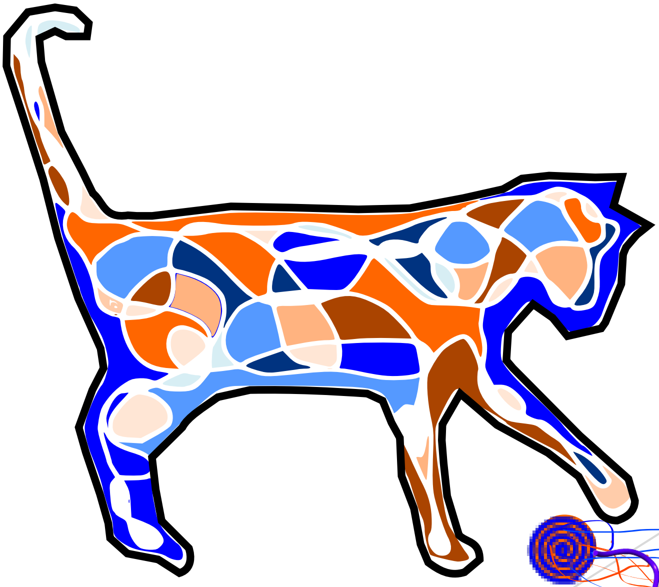
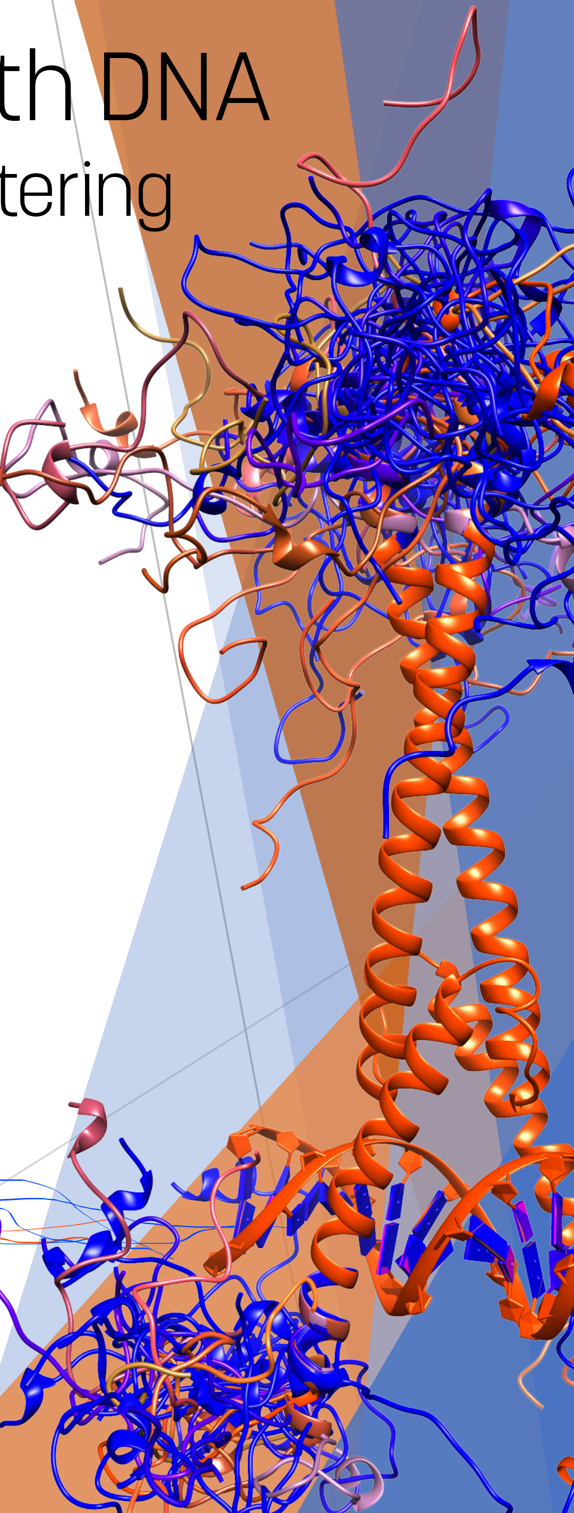
Linköping Studies in Science and Technology
Dissertation No. 2155

MYC and MexR interactions with DNA
A Small Angle Scattering perspective

Francesca Caporaletti





LINKÖPING UNIVERSITY

Linköping Studies in Science and Technology
Thesis No. 2155

MYC and MexR interactions with DNA - a Small Angle Scattering perspective

Francesca Caporaletti

Protein Chemistry
Department of Physics, Chemistry, and Biology (IFM)
Linköping University, SE-581 83 Linköping, Sweden

Linköping 2022

In the front page there is an ensemble pf MAX:MAX:DNA, and in the rear page a 2D image of the neutron detector at D22

During the course of research underlying this thesis, Francesca Caporaletti was enrolled in IGS (ILL Graduate School), a multidisciplinary doctoral program.

© Francesca Caporaletti
ISBN 978-91-7929-417-5 print
ISBN 978-91-7929-608-7 PDF
ISSN 0345-7524 No. 2155

Printed by LiU-Tryck 2022

Ma s’io avessi previsto tutto questo, dati causa e pretesto, le attuali conclusioni
Credete che per questi quattro soldi, questa gloria da stronzi, avrei scritto canzoni?
Va beh, lo ammetto e mi son sbagliato e accetto il "crucifige" e così sia
Chiedo tempo, son della razza mia, per quanto grande sia oh, il primo che ha studiato
Ma s’io avessi previsto tutto questo, dati causa e pretesto, forse farei lo stesso
Mi piace far canzoni e bere vino, mi piace far casino, poi sono nato fesso
E quindi tiro avanti e non mi svesto dei panni che son solito portare
Ho tante cose ancora da raccontare per chi vuole ascoltare e a culo tutto il resto

Francesco Guccini, L’avvelenata

TRAD: Frankly, if I had expected all of this, data reasons, pretexts and present conclusions. Do you think that for these four pennies, for this useless glory, would I have written songs? Okay, I admit I made a mistake, and I accept the Crucifixion and so be it. I ask for time, I am of my kind, as many as we are, the first who studied. But frankly, if I had expected all of these: data reasons and pretexts. Maybe I would make the same thing I love to make songs and to drink wine I love to party And also I am born a bloody idiot. And so I move on, and I don’t shuffle Off the clothes I usually wear I have so many things to tell to anyone who will listen.

Abstract

Protein-DNA complexes govern transcription, that is, the cellular mechanism that converts the information stored in the DNA into proteins. These complexes need to be highly dynamic to respond to external factors that regulate their functions in agreement with what the cells need at that time. Macromolecular X-ray crystallography is very useful for structural studies of large molecular assemblies, but its general application is limited by the difficulties in crystallising highly dynamic and transient complexes. Furthermore, crystal lattices constrain the macromolecular conformation and do not entirely reveal the conformational ensemble adopted by protein-DNA complexes in the solution.

Small Angle X-Ray Scattering (SAXS) and Small Angle Neutron Scattering (SANS) are two complementary techniques known jointly as Small-angle Scattering (SAS). SAS is a powerful tool for analysing the shape and changes of molecules in solution in their native state. It is beneficial if the variability of conformation or disorder complements high-resolution methods such as NMR or crystallography. With SANS, we can explore non-crystallisable protein-DNA complexes in solution without restrictions of artificially symmetrised DNA and limitations of a protein sequence. Neutrons are well-suited probes for studying protein-DNA complexes for the capability of the neutrons to scatter common atoms in biomolecules differentially. They can thereby distinguish between hydrogen and deuterium. Together with varying the solvent deuterium ratio, the contrast variation approach can reveal

shapes of distinct components within a macromolecular complex.

This thesis aims to explore unchartered territories of regulatory protein-DNA interactions by studying such complexes by SAS, with a specific focus on the flexibility of the complexes. In my study of the MexR-DNA complex, I try to elucidate the molecular mechanism by which the MexR repressor regulates the expression of the MexAB-OPrM efflux pump through DNA binding. This pump is one of the multidrug-resistant tools of the pathogen Pseudomonas Aeruginosa (P. Aer.) can extrude antibacterial drugs from the bacteria enabling them to survive in hostile environments. In the second project, I strive to explore the MYC:MAX:DNA complex. This heterodimer assembly functions as a central hub in cellular growth control by regulating many biological processes, including proliferation, apoptosis, differentiation and transformation. Overexpression or deregulation of MYC is observed in up to 70% of human aggressive cancer forms, including prostate and breast cancers. By combining SAS with biophysical methods, the work presented in this thesis reveals novel information on the shape and dynamics of biomolecular assemblies critical to health and disease.

This thesis comprises five chapters, each dealing with a different aspect of the work in those years. In the first chapter, I will introduce the reader to the motivations of this research, and I will give a brief state of the art of the two projects. In the second chapter, I will give the reader all the theoretical instruments to understand better all the methods used in this thesis; I write first to provide an overview regarding the proteins and their capability to bind other macromolecules. I then will exploit the basics of the small-angle technique, focusing on the neutron contrast variation: the fundamental method used throughout this thesis and the *ab initio* modelling.

The third chapter, Methods, discusses the SAS measurements and the requirements for the experiments themselves, the procedure for the data reduction and the data processing and analysis to obtain the structural information.

In the fourth chapter, I will summarise the results of the submitted papers and my contributions:

1. Small-angle X-ray and neutron scattering of MexR and its complex with DNA supports a conformational selection binding model

2. Resolving the DNA interaction of the MexR antibiotic resistance regulatory protein

3. Upgraded D22 SEC-SANS set-up dedicated to the biology community

4. SAS studies on the regulation of MYC303:MAX DNA and MAX:MAX DNA binding in cancer.

Preface

This thesis results from several years of my life, but I could not do everything by myself. It is a joint job of different people, countries, and organizations. First of all, ILL and LiU are where I have been physically working. The other facilities where I went for the data acquisition, such as ISIS and ANSTO. I found terrific people collaborating here, such as Dr Najet Mahmoudi, Dr Kathleen Wood, and Dr Anthony Duff. I want to thank them for the different points of view that they have given me. However, before everyone, I want to express gratitude to my Swedish supervisor Maria Sunnerhagen for choosing me for this project and allowing me to work on it. To the other supervisor: Björn Wallner, be kind and patient with me, mainly modelling and analysis. Then I want to thank the French side of this project. My ILL supervisor Anne Martel, helps me with all the beamtimes at D22, even during Easter and weekends. So thank you, Anne, for all your time in the D22 cabinet. Also, a special thanks to Frank Gabel was always helpful and present, even if he was just formally my supervisor at the beginning of my work. I also want to thank all the PhD students of ILL because we were a family abroad inside ILL. Another heartful thanks to the other people of Maria Sunnerhagen's group. They have helped me a lot with this project. Thank you, Zuzanna, Vamsi, Vivian, Marion and Johannes. Then, I want to thank my family for trying to be close to me even with hundreds of kilometres of distance and a pandemic in the middle. Furthermore, a special thank you to Flavio for always being my first supporter and helping me to survive those difficult years.

Contents

List of papers included in the thesis

I Caporaletti F., Pietras Z., Morad V., Mårtensson L.G., Gabel F., Wallner B., Martel A., and Sunnerhagen M.
Small-angle X-ray and neutron scattering of MexR and its complex with DNA supports a conformational selection binding model.
Biophysical Journal (2022)

II Pietras Z., Caporaletti F., Jeffries C.M., Morad V., Wallner B., Martel A. , SunnerhagenM.
Structural characterization of the Pseudomonas Aeruginosa MexR–mexR repressor-operator complex: a small-angle X-ray and neutron scattering perspective
Manuscript in preparation

III Martel A., Cocho C., Caporaletti F., Mark Jacques M., Aazzouzi A. E., Lapeyre F. and Porcar L.
Upgraded D22 SEC-SANS set-up dedicated to the biology community
Journal of Applied Crystallography, manuscript in submission

IV Caporaletti F., Moparthi V.K., Martel A., Wallner B., Sunnerhagen M. **Small-angle neutron and x-ray scattering and biophysical analysis of a MYC:MAX:DNA complex including the conserved MYC-MBIV region** Manuscript in preparation

Additional pubblications

1. Caporaletti F., Carbonaro M., Maselli P., Nucara A.
Hydrogen–Deuterium exchange kinetics in β-lactoglobulin (-)-epicatechin complexes studied by FTIR spectroscopy
International Journal of Biological Macromolecules, 2017, 104: 521-526

2. Caporaletti F. , Rubio-Magnieto J. , Lo M. , Longevial J.-F. and Rose C. , Clément S. , van der Lee A. , Surin, M. , Richeter, S.
Design of metalloporphyrins fused to imidazolium rings for binding DNA G-quadruplexes
Journal of Porphyrins and Phthalocyanines 24.01n03 (2020): 340-349

Contribution

Manuscript I: Francesca Caporaletti (FC) designed the research, performed, analyzed the data and wrote the paper.

Manuscript II: FC designed the initial project, performed the SANS data collection, and analysed the corresponding data.

Manuscript III: FC tested the set-up, produced figures and wrote corresponding parts of the manuscript.

Manuscript IV: FC designed the SAS project, executed the SAS data collection, analysed the data, wrote the manuscript and produced the relevant figures.

Main Supervisor

Maria Sunnerhagen, Professor, Department of Physics, Chemistry and Biology (IFM), Chemistry (KEMI), Linköping University, Linköping, Sweden.

Co-supervisors

Björn Wallner, Professor, Department of Physics, Chemistry and Biology (IFM), Bioinformatics (BIOIN), Linköping University, Linköping, Sweden.
Anne Martel , Resercher, Instrument responsible on D22, Large Scale Structures (LSS), Institut Laue Langevin (ILL), Grenoble, France.
Franck Gabel, Resercher, Petite Angle SAS, Institute de Biologie structurale (IBS), Grenoble, France. (Supervisor from 2017 - 2019)

Faculty Opponent

Ann Terry, Researcher and Group manager for Diffraction and Scattering at MAX IV, Lund University, Sweden.

Thesis Examination Board

Marie Skepö, Professor, Department of Chemistry, Division of Theoretical Chemistry, Lund University, Sweden.

Maria Selmer Professor, Department of Cell and Molecular Biology, Division of Structural Biology, Uppsala University, Sweden.

Ingemar André Associate Professor, Department of Biochemistry and Structural Biology, Lund University, Sweden.

CHAPTER 1

Introduction

Introduction

Since the discovery of the DNA structure by Franklin, Watson and Crick [1], it was clear that there was a correlation between structure and function. It was possible to define the structure using the powerful X-ray diffraction technique. X-ray diffraction became a well-established technique for small-molecule crystallography; in particular, Perutz and Crowfoot were pioneers in crystallising proteins [2, 3], and he made it possible to understand the biological events from the molecular level. Nowadays, X-ray diffraction of crystallised proteins is the most used and valuable technique to study the structure of proteins and biological complexes together with Nuclear Magnetic Resonance (NMR) and Cryogenic electron microscopy (CryoEM).

All of them are powerful techniques, but X-Ray crystallography can give great insight into the atomic structure of the macromolecules but requires its crystal; in contrast, NMR spectroscopy studies chemical properties by studying individual nuclei. These are the powers of the methods but sometimes also the weakness.

