THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN
NATURAL SCIENCE

Structural and Interaction Studies of the Human Protein Survivin

María-José García-Bonete

UNIVERSITY OF GOTHENBURG

Department of Chemistry and Molecular Biology
Gothenburg, 2019

Thesis for the degree of Doctor of Philosophy
in Natural Science

Structural and Interaction Studies of the Human Protein Survivin

María-José García-Bonete

Cover: Crystallography structure of the homodimer human survivin protein

If you never try, you'll never know

Coldplay

Abstract

Cell division and cell death (apoptosis) are two essential processes to maintain the specific number of cells in all multicellular organisms. In humans, the misregulation of these processes leads to severe diseases, such as cancer, and neurological, inflammatory or autoimmune diseases. Proteins are the most versatile macromolecules in all living organisms and are the orchestra directors of the majority of cellular processes. Their three-dimensional structure and the interaction with other molecules are essential for their correct biological function.

This work focus on small human protein survivin which plays an important role in cell division and apoptosis, and has been extensively reported in clinical research. Our aim was to discover new interaction partners of survivin, and to study their specific binding and structure to better understand its function. We successfully used microarray peptide technology to determine new possible interaction partners and microscale thermophoresis to confirm these interactions. The direct interaction between the shugoshin-like protein family and survivin has been reported and highlights its importance in cell division.

In addition, this thesis exhibits the powerful multivariate Bayesian inference approach for data analysis by focussing on addressing X-ray crystallography problems of experimental phasing for molecular structure determination. This approach has also been successfully applied to determine the binding curve and to calculate the interaction strength between two molecules, and avoids manual treatment and human subjective bias.

Swedish summary

Celldelning och programmerad celldöd är två viktiga processer för att bibehålla det specifika antalet celler i alla multicellulära organismer. I människan leder missreglering av proteiner kopplade till dessa processer till allvarliga sjukdomar så som cancer samt neurologiska, inflammatoriska eller autoimmuna sjukdomar. Proteiner är de mest mångsidiga makromolekylerna i alla levande organismer och de är dirigenter för de flesta cellulära processerna. Deras tredimensionella struktur och interaktioner med andra molekyler är avgörande för deras korrekta biologiska funktion.

Detta arbete fokuserar på det mänskliga proteinet survivin som spelar en viktig roll vid celldelning och programmerad celldöd vilket det i stor omfattning har rapporterats om i klinisk forskning. Vårat mål var att upptäcka nya interaktionspartners för survivin och att studera deras specifika bindning och struktur för att bättre förstå deras funktion. Vi har framgångsrikt använt tekniken mikromatriser för peptider för att bestämma nya möjliga interaktionspartners och termofores i mikroskala (MST) för att bekräfta dessa interaktioner. Direkt interaktion mellan proteiner ur shugoshin-liknande proteinfamiljen och survivin har rapporterats och framhäver dess betydelse vid celldelning.

Därutöver behandlar avhandlingen även datanalysmetoden Bayesiansk statistik med multivariata metoder för att lösa fasproblemet inom röntgenkristallografi vid strukturbestämning av proteiner. Metoden har framgångsrikt använts för att bestämma bindningskurvan och beräkna interaktionsstyrkan mellan två molekyler genom att undvika påverkan av manuella tillvägagångsätt samt mänskliga subjektiva bedömningar.

List of publications

This thesis is based on the following research publications:

Paper I   G. Katona, **M.J. Garcia-Bonete** and I. Lundholm. *Estimating the difference between structure-factor amplitudes using multivariate Bayesian inference*, Acta Cryst. A (2016) **A72**:406-411
doi.org/10.1107/S2053273316003430

Paper II   G. Gravina, C. Wasén, **M.J. Garcia-Bonete**, M. Turkkila, M.C. Erlandsson, S. Töyrä Silfverswärd, M. Brisslert, R. Pullerits, K.M. Andersson, G. Katona and M.I Bokarewa. *Survivin in autoimmune disease*, Autoimmunity Reviews (2017) **16**:845-855
doi.org/10.1016/j.autrev.2017.05.016

Paper III   **M.J. Garcia-Bonete**, M. Jensen, C.V. Recktenwald, S. Rocha, V. Stadler, M. Bokarewa and G. Katona. *Bayesian Analysis of MicroScale Thermophoresis Data to Quantify Affinity of Protein:Protein Interactions with Human Survivin*, Scientific Reports (2017) **7**:16816
doi: 10.1038/s41598-017-17071-0

Paper IV   **M.J. Garcia-Bonete** and G. Katona. *Bayesian machine learning improves single wavelength anomalous difference phasing*, [Manuscript] (2019)

Related Publications

Paper I   **M.J. Garcia-Bonete**, M. Jensen and G. Katona. *A practical guide to developing virtual and augmented reality exercises for teaching structural biology.*, Biochemistry and Molecular Biology Education (2019) **47**:16-24
doi:10.1002/bmb.21188

Paper II   V.A. Gagner, I. Lundholm, **M.J. Garcia-Bonete**, H. Rodilla, R. Friedman, V. Zhaunerchyk, G. Bourenkov, T. Schneider, J. Stake and G. Katona. *Observation of terahertz dynamics in bovine trypsin*, [Manuscript]

Contribution report

Paper I   I participated in the paper writing and I produced the figures.

Paper II   I participated in preparing the review.

Paper III   I was responsible for the entire project. I designed the microarray, purified the protein and performed the experiments. I took part in the data analysis, in writing the paper and I produced all the figures.

Paper IV   I was responsible for the entire project. I purified and crystallised the proteins. I participated in the data collection and analyses. I solved and refined the structures. I contributed to writing the paper and producing the figures.

Contents

Abbreviations

BIR	Baculovirus IAP Repeat domain
CARD	Caspase Recruitment Domain
CCD	Charge Coupled Device (type of detector)
cIAP$_1$	Cellular Inhibitor of Apoptosis Protein-1
cIAP$_2$	Cellular Inhibitor of Apoptosis Protein-2
CPC	Chromosomal Passenger Complex
D$_{max}$	Maximum Particle Diameter
DTT	Dithiothreitol
eGFP	enhancer Green Fluorescence Protein
EM	Electromagnetic
ESRF	European Synchrotron Radiation Facility (synchrotron radiation facility outside Grenoble)
FSEC	Fluorescence Size Exclusion Chromatography
IAP	Inhibitor Apoptosis Protein family
ID09b	Insertion Device 09b (beamline at the ESRF)
ILP2	IAP like protein 2
IMAC	Immobilised Metal Affinity Chromatography
NIAP	Nucleo-binding oligomerization domain-like receptor Inhibitor Apoptosis Protein
PTM	Postransductional Modifications
RFU	Relative Fluorescence Units
Rg	Radius of gyration
RING	Really Inserting New Gene zinc finger domain
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
SEC	Size Exclusion Chromatography

T$_m$	Melting Temperature
TSA	Thermal Shift Assays
UBA	Ubiquitin-Associated domain
UBC	Ubiquitin-Conjugating domain
WT	Wild Type
XIAP	X-linked Inhibitor of Apoptosis Protein

Chapter 1

Introduction

All living organisms are composed of the basic structural and functional unit called a cell. The cell consists of water, inorganic ions and organic molecules (carbohydrates, nucleic acids, lipid and proteins) and can be classified into prokaryotic and eukaryotic. Eukaryotic cells are more complex than prokaryotic ones, and are characterised by presenting not only different compartments (organelles) with specific functions, but also DNA in the nucleus [1, 2]. They are found in more complex organisms, which may also be multicellular like humans.

To be able to maintain the correct shape, size and functions, multicellular organisms need to balance the total number of cells by two essential physiological processes: cell division and cell death. On the one hand, cell division increases the number of cells and allows the organism to grow. On the other hand, cell death eliminates those cells no longer needed or are damaged. These two processes are crucial for correct cellular balance in organisms and they should be tightly controlled or regulated [3]. In humans, the misregulation of these processes is linked to severe diseases, such as cancer, neurological, inflammatory and immune diseases [4, 5].

This thesis focuses on studying the human survivin protein, which is involved in cell division and cell death regulation. As this protein has been related to chemotherapy resistance, recurrence and bad outcome cancers, a better understanding of its function can lead to better diagnostics and treatment [6].

1.1 Protein structure and function

Proteins are one of the most important macromolecules in the cell. They are involved in almost every cellular process and their structure is required for both their function and the regulation of these processes. They are encoded in the genes present in the DNA which is transcribed into mRNA and is translated into protein. Proteins are polymers composed of a combination of 20 different amino acids. The amino acid sequence is specific for each protein and determines their three-dimensional structure and function.

In eukaryotic cells, the number of synthesised proteins is much bigger than that of genes. This is mainly possible by two processes that occur during eukaryotic protein synthesis: alternative splicing during mRNA maturation and post-translational modification (PTM) [7]. mRNA maturation occurs at the nucleus before protein is exported to the cytoplasm to be translated. Eukaryotic genes are present as introns and exons, which are nucleotide sequences that carry information. During mRNA maturation, introns are removed by RNA splicing and only exons are present in the final mRNA that encodes protein. Alternative splicing allows multiple proteins to be encoded from a gene by including some introns in matured mRNA. These proteins are commonly called isoforms and normally present a similar function, but can also perform unique functions.

PTM comprises chemical modifications introduced into proteins after their translation in the cytoplasm. The presence of PTM in proteins plays an important role in their function as they can be involved in their regulation, localisation and interaction with other molecules in cellular pathways.

1.2 Protein interactions

Proteins perform their function by interacting with different molecules, ranging from small ligands or cofactors to big complexes (e.g. proteins, DNA, lipids, etc.). Understanding the different interactions of a protein provides relevant information about its function, regulation and role in the involved cellular processes.

One of the important reasons for improving our understanding about protein interactions and cellular pathways is the discovery of new chemicals with a therapeutic effect (drugs) to cure the disease or reduce its symptoms [8]. The discovery of new drugs is closely linked to protein structure and interactions because, by knowing where molecules bind and their affinity, more selective and efficient drugs can be developed [9]. By assuming that an interaction between two biomolecules is rapidly reversible in an equilibrium controlled by the law of mass action, it can be defined as follows [9]:

$$[A] + [B] \underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} [AB]$$
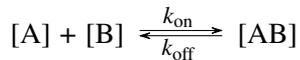
where:

[A] and [B] = is the concentration of the two interactions molecules,
              respectively

[AB]   = is the concentration of the complex

k_{on} = is the association rate constant

k_{off} = is the dissociation rate constant

Its binding affinity is defined as the strength of the interaction between two molecules, and it is physico-chemically described as the dissociation constant (K_D) when the system is in equilibrium (Eq.1.1).

$$K_D = \frac{[A][B]}{[AB]} = \frac{k_{off}}{k_{on}} \tag{1.1}$$

The dissociation constant can be used to calculate binding free energy by the van't Hoff formula (Eq.1.2).

$$\Delta G = RT \ln K_D \tag{1.2}$$

where:

R = is the universal gas constant

T = is the temperature expressed as Kelvin

1.3   Inhibitor of Apoptosis Protein family: Survivin

The Inhibitor of Apoptosis Protein (IAP) family is a group of human proteins that suppress programmed cell death (apoptosis) by different stimuli [10]. Although these proteins have several domains and functions, they have at least one Baculovirus IAP Repeat (BIR) domain [11]. This domain is characteristic of this family, gives its name to the genes that encode BIRC proteins (Figure 1.1) and is important for the direct interaction between pro-apoptotic proteins (e.g. caspases) [11, 12]. It is a globular domain that consists of approximately 70 residues and binds a Zn^{2+} ion coordinated by three cysteines and one histidine ($CX_2CX_6WX_3DX_5HX_6C$) [13]. The Human IAP family consists of eight members ($cIAP_1$, $cIAP_2$, XIAP, livin, ILP2, NAIP, survivin and Apollon/BRUCE) grouped into three different classes (Figure 1.1) depending on the presence of a RING (Really Interesting New Gene) zinc finger domain and the homology of their BIR domains (BIR1, BIR2 and BIR3) [14, 15].

Survivin is the smallest IAP family member and is encoded by the gene *BIRC5*, located at human chromosome 17 (locus 25.3) [17, 18]. Its normal expression is limited to developing embryos or rapidly dividing cells (e.g. haematopoietic, epithelial or gonadal cell lines) [6, 19]. The expression of survivin in differentiated tissues is usually linked to tumours or other diseases [20, 21]. Nonetheless, other IAPs can be expressed in differentiated tissues in normal cell lines. This protein is 16.6kDa big and presents a sequence of 142 amino acids. Survivin is located in the cytoplasm, the nucleus and in mitochondria, and is described as a dimer by several structural studies, including X-ray crystallography and nuclear magnetic resonance (NMR) [22–24]. The monomer structure consists of a BIR domain (residues 1-88), a linker (residues 89-97) and an extended carboxyl-terminal α-helix

Human Inhibitor Apoptosis Protein family
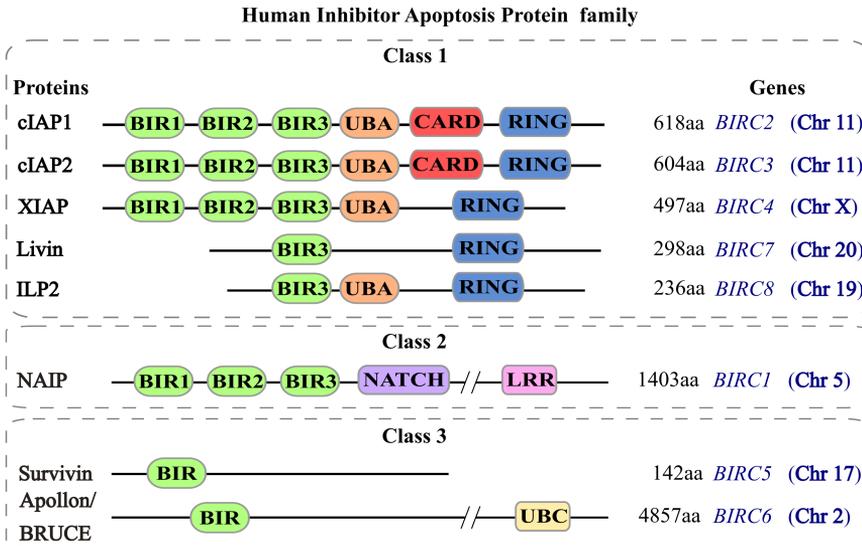
Figure 1.1. *Human IAP Family representation*. Three Human IAP Family classes are shown; including the proteins, genes and chromosomes that encode them. Class 1 consists of cIAP$_1$ and cIAP$_2$ (Cellular IAP 1 and 2), XIAP (X-linked IAP), livin and ILP2 (IAP-like Protein 2), presents one RING zinc finger domain and can have from one to three BIR domains in tandem. Class 2 is formed by an NAIP (Neuronal Apoptosis Inhibitor Protein), and presents three BIR domains and the characteristic NLR family domains (Nucleotide-binding Oligomerisation Domain (NOD)-like Receptor) family. Class 3 consists of the survivin and Apollon/BRUCE (BIR Repeat-containing Ubiquitin-Conjugating Enzyme) proteins that only contain one BIR domain similar to BIR1 or BIR2, depending on the classification used. The different domains are: BIR (Baculovirus IAP Repeat); CARD (Caspase Recruitment Domain; in green); LRR (Leucine-Rich Repeat; in pink); RING (Really Interesting New Gene; in purple); UBA, (UBiquitin-Associated; in orange); UBC (UBiquitin-Conjugating; in yellow) and NATCH (named as the protein that contains it; NAIP, C2TA, HET-E, and TEP1; in blue). This picture was created according to the domain information obtained from the "*Batch* Web CD-Search tool from NCBI" [16].

(residues 98-142) or a coil-coil domain (Figure 1.2). Like other IAPs, survivin binds Zn^{2+} in its BIR domain, and this binding is coordinated by Cys57, Cys60, His77 and Cys84 (Figure 1.2) [22,23]. The structure presents three separate and chemically different surfaces, including acidic and basic

patches in the BIR domain, as well as a hydrophobic helical cluster at the end of the C-terminal [23]. The basic patch includes part of the BIR domain, the linker and the beginning of the C-terminal α-helix. Part of this region (residues 89-102) participates together with N-terminal residues (6-10) in dimer formation. The hydrophobic cluster at the end of the C-terminal may play an important role in the survivin interactions with other proteins (Figure 1.2) [22, 23].
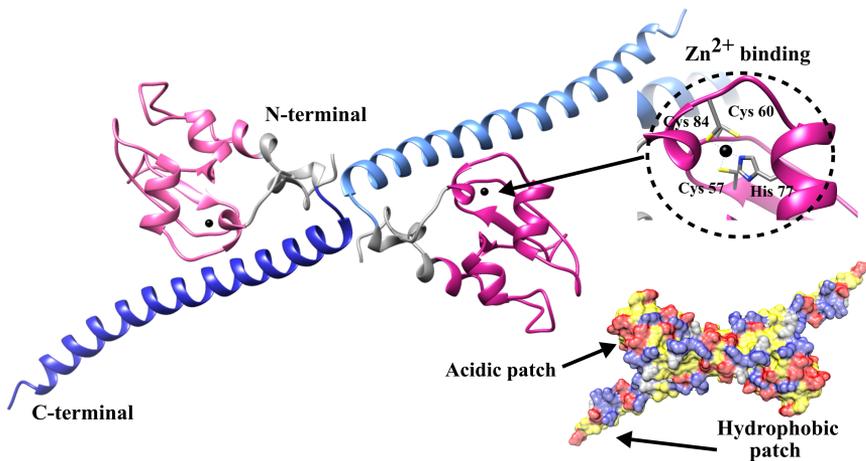
Figure 1.2. *Dimer survivin structure.* The BIR and C-terminal domains represented in pink and blue, respectively. The Zn^{2+} atoms are displayed as black spheres. At the right top corner of the figure the coordination of the Zn^{2+} binding is represented. At the right bottom corner, the molecular surface of the survivin dimer is displayed according to the local chemical properties: acidic (red), basic (grey), polar(grey) and hydrophobic (yellow). The figures were generated with UCSF chimera 1.11 software [25].

Survivin is involved in several important cellular processes, including cell division, apoptosis and the correct homeostasis of the immune system [5, 26]. Survivin can act as a transcription factor and regulates the synthesis of microRNA (non-coding short RNA molecules that provide epigenetic biological control by regulating gene expression at post-transcriptional

levels) [27]. The aberrant overexpression of survivin is tumour-related and is indicative of diminished overall survival, higher recurrence rates and resistance to therapy [6, 17, 20, 28].

1.4 Survivin functions

1.4.1 *Cell division*

The cell cycle is divided into four coordinated phases that consist of two gap phases (G_1 and G_2), one DNA synthesis phase (S) and the mitosis or division phase (M) (Figure 1.3). Mitosis is also divided into various phases; prophase, prometaphase, metaphase, anaphase and telophase, followed by cytokinesis (Figure 1.3) [1,29]. This whole process requires tight regulation and different checkpoints to ensure correct cell division.

Survivin plays an important role in cell division by forming part of the chromosomal passenger complex (CPC). The CPC is formed by four proteins; borealin, the inner centromere protein (INCENP), aurora kinase B (aurora B) and one monomer of survivin [30–32]. This complex is, in turn, divided into two modules linked by the INCENP protein: the localisation and regulation module and the activity module. The localisation module consists of borealin and survivin bound to the N-terminal of INCENP. The activity module is composed of aurora B and the C-terminal of INCENP, called IN-box, which is essential for full aurora B activity [32].

This protein complex is first observed at the nucleus in the late S phase, and presents its higher expression in phases G_2 and M. During mitosis, this complex is localised at different levels and is involved in many different processes. In the early prophase, the CPC is observed along chromosomal arms and is confined to the inner centromere region in the late prophase, prometaphase and metaphase. Its function in these phases includes, among others; the regulation of the chromosome structure, the removal of cohesin (a protein complex that keeps sister chromatids together) from chromosomal arms, mitotic spindle formation, the regulation of kinetochore-microtubule

attachments and the regulation of mitotic checkpoints. At the beginning of the anaphase, the CPC moves from the inner centromere to the microtubules of the central spindle, where it is involved in correct chromosome segregation. In the telophase, the CPC is localised in the mid-body and is involved in the physical separation of cells [32–34].

This cellular process becomes complicated because it requires the interaction of many different protein families and is tightly regulated [34]. The CPC has been described in interactions with several of these proteins, such as the interaction with histone H3 through survivin [35–37].

The human shugoshin-like protein family also plays a key role in cell division. It consists of two members (hSgo1 and hSgo2) and is related with the protection of the cohesin complex that keeps sister chromatids together before segregation [32, 38, 39]. Previous studies have shown that hSgol1 and hSgol2 directly interact with the CPC and are involved in the localisation of the CPC in centromeres [34, 40–44]. In **paper III**, the physical interaction of these proteins with human survivin is demonstrated.
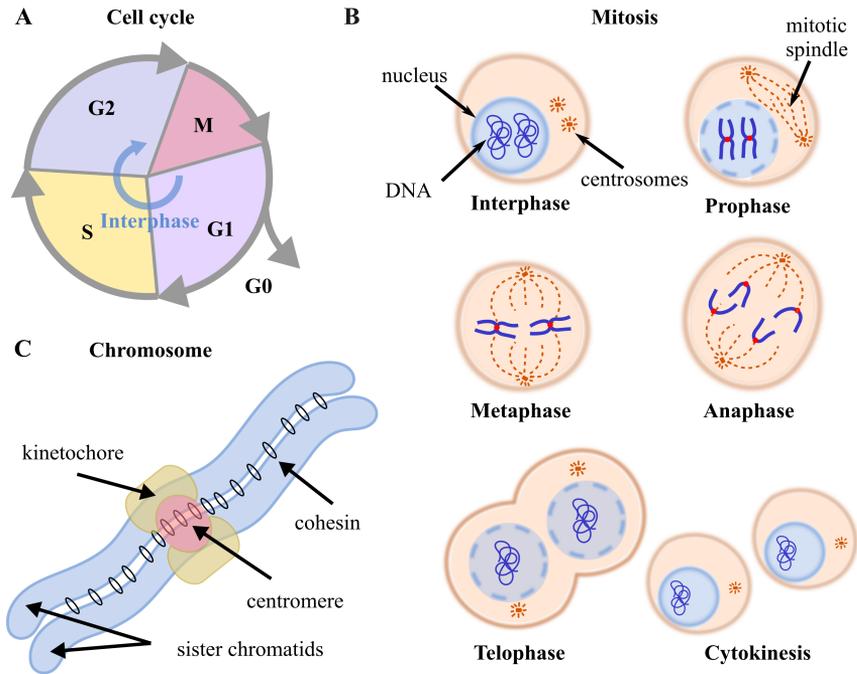
Figure 1.3. A. *Cell cycle*. In phase G_1, the cell is metabolically active and continuously grows, but DNA is not duplicated. From this phase, the cell can exit the cell cycle and go to the resting stage (Gap phase 0, G_0) or can continue with cell division and go to the phase S. DNA replication occurs in the phase S and is followed by the phase G_2, where the cell continues growing and produces the proteins needed for division. **B**. *Cell division or Mitosis.* In the prophase, chromosomes condense by presenting two sister chromatids linked by the centromere. In this phase, centrosomes (microtubules organising centre, MTOC) also build a cytoskeletal structure that is required for division, namely the mitotic spindle. In the prometaphase, the nuclear membrane is degraded and the mitotic spindle comes into contact with chromosomes. In addition, kinetochore (a complex protein structure) is associated with the centromere of each sister chromatid, which allows the connection of sister chromatids with the mitotic spindle by microtubules. In the metaphase, chromosomes are located along the equator zone of cells by microtubules. In the anaphase, sister chromatids are separated by a force generated by microtubules in opposite directions. Each chromatid gives a full new chromosome. The telophase is the last phase of mitosis, in which new chromosomes reach the mitotic spindle, the membrane is restored and the cell is prepared for cell division into two cells, known as cytokinesis. **C**. *Chromosome structure* [1, 2].

