulating Effects on PDGF- and Insulin-Dependent Signaling Pathways. *Oncogene*, 16(19), 2425–2434.

Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G. and Vuister, G. W. (2004). Validation of Protein Structures Derived by NMR Spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 45(3–4), 315–337.

Sun, X. J., Pons, S., Asano, T., Myers, M. G., Glasheen, E. and White, M. F. (1996). The Fyn Tyrosine Kinase Binds IRS-1 and Forms a Distinct Signaling Complex during Insulin Stimulation. *Journal of Biological Chemistry*, 271(18), 10583–10587.

Theillet, F. X., Smet-Nocca, C., Liokatis, S., Thongwichian, R., Kosten, J., Yoon, M. K., Kriwacki, R. W., Landrieu, I., Lippens, G. and Selenko, P. (2012). Cell Signaling, Post-Translational Protein Modifications and NMR Spectroscopy. *Journal of Biomolecular NMR*, 54(3), 217–236.

Tomlinson, J. H. and Williamson, M. P. (2012). Amide Temperature Coefficients in the Protein G B1 Domain. *Journal of Biomolecular NMR*, 52(1), 57–64.

Trudeau, T., Nassar, R., Cumberworth, A., Wong, E. T. C., Woollard, G. and Gsponer, J. (2013). Structure and Intrinsic Disorder in Protein Autoinhibition. *Structure*, 21(3), 332–341.

Uversky, V. N., Oldfield, C. J. and Dunker, A. K. (2008). Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. *Annual Review of Biophysics*, 37, 215–246.

Vögeli, B., Olsson, S., Güntert, P. and Riek, R. (2016). The Exact NOE as an Alternative in Ensemble Structure Determination. *Biophysical Journal*, 110(1), 113–126.

Waksman, G. (1992). Crystal Structure of the Phosphotyrosine Recognition Domain SH2 of the Src Oncogene Product Complexed with Tyrosine-Phosphorylated Peptides. *Cellular and molecular biology*, 358(5), 646–652.

Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. and Kuriyan, J. (1993). Binding of a High

Affinity Phosphotyrosyl Peptide to the Src SH2 Domain: Crystal Structures of the Complexed and Peptide-Free Forms. *Cell*, 72(5), 779–790.

Waltho, J. P. and Cavanagh, J. (1993). Practical Aspects of Recording Multidimensional NMR-Spectra in Water with Flat Base-lines. *Journal of Magnetic Resonance Series A*, 103(3), 338–348.

Williamson, M. P. (2012). How Proteins Work. *New York and London: Garland Science, Taylor and Francis Group.*

Williamson, M. P. (2013). Using Chemical Shift Perturbation to Characterise Ligand Binding. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 73, 1–16.

Wüthrich, K. and Wagner, G. (1975). NMR Investigations of the Dynamics of the Aromatic Amino Acid Residues in the Basic Pancreatic Trypsin Inhibitor. *FEBS Letters*, 50(2), 265–268.

Xu, H., Yang, D., Jiang, D. and Chen, H. Y. (2019). Phosphate Assay Kit in One Cell for Electrochemical Detection of Intracellular Phosphate Ions at Single Cells. *Frontiers in Chemistry*, 7(5), 1–6.

Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L. and Yao, X. (2008). GPS 2.0, a Tool to Predict Kinase-Specific Phosphorylation Sites in Hierarchy. *Molecular and Cellular Proteomics*, 7(9), 1598–1606.

Yamazaki, T., Forman-Kay, J. D. and Kay, L. E. (1993). Two-Dimensional NMR Experiments for Correlating ^{13}Cβ and ^{1}Hδ/ϵ Chemical Shifts of Aromatic Residues in ^{13}C-Labeled Proteins via Scalar Couplings. *Journal of the American Chemical Society*, 115(23), 11054–11055.

Yousaf, N., Deng, Y., Kang, Y. and Riedel, H. (2001). Four PSM/SH2-B Alternative Splice Variants and Their Differential Roles in Mitogenesis. *Journal of Biological Chemistry*, 276(44), 40940–40948.

Zhang, M., Deng, Y. and Riedel, H. (2008). PSM/SH2B1 Splice Variants: Critical Role in Src Catalytic Activation and the Resulting STAT3-Mediated Mitogenic Response. *Journal of Cellular Biochemistry*, 104(1), 105–118.

Zhao, D. and Jardetzky, O. (1994). An Assessment of the Precision and Accuracy of Protein Structures Determined by NMR: Dependence on Distance Errors. *Journal of Molecular Biology*, 239(5), 601–607.

Zuiderweg, E. R. P. (2002). Mapping Protein-Protein Interactions in Solution by NMR Spectroscopy. *Biochemistry*, 41(1), 1–7.

Appendix

Appendix A

Comparison table between the manual chemical shifts (Ref) and the automated CYANA chemical shifts values (Shift) of SH2 protein, the differences in chemical shift > 0.3 ppm are highlighted. Green highlighted shifts are the un-matched chemical shifts of amide sidechain of Arg, Asn, Gln. Grey highlighted shifts are the un-matched chemical shifts of aromatic sidechain of Tyr, Phe, Trp, His. Blue highlighted shifts are the amide chemical shifts. Purple highlighted shifts are un-matched chemical shifts of aliphatic sidechain of Leu. Dev is Ref-Shift, Extent is number of runs were completed to assign nuclei in FLYA, inside is the percentage of chemical shift value from independent runs which agree with the consensus shift value ,within the defined tolerance values. Inref is the percentage of shift value from 10 independent runs of FLYA which agree with the reference chemical shift value within the defined tolerances values. Total number of automated assigned shifts: 1441, out of them 1088 are strong.

Atom	Residue		Ref	Shift	Dev	Extent	inside	inref	
N	LYS	1		122.719		20	85	0	strong
H	LYS	1		7.63		20	89.6	0	strong
CA	LYS	1		56.315		20	100	0	strong
HA	LYS	1		4.393		20	99.9	0	strong
CB	LYS	1		33.17		20	99.9	0	strong
HB2	LYS	1		1.721		20	99.6	0	strong
HB3	LYS	1		1.725		20	99.6	0	strong
CG	LYS	1		24.745		20	100	0	strong
HG2	LYS	1		1.351		20	100	0	strong
HG3	LYS	1		1.352		20	99.9	0	strong
CD	LYS	1		29.069		20	100	0	strong
HD2	LYS	1		1.687		20	99.9	0	strong
HD3	LYS	1		1.689		20	99.7	0	strong
CE	LYS	1		42.174		20	99.8	0	strong
HE2	LYS	1		2.983		20	99.2	0	strong
HE3	LYS	1		2.985		20	95.2	0	strong
QZ	LYS	1		6.958		20	45	0	
C	LYS	1		175.678		20	84.8	0	strong

N	ILE	2		123.347		20	96.7	0	strong
H	ILE	2		8.318		20	99.6	0	strong
CA	ILE	2		60.727		20	100	0	strong
HA	ILE	2		4.09		20	100	0	strong
CB	ILE	2		38.794		20	100	0	strong
HB	ILE	2		1.732		20	99.9	0	strong
QG2	ILE	2		0.718		20	99.4	0	strong
CG2	ILE	2		17.321		20	99.9	0	strong
CG1	ILE	2		27.211		20	100	0	strong
HG12	ILE	2		1.123		20	99.9	0	strong
HG13	ILE	2		1.388		20	99.9	0	strong
QD1	ILE	2		0.809		20	100	0	strong
CD1	ILE	2		12.652		20	100	0	strong
C	ILE	2		175.726		20	100	0	strong
N	HIS	3		124.018		20	100	0	strong
H	HIS	3		8.559		20	100	0	strong
CA	HIS	3		55.867		20	74.5	0	
HA	HIS	3		4.644		20	75.8	0	
CB	HIS	3		29.912		20	91.4	0	strong
HB2	HIS	3		3.038		20	79.4	0	
HB3	HIS	3		3.12		20	70.7	0	
ND1	HIS	3		122.627		15	53.3	0	
CD2	HIS	3		120.196		20	35	0	
HD1	HIS	3		7.14		20	42.5	0	
CE1	HIS	3		137.719		20	60.2	0	
HD2	HIS	3		7.124		20	50.9	0	
HE1	HIS	3		8.22		20	35.3	0	
C	HIS	3		174.41		20	84.5	0	strong
N	HIS	4		120.945		20	26.7	0	
H	HIS	4		8.306		20	23.9	0	
CA	HIS	4		56.225		20	41	0	
HA	HIS	4		4.247		20	19.8	0	
CB	HIS	4		29.939		20	32.7	0	
HB2	HIS	4		3.13		20	24.9	0	
HB3	HIS	4		3.133		20	24.6	0	
ND1	HIS	4		119.804		7	28.5	0	
CD2	HIS	4		120.164		20	36.1	0	
HD1	HIS	4		1.091		20	19.1	0	
CE1	HIS	4		137.701		20	34	0	
HD2	HIS	4		6.834		20	35	0	
HE1	HIS	4		7.461		20	30.1	0	
C	HIS	4		174.987		20	53.7	0	

N	HIS	5		120.029		20	25.3	0	
H	HIS	5		7.697		20	20.8	0	
CA	HIS	5		56.444		20	59.7	0	
HA	HIS	5		5.369		20	30	0	
CB	HIS	5		30.021		20	24.9	0	
HB2	HIS	5		2.272		20	25.2	0	
HB3	HIS	5		2.416		20	24.5	0	
ND1	HIS	5		117.349		6	66.7	0	
CD2	HIS	5		120.432		20	31.2	0	
HD1	HIS	5		0.701		20	23.2	0	
CE1	HIS	5		137.165		19	48	0	
HD2	HIS	5		7.09		20	25.6	0	
HE1	HIS	5		7.391		20	31.5	0	
C	HIS	5		174.96		19	53.5	0	
N	HIS	6		120.087		20	33.3	0	
H	HIS	6		8.98		20	37.2	0	
CA	HIS	6		56.378		20	71.5	0	
HA	HIS	6		5.041		20	40.1	0	
CB	HIS	6		32.056		20	35	0	
HB2	HIS	6		2.277		20	26	0	
HB3	HIS	6		2.887		20	29.9	0	
ND1	HIS	6		110.054		15	42.1	0	
CD2	HIS	6		120.404		20	27.8	0	
HD1	HIS	6		6.923		20	34	0	
CE1	HIS	6		137.287		20	43.9	0	
HD2	HIS	6		7.058		20	22.2	0	
HE1	HIS	6		8.296		20	28.3	0	
C	HIS	6		175.664		19	48	0	
N	HIS	7		120.041		20	39.6	0	
H	HIS	7		8.977		20	56.6	0	
CA	HIS	7		56.386		20	71.1	0	
HA	HIS	7		4.394		20	30	0	
CB	HIS	7		29.616		20	35	0	
HB2	HIS	7		2.184		20	23.1	0	
HB3	HIS	7		3.125		20	23.9	0	
ND1	HIS	7		110.047		8	78.9	0	
CD2	HIS	7		120.215		19	31.6	0	
HD1	HIS	7		6.92		20	34.7	0	
CE1	HIS	7		137.267		20	47.7	0	
HD2	HIS	7		7.13		20	51.6	0	
HE1	HIS	7		8.298		20	28.5	0	
C	HIS	7		175.615		20	80	0	strong

N	HIS	8		123.421		20	56.3	0	
H	HIS	8		8.317		20	60	0	
CA	HIS	8		55.941		20	99.9	0	strong
HA	HIS	8		4.642		20	100	0	strong
CB	HIS	8		29.815		20	99.6	0	strong
HB2	HIS	8		3.135		20	99	0	strong
HB3	HIS	8		3.135		20	87.6	0	strong
ND1	HIS	8		119.881		7	80.1	0	
CD2	HIS	8		124.503		20	65	0	
HD1	HIS	8		3.644		20	64.9	0	
CE1	HIS	8		138.539		20	65	0	
HD2	HIS	8		7.14		20	96.6	0	strong
HE1	HIS	8		7.493		20	64.3	0	
C	HIS	8		174.411		20	79.9	0	
N	ASP	9	121.049	121.05	-0.001	20	100	100	strong=
H	ASP	9	8.549	8.547	0.002	20	99.3	100	strong=
CA	ASP	9	54.548	54.551	-0.003	20	100	100	strong=
HA	ASP	9	4.622	4.622	0	20	99.7	100	strong=
CB	ASP	9	41.226	41.245	-0.019	20	99.1	100	strong=
HB2	ASP	9	2.701	2.694	0.007	20	96.1	90	strong=
HB3	ASP	9	2.698	2.698	0	20	99.8	100	strong=
C	ASP	9	175.773	175.774	-0.001	20	100	100	strong=
N	GLN	10	120.213	120.212	0.001	20	99.9	100	strong=
H	GLN	10	8.289	8.289	0	20	99.9	100	strong=
CA	GLN	10	53.408	53.259	0.149	20	99.6	0	strong!
HA	GLN	10		4.743		20	99.7	0	strong
CB	GLN	10	29.133	29.133	0	20	69.9	0	!
HB2	GLN	10		2.143		20	98.4	0	strong
HB3	GLN	10		2.139		20	60	0	
CG	GLN	10		33.78		20	65	0	
HG2	GLN	10		2.627		20	49.8	0	
HG3	GLN	10		2.627		20	65	0	
NE2	GLN	10		112.235		20	94.6	0	strong
HE21	GLN	10		6.949		20	93.8	0	strong
HE22	GLN	10		7.539		20	89.6	0	strong
C	GLN	10	174.2	174.244	-0.044	20	100	100	strong=
CA	PRO	11	64.113	64.118	-0.005	20	99.4	100	strong=
HA	PRO	11	4.423	4.418	0.005	20	96.6	95	strong=
CB	PRO	11	31.933	31.902	0.031	20	94.6	95	strong=
HB2	PRO	11	2.338	2.338	0	20	76.3	75	=
HB3	PRO	11	2.34	2.34	0	20	96	0	strong!
CG	PRO	11	27.55	27.513	0.037	20	99.1	100	strong=