Those techniques have improved over the years, but there is some problem to overcome: the most prominent is the growth of protein crystals. There is an enormous number of cases in which a single crystal of a substance can be obtained, but unfortunately, in some cases, proteins or complexes are impossible to crystallise; in particular, intrinsically disordered proteins and complexes with DNA, or a combination of the two, like in our case. Consequently, studying those complexes with X-Ray diffraction is impossible, so we have to analyse them in a solution, even if we lose spatial resolution. Small-angle scattering is a powerful tool for analysing the shape and the conformational changes of macromolecules, and it is beneficial in the case of non-rigid structures, where there is the variability of conformation or intrinsically disordered structures, which precludes structural determination via NMR or crystallography [4]. In particular, protein-DNA complexes must be highly dynamic to respond to external factors and execute their function.

In the classic crystal structure, the DNA sequences are often made artificially palindromic. This artefact helps crystallisation and analysis and often creates more rigid complexes. However, they are losing their wholesomeness because *in vivo* DNA-binding proteins usually bind to DNA sequences that are not palindromic. They could serve different functionalities, like overlapping with other operators or regulating sequences. Consequently, in the crystal environment, the protein-DNA complexes poorly represent the authentic conformational ensemble: the group of all structures that the complex can have. With SANS, we can explore non-crystallisable protein-DNA complexes in solution without artificially symmetric DNA restrictions and protein sequence constraints because the proteins are often modified to remove the flexible parts, it allows us thus exploring unknown territories of protein-DNA interactions.

In particular, we will try in this thesis to elucidate the molecular mechanism by which the MexR repressor regulates the expression of the MexAB-OPrM efflux pump through the protein-DNA binding. Moreover, we will strive to explore the MYC:MAX heterodimer assembly functions as a central hub in cellular growth control and explore the difference with MAX:MAX homodimer. The next part will explain deeper the state of the art of those projects.

1.1 Motivation

It is established that proteins and DNA interactions are fundamental for different cellular functions such as life regulation, apoptosis, reproduction, defence, and more. Transcription factors interact with DNA and other proteins involved in regulatory complexes and need to respond swiftly and accurately to changes in cellular stimuli. Nevertheless, those proteins are flexible to adapt to a signal and then interact with the DNA.

In particular, we will focus on two complexes responsible for regulating different mechanisms in different organisms. The first one that we will analyse is composed of two dimers[1] of MexR bond to two adjacent pieces of DNA. We want to exploit the reason for those two close domains and if an interaction between the two dimers is involved in these bindings.

The second project is the MYC:MAX dimer, and our goal for this project are to distinguish between the MYC:MAX heterodimer and the MAX:MAX homodimer. This project is particularly challenging because the proteins involved have "flanking regions" that are intrinsically disordered. However, the information we can have from those flanking regions and how they interact with the rest of the complex and with the DNA can give us essential input to comprehend how the heterodimer's overexpression over the homodimer is involved in carcinogenesis.

[1] a protein dimer is a macromolecular complex formed by two protein monomers which are usually non-covalently bound

1.2 Structures determination of the MexR-DNA complexes

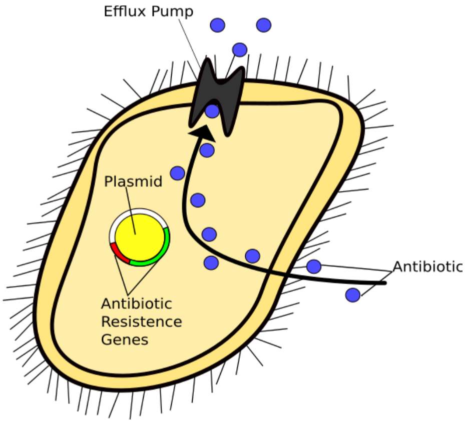


Figure 1.1: **MexAB-oprM efflux pump in *P. Aer.*** The MexAB-oprM efflux pump is one of the defensive mechanisms of *P. Aer.*

The common gram-negative[2] bacterium *Pseudomonas aeruginosa* is a conditional pathogen that causes severe nosocomial[3] infections [5]. *Pseudomonas aeruginosa* express a low level of MexAB-OprM transporter, which provides decreased susceptibility to multiple species of antibiotics [6]. Moreover, it has been described as a 'priority pathogen' by the World Health Organization (WHO) [7]. Infection of immunocompromised patients by *P. Aer.* is of increasing concern in hospitals with a particular interest in the actual situation where we have reached the total capacity of intensive care in hospitals worldwide due to the COVID-19 [8]. The problem with this organism's infections is its broad resistance to structurally and functionally diverse antibiotics. This type of multidrug resistance is mainly attributable to an interplay of low outer membrane permeability and expression of efflux pumps such as MexAB-OprM, MexAB-OprM, MexJK-OprM, MexEF-OprN,

[2] Gram-Negative bacteria are characterised by their cell envelopes, which are composed of a thin cell wall sandwiched between an inner cytoplasmic cell membrane and a bacterial outer membrane

[3] Hospital-acquired infection

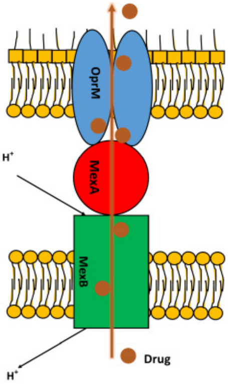


Figure 1.2: **The structure of MexAB-oprM efflux pump in *P. Aer.*** The pump extends from the extracellular surface to the cytoplasm, and it can efflux antibiotics from the cytoplasm outside the cell in exchange for protons.

MexXY-OprM, MexCD-OprJ, and MexVW-OprM, which recognise harming substances and expulse them from the cell [9]. The MexAB-oprM operon encodes three pump subunits [9]: the intrinsic inner membrane protein, MexB, the inner membrane-associated periplasmic lipoprotein, MexA, and the outer membrane-associated lipoprotein OprM.

MexR is a member of the MarR family of bacterial transcriptional regulators. It is the repressor for the MexAB-OprM efflux pump and inhibits the expressions of those subunits. Consequently, dysfunction of MexR, leading to a derepression[4] of MexAB-oprM expression can contribute to rendering *P. Aer.* resistant to multiple clinical essential antibiotics, such as quinolones, β-lactams, tetracycline, chloramphenicol, and novobiocin [10, 11].

Structurally and functionally, MexR is a member of the large MarR family of transcriptional regulators. MarR homologs are winged-Helix-Turn-Helix (HTH) DNA-binding proteins that exist as dimers and bind palindromic sequences within cognate promoters. They gener-

[4] Derepression is the removal of the repression. The repression is a mechanism often used to decrease or inhibit the expression of a gene.

ally have a triangular shape with pseudo-two-fold symmetry. The MexR dimer structure is shown in figure 1.3. The amino- and carboxy-terminal helices interdigitate to create a dimerisation interface that dictates the distance between the DNA recognition helices, thus controlling DNA-binding specificity and affinity [12]. Furthermore, to rapidly respond to external stimuli, MarR proteins need to react quickly to external regulators (ions, small molecules) and bind DNA at sites that overlap the promoter initiation sites [13, 14]. These inherent functional dynamics within the MarR family may contribute to their surprising resistance to crystallisation with DNA.

MexR, like the other MarR proteins, is a dimer composed principally of α-helix, as shown in figure 1.3, and its characteristic is to be significantly flexible in the dimerisation domain. It has for each monomer a winged-helix domain that consists of $\alpha 2$ (H1)-$\beta 1$ (S1)-$\alpha 3$ (H2)-$\alpha 4$ (H3, recognition helix)-$\beta 2$ (S2)-W1 (wing)-$\beta 3$ (S3) [11]. A range of crystal structure conformations has been observed for MexR, suggesting extensive structural variability in its apo form. The first findings of four copies of MexR dimer in multiple conformations suggested that an effector-induced conformational change may inhibit DNA binding by reducing the spacing of the DNA binding domains [11]. In the crystal structure of MexR dimer with its antirepressor ArmR, the latter occupies the cavity and the centre of MexR dimer, stabilising a conformation incompatible with DNA-binding [15]. Meanwhile, in the same year, Chen *et al.* have shown computationally that the formation of disulfide bonds is consistent with a conformational change that prevents the oxidates MexR from binding to the DNA. Consequently, the regulation of DNA binding also responds to stress-induced cellular changes, which promote CYS-crosslinked inactive conformations for both MarR and MexR [16]. After a couple of years, they have also crystallised the oxidated MexR. It is shown that the distance between the Arg73 - Arg'73 is practically identical between the apo-MexR and the oxidated one ($\approx$ 29 Å), but still, it prevents the DNA binding. They have deduced that MexR reacts to the oxidative stress produced by many antibiotics resulting in the oxidation of the redox-sensing cysteine. Disulphide bond formation induces a conformational change in the HTH DNA-binding domain, abolishing MexR's ability to bind the operator [17].

The HTH DNA-binding motif is extensively used by transcriptional regulators in prokaryotes and eukaryotes, enabling varied, efficient and versatile DNA binding. Crystallography has broadly determined their structures, particularly as homodimers to palindromic DNA sequences. Extended HTH motifs, by a third helix and an additional b-finger, have been named winged-HTH motifs, adding specificity by extending the DNA contact surface [18].

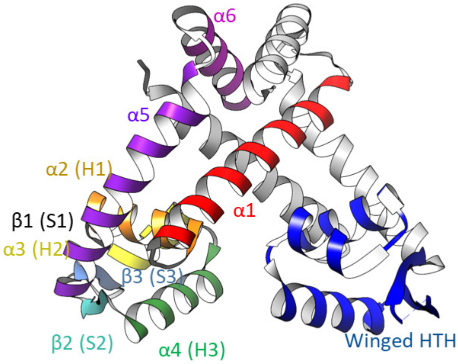


Figure 1.3: **Structure of the MexR dimer.** MexR ribbon in flat ribbon representation (PDB:1LNW chain:CD). Each secondary structure element is individually labelled and coloured on the left monomer. The winged HTH-motif is coloured in blue in the monomer on the right, and it consists in $\alpha2$ (H1)-$\beta1$ (S1)-$\alpha3$ (H2)-$\alpha4$ (H3, recognition helix)-$\beta2$ (S2)-W1 (wing)-$\beta3$ (S3).

It is shown that a multidrug-resistance-conferring MexR mutant with poor DNA binding stabilises a "closed state" of the MexR homodimer [19]. However, our solution and molecular dynamics simulations show that MexR can access a much wider family of states in the absence of DNA than in the crystal structures [19]. This larger unbound ensemble includes states that resemble one of the few high-resolution MarR-DNA complexes with DNA – the OhrR-DNA complex [19, 20]. However, whether the MexR-DNA complex resembles the OhrR-DNA complex is unknown. MarR-family DNA complexes do not readily crystallise, possibly due to the dynamics these proteins need to retain to respond to external factors. Their DNA target sequences overlap with non-symmetric promotor regions, and NMR spectra of the

MexR-DNA complex were insufficient for structural analysis.

According to *Evans et al.*, [21], the DNA binding site for MexR dimer appears to involve an inverted repeat of the sequence GTTGA as we can see from figure 1.4. Furthermore, it is possible to identify two closely located bindings fragments, dubbed PI and PII [21]. In this work, we will use Small Angle Scattering (SAS) to reveal the structure of the protein-DNA complex of MexR with its PII and entire FL (PI+PII) operator regions [22]. This analysis is essential for comprehending the binding of MarR family proteins with DNA because very few of those proteins have crystallised with DNA and none with two consecutive dimer sites [23–25]. The ongoing analysis will apply to the entire MarR family and provide critical knowledge on pathogenic bacteria relevant to clinical understanding and future drug design.

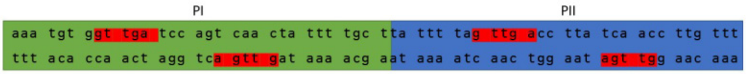

PI
PII
aaa tgt ggt tga tcc agt caa cta ttt tgc tta ttt tag ttg acc tta tca acc ttg ttt
ttt aca cca act agg tca gtt gat aaa acg aat aaa atc aac tgg aat agt tgg aac aaa

Figure 1.4: **Scheme of DNA FL fragment.** This figure represents the FL DNA that binds with the MexR proteins. Green is the PI fragment, and blue is the PII fragment. In red is highlighted the binding site GTTGA.

1.3 Regulation of MYC-MAX DNA binding in cancer

The MYC:MAX heterodimer works as a central hub in cellular growth control by regulating a wealth of biological functions: including proliferation, apoptosis, differentiation and transformation [26, 27]. Mutations increase MYC levels in the cell, disrupting ubiquitination and translocation and increasing MYC:MAX heterodimer formation over the prevailing MAX homodimer. Uncontrolled MYC expression disturbs the carefully tuned balance of cell growth regulation [28], which turns the MYC:MAX heterodimer into an oncoprotein multimodular platform and a key contributor to the development of many, if not most, human cancers [29]. Unlike MYC, MAX can homodimers and bind the E-boxes (Fig. 1.5). At the current status, the precise

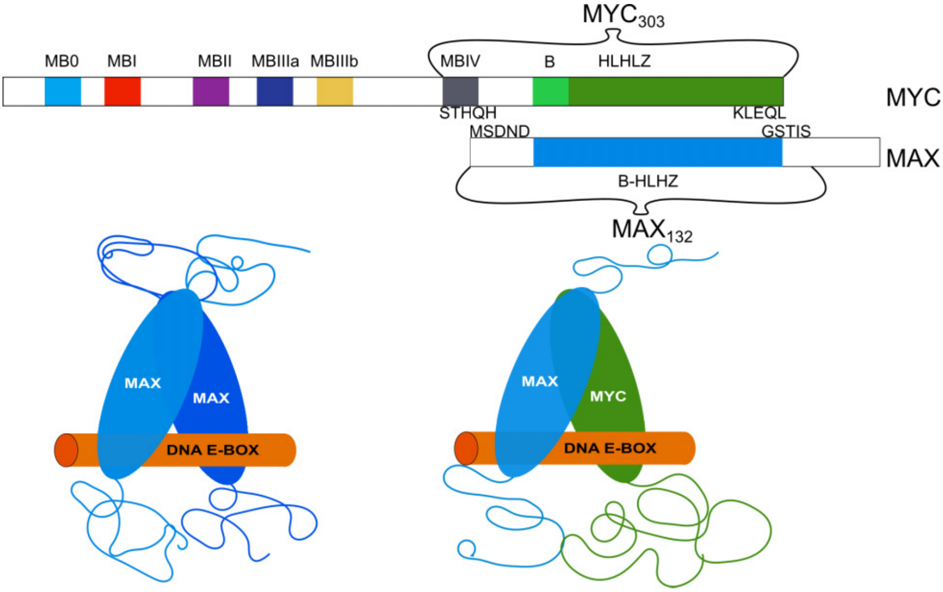


Figure 1.5: **MYC and MAX sequences and DNA bindings.** Sketch of MYC and MAX sequences and heterodimer and homodimer bind with E-box DNA.

role of the MAX homodimer is still unknown; there are suggestions [30] that the induction of MYC upon mitogenic stimulation and the subsequent formation of MYC-MAX heterodimers may activate promoters under transcriptional repression by constitutively expressed MAX homodimers. To bind DNA, the c-terminal region of MYC must form a heterodimer with the protein MAX. Crystal structures describing the MYC:MAX and MAX dimers have only included the core DNA-binding motif, including the bHLHzip region [31–33]. In combined circular dichroism spectroscopy and limited proteolysis approach, we have shown that the flanking regions of the MAX bHLHzip core add helical propensity to the fold, which does not agree with a MAX bHLHzip dimer motif flanked by disordered regions. Furthermore, we found that the full-length MAX homodimer, comprising MAX residues 1-132, is significantly more stable than MAX 18-106 fragment comprising only the bHLHzip region. However, since X-Ray crystallography and NMR have failed to describe entire MAX:MAX or MYC:MAX protein assemblies, structural contributions by regions flanking the core DNA binding motif of the MAX:MAX or MYC:MAX dimers remain unknown. We aim to understand the contribution of those flanking regions.