11

1.4.2   *Apoptosis*

Apoptosis is a process by means of which harmed cells or those no longer needed, are degraded by activating programmed cellular death. It is also called programmed cell death and is essential for maintaining tissue homeostasis in multicellular organisms, embryo development and immune system function [2, 45]. Apoptosis is mediated by a cysteine-aspartic protease family called caspases, which includes caspase-3, -6, -7, -8 and -9. These proteins are synthesised as an inactivated form (procaspase precursors) and can be classified as initiator caspases (caspase-8 and -9) or executioner caspases (caspase-3, -6 and -7). Caspases activate one another during the process called the caspase cascade, which leads to the proteolytic cleavage of several cellular targets and cell death. Apoptosis consists in two pathways that can be extrinsic (initiated by an external cellular signal) and intrinsic (initiated by an intracellular signal; e.g. cellular stress) (Figure 1.4) [46]. Misregulation in apoptosis is linked with several diseases, such as cancer and autoimmune diseases (described in **paper I**) [3, 5, 45].

Survivin can inhibit both apoptosis pathways, but it cannot bind caspases directly as it lacks the linker sequence upstream of the BIR domain present in other IAPs (e.g. XIAP) [23, 24]. The interaction of survivin with the XIAP protein enhances XIAP stability, which results in the inactivation of caspases-3 and -9 [5, 47]. Survivin can also bind pro-apoptotic protein Smac/DIABLO, which is an antagonist of the XIAP protein [36, 48]. This interaction prevents the Smac/DIABLO being released from mitochondria, and inhibits caspase activation. Structural studies have shown the interaction of survivin with the N-terminal of Smac/DIABLO [36].
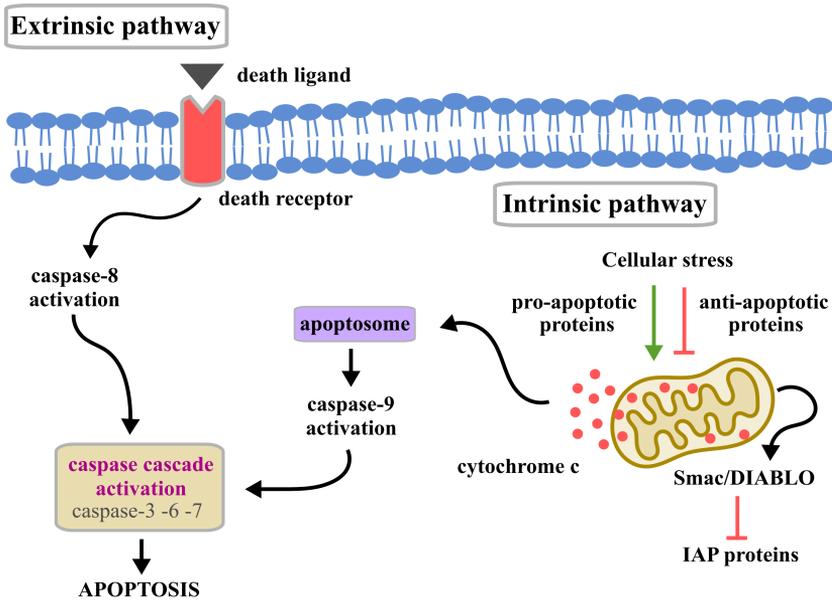
Figure 1.4. *Apoptosis pathways.* The extrinsic pathway is activated by an external ligand binding to the death receptor placed on the cell surface. This brings about the activation of the intrinsic pathway and the caspase cascade, which lead to cell death. The intrinsic pathway is also recognised as mitochondrial apoptosis and is activated by different cellular stresses that lead to cytochrome c release from mitochondria and the formation of apoptosome (a protein complex), which activates caspases to cause cell death [46].

1.5   Survivin isoforms

The *BIRC5* gene consists in four dominant exons (1, 2, 3 and 4) and two cryptic exons (2B and 3B) that lead to the expression of different alternative spliced variants of survivin, as shown in Figure 1.5 [28]. Many different survivin isoforms have been reported, and at least six of them have been seen to be of biological significance; survivin, survivin-2B, survivin-ΔEx3, survivin-3B, survivin-2α and survivin-3α [28, 49]. The majority of the mRNA expressions of the *BIRC5* gene comprise survivin, survivin-2B

13

and survivin-ΔEx3 [5].

Human Survivin Isoforms
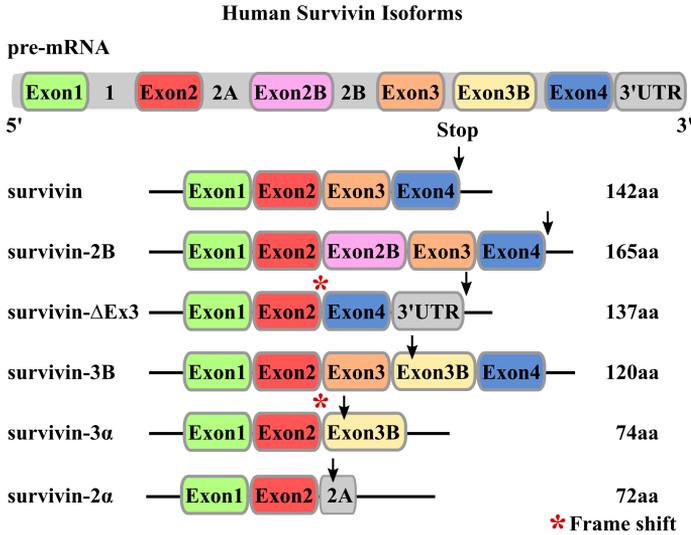
Figure 1.5. *Alternative BIRC5 splicing.*

The expression of survivin isoforms has been related mainly to malignant cells and is almost undetectable in normal cell lines [50–52]. While survivin expression is constantly high expressed in several cancers, the expression of survivin isoforms is variable and depends on specific cancer types and stages. Survivin isoforms are involved in various carcinogenic processes, including proliferation, apoptosis and metastasis [52]. This difference in expression and its relation with tumour development and patient survival suggest that some isoforms may have regulatory mechanisms and might be a better marker of tumour prognostics and diagnostics [53, 54].

Survivin-2B is the longest isoform with 165 residues. It presents the insertion of cryptic exon 2B, which interrupts the BIR domain by adding 23 residues (between the Ile74 and Gln75 residues) that affect Zn^{2+} binding [28]. This isoform has been located mainly in the cytoplasm, while low

expression levels have been seen at the nucleus and in mitochondria [49, 55, 56]. The survivin-2B function is unclear and controversial. Some studies show that it presents pro-apoptotic activity (by promoting cell death), or its expression is correlated inversely with tumour stages, and is more expressed in well-differentiated tumours. However, other studies reveal that expression is related more to treatment-resistant cancer cells or to other diseases (e.g. high survivin-2B expression in the serum of rheumatoid arthritis patients) [20, 49, 57, 58].

Survivin-ΔEx3 lacks exon-3 and contains only 137 residues [59]. This exclusion of exon-3 generates a unique carboxyl terminal that includes features that are not present in other isoforms [28]. This terminal consists of a mitochondrial localisation signal sequence, a nuclear localisation signal sequence and a Bcl2 homology domain (BH2). Survivin-ΔEx3 has been found mainly expressed in the nucleus of malignant cell lines [54, 56, 59]. The BH2 domain is characteristic of another apoptosis regulatory family, the Bcl2 family [60], and it confers survivin-ΔEx3 a specific anti-apoptotic function [61]. This isoform can be associated with Bcl2 (anti-apoptotic protein) and inhibits caspase-3 activity to result in apoptosis inhibition [62].

Noton et al. have demonstrated that survivin can form heterodimers with other isoforms (survivin-2B and survivin-ΔEx3), and that these isoforms do not play a role in either mitosis or the complex formation with the CPC [52, 63]. The formation of survivin heterodimers might play an important role in survivin regulation as isoforms exhibit various apoptotic properties and can affect the apoptotic activity of survivin [20, 28, 64].

Although there are many studies about the isoforms of survivin, their possible functions and how they interact with survivin remain to be elucidated. This thesis tests initial expression and purification trials of recombinant survivin-2B and survivin-ΔEx3 (Chapter 4).

1.6 Scope of the thesis

This thesis focus on the characterisation and structural studies of human survivin interactions that were used as a test system for Bayesian inference of data analyses.

Chapter 2 describes the methodology followed in this thesis for protein production and characterisation, protein:protein interactions, X-ray crystallography and structure determination.

Chapter 3 briefly introduces the statistical inference focussing of the Bayesian inference. The different Bayesian approaches used for the data analyses of **paper I/IV** and **paper III** are described. **Paper I** and **IV** investigate how the Bayesian inference can improve the calculation of structure factor differences in X-ray crystallography.

Chapter 4 focuses on the characterisation of human survivin and the interaction experiments of survivin:borealin and survivin:shugoshin proteins published in **paper III**. At the end of this chapter, the initial expression and purification trials of the recombinant survivin isoforms are described.

Chapter 5 summarises the conclusions and future work.

Paper II reviews current knowledge about survivin in autoimmune diseases. It describes the usefulness of survivin measures for clinical applications, and provides survivin inhibiting strategies and recent results about survivin inhibition in modern therapies for cancer and autoimmune diseases.

Chapter 2

Methodology

2.1 Protein production

2.1.1 *Expression system*

Biochemical analyses often require large amounts of pure sample (in the order of mg). Obtaining these amounts from a natural source can be arduous work, and is sometimes even impossible. The overexpression of recombinant proteins in heterologous systems is normally used. Recombinant proteins are synthesised in a cell from exogenous DNA (vector) containing the gene of interest, generated by genetic engineering. This DNA is artificially introduced into the cell during a process called transformation. However, the expression of recombinant proteins can also involve problems, such as poor host growth, the formation of inclusion bodies (non-soluble proteins), protein inactivation, no expression, misfolding, etc [65]. These problems can be overcome by optimising the expression conditions, changing the expression system, or even modifying the protein of interest (e.g. truncated proteins) [66, 67].

Many different host systems are available for recombinant protein expression (bacteria, yeast, insect cells, mammalian cells, etc.). *Escherichia coli* (*E. coli*) bacteria have been highly used for the overexpression of many heterologous recombinant proteins [67]. It is a prokaryotic organism that is easy to culture and allows high cell density, does not require expensive media to grow (e.g. Luria Bertani medium, LB) and can be easily transformed with exogenous DNA. The expression in prokaryotic cells involves certain disadvantages like post-translational modification (PTM) (e.g. glycosylation, phosphorylation) not being included [68]. Some eukaryotic proteins require such modifications for their correct folding and function. However, if no knowledge is available about these requirements, it is worth attempting overexpression in *E.coli* before engaging in more difficult expression hosts.

The most common *E.coli* expression strains present the T7 system (e.g. DE3 strains) [67]. In their chromosomal DNA, these strains have a copy of the phage T7 RNA polymerase gene, controlled by the lac promoter, which allows the expression of the genes cloned downstream of the T7 promoter. In the presence of an inducer (e.g. Isopropyl-β-D-1-thiogalactopyranoside, IPTG), T7 RNA polymerase is expressed and allows the expression of the gene of interest [69]. In this thesis, three different strains were used for protein expression (One Shot BL21 Star (DE3), Rossetta^{TM} (DE3) pLysS and Lemo21 (DE3)). They are all derivatives of the BL21 (DE3) strain, which presents the T7 system approach and is deficient in proteases Lon and OmpT, which allows higher protein expression [70,71]. One Shot BL21 Star (DE3) (Invitrogen, Thermo Fisher Scientific) also presents a mutation in the RNase gene by providing greater mRNA stability and better protein yields. Rossetta^{TM} (DE3) pLysS (Novagen-Merck) is designed to improve eukaryotic protein expression by presenting codons that are rarely used in

E.coli. This strain also expresses the natural inhibitor of T7 RNA polymerase, T7 lysozyme (pLysS), by suppressing its basal expression and stabilising the vectors that encode the proteins which affect cell growth and viability. Lemo21 (DE3) (New England BioLabs) has an extra plasmid with the Lemo system (pLemo), which allows a tuneable expression of difficult soluble proteins (e.g. toxic proteins). pLemo encodes the T7 lysozyme gene controlled by a L-rhamnose inducible promoter [72, 73].

The pET expression system is the most widespread vector system used for the expression of recombinant proteins [74]. These vectors contain the T7 promoter and the translational signals required for protein expression [75, 76]. They might also present other features [66], such as selection markers (provide resistance to antibiotics to ensure that the vector remains inside the cell), fusion proteins (to improve target protein solubility) [77], fusion tags (to improve protein purification) and cleavage sites (the amino acid sequences recognised by proteases, e.g. thrombin or Human Rhinovirus 3C protease (HRV3C), and are used to remove different tags after purification) [78]. In this thesis, several pET vectors were selected to improve both the expression and solubility of the target proteins (Table A.1).

2.1.2 *Cloning and expression*

The different protein target genes were commercially obtained (Thermo Fisher Scientific) and cloned into the specific vector. Cloning was performed with the InFusion HD Cloning kit (Clontech Takara) [81] and the new constructs containing the target gene were confirmed by sequencing (GATC Biotech). Appendix A describes all the cloned genes and the used primers.

Protein expression can be performed very differently. In this thesis, expression optimisation (temperature, induction time, IPTG concentration, etc.)

Table 2.1. pET vectors description

Vector	Selection marker	Fusion tag	Cleavage site	Advantage
pHis8 [79]	Kan^R	8xHis-tag (N-terminal)	Thrombin	Affinity chromatography
pET28b+8His	Kan^R	8xHis-tag (N-terminal)	Thrombin	Affinity chromatography
pET29b	Kan^R	6xHis-tag (C-terminal)	Thrombin	Affinity chromatography
pWarf(-) [80]	Kan^R	eGFP protein +8His-tag (C-terminal)	HRV3C	Fluorescence Affinity chromatography
pET48b	Kan^R	Thioredoxin(TRX) +6xHis-tag (N-terminal)	HRV3C	Solubility Affinity chromatography
pET49b	Kan^R	Glutathione S transferase (GST) +6xHis-tag (N-terminal)	HRV3C	Solubility Affinity chromatography

on a small scale was done following a large-scale overexpression under the best conditions [82]. After protein overexpression, cells were harvested, re-suspended in lysis buffer and disrupted by a high pressure homogeniser (EmulsiFlex-C3, Avestin).

Lysis buffer composition strongly depends on the target protein (e.g. isoelectric point of the protein) and the chosen purification approach. Additives are often included in the lysis buffer to improve target protein stability and solubility (e.g. salt, detergents, reducing agents, ligands, etc.) [83]. In this thesis, lysozyme, deoxyribonuclease (DNAse) and pefabloc (Sigma-Aldrich) were always included in lysis buffers. Lysozyme is an enzyme that

affects the bacteria cell wall and ensures the complete lysis of cells. DNAse is an enzyme that degrades DNA and improves a sample's viscosity. Many different proteins are released during cell disruption, including proteases (enzymes that degrade other proteins). Pefabloc SC (Sigma-Aldrich) is an irreversible protease inhibitor that is added to avoid recombinant protein degradation.

2.1.3 *Protein purification*

In this thesis, two-step purification was done, including affinity chromatography and size exclusion chromatography (SEC) [84]. Affinity chromatography consists of immobilising the target protein in a matrix and separating it from the protein mixture. In immobilised metal affinity chromatography (IMAC), the matrix is charged with metal ions, such as Ni^{2+} [85, 86]. These metals have an affinity for those proteins by presenting poly-histidine tags and bind them. After several wash steps to remove non-specific bindings, the target protein is eluted with imidazole, which competes with the histidine-tag for Ni^{2+} [85, 86]. This approach was used as the initial purification step because it normally leads to a reasonably pure sample.

SEC separates proteins into their molecular weights, where the larger the protein, the faster elution becomes. This approach can be used to remove different sized impurities, and to analyse the protein's homogeneity and oligomerisation state [87]. This chromatographic step is commonly used as a final purification step.

In some techniques, the presence of tags can affect the results, and they are sometimes removed after purification [78]. For example in protein crystallography, the presence of tags can increase the protein disorder level, which makes it more difficult to crystallise. Reverse chromatography

was also used to remove purification tags and fusion proteins after cleavage. IMAC was used to bind the free histidine-tags and fusion proteins or proteases that also contain histidine-tag [86]. The free-tag target protein is no longer able to bind to the matrix and is eluted directly to allow separation. The benzamidine column was used to remove thrombin after survivin histidine-tag cleavage. As thrombin is a serine protease and benzamidine is a reversible inhibitor of serine proteases, they interacted to allow thrombin to be separated from the target protein.

Protein purity was analysed by denatured protein electrophoresis (SDS-PAGE) [88]. The protein concentration was estimated by both the BCA assay [89] and absorbance at 280 nm with a spectrophotometer. The extinction coefficient of each target protein was theoretically calculated using the protein sequence in ProtParam [90, 91].

2.2 Peptide microarray

Microarray technology is a laboratory tool that provides high-throughput and rapid information about gene expression or biomolecular interactions. This technique was developed by Tse Wen Chang in 1983 using antibody microarrays to identify those cells carrying specific antigens [92–95]. However, DNA microarray (or DNA chip) technology became more popular after Davis and Brown published their work in Science in 1995 after studying *Arabidopsis thaliana* gene expression [94, 96].

A microarray consists of a solid support (membrane, plastic or glass) of a few cm^2, where biological probes (like DNA, proteins, etc.) are immobilised and exposed to a target molecule or a sample (e.g. a purified

protein, cDNA, cell lysates, etc.) [97]. The probes and target molecule interaction can be recorded by different methods (e.g. chemoluminescence, chromogenic enzymatic reactions, radioactive isotopes, etc.) [93, 98], but the most common one is fluorescent-labelled. The results are quantified by an image analysis. Nowadays, there is a wide range of microarrays approaches depending on the employed biological probe (DNA, protein, peptides, chemical compound, tissues, phenotypes, antibodies, etc.) [93].

The peptide microarray technology was introduced by Ronald Frank in 1992. He used spot synthesis microarrays to analyse those antibodies binding to peptides [99]. This technique allows several thousands of peptides to be simultaneously screened in a single experiment and different aspects of protein-protein interactions to be studied. The peptide microarray is also useful for studying the influence of PTMs on protein interactions, such as histone studies [100, 101].

In this thesis, the peptide microarray approach (**paper III**) was used to discover new interactions with human survivin using PEPperCHIP Peptide Microarray (PEPperPRINT, GmbH, Heidelberg, Germany) (Figure 2.1). This technique can be divided into several steps, including microarray preparation, blocking, pre-staining with the secondary antibody, sample incubation, staining with secondary antibodies and detection.

The microarray was designed by including the sequences of 19 different (partial or complete) proteins that are able to interact with survivin. They were converted into linear peptide segments (15 amino acids long) with a peptide-peptide overlap of 10 amino acids, and were printed in duplicate in the microarray.

Peptide Microarray technology
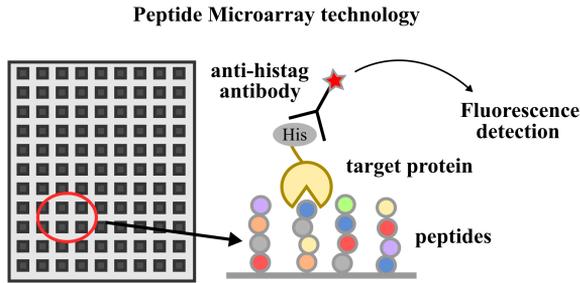
Figure 2.1. *Microarray sketch*

The microarray was synthesised *in-situ* on a glass support by the PEP-perPRINT technology, which consists of a laser printer with 24 different amino acid printing units [102]. The target protein of this experiment was human recombinant survivin, which presents a 6XHis-tag that was used for detection. Survivin binding detection was performed by a fluorescent antibody against His-tag (6XHis -Tag Antibody DyLight680). Two types of positive control peptides were also included as the internal quality control. 6XHis-tag (HHHHHH) peptides were included to evaluate anti-His tag antibody binding and to ensure survivin recognition. Human influenza haemagglutinin (YPYDVPDYAG) peptides were included as a quality control. These peptides were recognised by the anti-HA antibody.

Firstly, the microarray was incubated with a blocking solution (normally containing bovine serum albumin) to reduce non-specific binding. Since secondary antibodies can sometimes interact with synthesised peptides by giving background interactions, pre-staining with the secondary antibodies was done to detect and discriminate the signals that did not come from the interactions. Afterwards, the microarray was incubated with the target protein (survivin), followed by the secondary antibodies and detection.

Peptide microarrays have the advantage of being able to screen many different peptides in a single experiment with a small amount of sample and are, thus, more stable than protein microarrays. However, some limitations exist. Peptides are a small fraction of the protein and do not always contain the full interaction region or the specific secondary structure, which can lead to false-positive and negative interactions. As they are immobilised in a solid support, the interaction might also be affected. In addition, binding parameters cannot be obtained from such experiments, and qualitative information about the interaction is mainly obtained.

This technique is very useful for initially recognising possible interaction partners. However, cross-validations with other techniques that provide binding information (e.g. MST, ITC, etc.) should also be used.