HG2	PRO	11	1.946	1.939	0.007	20	91.4	95	strong=
HG3	PRO	11	1.956	1.968	-0.012	20	87.2	85	strong=
CD	PRO	11	50.657	50.686	-0.029	20	99.9	100	strong=
HD2	PRO	11	3.735	3.728	0.007	20	94.9	95	strong=
HD3	PRO	11	3.732	3.73	0.002	20	95	95	strong=
C	PRO	11	177.721	177.72	0.001	20	99.9	100	strong=
N	LEU	12	118.565	118.539	0.026	20	99.9	100	strong=
H	LEU	12	8.227	8.222	0.005	20	99.6	100	strong=
CA	LEU	12	55.179	55.16	0.019	20	100	100	strong=
HA	LEU	12	4.25	4.245	0.005	20	99.9	100	strong=
CB	LEU	12	39.615	39.61	0.005	20	99.6	100	strong=
HB2	LEU	12	0.359	0.36	-0.001	20	65	35	! (HB3)
HB3	LEU	12	0.561	0.556	0.005	20	89.4	90	strong=
CG	LEU	12	26.927	26.935	-0.008	20	89.6	90	strong=
HG	LEU	12	1.139	1.138	0.001	20	89.9	90	strong=
QD1	LEU	12	0.575	0.571	0.004	20	89.7	90	strong=
QD2	LEU	12		0.25		20	90	0	strong
CD1	LEU	12	23.114	23.094	0.02	20	100	100	strong=
CD2	LEU	12		25.711		20	89.9	0	strong
C	LEU	12	176.13	176.13	0	20	100	100	strong=
N	SER	13	110.914	110.913	0.001	20	100	100	strong=
H	SER	13	7.439	7.437	0.002	20	100	100	strong=
CA	SER	13	60.378	60.389	-0.011	20	100	100	strong=
HA	SER	13	4.199	4.195	0.004	20	99.9	100	strong=
CB	SER	13	63.297	63.272	0.025	20	98.9	100	strong=
HB2	SER	13	3.951	3.951	0	20	99.9	100	strong=
HB3	SER	13	3.953	3.952	0.001	20	55	55	=
C	SER	13	175.382	175.382	0	20	100	100	strong=
N	GLY	14	110.819	110.816	0.003	20	100	100	strong=
H	GLY	14	8.496	8.491	0.005	20	100	100	strong=
CA	GLY	14	44.968	44.96	0.008	20	100	100	strong=
HA2	GLY	14	3.736	3.735	0.001	20	99.7	100	strong=
HA3	GLY	14	3.74	3.739	0.001	20	99.8	100	strong=
C	GLY	14	174.309	174.309	0	20	100	100	strong=
N	TYR	15	120.017	120.02	-0.003	20	100	100	strong=
H	TYR	15	7.428	7.429	-0.001	20	99.9	100	strong=
CA	TYR	15	56.293	56.318	-0.025	19	63.1	63.2	=
HA	TYR	15		4.392		20	49.6	0	
CB	TYR	15	36.193	36.153	0.04	20	49.7	50	=
HB2	TYR	15		1.644		20	32.6	0	
HB3	TYR	15		2.273		20	34.4	0	
CD1	TYR	15	133.153	133.209	-0.056	20	95.8	95	strong=

HD1	TYR	15	7.009	7.008	0.001	20	96	95	strong=
CE1	TYR	15	118.211	121.361	-3.15	17	56.7	0	!
HE1	TYR	15	6.838	6.81	0.028	20	50.1	15	=
CE2	TYR	15	118.212	118.848	-0.636	20	73.2	15	!
HE2	TYR	15	6.841	7.184	-0.343	19	42.8	10.5	!
CD2	TYR	15	132.963	133.064	-0.101	20	73.8	75	=
HD2	TYR	15	7.012	7.011	0.001	20	76.8	75	=
HH	TYR	15		10.698		20	39.9	0	
C	TYR	15	176.553	176.553	0	20	100	100	strong=
CA	PRO	16		64.438		20	88.2	0	strong
HA	PRO	16		4.381		20	86.4	0	strong
CB	PRO	16		33.947		20	49.9	0	
HB2	PRO	16		2.067		20	62.9	0	
HB3	PRO	16		2.298		20	54.9	0	
CG	PRO	16		27.361		20	82.6	0	strong
HG2	PRO	16		1.763		20	85.1	0	strong
HG3	PRO	16		2.279		20	68.2	0	
CD	PRO	16		50.711		20	99.3	0	strong
HD2	PRO	16		3.525		20	94.5	0	strong
HD3	PRO	16		3.796		20	76.5	0	
C	PRO	16		172.433		14	77.1	0	
N	TRP	17	109.851	118.81	-8.959	20	64.2	20	! (CZ3)
H	TRP	17	5.571	7.359	-1.788	20	59.9	20	!
CA	TRP	17	52.93	52.909	0.021	20	79.9	80	=
HA	TRP	17	4.527	4.523	0.004	20	73.5	75	=
CB	TRP	17	28.893	28.747	0.146	20	80.1	80	strong=
HB2	TRP	17	2.281	2.278	0.003	20	56.7	60	=
HB3	TRP	17	3.536	3.53	0.006	20	83.9	85	strong=
CD1	TRP	17	128.753	128.84	-0.087	20	80	80	=
CE3	TRP	17	120.389	121.453	-1.064	20	44.2	10	!
NE1	TRP	17	131.359	131.356	0.003	20	90	90	strong=
HD1	TRP	17	6.817	6.678	0.139	20	80	15	!
HE3	TRP TRP	17	6.812	6.692	0.12	20	29.5	20	!
CZ3	TRP	17	118.571	120.292	-1.721	20	32.5	25	! (CE3)
CZ2	TRP	17	115.584	115.596	-0.012	19	42.1	42.1	=
HE1	TRP	17	10.678	10.677	0.001	20	90	90	strong=
HZ3	TRP	17	6.451	7.196	-0.745	20	29.7	0	! (HE1 18)
CH2	TRP	17	124.171	120.049	4.122	20	53.1	0	! (CE3)
HZ2	TRP	17	6.73	6.74	-0.01	20	43.2	40	=
HH2	TRP	17	6.454	6.691	-0.237	20	44.2	0	!

C	TRP	17	175.567	175.691	-0.124	20	80.5	80	strong=
N	PHE	18	122.753	122.751	0.002	20	80	80	strong=
H	PHE	18	7.642	7.637	0.005	20	80.2	80	strong=
CA	PHE	18	57.186	57.169	0.017	20	89.9	90	strong=
HA	PHE	18	5.641	5.644	-0.003	20	81.9	80	strong=
CB	PHE	18	39.987	39.938	0.049	20	79.8	80	=
HB2	PHE	18	2.659	2.643	0.016	20	81.5	80	strong=
HB3	PHE	18	3.054	3.052	0.002	20	79.6	80	=
CD1	PHE	18	133.163	133.168	-0.005	20	79.1	80	=
HD1	PHE	18	7.229	7.227	0.002	20	88.1	90	strong=
CE1	PHE	18	133.226	133.163	0.063	19	82.4	84.2	strong=
HE1	PHE	18	7.225	7.229	-0.004	20	44.7	45	=
CZ	PHE	18	130.937	125.612	5.325	20	55	0	!
HZ	PHE	18	7.419	5.648	1.771	20	63.4	0	! (HA)
CE2	PHE	18	131.903	133.108	-1.205	20	34.2	0	! (CD2)
HE2	PHE	18	7.413	7.231	0.182	20	59.5	5	! (HD2)
CD2	PHE	18	133.09	133.264	-0.174	20	80	80	=
HD2	PHE	18	7.23	7.243	-0.013	20	72.6	65	=
C	PHE	18	175.817	175.813	0.004	20	89.7	90	strong=
N	HIS	19	126.039	125.696	0.343	20	80	80	=
H	HIS	19	8.982	8.918	0.064	20	79.7	0	!
CA	HIS	19	56.249	56.301	-0.052	20	79.8	80	=
HA	HIS	19		4.395		20	79.9	0	
CB	HIS	19	33.909	33.91	-0.001	20	79.9	80	=
HB2	HIS	19		2.563		20	57.8	0	
HB3	HIS	19		3.481		20	74.8	0	
ND1	HIS	19		123.409		1	100	0	
CD2	HIS	19	116.902	116.888	0.014	20	79.7	80	=
HD1	HIS	19		0.749		20	78.2	0	
CE1	HIS	19	136.784	134.47	2.314	20	39.7	0	!
HD2	HIS	19	7.303	7.304	-0.001	20	79.3	80	=
HE1	HIS	19	7.383	6.844	0.539	20	44.9	0	!
C	HIS	19	174.792	174.989	-0.197	20	57.6	55	=
N	GLY	20		112.036		20	40.8	0	
H	GLY	20		6.943		20	40.9	0	
CA	GLY	20	47.186	47.181	0.005	20	100	100	strong=
HA2	GLY	20	3.5	3.5	0	20	99.7	100	strong=
HA3	GLY	20	3.768	3.764	0.004	20	98.9	100	strong=
C	GLY	20	174.306	174.306	0	20	99	100	strong=
N	MET	21	127.058	127.056	0.002	20	100	100	strong=
H	MET	21	8.788	8.787	0.001	20	100	100	strong=
CA	MET	21	55.63	55.63	0	20	99.9	100	strong=

HA	MET	21	4.561	4.559	0.002	20	99.9	100	strong=
CB	MET	21	30.81	30.876	-0.066	20	99.7	100	strong=
HB2	MET	21	1.909	1.897	0.012	20	96.7	100	strong=
HB3	MET	21	2.183	2.183	0	20	79.4	80	=
CG	MET	21	32.094	32.125	-0.031	20	99.5	100	strong=
HG2	MET	21	2.524	2.472	0.052	20	99.2	0	strong!
HG3	MET	21	2.475	2.475	0	20	94.4	95	strong=
QE	MET	21		2.082		20	95.4	0	strong
CE	MET	21		16.907		20	99.9	0	strong
C	MET	21	174.207	174.207	0	20	99.9	100	strong=
N	LEU	22	128.282	128.279	0.003	20	100	100	strong=
H	LEU	22	7.496	7.493	0.003	20	100	100	strong=
CA	LEU	22	54.091	54.127	-0.036	20	99.9	100	strong=
HA	LEU	22	4.528	4.527	0.001	20	99.3	100	strong=
CB	LEU	22	45.361	45.371	-0.01	20	99.9	100	strong=
HB2	LEU	22	1.473	1.48	-0.007	20	89.3	90	strong=
HB3	LEU	22	1.489	1.487	0.002	20	99.2	100	strong=
CG	LEU	22	26.768	26.955	-0.187	20	91.2	90	strong=
HG	LEU	22	1.791	1.778	0.013	20	45.6	35	=
QD1	LEU	22	1.05	1.046	0.004	20	79.9	80	=
QD2	LEU	22	0.801	0.803	-0.002	20	59.8	60	=
CD1	LEU	22	24.496	24.504	-0.008	20	90	90	strong=
CD2	LEU	22	24.742	24.742	0	20	71.8	25	! (CG)
C	LEU	22	175.385	175.385	0	20	100	100	strong=
N	SER	23	121.944	121.94	0.004	20	100	100	strong=
H	SER	23	8.406	8.404	0.002	20	100	100	strong=
CA	SER	23	57.764	57.755	0.009	20	100	100	strong=
HA	SER	23	4.302	4.299	0.003	20	99.9	100	strong=
CB	SER	23	64.819	64.802	0.017	20	94.9	95	strong=
HB2	SER	23	4.027	4.027	0	20	94.9	95	strong=
HB3	SER	23	4.346	4.336	0.01	20	85.8	80	strong=
C	SER	23	173.828	173.827	0.001	20	100	100	strong=
N	ARG	24	124.043	124.043	0	20	100	100	strong=
H	ARG	24	9.052	9.051	0.001	20	99.7	100	strong=
CA	ARG	24	60.095	60.125	-0.03	20	100	100	strong=
HA	ARG	24	3.565	3.56	0.005	20	100	100	strong=
CB	ARG	24	30.87	30.819	0.051	20	99.9	100	strong=
HB2	ARG	24	1.644	1.651	-0.007	20	89.9	90	strong=
HB3	ARG	24	2.014	2.012	0.002	20	98.1	100	strong=
CG	ARG	24	27.173	27.174	-0.001	20	99.8	100	strong=
HG2	ARG	24	0.355	0.354	0.001	20	99.8	100	strong=
HG3	ARG	24	1.142	1.149	-0.007	20	94.5	100	strong=

CD	ARG	24		43.194		20	94	0	strong
HD2	ARG	24		2.924		20	84.6	0	strong
HD3	ARG	24		3.002		20	81.4	0	strong
NE	ARG	24	121.002						
HE	ARG	24	7.756	7.216	0.54	20	70	0	
C	ARG	24	177.991	177.991	0	20	100	100	strong=
N	LEU	25	116.555	116.554	0.001	20	100	100	strong=
H	LEU	25	8.115	8.115	0	20	100	100	strong=
CA	LEU	25	57.851	57.873	-0.022	20	100	100	strong=
HA	LEU	25	4.136	4.129	0.007	20	100	100	strong=
CB	LEU	25	42.112	42.114	-0.002	20	100	100	strong=
HB2	LEU	25	1.556	1.563	-0.007	20	80.9	65	strong=
HB3	LEU	25	1.665	1.609	0.056	20	69	30	! (HG)
CG	LEU	25	27.07	27.099	-0.029	20	99.3	100	strong=
HG	LEU	25	1.628	1.624	0.004	20	99.9	100	strong=
QD1	LEU	25	0.905	0.91	-0.005	20	97.4	100	strong=
QD2	LEU	25	0.94	0.932	0.008	20	97.5	100	strong=
CD1	LEU	25	24.294	24.273	0.021	20	99.8	100	strong=
CD2	LEU	25	24.555	24.376	0.179	20	99.5	100	strong=
C	LEU	25	179.75	179.75	0	20	100	100	strong=
N	LYS	26	120.385	120.411	-0.026	20	99.9	100	strong=
H	LYS	26	7.732	7.723	0.009	20	99.9	100	strong=
CA	LYS	26	58.369	58.366	0.003	20	99.9	100	strong=
HA	LYS	26	4.121	4.121	0	20	99.9	100	strong=
CB	LYS	26	31.9	31.995	-0.095	20	99.9	100	strong=
HB2	LYS	26	1.831	1.835	-0.004	20	99.6	100	strong=
HB3	LYS	26	1.839	1.84	-0.001	20	99.1	100	strong=
CG	LYS	26	25.745	25.757	-0.012	20	92.6	95	strong=
HG2	LYS	26	1.473	1.464	0.009	20	83.1	85	strong=
HG3	LYS	26	1.471	1.47	0.001	20	97.1	100	strong=
CD	LYS	26	28.859	28.923	-0.064	20	74.6	75	=
HD2	LYS	26	1.82	1.693	0.127	20	39.7	5	! (HB3 25)
HD3	LYS	26	1.836	1.834	0.002	20	59.9	60	=
CE	LYS	26		42.263		20	99.6	0	strong
HE2	LYS	26		2.695		20	79.5	0	
HE3	LYS	26		2.992		20	91	0	strong
C	LYS	26	178.669	178.669	0	20	100	100	strong=
N	ALA	27	120.819	120.817	0.002	20	100	100	strong=
H	ALA	27	8.692	8.691	0.001	20	100	100	strong=
CA	ALA	27	55.128	55.113	0.015	20	100	100	strong=
HA	ALA	27	3.91	3.902	0.008	20	99.8	100	strong=