1.4 Research questions

As we already discuss, there are research questions that we have to formulate to resolve those projects. A complex structure is related to its function, so it is fundamental to understand the structure to relate it to the function. In particular, we have different questions that we want to answer:

1. Does the structure of MexR change from the crystallised form to adapt to the DNA? and if it does, how? Moreover, how is this affecting its flexibility?

2. Whit in the complete sequence of the DNA with the two dimers bound to it. Does the DNA bend? Moreover, how and for which reason? Is DNA's intrinsic flexibility, or is it due to the MexR-dimers' interaction?

3. What is the role of the flanking regions? Do they interact strongly with the DNA? Do they have a structure? Do they have a preferred space region to locate if they are flexible?

4. There is a difference in the MAX:MAX with the DNA's presence? With a focus on the differences between the flanking regions

5. What are the differences between the MAX:MAX:DNA and MYC:MAX:DNA complexes? Is there an explanation for over-expression leading to cancer?

Let us hope that we can answer all those questions through this thesis.

1.5 Delimitations

We have understood that SAS is a powerful technique to resolve the structure of complexes; section 3.1 will explain this in detail. Furthermore, we also understand that it was not possible to use X-ray or neutron crystallography because, due to the flexibility of the complexes, it is not possible to crystallize them. How"ver, why have we chosen SAS and not CryoEM or NMR? CryoEM is an electron microscopy technique applied on an aqueous sample solution that is flash-frozen [34]. It is a different approach from the scattering technique. It is an imaging technique it requires acquiring several 2D "images" of the same complex at different angulations. From the sum of those pictures, we can reconstruct the 3D structure of the complex. We need enough images to resolve several forms of a flexible complex and enough photos to recognize each step. The fundamental limitation of this technique is the wavelength of the electrons. In principle, they can be tuned to any desired wavelength; however, shorter wavelengths are progressively higher in energy. At some point, the energy input into the protein will quickly destroy the sample. Furthermore, it is a pretty new technique, and just recently [35], it was possible to study protein below 100 kDa, in this case, myoglobin. This well-known protein is not intrinsically flexible. Just recently, the wall of the 100 kDa is broken [36]; our complexes are all around this limit. This scientific topic can be a fascinating topic for the future to understand how much CryoEM can be pushed, but it was not the case and the goal of this PhD project. Maybe it could be an interesting next step for the future of those projects. Moreover, I want to remind the reader that I was a joint PhD between LiU and ILL, the most powerful neutron source globally. Being privileged to work directly on the site helps direct the experiment on the neutron site. In the case of NMR, we have the opposite problem. The focus of those experiments is to study something more extensive than what was done before: add the DNA, create the complex with all the DNA bindings sites, and extend as much as possible: we have completed MexR and MAX and elongated MYC, and this was not possible to do it with NMR. For a better understanding of what was already done [19]. In summary:

1. CryoEM is time-consuming in data acquisition and analysis. Furthermore, those complexes are at the actual size limit of the

technique.

2. Nowadays is not a well-used technique for this type of analysis, like more enormous complexes and transmembrane proteins. Using a non-standard method could lead to misinterpretation because there is not enough experience to completely trust the data. Meanwhile, for the NMR, the study of those proteins was already performed, and we intended to extend the knowledge. Furthermore, we were focused on the tertiary and quaternary structures. Moreover, in more enormous complexes, the shift was a complex analysis.

3. for NMR, our complexes are too big and too flexible. Also, it was impossible to study the apoMexR because it is too loose to be interpreted with NMR, and what was possible to do was already done.

4. My PhD was a joint project between Institut Laue Langevin and Linköping University; it was unreasonable not using the privilege of being in the most powerful neutron source until the European Spallation Source (ESS) will enter into function.

CHAPTER 2

Theory

This chapter will focus on understanding the fundamental bricks' entire thesis, particularly the SAS theory. Still it will also explain the other techniques we have used to assess the quality of the sample before irradiating the sample with the beam and complementing the analysis.

2.1 Protein functions and interactions with partners

It is common knowledge that proteins are composed of a precise sequence of amino acids that fold up in a particular 3D shape. However, proteins are not rigid structures, they are dynamics, and some have moving parts whose mechanical actions are connected to chemical events [37]. That allows them to act as catalysts, receptors, switches, motors, or pumps.

The ability of a protein to bind selectively and with high affinity to a ligand depends on forming a set of weak, non-covalent bonds: hydrogen bonds, ionic bonds, van der Waals attractions and hydrophobic interactions. Each of those bonds is weak, so t is required to have an effective binding that many weak bonds have to be formed simultaneously. The binding can happen if the surface of the ligand fits the protein. The region of a protein that binds a ligand is known as the "binding site", which consists of a portion of the protein surface formed by a particular disposition of the amino acids; those amino acids could belong to different parts of the polypeptides' chain, that are located in a defined region for the folding of the protein [38]. Even atoms buried in the protein's interior that seems not to have direct contact with the ligands can provide an essential structure to the protein surface, and its chemical properties [37]. Some changes in the amino acid chain, even in the protein's interior, can change the shape and destroy the binding site [37]. Therefore, the surface of each protein molecule has a unique chemical reactivity that depends not only on which amino acid side chains are exposed but also on their exact orientation relative to one another. For this reason, even two slightly different conformations of the same protein molecule may differ significantly in their chemistry. However, many protein domains can be grouped into families showing their evolution from a common ancestor. The three-dimensional structures of the members of the same domain family are remarkably similar. These facts allow a method called "evolutionary tracing" to identify those sites in a protein domain that are the most crucial to the domain's function [39]. For this purpose, those amino acids are almost unchanged in all known protein family members. The most invariant positions often form one or more clusters on the protein surface, which generally correspond to ligand binding sites.

2.1.1 Binding strength

Molecules in the cell encounter each other frequently because of their continual random thermal movements. When colliding molecules have poorly matching surfaces, few non-covalent bonds form, and the two molecules dissociate as rapidly as they come together. At the other extreme, when many non-covalent bonds form, the association can persist for a long time. Strong interactions occur in cells whenever a biological function requires that molecules remain associated for a long time. The strength with which any two molecules bind can be measured directly. Suppose there is a population of identical proteins with their ligands in a solution at a random time interval. In that case, one of the ligands will interact with the protein and form via the binding site a protein-ligand complex. Consequently, the complex population will increase over time, but due to the thermal motions, there is also a possibility that the complex could break. Eventually, this creation and destruction of complexes will reach a statistical equilibrium state, in which the rate over time of complex creations, which is called association, is the same as complex destruction, which is called dissociation. Suppose the concentrations of protein and ligand are known. In that case, it is possible to calculate the equilibrium constant K, which is a quantitative way to measure the strength of the binding. It is a direct measure of the free-energy difference between the bound and unbound state.

For this thesis, we have used Isothermal titration Calorimetry (ITC). This technique is exploited in chapter 2.5.

2.2 Basics of small-angle scattering

Using X-rays and Neutrons as a probe to investigate the secrets of the matter is a powerful instrument because photons and neutrons interact within the atoms in the sample: they interact with the respective targets' electrons or the target's nuclei. It is possible to consider no energy exchange between the radiation and the atoms if the experiment focuses on studying the macromolecules' structural components. Those experiments are called elastic (or quasi-elastic). Even if the physical mechanism of the elastic X-ray and neutron scattering are different, the mathematical formalism behind the description is the same. Section 2.2.1 will describe the basis of Elastic scattering theory. With the differences between X-ray and neutron, if necessary, we will extend our primary interest in applying the theory to macromolecules. Suppose there is an interest in focusing on an analytical point of view. In that case, it is possible to skip this chapter and go directly to the method part 3.1, where it will be an extended explanation of how to perform the analysis.

2.2.1 Elastic Scattering theory

The neutron beam is an excellent probe for investigating structures. It is possible to modulate the wavelength, and neutrons interact with the

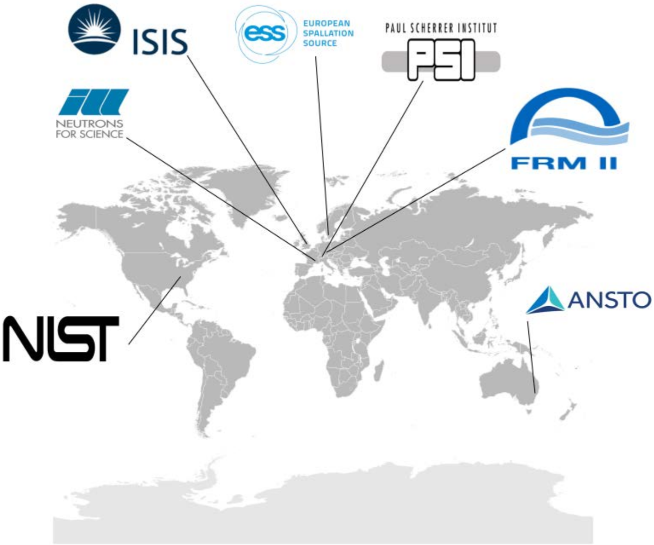


Figure 2.1: **Maps of some major neutron facilities worldwide.** With a particular focus on the European one and the ones used for an experiment or it was developed by the used software.

matter through strong nuclear interaction with the atomistic nuclei of the atoms. They are not charged particles, but they have a magnetic moment [40] (but it is not interesting for this thesis's goal, so we will not focus on that). Neutrons are powerful but difficult to be produced; there are just two ways: spallation and nuclear reactor; and there are few facilities around the world where they can be produced (Fig. 2.1):

- Nuclear Reactor (ILL, NIST, FRMII and ANSTO)
- Spallation Source (ISIS, ESS and PSI)

X-ray is an electromagnet wave, and for the experiment, it has to be monochromatic and plane. For that reason, there are modulators on the synchrotron. Neutrons are particles, but for the dualism particle waves, they can also be considered a plane wave with a wavelength much larger than the atom nucleus's size $\lambda = h/mv$, where m is the mass of the neutron and v is the speed. For example, in a characteristic neutron experiment, the λ is 6 Å, so the velocity of a neutron whose mass is $m = 1.67 \cdot 10^{-27} Kg$ is $v = 660 m/s = 2376 Km/h$. At this speed, we do not have any relativistic effects.

When we perform a scattering experiment, we are doing a transformation from a space in $\vec{r}$ coordinates (the real space) to the scattering vectors $\vec{q}$ of the reciprocal space. The Fourier transformation describes this process. It is essential to focus on the reciprocity between the dimensions; something small in real space is ample in reciprocal space and vice versa. For example, Dirac's delta distribution in real space becomes a constant in reciprocal space. In a simple scattering experiment, we can consider that the target is a single atom; both the distances between the source-sample and between sample-detector are considerable; consequently, both the probe will scatter as the nucleus is a point. The incident wave is represented as:

$$\phi(\vec{z}) = e^{i\vec{k_0}\vec{z}} \tag{2.1}$$

Where $\vec{k_0}$ is the vector of the probe and is related to its wavelength by:

$$|k_0| = \frac{2\pi}{\lambda}$$

Because the target is seen as punctiform, the scattered wave is spherical, as it is shown in figure 2.2:

$$\phi(\vec{r}) = -\frac{b}{\vec{r}}e^{i\vec{k}\cdot\vec{r}} \tag{2.2}$$

Where b is the scattering length for the atom, to consider the absorbance of the neutron negligible because the experiment's energies are not enough in consequence, b is a real number[1]; meanwhile, it is not possible to consider the negligible absorbance component of the X-ray. It is possible to relate the scattering length with the cross-section, that is, the ratio between the scattered beam and the incident beam; more information could be found on [41].

$$\sigma = 4\pi b^2 \tag{2.3}$$

So far, we have considered an only-atom system, but an exciting experiment involves several atoms bound to each other. Having more than one atom introduces several features.

[1] It is valid for almost all atoms. One example is Cadmium. This toxic material is a good absorber of neutrons and is used for neutron shields.

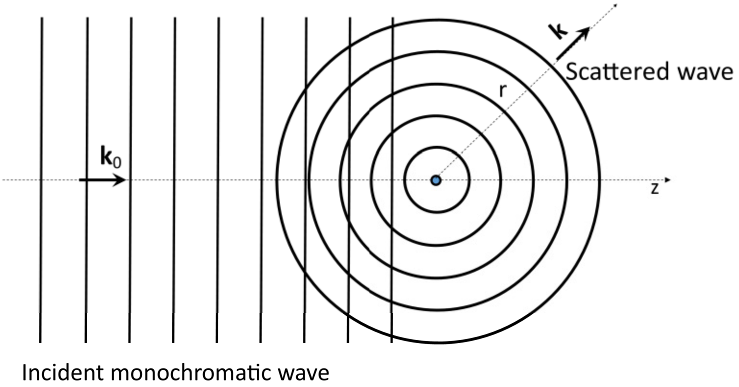


Figure 2.2: **A typical scattering experiment with a punctiform sample.** In the figure, the k_0 represent the plane incident wave, and k represents the spherical diffracted wave. For the principle of Huygens-Fresnel, each point is a spherical wave emitter.

1. There is interference between waves scattered by the nuclei of the sample

2. Only one part of the scattered waves will take part in the interference. Consequently, σ should be divided into two parts: coherent and incoherent cross-section. For the scattering experiment, the incoherent part participates in the background because the scattering is isotropic (for more information, it is possible to look at Allen and Higgins 1973 [42]); meanwhile, the coherent part is the one that causes the interference.

3. The atoms are not rigidly bound by each other. The system is flexible, which can cause a change in energy, which can cause a part of the scattering to become inelastic.

Each atom has different property; from Table 2.1, plus some considerations that we have done previously, we can reveal some crucial factors:

1. The neutron scattering length does not depend on the scattering angle. It is a consequence of the fact that nuclei can be considered points.

Atom	b_{coh} $(10^{-13}cm)$	σ_{inc} $(10^{-24}cm^2)$	$\sigma_{capture}$ $(10^{-24}cm^2)$	$f_{x-ray}(\theta = 0°)$ $(10^{-12}cm)$
^{1}H	-3.74	80.27	0.18	0.28
^{2}H (D)	6.67	2.05	0.00	0.28
^{12}C	6.65	0.00	0.00	1.69
^{14}N	9.37	0.50	0.99	1.97
^{16}O	5.80	0.00	0.00	2.25
^{31}P	5.17	0.3	0.11	4.23
^{32}S	2.80	0.00	0.07	4.5

Table 2.1: **Neutron scattering length** The atoms involved in biological macromolecules are listed in the table. The first column is the neutron coherent length scattering. The second is the incoherent neutron cross-section, the third is the neutron capturing cross-section, and the fourth and last is the cross-section for x-Ray at 0° angle. $\sigma_{capture}$ is proportional to wavelength and here is given for 1 Å[43].

2. The neutron scattering length has the same order of magnitude, pretty different from X-Ray, where H is much smaller than C, O or N, so neutrons are more sensitive to the presence of H than X-ray.

3. There is a big difference between H and D scattering lengths. As we can see from the Table, H's scattering length is negative; meanwhile, D is positive. So it is possible to replace H with D on the label without adding heavy metal different macro-molecules. It is a crucial point of the application of neutron scattering in biological studies. We will explain more in chapter 2.2.3.