2.3 MicroScale Thermophoresis

MST is a versatile and sensitive technique for studying biomolecule interactions and estimating their binding affinity. It is based on a physical phenomenon called thermophoresis, described by Carl Ludwig in 1856 [103], which consists of the directed movement of particles over a temperature gradient. This movement depends on the size, charge and hydration shell of the studied molecule [104]. An MST measurement consists of heating a sample inside a thin glass capillary by an infrared laser and recording the movement of molecules by monitoring sample fluorescence [105]. This generates a microscale temperature gradient (maximum temperature increase of 2-6K) inside the capillary, which makes the molecules to quickly move away from the heated spot (depletion). Simultaneously, thermophoresis is followed by measuring fluorescence at the heated spot. This fluorescence derives from

a fluorophore that is intrinsic or covalently attached to the molecule of interest. Movement of molecules leads to a concentration change between the heated spot and the bulk liquid, which can be quantified by the Soret coefficient (S_T) (Eq. 2.1) of the studied molecule [104, 105].

$$\frac{C_{hot}}{C_{cold}} = e^{-S_T \cdot (T - T_0)} \tag{2.1}$$

where:

C_{hot} = molecule concentration at the localised spot when the laser is on
C_{cold} = molecule concentration at the localised spot when the laser is off
T = final temperature
T_0 = initial temperature

The Soret coefficient is characteristic of the studied molecule and depends on the size, charge and hydration shell of the molecule [105]. When two molecules interact and form a complex, their size, charge or hydration shell can be affected. This makes MST a suitable technique for studying binding formation and obtaining binding affinities. MST measurements can be divided into different events (initial fluorescence, T-Jump, thermophoresis, inverse T-jump and back diffusion), as described in Figure 2.2 [106]. Initial fluorescence is the sample fluorescence before the experiment starts, done without laser heating. The T-Jump is the fluorescence change that occurs when the laser is heated up and before the thermophoretic effect takes place. Thermophoresis is the fluorescence change caused by molecules moving when the laser is heated up. Inverse T-Jump is the fluorescence change caused by the sample cooling after turning off the laser. Back-diffusion is the fluorescence recovery by the mass diffusion of the molecule after turning off the laser [106].
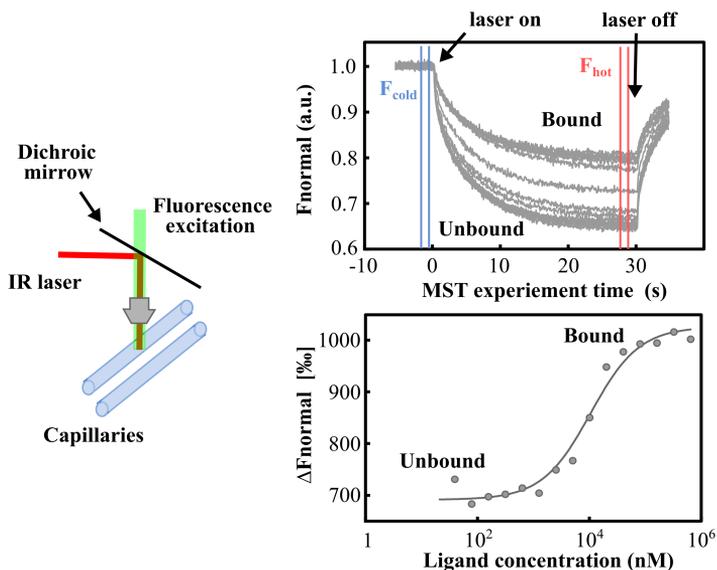
Figure 2.2. A. *MST optics representation*. The IR laser locally heats the sample inside capillaries, and sample fluorescence is excited and detected through the same objective. **B**. MST traces of a standard binding experiment for different concentrations of the non-labelled protein, including bound and unbound states. F_{cold} and F_{hot} represent the fluorescence region when the IR laser is off and on, respectively. They are used to calculate the ΔF_{norm} utilised for the binding curve calculation. **C**. Binding curve. Each point represents the ΔF_{norm} of each MST trace for the specific concentration of the non-labelled protein. [106]

In a binding experiment between two proteins, one of the partners is fluorescent-labelled, while the other is unlabelled. To estimate their binding affinity, a serial dilution of the unlabelled protein is mixed with a constant concentration of the labelled protein. The MST traces of each sample are recorded and binding is analysed by comparing different concentrations. To obtain good quality binding curves, unlabelled protein titration aims to maintain the completely bound and unbound states at the higher and lower concentrations, respectively. In addition, the labelled protein concentration

should be lower than the expected K_D. In thermophoresis, it is also noteworthy that the initial fluorescence does not vary with the different concentrations of the unlabelled protein. The recorded fluorescence is normalised against the initial fluorescence, F_{norm} and used to calculate the normalised fluorescence difference (ΔF_{norm}) for each MST trace (Eq. 2.2) [104].

$$F_{norm} = \frac{F}{F_0} = (1-x)F_{norm}(U) + x \cdot F_{norm}(B) \qquad (2.2)$$

where:

F = fluorescence values after IR laser activation

F_0 = fluorescence values prior to laser activation

x = fraction of fluorescence molecules bound to their targets

T_0 = initial temperature

$F_{norm}(U)$ = contribution of the unbound fluorescence molecule after IR laser activation

$F_{norm}(B)$ = contribution of the complex after IR laser activation

The binding curve can be fitted by plotting the ΔF_{norm} against the unlabelled protein concentrations (on a log10 scale). The dissociation constant is calculated from the law of mass action described in Eq. 2.3 [107].

$$x = \frac{c_f + c + K_D - \sqrt{(c_f + c + K_D)^2 - 4c_f c}}{2c_f} \qquad (2.3)$$

where:

x = bound fraction of the labelled protein reported by the F$_{norm}$ of the MST measurements

c_f = unlabelled protein concentration

c = labelled a protein concentrations

K_D = dissociation constant

MST offers several advantages: it is easy to implement, the experiment can be done in a short time, it does not require large amounts of sample, there are almost no buffer restrictions, it is immobilisation- and temperature-free, and there are no molecular weight limitations [108]. However, it is essential for one of the molecular partners to be fluorescent-labelled, which can affect the native state of the protein and, therefore, the formation of the complex.

MST was used to characterise the interaction between the survivin:borealin and survivin:shugoshin proteins to estimate K$_D$ and other binding parameters. These experiments were performed according to **paper III** and following the NanoTemper technologies protocol in a Monolith NT.115 (green/blue) instrument (NanoTemper, Germany) [109]. Survivin was chemically labelled using the cysteine reactive dye (NT-495-maleimide and NT-547-maleimide) kit (NanoTemper, Germany). Measurements were taken using Monolith NT.115 premium capillaries (NanoTemper, Germany).

2.4 Thermal shift assay

The melting temperature (T$_m$) of a protein is the temperature value at which half the protein loses its structure and is partially denatured. This information can be important for the characterisation of a target protein, for

choosing the optimal buffer for further experiments and for the hit identification of new drugs [110, 111]. The T_m can also be affected by ligands binding to the protein [112].

Thermal shift assays (TSA) allow the easy determination of the T_m shift of a target protein under different conditions (e.g. with different buffers, ligand binding, etc.) and are normally measured by light scattering or fluorescence techniques [112].

The thermofluor assay is a fluorescent technique developed by Semisotnov et al and published in 1991 [113]. These authors studied the binding of the hydrophobic fluorescent probe, 1-anilino-naphthalene-8-sulfonate (ANS), to proteins with different structural organisations.

This technique consists of adding a fluorescence dye (sensitive to the environment) to the protein solution and monitoring protein unfolding when temperature rises. This is possible because this dye type has a low fluorescence signal in polar environments (e.g. an aqueous solution), but presents a high fluorescence signal when exposed to non-polar environments (e.g. a denatured protein) (Figure 2.3) [114].

There are several dye types for such assays. In this thesis, SYPRO Orange (Thermo Fisher Scientific. λ_{ex} 470 nm /λ_{em} 570 nm [115]) was used because it can be easily measured with filters from standard quantitative PCR instruments [114]. However, sometimes some samples can give unclear signals or have a high background, because the dye binds to the native protein state.

At the beginning of a thermal shift experiment, the protein is in its native state (folded) and SYPRO Orange presents a low fluorescence signal. However when temperature rises and the protein starts to unfold, its hydrophobic core is exposed to the solution, which allows SYPRO Orange to bind

it [116]. This increases dye fluorescence until all the molecules are dena-
tured. At the end, the unfolded protein molecules start to aggregate by pro-
ducing dye dissociation and the fluorescence signal starts dropping [117].
From these curves, it is easy to estimate the T_m of the target protein by plot-
ting the first derivative of fluorescence emission according to temperature
(-d(RFU)/dT) [115]. Emission fluorescence is represented in relative fluo-
rescence units (RFU) according to the measured samples and the instrument
used. Figure 2.3 is a sketch of the thermofluor assay.
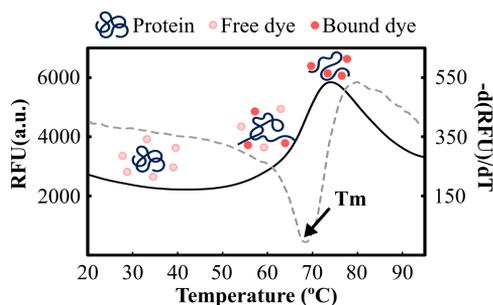
Figure 2.3. *Thermofluor assay*. This image shows the lysozyme melting point curve
(*black line*) and its derivative curve, -d(RFU)/dT (grey dashed line). The minimum
derivative curve point corresponds to the T_m. [117]

A thermofluor experiment was done to analyse the different buffer con-
ditions for human survivin recombinant (Chapter 4). The experiment was
performed in a BioRad CFX96 instrument using a HEX filter (excitation:
515-535 nm, detection: 560-580 nm). The sample was prepared after includ-
ing 2X of SYPRO Orange and 50μM of the protein in the specific buffer.
The protocol consisted of a 30-minute incubation at 20ºC, followed by in-
creasing the temperature by 0.5ºC and a 30-seconds incubation before mea-
suring fluorescence.

2.5   X-ray Crystallography

To understand how molecules behave, their structure is an important factor. By knowing their atomic structure, it is possible to better understand their function and the cellular process where molecules are involved.

Proteins are molecules whose shape and size can vastly differ, approximately 1-10 nm, which makes it impossible to study them by the naked eye (>100 μm) or with light microscopes (1 mm-100 nm). X-ray crystallography is a useful technique for studying the atomic structure of molecules such as proteins, and there are more than 20 Nobel prizes associated with this technique (e.g. "The G-protein-couple receptors studies" by Lefkowitz R.J. or Kobilka B.K. in 2012 [118]).

X-ray (0.01-10 nm) is an electromagnetic (EM) radiation type with a shorter wavelength than visible (400-700 nm) and ultraviolet (10-400 nm) light. As atom bond lengths fall within the range of a few Ångströms (e.g. C-C bond = 1.54 Å), X-ray radiation offers a suitable wavelength for studying the atomic structure of molecules (0.1 nm = 1Å).

X-ray crystallography consists of irradiating a crystal with an X-ray beam and collecting the intensities of the relative reflections from the recorded diffraction patterns. X-ray crystallography requires a crystal to be performed.

A crystal is an ordered three-dimensional array of a specific motif (e.g. atoms, molecules, proteins, etc.) in a lattice. The motif and lattice form what is known as the unit cell. The full crystal can be built by translating only the unit cell into three dimensions. However in the unit cell, there are other symmetry operators (e.g. inversion, reflection, rotoinversion, glide plane, rototranslation, translation and rotation). The asymmetric unit is the smallest portion of a crystal that can generate the unit cell by applying symmetry

operators. Given macromolecules' chiral nature, only rotation, translation and rototranslation can be applied in their crystals, which reduce the number of possible space groups from 230 to 65 [119, 120].

In a protein crystal, molecules are held together by non-covalent interactions that take place between protein molecules and the solvent, which makes protein crystals more fragile than small molecule crystals [121].

2.5.1 *Crystal formation*

Crystal formation can be one of the limitations of protein crystallography because it depends on many different factors (protein concentration, purity, temperature, buffer composition, ligands, protein:precipitant ratio, etc.), and also on the protein's own specific properties.

The use of crystallisation screening (e.g. sparse matrix screen) allows different reagents and methods to be evaluated, which provides information about the conditions leading to crystal formation. The majority of protein crystals need further optimisation, which can be guided by the "crystallisation phase diagram" (Figure 2.4A). This 2D diagram is a simplification to help explain how two variables can affect crystal formation. The vertical axis represents the protein concentration, while the horizontal axis represents the precipitant concentration. Different zones can be described between these two variables: undersaturated, saturated, metastable or growth, labile or nucleation and precipitation [122, 123].

The goal is to create a supersaturation solution of the protein and precipitant, where the protein solution dehydrates in a very controlled manner to create large enough crystals for single crystal X-ray diffraction. Different methods for crystallisation are available, but vapour diffusion is one of the most widespread [119, 120, 123]. In this method, a small volume of

protein solution and precipitant solution are mixed together and placed inside a closed chamber containing a reservoir with the precipitant solution. As the precipitant solution in the reservoir has a higher concentration, the drop slowly dehydrates by vapour diffusion until both the drop and reservoir reach an equilibrium. In this thesis, hanging and sitting drop vapour diffusion methods were used to obtain lysozyme and survivin crystals. The hanging drop is placed on an inverted cover slip, which also acts as seal at the top of the reservoir together with oil or vacuum grease. In the sitting drop, the drop is placed on a pedestal separated from the reservoir [119,120,123].

To crystallise a protein, it is necessary to overcome a similar energy barrier to a chemical reaction (Figure 2.4B), where the protein molecules aggregate in an ordered manner to form crystal nuclei. Even though it appears to be a straightforward process, obtaining well-diffracted crystals is a trial-and-error process that often proves unsuccessful. So why are crystals needed?
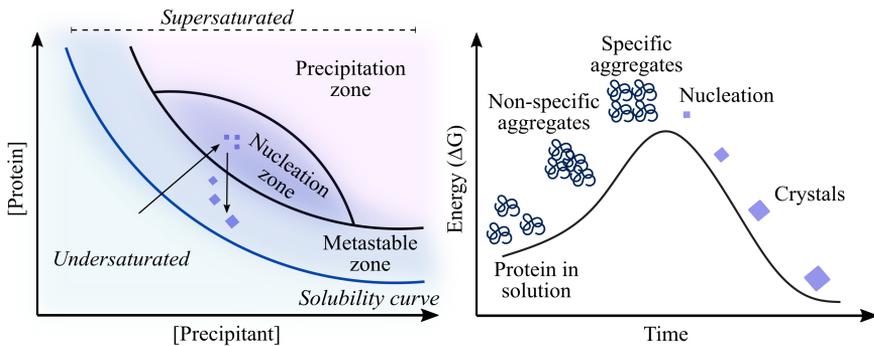
Figure 2.4. *Phase diagram.* **A**. Crystallisation phase diagram. Under the undersaturation condition, the amount of protein and precipitant is so small that the solution remains in a single liquid phase (clear drops). The solubility curve consists of the region where a crystal is in equilibrium with the solution (saturation). The other zones are considered to be supersaturated zones and differ by the protein:precipitant ratio. Nucleation is a zone where supersaturation is high enough for crystal formation. The metastable zone is the perfect zone for growing crystals once crystal formation has started (nucleation). The precipitant zone is where the protein concentration is too high and produces amorphous precipitation (non-specific or unorganised aggregation). The arrows represent the crystal formation steps. **B**. To initialise crystal formation (nucleation), the protein concentration needs to overcome a similar energy barrier (specific aggregation or organised aggregation) to a chemical reaction. [122, 123]

2.5.2 *X-ray diffraction theory*

When a protein solution (non-ordered sample) is irradiated by X-rays, it scatters as waves with different directions and intensities. The total intensity of a wave in a specific direction consists of the interference of constructive and destructive waves. This means that the intensity obtained from either a single molecule or a protein solution is not strong enough to obtain high-resolution data. However, the scatter signal is amplified when a crystal is irradiated by X-rays. Notwithstanding, many interferences are destructive under specific conditions (when Bragg's condition is met) and scattering waves

involve a constructive interference (coherent) by allowing Bragg peaks or reflections to be collected. This phenomenon is called diffraction [119, 121].

Bragg's law and Ewald construction

Bragg's law (Eq. 2.4) was published by W.H. Bragg and his son L. Bragg in 1913 [124]. It explains how diffraction occurs. These authors considered diffraction to be a reflection of X-rays caused by sets of equivalent and parallel planes of atoms in a crystal (Figure 2.5). When reflected rays are in phase (n is integer), a constructive interference of the reflected waves occurs and leads to diffraction [119, 121, 125].

$$n \cdot \lambda = 2 \cdot sin\theta \qquad (2.4)$$

where:

d = distance between planes in the lattice.
θ = angle of the incident and scatter X-ray beam.
n = an integer.
λ = the wavelength of X-ray beam.

Each Bragg's peak (hkl) corresponds to the diffraction from a set of crystal planes defined by Miller indices, hkl (integers). Its intensity is proportional to the electrons present in that plane. The position of reflections depends on the crystal lattice (the space group and cell dimension) and their intensity provides information about the content in the unit cell [119]. The diffraction pattern of a crystal is a representation of the reciprocal lattice, which has the same Laue symmetry as the real lattice. By knowing the reciprocal lattice of a crystal, the real lattice can be calculated using Fourier transform. Ewald's sphere is a geometrical construction used to explain the

relation between the reciprocal and the real lattice in crystal diffraction and is described in Figure 2.5 [121, 125].
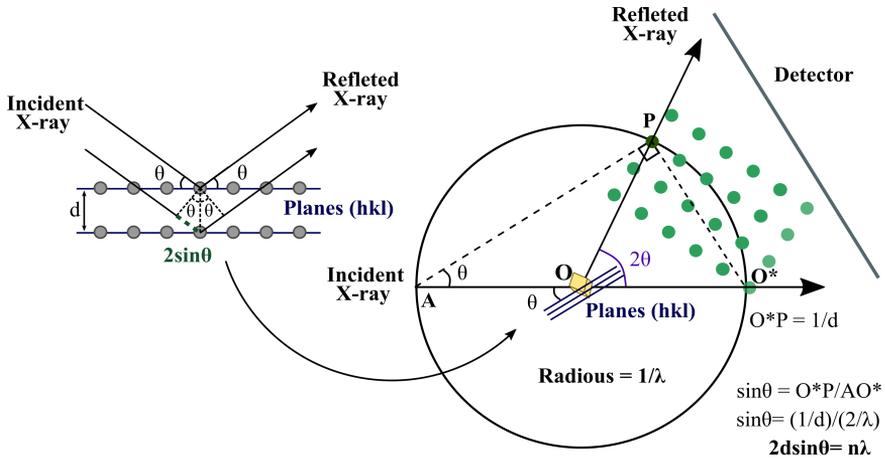
Figure 2.5. *Left*. Representation of the Bragg's law. *Right*. Ewald's sphere construction It is a geometric construction with radius $1/\lambda$ that relates the real lattice (crystal planes (hkl), blue) and the reciprocal lattice (green), and theoretically explains the diffraction of a crystal. This 2D representation represents a set of planes (hkl, blue) in the real space (crystal) with a plane separation, d. The real lattice origin and the reciprocal lattice origin are represented by O and O*, respectively. When the crystal is irradiated by an incident X-ray beam (AO*) with wavelength λ and Bragg´s law is satisfied, the reflected rays that produced diffraction (OP) will cross the Ewald sphere at a specific point of the reciprocal lattice, P. The reciprocal vector O*P is normal to the specific set of planes (hkl) and presents a length of 1/d. By using the triangle's properties, Bragg´s law can be extracted from this representation. Crystal rotation also rotates the reciprocal lattice and allows more reciprocal lattice points to cross the Ewald sphere. [121]

Each reflection (hkl) is described as the sum of the contributions of all the scatters (atoms) in the unit cell, and can be computed by the structure factor equation (F_{hkl}). The structure factor of a specific reflection (hkl) depends on the electronic properties of the atoms in the unit cell (fj), the amplitude of the contribution (hkl) and their position in the unit cell (*xj, yj, zj*) (phase) (Eq. 2.5). At a specific position of the unit cell (x,y,z), the electron

density (ρ(xyz)) can be calculated by the Fourier transform of the structure factor using Eq. 2.6 [119, 125].

$$F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i}(hx_j + ky_j + lz_j) \tag{2.5}$$

$$\rho(x,y,z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| e^{-2\Pi i(hx+ky+lz)+i\phi(hkl)} \tag{2.6}$$

where:

V = unit cell volume.

$|F_{hkl}|$ = amplitude of the structure factors.

$\phi(hkl)$ = phase.

Phase problem

One of the main difficulties of X-ray diffraction is the *phase problem* [119, 121]. During diffraction, the intensities of the reflected waves are collected, but their phases are lost [126]. The phase is essential to determine the structure factor and, from it, the density map that leads to solve the structure.

Phase provides more important information than what intensity gives to solve a crystal structure. One good example is the "animal magic" image published in Kevin Cowtan´s Book of Fourier, where the phases and intensity of two images (one duck and one cat) are mixed and, by inverse Fourier transform, the resulting images are dominated by the phase used in each case [119, 126]. The phase problem can be solved by different methods, but can still sometimes be a challenge.

In small molecules with a few atoms in the unit cell, direct methods work well to obtain the phase. Direct methods assume that there are statistical phase relations between certain sets of structure factors and exploit these relationships to calculate the initial phases [119, 126]. In proteins, thousands of atoms are often found in the unit cell, which makes these relations weaker. All this very much complicates the initial phase calculations. However, such methods are very useful for determining the initial phase of the marker atom substructure in experimental phasing methods, such as isomorphous replacement and anomalous dispersion [119, 126].

Patterson method was introduced in 1934 by Arthur Lindo Patterson and was the first direct method to obtain phase information by using only experimental information (intensities) [119, 125]. This method consists of the Patterson function calculation ($P(u,v,w)$), based on the autocorrelation of the electron density, which is computed directly from the intensity recorded by Fourier transform with no phase information (Eq. 2.7). The Patterson map is built by the maxima of the Patterson function, and contains the interatomic distance vectors between the atoms in the unit cell. The maxima of the Patterson function are directly proportional to the implied atomic numbers, which makes the Patterson map suitable for detecting "heavy" atom substructures. This map is used as part of other methods to determine heavy atom positions, in non-crystallography symmetry searches and map averaging, and also in the orientation step of a model search in molecular replacement [119, 127].

$$P(u,v,w) = \frac{1}{V} \sum_{-\infty}^{\infty}{}_{hkl} |F_{hkl}|^2 \cdot cos2\pi[hu + kv + lw] \qquad (2.7)$$

Molecular replacement is one of the commonest methods used to solve the structure of macromolecules, and it requires a model that is structurally

related to the target molecule in the crystal [119, 127]. This method consists of placing the atomic model phases from a related known structure in the unit cell of the unknown structure to be used as initial phases for map reconstruction. The calculation is done in the Patterson space and involves six dimensional searches over all the possible orientations (rotation) and translation by considering the intramolecular and intermolecular vectors, respectively. This method is relatively straightforward for solving macromolecule structure, but has been limited to the macromolecules with structural relatives. Arcimboldo's approach extends these limits by using small model fragments like small a-helices or coiled coils to obtain initial phases in protein structures [128, 129].

2.5.3 *Anomalous dispersion*

Before explaining how anomalous dispersion techniques work, it is important to introduce some reflection symmetry concepts.

Reflections can be classified as centric and acentric reflections. Centric reflections are pairs of reflections that are centrosymmetrically related through Laue symmetry (space group specific). These special reflections are found only in centrosymmetrically space groups, their intensity is not affected by anomalous scattering and the phase is restricted. However, acentric reflections are pairs of reflections that are non-centrosymmetrically related by Laue symmetry and have no phase restriction [119, 125, 130].

Friedel's pair is formed between the acentric reflections that are related by centrosymmetry inversion: F(h,k,l) and F(-h,-k,-l). Their amplitudes are the same, but phases are opposite. This is described as Friedel´s law. Bijvoet pairs are acentric reflection pairs generated by applying the point group symmetry operator to one of Friedel's pair mates. Friedel's pairs are a special

pair of Bijvoet's pairs. For example, having the Friedel pair hkl (F+) and -h-k-l (F-) and applying a 2-fold symmetry operator to Friedel's mate -h-k-l (F-), the reflection obtained, h-kl (F-), is Bijvoet´s pair mate of the hkl (F+) reflection. According to this example, reflection hkl (F+) has two Bijvoet's pair mates: –h-k-l (F-) and h-k-l (F-). In addition, the hkl and –h-k-l pair is called Friedel's pair [119, 125, 130–132].

Anomalous dispersion techniques (Single/Multi-wavelength Anomalous Dispersion; SAD and MAD) are the most widely used methods for *de novo* structure determinations [119, 126, 133]. These techniques are based on the capacity of atoms to absorb X-rays at a certain wavelength (by promoting an electron transition from a lower energy level to a higher one). The change in absorption in accordance with the wavelength is called the absorption edge, and is characteristic of each atom. An atom exhibits anomalous dispersion when the incident X-ray wavelength comes close to its absorption edge. The standard atoms present in biomolecules do not contribute to anomalous dispersion because their absorption edges are far from the X-ray wavelength used in X-ray crystallography (apart from S, which can also be used). However, the presence of heavy atoms in the structure produces anomalous X-ray dispersion, which is used to solve the phase problem [119, 126].

When an anomalous dispersion effect occurs for certain atoms in the crystal, the phase and the intensity of Friedel's pairs shift. Friedel's law is broken and the intensity for each mate is no longer equal (Figure 2.6. Anomalous dispersion methods use this anomalous intensity difference to localise the scatter atom on the electron density map by direct methods. The rest of the structure can be determined from this initial phase [119, 125, 126, 133].