QB	ALA	27	1.507	1.506	0.001	20	99.9	100	strong=
CB	ALA	27	19.941	19.95	-0.009	20	99.9	100	strong=
C	ALA	27	178.699	178.699	0	20	100	100	strong=
N	ALA	28	117.589	117.58	0.009	20	99.9	100	strong=
H	ALA	28	8.205	8.209	-0.004	20	100	100	strong=
CA	ALA	28	54.906	54.896	0.01	20	100	100	strong=
HA	ALA	28	3.717	3.716	0.001	20	99.5	100	strong=
QB	ALA	28	1.508	1.505	0.003	20	100	100	strong=
CB	ALA	28	18.636	18.648	-0.012	20	99.9	100	strong=
C	ALA	28	177.953	177.953	0	20	100	100	strong=
N	GLN	29	114.892	114.889	0.003	20	100	100	strong=
H	GLN	29	7.664	7.666	-0.002	20	99.9	100	strong=
CA	GLN	29	59.168	59.138	0.03	20	99.8	100	strong=
HA	GLN	29	3.848	3.848	0	20	99.6	100	strong=
CB	GLN	29	28.119	28.156	-0.037	20	98.7	100	strong=
HB2	GLN	29	2.246	2.245	0.001	20	81.9	80	strong=
HB3	GLN	29	2.261	2.26	0.001	20	72.9	75	=
CG	GLN	29	33.913	33.984	-0.071	20	99.9	100	strong=
HG2	GLN	29	2.383	2.375	0.008	20	69.4	70	=
HG3	GLN	29	2.617	2.615	0.002	20	94.7	95	strong=
NE2	GLN	29	112.209	111.192	1.017	20	66.8	5	!
HE21	GLN	29	6.83	6.817	0.013	20	80.1	0	strong!
HE22	GLN	29	7.735	7.456	-0.279	20	54.4	0	!
C	GLN	29	179.535	179.535	0	20	100	100	strong=
N	LEU	30	116.906	116.907	-0.001	20	100	100	strong=
H	LEU	30	7.97	7.961	0.009	20	100	100	strong=
CA	LEU	30	58.103	58.065	0.038	20	99.9	100	strong=
HA	LEU	30	4.162	4.154	0.008	20	99.7	100	strong=
CB	LEU	30	42.224	42.234	-0.01	20	100	100	strong=
HB2	LEU	30	1.266	1.264	0.002	20	100	100	strong=
HB3	LEU	30	1.915	1.911	0.004	20	89.9	90	strong=
CG	LEU	30	26.586	26.562	0.024	20	97.8	100	strong=
HG	LEU	30	0.827	0.821	0.006	20	65	65	=
QD1	LEU	30	1.915	1.915	0	20	58.2	25	! (QG2 31)
QD2	LEU	30	0.244	0.243	0.001	20	50	0	
CD1	LEU	30	28.169	26.632	1.535	20	65.1	0	
CD2	LEU	30	26.625	28.17	1.545	20	50	45	! (CB 29)
C	LEU	30	180.506	180.506	0	20	100	100	strong=
N	VAL	31	109.067	109.07	-0.003	20	100	100	strong=
H	VAL	31	7.615	7.614	0.001	20	100	100	strong=

CA	VAL	31	63.5	63.505	-0.005	20	100	100	strong=
HA	VAL	31	4.167	4.165	0.002	20	99.9	100	strong=
CB	VAL	31	31.178	31.179	-0.001	20	100	100	strong=
HB	VAL	31	2.339	2.334	0.005	20	100	100	strong=
QG1	VAL	31	0.762	0.76	0.002	20	99.6	100	strong=
QG2	VAL	31	0.892	0.894	-0.002	20	90.3	90	strong=
CG1	VAL	31	18.455	18.428	0.027	20	99.5	100	strong=
CG2	VAL	31	21.107	21.162	-0.055	20	78.7	80	=
C	VAL	31	176.516	176.516	0	20	100	100	strong=
N	LEU	32	118.282	118.284	-0.002	20	100	100	strong=
H	LEU	32	7.675	7.679	-0.004	20	99.9	100	strong=
CA	LEU	32	55.129	55.094	0.035	20	100	100	strong=
HA	LEU	32	4.164	4.167	-0.003	20	99.8	100	strong=
CB	LEU	32	41.683	41.664	0.019	20	99.7	100	strong=
HB2	LEU	32	1.495	1.491	0.004	20	95	95	strong=
HB3	LEU	32	1.819	1.817	0.002	20	94.9	95	strong=
CG	LEU	32	26.381	26.4	-0.019	20	74.2	75	=
HG	LEU	32	1.828	1.824	0.004	20	69.8	70	=
QD1	LEU	32	0.49	0.479	0.011	20	100	100	strong=
QD2	LEU	32	0.793	0.787	0.006	20	82.5	85	strong=
CD1	LEU	32	21.766	21.738	0.028	20	99.9	100	strong=
CD2	LEU	32	24.833	24.875	-0.042	20	97.7	100	strong=
C	LEU	32	178.725	178.725	0	20	100	100	strong=
N	GLU	33	123.375	123.376	-0.001	20	100	100	strong=
H	GLU	33	7.134	7.13	0.004	20	99.9	100	strong=
CA	GLU	33	59.05	59.039	0.011	20	99.9	100	strong=
HA	GLU	33	4.12	4.121	-0.001	20	99.5	100	strong=
CB	GLU	33	29.244	29.19	0.054	20	99.1	100	strong=
HB2	GLU	33	2.08	2.095	-0.015	20	88.5	100	strong=
HB3	GLU	33	2.114	2.11	0.004	20	97.5	100	strong=
CG	GLU	33	35.768	35.834	-0.066	20	63.3	55	=
HG2	GLU	33	2.369	2.374	-0.005	20	79	80	=
HG3	GLU	33	2.366	2.42	-0.054	20	70.2	30	!
C	GLU	33	177.716	177.716	0	20	100	100	strong=
N	GLY	34	112.814	112.812	0.002	20	100	100	strong=
H	GLY	34	8.958	8.954	0.004	20	99.9	100	strong=
CA	GLY	34	45.121	45.123	-0.002	20	98.5	100	strong=
HA2	GLY	34	3.809	3.859	-0.05	20	55	5	!
HA3	GLY	34	4.125	4.137	-0.012	20	62.3	65	=
C	GLY	34	175.482	175.482	0	20	60	60	=
N	GLY	35	109.557	109.532	0.025	20	99.8	100	strong=
H	GLY	35	8.448	8.451	-0.003	20	99.9	100	strong=

CA	GLY	35	46.116	46.14	-0.024	20	99.9	100	strong=
HA2	GLY	35	3.944	3.937	0.007	20	94.9	95	strong=
HA3	GLY	35	4.176	4.177	-0.001	20	58.1	30	!
C	GLY	35	177.269	177.269	0	20	100	100	strong=
N	THR	36	119.791	119.791	0	20	100	100	strong=
H	THR	36	9.091	9.089	0.002	20	100	100	strong=
CA	THR	36	65.149	65.138	0.011	20	100	100	strong=
HA	THR	36	3.701	3.699	0.002	20	99.9	100	strong=
CB	THR	36	68.191	68.176	0.015	20	100	100	strong=
HB	THR	36	4.211	4.208	0.003	20	100	100	strong=
QG2	THR	36	1.192	1.187	0.005	20	100	100	strong=
CG2	THR	36	22.186	22.192	-0.006	20	99.9	100	strong=
C	THR	36	177.122	177.11	0.012	20	60	60	=
N	GLY	37	109.715	109.722	-0.007	20	99.9	100	strong=
H	GLY	37	8.439	8.436	0.003	20	99.2	100	strong=
CA	GLY	37	46.227	46.295	-0.068	20	99.7	100	strong=
HA2	GLY	37	3.929	3.928	0.001	20	94.6	5	strong!
HA3	GLY	37	4.058	4.059	-0.001	20	59.9	5	!
C	GLY	37	175.035	175.035	0	20	100	100	strong=
N	SER	38	116.388	116.377	0.011	20	100	100	strong=
H	SER	38	7.314	7.311	0.003	20	100	100	strong=
CA	SER	38	57.343	57.362	-0.019	20	100	100	strong=
HA	SER	38	4.623	4.625	-0.002	20	99.9	100	strong=
CB	SER	38	64.156	64.148	0.008	20	99.7	100	strong=
HB2	SER	38	3.642	3.644	-0.002	20	92	95	strong=
HB3	SER	38	4.308	4.306	0.002	20	100	100	strong=
C	SER	38	171.835	171.835	0	20	100	100	strong=
N	HIS	39	121.234	121.235	-0.001	20	100	100	strong=
H	HIS	39	7.125	7.124	0.001	20	98	100	strong=
CA	HIS	39	57.926	57.931	-0.005	20	99.9	100	strong=
HA	HIS	39	4.405	4.399	0.006	20	98.5	100	strong=
CB	HIS	39	31.406	31.403	0.003	20	95	95	strong=
HB2	HIS	39	3.099	3.098	0.001	20	99.1	100	strong=
HB3	HIS	39	3.102	3.1	0.002	20	99.3	100	strong=
ND1	HIS	39		121.25		9	66.7	0	
CD2	HIS	39	120.04	120.114	-0.074	20	74.4	75	=
HD1	HIS	39		7.132		20	43.5	0	
CE1	HIS	39	137.25	137.353	-0.103	20	85	85	strong=
HD2	HIS	39	7.122	7.125	-0.003	20	89.3	90	strong=
HE1	HIS	39	8.302	8.333	-0.031	20	76.6	15	!
C	HIS	39	176.881	176.881	0	20	100	100	strong=
N	GLY	40	116.135	116.132	0.003	20	100	100	strong=

H	GLY	40	9.584	9.584	0	20	100	100	strong=
CA	GLY	40	45.878	45.861	0.017	20	100	100	strong=
HA2	GLY	40	3.516	3.518	-0.002	20	95.1	95	strong=
HA3	GLY	40	4.577	4.57	0.007	20	95	95	strong=
C	GLY	40	174.767	174.763	0.004	20	95	95	strong=
N	VAL	41	123.341	123.327	0.014	20	100	100	strong=
H	VAL	41	8.481	8.476	0.005	20	99.7	100	strong=
CA	VAL	41	62.793	62.784	0.009	20	100	100	strong=
HA	VAL	41	5.183	5.176	0.007	20	100	100	strong=
CB	VAL	41	31.544	31.523	0.021	20	99.9	100	strong=
HB	VAL	41	2.372	2.367	0.005	20	99.3	100	strong=
QG1	VAL	41	1.017	1.012	0.005	20	65	65	=
QG2	VAL	41	1.016	1.016	0	20	99.6	100	strong=
CG1	VAL	41	22.722	22.699	0.023	20	99.6	100	strong=
CG2	VAL	41	22.723	22.73	-0.007	20	80	80	=
C	VAL	41	176.376	176.376	0	20	100	100	strong=
N	PHE	42	121.368	121.366	0.002	20	100	100	strong=
H	PHE	42	8.488	8.482	0.006	20	100	100	strong=
CA	PHE	42	55.67	55.664	0.006	20	100	100	strong=
HA	PHE	42	6.257	6.254	0.003	20	100	100	strong=
CB	PHE	42	45.94	45.962	-0.022	20	100	100	strong=
HB2	PHE	42	2.882	2.877	0.005	20	99.9	100	strong=
HB3	PHE	42	3.241	3.234	0.007	20	99.3	100	strong=
CD1	PHE	42		129.295		20	80	0	
HD1	PHE	42		6.245		20	89.9	0	strong
CE1	PHE	42		123.847		20	83.8	0	strong
HE1	PHE	42		6.246		20	74.9	0	
CZ	PHE	42		123.686		20	69.4	0	
HZ	PHE	42		5.872		20	64.9	0	
CE2	PHE	42		132.435		20	45	0	
HE2	PHE	42		7.074		20	39.8	0	
CD2	PHE	42		131.395		20	94.9	0	strong
HD2	PHE	42		7.226		20	94.8	0	strong
C	PHE	42	172.701	172.701	0	20	100	100	strong=
N	LEU	43	113.423	113.423	0	20	100	100	strong=
H	LEU	43	9.391	9.389	0.002	20	100	100	strong=
CA	LEU	43	55.634	55.603	0.031	20	100	100	strong=
HA	LEU	43	4.754	4.758	-0.004	20	99.8	100	strong=
CB	LEU	43	44.608	44.69	-0.082	20	99.3	100	strong=
HB2	LEU	43	1.975	1.975	0	20	57.8	0	!
HB3	LEU	43	1.705	1.708	-0.003	20	94.8	95	strong=
CG	LEU	43	24.931	24.835	0.096	20	89.3	90	strong=

HG	LEU	43	1.618	1.613	0.005	20	89.9	90	strong=
QD1	LEU	43	0.237	0.234	0.003	20	85	85	strong=
QD2	LEU	43	0.746	0.745	0.001	20	50.4	0	!
CD1	LEU	43	28.152	25.129	3.023	20	55	55	=
CD2	LEU	43	25.129	28.192	-3.063	20	88	10	strong!
C	LEU	43	174.869	174.869	0	20	100	100	strong=
N	VAL	44	119.826	119.824	0.002	20	100	100	strong=
H	VAL	44	9.193	9.192	0.001	20	100	100	strong=
CA	VAL	44	61.215	61.249	-0.034	20	100	100	strong=
HA	VAL	44	5.66	5.654	0.006	20	100	100	strong=
CB	VAL	44	34.378	34.363	0.015	20	99.9	100	strong=
HB	VAL	44	2.631	2.627	0.004	20	100	100	strong=
QG1	VAL	44	1.397	1.396	0.001	20	98.3	10	strong!
QG2	VAL	44	1.106	1.102	0.004	20	90	90	strong=
CG1	VAL	44	21.991	21.983	0.008	20	100	100	strong=
CG2	VAL	44	22.704	22.873	-0.169	20	94.8	95	strong=
C	VAL	44	173.386	173.314	0.072	20	100	100	strong=
N	ARG	45	123.902	123.887	0.015	20	100	100	strong=
H	ARG	45	9.494	9.488	0.006	20	100	100	strong=
CA	ARG	45	53.468	53.558	-0.09	20	98	100	strong=
HA	ARG	45	5.163	5.125	0.038	20	83.3	30	strong!
CB	ARG	45	34.247	34.245	0.002	20	90.3	5	strong!
HB2	ARG	45	1.403	1.401	0.002	20	89.7	80	strong=
HB3	ARG	45	1.354	1.37	-0.016	20	81.7	75	strong=
CG	ARG	45	26.803	27.1	-0.297	20	99.5	100	strong=
HG2	ARG	45	0.738	0.738	0	20	72.9	0	!
HG3	ARG	45	0.733	0.732	0.001	20	99.8	0	strong!
CD	ARG	45	43.204	43.184	0.02	20	90	90	strong=
HD2	ARG	45	2.868	2.691	0.177	20	54.9	55	=
HD3	ARG	45	2.691	2.867	-0.176	20	84.8	10	strong!
NE	ARG	45	120.584						
HE	ARG	45	6.997	7.215	-0.218	20	40	0	
C	ARG	45	174.606	174.587	0.019	20	99.9	100	strong=
N	GLN	46	120.594	120.6	-0.006	20	100	100	strong=
H	GLN	46	8.822	8.815	0.007	20	99.7	100	strong=
CA	GLN	46	56.194	56.095	0.099	20	99.6	100	strong=
HA	GLN	46	4.566	4.563	0.003	20	99.8	100	strong=
CB	GLN	46	30.971	31.013	-0.042	20	99.1	100	strong=
HB2	GLN	46	1.918	1.906	0.012	20	85.8	85	strong=
HB3	GLN	46	2.17	2.143	0.027	20	81.5	70	strong=
CG	GLN	46	36.178	36.02	0.158	20	64	65	=
HG2	GLN	46	2.203	2.48	-0.277	20	69	20	! (HG3)