Interference effect

As we said in the previous chapters, the atom's interference of the incoherent part is not angle-dependent so the incoherent cross-section will be averaged, resulting in anisotropic scattering. Consequently for an sample with an incoherent cross-section σ_{inc} the scatter intensity I_s is:

$$I_s = I_0 N S \sigma_{inc} \frac{d\Omega}{4\pi} \tag{2.4}$$

where I_0 is the incident neutron flux, $d\Omega$ is the solid angle, S is the surface, and N is the number of atoms per unit:

$$N = N_a \frac{d}{M} e \tag{2.5}$$

Where N_a is the Avogadro number, d is the sample's density, e is the thickness of the sample, and M is the molecular weight. We can manifest that the intensity is not defined from the angle, which can be considered background, so we can assume that there is no interference, and the interference effect can be applied to just the coherent part. If we want to extend this to several atoms, we must sum each atom.

Now we can define the partial coherent cross-section as:

$$\frac{d\sigma_{coh}}{d\Omega} = \frac{I_s(\Omega)}{I_0 N S} \tag{2.6}$$

Then:

$$\frac{d\sigma_{coh}}{d\Omega} = \sum_{i,j} b_i b_j exp[i\vec{Q} \cdot (\vec{r_i} - \vec{r_j})] \tag{2.7}$$

Where $\vec{Q}$ is the scattering vector, and its module is defined as:

$$|\vec{Q}| = \frac{4\pi}{\lambda} sin\theta \tag{2.8}$$

Where 2θ is the scattering angle and λ is the wavelength. The equation 2.7 is the basis of the structural determination using the SAS. In this case, we have approximated that the inelastic effect is negligible; consequently, we never consider that the probe can change energy. Moreover, this is the basis of the SAS technique, which is an elastic experiment. When we have the sample in solution, there is no preferred orientation; with those conditions, it is impossible to get information at an atomistic level. So we must consider the molecule as a continuous medium, and we have to define a local Scattering Length Density (SLD):

$$\rho(r) = \frac{1}{v} \int_{v} b_i(r) d^3 r \tag{2.9}$$

The understanding of how for each macromolecule, the SLD is changing at different percentages of D_2O is fundamental for the comprehension of how it is possible to resolve the structure of a complex separating the various components and it is what makes SANS so powerful. In fact, on these values, there is the most significant difference between neutron and X-rays in SLD because it depends on the physic of the interaction of the two types of probes. The x-rays interact only with the electrons because they have a smaller mass than protons and neutrons and have a negative charge. So if we consider an atom with a radial density $\rho(r)$:

$$b_x(q) = r_0 4\pi \int \rho(r) r^2 \frac{sin(qr)}{qr} dr \tag{2.10}$$

Since $lim(sin(x)/x)_{x\to 0} = 1$, so $b_x(0) = Zr_0$ where Z is the atomic number. consequently, the heavier the atom, the stronger the X-ray will interact. Instead of b, the scattering factor f is used, which is the ratio between the scattered amplitude of the scattered object and the single electron's scattering at the same condition. On the other hand, neutrons interact with nuclear potential and spin. In the case of biological interest, the spin component is negligible. The scattering length $b_{neutron}$ does not increase with the atomic number, but it is sensible to the isotopes, focusing on the hydrogen isotopes. The peculiarity of the neutrons to recognise also the light atoms like hydrogen is well-used in neutron crystallography. Combined with X-Ray crystallography, it can give complete information regarding the structure. Nevertheless, the most exciting feature of the neutrons for structural research is the difference in b between H ($-0.376 \cdot 10^{-12}$ cm) and D ($0.667 \cdot 10^{-12}$ cm). This difference will provide an effective tool for selective labelling and contrast variation in neutron scattering and diffraction.

2.2.2 Scattering of solution of macro-molecules

It is easier to describe an assembly of atoms in terms of their scattering length density distribution $\rho(r)$. It is possible to assume that the solvent is described with a constant scattering density ρ_s. The scattering amplitude is defined as:

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

Where the integration is performed over the particle volume V and $\Delta\rho(r) = \rho(r) - \rho_s$, from the equation 2.11 is possible to reveal that the amplitude is a complex number, so it is impossible to measure the amplitude directly, it is possible to measure just the intensity that is the squared module of the amplitude $I(q) = A(q)A^*(q)$. Consequently, the information regarding the phase is lost. If one considers identical particles' ensemble, the total scattering will depend on their distribution in the solution. The total scattering density of the irradiated sample will be the convolution between the particle density distribution $\Delta\rho(r)$ and a function that describes the position and the orientation of the particles, which we can call $d(r)$, so $\Delta\rho_{total}(r) = \Delta\rho(r) * d(r)$. In the reciprocal space, a convolution becomes a simple multiplication so: $A_{total}(q) = A(q)F(q)$, and for the same consequence also, the intensity is a simple multiplication of two elements: $I_{total}(q) = I(q)S(q)$ where the first term is describing the structure of the macromolecule. Meanwhile, the second is their distribution into the reciprocal space. Two extreme cases will be considered: the first is a perfect crystal, with all the particles perfectly distributed in the space; the second is when the particles are randomly distributed in both position and orientation. In the crystal's case, the scattering amplitude of the individual particles is summed, so if the crystal gets irradiated by coherent radiation, it will scatter coherently. The total scattering intensity is redistributed along the direction of the crystal's reciprocal lattice on the detector. Those images are called diffraction patterns. The second case (the most interesting for this thesis because it is the one used) is the scattering intensity rather than the amplitude that gets summed; there is no coherent scattering between the wave emitted from the individual molecules. Consequently, on the detector, the intensity of the entire ensemble is a continuous isotropic function proportional to a singular particle averaged over all the orientations with $I_{total}(q) = I(Q) \times S(q)$. Where $I(q)$ represent the particle scattering and $S(q)$ is the interference term, they are also called respectively "form factor" and "structure factor". In SAS is used to study the structure of macromolecules in solution.

2.2.3 Contrast

As we have discussed in the chapter 2.2.2, SAS arises from fluctuations in the density of electron or neutron scattering length in solutions. The difference between the average density of a macromolecule in solution and its solvent is called contrast: $\Delta\rho =< \Delta\rho(r) >=< \rho(r) > -\rho_s$ The scattering density of a macro-molecule can be represented as $\rho(r) = \Delta\rho \times \rho_C(r) + \rho_F(r)$, where $\rho_C(r)$ is a function that is equal to one inside the molecule and zero outside, and $\rho_F(r) = \rho(r) - < \rho(r) >$ represents the fluctuations of the scattering density [44]. If we insert this expression into equation 2.11, the amplitude has two terms: $A(q) = \Delta\rho \times A_C(q) + A_F(q)$, so the intensity can be rewritten as:

$$I(q) = \Delta\rho^2 I_C(q) + 2\Delta\rho I_{CF}(q) + I_F(q) \tag{2.12}$$

Where $I_C(q)$, $I_F(q)$ and $I_{CF}(q)$ are the scattering from the shape, the fluctuations and the cross-term; this equation explains how the intensity is a result of different part and that the shape and the internal part of the structure of the macromolecules could be separated using measurements at different solvent densities $\Delta\rho$. This technique is called contrast variation. In a SAS experiment, the solvent, in general, is H_2O or aqueous solutions containing salt. It is possible to manipulate X-Ray contrast with the labelling by heavy atoms, but it is a difficult task, and sometimes, it ends up in failure. Alternatively, adding a large quantity of salt in the solvent but changing the buffer can also change the structure or lead to aggregation. So X-Ray contrast variation is, in general, challenging, and it may influence the conformation of the macromolecules. Contrast variation is most effective and reliable in SANS. It is possible to module the ρ simply by changing Hydrogen with Deuterium either in the solvent or in the macromolecules. Usually, those modifications are minimal, and there is no influence on the molecular structure or the interaction between the complex components. However, there are some cases where the deuteration has altered the macromolecule's structure, and in general, adding D_2O in the solvent tends to increase the probability of proteins' aggregation. In figure 2.3, it is possible to see how the ρ of the different macromolecules is changing as a function of the deuterium in the aqueous environment. The slope is due to the exchange of the labile proton with the ones in the water. The protons that tend to exchange are bound to electronegative atoms such as oxygen, nitrogen and sulphur. Besides, to change ρ of the buffer, it is possible to highlight a selected structural fragment of the complex particle with the deuteration. The deuteration can be done by growing cells and bacteria in the deuterated medium; in those conditions, both the exchangeable and the not exchangeable hydrogens can be replaced by deuteriums. In figure 2.3, it is evident that except perdeuterated (that means that

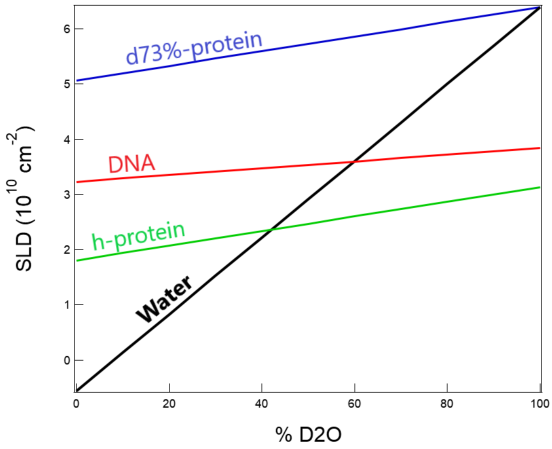


Figure 2.3: **Contrast variation experiment in SANS.** Schematic of a contrast variation experiment in SANS for a protein–DNA complex using different ratios of H_2O/D_2O. We can see that the biomolecules scatter a specific %D_2O the same as the buffer. The graph represents how the scattering length densities over the percentage of D_2O vary for various biomolecules [45].

the totality of the H is substituted with D) protein and DNA that exist a H_2O/D_2O ratio at which the macromolecule has the same SLD at the buffer. This point is known as Match Point (MP), where the contrast is approximately 0.

2.3 Modelling and *ab initio* recognition

Protein modelling and experimental protein structure determination are strictly related; our primary goal is reconstructing the 3D atomic-level structure. However, in our case, we do not have this resolution for the experiments ($\approx$ 3Å), but it is still possible to recognise the structure of the particle in the study. Still, in this case, the modelling is a pivotal point to be able to recognise the structure. Furthermore, it is possible to evaluate *ab initio* the structure with different software: Polynomial Expansions of Protein Structures and Interactions (PEPSI), CRYSON and CRYSOL, and it is possible to evaluate with the modelling done separately.

2.3.1 Multipole expansion

The SLD can be expressed as a sum of spherical harmonic (multipole expansion) as:

$$\rho(r) \approx \sum_{l=0}^{L} \sum_{m=-1}^{l} \rho_{lm}(r) \Upsilon_{lm}(\omega) \qquad (2.13)$$

$$\rho_{lm}(r) = \int_{\omega} \rho(r) \Upsilon^*_{lm}(\omega)\, d\omega$$

where r and ω are spherical coordinates and $\rho_{lm}(r)$ are the radial function. The value L define the accuracy of the expansion. Consequently, also the scattering amplitude can be represented by the spherical harmonics:

$$A(q) = \sum_{l=0}^{L} \sum_{m=-1}^{l} A_{lm}(q) \Upsilon_{lm}(\omega) \qquad (2.14)$$

$$A_{lm}(q) = i^l \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} j_l(qr) \rho^*_{lm}(r)\, dr$$

So the intensity will become [2]:

$$I(q) = \sum_{l=0}^{L} I_l(q) = 2\pi^2 \sum_{m=-1}^{l} \sum_{l=0}^{L} |A_{lm}(q)|^2 \qquad (2.15)$$

So the scattering intensity of a macromolecule is a sum of the orthogonal contribution from different spherical harmonics. The property of the spherical harmonics to be orthogonal one each other results in the absence of the cross-terms is an essential property of multipole expansion. The equation 2.15 allows a rapid compute scattering pattern from a given structure, but also, for the inverse problem, reconstructing the structure from a given scattering profile. In figure 2.4,

[2] the average of the cross term are 0 for the orthogonality of the Υ functions, it is a characteristic of the spherical harmonics.

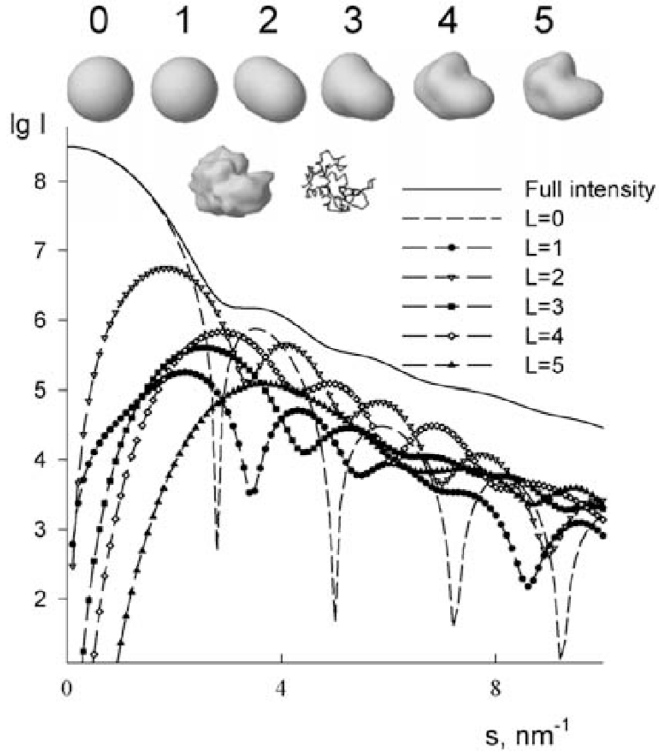


Figure 2.4: **Shape representation using spherical harmonics.** On the top row are shown the surface representations of the truncated envelope functions of lysozyme. The graph represents lysozyme's associated shape scattering intensity and the contributions from different multipoles. [46]

there is a vivid representation of how the sum of spherical expansion becomes the scattering profile.

The spherically averaged scattering intensity $I(q)$ from a single molecule immersed in a solvent with a Length Scattering density (LSD) ρ can be written as:

$$I(q) =< |A_a(q) - \rho A_c(q) + \delta\rho A_b(q)|^2 >_\Omega \tag{2.16}$$

where $A_a(q)$ is the amplitude from the molecule in vacuum, $A_c(q)$ is the amplitude from the excluded volume and $A_b(q)$ is that from the hydration shell.

2.3.2 *Ab initio* shape analysis

In 1970, Stuhrmann [44] proposed to represent the particle shape by spherical coordinates. The particle density was unity inside the envelope and zero outside.

$$\rho(r) = \begin{cases} 1 & if \quad 0 \leq r < F(\omega) \\ 0 & if \quad r \geq F(\omega) \end{cases} \tag{2.17}$$

The envelope was described by a series of spherical harmonics, as we have seen in the previous paragraph:

$$F(\omega) = \sum_{l=0}^{L} \sum_{m=-1}^{l} f_{lm}(q) \Upsilon_{lm}(\omega) \tag{2.18}$$

where the maximum order of harmonic L defines the resolution. This approach was further developed by Svergun [47], who proposed algorithms for rapid computation of scattering intensities from such a model and implemented them in the program SASHA [48]. The angular envelope function's modelling has limitations in describing complicated, e.g. very anisometric, particles or internal cavities. A more comprehensive description is achieved in the bead modelling methods, which use the improved speed of modern computers to revive many parameter modelling strategies with a Monte Carlo-type search. The *ab initio* bead modelling in a confined volume was first proposed by Chacon *et al.* [49]. The maximum dimension D_{max} of a particle is readily obtained from the scattering pattern, and the particle must fit inside a sphere of this diameter. If one fills the sphere with densely packed beads that are spheres of radius $r_0 \ll D_{max}$, each of these beads may belong either to the particle or to the solvent, and the particle shape is described by a string, X, of M bits. Scattering intensity from the bead model is computed:

$$I(s) = f^2(s) \sum_{i=1}^{M} \sum_{j=1}^{M} X_i X_j \frac{sin(sr_{ij})}{sr_{ij}} \tag{2.19}$$

Where r_{ij} is the distance between the beads and $f(s)$ is the form factor of a sphere of radius r_0. Starting from a random distribution of 1 and

0, the model is modified to find the binary array that fits the experimental data using a genetic algorithm. In a more general approach, the beads may belong to different components so that the shape and internal structure of multi-component particles can be reconstructed by simultaneously fitting scattering data at different contrasts (MONSA) [50]. The procedure degenerates to an *ab initio* shape determination for single-component particles. The model intensity is computed using spherical harmonics to speed up the computations. Compactness and connectivity constraints are imposed in the search, implemented in the simulated annealing program DAMMIF [51]. Those programs are all Monte-Carlo-based approaches, so running these programs several times on the same data starting from different initial approximations may yield different final models.