However, different effects can also affect diffraction intensities, which makes it difficult to discriminate the intensity difference that belongs to the atom absorption and to obtain accurate phases. Radiation damage is one of

the commonest effects [119, 125, 126, 133].

The effect of anomalous scattering on a given structure factor behaves as a complex number composed of two correction terms: a real number (*f'*; dispersive term) and an imaginary number (*f"*; absorption term) Eq. 2.8. The dispersive term (*f'*) modifies the normal scattering factor (f_0), whereas the absorption term (*f"*) is 90º advanced in the phase [119, 125, 126].

$$f = f_0 + f' + if''$$ (2.8)

where:

f = anomalous scattering structure factor
f_0 = contribution of the normal scattering factor
f' and f'' = contribution of anomalous scattering factors

Having determined the heavy atom substructure, the calculated amplitude and phase of this contribution can be represented by Harker construction (Figure 2.6). In a SAD experiment, there are two possibilities for the phase of each reflection. This phase ambiguity needs to be broken to determine the structure by using the MAD technique or density modification and direct methods [119, 126].

Density modification considerably improves initial phases and electron density maps, which are useful for determining the correct phase from a SAD experiment. It consists of a group of different techniques that modify the electron density map according to the common features that a protein electron density map has such as the fact that it should be possitive. Another approach is the solvent flattening. A protein crystal contains approximately half the volume occupied by the ordered protein molecules and the other half with a solvent. The solvent is disordered and should not present any features

No anomalous dispersion

$|\mathbf{F}_H+|$

$|\mathbf{F}_P+|$

$|\mathbf{F}_{PH}+|$

$|\mathbf{F}_{PH}-|$

$|\mathbf{F}_P-|$

$|\mathbf{F}_H-|$

Friedel's law

$|\mathbf{F}_{PH}+|=|\mathbf{F}_{PH}-|$

Anomalous dispersion

$|\mathbf{F}_H+|$

$|\mathbf{F}_P+|$ f''

$|\mathbf{F}_{PH}+|$

$|\mathbf{F}_{PH}-|$

$|\mathbf{F}_P-|$ f''

$|\mathbf{F}_H-|$

Breaking down Friedel's law

$|\mathbf{F}_{PH}+|\neq|\mathbf{F}_{PH}-|$

$F_{PH}-$ $F_{PH}-$

f'' F_H

$F_{PH}+$ f'' $F_{PH}+$

Harker construction SAD
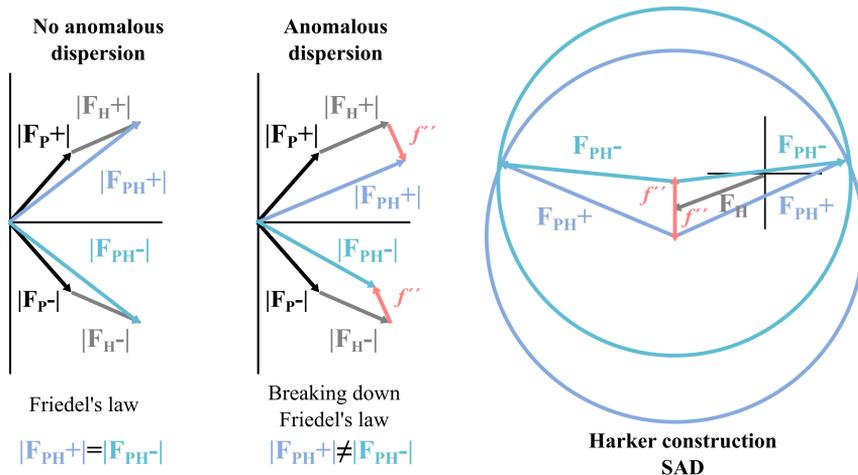
Figure 2.6. *Left*. Representation of Friedel pair's amplitude (|F|) in the absence and presence of anomalous dispersion. FP, FH and FPH correspond to the structure factor amplitude of the protein, the scattering atom, and the protein+scattering atom, respectively. f" is the contribution of anomalous scattering and +/- symbols correspond to the phase direction. *Right*. Harker construction for a hypothetical SAD experiment. It illustrates the phase ambiguity of protein structure factor (FPH), even when the scattering atom substructure (FH) is known [126].

on the electron density map. By bearing this in mind, accurate phases can be obtained by flattening solvent regions [119, 125, 132, 134].

2.5.4 *Data collection*

Data collection is one of the most important parts of crystallography experiments and successfully solving the structure depends on it. Protein crystal diffraction data are often collected mainly in large facilities called synchrotrons. Synchrotron science relies on the physical phenomenon of when a charged particle is accelerated, it emits EM radiation. EM radiation is produced as electrons pass through a magnetic field. The relativistic speed of electrons leads to the high flux and brilliance of X-rays because EM

radiation goes through Lorenz transformation. This is necessary to obtain good diffraction data in reasonable exposure times, compared to in-house sources [135].

A synchrotron facility consists of an electron gun, linac (lineal accelerator), a booster ring (a small ring where electrons start to be accelerated, optional), a storage ring (where electrons are accelerated to almost reach the speed of the light) and beamlines (where X-rays are focused, possibly monochromatised, and delivered to the experimental hutch). The storage ring contains bending magnets and insertion devices (wigglers and undulators) that accelerate and direct electrons [121].

Once protein crystals are obtained, they are harvested ("fished") with small loops and mounted onto the goniometer to be oriented and exposed to X-rays. Diffraction collection can be carried out at room temperature, but cryo-cooling the crystal in liquid nitrogen (approx. 100K) is more commonly used to avoid radiation damage (by ionising X-ray radiation).

To be able to solve a crystal structure, all the unique reflections in the reciprocal lattice must be collected. This can be achieved by rotating the crystal during collection. The total rotation depends on the symmetry of the crystal. However, longer exposure times increase radiation damage by creating disruptions to the crystals lattice, as well as chemical changes that reduce the intensity and quality of data. Cryo-cooling, including cryo-protectants (e.g. glycerol, PEG...), cuts exposure times (collecting only minimum angles to ensure that all unique reflections are collected, which depends on the crystal symmetry) or reduces transmission, which are ways to diminish damage by radiation [121].

Nowadays, almost all synchrotrons use Single-Photon Counting (SPC) pixel detectors (e.g. Pilatus or Eiger detectors) for protein crystallography

collection instead of Charge Coupled Device detectors (CCD) [136]. These detectors consist in an arrangement of several modules (with Hybrid Photon Counting Technology) and are characterised by having faster readout times, single-pixel point-spread functions and good signal-to-noise ratios. They are useful for fine-slicing data collection and time-resolve experiments.

In **paper IV**, two data collection types (continuous rotation and inverse-beam geometry collection) are described and analysed.

2.5.5 Data processing

Data processing involves the following steps: indexing, refining the unit cell and detector parameters, integration, scaling and merging [119, 137].

Indexing consists of finding the reflection spots of the diffraction pattern and assigning a consistent set of three integers (or Miller indices) to extend the reciprocal lattice and determine the crystal symmetry to obtain information about cell parameters, the space group and the crystal's orientation. The integration step consists of measuring the intensity of each reflection by considering not only the background signal, but also the crystal and detector parameters [137].

The data reduction step consists of scaling and merging. Scaling places all the data (different images, different crystals datasets, etc.) on a common scale by avoiding experimental variations, such as X-rays intensity, beam decay, variation in diffracting volume, exposure time and radiation damage. Afterwards, outliers are removed and intensity data, which are obtained from the different observations made of the same reflection, are averaged. The unique dataset of the averaged intensities is converted into structure factor amplitudes [137].

Different indicators exist to evaluate data quality [137, 138]. Completeness is the percentage of the number of unique reflections recorded versus the total number of reflections. R_{merge} is a traditional internal data consistency measure, which measures the spread of independent reflection intensity measurements. I/sigma represents the ratio between the mean intensity and its standard error. $CC_{1/2}$ measures the correlation coefficient between random half datasets. Redundancy is the ratio of the total number of reflections and the number of unique reflections. Anomalous diffraction experiments also include other quality indicators. An anomalous correlation ($Anomal_{corr}$) is the percentage of the correlation between the random halfsets of anomalous intensity differences. SigAno is the amplitude difference divided by the error (standard error of the anomalous amplitudes).

Many different programmes can be used for data processing (e.g. Mosflm [139], XDS [140], Aimless [138], etc.). In this thesis (**paper IV**), all the data processing steps were done using the XDS package (including XSCALE and XSCONV) [140].

2.5.6 *Model building, refinement and validation*

After data processing, it is necessary to determine phases, calculate the electron density map and build the initial structural model. Reliable structural models require further model refinement. Refinement consists of modifying the structural model, to obtain a better agreement between the structure factors calculated from the model, F_{calc}, and the structure factors observed from the experimental data, F_{obs} [119].

The F_{obs}-F_{calc} map depicts the regions where the model does not agree with the data. These maps represent positive and negative electron density areas. The positive electron density indicates that one part of the model is

missing in that specific area, while the negative electron density indicates that the model is not correct in that specific area. These map types are very useful for detecting errors in models. The structural model is built on a $2F_{obs}$-F_{calc} map as the model bias is less than on a F_{obs} map [119].

After each refinement cycle, different parameters are calculated to evaluate the progress of refinement. The commonest and most important is R value (R_{work}), calculated using Eq. 2.9 [119, 141].

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{2.9}$$

According to this formula, the difference between the experimental observations and the calculated ones (the model) diminishes when Rvalue (R_{work}) also decreases. This is an indication that the model better represents data. Cross-validation should be included to avoid model bias. This could be done by calculating the R_{free} value, the Rvalue from a small set of data that was not included during refinement. Another possible validation is the Ramachandran plot, a 2D plot that shows the regions allowed for amino acid residues according to dihedral angles ψ and ϕ of the protein backbone [119]. In this thesis, the Phenix package [142] was used to obtain initial phases (Phaser [143] and Autosol [144] for experimental phasing), automated building and density modification (Autobuild) [145], refinement (Refine) [146] and validation (MolProbity) [147]. Coot was also used to manually build the model in order to improve refinement [148].

2.6 Small Angle X-ray Scattering

Small angle X-ray scattering (SAXS) is a useful technique for obtaining low-resolution structural information of a macromolecule in solution

[149, 150]. It is often combined with other high-resolution structural determination methods, such as X-ray crystallography, NMR and electron microscopy (EM). In addition, this technique is most important for obtaining structural information about flexible or disorder proteins, which is difficult to acquire by other techniques [151].

For SAXS measurements, the sample solution (protein solution) is placed inside a quartz capillary and is irradiated with a collimated monochromatic X-ray beam (approx. 0.1 nm). Scattered rays are recorded in the detector by producing a scattering pattern of the sample (Figure 2.7). As the X-ray diffraction section previously mentions, when a sample solution is irradiated with an X-ray beam, waves scatter in different directions given the random orientation of the molecules in solution. This produces an isotropic scattering pattern that can be recorded and radially averaged. Scattering intensity (I) is represented according to the module of the momentum transfer,q, (Eq. 2.10), which depends on the scattering angle (2θ) and beam wavelength (λ) [149, 151].

$$|q| = \frac{4\pi}{\lambda} sin\theta \tag{2.10}$$

where:

q = scattering vector
λ = beam wavelength
2θ = scattering angle

The X-ray scattering of a macromolecule in solution depends on the number of molecules in the specific analysed volume (concentration) and the excess scattering length density (represented as the contrast) ($\Delta\rho(r)$). SAXS is a contrast technique where the excess scattering length density comes from the difference between the electron density of the macromolecule

and the solvent ($\Delta\rho(r) = \rho_p(r) - \rho_s$). Therefore, a standard SAXS experiment consists in two measurements: one of the macromolecule (protein) in solution and another of the solvent (buffer) (Figure 2.7) [150, 151].
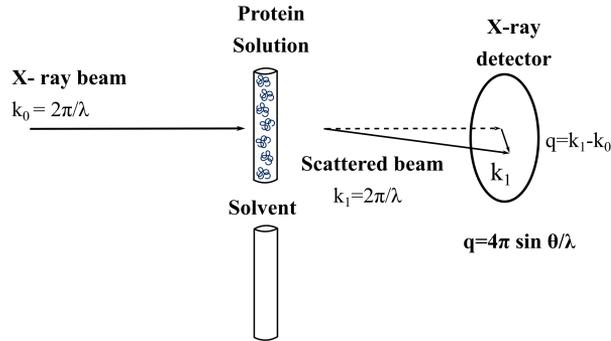
Figure 2.7. *Schematic illustration of SAXS experiment of molecules in solution [151].*

The excess scattering length density ($\Delta\rho(r)$) is related with the scattering amplitude of particles (A(q)) by Fourier transform (Eq. 2.11) .

$$A(q) = \int \Delta\rho(r)e^{(iqr)}dr \qquad (2.11)$$

where:

q = scattering vector

r = interactomic distance

V = particle volume

$\Delta\rho(r)$ = excess scattering length

In an SAXS experiment, what is measured is the scattering intensity, defined as the product of the amplitude and its complex conjugate averaged over all the orientations (I(q)=A(q)A(q)*) [150].

The total scattering intensity recorded equals the form factor of the molecule (I(q), intramolecular interactions) times the structure factor of solution (S(q), intermolecular interaction, repulsion and attraction); Itotal(q)=I(q)S(q). The form factor is that which provides information about the size and shape of the molecule. Therefore, it is the one we are interested in measuring. To achieve this, one of the requirements of an SAXS measurement for protein solution is that the sample is monodisperse, and no intermolecular interactions are present (it allows us to assume that the structure factor of the solution is 1) [149].

From the scattering curve, different parameters related with the size (at low q values) and shape of the molecule (at higher q values) can be obtained, such as the radius of gyration (Rg), particle volume, maximum particle diameter (D_{max}) and molecular weight [149, 150].

Rg is the square root mean of the distance of all the electrons from the particle centre that provides information about the overall molecule size and, therefore, about its molecular weight and oligomerisation stage. This parameter can be extracted from the scattering intensity at q=0 using the Guinier approximation [152]. The Guinier plot is a linear representation (lnI(q) vs.q^2) of the low q region of the scattering curve that allows I(0) to be obtained. This linear representation also provides information about the quality and monodispersity of data. Lack of linearity might suggest intermolecular interaction (e.g. aggregation) or radiation damage in the sample [151, 153].

The maximum particle diameter (D_{max}) can be obtained from the distance distribution function (or pr function), which is a histogram of the interatomic vectors within a macromolecule. It can be determined by an indirect Fourier transform of the scattering intensity curve. Rg, I(0) and initial shape information can be obtained from this function [149, 151].

The flexibility and folding state of a protein can also be analysed by the Kratky plot ($q^2I(q)$vs q). where a bell shape curve represents a globular protein and a plateau curve suggests an extended or unfolded protein [150,151].

In this thesis, SEC-SAXS was used to characterise human survivin in solution and to compare the results with previously published structure data. Data were collected at the BM29 beamline [154] of ESRF in France in a Superdex-75 10/300 GL column. Table 4.1 in Chapter 4 describes the data collection and data analysis. The ATSAS 2.8.4 package [155] was used to analyse the scattering curve and obtain structural information. Primus qt [156] was employed to analyse the scattering curve and estimate the radius of gyration Rg, D_{max} and molecular weight from the Guinier plot and the distance distribution function p(r).

The DAMMIF [157] and DAMAVER [158] programmes were used to generate an *ab initio* shape model from the experimental scattering curve, which was superimposed to the crystallographic model of survivin. The CRYSOL [159] software was used to generate a theoretical scattering curve from the survivin crystallography data to be compared with the experimental scattering data.

Chapter 3

Bayesian inference

3.1 Background

The origin of statistics dates back to the distant past. There is evidence that its origins goes back to the Xia Dynasty (2070-1600 BC) in China, with the first Chinese censuses. Until the 17th century, statistics was almost descriptive, and a system of data enumeration and organisation. Today, the use of statistics is much more extensive (from mathematics to social sciences). Statistics can take two different approaches: frequentist (classic) and Bayesian, depending on the concept of the adopted probability [160]. However, Bayesian statistics had emerged much earlier than the frequentist type, which developed much later due to lack of computers. Bayesian statistics presents complex mathematical problems, which require the use of computers with a high computation capacity, and also big data software [161, 162].

Statistical inference is the set of statistical techniques that allows us to draw conclusions about a population, usually from a sample of that population.

To be able to understand how frequentist and Bayesian inference work, and their differences, a small journey through the probability theory and concepts is included.

Probability focuses on assigning numbers to uncertainty, and it has not only encroached science, but also culture, and the way that society thinks. Thomas Bayes (1702-1761) was the first person to introduce the Probability Theory based on previous observations, and established the famous Bayes Theorem (Eq. 3.1), which was refined and published by Richard Price in 1763 [162–165].

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3.1}$$

where:

$P(A)$ = probability of event A occurring.
$P(B)$ = probability of event B occurring.
$P(A|B)$ = conditional probability of event A occurring once event B has already occurred.
$P(B|A)$ = conditional probability of event B occurring once event A has already occurred.

This theorem is the basis of Bayesian inference. Frequentist inference, however, was constructed as an independent science at the beginning of the 20th century, driven by the works of the English-born Karl Pearson and Ronald A. Fisher [166], and it has been, and still is, widely used [167].

The beginning of the modern Probability Theory lies in the correspondence between two French mathematicians, Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1665). In 1713, Swiss mathematician Jacob Bernouilli

established the "*Law of large numbers*" which states that the relative frequency of an event stabilises around a number after the number of observations of the event increases. This approximate number is defined as the frequentist probability of an event occurring. In his book "*Théorie analytique des probabilités*", published in 1774, Pierre Simon de Laplace defines the probability of an event occurring as the ratio of the number of favourable cases divided by the number of possible cases (when nothing allows us to expect that one of these cases should occur more than any other; i.e., they are equally possible). This way to define the probability of an event was used to develop statistical inference in the 20th century; frequentist inference [167].

The axiomatic definition of probability was established by Russian mathematician Andréi Nicoláyevich Kolmogórov (1903-1987) based on the properties of relative frequencies. From these axioms some probability theorems have been deduced Figure 3.1, including Bayes Theorem Eq. 3.1 [167].

Marginal Probability	**Union Probability**	**Conditional probability**	**Joint Probability**
The probability of "A" occurring. **P(A)**	The probability of "A" or "B" occurring. **P(AUB)= P(A)+P(B)-P(A∩B)**	The probability of "A" occurring given that "B" has already occurred. **P(A|B) =P(A∩B) / P(B)**	The probability of "A" and "B" occurring simultaneously. **P(A∩B)=P(A|B)*P(B) P(A∩C) = 0**
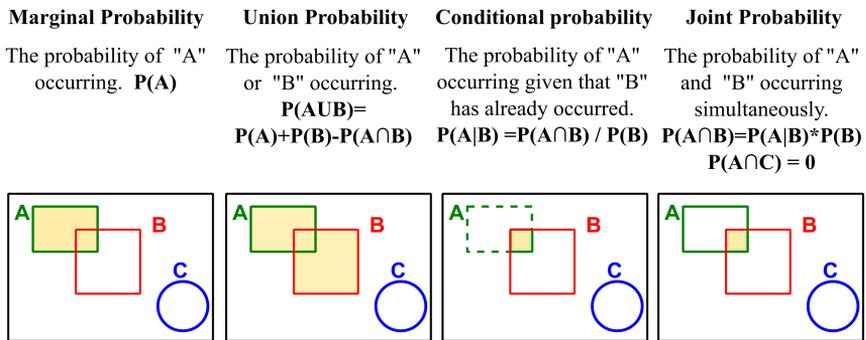
Figure 3.1. *Basic concepts of probability*.

Probability and likelihood are two terms that can be confused. Likelihood is not a probability, but is proportional to probability. In fact likelihood does not obey all probability rules (e.g. likelihoods do not need to add up to 1) [168–170].

3.2 Frequentist inference

Frequentist statistics calculates the probability of an event occurring, or not, by the repeated sampling of an experiment, and by taking the limit of relative frequencies as the probability of the event. For example; if an experiment is repeated 100 times and 20 of these times fail, the frequentist probability of failing (*P(F)*) is 0.2. In frequentist inference, parameter values are normally unknown and fixed, and data are random. Parameter estimation is based on choosing the values with the highest likelihood [161, 162, 165, 168, 171].

One of the main inconveniences of frequentist inference consists in needing to have a large number of observations to obtain a good estimation. Unfortunately in science, as in other disciplines, this is not always feasible and can be limited by many factors, such as the amount of sample, time or budget.

3.3 Bayesian inference

Bayesian inference is based on improving the predictions that an event will occur by adding new information or evidence. It is a statistical approach used to obtain sharper parameter estimations when the number of observations is not big enough. In this case, data are fixed and parameters are unknown random variables whose probability distribution is calculated by Bayes' Theorem [161, 162, 168, 171–173].

First of all, the studied parameter (or parameters) should be defined (θ), and its/their marginal probability should be calculated including prior beliefs (*prior* probability), without considering experimental data. Secondly,

the likelihood function of experimental data (D), including the previous beliefs of the parameter, should also be quantified. For parameter estimations, the marginal probability of experimental data should be a proportional constant as data are fixed. The *posterior* probability can be calculated by combining the *prior* probability and the data likelihood function using the proportional expression Eq. 3.2 [161, 162, 171, 173].

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta) \qquad (3.2)$$

where:

$P(\theta|D)$ = probability of θ value given the D *Posterior distribution*.
$P(D|\theta)$ = The likelihood of θ given the obtained D. *Likelihood function*.
$P(\theta)$ = marginal probability of θ. *Prior probability*.

Bayesian inference uses credible regions (CR) instead of confidence intervals (CI) (frequentist inference). CR are calculated from the posterior density function. Even if CR look very similar to CI, they essentially answer different things. A 95% CR states that, given the observed data there is a 95% probability that the true parameter value falls inside the credible region. A 95% CI states that if an experiment is repeated 100 times, approximately only 5 of the times will the parameter value fall outside the confidence interval as the experiments are all done in the same way [161, 162, 173, 174].

This approach has been used in **papers I**, **III** and **IV** to estimate different parameters that enhance the merging step of X-ray crystallographic data reduction (**paper I/IV**), which better estimate the binding curve and binding kinetics parameters from the MST binding assays (**paper III**).

In **paper I**, the multivariate Bayesian approach is used to estimate the structure factor amplitudes of a pair of observations from a specific reflection, and to calculate their differences (from synthetic data). This approach can have a major impact on the data analysis of the experimental phasing for X-ray crystallography (**paper IV**) or on the pump-probe diffraction experiments, where the calculation of accurate amplitude differences is essential for further analyses. In conclusion, with **paper I** multivariate analyses provide a better structure factor amplitude difference estimation than univariate analyses, which can strongly influence the results, as shown in **paper IV**. Figure 3.2 describes the multivariate Bayesian approach used in **papers I** and **IV**.
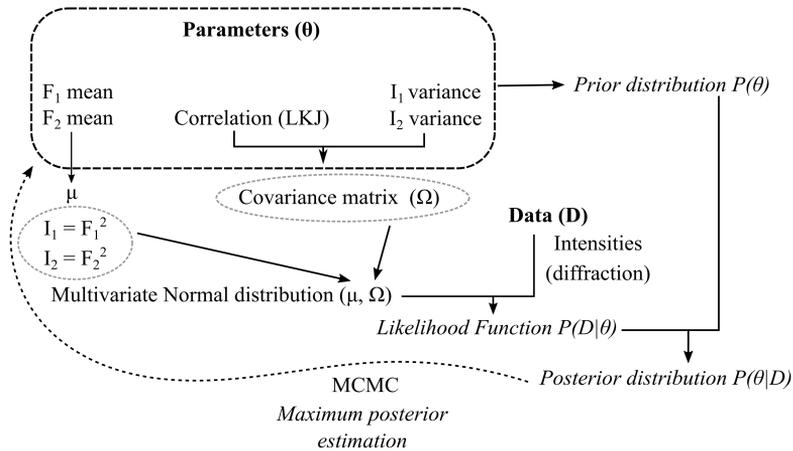
Figure 3.2. *The multivariate Bayesian approach for structure-factor amplitude estimations (**paper I**). I and F represent the structure factors' amplitude and intensity, respectively. LKJ is the (Lewandowski, Kurowicka and Joe) log-likelihood correlation [175]. MCMC is the Markov chain Monte Carlo method.*

In **paper IV**, the multivariate Bayesian approach, described in **paper I**, is applied to estimate the anomalous difference in the experimental SAD data, (collected from lysozyme and survivin crystals), and is compared with the commonly used univariate approach. According to these results (**paper**

IV), a multivariate approach enhances experimental phasing. With survivin, this was essential to obtain initial phases and to solve the structure. Moreover, two different data collection methods (continuous rotation and inverse-beam geometry) were tested. According to the results, better phases were obtained by the continuous rotation method. However, as the data quality of continuous rotation was also slightly better that the inverse-beam geometry, it was not possible to accept that one method was better than the other.