HG3	GLN	46	2.48	2.203	0.277	20	89	100	strong=
NE2	GLN	46	109.953	110.493	-0.319	20	96.3	75	strong=
HE21	GLN	46	7.339	6.559	0.823	20	84.9	0	strong!
HE22	GLN	46	6.905	6.939	-0.013	20	94.3	95	strong=
C	GLN	46	175.193	175.187	0.006	20	99.7	100	strong=
N	SER	47	114.948	114.949	-0.001	20	100	100	strong=
H	SER	47	7.755	7.753	0.002	20	99.3	100	strong=
CA	SER	47	57.296	57.311	-0.015	20	95.7	95	strong=
HA	SER	47	4.491	4.493	-0.002	20	89.2	90	strong=
CB	SER	47	63.708	63.708	0	20	50	50	!
HB2	SER	47	3.845	3.647	0.198	20	76.5	0	! (HB3)
HB3	SER	47	3.647	3.845	-0.198	20	75.6	55	=
C	SER	47	176.075	176.075	0	20	100	100	strong=
N	GLU	48	129.121	129.121	0	20	100	100	strong=
H	GLU	48	10.481	10.477	0.004	20	99.9	100	strong=
CA	GLU	48	57.876	57.861	0.015	20	99.9	100	strong=
HA	GLU	48	4.303	4.299	0.004	20	94.9	95	strong=
CB	GLU	48	31.246	31.346	-0.1	20	98.8	100	strong=
HB2	GLU	48	2.13	2.124	0.006	20	97.9	100	strong=
HB3	GLU	48	2.161	2.133	0.028	20	91.6	30	strong=
CG	GLU	48	36.88	36.807	0.073	20	94	100	strong=
HG2	GLU	48	2.15	2.339	-0.189	20	83.7	5	strong!
HG3	GLU	48	2.402	2.407	-0.005	20	45.1	45	=
C	GLU	48	177.977	177.977	0	20	100	100	strong=
N	THR	49	109.721	109.719	0.002	20	100	100	strong=
H	THR	49	8.157	8.157	0	20	99.9	100	strong=
CA	THR	49	63.263	63.251	0.012	20	100	100	strong=
HA	THR	49	4.305	4.3	0.005	20	100	100	strong=
CB	THR	49	70.379	70.387	-0.008	20	100	100	strong=
HB	THR	49	4.119	4.111	0.008	20	100	100	strong=
QG2	THR	49	1.245	1.24	0.005	20	100	100	strong=
CG2	THR	49	22.167	22.158	0.009	20	100	100	strong=
C	THR	49	175.265	175.233	0.032	20	99.1	100	strong=
N	ARG	50	124.881	124.88	0.001	20	100	100	strong=
H	ARG	50	8.184	8.187	-0.003	20	98.9	100	strong=
CA	ARG	50	54.464	54.502	-0.038	20	94.9	95	strong=
HA	ARG	50	4.538	4.538	0	20	94.8	95	strong=
CB	ARG	50	29.666	29.666	0	20	59.8	40	!
HB2	ARG	50	1.518	1.519	-0.001	20	98.3	100	strong=
HB3	ARG	50	1.69	1.69	0	20	72	50	=
CG	ARG	50	26.875	26.81	0.065	20	99.1	100	strong=
HG2	ARG	50	1.504	1.504	0	20	98.8	100	strong=

HG3	ARG	50	1.507	1.51	-0.003	20	73.3	75	=
CD	ARG	50	43.271	43.239	0.032	20	95.5	95	strong=
HD2	ARG	50	3.129	3.129	0	20	95.1	95	strong=
HD3	ARG	50	3.134	3.133	0.001	20	99.8	100	strong=
NE	ARG	50	121.024						
HE	ARG	50	7.432	7.204	0.228	20	49.3	0	
C	ARG	50	175.152	175.052	0.1	20	87.6	100	strong=
N	ARG	51	123.424	123.362	0.062	20	99.9	100	strong=
H	ARG	51	8.302	8.304	-0.002	20	98.9	100	strong=
CA	ARG	51	57.981	57.997	-0.016	20	100	100	strong=
HA	ARG	51	4.09	4.084	0.006	20	100	100	strong=
CB	ARG	51	29.584	29.547	0.037	20	99.7	100	strong=
HB2	ARG	51	1.803	1.791	0.012	20	94.7	95	strong=
HB3	ARG	51	1.794	1.796	-0.002	20	99.5	100	strong=
CG	ARG	51	27.004	27.017	-0.013	20	100	100	strong=
HG2	ARG	51	1.709	1.707	0.002	20	98.1	100	strong=
HG3	ARG	51	1.711	1.713	-0.002	20	93.3	95	strong=
CD	ARG	51	43.244	43.263	-0.019	20	100	100	strong=
HD2	ARG	51	3.241	3.236	0.005	20	100	100	strong=
HD3	ARG	51	3.241	3.237	0.004	20	100	100	strong=
NE	ARG	51	120.345						
HE	ARG	51	7.855	6.959	0.896	20	80	0	
C	ARG	51	178.046	178.046	0	20	100	100	strong=
N	GLY	52	113.064	113.068	-0.004	20	100	100	strong=
H	GLY	52	8.859	8.858	0.001	20	100	100	strong=
CA	GLY	52	45.377	45.387	-0.01	20	100	100	strong=
HA2	GLY	52	3.792	3.789	0.003	20	98.3	100	strong=
HA3	GLY	52	4.28	4.279	0.001	20	99.8	100	strong=
C	GLY	52	173.89	173.89	0	20	100	100	strong=
N	GLU	53	118.594	118.592	0.002	20	100	100	strong=
H	GLU	53	7.828	7.826	0.002	20	99.6	100	strong=
CA	GLU	53	55.6	55.604	-0.004	20	100	100	strong=
HA	GLU	53	4.846	4.846	0	20	98.9	100	strong=
CB	GLU	53	32.008	32.051	-0.043	20	97.6	100	strong=
HB2	GLU	53	2.066	2.057	0.009	20	99.6	100	strong=
HB3	GLU	53	2.466	2.459	0.007	20	85.3	85	strong=
CG	GLU	53	36.972	36.955	0.017	20	97	100	strong=
HG2	GLU	53	2.276	2.284	-0.008	20	67.3	55	=
HG3	GLU	53	2.327	2.317	0.01	20	83.7	85	strong=
C	GLU	53	175.856	175.856	0	20	100	100	strong=
N	TYR	54	119.848	119.853	-0.005	20	100	100	strong=
H	TYR	54	8.783	8.783	0	20	100	100	strong=

CA	TYR	54	56.896	56.888	0.008	20	100	100	strong=
HA	TYR	54	5.327	5.33	-0.003	20	99.9	100	strong=
CB	TYR	54	42.131	42.062	0.069	20	99.7	100	strong=
HB2	TYR	54	2.776	2.771	0.005	20	99.9	100	strong=
HB3	TYR	54	3.143	3.144	-0.001	20	99.9	100	strong=
CD1	TYR	54	132.539	132.558	-0.019	20	74.7	75	=
HD1	TYR	54	7.058	7.058	0	20	99.9	100	strong=
CE1	TYR	54	118.3	120.321	-2.021	20	92	0	strong!
HE1	TYR	54	6.842	7.067	-0.225	20	99.4	0	strong!
CE2	TYR	54	118.304	120.278	-1.974	20	70.1	0	!
HE2	TYR	54	6.841	7.092	-0.251	20	48.6	0	!
CD2	TYR	54	132.593	132.608	-0.015	20	99.6	100	strong=
HD2	TYR	54	7.055	7.059	-0.004	20	100	100	strong=
HH	TYR	54		8.285		20	67.9	0	
C	TYR	54	174.511	174.519	-0.008	20	99.2	100	strong=
N	VAL	55	121.179	121.179	0	20	100	100	strong=
H	VAL	55	9.641	9.64	0.001	20	100	100	strong=
CA	VAL	55	61.584	61.564	0.02	20	100	100	strong=
HA	VAL	55	4.872	4.871	0.001	20	99.7	100	strong=
CB	VAL	55	36.64	36.692	-0.052	20	100	100	strong=
HB	VAL	55	1.792	1.788	0.004	20	100	100	strong=
QG1	VAL	55	0.974	0.974	0	20	99.6	100	strong=
QG2	VAL	55	1.094	1.092	0.002	20	94.8	95	strong=
CG1	VAL	55	21.628	21.713	-0.085	20	97.7	100	strong=
CG2	VAL	55	23.112	23.114	-0.002	20	76.2	10	!
C	VAL	55	174.203	174.195	0.008	20	99.2	100	strong=
N	LEU	56	129.43	129.424	0.006	20	100	100	strong=
H	LEU	56	9.38	9.38	0	20	100	100	strong=
CA	LEU	56	53.862	53.854	0.008	20	100	100	strong=
HA	LEU	56	5.258	5.253	0.005	20	100	100	strong=
CB	LEU	56	44.65	44.634	0.016	20	100	100	strong=
HB2	LEU	56	1.299	1.296	0.003	20	100	100	strong=
HB3	LEU	56	2.165	2.165	0	20	100	100	strong=
CG	LEU	56	28.165	28.081	0.084	20	98.4	100	strong=
HG	LEU	56	1.611	1.612	-0.001	20	99.5	100	strong=
QD1	LEU	56	0.703	0.695	0.008	20	99.6	100	strong=
QD2	LEU	56	0.619	0.612	0.007	20	98.9	100	strong=
CD1	LEU	56	25.559	25.874	-0.315	20	75.9	55	=
CD2	LEU	56	25.868	25.93	-0.062	20	99.8	100	strong=
C	LEU	56	174.979	174.979	0	20	100	100	strong=
N	THR	57	128.481	128.481	0	20	100	100	strong=
H	THR	57	9.093	9.09	0.003	20	100	100	strong=

CA	THR	57	62.727	62.719	0.008	20	100	100	strong=
HA	THR	57	5.565	5.559	0.006	20	100	100	strong=
CB	THR	57	71.762	71.763	-0.001	20	100	100	strong=
HB	THR	57	3.661	3.657	0.004	20	99.9	100	strong=
QG2	THR	57	0.97	0.965	0.005	20	99.8	100	strong=
CG2	THR	57	22.693	22.687	0.006	20	100	100	strong=
C	THR	57	173.529	173.529	0	20	100	100	strong=
N	PHE	58	121.074	121.074	0	20	100	100	strong=
H	PHE	58	8.942	8.942	0	20	100	100	strong=
CA	PHE	58	55.376	55.361	0.015	20	100	100	strong=
HA	PHE	58	5.885	5.878	0.007	20	100	100	strong=
CB	PHE	58	42.511	42.523	-0.012	20	100	100	strong=
HB2	PHE	58	2.838	2.842	-0.004	20	99	100	strong=
HB3	PHE	58	2.859	2.855	0.004	20	98.9	100	strong=
CD1	PHE	58	132.093	132.348	-0.255	20	65	65	=
HD1	PHE	58	6.846	6.853	-0.007	20	69.4	70	=
CE1	PHE	58	132.112	129.189	2.923	20	35	15	! (CZ)
HE1	PHE	58	6.85	6.865	-0.015	20	71.4	55	=
CZ	PHE	58	128.915	125.49	3.425	20	45	10	!
HZ	PHE	58	6.847	6.86	-0.013	20	84	80	strong=
CE2	PHE	58	132.122	134.288	-2.166	20	59.9	15	!
HE2	PHE	58	6.842	6.892	-0.05	20	76	10	! (HD22
CD2	PHE	58	132.089	132.395	-0.306	20	65	65	=
HD2	PHE	58	6.84	6.854	-0.014	20	93.3	85	strong=
C	PHE	58	171.864	171.864	0	20	100	100	strong=
N	ASN	59	119.488	119.49	-0.002	20	100	100	strong=
H	ASN	59	8.461	8.46	0.001	20	100	100	strong=
CA	ASN	59	52.018	52.02	-0.002	20	100	100	strong=
HA	ASN	59	4.432	4.427	0.005	20	100	100	strong=
CB	ASN	59	39.046	39.043	0.003	20	100	100	strong=
HB2	ASN	59	2.396	2.392	0.004	20	99.6	100	strong=
HB3	ASN	59	3.464	3.46	0.004	20	99.9	100	strong=
ND2	ASN	59	110.723	110.802	-0.007	20	49.6	45	=
HD21	ASN	59	7.45	6.83	0.644	20	81.9	0	strong!
HD22	ASN	59	6.898	6.917	-0.028	20	37.9	25	=
C	ASN	59	174.146	174.146	0	20	100	100	strong=
N	PHE	60	128.735	128.736	-0.001	20	100	100	strong=
H	PHE	60	9.31	9.31	0	20	100	100	strong=
CA	PHE	60	55.479	55.395	0.084	20	99.4	100	strong=
HA	PHE	60		5.208		20	100	0	strong
CB	PHE	60	40.859	40.803	0.056	20	98.7	100	strong=
HB2	PHE	60		2.689		20	94.9	0	strong