2.4 Dynamic Light Scattering

Dynamic Light Scattering (DLS) is based on the Brownian motion of dispersed particles. When particles are dispersed in a liquid, they move randomly in all directions. The principle of Brownian motion is that particles constantly collide with solvent molecules. These collisions cause a certain amount of energy to be transferred, which induces particle movement. The energy transfer is more or less constant and therefore has a more significant effect on smaller particles. As a result, smaller particles move at higher speeds than larger particles. If all the parameters are known, it is possible to determine the hydrodynamic diameter by measuring the speed of the particles.

In a dynamic light scattering instrument, the laser light encounters the macromolecules in solution, the incident light scatters in all directions, and the detector located at 90° degrees will collect a part of the scattered light. The Doppler effect affects the scattered light because the molecules are in a constant motion [52]. Consequently, the scattered light will result in a mutually destructive or constructive phase. On a DLS experiment, the $G_2(\tau)$ [3] is acquired; that describes the motion of the macromolecules and can be expressed as an integral over the product of the intensity at t and $t+\tau$:

$$G_2(\tau) = \langle I(t)I(t+\tau)\rangle \tag{2.20}$$

[3] second order correlation function

Where τ is the lag time. the normalization of $G_2(\tau)$ is:

$$g2(\tau) = \frac{\langle I(t)I(t+\tau)\rangle}{\langle I(t)\rangle^2} \tag{2.21}$$

Unfortunately, it is impossible to know how each particle moves in solution. However, the motion of the particles relative to each other is correlated employing an electric field correlation function $G_1(\tau)$[4], which describes the correlated particle movent, similarly to the equation 2.21 can be normalised:

$$g1(\tau) = \frac{\langle E(t)E(t+\tau)\rangle}{\langle E(t)E(t)\rangle} \tag{2.22}$$

where $E(t)$ and $E(t+\tau)$ are the scattered electric field at t and $t+\tau$ The $g_1(t)$ and $g_2(t)$ are related with the Siegert relation [53]:

$$g2(\tau) = B + \beta|g_1(\tau)| \tag{2.23}$$

Where B is the baseline (in general, its value is 1) and β is the coherence factor, it depends on the detector, the machine's alignment and some sample properties.

For the perfect monodisperse particles, $g_1(\tau)$ is an exponential decay, so the intensity correlation function is:

$$g2(\tau) = 1 + \beta e^{-2\Gamma\tau} \tag{2.24}$$

Nevertheless, usually, the systems are not monodisperse but polydisperse, so $g_1(\tau)$ cannot be represented as a single exponential but with an integral over the distribution of the decay rates $\rho(\Gamma)$:

$$g_1(\tau) = \int_0^{\inf} \rho(\Gamma)e^{-\Gamma\tau}d\Gamma \tag{2.25}$$

Where:

$$\Gamma = -Dq^2 \tag{2.26}$$

And:

[4] first order correlation function

$$q = \frac{4\pi\eta}{\lambda} sin\left(\frac{\theta}{2}\right) \tag{2.27}$$

where θ is the angle where the detector is located and λ is the laser wavelength.

Nevertheless, how is the diffusion coefficient related to the dimensions of the hydrodynamic radius?

The Stokes-Einstein equation gives the relation between the speed of the particles and the particle size (Equation 2.28). The translational diffusion coefficient gives the speed of the particles D. The equation includes the viscosity of the dispersant and the temperature because both parameters directly influence particle movement. A fundamental requirement for the Stokes-Einstein equation is that the particles' movement must be solely based on Brownian motion. If there is sedimentation, there is no random movement, which would lead to inaccurate results. Therefore, the onset of sedimentation indicates the upper size limit for DLS measurements. In contrast, the lower size limit is defined by the signal-to-noise ratio. Small particles do not scatter much light, which leads to a poor measurement signal.

$$D = \frac{k_B T}{6\pi\eta R_H} \tag{2.28}$$

D is the translational diffusion coefficient, k_B is the Boltzman constant, η is the viscosity, and R_H is the hydrodynamic radius. So it is possible from this relation to obtain the average radius of the macromolecule dissolved.

2.5 Isothermal Titration Calorimetry (ITC)

ITC is used to determine the thermodynamic parameter of the interaction of macromolecules in solutions [54]. It provides the binding equilibrium by measuring the heat evolved on the association of a ligand with its partner. In a single experiment, the values of the binding constant ($K_a = 1/K_d$), the stoichiometry (n), and the enthalpy of binding (ΔH_b) between two or more molecules in solution are determined. The association constant determines the free energy and the entropy of binding. The temperature dependence of the ΔH_b parameter, measured by performing the titration at several temperatures, provides the heat capacity (ΔC_p) [55]. From these initial measurements, Gibbs free energy changes ΔG and entropy changes ΔS can be determined using the relationship:

$$\Delta G = -RT \ln K_a = \Delta H - T\Delta S \tag{2.29}$$

Where R is the gas constant, and T is the absolute temperature). The advantage of ITC is that the visual signal is the change of heat on complex formation; this measure, contrary to the other techniques, does not require any labelling.

To optimize the ITC measurement, the receptor concentration in the cell of the instrument and the ligand loaded in the syringe should be well chosen. The receptor concentration appropriate for an ITC experiment can be estimated using the dimensionless constant c:

$$c = N \cdot K_a \cdot C_0 \tag{2.30}$$

Where N is the stoichiometry of the reaction, K_a is the association constant (M^{-1}), and C_0 is the initial concentration of the receptor in the sample cell (M). , it is recommended to use an initial concentration of the receptor C_0 that results in a c value between 20 and 200 to obtain a well-defined binding isotherm with two plateaux [56].

An isothermal titration calorimeter comprises two identical, highly efficient, thermally conducting and chemically inert material cells surrounded by an adiabatic jacket. Thermocouple circuits detect temperature differences between the reference cell, filled with buffer or water, and the sample cell containing the macromolecule. The ligand is titrated into the sample cell during the experiment, causing heat to be either taken up or released. Measurements consist of the time-dependent input of power required to maintain equal temperatures between the sample and reference cells.

In an exothermic reaction, the temperature in the sample cell increases upon the addition of the ligand. Consequently, the feedback power to the sample cell has to be decreased to maintain an equal temperature between the two cells. In an endothermic reaction, the opposite occurs; the feedback circuit increases the power to maintain a constant temperature.

The raw data is the power needed to sustain the reference and the sample cell at an identical temperature against time. As a result, the graph consists of a series of spikes of heat flow, with every spike corresponding to one injection. These heat flow spikes are integrated concerning time, giving the total heat exchanged per injection — the pattern of these heat effects as a function of the molar ratio between ligand and macromolecule. Degassing samples is often necessary to obtain good measurements, as gas bubbles within the sample cell will lead to abnormal data plots in the recorded results.

CHAPTER 3

Methods

In this chapter, we will discuss the SAS measurements, the requirements for the experiments themselves, the procedure for data reduction and the data processing and analysis to obtain the structural information.

3.1 Small Angle Scattering experiments

A parallel beam illuminates a sample in solution in a typical SAS experiment. At the end of the line, the detector measures the scattered beam removing the direct beam with a stopper. The diffraction pattern is circularly symmetric because the specimen is isotropic about the beam, and unlike crystallography, it is independent of the orientation of the sample. The function I(q) contains the primary data of the experiment, that is, the intensity of $q = 4\pi sin(\theta)/\lambda$. Here 2θ is the angle between the transmitted and the scattered beams, and λ is the wavelength; in figure 3.1, it is possible to see a schematic representation of SAS setup.

Most SAXS and SANS experiments require pure samples in a mono-disperse solution. For both probes, the requirements of the sample are the same.

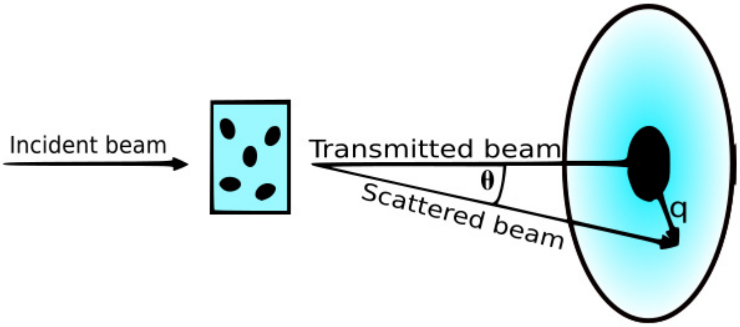


Figure 3.1: **Schematic representation of Small-Angle Scattering setup.** It represents a schematic representation of a Small-Angle Scattering experiment. The incident beam (that could be neutron or photons) goes through the sample in solution; most of it does not interact with the samples and pass through it; this is the transmitted beam; if instead interact with the sample, it gets scattered, the difference between the scattered beam and the transmitted beam is the vector q.

3.1.1 Monodispersity

Monodispersity is an essential prerequisite for structural modelling based on SAS. The scattering of the particles is related to the particles' dimension, particularly at small q; aggregates will render the data almost entirely uninterpreted and increase the scattering at a low q value. In practice, the sample should be at least 95% monodisperse. There are several ways to check the monodispersity:

- Multi Angle Light Scattering (MALLS)
- Dynamic Light Scattering (DLS)
- Size Exclusion Chromatography (SEC)
- Analytical Ultra-Centrifugation (AUC)

We will focus on DLS and SEC because they are used for this thesis.

Dynamic Light Scattering (DLS)

DLS is a powerful tool for studying the diffusion behaviour of macromolecules in solution [57]. When a monochromatic beam of light encounters a solution containing macromolecules, light scatters in all directions as a function of the size and shape of the macromolecules. For the Brownian motion, the intensity over the time scale of the scattered light has fluctuations. From those, it is possible to extrapolate the diffusion coefficient D_τ related to the macromolecule's hydrodynamic radius. The diffusion coefficient is highly dependent on the size of the macromolecule, the temperature, and the solvent's viscosity. In general, the more giant macromolecules are, the slower is going to be their diffusion.

Nowadays, DLS modern instruments already have software that automatically performs the data analysis. Nevertheless, this does not have to stop us from understanding how this analysis is performed to understand the results better. The function 2.23 acquired from the detector contains the information regarding the macromolecule in the exam, with a particular focus on the R_h. To do so, it is possible to fit the function with the monomodal distribution. This method does not require any *a priori* information. Still, it cannot give any information regarding the diffusion of the coefficients, so it is suitable for a sum of Gaussian distribution around the mean values. As we said in the theory chapter 2.4; $g_1(\tau)$ is the integral of exponential decay with a constant decay value Γ. For a polydisperse system, $g_1(\tau)$ is a sum of several exponential decays. We can define the cumulative generating function $K(-\tau, \Gamma)$ that is related to the $ln(g_1(\tau))$ and function is now define as:

$$K(-\tau, \Gamma) = ln\, g_1(\tau) = -\overline{\Gamma}\tau + \sum_{n=2}^{\infty}(-1)^n \frac{k_n}{n!}\tau^n \tag{3.1}$$

The k_1 is the average, k_2 is the variance, and k_3 is the skewness of the Gaussian distribution of the decay rates. We can truncate the sum at $n = 3$ to prevent overfitting. For more profound knowledge of this method, Koppel *et al.* [58] is suggested.

I used the DLS technique to assess the best buffer to reach sample monodispersity. We also analysed the hydrodynamic radius of the protein alone and the protein in complex with each DNA fragment.

Size Exclusion Chromatography (SEC)

SEC is widely employed for characterising proteins. It can be used to purify the protein as a technique for the qualitative and quantitative evaluation of the aggregates. The main advantage of the SEC is in the mild mobile phase conditions that permit the characterisation of proteins with a minimal impact on the conformational structure of proteins and complexes (if the binding is strong enough for keeping the complex) [59].

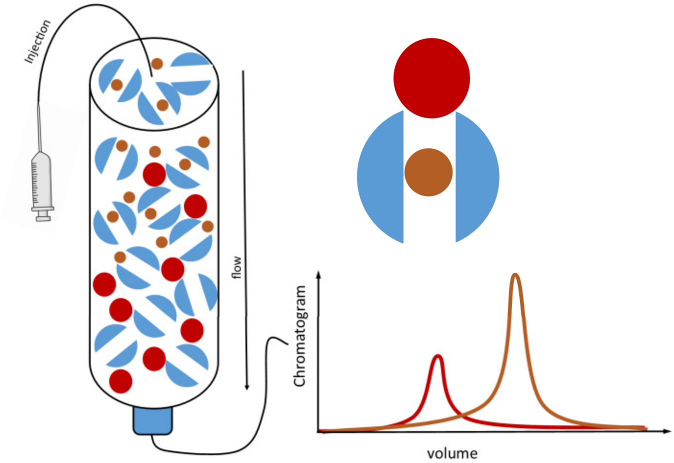


Figure 3.2: **Size exclusion chromatography** On the left is the schematic representation of the beads inside the column. On the right is the explanation of the stationary phase. The larger particle cannot enter the gel beads, and they are excluded faster than the smaller ones.

SEC can separate the biomolecules according to their hydrodynamic radius. Inside the columns, there is a stationary phase (gel beads) consisting of spherical porous particles thought that the biomolecules diffuse differently based on their size. Figure 3.2 explains a straightforward graphical representation of how SEC works.

3.1.2 Concentration

Concentration for SAS experiments are usually in the range of 1-10 mg/ml. The sample must have the characteristic that it should be diluted to avoid interparticle interactions. It is also essential to determine the concentration because it is linear with the intensity at 0 angles. Usually, the best way to measure the concentration of proteins is via the absorbance at 280 nm (A_{280}), but this works if the protein has tryptophans. Unfortunately, in our case, none of the proteins that we had analysed had any tryptophan. So we had to use another method to measure the concentration. We have used the Bradford Protein Assay to measure the concentration.

Bradford Protein Assay

The Bradford protein assay measures the total protein concentration in a sample in solution. The principle of this essay is that the binding of protein to Comassie-dye under acidic conditions results in a colour change from brown to blue. This method measures the presence of the essential amino acid residues: arginine, lysine and histidine, contributing to the formation of the protein-dye complex. It was chosen because reducing agents (DTT and β-ME) do not cause interference. This technique was invented by Bradford [60].

Because most of the protein we have used for this project does not have any tryptophane, it is necessary to find another way to measure the concentration. To do so, a calibration curve was created using Bovine Serum Albumin (BSA). The same solution measured the concentration at 280 nm of BSA at different concentrations between 0 and 2 mg/ml, and then it was mixed with the Bradford reagent and measured the absorbance at 595 nm. The graph 3.3 shows the absorbance at 595 nm vs the concentration calculated with the absorbance at 290 nm.