In **paper III**, the multivariate Bayesian inference is used to estimate the parameters of the binding curves obtained from the collected MST data in the survivin interactions (with borealin and shugoshin proteins), and to compare this method with the standard non-linear least square (NLLSQ) regression method [176]. The NLLSQ method analysis consists of a single point binding estimation curve which, according to the **paper III** results, appears to be considerably influenced by the outliers present in the data. The multivariate Bayesian inference approach represents the binding estimate curve as a multitude of curves corresponding to the joint posterior probability distribution calculated for each studied parameter. This approach showed a weaker impact by data outliers, which led to a better estimation of the binding curve and, therefore, of the binding kinetic parameters (e.g. K_D). Figure 3.3 shows a diagram of the multivariate Bayesian approach used in **paper III** to estimate the K_D of the binding.

In these papers, the power of Bayesian inference for obtaining a better estimation of the collected data was probed. However, it is important to be aware of the main pitfalls of Bayesian inference: prior distribution. As previously described, Bayesian inference relies on the prior beliefs of the studied parameter value to obtain a better parameter estimation from the collected data. If prior distribution is too conservative, experimental data can be censored and affect the posterior distribution calculation.
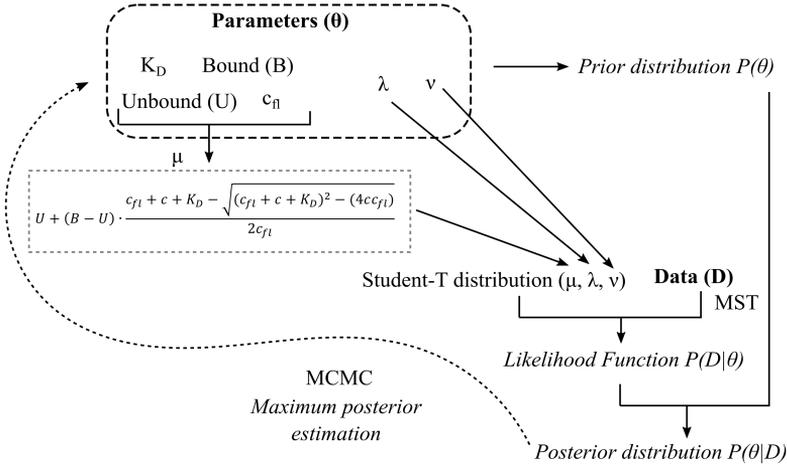
Figure 3.3. *Multivariate Bayesian approach for K_D estimation (**paper III**).* K_D, bound (B) and unbound (U) fluorescence, concentration of the fluorescent-labelled protein (c_{fl}) and the unlabelled protein concentration (c) are kinetic parameters. They are all random variables, except for the unlabelled protein concentration (c) that is constant. λ and ν are the parameters for the Student-T distribution, the degrees of freedom and scale parameter, respectively, and are also variable. MCMC is the Markov chain Monte Carlo method

Summary

The Bayesian inference approach was applied for the data analyses in **papers I**, **III** and **IV**. The main goal of using this approach is to obtain better parameter estimations from the collected data and to compare these results with previous analysis methods. This chapter includes an introduction to statistical inference, followed by a comparison made between Bayesian inference and frequentist inference, and the different approaches used in papers **I**, **III** and **IV**.

Chapter 4

Survivin interactions

Human survivin is the protein of interest in this work. In **paper III**, the interactions between survivin and human shugoshin proteins, hSgol1 and hSgol2, was analysed. Recombinant human survivin was overexpressed in *E.coli*, and the purified protein was characterised before the interaction and structural studies (**paper III**). This chapter focuses on the characterisation of human survivin protein and protein-protein interactions studies that led to the publication of **paper III**.

4.1 Human survivin production and characterization

4.1.1 *Survivin production*

Survivin was overexpressed in the *E.coli* BL21 star (DE3) strain using a modified pET28a vector (pHIS8) containing kanamycin selection and eight histidine tags (to facilitate protein purification), followed by a thrombin cleavage site (LVPR'GS) and the full sequence of the human survivin gene (BIRC5). This construct (vector) was obtained from Verdecia's lab [22] and was used for the protein expression of survivin for **papers III** and **IV**.

To obtain large amounts of protein, survivin expression was optimised on a small scale by varying the temperature (30ºC and 37ºC), IPTG concentrations (0.5, 0.75 and 1 mM) and induction times (2 h, 4 h and 6 h). High expression levels are not the only important fact in protein production optimisation, but protein solubility is also valuable. The protein expression level and solubility were analysed by SDS-PAGE electrophoresis by comparing the induced and non-induced samples, and the soluble (supernatant) and insoluble (pellet) fractions after cell disruption. Although all the tested conditions showed part of the protein in the insoluble fraction (inclusion bodies), the induction performed at 30ºC for 4 h also led to large amounts of soluble survivin protein (**paper III**).

Survivin production was performed according to the protocol published in paper III, including some of the improvements described below. Expression increased by the addition of 80 μM $ZnCl_2$ to the media before induction. Since survivin contains a zinc-binding site in the BIR domain, it is not surprising that the addition of zinc to media helps protein folding and leads to a bigger amount of soluble protein. Survivin was purified in two steps: affinity chromatography and SEC (**paper III**). Nickel affinity purification was ameliorated by including 20 mM imidazole in the binding buffer (50 mM Tris, pH 8, 500 mM NaCl and 20 mM imidazole), two washing steps at low imidazole concentrations before eluting (30 mM and 40 mM) and a final elution gradient (40 mM to 250 mM). In addition, overnight dialysis in SEC buffer (50 mM, 150 mM NaCl and 1 mM DTT) was an essential step for protein stability as the protein precipitated when exposed to high concentrations of imidazole over an extended period.

The presence of small tags in proteins (e.g. His-tag) does not usually interfere with their characterisation. However, as His-tags are commonly disorder regions, tag removal can improve crystal formation and packing (ordered structure), which leads to better diffraction and higher resolution

crystals. For crystallisation purposes (**paper IV**), the His-tag was cleaved off from survivin by incubating at 10-12ºC overnight with the thrombin protease (GE Healthcare). After cleavage, thrombin was removed by binding it to a HiTrap Benzamidine FF column (GE Healthcare) (Chapter 2). Crystal quality was also improved by adding an extra SEC purification as a final step.
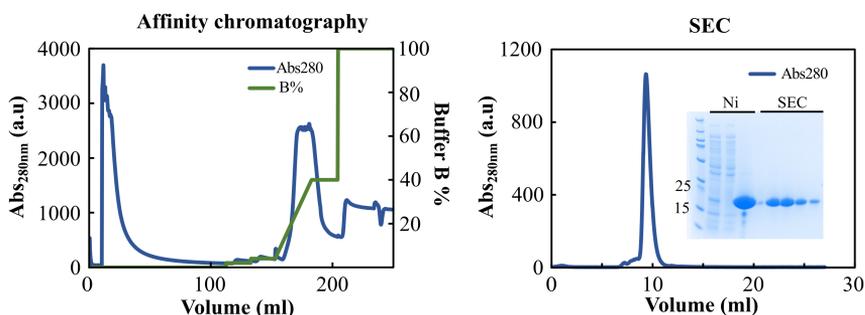
Figure 4.1. *Survivin purification profiles and SDS-PAGE gel..Left*. The IMAC survivin profile where 100% buffer B corresponds to 500 mM Imidazole. *Right*. The SEC survivin profile. The SDS-gel shows samples from the IMAC and the SEC purification. The IMAC wells correspond to the initial diluted sample, the diluted non-specific binding sample (0-100 ml) and the elution peak (approx. 160-200 ml), respectively.

Figure 4.1 provides examples of survivin purification and sample purity. According to these results, high-purity survivin samples can be obtained after nickel purification. However, the SEC step allows the elimination of small contaminants and more homogenous protein sample can be obtained Figure 4.1.

4.1.2 *Buffer screening*

Protein stability is an important factor for protein studies and may be strongly linked to the buffer conditions. A thermofluor assay (Chapter2) was

used to test different buffer conditions and to monitor the melting temperature (T_m) shift of survivin. Thirteen of the commonest buffers used in protein purification and crystallisation were selected and tested (Figure 4.2) [177]. These buffers cover a wide range of pH values as protein charge and stability are closed related to pH. Sodium chloride (NaCl) and dithiothreitol (DTT) were also added to increase the ionic strength of the buffer and to avoid damage by oxidation, respectively.
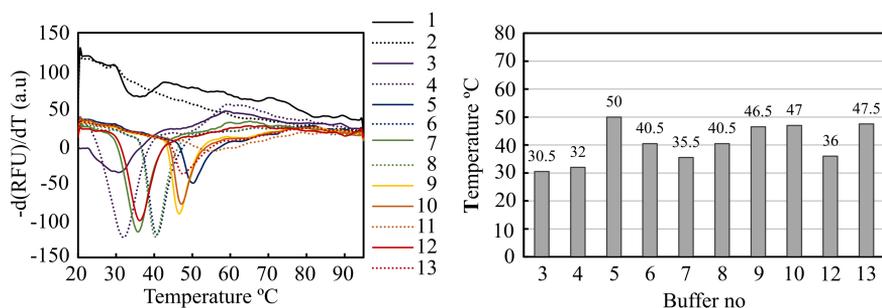
Figure 4.2. *Survivin buffer screening results using a thermofluor assay.* **A.** Representation of the first derivative of fluorescence emission (–d(RFU)/dT) in accordance with temperature, which was calculated from the initial melting curves using the Bio-Rad CXF software. B. Temperature melting of survivin in specific buffers. Buffers 1, 2 and 11 were not included as no proper information was obtained from the curves. The analysed buffers were: sodium citrate pH 5.0 (buffer 1), sodium acetate pH 4.5 (buffer 2), MES pH 6.5 (buffer 3), potassium phosphate pH 6.0 (buffer 4), sodium phosphate pH 7.5 (buffer 5), HEPES pH 7.5 (buffer 6), MOPS pH 7.0 (buffer 7), ammonium acetate pH 7.3 (buffer 8), Tris pH 8.0 (buffer 9), Tricine pH 8.0 (buffer 10), Bis-Tris propane pH 8.5 (buffer 11), PIPES pH 7.0 (buffer 12), Glycine pH 9.0 (buffer 13). All the buffers were prepared at a final concentration of 100 mM, including 150 mM NaCl and 1 mM DTT. Each condition was replicated 3 times and the results gave the average.

In line with these results, basic pH values (over 7) gave a higher T_ms than acidic pH values (below than 7), which suggests that protein stability would also be greater at a basic pH. Four of the tested buffers can be considered the most recommended ones, including sodium phosphate pH 7.5, glycine pH 9, Tricine pH 8 and Tris pH 8. As there are no large differences in the survivin

T$_m$ among these buffers, and as Tris buffer has been used satisfactorily with survivin in previous studies, further experiments were done in 50 mM Tris-HCl pH 8, 150 mM NaCl and 1 mM DTT. Phosphate buffer was not chosen because the initial purifications trial did not succeed with it.

4.1.3   *Survivin: dimer in solution*

Previous crystallography and NMR publications have described survivin as a homodimer, but the function of the protein in this oligomer state is still not fully clear [22–24]. Survivin homodimer has been previously related with anti-apoptosis function at the cytoplasm because it is well-known that during cell division, survivin acts as a monomer by forming part of the CPC in the nucleus. In this thesis, the size exclusion chromatography (Figure 4.1) and small angle X-ray scattering (Figure 4.3) results were combined to determine sample quality, and the size and oligomer stage of the survivin protein used in our experiments (**papers III** and **IV**).

The SEC result shows a homogenous peak around 40 kDa, which suggests that the survivin sample is a dimer (a theoretical monomer value of 18.74 kDa) and there was no obvious aggregation in the sample. These results were also confirmed by the SAXS analysis, as summarised in Table 4.1. The Guinier plot did not show any sample aggregation, and the obtained Rg and D$_{max}$ values were characteristic of small proteins. Molecular weight was also estimated to be approximately 33 kDa. The Kratky plot showed partial flexibility in the protein (Figure 4.3C), which could be associated with the linker between the BIR domain and the C-terminal.

The comparison made of the experimental and theoretical scattering curves (calculated from a crystallography model, PDB id 1F3H [22]) gave a χ^2 value that came closer to 1 ($\chi^2 = 0.93$). This result indicates a good

agreement between the protein crystal structure model and the protein in solution (Figure 4.3D). This was also represented by superposing the crystal structure with the *ab initio* model shape calculated from the scattering curve.
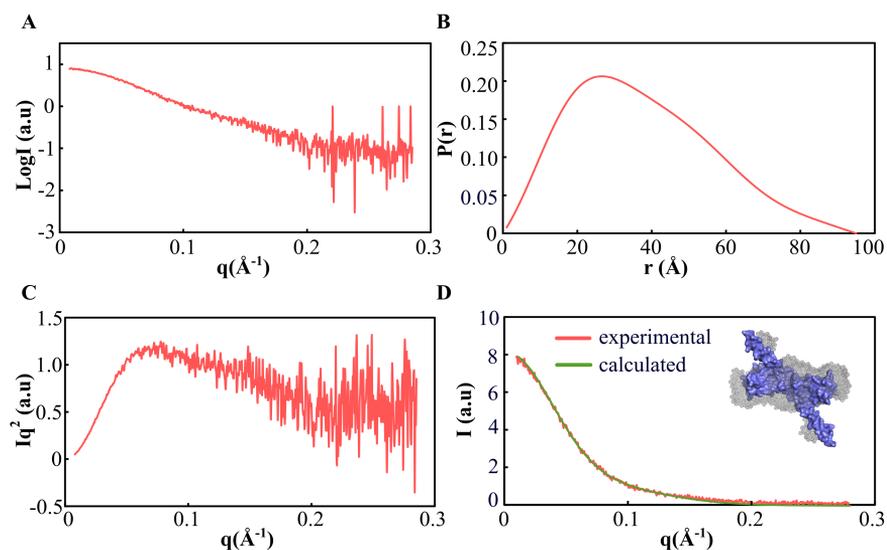
Figure 4.3. *SAXS analysis of survivin* **A**. Scattering curve (logI vs q(Å^{-1})). **B**. Distance distrubution function (pr function). **C** Kratky plot (Iq^2 vs q(Å^{-1})). **D** Comparison of the experimental scattering data (red) and the calculated scattering curve from the crystal structure PBD id:1F3H (green) [22]. Superposition of the 1F3H structure (blue) [22] into the *ab initio* envelop (grey) calculated from the SAXS data.

Table 4.1. SAXS results of survivin.

Data collection parameters	
Beamline	BM29, ESRF, Grenoble, France
Wavelength (Å)	0.99
q range (nm^{-1})	0.025-5
Injected sample (mg)	$\approx$3
SEC column	Superdex 75 100/300GL
Temperature (K)	283.15
Structural parameters	
I(0) (cm^{-1})[from Guinier]	8.05 $\pm$ 0.02
Rg (Å)[from Guinier]	28.7 $\pm$ 0.2
I(0) (cm^{-1})[p(r) function]	8.10 $\pm$ 0.03
Rg (Å)[p(r) function]	29.5 $\pm$ 0.3
D$_{max}$(Å)	95
Porod volume estimate, Vp (Å3)	56040
Molecular weight(MW) determination	
MW [from Porod volume (Vp/1.6)] (kDa)	35.02
MW (kDa) – Bayesian inference primus [178]	33.10 (31.30-34.95)
Modelling parameters	
Shape reconstruction	DaMMIF [157]
Symmetry	P2
No. of models averaged/total	20
DAMAVER NSD (var) [158]	1.26$\pm$ 0.14
Software employed	
Data evaluation	Primus qt, GNOM [156]
Computation of model intensities	CRYSOL [159]
3D graphics representations	UCSF Chimera [25]

4.2   Microarray peptide analysis

Paper III describes two new interaction partners of the human survivin protein, hSgol1 and hSgol2, which belong to the shugoshin-like protein family. These interactions were established by using shugoshin peptides and the specific sequence was determined by the microarray experiments described in **paper III**. In this microarray, the amino acid sequence of 19 different human proteins (complete or partial) (Table 4.2) was analysed in small 15 residue-long peptides with an overlap of 10 residues (**paper III**). There were three main reasons for selecting these proteins:

1. They are involved in similar cellular processes, such as survivin, cell division and programmed cell death

2. Previous studies have already shown interactions with survivin, which can be used as positive controls

3. They present similar amino acid sequences to proteins, which have already been known to interact with survivin

Table 4.2 summarises the different analysed proteins and their characteristics to be selected. The amino acid sequences of some were not fully studied, and only a few residues were included in the analysis. The truncated sequences are represented in the table by the number of residues included in the analysis, next to the protein name. The microarray results indicated two basic peptides as the principal interaction candidates. These two peptides correspond to the C-terminal region of the shugoshin-like protein family (hSgol1 and hSgol2) and were used to design the peptides for the MST analysis (**paperIII**). In addition, other peptides with weaker intensities were also selected for further analyses, such as borealin (**paper III**), BCL2, JADE-3 and BCL6.

Table 4.2. Proteins for the microarray peptide analysis.

Protein name	Characteristics/Function
p53	Involved in apoptosis and cell division
BCL-2	Involved in apoptosis
BCL-6	Involved in apoptosis
Aurora kinase A	Involved in chromosomal instability
Aurora kinase B	A CPC member
Smac/DIABLO$^{1-134}$	Involved in apoptosis (+ control)
Borealin$^{1-93}$	A CPC member (+ control)
Histone H3	Involved in the chromatin structure (+ control)
JADE-3$^{105-173}$	Sequence similarity with INCEPT interaction region
FLJ44060 isoform CRA_$b^{268-514}$	High sequence similarity with the borealin interaction region
LysM3$^{143-204}$	High sequence similarity with the borealin interaction region
hSgol1	Involved in cell division and previously studied
hSgol2	Involved in cell division and is hSgol1-related
Tubulin y	Involved in cell division (microtubules)
Tubulin a	Involved in cell division (microtubules)
Tubulin b	Involved in cell division (microtubules)
Anx13A intestine$^{94-149}$	Sequence similarity with the INCEPT interaction region
Anx13$^{102-141}$	Sequence similarity with the INCEPT interaction region
SYCP2L$^{582-628}$	Sequence similarity with the INCEPT interaction region

Even though many of these proteins do not show any interaction with survivin in the microarray assay, possible interactions taking place under other conditions cannot be ruled out. This technique analyses protein interactions using the small peptides attached to a surface. This means that any interaction requiring a protein secondary/tertiary structure, or is affected by the immobilisation of the peptide, is not reported. Some specific interactions can also depend on PTM in the proteins and are not reported either.

4.3 Borealin interaction with survivin

The borealin and survivin interaction has been previously studied as part of the CPC. In **paper III**, we chose borealin$^{6-20}$ (GSSRVAKTNSLRRRK) as a positive control of the interaction with survivin. This short peptide, including part of N-terminal borealin (residues 6-20), was selected from the microarray analysis and sufficed to weakly interact with survivin. It presented a 10.8 μM K$_D$ (**paper III**). Previous studies have reported that the main interaction between these two proteins included borealin$^{20-78}$ and the survivin region that followed the BIR domain and the first part of the C-terminal (homodimerisation region). However, no dissociation constant information has been reported.

The survivin and borealin$^{6-20}$ interaction can be involved to increase the stability of the previously described complex. The CPC structural studies have shown that the borealin$^{6-20}$ region is located at the end of the survivin C-terminal (Figure 4.4). However, only a few residues (LRRRK) from the borealin peptide have been modelled in these structures and are located at the end of the survivin C-terminal (139-141). This suggests that the rest of the peptide (6-14) can be folded around the survivin C-terminal.

4.4 Shugoshin and survivin interactions

4.4.1 *Microscale Thermophoresis*

The new interactions between hSgol1$^{291-312}$/ hSgol2$^{1066-1085}$ and survivin were validated by the MST experiments (**paper III**), which reported dissociation constants valued at 80.7 and 1.6 μM, respectively.
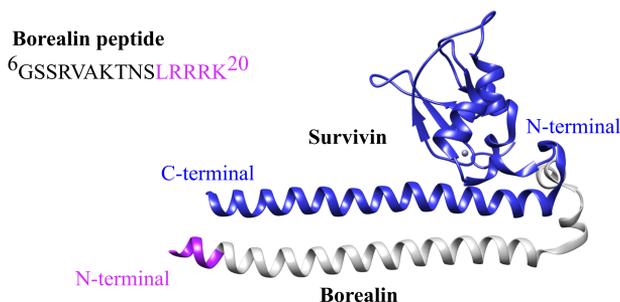
Figure 4.4. *The borealin and survivin crystal structure, PDB id 2QFA [30].* Survivin is shown in blue and borealin in grey and pink. The pink part of the helix shows the five last residues of the peptide used in the MST experiments (borealin[6−20]).

Previous studies have shown that survivin can interact with proteins by presenting an N-terminal that mimics the phosphorylated N-terminal of histone H3 (Ala1-Arg2-Thr3ph-Lys4). hSgol1 presents such an N-terminal region (Ala1-Lys2-Glu3-Arg4), and its interaction with survivin and with the CPC has been previously reported. The new interaction between hSgol1[291−312] and survivin (**paper III**) was more than 100 residues far from the previous hSgol1 N-terminal interaction, which suggests that this new interaction does not necessarily interfere with the previous one. On the contrary, hSgol2 does not have such an N-terminal, and no direct interaction with survivin has been previously reported. However, both shugoshin proteins have been required for the centromeric localisation of the CPC during cell division via the direct interaction with other CPC members. This scenario suggests that this new interaction with survivin might also be involved in the shugoshin:CPC complex.

The multiple sequence alignment of the shugoshin proteins C-terminal region from different mammalian species (**paper III**) showed that the peptide region was highly conserved between them, which could indicate its usefulness for some of their functions.

Peptides hSgol1$^{291-312}$ and hSgol2$^{1066-1085}$ are highly positive-charged and their secondary structure was predicted as coiled coil and α-helix, respectively. However, the results of the circular dichroism described in **paper III** revealed that their structure looked more like a random coil, and that hSgol1$^{291-312}$ presented a more local order than hSgol2$^{1066-1085}$.