HB3	PHE	60		3.02		20	94.3	0	strong
CD1	PHE	60	128.609	130.831	-2.222	20	49.7	0	!
HD1	PHE	60	7.172	6.948	0.224	20	65.6	0	! (HE1)
CE1	PHE	60	128.593	131.199	-2.606	20	47.4	0	!
HE1	PHE	60	6.933	7.224	-0.291	20	48.9	40	! (HE2)
CZ	PHE	60	124.221	123.842	0.379	20	44.9	35	=
HZ	PHE	60	7.126	5.199	1.927	20	44.9	0	!
CE2	PHE	60	133.029	133.177	-0.148	20	58.1	60	=
HE2	PHE	60	7.233	7.226	0.007	20	49.2	50	=
CD2	PHE	60	128.595	132.169	-3.574	19	91.9	0	strong!
HD2	PHE	60	7.184	7.211	-0.027	20	98.1	70	strong=
C	PHE	60	174.158	174.158	0	16	75	6.2	!
N	GLN	61	127.582	110.799	16.783	20	70.1	15	! (NE2)
H	GLN	61	11.411	7.446	3.965	20	71.7	15	!
CA	GLN	61	56.068	56.062	0.006	20	100	100	strong=
HA	GLN	61	3.565	3.561	0.004	20	99.7	100	strong=
CB	GLN	61	26.59	26.558	0.032	20	99.8	100	strong=
HB2	GLN	61	1.594	1.601	-0.007	20	83	85	strong=
HB3	GLN	61	1.981	1.983	-0.002	20	99.8	100	strong=
CG	GLN	61	32.864	32.867	-0.003	20	99.6	100	strong=
HG2	GLN	61	1.525	1.521	0.004	20	99.8	100	strong=
HG3	GLN	61	1.723	1.72	0.003	20	64	65	=
NE2	GLN	61	110.593	112.185	-1.463	20	94.2	5	strong!
HE21	GLN	61	6.941	6.946	-0.005	20	73.4	80	=
HE22	GLN	61	6.575	7.494	-0.919	20	59.6	0	!
C	GLN	61	175.832	175.888	-0.056	20	99.2	100	strong=
N	GLY	62	103.813	103.814	-0.001	20	100	100	strong=
H	GLY	62	9.058	9.059	-0.001	20	99.6	100	strong=
CA	GLY	62	45.345	45.339	0.006	20	99.7	100	strong=
HA2	GLY	62	3.183	3.183	0	20	99.9	100	strong=
HA3	GLY	62	4.093	4.09	0.003	20	99.3	100	strong=
C	GLY	62	172.743	172.743	0	20	95	95	strong=
N	LYS	63	120.383	120.405	-0.022	20	100	100	strong=
H	LYS	63	7.912	7.92	-0.008	20	99.9	100	strong=
CA	LYS	63	54.489	54.441	0.048	20	99.8	100	strong=
HA	LYS	63	4.532	4.533	-0.001	20	99.9	100	strong=
CB	LYS	63	34.558	34.554	0.004	20	99.9	100	strong=
HB2	LYS	63	1.834	1.833	0.001	20	94.7	95	strong=
HB3	LYS	63	1.843	1.85	-0.007	20	92.1	100	strong=
CG	LYS	63	28.887	25.324	3.565	20	99.8	0	strong
HG2	LYS	63	1.714	1.441	0.273	20	99.9	0	strong
HG3	LYS	63	1.452	1.443	0.009	20	99.7	100	strong=

CD	LYS	63	29	28.933	0.067	20	99.8	0	strong!
HD2	LYS	63	1.443	1.439	0.004	20	64.9	65	=
HD3	LYS	63	1.7	1.719	-0.019	20	85.1	0	strong!
CE	LYS	63		42.312		20	99.9	0	strong
HE2	LYS	63		3.02		20	63.2	0	
HE3	LYS	63		3.023		20	98.1	0	strong
C	LYS	63	174.655	174.627	0.028	20	99.3	100	strong=
N	ALA	64	124.468	124.468	0	20	100	100	strong=
H	ALA	64	8.781	8.781	0	20	100	100	strong=
CA	ALA	64	51.511	51.526	-0.015	20	100	100	strong=
HA	ALA	64	4.474	4.47	0.004	20	99.9	100	strong=
QB	ALA	64	1.032	1.033	-0.001	20	99.9	100	strong=
CB	ALA	64	19.244	19.267	-0.023	20	100	100	strong=
C	ALA	64	176.328	176.328	0	20	100	100	strong=
N	LYS	65	123.298	123.301	-0.003	20	100	100	strong=
H	LYS	65	8.563	8.56	0.003	20	100	100	strong=
CA	LYS	65	53.008	52.964	0.044	20	99.9	100	strong=
HA	LYS	65	4.33	4.324	0.006	20	99.9	100	strong=
CB	LYS	65	34.836	34.813	0.023	20	100	100	strong=
HB2	LYS	65	0.293	0.293	0	20	95	95	strong=
HB3	LYS	65	0.985	0.981	0.004	20	99.7	100	strong=
CG	LYS	65	25.121	25.065	0.056	20	93.8	95	strong=
HG2	LYS	65	0.993	0.99	0.003	20	62.8	65	=
HG3	LYS	65	1.023	1.015	0.008	20	96.5	0	strong!
CD	LYS	65	28.371	28.359	0.012	20	100	100	strong=
HD2	LYS	65	1.461	1.452	0.009	20	90.1	90	strong=
HD3	LYS	65	1.615	1.609	0.006	20	85	85	strong=
CE	LYS	65		42.384		20	94.9	0	strong
HE2	LYS	65		2.697		20	99.9	0	strong
HE3	LYS	65		2.699		20	94.9	0	strong
C	LYS	65	173.797	173.797	0	20	100	100	strong=
N	HIS	66	118.577	118.668	-0.091	20	100	100	strong=
H	HIS	66	8.297	8.305	-0.008	20	99.6	100	strong=
CA	HIS	66	54.631	54.634	-0.003	20	99.8	100	strong=
HA	HIS	66	5.142	5.138	0.004	20	99.9	100	strong=
CB	HIS	66	32.928	32.885	0.043	20	97.8	100	strong=
HB2	HIS	66	2.691	2.649	0.042	20	54.9	0	!
HB3	HIS	66	2.859	2.856	0.003	20	77.9	85	=
ND1	HIS	66		109.885		18	100	0	strong
CD2	HIS	66	120.246	120.245	0.001	20	99.4	100	strong=
HD1	HIS	66		5.558		20	89.9	0	strong
CE1	HIS	66	137.327	132.379	4.948	20	55	0	!

HD2	HIS	66	6.968	6.963	0.005	20	99.8	100	strong=
HE1	HIS	66	8.318	6.852	1.466	20	55	0	!
C	HIS	66	175.021	175.021	0	20	99.6	100	strong=
N	LEU	67	127.284	127.327	-0.043	20	100	100	strong=
H	LEU	67	9.545	9.548	-0.003	20	100	100	strong=
CA	LEU	67	53.468	53.493	-0.025	20	100	100	strong=
HA	LEU	67	4.856	4.856	0	20	100	100	strong=
CB	LEU	67	44.768	44.802	-0.034	20	100	100	strong=
HB2	LEU	67	1.714	1.719	-0.005	20	77.1	80	=
HB3	LEU	67	1.841	1.836	0.005	20	99.9	100	strong=
CG	LEU	67	24.703	24.702	0.001	20	99.2	0	strong!
HG	LEU	67	0.814	0.819	0.004	20	94.7	0	strong!
QD1	LEU	67	0.82	0.806	0.014	20	70.1	25	!
QD2	LEU	67	0.806	0.821	-0.014	20	94.9	0	strong!
CD1	LEU	67	24.701	24.701	0	20	44.5	30	!
CD2	LEU	67	24.764	24.766	-0.002	20	94.8	5	strong!
C	LEU	67	175.371	175.375	-0.004	20	99.5	100	strong=
N	ARG	68	125.469	125.471	-0.002	20	100	100	strong=
H	ARG	68	8.836	8.838	-0.002	20	99.1	100	strong=
CA	ARG	68	58.046	58.001	0.045	20	99.9	100	strong=
HA	ARG	68	4.512	4.505	0.007	20	99.7	100	strong=
CB	ARG	68	30.804	30.72	0.084	20	95.6	95	strong=
HB2	ARG	68	1.837	1.821	0.016	20	47.2	45	=
HB3	ARG	68	1.941	1.945	-0.004	20	93.9	95	strong=
CG	ARG	68	27.618	27.619	-0.001	20	75.3	5	!
HG2	ARG	68	1.939	1.939	0	20	40	0	! (HG 67)
HG3	ARG	68	1.941	1.945	-0.004	20	79.2	0	! (HB3)
CD	ARG	68	43.422	42.422	0	20	70	30	!
HD2	ARG	68	3.143	3.143	0	20	69.9	0	!
HD3	ARG	68	3.263	3.263	0	20	74.9	45	!
NE	ARG	68	120.577						
HE	ARG	68	7.576	6.841	0.735	20	73.8	0	
C	ARG	68	175.462	175.466	-0.004	20	99.9	100	strong=
N	LEU	69	122.102	122.1	0.002	20	100	100	strong=
H	LEU	69	8.643	8.637	0.006	20	99.9	100	strong=
CA	LEU	69	53.762	53.753	0.009	20	100	100	strong=
HA	LEU	69	5.074	5.07	0.004	20	100	100	strong=
CB	LEU	69	44.635	44.622	0.013	20	100	100	strong=
HB2	LEU	69	1.382	1.38	0.002	20	99.9	100	strong=
HB3	LEU	69	1.71	1.708	0.002	20	90	90	strong=
CG	LEU	69	26.778	26.713	0.065	20	69	70	=

HG	LEU	69	0.74	0.736	0.004	20	84.9	85	strong=
QD1	LEU	69		1.063		20	45	0	
QD2	LEU	69	0.721	0.735	-0.014	20	99.3	100	strong=
CD1	LEU	69		25.255		20	53.7	0	
CD2	LEU	69	26.627	26.65	-0.023	20	58.4	60	=
C	LEU	69	175.374	175.374	0	20	100	100	strong=
N	SER	70	116.368	116.37	-0.002	20	100	100	strong=
H	SER	70	8.39	8.388	0.002	20	98.7	100	strong=
CA	SER	70	56.859	56.864	-0.005	20	99.9	100	strong=
HA	SER	70	5.176	5.173	0.003	20	99.6	100	strong=
CB	SER	70	65.038	65.031	0.007	20	99.5	100	strong=
HB2	SER	70	3.793	3.788	0.005	20	97	100	strong=
HB3	SER	70	3.824	3.826	-0.002	20	95.2	95	strong=
C	SER	70	173.082	173.082	0	20	100	100	strong=
N	LEU	71	126.057	126.057	0	20	100	100	strong=
H	LEU	71	8.986	8.98	0.006	20	99.7	100	strong=
CA	LEU	71	53.453	53.438	0.015	20	100	100	strong=
HA	LEU	71	5.698	5.695	0.003	20	100	100	strong=
CB	LEU	71	45.013	45.003	0.01	20	99.9	100	strong=
HB2	LEU	71	1.7	1.699	0.001	20	92.1	90	strong=
HB3	LEU	71	1.694	1.699	-0.005	20	99.4	100	strong=
CG	LEU	71	28.264	28.267	-0.003	20	80	80	=
HG	LEU	71	1.609	1.604	0.005	20	74.8	75	=
QD1	LEU	71	0.691	0.69	0.001	20	90	90	strong=
QD2	LEU	71	0.693	0.69	0.003	20	75	75	=
CD1	LEU	71	25.111	25.101	0.01	20	94.3	95	strong=
CD2	LEU	71	25.13	25.126	0.004	20	65	65	=
C	LEU	71	177.95	177.95	0	20	100	100	strong=
N	ASN	72	120.657	120.655	0.002	20	100	100	strong=
H	ASN	72	8.214	8.213	0.001	20	100	100	strong=
CA	ASN	72	50.612	50.578	0.034	20	100	100	strong=
HA	ASN	72	5.141	5.138	0.003	20	100	100	strong=
CB	ASN	72	39.23	39.223	0.007	20	100	100	strong=
HB2	ASN	72	3.064	3.063	0.001	20	99.9	100	strong=
HB3	ASN	72	3.672	3.666	0.006	20	99.4	100	strong=
ND2	ASN	72		111.913		20	98.6	0	strong
HD21	ASN	72		7.423		20	79.1	0	
HD22	ASN	72		7.656		20	85.4	0	strong
C	ASN	72	177.967	177.967	0	20	99.7	100	strong=
N	GLU	73	119.429	119.428	0.001	20	100	100	strong=
H	GLU	73	9.152	9.152	0	20	100	100	strong=
CA	GLU	73	59.606	59.601	0.005	20	100	100	strong=

HA	GLU	73	4.131	4.128	0.003	20	100	100	strong=
CB	GLU	73	29.044	29.055	-0.011	20	100	100	strong=
HB2	GLU	73	2.123	2.119	0.004	20	99.9	100	strong=
HB3	GLU	73	2.125	2.119	0.006	20	95	95	strong=
CG	GLU	73	36.493	36.474	0.019	20	100	100	strong=
HG2	GLU	73	2.434	2.427	0.007	20	99.8	100	strong=
HG3	GLU	73	2.434	2.427	0.007	20	100	100	strong=
C	GLU	73	177.586	177.586	0	20	99.5	100	strong=
N	GLU	74	116.474	116.483	-0.009	20	100	100	strong=
H	GLU	74	7.632	7.628	0.004	20	100	100	strong=
CA	GLU	74	56.375	56.382	-0.007	20	100	100	strong=
HA	GLU	74	4.44	4.442	-0.002	20	100	100	strong=
CB	GLU	74	30.057	30.043	0.014	20	99.6	100	strong=
HB2	GLU	74	1.994	1.992	0.002	20	80	80	=
HB3	GLU	74	2.328	2.325	0.003	20	99.6	100	strong=
CG	GLU	74	36.703	36.669	0.034	20	99.6	100	strong=
HG2	GLU	74	2.333	2.335	-0.003	20	45	35	! (HG3)
HG3	GLU	74	2.332	2.334	-0.002	20	99.7	100	strong=
C	GLU	74	176.724	176.725	-0.001	20	100	100	strong=
N	GLY	75	108.564	108.565	-0.001	20	100	100	strong=
H	GLY	75	8.255	8.252	0.003	20	99.9	100	strong=
CA	GLY	75	45.509	45.496	0.013	20	100	100	strong=
HA2	GLY	75	3.608	3.605	0.003	20	99.6	100	strong=
HA3	GLY	75	4.333	4.329	0.004	20	100	100	strong=
C	GLY	75	174.275	174.283	-0.008	20	98.7	100	strong=
N	GLN	76	118.74	118.744	-0.004	20	100	100	strong=
H	GLN	76	7.94	7.937	0.003	20	99.9	100	strong=
CA	GLN	76	56.521	56.521	0	20	95.3	95	strong=
HA	GLN	76	4.13	4.128	0.002	20	93.9	95	strong=
CB	GLN	76	27.969	27.83	0.139	20	84.5	85	strong=
HB2	GLN	76	2.065	2.064	0.001	20	54.6	55	=
HB3	GLN	76	2.062	2.068	-0.006	20	74.1	75	=
CG	GLN	76	34.007	34.096	-0.089	20	91.8	85	strong=
HG2	GLN	76	2.378	2.371	0.007	20	85.9	90	strong=
HG3	GLN	76	2.619	2.622	-0.003	20	59.3	60	=
NE2	GLN	76	111.962	111.187	0.775	20	59.4	40	!
HE21	GLN	76	6.801	6.822	-0.021	20	87.1	85	strong=
HE22	GLN	76	7.489	7.467	0.022	20	45.7	10	!
C	GLN	76	175.471	175.506	-0.035	20	99.5	100	strong=
N	CYS	77	120.004	119.985	0.019	20	95	95	strong=
H	CYS	77	8.363	8.363	0	20	94.8	95	strong=
CA	CYS	77	56.43	56.501	-0.071	20	89.9	90	strong=