With a NanoDrop^{TM}, it is possible to measure the concentration of the sample without redoing the calibration curve, but there are a few steps to follow:

1. Take the protein and make different dilutions so that at least one of the dilutions is less than 2 mg/ml.

2. Take 5 μl of each dilution of protein and the buffer and add 100 μl of 1x Bradford Reagent (do not do simultaneously, because it is better to take the measurements for every sample at the same time from the mixing).

3. Take at least ten repetitions for each dilution (buffer included) at 595 nm with the option wavelength of the nanodrop (see below).

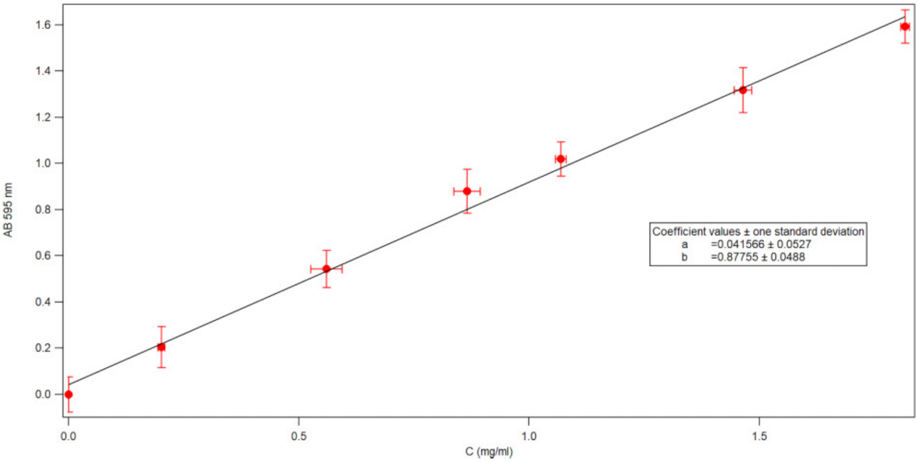


Figure 3.3: **Calibration curve of Bradford essay with BSA.** In the x asses, the concentration of the BSA was measured, and in the y asses, the absorbance of the same sample with the addition of Bradford reagent. The red dots are the experimental acquisition points, and the black line is the linear fit over the points.

4. Calculate average and deviation standard.
5. Subtract the buffer average to all the valid dilutions (when the concentration exceeds 2 mg/ml, the absorption at 595 nm could decrease)
6. Use the results of the linear fit for calculating the concentration and its error:

$$C = \frac{Ab_{595nm} - a}{b} \tag{3.2}$$

$$\Delta(C) = \sqrt{\frac{(\Delta(Ab_{595nm}))^2 + (\Delta(a))^2}{b^2} + \frac{(Ab_{595nm} - a)^2}{b^4}(\Delta(b))^2}$$

3.1.3 The buffers problem

A typical SAS experiment consists of acquiring the macromolecule in solution followed by the same acquisition of the buffer. The most

Table 3.1: **Bradford's linear fit results** Table with the result of the linear fit of the Bradford calibration curve.

coefficient	$y = a + bx$
a	0.042 ± 0.053
b	$(0.878 \pm 0.049)\frac{ml}{mg}$

straightforward way of preparation is to use the buffer obtained after dialysis or High-Performance Liquid Chromatography (HPLC). The solvent composition must be the same in the sample and the buffer. For SANS contrast variation, the composition of H/D is particularly critical, and the best way to be sure to have the exact composition of H/D in the buffer and the sample is with the dialysis. Fundamentally, not only all H/D atoms in the buffer have reached the exchange equilibrium, but also labile H/D of macro-molecule that have to be exchanged. Typically, two dialysis changes are enough, and the total dialysis time should be 24 hours to ensure that the H/D ratio has reached the plateau. For globular protein, the full H/D exchange can take many months, but in 24 hours, the totality of the isotopes in the buffer and 70-80% of the labile isotopes have exchanged. Another point to keep in mind is that the incoherent scattering of the Hydrogen dominates the SANS's background, so even if the contrast for a hydrogenated protein in 100% H_2O and 100% D_2O is not very different, the signal-to-noise ratio in 100% D_2O is an order of magnitude less than 100% H_2O because of the absence of H. Another critical point is that the pH should be the same in both the H_2O and D_2O solutions. When measuring the pH with a well-calibrated pHmeter in H_2O, it is a good idea to follow the formula of Glasoe and Long [61]:

$$pD = pH + 0.4X \tag{3.3}$$

Where pH is the reading on the pHmeter, and X is the mole fraction of D_2O in the solution.

3.1.4 Data reduction

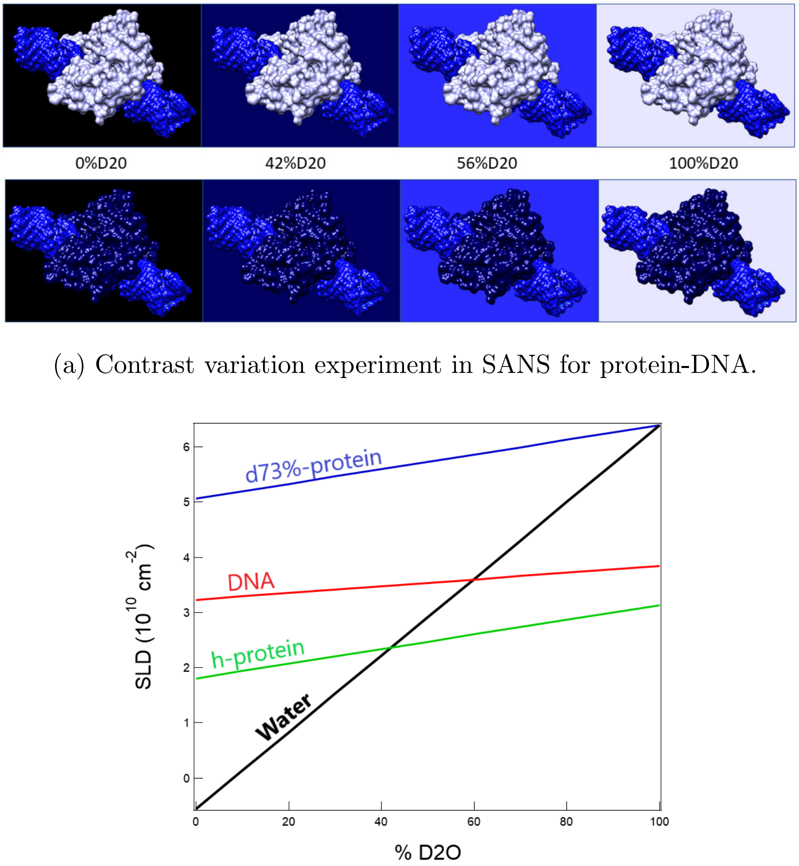
The molecules in the solution may take all the possible orientations; consequently, the sample scatters with radial symmetry. So it is possible to do a circular average about the direct beam, reducing the two-dimensional matrix of the detector in one dimension scattering profile. The basic protocol for the data reduction in SAS involves subtracting the buffer, empty cell and extraneous background (usually called blocked beam because it is an acquisition where the beam is blocked by a piece of material that absorbs the probe) from the sample measurement. The buffer is the most critical component in the subtraction, so treating the buffer as a routine measurement and then performing the subtraction later in the reduction process is advisable. The final intensities should be normalised to sample concentration and transmission.

3.1.5 Difference between SAXS and SANS

As we know so far, SAXS provides comprehensive envelope information on protein/complex structure; SANS delivers the same type of information but with less resolution due to the nature of the probe (x-ray radiation/neutron) to the flux (in an SR source is higher) and the detectors. Otherwise, it also offers the ability to use solvent contrast variation to distinguish and model different parts of a complex, either using the possibility of natural contrast (e.g. protein/nucleic acid) or by selective deuteration (e.g. protein/protein or increasing the $\Delta\rho$ of each component). Furthermore, using neutrons for data collection can remove the damage-radiation, which is often a problem in X-ray data acquisition because the samples do not absorb the neutrons [62]. For example, suppose the energy of the X-Ray beam is about 12 keV, which is typical energy for synchrotron radiation, only 10% of the radiation gets scattered. In that case, the rest is absorbed, leading to a photoelectric effect that leads to Auger emissions that causes excitations and a secondary electron cascade, or most of it gets transmitted without interaction with the sample. All of these phenomena cause damage to the sample. These effects can be reduced by cryo-freezing the samples. Still, the significant advantage of SAXS is that it is possible to study those samples in solution, so above 0 °C, it is impossible to perform the freezing as a solution for radiation damage. In particular, in a biological sample, what is dangerous regarding radiation damage is aggregation formation. The deterioration of the structure itself is not visible for the intrinsic resolution of the experiments. Still, the interaction with the X-ray with the water creates radicals that lead to the aggregations [63]. One way to reduce the radiation damage is to use a flow cell where the solution constantly flows. A disadvantage of this technique is the increased amount of material needed, but still, this amount is comparable with the one required for SANS measurements.

Contrast matching

The contrast matching technique (or contrast variation) relies on scattering between hydrogen and deuterium. Figure 3.4 shows a percentage of D_2O in the solvent for which $\Delta\rho = 0$, where ρ is the SLD, that depends on the type of biomolecule. The Contrast Match Point is the point where $\Delta\rho = \rho_{sample} - \rho_{buffer} = 0$, where it is impossible to detect the molecule with neutrons, so it is possible to separate and analyse the different compounds of the complex separately. Usually, the sample should be measured in pure H_2O buffer, pure D_2O buffer, and other H_2O/D_2O mixtures. The more contrast conditions in the analysis, better will better the result; usually, it is 4-5 contrast conditions. Plotting the $\sqrt{I_0}$ vs $\%D_2O$, we can obtain the intersection of the linear fit with zero, and this particular percentage is the Match Point of this specific protein. It is possible to perform neutron transmission measurements To find the exact mixture of H_2O/D_2O in the solvent. The transmission is defined as the flux ratio through the sample and the incident flux. The transmission can be measured from the intensity of the direct beam taught to the sell filled and over the intensity of the empty cell. Most nuclei do not absorb neutrons at SANS experiments' typical energy (thermic neutron); the beam is primarily attenuated for the incoherent scattering from hydrogen. We must replace the natural abundance of hydrogen with deuterium to achieve the desired labelling for the macromolecules. The replacement is done by growing microorganisms (in particular bacteria) in culture solutions containing D_2O and deuterated substrates. As we know so far, SAXS provides comprehensive envelope information on protein/complex structure; SANS delivers the same type of information but with less resolution due to the nature of the probe (x-ray radiation/neutron) to the flux (in an SR source is higher) and the detectors. Otherwise, it also offers the ability to use solvent contrast variation to distinguish and model different parts of a complex, either



(a) Contrast variation experiment in SANS for protein-DNA.

(b) Scattering length densities

Figure 3.4: **Contrast variation experiment in SANS.** Schematic of a contrast variation experiment in SANS for a protein–DNA complex using different ratios of H_2O/D_2O. We can see that the biomolecules scatter a specific %D_2O the same as the buffer. The graph represents how the scattering length densities over the percentage of D_2O vary for various biomolecules [45].

using the possibility of natural contrast (e.g. protein/nucleic acid) or by selective deuteration (e.g. protein/protein or enchanting of the $\Delta\rho$ of each component). Furthermore, using neutrons for data collection can remove the damage-radiation, which is often a problem in X-ray data acquisition because the samples do not absorb the neutrons [62]. For example, suppose the energy of the X-Ray beam is about 12 keV, which is typical energy for synchrotron radiation, only 10% of the radiation gets scattered. Still, a tiny part of the non-scattered light is absorbed, leading to a photoelectric effect that leads to Auger emissions. This absorption causes damage to the sample. These effects can be reduced by cryo-freezing the samples. In particular, in a biological sample, what is dangerous regarding radiation damage is aggregation formation. The deterioration of the structure itself is not visible for the intrinsic resolution of the experiments. Still, the interaction with the X-ray with the water creates radicals that lead to the aggregations [63]. One way to reduce the radiation damage is to use a flow cell where the solution constantly flows. A disadvantage of this technique is the increased amount of material needed, but still, this amount is comparable with the one required for SANS measurements.

3.1.6 Experimental protocol

We want to obtain the scattering profile from the isolated macromolecule in solution with the subtraction of the buffer from a SAS experiment. The basic SAS protocol experiment is:

1. The transmission acquisition of samples, buffers, empty cells and empty diaphragm. These measurements are done by recording an image of the attenuated direct beam without the beam stop.

2. The scattering of each sample, corresponding buffer, an Empty cell, the blocked beam (usually a piece of Cadmium or Boron[1])

Unfortunately, life is more complicated than theory, and sometimes the background does not match the samples. There are several ways to estimate the background if a direct measurement is not available:

[1] Both those materials are good neutron absorbers

1. Assume that all the scattering at high q is due to the buffer and subtract the flat background at the same level of this high q region.
2. Remove a flat background when the scattering falls in the Porod asymptotic (q^{-4}).
3. adjusts the buffer scattering acquired with the sum of the other buffers.

3.2 Structural information

From the scattering profile of a non-interacting particle, it is possible to obtain some basic information about the sample. Some of those pieces of information are model-independent to give us information regarding the quality of the data.

3.2.1 Intensity at 0 angles and radius of gyration

It is possible to extract only a few parameters from I(Q) curves, according to *Guinier et al.*[64] that, for small values of Q, the curve is independent of any details of the structure and the equation 3.4

$$I(Q) = I(0)e^{-q^2 R_g^2/3} \tag{3.4}$$

R_g is the radius of gyration, and I(0) is the intensity at 0 scattering angle. The equation 3.4 is valid in a range of $0 < q < 1/R_g$, also defined as the Guinier region. Whenever the Guinier plot is not linear at low q, that could mean that it could be aggregation if the trend is super-linear and interparticle repulsion if it is sub-linear. Macromolecules composed of different particles with different SLD can display strange behaviour close to the MP of the single particle in the Guinier region. From the equation 3.4 it is possible to extrapolate also $I(0)$ that is proportional:

$$I(0) \propto \frac{N}{V} V_P \Delta\rho^2 \tag{3.5}$$

So there is a relationship between $I(0)$ and the contrast. $I(0)$ is a linear function of $\Delta\rho^2$ and a plot of $I(0)^{\frac{1}{2}}$ versus % D_2O in the buffer should be linear and should going at zero at the MatchPoint of the biomolecule.