4.4.2 *Survivin-eGFP and hSgol1$^{195-312}$ interaction using FSEC*

To improve our knowledge about the survivin and hSgol1 protein interaction, a longer peptide that included the previous peptide was designed, cloned and overexpressed, namely hSgol1$^{195-312}$. This region was selected by using the PHD secondary structure predictor tool [179] and the RADAR tool [180], which detected the repeating regions inside the protein sequence. The aim of this design was to extend the small peptide (hSgol1$^{291-312}$), including a specific secondary structure (coiled coils) and possible repetitive regions in the protein sequence. As this new hSgol1 peptide, hSgol1$^{195-312}$, has the molecular weight of a small protein (approx. 14 kDa), the complex formation between survivin and hSgol1$^{195-312}$ was preliminarily analysed using SEC.

When two proteins interact and form a complex, the molecular weight of this complex should be larger than the individual proteins, and elutes earlier from SEC by being able to easily analyse possible complex formation. As absorbance at 280 nm is common for almost all proteins, enhancer green fluorescence protein (eGFP, λ_{ex} 485 nm /λ_{em} 512 nm [80]) was fused to survivin to monitor the complex elution by fluorescence size exclusion chromatography (FSEC) by an FP-2020 Plus fluorescence detector. By considering that fluorescence is more sensitive than absorbance at 280 nm, FSEC permits small amounts of labelled protein (survivin-eGFP) to be used for detection. This has the advantage of the amount of unlabelled protein

(hSgol1$^{195-312}$) being several times bigger to ensure full complex formation (survivin-eGFP:hSgol1$^{195-312}$). Proteins survivin-eGFP and hSgol1$^{195-312}$ were expressed and purified by standard methods, and the specific protocols are described in Appendix B.

The complex formation was analysed by incubating survivin-eGFP (2.5 μM) and hSgol1$^{195-312}$ (68 μM) for 1 h at 12°C, injecting them into the FSEC and eluting at 0.4 ml/min in 50 mM Tris, pH 7.5, 300 mM NaCl and 1 mM DTT. Similar experiments were run with the individual proteins and a negative control using eGFP (2.5 μM) instead of survivin-eGFP. To estimate the molecular weight of the complex and the other proteins, previous calibrations using a Gel Filtration Calibration Kit (GE Healthcare) under the same buffer conditions were also done.

Earlier survivin-eGFP elution in the presence of hSgol1$^{195-312}$ was observed, which suggests complex formation (Figure 4.5A). Survivin-eGFP is the heaviest component of the complex with a theoretical MW of 46 kDa (vs. 14.7 kDa for hSgol1$^{195-312}$). According to the Superdex 200 calibration experiment (Figure 4.5E), survivin-eGFP eluted later than what its theoretical MW suggested (an apparent MW of 33 kDa). hSgol1$^{195-312}$ eluted at a volume corresponding to a protein with an MW of 32 kDa, which suggests that hSgol1$^{195-312}$ is in a dimeric form (Figure 4.5C). The simplest explanation for the apparent MW of the complex formed between survivin-eGFP and hSgol1$^{195-312}$ (49 kDa) is 1:1 stoichiometry. This faster elution was not observed in the negative control gel filtration experiment using eGFP (30kDa; the apparent MW in gel filtration was 26.5 kDa) and hSgol1$^{195-312}$ (Figure 4.5B). However, this interaction needs to be confirmed by other methods, such as MST.
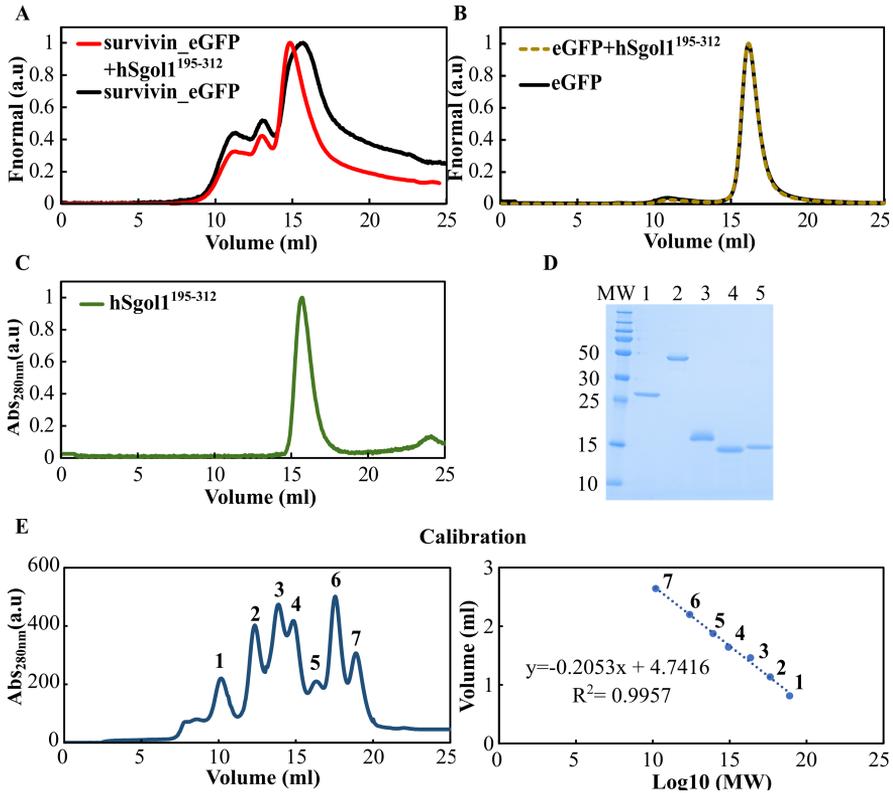
Figure 4.5. *FSEC experiment results.* **A.** Chromatography of the survivin-eGFP construct alone (black) and co-injected with hSgol1$^{195-312}$ (red). **B**. *Negative control.* Chromatography of the eGFP construct alone (black) and co-injected with hSgol1$^{195-312}$ (yellow dashed line). **C.** hSgol1$^{195-312}$ chromatogram. **D**. SDS – PAGE of the studied proteins. The numbers on the left correspond to the different molecular weights (kDa) of the protein ladder. The numbers at the top represent the different loaded samples; 1. eGFP (30kDa), 3. survivin-eGFP (46kDa), 3. survivin (16.4kDa), 4. hSgol1$^{195-312}$ (14.7kDa), 5. Trx-tag afterhSgol1$^{195-312}$ cleavage (15.9kDa). **E**. Superdex200 Increase 10/300GL calibration using the followed mixture of different proteins: ferritin (440kDa), aldolase (158kDa), conalbumin (75kDa), ovalbumin (44kDa), carbonic anhydrase (29kDa), ribonuclease A (13.7kDa), aprotinin (6.5kDa). They are labelled from 1 to 7, respectively.

4.5 Co-crystallisation trials of survivin and shugoshin peptides.

One important factor of studying protein-protein interactions is to know the specific region where the interaction occurs. This provides functional information that helps to better understand the cellular pathways where these proteins are involved. Cell division requires the interaction of many different proteins at the same time (e.g. the CPC formation), which makes it difficult to fully understand how it works. By knowing that survivin and shugoshin proteins interact, many other questions appear, such as how does this interaction occur? Do shugoshin proteins bind to the dimeric or the monomeric form of survivin? Can shugoshin bind survivin when it interacts with the CPC? Some of these questions can be answered by studying the atomic structure of the complex.

The co-crystallisation of survivin and $hSgol1^{291-312}$ and $hSgol2^{1066-1085}$ was tested by the sitting drop vapour diffusion approach and commercial crystallisation screens (e.g. PACT, Morpheus, etc.). The drops containing the complex solution (1 mM survivin + 5 mM shugoshin peptides, respectively) and the precipitant solution at a ratio of 1:1, were set up using a mosquito crystallisation robot on MRC2 plates. Plates were incubated at 20ºC for 1-3 weeks.

Crystals with different shapes and sizes were found under several conditions (Figure 4.6). Some of the conditions presented crystals that differed in size and shape in the presence of $hSgol1^{291-312}$ or $hSgol2^{1066-1085}$, respectively, which could indicate complex formation. These crystals were harvested, cryo-cooled and tested in ESRF. A few of them diffracted to high resolution (e.g. 2.5Å), but no peptide was found in the structure.
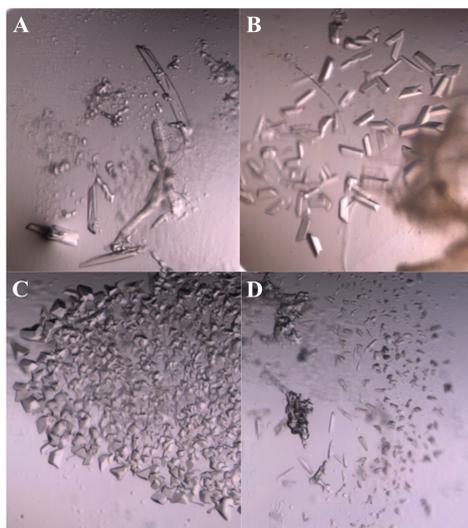
Figure 4.6. *Crystals obtained from the co-crystallisation of Survivin and shugoshin peptides.*A and B show the crystals from the condition 0.2 M Potassium sodium tartrate tetrahydrate 20% w/v PEG 3350 (PACT premier – Mol dimension), including peptides hSgol1 and hSgol2, respectively. C and D show crystals from the condition D3 0.1 M Potassium chloride 0.1 M Sodium HEPES 7.0 15% w/v PEG 5000 MME (Proplex – Mol. dimension), including peptides hSgol1 and hSgol2, respectively.

4.6 Survivin isoforms

The expression of survivin isoforms has been related to several diseases, and some studies have suggested that the formation of heterodimers with survivin can play an important role towards survivin regulation. However, no information is available about their structure, function or how they can interact with survivin. In this thesis, preliminary expression and purification studies have been done in the survivin-2B and survivin-ΔEx3 isoforms in order to obtain pure and stable proteins for structural and interactional studies.

Four different constructs were designed and tested for the expression of survivin isoforms. They included different localisations of the his-tag (pET28b+8xHis and pET29b) and two different protein fusion tags (pET48b and pET49b) to increase protein solubility. The survivin-2B and survivin-ΔEx3 genes were obtained from Mahotka's lab [56] in the pET29b vector, and the other constructs were generated using the InFusion HD Cloning kit. Two different *E.coli* strains were used to analyse the expression: One Shot BL21 Star (DE3) and Rossetta^{TM} (DE3) pLysS. This last one was selected to improve the expression of the eukaryotic proteins in *E.coli*. A small-scale test expression was performed according to the pET manual protocol [82]. Figure 4.7 summarises the best expression results for each construct in Rossetta^{TM} (DE3) pLysS .

Survivin-2B overexpression was significantly better than survivin-ΔEx3. However, the majority of the protein formed inclusion bodies. Survivin-2B expression gave small fractions of soluble protein using pET29b and pET28b+8His. Survivin-ΔEx3 was almost not expressed in any construct or strain, which suggests that either it may be toxic for cells or codon optimisation is needed. Survivin-2B purification using the his-tag construct was

not successful because the protein was unstable and did not bind properly to nickel affinity column. The survivin-2B fusion to TRX (pET48b) gave good purified protein yields, but survivin-2B was unstable after cleavage of the TRX.

Summary

The microarray peptide technology allowed us to easily identify some
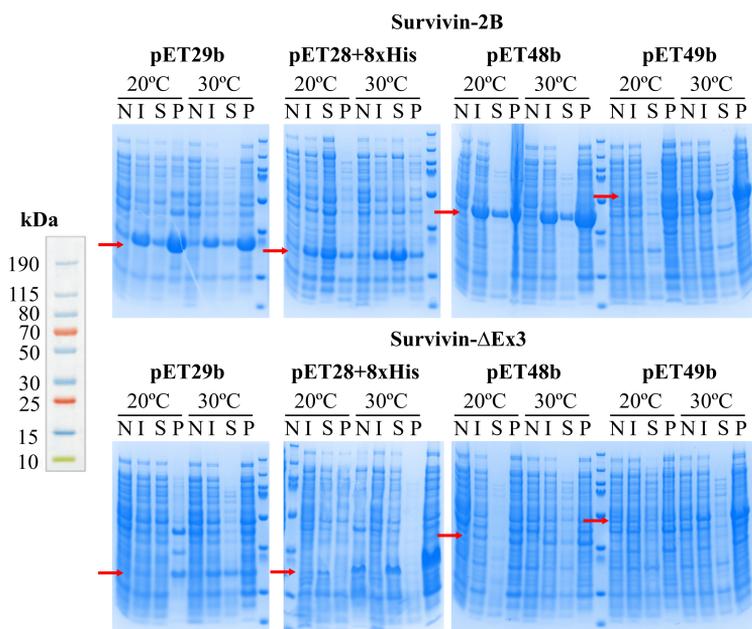
Figure 4.7. *The test expression of survivin isoforms in different pET vectors* (pET29b, pET28b, pET48b and pET49b). The top and bottom gels correspond to survivin-2B and survivin-ΔEx3, respectively. The different samples are described by; N: non-induced, I: induced, S: supernatant and P: pellet or inclusion bodies. The red arrow marks the position where the survivin isoform expression is expected. The molecular marker corresponds to the PageRuler Plus Prestained Protein Ladder (Thermo Fisher Scientific).

possible interaction partners for survivin. Microscale thermophoresis was used to evaluate these interactions, which led to two novel interactions between survivin and the shugoshin family, as published in **paper III**. Recombinant survivin isoforms appeared less stable than survivin, which suggests that testing codon optimisation or other expression systems may be necessary.

Chapter 5

Concluding remarks

The work presented in this thesis focuses on the discovery of new interaction partners of the human survivin protein to improve our understanding of its function and its involvement in different diseases. This work also shows the strength of Bayesian inference in data analyses by focusing on crystallography data and binding affinity calculations.

Paper I demonstrates how the multivariate Bayesian model yields more precise structure factor amplitude estimates from pairs of recorded diffraction intensities than the common univariate model. Structure factor amplitude difference calculations play an important role in solving experimental problems or for evaluating time-dependent structural changes in pump-probe experiments. Therefore, a better estimation can improve these analyses.

Paper II provides a detailed review of current knowledge about the non-malignant properties of survivin and focuses on its implication to autoimmune diseases. Survivin has been reported to be a promising biomarker and a therapeutic target in autoimmune diseases. In clinical applications, survivin measurements have been shown to be important for diagnostic and therapeutic treatments.

Paper III reports two novel interactions between human survivin and the shugoshin-like protein family. These interactions provide information to help us better understand how survivin or the CPC complex is localised at the inner centromere during cell division. As shugoshin proteins protect the cohesion of sister chromatids (cohesin complex) during cell division, the interaction with survivin at the centromere can also play a role in the removal of centromeric cohesin before chromosomal segregation. **Paper III** also demonstrates how the Bayesian inference analysis can greatly improve the calculation of binding curves and the dissociation constant estimation. This approach is less sensitive to the presence of data outliers than standard non-linear least square (NLLSQ) regression methods and provides an independent binding data analysis without involving any manual intervention, which helps avoid any human subjective bias.

Paper IV confirms that the multivariate Bayesian model described in **paper I** very much improves the anomalous difference calculations that are essential for SAD experimental phasing in difficult cases (e.g. survivin). It also suggests that as long as the X-ray beam flux is attenuated at a reasonable level, radiation damage does not apparently influence the chosen data collection approach (continuous rotation and inverse-beam geometry) and the results. An alternative pairing method that employs the multivariate Bayesian model concluded that pairing closer Bijvoet's reflections presents a much lower correlation than pairing Friedel's reflections, having an effect in the anomalous difference estimation.

Future view

The rapid development of high computational capacity machines have allowed the complex mathematical problems that Bayesian inference exhibits to be solved. Bayesian inference is a powerful technique that helps

obtain more precise estimations and improved data analysis. However, the time spent on these calculations is still one of the main limitations. Technology improvements may lead to implement Bayesian inference into more standard data analysis procedures.

Why is it relevant to continue studying survivin at the molecular level? Survivin is involved in quite a few severe diseases, such as autoimmune disease and cancer. Cancer is one of the main causes of global death where survivin has been related with more critical cancer stages. Studies conducted at the molecular level reveal important information about the protein function and specific interaction regions with other proteins. Elucidating new interactions can lead to a better understanding of cell division and apoptosis, and to an improvement in diagnosing or treating related diseases.

Why is it important to obtain more robust knowledge about survivin isoforms?
These isoforms have been related with different disease stages, but there is still a world to be discovered at the molecular level. As they have been suggested to regulate and interact with survivin, elucidating their structure and how these interactions take place will also be important to fully understand the survivin function.

Appendix A

Survivin gene

ATGGGTGCCCCGACGTTGCCCCCTGCCTGGCAGCCCTTTCTC
AAGGACCACCGCATCTCTACATTCAAGAACTGGCCCTTCTTGGA
GGGCTGCGCCTGCACCCCGGAGCGGATGGCCGAGGCTGGCTTC
ATCCACTGCCCCACTGAGAACGAGCCAGACTTGGCCCAGTGTTT
CTTCTGCTTCAAGGAGCTGGAAGGCTGGGAGCCAGATGACGAC
CCCATAGAGGAACATAAAAAGCATTCGTCCGGTTGCGCTTTCCT
TTCTGTCAAGAAGCAGTTTGAAGAATTAACCCTTGGTGAATTTT
TGAAACTGGACAGAGAAAGAGCCAAGAACAAAATTGCAAAGG
AAACCAACAATAAGAAGAAAGAATTTGAGGAAACTGCGAAGA
AAGTGCGCCGTGCCATCGAGCAGCTGGCTGCCATGGATTGA

hSgol1$^{195-312}$ gene

CTGAAAAAACATTGTAATAGCATCTGCCAGTTCGATAGCCT
GGATGATTTTGAAACCAGCCATCTGGCAGGTAAAAGCTTTGAAT
TTGAACGTGTTGGTTTTCTGGATCCGCTGGTTAACATGCATATTC
CGGAAAATGTTCAGCATAATGCATGCCAGTGGTCAAAAGATCA
GGTTAATCTGAGCCCGAAACTGATTCAGCCTGGCACCTTTACCA
AAACCAAAGAAGATATTCTGGAAAGCAAAAGCGAACAGACCA
AAAGCAAACAGCGTGATACCCAAGAACGTAAACGTGAAGAGA
AACGTAAAGCCAATCGTCGTAAAAGTAAACGTATGAGCAAATA

85

CAAAGAGAAC

Survivin-2B gene

ATGGGTGCCCCGACGTTGCCCCCTGCCTGGCAGCCCTTTCTC
AAGGACCACCGCATCTCTACATTCAAGAACTGGCCCTTCTTGGA
GGGCTGCGCCTGCACCCCGGAGCGGATGGCCGAGGCTGGCTTC
ATCCACTGCCCCACTGAGAACGAGCCAGACTTGGCCCAGTGTTT
CTTCTGCTTCAAGGAGCTGGAAGGCTGGGAGCCAGATGACGAC
CCCATTGGGCCGGGCACGGTGGCTTACGCCTGTAATACCAGCAC
TTTGGGAGGCCGAGGCGGGCGGATCACGAGAGAGGAACATAAA
AAGCATTCGTCCGGTTGCGCTTTCCTTTCTGTCAAGAAGCAGTTT
GAAGAATTAACCCTTGGTGAATTTTTGAAACTGGACAGAGAAA
GAGCCAAGAACAAAATTGCAAAGGAAACCAACAATAAGAAGA
AAGAATTTGAGGAAACTGCGAAGAAAGTGCGCCGTGCCATCGA
GCAGCTGGCTGCCAaTGGATTGA

Survivin-ΔEx3 gene

ATGGGTGCCCCGACGTTGCCCCCTGCCTGGCAGCCCTTTCTC
AAGGACCACCGCATCTCTACATTCAAGAACTGGCCCTTCTTGGA
GGGCTGCGCCTGCACCCCGGAGCGGATGGCCGAGGCTGGCTTC
ATCCACTGCCCCACTGAGAACGAGCCAGACTTGGCCCAGTGTTT
CTTCTGCTTCAAGGAGCTGGAAGGCTGGGAGCCAGATGACGAC
CCCATGCAAAGGAAACCAACAATAAGAAGAAAGAATTTGAGGA
AACTGCGAAGAAAGTGCGCCGTGCCATCGAGCAGCTGGCTGCC
ATGGATTGAGGCCTCTGGCCGGAGCTGCCTGGTCCCAGAGTGGC
TGCACCACTTCCAGGGTTTATTCCCTGGTGCCACCAGCCTTCCTG
TGGGCCCCTTAGCAATGTCTTGA

Table A.1. Primers used

Primer sequences (5' to 3')
pWarf-survivin Forward
AAGAAGGAGACTCGAGATGGGTGCCCCGACG
pWarf-survivin Reverse
CAGCTGGCTGCCATGGATGGATCCTTGGAAGTCT
pET48b hSgol1$^{195-312}$ Forward
CGGGTACCAGGATCCTCTGAAAAAAACATTGTAATAGCATCTGC
pET48b hSgol1$^{195-312}$ Reverse
TGCGGCCGCAAGCTTAGTTCTCTTTGTATTTGCTCATACGT
pET28+8His survivin-2B Forward
CGCGCGGCAGCCATATGATGGGTGCCCCGACGTTG
pET28+8His survivin-2B Reverse
GCTCGAATTCGGATCCTCAATCCATGGCAGCCAGCTG
pET48b survivin-2B Forward
CGGGTACCAGGATCCGATGGGTGCCCCGACGTT
pET48b survivin-2B Reverse
TGCGGCCGCAAGCTTTCAATCCATGGCAGCCAGCT
pET49b survivin-2B Forward
CGGGTACCAGGATCCGATGGGTGCCCCGACGTT
pET49b survivin-2B Reverse
TGCGGCCGCAAGCTTTCAATCCATGGCAGCCAGCT
pET28+8His survivin-ΔEx3 Forward
CGCGCGGCAGCCATATGATGGGTGCCCCGACGTTG
pET28+8His survivin-ΔEx3 Reverse
TGTGGGCCCCTTAGCAATGTCTTAGGGATCCGAATTCGAG
pET48b survivin-ΔEx3 Forward
CGGGTACCAGGATCCGATGGGTGCCCCGACGTT
pET48b survivin-ΔEx3 Reverse
TGCGGCCGCAAGCTTCAAGACATTGCTAAGGGGCCCA
pET49b survivin-ΔEx3 Forward
CGGGTACCAGGATCCGATGGGTGCCCCGACGTT
pET49b survivin-ΔEx3 Reverse
TGCGGCCGCAAGCTTCAAGACATTGCTAAGGGGCCCA

Appendix B

hSgol1$^{195-312}$ production

The hSgol1$^{195-312}$ sequence was cloned into a pET48+ vector and the peptide was expressed in *E.coli* BL21 star (DE3) strain cells. This construct contains kanamycin selection and the thioredoxin (TRX) protein sequence, followed by a 6-histidine tag, the HRV3C cleavage site (LEVLFQ'GP) and the hSgol1$_{195-312}$ sequence. hSgol1$^{195-312}$-TRX was overexpressed at 180 rpm and 30ºC with 0.5 mM IPTG for 3-4 h. Cells were harvested and lysed by the standard method described in Chapter 2. This peptide was purified in three steps: nickel affinity chromatography, reverse chromatography and SEC. Affinity chromatography was performed by binding the protein to the column with 50 mM Tris, pH 8, 300 mM NaCl buffer and by eluting the protein in two steps using the same buffer that contained imidazole (10 mM step, and a gradient step from 10 mM to 200 mM). After affinity purification, the TRX-6His tag was cleaved by overnight dialysis (dialysis buffer: 50 mM Tris, pH7.5, 300 mM NaCl and 14 mM β6-mercaptoethanol) with HRV3C protease (1:100). TRX was removed by reverse his-tag purification, followed by a SEC step in a Superdex 75 100/300 GL column (SEC buffer: 50 mM Tris, pH 7.5, 300 mM NaCl, 5 mM DTT buffer). The purified protein was analysed by SDS-PAGE electrophoresis and protein concentration was estimated by absorbance at 280 nm.