HA	CYS	77	5.366	5.364	0.002	20	89.6	90	strong=
CB	CYS	77	30.157	30.029	0.128	20	74.7	75	=
HB2	CYS	77	2.308	2.28	0.028	20	52.5	20	=
HB3	CYS	77	2.364	2.378	-0.014	20	41.7	30	=
HG	CYS	77	0.728	0.723	0.005	20	83.6	85	strong=
C	CYS	77	171.952	171.952	0	20	90	90	strong=
N	ARG	78	129.634	129.642	-0.008	20	90	90	strong=
H	ARG	78	8.982	8.984	-0.002	20	88.9	90	strong=
CA	ARG	78	54.229	54.201	0.028	20	88.4	90	strong=
HA	ARG	78	5.119	5.111	0.008	20	85.4	90	strong=
CB	ARG	78	31.794	31.813	-0.019	20	90.2	90	strong=
HB2	ARG	78	0.387	0.384	0.003	20	90	90	strong=
HB3	ARG	78	1.384	1.38	0.004	20	90	90	strong=
CG	ARG	78	27.102	27.178	-0.076	20	94.1	90	strong=
HG2	ARG	78	0.859	0.856	0.003	20	63.1	60	=
HG3	ARG	78	1.128	1.121	0.007	20	89.9	90	strong=
CD	ARG	78	43.217	43.178	0.039	20	99.2	100	strong=
HD2	ARG	78	2.717	2.689	0.028	20	49.7	30	=
HD3	ARG	78	2.872	2.868	0.004	20	89.9	90	strong=
NE	ARG	78	117.37						
HE	ARG	78	6.829	5.122	1.707	20	30	0	
C	ARG	78	175.729	175.731	-0.002	20	90	90	strong=
N	VAL	79	124.277	124.288	-0.011	20	90	90	strong=
H	VAL	79	8.476	8.479	-0.003	20	89.4	90	strong=
CA	VAL	79	60.21	60.198	0.012	20	89.5	90	strong=
HA	VAL	79	4.465	4.462	0.003	20	89.8	90	strong=
CB	VAL	79	34.119	34.104	0.015	20	94.2	90	strong=
HB	VAL	79	2.142	2.136	0.006	20	84.8	85	strong=
QG1	VAL	79	1.047	1.043	0.004	20	76.9	80	=
QG2	VAL	79	1.049	1.067	-0.018	20	87.6	85	strong=
CG1	VAL	79	20.39	20.401	-0.011	20	70	25	! (CG2)
CG2	VAL	79	21.769	21.725	0.044	20	82.8	85	strong=
C	VAL	79	173.814	173.795	0.019	20	92.7	90	strong=
N	GLN	80	125.247	125.235	0.012	20	90	90	strong=
H	GLN	80	9.556	9.559	-0.003	20	90	90	strong=
CA	GLN	80	57.556	57.601	-0.045	20	100	100	strong=
HA	GLN	80		3.777		20	99.8	0	strong
CB	GLN	80	26.833	26.964	-0.131	20	69.1	70	=
HB2	GLN	80		2.273		20	60	0	
HB3	GLN	80		2.31		20	96.2	0	strong
CG	GLN	80		34.165		20	96.6	0	strong
HG2	GLN	80		2.307		20	99.1	0	strong

HG3	GLN	80		2.311		20	54	0	
NE2	GLN	80	111.264	110.734	0.53	20	70	70	=
HE21	GLN	80	6.804	6.91	-0.106	20	82.9	70	strong=
HE22	GLN	80	7.491	6.911	0.58	20	58.1	35	! (HE21)
C	GLN	80	174.99	175.026	-0.036	20	90	90	strong=
N	HIS	81	116.31	116.313	-0.003	20	100	100	strong=
H	HIS	81	8.584	8.579	0.005	20	99.9	100	strong=
CA	HIS	81	56.471	56.461	0.01	20	99.9	100	strong=
HA	HIS	81	4.527	4.527	0	20	94.9	95	strong=
CB	HIS	81	28.884	28.951	-0.067	20	99.5	100	strong=
HB2	HIS	81	3.353	3.346	0.007	20	99.9	100	strong=
HB3	HIS	81	3.349	3.348	0.001	20	99.9	100	strong=
CD2	HIS	81	118.411	120.261	-1.85	20	64.7	25	!
HD1	HIS	81		1.032		20	99.4	0	strong
CE1	HIS	81	137.553	137.924	-0.371	20	43.2	40	=
HD2	HIS	81	7.114	7.206	-0.092	20	69	5	!
HE1	HIS	81	8.267	8.226	0.041	20	43.9	5	!
C	HIS	81	174.049	174.145	-0.096	20	95	95	strong=
N	LEU	82	123.254	123.254	0	20	100	100	strong=
H	LEU	82	8.455	8.456	-0.001	20	99.9	100	strong=
CA	LEU	82	54.273	54.191	0.082	20	99.7	100	strong=
HA	LEU	82	4.424	4.415	0.009	20	100	100	strong=
CB	LEU	82	44.503	44.503	0	20	100	100	strong=
HB2	LEU	82	1.015	1.014	0.001	20	99.8	100	strong=
HB3	LEU	82	1.996	1.996	0	20	95	95	strong=
CG	LEU	82	26.735	26.849	-0.114	20	89.6	90	strong=
HG	LEU	82	1.41	1.412	-0.002	20	89.8	90	strong=
QD1	LEU	82	0.751	0.746	0.005	20	100	100	strong=
QD2	LEU	82	1.024	1.022	0.002	20	84.9	85	strong=
CD1	LEU	82	22.778	22.792	-0.014	20	99.9	100	strong=
CD2	LEU	82	25.857	25.774	0.083	20	79.9	80	=
C	LEU	82	174.896	174.959	-0.063	20	99.1	100	strong=
N	TRP	83	120.423	120.428	-0.005	20	100	100	strong=
H	TRP	83	7.912	7.914	-0.002	20	100	100	strong=
CA	TRP	83	56.428	56.421	0.007	20	100	100	strong=
HA	TRP	83	5.046	5.045	0.001	20	99.8	100	strong=
CB	TRP	83	32.075	32.088	-0.013	20	99.8	100	strong=
HB2	TRP	83	2.82	2.829	-0.009	20	95.6	100	strong=
HB3	TRP	83	2.887	2.883	0.004	20	92.6	95	strong=
CD1	TRP	83	126.973	126.99	-0.017	20	100	100	strong=
CE3	TRP	83	120.703	126.976	-6.273	20	50	0	

NE1	TRP	83	128.398	128.403	-0.005	20	95	95	strong=
HD1	TRP	83	7.036	7.032	0.004	20	99.9	100	strong=
HE3	TRP	83	7.3	7.033	0.267	20	49.9	0	
CZ3	TRP	83	121.767	120.332	1.435	20	33.7	15	
CZ2	TRP	83	114.067	114.528	-0.461	19	94.2	21.1	strong!
HE1	TRP	83	9.926	9.93	-0.004	20	94.9	95	strong=
HZ3	TRP	83	6.84	6.799	0.041	20	20	5	!
CH2	TRP	83	124.172	120.34	3.832	18	33.2	0	
HZ2	TRP	83	7.342	7.349	-0.007	20	90	90	strong=
HH2	TRP	83	7.12	7.192	-0.072	20	16.6	0	!
C	TRP	83	175.061	175.061	0	20	81.3	20	strong!
N	PHE	84	120.071	120.071	0	20	100	100	strong=
H	PHE	84	8.99	8.994	-0.004	20	99.2	100	strong=
CA	PHE	84	56.304	56.365	-0.061	20	98	100	strong=
HA	PHE	84	4.557	4.552	0.005	20	99	100	strong=
CB	PHE	84	44.076	44.077	-0.001	20	99.9	100	strong=
HB2	PHE	84	2.52	2.519	0.001	20	99.9	100	strong=
HB3	PHE	84	3.131	3.122	0.009	20	86.7	85	strong=
CD1	PHE	84	131.752	131.434	0.318	20	72.6	65	=
HD1	PHE	84	6.919	6.948	-0.029	20	99.7	70	strong=
CE1	PHE	84	130.332	130.663	-0.331	20	75.2	70	=
HE1	PHE	84	6.519	6.904	-0.385	20	63.5	30	
CZ	PHE	84	130.325	124.713	5.612	19	30.3	0	!
HZ	PHE	84	6.938	6.917	0.021	20	28.6	20	=
CE2	PHE	84	130.256	131.162	-0.906	20	56.2	5	!
HE2	PHE	84	6.518	6.937	-0.419	20	92.5	0	strong!
CD2	PHE	84	131.772	131.234	0.538	20	80	45	!
HD2	PHE	84	6.921	6.949	-0.028	20	94.8	95	strong=
C	PHE	84	175.734	175.71	0.024	20	86.5	85	strong=
N	GLN	85	120.816	120.816	0	20	100	100	strong=
H	GLN	85	9.561	9.557	0.004	20	100	100	strong=
CA	GLN	85	58.074	58.1	-0.026	20	100	100	strong=
HA	GLN	85	3.886	3.882	0.004	20	100	100	strong=
CB	GLN	85	28.621	28.603	0.018	20	99.9	100	strong=
HB2	GLN	85	2.201	2.195	0.006	20	100	100	strong=
HB3	GLN	85	2.202	2.195	0.007	20	100	100	strong=
CG	GLN	85	33.778	33.763	0.015	20	100	100	strong=
HG2	GLN	85	2.522	2.518	0.004	20	99.7	100	strong=
HG3	GLN	85	2.521	2.52	0.001	20	99.9	100	strong=
NE2	GLN	85	112.281	112.294	-0.013	20	99.4	100	strong=
HE21	GLN	85	6.959	6.952	0.007	20	98.6	100	strong=
HE22	GLN	85	7.617	7.618	-0.001	20	100	100	strong=

C	GLN	85	174.22	174.22	0	20	100	100	strong=
N	SER	86	105.494	105.49	0.004	20	100	100	strong=
H	SER	86	7.392	7.386	0.006	20	100	100	strong=
CA	SER	86	56.535	56.51	0.025	20	100	100	strong=
HA	SER	86	5.125	5.125	0	20	99.9	100	strong=
CB	SER	86	66.881	66.99	-0.109	20	100	100	strong=
HB2	SER	86	4.198	4.173	0.025	20	91.8	20	strong=
HB3	SER	86	4.206	4.208	-0.002	20	84.9	85	strong=
C	SER	86	175.236	175.236	0	20	100	100	strong=
N	ILE	87	122.052	122.051	0.001	20	100	100	strong=
H	ILE	87	9.64	9.638	0.002	20	100	100	strong=
CA	ILE	87	63.964	63.955	0.009	20	100	100	strong=
HA	ILE	87	3.8	3.791	0.009	20	99.8	100	strong=
CB	ILE	87	37.741	37.743	-0.002	20	100	100	strong=
HB	ILE	87	1.479	1.476	0.003	20	99.9	100	strong=
QG2	ILE	87	0.171	0.163	0.008	20	100	100	strong=
CG2	ILE	87	15.664	15.655	0.009	20	100	100	strong=
CG1	ILE	87	29.351	29.383	-0.032	20	100	100	strong=
HG12	ILE	87	0.805	0.804	0.001	20	99.7	100	strong=
HG13	ILE	87	1.391	1.383	0.008	20	100	100	strong=
QD1	ILE	87	0.569	0.562	0.007	20	100	100	strong=
CD1	ILE	87	14.435	14.422	0.013	20	100	100	strong=
C	ILE	87	175.219	175.219	0	20	100	100	strong=
N	PHE	88	121.239	121.235	0.004	20	100	100	strong=
H	PHE	88	6.812	6.82	-0.008	20	99.9	100	strong=
CA	PHE	88	60.034	60.024	0.01	20	100	100	strong=
HA	PHE	88	3.838	3.837	0.001	20	99.2	100	strong=
CB	PHE	88	37.896	37.884	0.012	20	99.9	100	strong=
HB2	PHE	88	2.839	2.833	0.006	20	98.8	100	strong=
HB3	PHE	88	3.308	3.305	0.003	20	100	100	strong=
CD1	PHE	88	131.02	131.574	-0.554	20	81.7	15	strong!
HD1	PHE	88	7.202	7.223	-0.021	20	88	85	strong=
CE1	PHE	88	132.343	131.192	1.151	20	80.3	5	strong!
HE1	PHE	88	7.035	7.23	-0.195	20	59.5	25	! (HD1)
CZ	PHE	88	132.054	131.347	0.707	20	36.8	20	! (CD2)
HZ	PHE	88	7.02	7.233	-0.213	20	39.2	0	!
CE2	PHE	88	132.34	131.387	0.953	20	50.4	25	! (CD2)
HE2	PHE	88	7.032	7.853	-0.821	20	44.7	0	
CD2	PHE	88	131.032	131.099	-0.067	20	90.7	80	strong=
HD2	PHE	88	7.201	7.23	-0.029	20	91.9	50	strong=
C	PHE	88	178.104	178.104	0	20	100	100	strong=
N	ASP	89	119.062	119.066	-0.004	20	100	100	strong=

H	ASP	89	7.829	7.825	0.004	20	99.5	100	strong=
CA	ASP	89	57.104	57.097	0.007	20	100	100	strong=
HA	ASP	89	4.314	4.314	0	20	100	100	strong=
CB	ASP	89	41.498	41.514	-0.016	20	99.7	100	strong=
HB2	ASP	89	2.727	2.731	-0.004	20	96.9	100	strong=
HB3	ASP	89	2.895	2.891	0.004	20	99.9	100	strong=
C	ASP	89	178.103	178.103	0	20	100	100	strong=
N	MET	90	121.617	121.625	-0.008	20	100	100	strong=
H	MET	90	7	7.007	-0.007	20	99.9	100	strong=
CA	MET	90	58.737	58.732	0.005	20	99.7	100	strong=
HA	MET	90	2.002	1.998	0.004	20	79.9	80	=
CB	MET	90	30.402	30.407	-0.005	20	98.3	100	strong=
HB2	MET	90	1.115	1.114	0.001	20	99.1	100	strong=
HB3	MET	90	1.119	1.118	0.001	20	59.4	40	!
CG	MET	90	31.932	31.924	0.008	20	75	75	=
HG2	MET	90	1.376	1.377	-0.001	20	74.6	75	=
HG3	MET	90	1.873	1.87	0.003	20	60	60	=
QE	MET	90		1.678		20	89.9	0	strong
CE	MET	90		17.551		20	89.9	0	strong
C	MET	90	176.119	176.119	0	20	100	100	strong=
N	LEU	91	117.926	117.924	0.002	20	100	100	strong=
H	LEU	91	7.583	7.583	0	20	99.9	100	strong=
CA	LEU	91	57.606	57.611	-0.005	20	100	100	strong=
HA	LEU	91	3.385	3.381	0.004	20	100	100	strong=
CB	LEU	91	41.18	41.173	0.007	20	100	100	strong=
HB2	LEU	91	0.822	0.819	0.003	20	100	100	strong=
HB3	LEU	91	1.487	1.487	0	20	94.8	95	strong=
CG	LEU	91	25.856	25.865	-0.009	20	100	100	strong=
HG	LEU	91	1.201	1.198	0.003	20	99.7	100	strong=
QD1	LEU	91	-0.07	-0.078	0.008	20	99.8	100	strong=
QD2	LEU	91	-0.533	-0.537	0.004	20	100	100	strong=
CD1	LEU	91	22.577	22.537	0.04	20	100	100	strong=
CD2	LEU	91	23.982	23.985	-0.003	20	99.9	100	strong=
C	LEU	91	179.323	179.288	0.035	20	99.6	100	strong=
N	GLU	92	116.207	116.198	0.009	20	100	100	strong=
H	GLU	92	7.478	7.478	0	20	100	100	strong=
CA	GLU	92	58.456	58.464	-0.008	20	100	100	strong=
HA	GLU	92	3.987	3.982	0.005	20	100	100	strong=
CB	GLU	92	29.449	29.533	-0.084	20	99.8	100	strong=
HB2	GLU	92	1.998	1.998	0	20	99.9	100	strong=
HB3	GLU	92	2.001	1.999	0.002	20	100	100	strong=
CG	GLU	92	35.81	35.851	-0.041	20	100	100	strong=