3.2.2 Pair-distance distribution function

The determination of the Pair Distance Distribution Function (PDDF) P(r) can give more structural information[65]. The P(r) reveals information about molecules' shape, allowing a more intuitive interpretation of I(q). Furthermore, the P(r) is also critical for the 3D space model reconstruction, software such as DAMMIF [51]. The I(q) and P(r) are closely related:

$$I(q) = 4\pi \int_{0}^{d_{max}} P(r) \frac{sin(qr)}{qr} dr \tag{3.6}$$

where D_{max} is the maximal chord length. The PDDF can be obtained from the equation 3.6 via an Inverse Fourier Transform (IFT) procedure. The PDDF represents the distribution of distances between two points in the macromolecule weighted by the product of the excess SLD at those two points. This function can exist between 0 and D_{max} because it is the maximum distance between any two points inside the macromolecule. As we have said already, we can obtain several structural information from P(r) because this is a function in real space, and it is easier to recognise the features of the macromolecule. It is possible to extract only a few parameters from I(Q) curves, according to *Guinier et al.*[64] that, for small values of Q, the curve is independent of any details of the structure and the equation 3.4. Rg is the radius of gyration, and I(0) is the intensity at 0 scattering angle. The equation 3.4 is valid in a range of $0 < q < 1/R_g$, also defined as the Guinier region. Whenever the Guinier plot is not linear at low q, that could mean that it could be aggregation if the trend is superlinear and interparticle repulsion if it is sub-linear. Macromolecules composed of different particles with different SLD can display strange behaviour close to the MP of the single particle in the Guinier region. From the equation 3.4 it is possible to extrapolate also $I(0)$ that is proportional:

$$I(0) \propto \frac{N}{V} V_P \Delta\rho^2 \tag{3.7}$$

So there is a relationship between $I(0)$ and the contrast. $I(0)$ is a linear function of $\Delta\rho^2$ and a plot of $I(0)^{\frac{1}{2}}$ versus % D_2O in the buffer should be linear and should going at zero at the MatchPoint of the

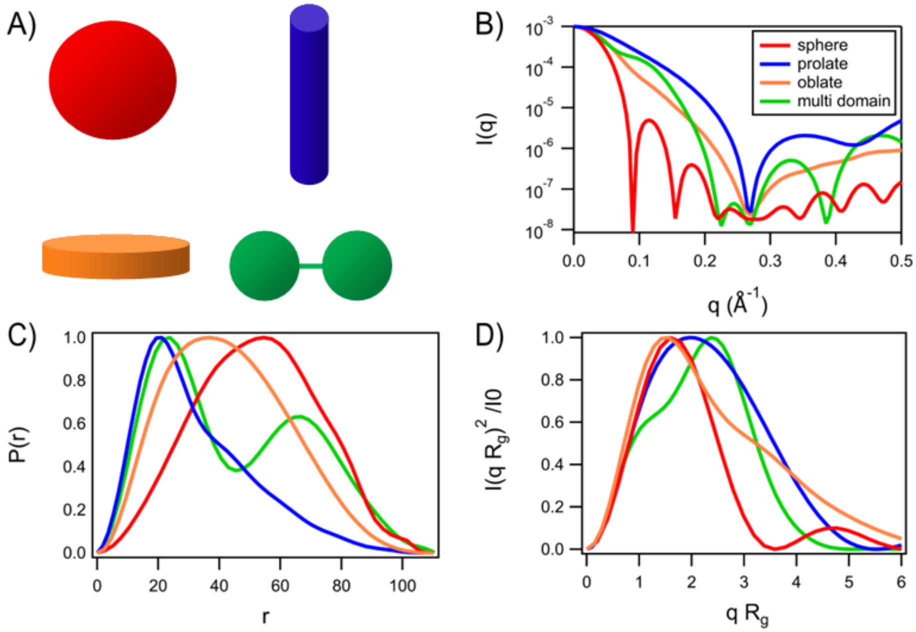


Figure 3.5: **Scattering profiles, distance distribution functions, and normalised Kratky plots of some geometrical bodies with the same maximum size.** A) Four different geometrical bodies with the same maxim size ($D_{max} = 110$Å), the colours are the same in the other panels. B) The Scattering profiles of the corresponding shapes. C) The distance distribution functions. D) The normalised Kratky plots [66].

biomolecule. Figure 3.5 shows the scattering curve of different bodies that have the same D_{max} and the relative P(r). By definition P(r) start with $P(0) = 0$ and then should terminate smoothly at zero at $P(D_{max}) = 0$. If $P(0) \neq 0$ is a sign of an error in the background subtraction, this can be used to estimate the correct background. A shoulder o a long tail in the high-r region could mean aggregation. Meanwhile, negative values can be due to the difference in contrast between components (only observed in SANS) or repulsive interaction between particles.

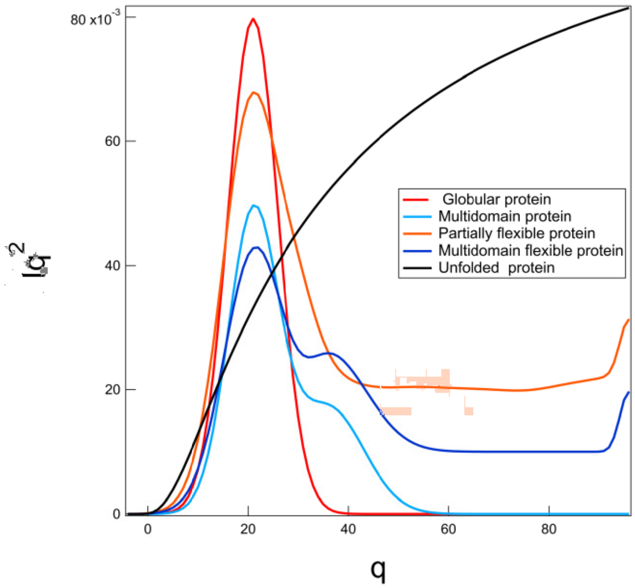


Figure 3.6: **Schematic representation of typical Kratky plots.** The curvature of each macromolecule depends on various characteristics of the sample, such as the molecular shape, degree of flexibility, number of domains, and globularity.

3.2.3 Kratky Plot

The Kratky plot is another way to present the data; this plot is $I(q)q^2$ vs q plot. This way to show the data has a practical qualitative difference for molecules with different chain densities and can assess the samples' flexibility and unfolding. Unfolded protein should have a plateau at high q in the Kratky plot. Meanwhile, a globular protein should have a Gaussian-shaped peak where the centre depends on the protein's radius. A partially unfolded protein may have a plateau and the Gaussian peak, and a multi-domain complex will appear with multiple peaks. To have a semi-quantitative analysis over the Kratky plot, it is possible to normalise over the intensity at q=0 (I(0)) and the dimension ($\mathrm{R_g}$). Dimensionless Kratky plot is present as:

$$\frac{(qR_g)^2 I(q)}{I(0)} \, vs. \, qR_g \tag{3.8}$$

The Kratky plot is routinely used in SAS data analysis and provides the first estimate of the folded state of the macromolecule.

3.2.4 Stuhrumann plot

As we have already discussed 3.1.5, the contrast between solvent and the dissolved macromolecules can vary in a gargantuan way, and also the R_g that is dependent on the LSD (ρ) can vary.

$$R_g^2 = \frac{\int_{V_p} (\rho(r) - \rho_s) r^2 \, dr}{\int_{V_p} (\rho(r) - \rho_s) \, dr} \tag{3.9}$$

as a consequence the R_g^2 varies with the contrast:

$$R_g^2 = R_v^2 + \frac{\alpha}{\rho} + \frac{\beta}{\rho^2} \tag{3.10}$$

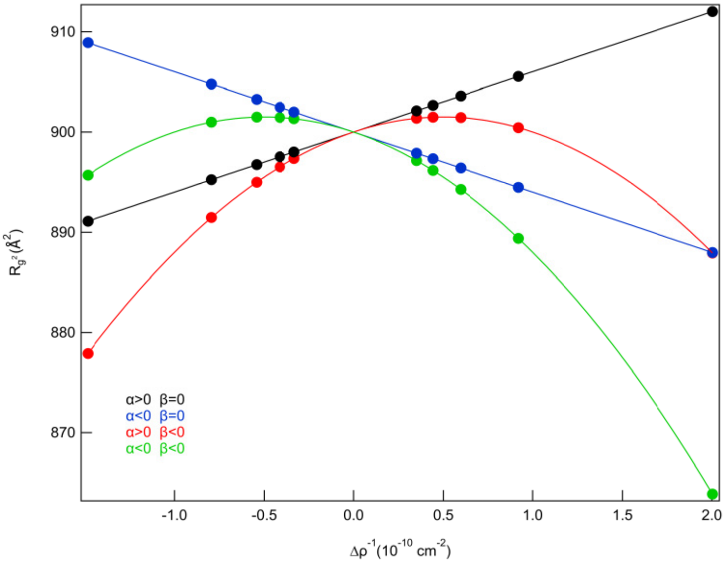


Figure 3.7: **Stuhrmann plot** Example of Stuhrmann plot's examples for four different sets of data of a complex formed by two different components and their parabolic fits

The equation above 3.10 was expressed by Ibel and Stuhrmann in the 1975 [67]. This equation is a quadratic function (parabola) with negative concavity (β must be negative), as it is possible to evince in figure 3.7. This type of plot can give pieces of information without *a priori* knowledge of the complex. For a complex with two different components, the value β indicates the distance between the two centres of mass:

$$d = \left(\frac{\beta}{\overline{\rho_1}^2}\right)^{\frac{1}{2}} + \left(\frac{\beta}{\overline{\rho_2}^2}\right)^{\frac{1}{2}} \tag{3.11}$$

3.3 *Ab initio* Methods: Dummy atom modelling

In this thesis, we used the software DAMMIF [51] and MONSA [68] to generate the low-resolution envelopes of the complexes using the SANS and SAXS data sets. Those programs can produce envelopes of the macromolecules in solution from the SANS and SAXS experimental data so that those *ab initio* approaches can provide valuable information about the shape and relative positioning without *a priori* structural knowledge, *e.g.* crystal structures and NMR models. The main difference between the two programs (DAMMIF and MONSA) describes the heterogeneity between fully hydrogenated protein, partially deuterated protein and DNA. The DAMMIF inputs were generated using the program GNOM [69] with the pair distance distribution constrained to 0 at $q = 0$ and $q = D_{Max}$. As a quality control, we compared the radii of gyration determined from the pair distribution analyses with those defined by the Guinier analysis.

3.3.1 MONSA

MONSA is an extended version of DAMMIN and works best for multiphase bead modelling. Furthermore, it allows the fitting of multiple scattering curves simultaneously. It can benefit from both neutron and X-Ray scattering. The MONSA input consists of the SLD and volume fraction estimated from the chemical composition using the software MULCH [70]. The initial volume could be a sphere, a prolate or an oblate with different radii, containing a fixed number of beads with a fixed radius. Those pieces of information are stored in different files:

1. **Master file (*.mst).** It contains the general phase information like volumes, radii of gyrations and connectivity.

2. **Control file (*.con).** It contains information regarding the different SLD for each phase associated with the data files and the relative weight of the data set.

3. **Data files (*.dat).** They contain the raw experimental data at different contrasts.

4. **DAM file (*.pdb).** It is a file that defines the search volume and the number of phases. To generate this file, the software DAMESV is necessary.

After those files are prepared, it is possible to launch MONSA and open the interactive configuration where the files need to be called, and it is possible to define more factors and penalties. The results are a PDB file with all the phases and a series of associated fits for the data set. For more information, Svergun *et al.* 2000 [50].

3.4 Recreation of the scattering curve from model

Often in this thesis, I have to compare the experimental data with a model created. So, a tool that can build the corresponding scattering curve from an atomic model for both SAXS and SANS data sets was necessary. Different software comes in hand for this task:

1. CRYSOL

2. CRYSON

3. PEPSI

3.4.1 CRYSON and CRYSOL

CRYSOL is a program that evaluates the scattering of macromolecules with an atomic structure. This program uses multipole expansion for the calculation of the scattering pattern, and it also takes into account the hydration shell. With a PDB file, CRYSOL can predict the solution scattering or perform the experimental curve's fit using only two free parameters, the average displaced solvent volume per atomic group and the contrast of the hydration layer. The particle with scattering density $p_\alpha(r) asolventsurrounds$ with an average scattering densityρ_0. The hydration shell is approximated by a border layer of adequate thickness A and density ρ_b that may differ from ρ_0. For further knowledge regarding this theory, I invite you to read paragraph 2.3.1. The software input is the PDB(s) and the experimental curve. It is also possible to tune the other information, such as the number of harmonics (the default value is 15, the more significant the value more accurate the curve will be, and the computational time will increase). The output will be the theoretical curve and, if experimental data were given, the reduced χ^2. CRYSON is a re-implementation of CRYSOL, adapted to work with SANS data. There are some differences between CRYSON and CRYSOL:

- CRYSON asks the concentration of D_2O in the solvent y. The solvent density is evaluated as follows:

  $$\rho_{solvent} = \rho_{H_2O} \cdot (1 - y) + \rho_{D_2O} \cdot y \tag{3.12}$$

  where $\rho_{H_2O} = 0.562 \cdot 10^{10} cm^{-2}$ and $\rho_{D_2O} = 6.404 \cdot 10^{10} cm^{-2}$

- CRYSON prompts for the fraction of non-exchanged NH peptide: $DNEXCH$. It is assumed that all exchangeable H (which belong to NH, NH_2, NH_3, OH and SH groups) are replaced by D in proportion y, except for the main-chain NH groups, for which the probability is $y \cdot (1 - DNEXCH)$. Default value $DNEXCH = 0.1$ is generally accepted. The SLD ρ of any atomic group at the D_2O concentration y is evaluated as:

  $$\rho(y) = \rho(0) + EXCH \cdot y \cdot (\rho_D - \rho_H)(1 - DNEXCH) \tag{3.13}$$

  where $\rho(0)$ is the SLD of the atomic group at 0% D_2O, ρ_D is the SLD of deuterium $(0.6671 \cdot 10^{-12} cm)$, ρ_H is the SLD of hydrogen $(-0.3742 \cdot 10^{-12} cm)$.

Table 3.2: List of SLD ($\rho(0)$ of each atomic group and neutron exchange)

atomic group	$\rho(0)$ $(10^{-12}cm)$	EXCH
C	0.6651	0
CH	0.2909	0
CH_2	0.083	0
CH_3	0.458	0
N	0.94	0
NH	0.5658	1
NH_2	0.1916	2
NH_3	0.1826	3
O	0.5804	0
OH	0.2062	1
S	0.2847	0
SH	0.9	1
P	0.517	0
Fe	0.95	0
Cu	0.76	0
Ca	0.47	0
Mg	0.52	0
Mn	0.39	0
Zn	0.57	0
N histidine	0.7529	0.5

- CRYSON also fit the background as a flat curve with the last points of the experimental data.
- CRYSON searches for the contrast in the hydration shell in the range $(0 < DroShell < 0.2 \cdot |\rho_{solvent}|) \cdot 10^{10}\, cm^{-2}$.

3.4.2 PEPSI

PEPSI is a software that calculates small-angle scattering profiles from atomistic models. Such as CRYSOL and CRYSON; the method is based on the multipole expansion 2.15. Still, it is significantly faster than its opponents because it uses the Nyquist-Shannon-Kotelnikov sampling theorem. The initial implementation of CRYSOL had a linear dependence of both the number of atoms in the molecules (N)

and the number of points of the experimental curves (M) as $O(NM)$, with more recent versions of the computational cost it is lowered to $O(N+M)$. This software is fast compared to other like SASSIM or FoXS. Still, it has the disadvantage of the representation being too simplistic of the hydration shell using a two dimension angular function. PEPSI uses the three-dimension Zernike polynomials instead the multipole expansion methods, so the computational complexity of this method is $O(N+M)$ as CRYSOL, but the hidden time-limiting step is the computation of the three-dimensional Zernike moments. To calculate them, atomic models are mapped onto a three-dimensional grid. To increase the speed, PEPSI uses a fast model for the simulation of the hydration shell that is a uniform grid of points, and then it applies an adaptive order of multipole expansion. With the Nyquist-Shannon-Kotelnikov theorem, it is possible to determine the required expansion order using the radius of gyration of the hydration's shell and the maximum scattering vector. Furthermore, it represents the scattering curve using cubic spline interpolation, and finally, it uses partial scattering intensities to fit the theoretical curve to the experimental profile rapidly.

For all the projects, we have generated thousands of all atoms models and compared all to the experimental scattering curves. We preferred to use PEPSI-SANS and SAXS because they were faster than the other software with similar results as it demonstrated in [71], and for its Command Line Interface (CLI) that allows seamless integration with batch job scheduling systems.