Survivin-eGFP production

The survivin gene was cloned into the pWarf(-) vector and the protein was overexpressed in *E.coli* Lemo21 (DE3) strain cells. This construct contains kanamycin selection and the full human survivin protein sequence, followed by the HRV3C protease cleavage site, the eGFP gene and a C-terminal 8xHis-tag. Expression was optimised at 180 rpm and 25ºC with different concentrations of IPTG (0 mM, 0.1 mM, 0.2 mM and 0.4 mM) and of L-rhamnose (0 mM, 0.10 mM, 0.25 mM, 0.50 mM, 0.75 M and 1 mM). Twenty-four different conditions were tested. The protein expression level was analysed by measuring the relative fluorescence (RFU) in the complete cells and normalising the results with the optical density (OD600) of each culture (Figure B.1).
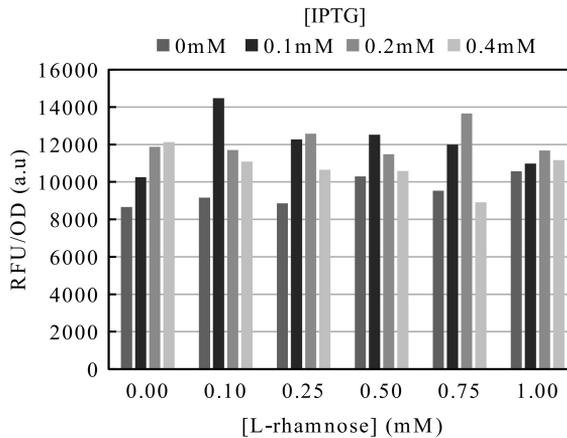
Figure B.1. This histogram represents the survivin-eGFP expression test results. The x-axis depicts the different tested L-rhamnose concentrations, while the y-axis is the relative fluorescence normalised by the OD600 of each culture. Each series represents the different tested IPTG concentrations.

According to these results (FigureB.1), survivin-eGFP was higher expressed with 0.1mM L-Rhamnose and 0.1 mM IPTG concentrations. Big-scale expression was performed in LB media containing 50 ug/ml Kan, 34 μg/ml chloramphenicol (Cam) 80 μM ZnCl2 and 0.1 mM L-Rhamnose. Induction was started at OD600nm 0.4-0.6 with 0.1 mM IPTG. Cells were incubated at 180 rpm and 25ºC overnight.

Survivin-eGFP was purified by the same protocol described for survivin in **paper III**, but a Superdex-200 10/300 GL column was used. Apart from the standard protein detection at 280 nm, eGFP fluorescence was also monitored during elution.

An eGFP construct was also generated from the pWarf(-) vector to express the eGFP protein for the negative control. The protein was expressed in the *E.coli* BL21 star (DE3) strain and purified according **paper III** protocol .

Acknowledgements

It feels like it was only yesterday when I was walking along the Lundberg lab corridor for the first time. However, almost 5 years since that day have gone by. The journey has been great, but also hard, and it would not have been possible without the support of many people who I wish to thank.

First of all, I would like to thank you, **Gergely**, for giving me the chance of joining your group and introducing me to the crystallography world. Your knowledge is endless and I am glad to have been able to learn it from you. Thank you for all the questions that led me away from my comfort zone, and made me think and learn. If I have to pick up something from all these years, it would be the beamtimes. They have been exhausting but also great fun and inspiring. I will never forget the fateful beamtime when it took us forever to arrive at the ESRF and we got 1.7Å of what we thought was survivin-2B.

Maria B., for being my co-supervisor and for all the inspiring and enjoyable survivin discussions, which I really enjoyed.

To my examiner, **Richard**, for guiding and helping me over the years and for giving me the chance to participate in an XFEL beamtime with his group. It was a great experience.

To all the members of Katona´s group as it has been a pleasure to work with you all. **Annette**, for being so kind to me and for showing me around my first months in the lab. **Ida**, for all the fun in the THz beamtimes. **Viktor**, I really enjoyed the Oviedo conference. Good luck, you are the next! **Hassan** and **Stan**, it was fantastic having you in the group. **Sofia**, **Nauras**, **Nathalie**, **Maja** and **Torbjorn**, thank you all for working as bachelor/master students

with me and bringing cheeriness to the survivin's project. **Maja**, for being my travel partner in the survivin project in the last years, for all your help and your huge support in the hardest times. We had a lot of fun!

To all the people who form part of the Lundberg lab; Neutze´s group (**Rebecka**, **Rob B**, **Petra B**, **Cecilia W.**, **Greger**, **Per**, **Giorgia**, **Daniel**, etc.), Westenhoff´s group (**Sebastian**, **Elin C.**, **Leo**, **Andrea**, **Amke**, **Linnéa**, **Joachim**, **Matthijs**, **Weixiao**, **Leticia**, **Michal**, **Ann**, **Laras**, etc.), Brändén´s group (**Gisela**, **Cecilia S.**, **Andreas**, **Swagatha**, **Jonatan**, etc.), Höög´s group (**Johanna**, **Davide**, **Jake**, **Dimitra**, etc.), Hedfalk´s group (**Kristina**, **Florian**, **Jessica**, etc.), Burmann's group (**Bjorn**,**Darius**, **Emilie**, **Laura**, **Josh**, **Lisa**, **Damasus**, **Ashis**, **Irena**, **Ylber**, etc.); thank you all for the good times, fika times and beer clubs! I feel really glad to have met you all.

Rebecka thanks for all the laughs and good times. I am looking forward to being your neighbour in Bollelingsås. Leo, my dancing partner and a great friend! Thanks for all your positivity and support. **Petra B.**, for all the teaching moments and dances during the thesis parties. The chocolate really helped me on these long days. Tina, **Per** and **Giorgia**, without whom I couldn't imagine a work trip being so fun as it was in Japan. Thank you for the amazing experience. I think it is time to be a unicorn again! **Weixiao**, I really appreciate you having been so patient and supporting me in the difficult times. I really enjoyed discussing science with you and learning from you. **Elin C.**, I really enjoyed having you around and the funny Tällberg. **Cecilia S.**, thank you for being so sweet with me. El Campello is waiting for us! **Florian**, for all the laughs and fun. I am looking forward to the next opera. **Davide**, it was awesome to have you as roomie! **Kristina** and **Gisela**, it was such a wonderful experience teaching and learning in your courses.

To all people who are no longer in the Lundberg lab, but have also

formed part of this journey; thank you for all the fun (**Elin D.**, **Rhawnie**, **Rob D.**, **David**, **Mike**, **Stefan**, **Stephan**, **Rob D.** (I love your Halloween parties), **Petra E.**, **Oskar**, **Rajiv** (another Australia?), **Alex**, **Rosie**, **Parveen**, **Nasha**, **Rachel**, **Jenni**, **Jenny**, **Emil**, **Ash**, **Amit**, etc.). **Rosie**, for making crystallography so much fun!; for all your help and positivity. **Nasha** and **Rachel**, for all the good times in the moon and the crazy lyase experience! **Parveen**, thank you for all your help when I was a newbie in the lab and for the afterworks.

I thank all the collaborators and co-authors who made this possible, especially **Christian** and **Malin**.

Lars, **Bruno** and **Valida**, the lab would not work without you!

To my family, especially my parents, **Fernando** and **María José**. There are no words for saying thank you to you. You have been the best teachers that I could have, and your hard work and perseverance have been my inspiration. To my big brother **Fer**, for taking care of me and supporting me. Thanks for always making me smile.

To all my friends in Sweden and back at home. **Belén**, **Maryam** and **Mario**, for taking care of me for all these years and never letting me fall. **Cosme**, **Moi**, **Sandra** and **Usua**, those unconditional friends who have always been there, for both good and bad. To the "**Biologuillos group**", because my journey in science started with you, and the genetic exercises! I am so pleased to have such good friends.

Simón, thank you for being at my side throughout this long journey and making it funnier. For your support and help in the hardest times. For all the times that you have picked me up late from the lab and listened to me talking about science and my crazy experiments. For helping me feel calm when I did not feel it, and for not running away. I love you.

Bibliography

[1] G. M. Cooper, *The Cell: A Molecular Approach*. Sunderland (MA): Sinauer Associates, 2nd ed., 2000.

[2] B. Alberts, A. Johnson, J. Lewis, et, and al, *Biology of the Cell*. New York: Garland Science, 4th ed., 2002.

[3] B. Pucci, M. Kasten, and A. Giordano, "Cell cycle and apoptosis," *Neoplasia*, vol. 2, no. 4, pp. 291–9, 2000.

[4] Q. L. Deveraux and J. C. Reed, "IAP family proteins–suppressors of apoptosis," *Genes Dev*, vol. 13, no. 3, pp. 239–52, 1999.

[5] G. Gravina, C. Wasen, M. J. Garcia-Bonete, M. Turkkila, M. C. Erlandsson, S. Toyra Silfversward, M. Brisslert, R. Pullerits, K. M. Andersson, G. Katona, and M. I. Bokarewa, "Survivin in autoimmune diseases," *Autoimmun Rev*, vol. 16, no. 8, pp. 845–855, 2017.

[6] H. Doolittle, A. Morel, and D. Talbot, "Survivin-directed Anticancer Therapies – A Review of Pre-clinical Data and Early-phase Clinical Trials," *European Oncology*, vol. 6, no. 1, pp. 10–14, 2010.

[7] G. Duan and D. Walther, "The roles of post-translational modifications in the context of protein interaction networks," *PLoS Comput Biol*, vol. 11, no. 2, p. e1004049, 2015.

[8] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *Br J Pharmacol*, vol. 162, no. 6, pp. 1239–49, 2011.

[9] P. L. Kastritis and A. M. Bonvin, "On the binding affinity of macromolecular interactions: daring to ask why proteins interact," *J R Soc Interface*, vol. 10, no. 79, p. 20120835, 2013.

[10] D. Vasudevan and H. D. Ryoo, "Regulation of Cell Death by IAPs and Their Antagonists," *Curr Top Dev Biol*, vol. 114, pp. 185–208, 2015.

[11] J. Silke and P. Meier, "Inhibitor of apoptosis (IAP) proteins-modulators of cell death and inflammation," *Cold Spring Harb Perspect Biol*, vol. 5, no. 2, 2013.

[12] G. Wu, J. Chai, T. L. Suber, J. W. Wu, C. Du, X. Wang, and Y. Shi, "Structural basis of IAP recognition by Smac/DIABLO," *Nature*, vol. 408, no. 6815, pp. 1008–12, 2000.

[13] D. Finlay, P. Teriete, M. Vamos, N. D. P. Cosford, and K. Vuori, "Inducing death in tumor cells: roles of the inhibitor of apoptosis proteins," *F1000Res*, vol. 6, p. 587, 2017.

[14] B. Yan, "Research progress on Livin protein: an inhibitor of apoptosis," *Mol Cell Biochem*, vol. 357, no. 1-2, pp. 39–45, 2011.

[15] P. Obexer and M. J. Ausserlechner, "X-linked inhibitor of apoptosis protein - a critical death resistance regulator and therapeutic target for personalized cancer therapy," *Front Oncol*, vol. 4, p. 197, 2014.

[16] A. Marchler-Bauer and S. H. Bryant, "CD-Search: protein domain annotations on the fly," *Nucleic Acids Res*, vol. 32, no. Web Server issue, pp. W327–31, 2004.

[17] G. Ambrosini, C. Adida, and D. C. Altieri, "A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma," *Nat Med*, vol. 3, no. 8, pp. 917–21, 1997.

[18] D. C. Altieri, "Survivin - The inconvenient IAP," *Semin Cell Dev Biol*, vol. 39, pp. 91–6, 2015.

[19] N. K. Sah, Z. Khan, G. J. Khan, and P. S. Bisen, "Structural, functional and therapeutic biology of survivin," *Cancer Lett*, vol. 244, no. 2, pp. 164–71, 2006.

[20] S. Mokuda, T. Miyazaki, Y. Ito, S. Yamasaki, H. Inoue, Y. Guo, W. S. Kong, M. Kanno, K. Takasugi, E. Sugiyama, and J. Masumoto, "The proto-oncogene survivin splice variant 2B is induced by PDGF and leads to cell proliferation in rheumatoid arthritis fibroblast-like synoviocytes," *Sci Rep*, vol. 5, p. 9795, 2015.

[21] E. C. LaCasse, S. Baird, R. G. Korneluk, and A. E. MacKenzie, "The inhibitors of apoptosis (IAPs) and their emerging role in cancer," *Oncogene*, vol. 17, no. 25, pp. 3247–59, 1998.

[22] M. A. Verdecia, H. Huang, E. Dutil, D. A. Kaiser, T. Hunter, and J. P. Noel, "Structure of the human anti-apoptotic protein survivin reveals a dimeric arrangement," *Nat Struct Biol*, vol. 7, no. 7, pp. 602–8, 2000.

[23] L. Chantalat, D. A. Skoufias, J. P. Kleman, B. Jung, O. Dideberg, and R. L. Margolis, "Crystal structure of human survivin reveals a bow tie-shaped dimer with two unusual alpha-helical extensions," *Mol Cell*, vol. 6, no. 1, pp. 183–9, 2000.

[24] C. Sun, D. Nettesheim, Z. Liu, and E. T. Olejniczak, "Solution structure of human survivin and its binding interface with Smac/Diablo," *Biochemistry*, vol. 44, no. 1, pp. 11–7, 2005.

[25] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera–a visualization system for exploratory research and analysis," *J Comput Chem*, vol. 25, no. 13, pp. 1605–12, 2004.

[26] D. C. Altieri, "Survivin and IAP proteins in cell-death mechanisms," *Biochem J*, vol. 430, no. 2, pp. 199–205, 2010.

[27] K. M. Andersson, M. Turkkila, M. C. Erlandsson, A. Bossios, S. T. Silfversward, D. Hu, L. Ekerljung, C. Malmhall, H. L. Weiner, B. Lundback, and M. I. Bokarewa, "Survivin controls biogenesis of microRNA in smokers: A link to pathogenesis of rheumatoid arthritis," *Biochim Biophys Acta Mol Basis Dis*, vol. 1863, no. 3, pp. 663–673, 2017.

[28] D. Trnski, M. Gregoric, S. Levanat, P. Ozretic, N. Rincic, T. M. Vidakovic, D. Kalafatic, I. Maurac, S. Oreskovic, M. Sabol, and V. Musani, "Regulation of Survivin Isoform Expression by GLI Proteins in Ovarian Cancer," *Cells*, vol. 8, no. 2, 2019.

[29] C. M. O'Connor and J. U. Adams, *Essentials of Cell Biology*. Cambridge, MA: NPG Education, 2010.

[30] A. A. Jeyaprakash, U. R. Klein, D. Lindner, J. Ebert, E. A. Nigg, and E. Conti, "Structure of a Survivin-Borealin-INCENP core complex reveals how chromosomal passengers travel together," *Cell*, vol. 131, no. 2, pp. 271–85, 2007.

[31] S. Ruchaud, M. Carmena, and W. C. Earnshaw, "The chromosomal passenger complex: one for all and all for one," *Cell*, vol. 131, no. 2, pp. 230–1, 2007.

[32] M. Carmena, M. Wheelock, H. Funabiki, and W. C. Earnshaw, "The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis," *Nat Rev Mol Cell Biol*, vol. 13, no. 12, pp. 789–803, 2012.

[33] M. S. van der Waal, R. C. Hengeveld, A. van der Horst, and S. M. Lens, "Cell division control by the Chromosomal Passenger Complex," *Exp Cell Res*, vol. 318, no. 12, pp. 1407–20, 2012.

[34] S. Hindriksen, S. M. A. Lens, and M. A. Hadders, "The Ins and Outs of Aurora B Inner Centromere Localization," *Front Cell Dev Biol*, vol. 5, p. 112, 2017.

[35] E. Niedzialkowska, F. Wang, P. J. Porebski, W. Minor, J. M. Higgins, and P. T. Stukenberg, "Molecular basis for phosphospecific recognition of histone H3 tails by Survivin paralogues at inner centromeres," *Mol Biol Cell*, vol. 23, no. 8, pp. 1457–66, 2012.

[36] J. Du, A. E. Kelly, H. Funabiki, and D. J. Patel, "Structural basis for recognition of H3T3ph and Smac/DIABLO N-terminal peptides by human Survivin," *Structure*, vol. 20, no. 1, pp. 185–95, 2012.

[37] A. E. Kelly, C. Ghenoiu, J. Z. Xue, C. Zierhut, H. Kimura, and H. Funabiki, "Survivin reads phosphorylated histone H3 threonine 3 to activate the mitotic kinase Aurora B," *Science*, vol. 330, no. 6001, pp. 235–9, 2010.

[38] D. Clift and A. L. Marston, "The role of shugoshin in meiotic chromosome segregation," *Cytogenet Genome Res*, vol. 133, no. 2-4, pp. 234–42, 2011.

[39] S. A. Kawashima, Y. Yamagishi, T. Honda, K. Ishiguro, and Y. Watanabe, "Phosphorylation of H2A by Bub1 prevents chromosomal instability through localizing shugoshin," *Science*, vol. 327, no. 5962, pp. 172–7, 2010.

[40] S. A. Kawashima, T. Tsukahara, M. Langegger, S. Hauf, T. S. Kitajima, and Y. Watanabe, "Shugoshin enables tension-generating attachment of kinetochores by loading Aurora to centromeres," *Genes Dev*, vol. 21, no. 4, pp. 420–35, 2007.

[41] Y. Yamagishi, T. Honda, Y. Tanno, and Y. Watanabe, "Two histone marks establish the inner centromere and chromosome bi-orientation," *Science*, vol. 330, no. 6001, pp. 239–43, 2010.

[42] A. A. Jeyaprakash, C. Basquin, U. Jayachandran, and E. Conti, "Structural basis for the recognition of phosphorylated histone h3 by the survivin subunit of the chromosomal passenger complex," *Structure*, vol. 19, no. 11, pp. 1625–34, 2011.

[43] T. Tsukahara, Y. Tanno, and Y. Watanabe, "Phosphorylation of the CPC by Cdk1 promotes chromosome bi-orientation," *Nature*, vol. 467, no. 7316, pp. 719–23, 2010.

[44] A. P. Baron, C. von Schubert, F. Cubizolles, G. Siemeister, M. Hitchcock, A. Mengel, J. Schroder, A. Fernandez-Montalvan, F. von Nussbaum, D. Mumberg, and E. A. Nigg, "Probing the catalytic functions of Bub1 kinase using the small molecule inhibitors BAY-320 and BAY-524," *Elife*, vol. 5, 2016.

[45] H. Okada and T. W. Mak, "Pathways of apoptotic and non-apoptotic death in tumour cells," *Nat Rev Cancer*, vol. 4, no. 8, pp. 592–603, 2004.

[46] D. R. McIlwain, T. Berger, and T. W. Mak, "Caspase functions in cell death and disease," *Cold Spring Harb Perspect Biol*, vol. 5, no. 4, p. a008656, 2013.

[47] S. Fulda and D. Vucic, "Targeting IAP proteins for therapeutic intervention in cancer," *Nat Rev Drug Discov*, vol. 11, no. 2, pp. 109–24, 2012.

[48] Z. Song, X. Yao, and M. Wu, "Direct interaction between survivin and Smac/DIABLO is essential for the anti-apoptotic activity of survivin during taxol-induced apoptosis," *J Biol Chem*, vol. 278, no. 25, pp. 23130–40, 2003.

[49] N. K. Sah and C. Seniya, "Survivin splice variants and their diagnostic significance," *Tumour Biol*, vol. 36, no. 9, pp. 6623–31, 2015.

[50] B. Ryan, N. O'Donovan, B. Browne, C. O'Shea, J. Crown, A. D. Hill, E. McDermott, N. O'Higgins, and M. J. Duffy, "Expression of survivin and its splice variants survivin-2B and survivin-DeltaEx3 in breast cancer," *Br J Cancer*, vol. 92, no. 1, pp. 120–4, 2005.

[51] A. Pavlidou, C. Kroupis, and K. Dimas, "Association of survivin splice variants with prognosis and treatment of breast cancer," *World J Clin Oncol*, vol. 5, no. 5, pp. 883–94, 2014.

[52] J. Gaytan-Cervantes, C. Gonzalez-Torres, V. Maldonado, C. Zampedri, G. Ceballos-Cancino, and J. Melendez-Zajgla, "Protein Sam68 regulates the alternative splicing of survivin DEx3," *J Biol Chem*, vol. 292, no. 33, pp. 13745–13757, 2017.

[53] A. G. Antonacopoulou, K. Floratou, V. Bravou, A. Kottorou, F. I. Dimitrakopoulos, S. Marousi, M. Stavropoulos, A. K. Koutras, C. D. Scopa, and H. P. Kalofonos, "The survivin -31 snp in human colorectal cancer correlates with survivin splice variant expression and improved overall survival," *Anal Cell Pathol (Amst)*, vol. 33, no. 5, pp. 177–89, 2010.

[54] M. J. Duffy, N. O'Donovan, D. J. Brennan, W. M. Gallagher, and B. M. Ryan, "Survivin: a promising tumor biomarker," *Cancer Lett*, vol. 249, no. 1, pp. 49–60, 2007.

[55] M. E. Johnson and E. W. Howerth, "Survivin: a bifunctional inhibitor of apoptosis protein," *Vet Pathol*, vol. 41, no. 6, pp. 599–607, 2004.

[56] C. Mahotka, J. Liebmann, M. Wenzel, C. V. Suschek, M. Schmitt, H. E. Gabbert, and C. D. Gerharz, "Differential subcellular localization of functionally divergent survivin splice variants," *Cell Death Differ*, vol. 9, no. 12, pp. 1334–42, 2002.

[57] A. Pavlidou, M. Dalamaga, C. Kroupis, G. Konstantoudakis, M. Belimezi, G. Athanasas, and K. Dimas, "Survivin isoforms and clinicopathological characteristics in colorectal adenocarcinomas using real-time qPCR," *World J Gastroenterol*, vol. 17, no. 12, pp. 1614–21, 2011.

[58] S. De Maria, G. Pannone, P. Bufo, A. Santoro, R. Serpico, S. Metafora, C. Rubini, D. Pasquali, S. M. Papagerakis, S. Staibano, G. De Rosa, E. Farina, M. Emanuelli, A. Santarelli, M. A. Mariggio, L. Lo Russo, and L. Lo Muzio, "Survivin gene-expression and splicing isoforms in oral squamous cell carcinoma," *J Cancer Res Clin Oncol*, vol. 135, no. 1, pp. 107–16, 2009.

[59] H. Caldas, L. E. Honsey, and R. A. Altura, "Survivin 2alpha: a novel Survivin splice variant expressed in human malignancies," *Mol Cancer*, vol. 4, no. 1, p. 11, 2005.

[60] S. Cory and J. M. Adams, "The Bcl2 family: regulators of the cellular life-or-death switch," *Nat Rev Cancer*, vol. 2, no. 9, pp. 647–56, 2002.

[61] J. Waligorska-Stachura, N. Sawicka-Gutaj, M. Zabel, W. Liebert, P. Gut, A. Czarnywojtek, and M. Ruchala, "Decreased expression of survivin 2B in human pituitary adenomas. A preliminary study," *Folia Histochem Cytobiol*, vol. 55, no. 1, pp. 21–25, 2017.

[62] M. H. Malcles, H. W. Wang, A. Koumi, Y. H. Tsai, M. Yu, A. Godfrey, and C. Boshoff, "Characterisation of the anti-apoptotic function of survivin-DeltaEx3 during TNFalpha-mediated cell death," *Br J Cancer*, vol. 96, no. 11, pp. 1659–66, 2007.