HG2	GLU	92	2.265	2.264	0.001	20	100	100	strong=
HG3	GLU	92	2.268	2.264	0.004	20	100	100	strong=
C	GLU	92	179.588	179.621	-0.033	20	99.6	100	strong=
N	HIS	93	120.577	120.588	-0.011	20	100	100	strong=
H	HIS	93	7.949	7.945	0.004	20	100	100	strong=
CA	HIS	93	60.689	60.684	0.005	20	99.9	100	strong=
HA	HIS	93	4.254	4.253	0.001	20	98.8	100	strong=
CB	HIS	93	30.356	30.347	0.009	20	99.8	100	strong=
HB2	HIS	93	2.714	2.71	0.004	20	94.7	95	strong=
HB3	HIS	93	2.967	2.963	0.004	20	90	90	strong=
ND1	HIS	93		112.292		14	57	0	
CD2	HIS	93	120.151	120.399	-0.248	20	84.9	85	strong=
HD1	HIS	93		7.497		20	39.9	0	
CE1	HIS	93	137.548	131.734	5.814	20	36	0	!
HD2	HIS	93	5.788	5.805	-0.017	20	84.4	85	strong=
HE1	HIS	93	8.239	7.448	0.791	20	39.9	25	!
C	HIS	93	178.745	178.744	0.001	20	100	100	strong=
N	PHE	94	115.655	115.654	0.001	19	100	100	strong=
H	PHE	94	7.707	7.704	0.003	20	94.9	95	strong=
CA	PHE	94	57.255	57.212	0.043	20	100	100	strong=
HA	PHE	94	5.68	5.68	0	20	99.9	100	strong=
CB	PHE	94	37.666	37.679	-0.013	20	94.9	95	strong=
HB2	PHE	94	2.488	2.482	0.006	20	87	85	strong=
HB3	PHE	94	3.351	3.342	0.009	20	96.6	100	strong=
CD1	PHE	94	131.686	131.596	0.09	20	65	65	=
HD1	PHE	94	7.173	7.215	-0.042	20	54.8	0	!
CE1	PHE	94	130.68	131.433	-0.753	20	57.3	25	!
HE1	PHE	94	7.049	7.219	-0.17	20	55	0	!
CZ	PHE	94	128.141	125.521	2.62	19	53.4	0	!
HZ	PHE	94	7.071	5.67	1.401	20	49.6	0	!
CE2	PHE	94	130.641	131.449	-0.808	20	45.5	15	!
HE2	PHE	94	7.059	7.218	-0.159	20	68.2	0	!
CD2	PHE	94	131.793	131.431	0.362	20	56.8	55	=
HD2	PHE	94	7.163	7.219	-0.056	20	99.3	0	strong!
C	PHE	94	174.873	174.811	0.062	20	100	100	strong=
N	ARG	95	117.386	117.388	-0.002	20	100	100	strong=
H	ARG	95	7.315	7.308	0.007	20	99.9	100	strong=
CA	ARG	95	57.973	58.034	-0.061	20	88.3	90	strong=
HA	ARG	95	4.58	4.593	-0.013	20	82.9	80	strong=
CB	ARG	95	30.889	30.991	-0.102	20	90.1	90	strong=
HB2	ARG	95	1.939	1.925	0.014	20	85.3	90	strong=
HB3	ARG	95	1.943	2.012	-0.069	20	55.3	35	!

CG	ARG	95	27.548	27.601	-0.053	20	92.1	90	strong=
HG2	ARG	95	1.646	1.629	0.017	20	84.7	85	strong=
HG3	ARG	95	1.941	1.939	0.002	20	64.7	65	=
CD	ARG	95	43.365	43.344	0.021	20	97.3	100	strong=
HD2	ARG	95	3.233	3.231	0.002	20	89.6	90	strong=
HD3	ARG	95	3.233	3.234	-0.001	20	70	65	=
NE	ARG	95	120.533						
HE	ARG	95	7.283	7.349	-0.066	20	37	0	
C	ARG	95	177.299	177.292	0.007	20	90	90	strong=
N	VAL	96	111.78	111.777	0.003	20	90	90	strong=
H	VAL	96	6.972	6.968	0.004	20	89.9	90	strong=
CA	VAL	96	61.919	61.991	-0.072	20	90	90	strong=
HA	VAL	96	4.123	4.118	0.005	20	90	90	strong=
CB	VAL	96	34.521	34.553	-0.032	20	95.3	100	strong=
HB	VAL	96	1.827	1.826	0.001	20	90	90	strong=
QG1	VAL	96	0.758	0.753	0.005	20	90	90	strong=
QG2	VAL	96	0.57	0.569	0.001	20	89.5	90	strong=
CG1	VAL	96	20.278	20.287	-0.009	20	94.9	95	strong=
CG2	VAL	96	20.865	20.88	-0.015	20	90.2	90	strong=
C	VAL	96	174.314	174.335	-0.021	20	92.6	90	strong=
N	HIS	97	121.544	121.545	-0.001	20	90	90	strong=
H	HIS	97	8.261	8.266	-0.005	20	89.7	90	strong=
CA	HIS	97	52.497	52.505	-0.008	20	89.7	90	strong=
HA	HIS	97		4.837		20	86.8	0	strong
CB	HIS	97	29.175	29.253	-0.078	20	84.7	85	strong=
HB2	HIS	97		2.307		20	84.7	0	strong
HB3	HIS	97		3.103		20	88.5	0	strong
ND1	HIS	97		110.981		6	33.3	0	
CD2	HIS	97	121.218	120.227	0.991	20	70	5	!
HD1	HIS	97		2.301		20	49.7	0	
CE1	HIS	97	138.417	137.297	1.12	19	90	0	strong
HD2	HIS	97	6.777	7.119	-0.342	20	69.7	5	!
HE1	HIS	97	7.47	8.296	-0.826	20	84.9	0	strong
C	HIS	97	170.854	170.854	0	20	90	90	strong=
CA	PRO	98	61.95	61.951	-0.001	20	100	100	strong=
HA	PRO	98	4.541	4.539	0.002	20	99.9	100	strong=
CB	PRO	98	32.789	32.894	-0.105	20	95.1	100	strong=
HB2	PRO	98	1.869	1.869	0	20	55.1	15	!
HB3	PRO	98	2.066	2.055	0.011	20	99.6	100	strong=
CG	PRO	98	27.648	27.642	0.006	20	99.5	100	strong=
HG2	PRO	98	1.925	1.939	-0.014	20	75	70	=
HG3	PRO	98	2.061	2.057	0.004	20	89.1	90	strong=

CD	PRO	98	50.624	50.643	-0.019	20	99	100	strong=
HD2	PRO	98	3.74	3.739	0.001	20	99.4	100	strong=
HD3	PRO	98	3.742	3.741	0.001	20	94.8	95	strong=
C	PRO	98	177.429	177.429	0	20	100	100	strong=
N	ILE	99	123.149	123.147	0.002	20	100	100	strong=
H	ILE	99	8.127	8.127	0	20	99.6	100	strong=
CA	ILE	99	59.382	59.355	0.027	20	100	100	strong=
HA	ILE	99		3.598		20	99.9	0	strong
CB	ILE	99	39.344	39.343	0.001	20	100	100	strong=
HB	ILE	99		1.132		20	100	0	strong
QG2	ILE	99		0.235		20	95	0	strong
CG2	ILE	99		16.379		20	95	0	strong
CG1	ILE	99		27.26		20	99.8	0	strong
HG12	ILE	99		-0.367		20	99.2	0	strong
HG13	ILE	99		1.128		20	100	0	strong
QD1	ILE	99		-0.376		20	90	0	strong
CD1	ILE	99		12.24		20	90	0	strong
C	ILE	99	175.249	175.249	0	20	100	100	strong=
CA	PRO	100	62.681	62.668	0.013	20	100	100	strong=
HA	PRO	100	4.688	4.68	0.008	20	99.9	100	strong=
CB	PRO	100	31.21	30.964	0.246	20	84.1	85	strong=
HB2	PRO	100	2.134	2.14	-0.006	20	86.4	90	strong=
HB3	PRO	100	2.156	2.154	0.002	20	67.7	65	=
CG	PRO	100	27.317	27.361	-0.044	20	100	100	strong=
HG2	PRO	100	2.148	2.135	0.013	20	99.9	100	strong=
HG3	PRO	100	2.143	2.136	0.007	20	99.7	100	strong=
CD	PRO	100	51.141	51.209	-0.068	20	100	100	strong=
HD2	PRO	100	3.314	3.31	0.004	20	100	100	strong=
HD3	PRO	100	4.044	4.038	0.006	20	100	100	strong=
C	PRO	100	175.903	175.891	0.012	20	100	100	strong=
N	LEU	101	122.437	122.438	-0.001	20	100	100	strong=
H	LEU	101	8.188	8.191	-0.003	20	99.6	100	strong=
CA	LEU	101	54.113	54.12	-0.007	20	100	100	strong=
HA	LEU	101	4.435	4.432	0.003	20	99.8	100	strong=
CB	LEU	101	43.389	43.359	0.03	20	100	100	strong=
HB2	LEU	101	1.507	1.509	-0.002	20	87.9	90	strong=
HB3	LEU	101	1.515	1.517	-0.002	20	99.9	100	strong=
CG	LEU	101	26.818	26.922	-0.104	20	65	65	=
HG	LEU	101	1.513	1.506	0.007	20	65	65	=
QD1	LEU	101	0.701	0.705	-0.004	20	74.9	75	=
QD2	LEU	101	0.649	0.638	0.011	20	74.9	75	=
CD1	LEU	101	23.208	23.226	-0.018	20	97.9	100	strong=

CD2	LEU	101	25.603	25.552	0.051	20	69.8	70	=
C	LEU	101	178.273	178.289	-0.016	20	100	100	strong=
N	GLU	102	122	121.993	0.007	20	99.9	100	strong=
H	GLU	102	8.973	8.97	0.003	20	99.9	100	strong=
CA	GLU	102	58.248	58.241	0.007	20	100	100	strong=
HA	GLU	102	4.173	4.172	0.001	20	99.6	100	strong=
CB	GLU	102	29.417	29.384	0.033	20	99.8	100	strong=
HB2	GLU	102	2.096	2.083	0.013	20	95.2	90	strong=
HB3	GLU	102	2.089	2.092	-0.003	20	79.5	80	=
CG	GLU	102	36.504	36.568	-0.064	20	99.9	100	strong=
HG2	GLU	102	2.347	2.345	0.002	20	55	55	=
HG3	GLU	102	2.35	2.345	0.005	20	100	100	strong=
C	GLU	102	176.947	176.947	0	20	100	100	strong=
N	SER	103	112.734	112.737	-0.003	20	100	100	strong=
H	SER	103	8.02	8.024	-0.004	20	99.9	100	strong=
CA	SER	103	58.318	58.325	-0.007	20	99.9	100	strong=
HA	SER	103	4.434	4.432	0.002	20	99.5	100	strong=
CB	SER	103	63.649	63.576	0.073	20	99.8	100	strong=
HB2	SER	103	3.931	3.94	-0.009	20	81.5	80	strong=
HB3	SER	103	4.006	4.001	0.005	20	55	55	=
C	SER	103	175.159	175.156	0.003	20	100	100	strong=
N	GLY	104	110.75	110.753	-0.003	20	100	100	strong=
H	GLY	104	8.232	8.237	-0.005	20	98.3	100	strong=
CA	GLY	104	45.319	45.312	0.007	20	99.9	100	strong=
HA2	GLY	104	3.87	3.867	0.003	20	99.7	100	strong=
HA3	GLY	104	4.258	4.255	0.003	20	74.9	75	=
C	GLY	104	174.583	174.591	-0.008	20	99.4	100	strong=
N	GLY	105	108.692	108.695	-0.003	20	100	100	strong=
H	GLY	105	8.242	8.239	0.003	20	99	100	strong=
CA	GLY	105	45.084	45.11	-0.026	20	98.3	100	strong=
HA2	GLY	105	4.068	4.063	0.005	20	59.5	60	=
HA3	GLY	105	4.056	4.075	-0.019	20	77.8	80	=
C	GLY	105	173.718	173.718	0	20	100	100	strong=
N	SER	106	115.027	115.023	0.004	20	100	100	strong=
H	SER	106	8.479	8.475	0.004	20	99.8	100	strong=
CA	SER	106	58.443	58.424	0.019	20	100	100	strong=
HA	SER	106	4.593	4.589	0.004	20	100	100	strong=
CB	SER	106	64.166	64.257	-0.091	20	99.9	100	strong=
HB2	SER	106	3.915	3.919	-0.004	20	99.7	100	strong=
HB3	SER	106	3.923	3.923	0	20	99.6	100	strong=
C	SER	106	174.772	174.772	0	20	100	100	strong=
N	SER	107	118.311	118.3	0.011	20	100	100	strong=

H	SER	107	8.204	8.207	-0.003	20	100	100	strong=
CA	SER	107	58.436	58.389	0.047	20	100	100	strong=
HA	SER	107	4.596	4.596	0	20	99.8	100	strong=
CB	SER	107	64.614	64.797	-0.183	20	98	100	strong=
HB2	SER	107	3.927	3.795	0.132	20	87.7	0	strong
HB3	SER	107	3.812	3.818	-0.006	20	93.3	100	strong=
C	SER	107	173.408	173.408	0	20	95	95	strong=
N	ASP	108	119.85	119.852	-0.002	20	100	100	strong=
H	ASP	108	8.381	8.384	-0.003	20	99.7	100	strong=
CA	ASP	108	54.158	54.161	-0.003	20	99.7	100	strong=
HA	ASP	108	4.755	4.748	0.007	20	99.8	100	strong=
CB	ASP	108	41.47	41.525	-0.055	20	100	100	strong=
HB2	ASP	108	2.776	2.776	0	20	100	100	strong=
HB3	ASP	108	2.777	2.776	0.001	20	100	100	strong=
C	ASP	108	175.726	175.718	0.008	20	99.1	100	strong=
N	VAL	109	121.257	121.255	0.002	20	100	100	strong=
H	VAL	109	8.393	8.395	-0.002	20	99.5	100	strong=
CA	VAL	109	61.699	61.692	0.007	20	99.8	100	strong=
HA	VAL	109	4.479	4.472	0.007	20	95.2	100	strong=
CB	VAL	109	33.953	33.962	-0.009	20	100	100	strong=
HB	VAL	109	2.149	2.147	0.002	20	99.8	100	strong=
QG1	VAL	109	0.897	0.888	0.009	20	90.6	90	strong=
QG2	VAL	109	0.781	0.79	-0.009	20	99.3	100	strong=
CG1	VAL	109	20.618	20.568	0.05	20	70.1	70	=
CG2	VAL	109	21.484	21.588	-0.104	20	99.6	100	strong=
C	VAL	109	172.41	172.41	0	20	100	100	strong=
N	VAL	110	118.833	118.829	0.004	20	95	95	strong=
H	VAL	110	7.352	7.354	-0.002	20	94.8	95	strong=
CA	VAL	110	58.611	58.637	-0.026	20	99.6	100	strong=
HA	VAL	110	4.564	4.55	0.014	20	95.9	95	strong=
CB	VAL	110	34.034	34.028	0.006	20	100	100	strong=
HB	VAL	110	2.034	2.029	0.005	20	99.7	100	strong=
QG1	VAL	110	0.661	0.653	0.008	20	99.6	100	strong=
QG2	VAL	110	0.469	0.466	0.003	20	99.3	100	strong=
CG1	VAL	110	18.33	18.33	0	20	99.9	100	strong=
CG2	VAL	110	21.297	21.263	0.034	20	100	100	strong=
C	VAL	110	175.555	175.553	0.002	20	99.9	100	strong=
N	LEU	111	118.754	118.803	-0.049	20	99.9	100	strong=
H	LEU	111	8.225	8.233	-0.008	20	99.9	100	strong=
CA	LEU	111	54.089	54.001	0.088	20	100	100	strong=
HA	LEU	111	4.527	4.493	0.034	20	99.5	5	strong
CB	LEU	111	40.054	40.166	-0.112	20	94.9	95	strong=

HB2	LEU	111	1.001	1.009	-0.008	20	93.4	95	strong=
HB3	LEU	111	1.17	1.171	-0.001	20	97.6	100	strong=
CG	LEU	111	25.736	25.638	0.098	20	96.5	100	strong=
HG	LEU	111	1.148	1.147	0.001	20	65	30	! (QD2)
QD1	LEU	111	0.072	0.071	0.001	20	95	95	strong=
QD2	LEU	111	-0.05	-0.046	-0.004	20	85	85	strong=
CD1	LEU	111	20.953	20.955	-0.002	20	100	100	strong=
CD2	LEU	111	25.625	25.583	0.042	20	88.2	85	strong=
C	LEU	111	177.702	177.75	-0.048	20	100	100	strong=
N	VAL	112	122.307	122.304	0.003	20	99.9	100	strong=
H	VAL	112	8.96	8.962	-0.002	20	100	100	strong=
CA	VAL	112	63.928	63.914	0.014	20	100	100	strong=
HA	VAL	112	4.215	4.211	0.004	20	100	100	strong=
CB	VAL	112	34.063	34.054	0.009	20	100	100	strong=
HB	VAL	112	1.961	1.96	0.001	20	100	100	strong=
QG1	VAL	112	1.097	1.094	0.003	20	96.8	95	strong=
QG2	VAL	112	1.063	1.045	0.018	20	87.9	85	strong=
CG1	VAL	112	21.375	21.391	-0.016	20	99.2	100	strong=
CG2	VAL	112	21.519	21.587	-0.068	20	97.3	100	strong=
C	VAL	112	175.977	175.978	-0.001	20	100	100	strong=
N	SER	113	112.372	112.369	0.003	20	100	100	strong=
H	SER	113	7.566	7.568	-0.002	20	99.8	100	strong=
CA	SER	113	57.299	57.321	-0.022	20	100	100	strong=
HA	SER	113	4.682	4.679	0.003	20	100	100	strong=
CB	SER	113	64.38	64.39	-0.01	20	100	100	strong=
HB2	SER	113	3.832	3.828	0.004	20	100	100	strong=
HB3	SER	113	3.832	3.828	0.004	20	100	100	strong=
C	SER	113	171.01	171.01	0	20	100	100	strong=
N	TYR	114	114.116	114.144	-0.028	20	100	100	strong=
H	TYR	114	7.089	7.09	-0.001	20	100	100	strong=
CA	TYR	114	53.452	53.437	0.015	20	100	100	strong=
HA	TYR	114	5.39	5.388	0.002	20	100	100	strong=
CB	TYR	114	40.274	40.275	-0.001	20	100	100	strong=
HB2	TYR	114	1.869	1.868	0.001	20	100	100	strong=
HB3	TYR	114	2.611	2.611	0	20	99.6	100	strong=
CD1	TYR	114	134.004	134.239	-0.235	20	90	90	strong=
HD1	TYR	114	6.826	6.859	-0.033	20	99.9	0	strong!
CE1	TYR	114	118.312	120.268	-1.956	20	50.9	30	!
HE1	TYR	114	6.9	6.856	0.044	20	84.8	0	strong!
CE2	TYR	114	118.346	118.578	-0.232	20	94.4	95	strong=
HE2	TYR	114	6.9	6.868	0.032	20	83.2	5	strong!
CD2	TYR	114	134.002	134.26	-0.258	20	99.9	100	strong=

HD2	TYR	114	6.826	6.863	-0.037	20	84.9	0	strong!
HH	TYR	114		8.303		20	99.7	0	strong
C	TYR	114	176.586	176.586	0	20	100	100	strong=
N	VAL	115	122.64	122.676	-0.036	20	100	100	strong=
H	VAL	115	8.043	8.05	-0.007	20	100	100	strong=
CA	VAL	115	59.423	59.483	-0.06	20	99.7	100	strong=
HA	VAL	115		4.474		20	99.9	0	strong
CB	VAL	115	33.259	33.188	0.071	20	99.5	100	strong=
HB	VAL	115		1.953		20	100	0	strong
QG1	VAL	115		0.848		20	99.2	0	strong
QG2	VAL	115		0.863		20	50	0	
CG1	VAL	115		21.448		20	95.6	0	strong
CG2	VAL	115		21.554		20	60	0	
C	VAL	115	174.916	174.916	0	20	100	100	strong=
CA	PRO	116	62.562	62.562	0	20	100	100	strong=
HA	PRO	116	4.992	4.979	0.013	20	94.6	95	strong=
CB	PRO	116	32.4	32.4	0	20	99.1	100	strong=
HB2	PRO	116	2.191	2.19	0.001	20	80	80	strong=
HB3	PRO	116	2.486	2.479	0.007	20	99.6	100	strong=
CG	PRO	116	27.55	27.55	0	20	83	5	strong!
HG2	PRO	116	2.053	2.053	0	20	83.2	5	strong!
HG3	PRO	116	2.228	2.228	0	20	82.2	90	strong=
CD	PRO	116	51.196	51.176	0.02	20	100	100	strong=
HD2	PRO	116	3.87	3.868	0.002	20	99.3	100	strong=
HD3	PRO	116	4.181	4.177	0.004	20	99.9	100	strong=
C	PRO	116	177.276	177.276	0	20	90	90	strong=
N	SER	117	117.71	117.713	-0.003	20	100	100	strong=
H	SER	117	8.969	8.965	0.004	20	100	100	strong=
CA	SER	117	57.961	57.975	-0.014	20	100	100	strong=
HA	SER	117	3.695	3.693	0.002	20	99.9	100	strong=
CB	SER	117	63.534	63.543	-0.009	20	100	100	strong=
HB2	SER	117	2.668	2.662	0.006	20	75	75	=
HB3	SER	117	3.353	3.35	0.003	20	95	95	strong=
C	SER	117	174.009	174.009	0	20	95	95	strong=
N	GLN	118	126.159	126.159	0	20	100	100	strong=
H	GLN	118	7.827	7.827	0	20	100	100	strong=
CA	GLN	118	57.145	57.178	-0.033	20	100	100	strong=
HA	GLN	118		4.231		20	99.9	0	strong
CB	GLN	118	30.699	30.723	-0.024	20	89.9	90	strong=
HB2	GLN	118		1.953		20	99.5	0	strong
HB3	GLN	118		2.156		20	70.7	0	
CG	GLN	118		33.945		20	99.1	0	strong

HG2	GLN	118		2.266		20	79.8	0	
HG3	GLN	118		2.281		20	97.1	0	strong
NE2	GLN	118		112.067		20	99.4	0	strong
HE21	GLN	118		6.813		20	98.8	0	strong
HE22	GLN	118		7.49		20	89.3	0	strong
C	GLN	118	175.731	175.748	0.017	20	100	0	strong

Appendix B

NMR chemical shift changes table of SH2Bc Y114F and SH2 upon binding to tethered pY139 and free C-terminal ligand.

Amino acid	SH2c-tethered pTyr139, in phosphate buffer	SH2c-tethered pTyr139, in Tris buffer	SH2-C terminal ligand In Tris buffer
D9	0.003855		0.0241246
Q10			0.0082366
P11			
L12	0.003456		0.0136179
S13	0.002904		0.00619448
G14	0.003949		0.0189743
Y15	0.006212	0.00546	0.00100975
P16			
W17			
F18	0.002903	0.0123693	0.0105385
H19	0.002736	0.0551601	0.0397949
G20			
M21	0.001186		0.0232053
L22	0.004867	0.183515	0.080105
S23	0.002904		0.0354255
R24	0.00927		0.335786
L25	0.002341	0.0143989	0.0221439
K26	0.002024	0.0149251	0.0035089
A27	0.001258	0.13862	0.0614591
A28	0.001592	0.0357385	0.0298548
Q29	0.002904	0.0912153	0.0485879
L30	0.003577	0.0181366	0.010001
V31	0.001317	0.0663113	0.0329885
L32	0.000825	0.00515632	0.00418134
E33	0.005562	0.00699626	0.0180255
G34	0.008638	0.116241	0.0135558
G35	0.000312	0.0051923	0.0185468
T36			0.0215392
G37	0.000838	0.0110569	0.00716704

S38	0.003317	0.0053917	0.0102793
H39	0.005786	0.008	0.0270725
G40	0.003942	0.0160739	0.0252
V41	0.00778	0.0280661	0.00823624
F42	0.0061		0.00722775
L43	0.002714	0.0336125	0.00379057
V44	0.003319	0.00392	0.00320225
R45	0.004553	0.0271751	0.0217703
Q46	0.003001	0.131479	0.0614393
S47	0.005946		0.10501
E48	0.01324		
T49	0.011554	0.13037	0.0297364
R50	0.005936		0.135134
R51	0.026512	0.244979	0.0789314
G52	0.002903	0.191388	0.0148464
E53	0.001252	0.183	0.0720348
Y54	0.006207	0.24759	0.134873
V55	0.000317	0.064952	0.0179945
L56	0.006388	0.0368114	0.025242
T57	0.002094	0.0251225	0.0127465
F58	0.00783	0.0340288	0.00600586
N59	0.001391	0.0141549	0.00418134
F60	0.005048	0.0151093	0.00271116
Q61	0.003652		
G62	0.001949	0.00261197	0.0180918
K63	0.003427	0.00471695	0.0104548
A64	0.001555	0.00600163	
K65	0.004952	0.019998	0.00284176
H66	0.008428	0.0182274	0.0029
L67	0.008716	0.0263735	0.0447914
R68	0.006495	0.0634909	0.0185158
L69	0.008034	0.119171	0.0452645
S70	0.006875	0.0502327	0.0260185
L71	0.002605	0.0393848	0.0189105
N72	0.005423	0.0501905	0.0112446
E73	0.002353	0.0244183	0.0210298
E74	0.003562	0.0107532	0.00364

G75	0.000406	0.0146573	0.00364005
Q76	0.002494	0.0121366	0.00719947
C77	0.002903		0.0112394
R78	0.001028	0.00610364	0.00694622
V79	0.004044	0.00294	
Q80	0.00315	0.0140994	0.0259044
H81	0.002903		0.0373602
L82	0.005657	0.0215761	0.0412629
W83	0.009804	0.0166433	0.0197449
F84	0.007247		0.0143856
Q85	0.002303	0.00643441	0.00445439
S86	0.003306	0.0181041	0.00440073
I87	0.006787	0.0204949	0.00507007
F88	0.011722	0.0919324	0.0190707
D89	0.009474		0.00716704
M90	0.004092	0.0398651	0.00423792
L91	0.001839	0.00270414	0.0146164
E92	0.008885	0.0310851	0.009801
H93	0.007754	0.0531489	0.0148464
F94	0.004633	0.0119289	0.00770042
R95	0.024814	0.0670577	0.0103031
V96	0.013006	0.0144891	0.0356714
H97	0.007841	0.020501	0.0232
P98			
I99	0.006159	0.0193197	0.00867696
P100			
L101	0.007104	0.0405317	0.0409067
E102		0.0193691	0.0775526
S103			0.0473868
G104	0.006723		0.024345
G105	0.002098	0.00730821	0.00400979
S106	0.008392		0.00918096
S107	0.002904		0.0067261
D108	0.002903	0.0244382	0.0223337
V109	0.002904	0.0137053	0.011136
V110	0.010598	0.0343779	0.0157119
L111	0.004974	0.0370471	

V112	0.024425		0.00480371
S113	0.019252		0.00759234
F114	0.069337		0.0051614
V115	0.018847		0.00582646
P116			
S117	0.007517	0.0147426	0.00379057
Q118	0.00038	0.00662118	0.00701781
R119	0.005506	0.0765013	
Q120	0.000478	0.0140442	
Q121	0.000374	0.0102765	
G122	0.004129	0.008	
R123	0.002598	0.0340883	
E124	0.002904	0.0208125	
Q125	0.002904	0.0172047	
A126	0.004071	0.0135558	
G127	0.002175	0.00415384	
S128	0.004435	0.009	
H129	0.007643		
A130	0.003019	0.0211004	
G131	0.003648	0.0218687	
V132	0.002054	0.0345941	
C133	0.007554	0.0624331	
E134	0.006639		
G135	0.004169		
D136	0.013416	0.286027	
R137	0.063773		
C138	0.002903	0.703888	
Y139	0.111171	0.454868	
P140			
D141	0.088712	0.0391751	
A142	0.052351		
S143	0.047172	0.0123373	
S144	0.013808		
T145	0.007455	0.00488262	
L146	0.017731	0.0390361	
L147	0.007587	0.0254233	
P148			

F149	0.024329	0.0517573	
G150	0.013475	0.048935	
A151	0.005832	0.0221838	
S152	0.019573	0.00602608	
D153	0.005106	0.0122151	
C154	0.004661	0.0265372	
V155	0.007344	0.0684055	
T156	0.004311	0.0335618	
E157	0.002412	0.0172047	
H158	0.007182	0.024318	
L159	0.005743	0.0549004	
P160			