[63] E. A. Noton, R. Colnaghi, S. Tate, C. Starck, A. Carvalho, P. Ko Ferrigno, and S. P. Wheatley, "Molecular analysis of survivin isoforms: evidence that alternatively spliced variants do not play a role in mitosis," *J Biol Chem*, vol. 281, no. 2, pp. 1286–95, 2006.

[64] M. Espinosa, G. Ceballos-Cancino, R. Callaghan, V. Maldonado, N. Patino, V. Ruiz, and J. Melendez-Zajgla, "Survivin isoform Delta Ex3 regulates tumor spheroid formation," *Cancer Lett*, vol. 318, no. 1, pp. 61–7, 2012.

[65] G. Georgiou and P. Valax, "Expression of correctly folded proteins in Escherichia coli," *Current Opinion in Biotechnology*, vol. 7, no. 2, pp. 190–197, 1996.

[66] R. Vincentelli, A. Cimino, A. Geerlof, A. Kubo, Y. Satou, and C. Cambillau, "High-throughput protein expression screening and purification in Escherichia coli," *Methods*, vol. 55, no. 1, pp. 65–72, 2011.

[67] G. L. Rosano and E. A. Ceccarelli, "Recombinant protein expression in Escherichia coli: advances and challenges," *Frontiers in microbiology*, vol. 5, pp. 172–172, 2014.

[68] F. A. Marston, "The purification of eukaryotic polypeptides synthesized in Escherichia coli," *The Biochemical journal*, vol. 240, no. 1, pp. 1–12, 1986.

[69] I. Iost, J. Guillerez, and M. Dreyfus, "Bacteriophage T7 RNA polymerase travels far ahead of ribosomes in vivo," *Journal of bacteriology*, vol. 174, no. 2, pp. 619–622, 1992.

[70] F. W. Studier and B. A. Moffatt, "Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes," *Journal of Molecular Biology*, vol. 189, no. 1, pp. 113–130, 1986.

[71] H. Jeong, V. Barbe, C. H. Lee, D. Vallenet, D. S. Yu, S.-H. Choi, A. Couloux, S.-W. Lee, S. H. Yoon, L. Cattolico, C.-G. Hur, H.-S. Park, B. Ségurens, S. C. Kim, T. K. Oh, R. E. Lenski, F. W. Studier, P. Daegelen, and J. F. Kim, "Genome Sequences of Escherichia coli B strains REL606 and BL21(DE3)," *Journal of Molecular Biology*, vol. 394, no. 4, pp. 644–652, 2009.

[72] S. Wagner, M. M. Klepsch, S. Schlegel, A. Appel, R. Draheim, M. Tarry, M. Högbom, K. J. van Wijk, D. J. Slotboom, J. O. Persson, and J.-W. de Gier, "Tuning Escherichia coli for membrane protein overexpression ," *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, p. 14371, 2008.

[73] M. J. Giacalone, A. M. Gentile, B. T. Lovitt, N. L. Berkley, C. W. Gunderson, and M. W. Surber, "Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system," *BioTechniques*, vol. 40, no. 3, pp. 355–364, 2006.

[74] H. P. Sørensen and K. K. Mortensen, "Advanced genetic strategies for recombinant protein expression in Escherichia coli," *Journal of Biotechnology*, vol. 115, no. 2, pp. 113–128, 2005.

[75] J. W. Dubendorf and F. W. Studier, "Controlling basal expression in an inducible T7 expression system by blocking the target T7 promoter with lac repressor," *Journal of Molecular Biology*, vol. 219, no. 1, pp. 45–59, 1991.

[76] M. P. Calos, "DNA sequence for a low-level promoter of the lac repressor gene and an 'up' promoter mutation," *Nature*, vol. 274, no. 5673, pp. 762–765, 1978.

[77] S. Costa, A. Almeida, A. Castro, and L. Domingues, "Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system," *Frontiers in Microbiology*, vol. 5, no. 63, 2014.

[78] D. S. Waugh, "An overview of enzymatic reagents for the removal of affinity tags," *Protein expression and purification*, vol. 80, no. 2, pp. 283–293, 2011.

[79] J. M. Jez, J.-L. Ferrer, M. E. Bowman, R. A. Dixon, and J. P. Noel, "Dissection of Malonyl-Coenzyme A Decarboxylation from Polyketide Formation in the Reaction Mechanism of a Plant Polyketide Synthase," *Biochemistry*, vol. 39, no. 5, pp. 890–902, 2000.

[80] J. M. Hsieh, G. M. Besserer, M. G. Madej, H.-Q. Bui, S. Kwon, and J. Abramson, "Bridging the gap: a GFP-based strategy for overexpression and purification of membrane proteins with intra and extracellular C-termini," *Protein science : a publication of the Protein Society*, vol. 19, no. 4, pp. 868–880, 2010.

[81] A. L. Throop and J. LaBaer, "Site-specific recombinational cloning using gateway and in-fusion cloning schemes," *Current protocols in molecular biology*, vol. 110, pp. 3.20.1–3.20.23, 2015.

[82] *Novagen pET System Manual*. Merck KGaA, Darmstadt, Germany, User Protocol TB055 Rev. C 0611J, 11th ed., 2011.

[83] EMBL, "Protein expression and purification core facility," 2019. Available at: https://www.embl.de/pepcore/pepcore_services/index.html.

[84] *Strategies for Protein Purification*. Uppsala: GE Healthcare Bio-Sciences AB, 2010.

[85] J. A. Bornhorst and J. J. Falke, "Purification of proteins using polyhistidine affinity tags," *Methods in enzymology*, vol. 326, pp. 245–254, 2000.

[86] H. Block, B. Maertens, A. Spriestersbach, N. Brinker, J. Kubicek, R. Fabis, J. Labahn, and F. Schäfer, *Chapter 27 Immobilized-Metal Affinity Chromatography (IMAC): A Review*, vol. 463, pp. 439–473. Academic Press, 2009.

[87] S. Fekete, A. Beck, J.-L. Veuthey, and D. Guillarme, "Theory and practice of size exclusion chromatography for the analysis of protein aggregates," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 101, pp. 161–173, 2014.

[88] A. L. Shapiro, E. Viñuela, and J. V. Maizel, "Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels," *Biochemical and Biophysical Research Communications*, vol. 28, no. 5, pp. 815–820, 1967.

[89] J. M. Walker, *The Bicinchoninic Acid (BCA) Assay for Protein Quantitation*, pp. 11–14. Totowa, NJ: Humana Press, 2002.

[90] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, *Protein Identification and Analysis Tools on the ExPASy Server*, pp. 571–607. Totowa, NJ: Humana Press, 2005.

[91] C. The UniProt, "UniProt: a worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, 2018.

[92] T. W. Chang, "Binding of cells to matrixes of distinct antibodies coated on solid surface," *J Immunol Methods*, vol. 65, no. 1-2, pp. 217–23, 1983.

[93] P. V. Pham, *Chapter 19 - Medical Biotechnology: Techniques and Applications*, pp. 449–469. Academic Press, 2018.

[94] M. Abdollahzadeh, M. S. Saidi, and A. Sadeghi, "Enhancement of surface adsorption-desorption rates in microarrays invoking surface charge heterogeneity," *Sensors and Actuators B: Chemical*, vol. 242, pp. 956–964, 2017.

[95] A. L. Furst, S. H. Klass, and M. B. Francis, "DNA Hybridization to Control Cellular Interactions," *Trends in Biochemical Sciences*, 2018.

[96] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science*, vol. 270, no. 5235, pp. 467–70, 1995.

[97] M. B. Miller and Y. W. Tang, "Basic concepts of microarrays and potential applications in clinical microbiology," *Clin Microbiol Rev*, vol. 22, no. 4, pp. 611–33, 2009.

[98] M. P. . V. M. López M., "Microarrays y Biochips de ADN. ," *Informe de Vigilancia Tecnológica*, no. GEN-ES02001, p. 56, 2002.

[99] R. Frank, "Spot-synthesis: an easy technique for the positionally addressable, parallel chemical synthesis on a membrane support," *Tetrahedron*, vol. 48, no. 42, pp. 9217–9232, 1992.

[100] E. M. Cornett, B. M. Dickson, R. M. Vaughan, S. Krishnan, R. C. Trievel, B. D. Strahl, and S. B. Rothbart, *Chapter Two - Substrate Specificity Profiling of Histone-Modifying Enzymes by Peptide Microarray*, vol. 574, pp. 31–52. Academic Press, 2016.

[101] S. B. Rothbart, K. Krajewski, B. D. Strahl, and S. M. Fuchs, "Peptide microarrays to interrogate the "histone code"," *Methods Enzymol*, vol. 512, pp. 107–35, 2012.

[102] V. Stadler, T. Felgenhauer, M. Beyer, S. Fernandez, K. Leibe, S. Guttler, M. Groning, K. Konig, G. Torralba, M. Hausmann, V. Lindenstruth, A. Nesterov, I. Block, R. Pipkorn, A. Poustka, F. R. Bischoff, and F. Breitling, "Combinatorial synthesis of peptide arrays with a laser printer," *Angew Chem Int Ed Engl*, vol. 47, no. 37, pp. 7132–5, 2008.

[103] C. Ludwig, *Diffusion zwischen ungleich erwärmten Orten gleich zusammengesetzter Lösung*. Aus der K.K. Hof- und Staatsdruckerei, in Commission bei W. Braumüller, Buchhändler des K.K. Hofes und der K. Akademie der Wissenschaften, 1856.

[104] M. Jerabek-Willemsen, C. J. Wienken, D. Braun, P. Baaske, and S. Duhr, "Molecular interaction studies using microscale thermophoresis," *Assay and drug development technologies*, vol. 9, no. 4, pp. 342–353, 2011.

[105] C. J. Wienken, P. Baaske, U. Rothbauer, D. Braun, and S. Duhr, "Protein-binding assays in biological liquids using microscale thermophoresis," *Nature Communications*, vol. 1, p. 100, 2010.

[106] M. Jerabek-Willemsen, T. André, R. Wanner, H. M. Roth, S. Duhr, P. Baaske, and D. Breitsprecher, "MicroScale Thermophoresis: Interaction analysis and beyond," *Journal of Molecular Structure*, vol. 1077, pp. 101–113, 2014.

[107] S. A. I. Seidel, P. M. Dijkman, W. A. Lea, G. van den Bogaart, M. Jerabek-Willemsen, A. Lazic, J. S. Joseph, P. Srinivasan, P. Baaske, A. Simeonov, I. Katritch, F. A. Melo, J. E. Ladbury, G. Schreiber, A. Watts, D. Braun, and S. Duhr, "Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions," *Methods (San Diego, Calif.)*, vol. 59, no. 3, pp. 301–315, 2013.

[108] Y. Mao, L. Yu, R. Yang, L.-b. Qu, and P. d. B. Harrington, "A novel method for the study of molecular interaction by using microscale thermophoresis," *Talanta*, vol. 132, pp. 894–901, 2015.

[109] *MST StartingGuide –Monolith NT.115*. Munich: NanoTemper Technologies GmbH.

[110] U. B. Ericsson, B. M. Hallberg, G. T. DeTitta, N. Dekker, and P. Nordlund, "Thermofluor-based high-throughput stability optimization of proteins for structural studies," *Analytical Biochemistry*, vol. 357, no. 2, pp. 289–298, 2006.

[111] M.-C. Lo, A. Aulabaugh, G. Jin, R. Cowling, J. Bard, M. Malamas, and G. Ellestad, "Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery," *Analytical Biochemistry*, vol. 332, no. 1, pp. 153–159, 2004.

[112] M. K. Groftehauge, N. R. Hajizadeh, M. J. Swann, and E. Pohl, "Protein-ligand interactions investigated by thermal shift assays (TSA) and dual polarization interferometry (DPI)," *Acta Crystallographica Section D*, vol. 71, no. 1, pp. 36–44, 2015.

[113] G. V. Semisotnov, N. A. Rodionova, O. I. Razgulyaev, V. N. Uversky, A. F. Gripas', and R. I. Gilmanshin, "Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe," *Biopolymers*, vol. 31, no. 1, pp. 119–128, 1991.

[114] J. J. Lavinder, S. B. Hari, B. J. Sullivan, and T. J. Magliery, "High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering," *Journal of the American Chemical Society*, vol. 131, no. 11, pp. 3794–3795, 2009.

[115] K. Huynh and C. L. Partch, "Analysis of protein stability and ligand interactions by thermal shift assay," *Current protocols in protein science*, vol. 79, pp. 28.9.1–28.9.14, 2015.

[116] F. H. Niesen, H. Berglund, and M. Vedadi, "The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability," *Nature Protocols*, vol. 2, p. 2212, 2007.

[117] F. Vollrath, N. Hawkins, D. Porter, C. Holland, and M. Boulet-Audet, "Differential Scanning Fluorimetry provides high throughput data on silk protein transitions," *Scientific Reports*, vol. 4, p. 5625, 2014.

[118] "The Nobel Prize in Chemistry 2012," 2019. Available at: https://www.nobelprize.org/prizes/chemistry/2012/summary/.

[119] B. Rupp, *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, 2009.

[120] C. Giacovazzo, H. Monaco, G. Artioli, D. Viterbo, G. Ferraris, G. Gilli, G. Zanotti, and M. Catti, *Fundamentals of Crystallography*. New York: Oxford University Press, 2002.

[121] G. Rhodes, *Crystallography Made Crystal Clear*. Burlington: Academic Press, 2006.

[122] N. Asherie, "Protein crystallization and phase diagrams," *Methods*, vol. 34, no. 3, pp. 266–272, 2004.

[123] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta crystallographica. Section F, Structural biology communications*, vol. 70, no. Pt 1, pp. 2–20, 2014.

[124] W. H. Bragg and W. L. Bragg, "The reflection of X-rays by crystals," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 88, no. 605, pp. 428–438, 1913.

[125] Martinez-Ripoll, "Csic-crystallography," 2018. Available at: http://www.xtal.iqfr.csic.es/Cristalografia/welcome-en.html.

[126] G. Taylor, "The phase problem," *Acta Crystallographica Section D*, vol. 59, no. 11, pp. 1881–1890, 2003.

[127] P. Evans and A. McCoy, "An introduction to molecular replacement," *Acta crystallographica. Section D, Biological crystallography*, vol. 64, no. Pt 1, pp. 1–10, 2008.

[128] D. D. Rodríguez, C. Grosse, S. Himmel, C. González, I. M. de Ilarduya, S. Becker, G. M. Sheldrick, and I. Usón, "Crystallographic ab initio protein structure solution below atomic resolution," *Nature Methods*, vol. 6, p. 651, 2009.

[129] I. Caballero, M. Sammito, C. Millan, A. Lebedev, N. Soler, and I. Uson, "ARCIMBOLDO on coiled coils," *Acta Crystallographica Section D*, vol. 74, no. 3, pp. 194–204, 2018.

[130] S. Hovmöller, *Rotation Matrices and Translation Vectors in Crystallography*. Walles: International Union of Crystallography, University College Cardiff Press, 1981.

[131] E. A. Merritt, "X-ray anomalous scattering," 2012. Available at: http://skuld.bmsc.washington.edu/scatter/AS_index.html.

[132] R. J. Read and A. J. McCoy, "Using SAD data in Phaser," *Acta crystallographica. Section D, Biological crystallography*, vol. 67, no. Pt 4, pp. 338–344, 2011.

[133] W. A. Hendrickson, "Determination of macromolecular structures from anomalous diffraction of synchrotron radiation," *Science*, vol. 254, no. 5028, p. 51, 1991.

[134] K. D. Cowtan and K. Y. J. Zhang, "Density modification for macromolecular phase improvement," *Progress in Biophysics and Molecular Biology*, vol. 72, no. 3, pp. 245–270, 1999.

[135] A. Balerna and S. Mobilio, *Introduction to Synchrotron Radiation*, pp. 3–28. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[136] H. Toyokawa, C. Broennimann, E. F. Eikenberry, B. Henrich, M. Kawase, M. Kobas, P. Kraft, M. Sato, B. Schmitt, M. Suzuki, H. Tanida, and T. Uruga, "Single photon counting pixel detectors for synchrotron radiation experiments," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 623, no. 1, pp. 204–206, 2010.

[137] H. R. Powell, "X-ray data processing," *Bioscience reports*, vol. 37, no. 5, p. BSR20170227, 2017.

[138] P. R. Evans and G. N. Murshudov, "How good are my data and what is the resolution?," *Acta Crystallographica Section D*, vol. 69, no. 7, pp. 1204–1214, 2013.

[139] A. G. W. Leslie and H. R. Powell, "Processing diffraction data with mosflm," in *Evolving Methods for Macromolecular Crystallography* (R. J. Read and J. L. Sussman, eds.), pp. 41–51, Springer Netherlands.

[140] W. Kabsch, "XDS," *Acta Crystallographica Section D*, vol. 66, no. 2, pp. 125–132, 2010.

[141] P. A. Karplus and K. Diederichs, "Assessing and maximizing data quality in macromolecular crystallography," *Current opinion in structural biology*, vol. 34, pp. 60–68, 2015.

[142] P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "PHENIX: a comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D*, vol. 66, no. 2, pp. 213–221, 2010.

[143] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read, "Phaser crystallographic software," *Journal of Applied Crystallography*, vol. 40, no. 4, pp. 658–674, 2007.

[144] T. C. Terwilliger, P. D. Adams, R. J. Read, A. J. McCoy, N. W. Moriarty, R. W. Grosse-Kunstleve, P. V. Afonine, P. H. Zwart, and L.-W. Hung, "Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard," *Acta Crystallographica Section D*, vol. 65, no. 6, pp. 582–601, 2009.

[145] T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, N. W. Moriarty, P. H. Zwart, L.-W. Hung, R. J. Read, and P. D. Adams, "Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard," *Acta Crystallographica Section D*, vol. 64, no. 1, pp. 61–69, 2008.

[146] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams, "Towards automated crystallographic structure refinement with phenix.refine," *Acta Crystallographica Section D*, vol. 68, no. 4, pp. 352–367, 2012.

[147] V. B. Chen, I. Arendall, W. Bryan, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography," *Acta Crystallographica Section D*, vol. 66, no. 1, pp. 12–21, 2010.

[148] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of Coot," *Acta Crystallographica Section D*, vol. 66, no. 4, pp. 486–501, 2010.

[149] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution," *Quarterly Reviews of Biophysics*, vol. 40, no. 3, pp. 191–285, 2007.

[150] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering," *Journal of Structural Biology*, vol. 172, no. 1, pp. 128–141, 2010.

[151] A. G. Kikhney and D. I. Svergun, "A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins," *FEBS Letters*, vol. 589, no. 19PartA, pp. 2570–2577, 2015.

[152] A. Guinier, "La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques," *Ann. Phys.*, vol. 11, no. 12, pp. 161–237, 1939.

[153] J. Trewhella, A. P. Duff, D. Durand, F. Gabel, J. M. Guss, W. A. Hendrickson, G. L. Hura, D. A. Jacques, N. M. Kirby, A. H. Kwan, J. Pérez, L. Pollack, T. M. Ryan, A. Sali, D. Schneidman-Duhovny, T. Schwede, D. I. Svergun, M. Sugiyama, J. A. Tainer, P. Vachette, J. Westbrook, and A. E. Whitten, "2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update," *Acta crystallographica. Section D, Structural biology*, vol. 73, no. Pt 9, pp. 710–728, 2017.

[154] M. E. Brennich, J. Kieffer, G. Bonamis, A. De Maria Antolinos, S. Hutin, P. Pernot, and A. Round, "Online data analysis at the ESRF bioSAXS beamline, BM29," *Journal of Applied Crystallography*, vol. 49, no. 1, pp. 203–212, 2016.

[155] D. Franke, M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, and D. I. Svergun, "ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions," *Journal of applied crystallography*, vol. 50, no. Pt 4, pp. 1212–1225, 2017.

[156] P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, and D. I. Svergun, "PRIMUS: a Windows PC-based system for small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 36, no. 5, pp. 1277–1282, 2003.

[157] D. Franke and D. I. Svergun, "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering," *Journal of applied crystallography*, vol. 42, no. Pt 2, pp. 342–346, 2009.

[158] V. V. Volkov and D. I. Svergun, "Uniqueness of ab initio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 36, no. 3 Part 1, pp. 860–864, 2003.

[159] D. Svergun, C. Barberato, and M. H. J. Koch, "CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates," *Journal of Applied Crystallography*, vol. 28, no. 6, pp. 768–773, 1995.

[160] J. VanderPlas, *Frequentism and Bayesianism: A Python-driven Primer*. 2014.

[161] N. Gutierrez A., *Metodos Bayesianos en Estadistica Oficial*. 2017.

[162] W. M. Bolstad, *Introduction to Bayesian Statistics*. Wiley-Blackwell, 2nd ed., 2007.

[163] T. Bayes and n. Price, "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.

[164] A. Landro and M. González, "BERNOULLI, DE MOIVRE, BAYES, PRICE Y LOS FUNDAMENTOS DE LA INFERENCIA INDUCTIVA," *Cuadernos del CIMBAGE*, vol. 15, pp. 33–56, 2013.

[165] J. L. Devore, *Probability and Statistics for Engineering and the Sciences.* Belmont: Broks Cole Publishing Co, 6th ed., 2004.

[166] K. Pearson, R. A. Fisher, and H. F. Inman, "Karl Pearson and R. A. Fisher on Statistical Tests: A 1935 Exchange from Nature," *The American Statistician*, vol. 48, no. 1, pp. 2–11, 1994.

[167] C. B. Boyer, *A History of Mathematics*. New York, United States: Wiley International Edition, 1968.

[168] A. Etz, "Introduction to the Concept of Likelihood and Its Applications," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 60–69, 2018.

[169] L. J. Bain and M. Engelhardt, *Introduction To Probability And Mathematical Statistics*. Boston: Duxbury Press, 1987.

[170] G. Grimmett and D. Welsh, *PROBABILITY. An Introduction.* Oxford Science Publications, 1990.

[171] A. Eshky, "Bayesian methods of parameter estimation," *University of Edinburgh, School of Informatics*, 2008.

[172] C. Davidson-Pilon, *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Addison-Wesley, 2015.

[173] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles, *Bayesian Approaches To Clinical Trials And Health-Care Evaluation*. John Wiley and Sons, Ltd, 2004.

[174] T. Schweder and N. L. Hjort, *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press, 2016.

[175] D. Lewandowski, D. Kurowicka, and H. Joe, *Generating random correlation matrices based on vines and extended onion method*, vol. 100. 2009.

[176] T. H. Scheuermann, S. B. Padrick, K. H. Gardner, and C. A. Brautigam, "On the acquisition and analysis of microscale thermophoresis data," *Analytical Biochemistry*, vol. 496, pp. 79–93, 2016.

[177] L. Reinhard, H. Mayerhofer, A. Geerlof, J. Mueller-Dieckmann, and M. S. Weiss, "Optimization of protein buffer cocktails using Thermofluor," *Acta crystallographica. Section F, Structural biology and crystallization communications*, vol. 69, no. Pt 2, pp. 209–214, 2013.

[178] N. R. Hajizadeh, D. Franke, C. M. Jeffries, and D. I. Svergun, "Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data," *Scientific Reports*, vol. 8, no. 1, p. 7204, 2018.

[179] B. Rost and C. Sander, "Prediction of Protein Secondary Structure at Better than 70% Accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 584–599, 1993.

[180] S. Chojnacki, A. Cowley, J. Lee, A. Foix, and R. Lopez, "Programmatic access to bioinformatics tools from EMBL-EBI update: 2017," *Nucleic Acids Research*, vol. 45, no. W1, pp. W550–W553, 2017.