3.5 Homology Modelling

The structure of the protein can be obtained with three different methods:

1. within empirical information such as crystallography, NMR or cryEM, with the differences and limitations of those techniques. There is a more extensive explanation of differences and limitations in chapter 1.5

2. within a purely theoretical method, predicting the 3D structure of the protein purely from the aminoacidic sequence. The so-called Anfinsen's dogma [72] demonstrates that the most stable

conformation is the one with the lowest free energy, and it is the native structure of the protein. It is easy to imagine that researching the absolute minimum in the conformational space for large molecules and complexes is difficult and time-consuming.

3. within the combination of the homology model with SAS data can distinguish between the different molecules.

In our case, homology was the best and fastest way to obtain the complex structure; because there is already a partial crystal structure of the complex, in particular for MexR, the crystal structure of the protein was already known, but the complexes with the DNAs was unknown, and this was our goal. Meanwhile, for MYC-MAX, the core of the complex bound with the DNA E-box was already crystallized, and we could focus on the flanking regions. So in this thesis, we have made extensive use of Homology modelling. Homology modelling is currently the most accurate method to generate reliable three-dimensional protein structure models. Homology (or comparative) modelling methods construct a model of the "target" from experiments and homologous protein collected in the PDB libraries that are called "templates". Homology modelling relies on identifying the protein's structure that resembles the structure/sequence of the "target". Naturally, homologous proteins have similar stable tertiary structures, and this fact is used in homology modelling; in fact, it has proved that the 3D structure is evolutionary more conservated than the sequences [73], so even proteins that have differences in the sequence appreciably can have detectable similarity in the structures. The sequence liniments and template structures are used to produce the model of the "target" protein. This step is fundamental; in fact, the correctness of the alignment and choosing the best template structure is one of the most critical factors compared with the choice of the best software for the correct function[74]. The regions of the model that are reconstructed without a template are less accurate than the rest. Nevertheless, homology can give qualitative information regarding the biochemistry of the protein that can be useful for drug design and the comprehension of complex formation. The homology model can be schematized in several steps:

1. **Template selection and target-template alignment** This is a crucial step. The choice of the best template is the most

important because all the modelling is based on it. One of the simplest methods for the identification relies on the sequence alignment with some databases of already known protein structures such as FASTA [75].

2. **Model construction** After the template choice and its alignment, it is possible to create a 3D model of the target, and there are some ways to reconstruct the model:

    - **Fragment assembly** It relies on the conserved structural fragments; it is known that several proteins in different organisms, even if they have other primary structures, they have the same function and the same structure, and in general, those differences are located in the loops. So a good idea to resolve the structure of the unknown protein is like the first step to construct the conserved regions and then substitute the missing pieces from other proteins.
    - **Segment matching** In this case, it is possible to perform the alignment over small fragments, and each fragment have to be fitted with another fragment from the protein data bank
    - **Spatial restrain** This is the most common method, where the template is used to construct a set of geometrical criteria converted in probability density function for each restraint. This method is extensively used in loop modelling. The most used software (also the one that was extensively used in this thesis) is MODELLER [76]

3. **Loop modelling**: The regions on the target sequence that are not aligned with a template must be modelled with loop modelling. Those coordinates are less accurate than the ones "copied" from another structure. Most inaccuracies are due to two amino acids: lysine and arginine, because the dihedral angles are notoriously difficult to predict.

4. **Model assessment** It is possible to assess the model(s) obtained without any references from the actual target structure using statistical potential and minimization of energy calculations. The former is an empirical method where it is possible

to assign a probability to each pairwise interaction between two amino acids based on the frequency from the assembly of all the known proteins; the latter aims to replicate *in silico* all the interatomic interactions responsible for protein stability and formation. In particular, it is focused on Van Der Walls and electrostatic interaction. That calculation provides an energy landscape, and there is the strong assumption that the minimum is also the native state of the protein.

5. **Structural comparison methods** It is possible to gather information regarding the actual structure of the protein or complex from an experiment, like in our case, we have compared all the structures generated with the experimental SAS data, and we have made those comparisons with three software: CRYSOL, CRYSON and PEPSI; then we have select the most correct model.

3.5.1 Docking: ZDOCK

Protein interactions play essential roles in maintaining life because those interactions regulate different aspects of the cell. Moreover, molecular docking is used to predict the structure of complexes, and protein flexibility is one of the biggest challenges for binding mode predictions [77]. Prediction of a complex structure is difficult because of molecular flexibility and conformational changes. ZDock is one of the software that we have used for protein docking.

The rigid-body protein-protein docking program ZDOCK uses the Fast Fourier Transform (FFT) algorithm to enable an efficient docking search on a 3D grid where both the proteins are present and utilize a combination of shape complementarity, electrostatics and potential statistical terms for scoring on protein is fixed meanwhile the other is moving around. ZDOCK achieves high predictive accuracy on protein-protein docking benchmarks, with >70% success in the top 1000 predictions for rigid-body cases i [78], and consistent success in the international protein-protein docking experiment, Critical Assessment of PRotein Interactions (CAPRI) [79].

3.5.2 Docking: HDOCK

For this project was necessary to find a way to have a binding between the DNA and the protein, we used the software HDOCK (it was already used for a similar protein []). HDOCK is a docking software and (like the others) is based on sample and scoring, with two structures: protein, DNA, or RNA. In general, there is no information regarding the binging sites *ab initio* global docking usually is needed to sample assumed binding modes in six degrees of freedom: three rotations plus three translations. The programmers of HDOCK have developed a hybrid strategy to incorporate the binding interface information into traditional global docking. With this strategy, they have ranked top performers in the CAPRI sessions; furthermore, they have also developed web servers that accept both structure and sequences, and it extrapolates the pieces of information of the binding sites directly from the PDB. The docking process is fully automated, and the results can be presented interactively to users via a web page and email. For further information regarding how HDOCK works, I suggest reading their paper [80].

3.5.3 NOLB

NOLB is a computational method for non-linear Normal Mode Analysis (NMA). It relies on Rotations and Translations of Blocks (RTB), and it is possible to deepen the topic in the following paper [81]. It is demonstrated [82] how it is possible to interpret the eigenvalue of the RTB in terms of angular and linear velocities applied to the rigid block where the macromolecule is likely to be divided. Nowadays, Molecular Dynamics (MD) can accurately predict the movements of complexes, but it is computationally expensive and time-consuming. Meanwhile, NMA is cheaper and still allows the extraction of the essential collective motions of the biomolecule in the exam. NMA uses a quadratic approximation for the potential energy and produces a linear deformation of the initial structure, which is accurate only for small-amplitude motions. A larger amplitude could destroy secondary structures and bonds. In this work, they have used the harmonic oscillator method to change a registered PDB structure and then confront it with the experimental data. The basic idea of the NMA method is to represent the potential energy V in the vicinity of q_0 by its quadratic approximation and to solve Newton's equation of motion analytically:

$$M(\ddot{q}+\ddot{q}_0)+\nabla V(q_0+q)\approx M\ddot{q}+Kq=0 \tag{3.14}$$

M is the diagonal mass matrix, and K is the Hessian matrix of the potential energy V evaluated at the equilibrium position q_0. We should note that in classical mechanics, K is traditionally called the stiffness matrix.

3.5.4 Rosetta

Protein-protein interactions are fundamental in every aspect: in the intercellular or intracellular communication for morphology, mobility, cell growth, proliferation gene expression and infection and disease; for example, viruses need to contact protein on the cell surface to attach and also to enter a cell and also something that is not a pathogen like a neurodegenerative disease. Models obtained from *de novo* prediction have been demonstrated to help get insight from a biological point of view, either through the recognition of binding sites or functional annotation by folded identification [83]. The Rosetta algorithm attempts to mimic the interplay between local and global interactions that leads to protein structure formation. The method is based on the empirical observation that the local sequences bias but do not define the local structure of the protein in the exam. The final structure is obtained when the fluctuations of the local structure come together to get a compact structure, with favourable non-local interaction such as hydrophobic residues buried inside the protein, β-strands pair between each other, and the formation of disulfuric bonds. Rosetta is one of the most used software for forming protein complexes. The Rosetta algorithm uses the structures sampled by local sequences, and they are approximated by the distribution of structures already seen for those short sequences in already known proteins. Consequently, the library's presence, where all those fragments of local structures are derived from well-known proteins, is fundamental for the algorithm. Then the final compact structure is assembled by the random combination of those fragments using Montecarlo. The fitness of each conformation generated is evaluated based on a scoring function [2] derived

[2] Scoring functions are mathematical functions used to approximately predict the binding affinity between two molecules after they have been docked—

from the statistic of known protein structures. Rosetta programmers developed a protocol or multiple protocols for understanding and mimicking the protein-protein interaction called protein-protein docking. Rosetta can form the complex between two proteins and identify the binding sites, and it can do starting from an unbound structure. There are two main kinds of docking: global docking and local docking. The difference between the two is based on the knowledge about the system, so if there is no information on where the binding site the global docking is preferable; what happens during global docking is the beginning of it is possible to generate a random position in between the two proteins after the test of all the different combination of position between the two partners and discover the one with the minimum of free energy. It is less accurate than local docking and requires much more computational time, and nowadays, it works best only for small complexes of less than 450 amino acids. However, if there is an idea of where the binding site is, there is no need to have a crystal structure; it is enough to define the local docking site, so it is possible to start prepositioning the structures. Local docking is much more accurate than global docking and requires less computational time. Furthermore, there is the possibility of integrating experimental data to have other constraints to help the simulation.

3.6 Multi-state modelling

3.6.1 MultiFoXs

MultiFoXs is a software that could create an ensemble of different models relying on a SAXS profile. The program takes as an input the SAXS profile, an atomic structure, a list of flexible residues and the number of conformations. After the job submission, the server starts exploring the space of ϕ and ψ main chain dihedral angles of the user-submitted flexible residues, using Rapidly exploring Random Trees (RRT) algorithm [84]. This exploration will generate the required number of conformations, and then the SAXS profile is calculated for each generated confirmation. After this step, the multistate models [3]

[3] a multistate model is a model developed by the weighted sum of different models

are ordered from the best to the worst based on the χ:

$$\chi = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{I_{exp}(q_i) - c\sum_n I_n(q_i, c_1, c_2)}{\sigma(q_i)}\right)^2} \tag{3.15}$$

where $I_n(q_i, c_1, c_2)$ and w_n are the computed profiles and their corresponding weight. For further detail, herein is possible to find the associate paper of FoXs and MultiFoxs [85]

3.6.2 iBME

In this thesis, we have encountered disordered proteins, and it is impossible to describe them with just one mode, but we need an ensemble. To generate conformational ensembles representing the ID protein from the SAXS experiments, it is possible to use the iBME (iterative Bayesian Maximum entropy) algorithm [86]. In the BME [87] method the the functional L have to be minimised:

$$L(\omega_1, \cdots, \omega_m) = \frac{m}{2}\chi_r^2(\omega_1, \cdots, \omega_m) - \theta S_{rel}(\omega_1, \cdots, \omega_m) \tag{3.16}$$

Where m is the number of points of the experimental curve, χ is the measure of agreement between the model and the data, ω are the weights, S$_{rel}$ measures the divergence between the optimized weight and the initial weights, and θ is a tumble parameter balancing between the minimization steps of S$_{rel}$ and χ_r^2. The value $\phi = e^{S_{rel}}$ indicates the fraction of models that contribute effectively to the averages calculated with the optimized weights.

Meanwhile, iBME uses χ_r^2 as a cut-off to find the scale factor and the constant background until its convergence [86].

CHAPTER 4

Results

- In Paper I, we have used small angle X-Ray and neutron scattering to investigate the structural change of the MexR regulator in the presence of PII DNA. With the help of Molecular dynamic simulations, we have found that the MexR present a distinct asymmetry when bound to the DNA in a homo-dimer protein.

- Paper II is the extension of the first paper, and it is based on the characterisation of the complex formed by two MexR and the DNA comprising PI and PII boxes. The analysis suggests that the complex, particularly the DNA, has intrinsic flexibility that leads to a bending of the DNA between the two binding boxes.

- Paper III describes an upgrade dedicated to the neutron biology community done at D22: SEC-SANS. This add-on is increasing the flexibility of D22 and will help biologists collect even more data at D22.

- Paper IV is based on the characterisation with small angle X-Ray and neutron of two different complexes: MAX:MAX:DNA and MYC:MAX:DNA those are intrinsically disordered. In this paper, we have to exploit two ways to find a conformational ensemble that simultaneously fits SAXS and SANS data.

CHAPTER 5

Future perspectives

5.1 MexR and DNA

The Multidrug Resistance (MDR) is a global problem and will become predominant worldwide as we know *P. Aug* is one of several bacteria that present the same feature as a similar mechanism to resist antibiotics. And the understanding of the binding of the repressor, in our case MexR, to the cognate sites was just the first step. In my first paper, we focused principally on the protein and probably, and a focused experiment on the DNA will be necessary. I do not suggest any small angle scattering for this purpose, but NMR. Meanwhile, extending the complex with the addition of NalD, another MexAB-OprM repressor, is possible. The technique I can suggest for this task is SAS supported with CryoEM because now the complex is big enough to be studied with this technique. The structural difference between the two complexes bound with the DNA could give a great insight regarding the role of the repressor and how can the interaction with the antibiotics could lead the dissociation within the DNA, and maybe in the future, could also lead to the production of antibiotics that they will not initiate the dissociation process.

5.2 Upgrades in SANS

Small Angle Neutron Scattering is a powerful technique because of its versatility. Still, it has a significant downgrade and is a low-resolution method. Without any knowledge *a priori* structure, like crystal or NMR, it can not goes too far with the analysis. But right now, the SANS instruments need to increase the speed of acquisition, and it is possible o do this in just two ways: increase the flux, which I would say that it is not feasible, or acquire at the same time at a small and at a large angle, something that they have applied at D22 with the more accessible of a second detector closer to the sample. Another way to have better user-friendly acquisition that it can be helped by more accessible software to do the data analysis and not as they are doing right now with the Mantid project to unify all the neutron experiment analysis and acquisition in just one software, but to have a specific program for the particular experiment, even if we are talking about SANS it is different the analysis between,n skyrmion, ferrous crystals or biomolecules in solutions; I cannot speak for the other samples, but for biomolecules, in solutions, a few automation will grateful, in particular something that automatically stops the acquisition when enough statistic is reached or in spallations source such at ISIS that the counter doesn't continue if the spallation is down. From my side, I have written a small script for the autonomous calculation of the Radius of Gyration and for making the Kratky plot directly in IGOR and another script in Igor for the measure of the Match Point.

5.3 Intrinsically disordered protein

Small angle scattering is fundamental when we are talking about Intrinsically Disordered Protein (IDP), they are difficult to crystallise and to analyse with CryoEM, but SAXS is almost routine for those types of proteins, SANS it is not; for various reasons: the price, the difficulty to have beam-time at neutron facilities and the problems on the analysis. Treating several data simultaneously is a difficult task, and in paper IV, we resolve with some conformational ensembles with two different methods that could fit simultaneously. In the future, it is exciting to use the same techniques with other complexes (to legitimate the use of SANS) and to understand which cases are preferable to use one method or the other.

AUC Analytical Ultra-Centrifugation

CAPRI Critical Assessment of PRotein Interactions

CLI Command Line Interface

CryoEM Cryogenic electron microscopy

DLS Dynamic Light Scattering

HPLC High-Performance Liquid Chromatography

HTH Helix-Turn-Helix

IDP Intrinsically Disordered Protein

IFT Inverse Fourier Transform

ITC Isothermal titration Calorimetry

LSD Length Scattering density

MALLS Multi Angle Light Scattering

MD Molecular Dynamics

MDR Multidrug Resistance

MP Match Point

NMA Normal Mode Analysis

NMR Nuclear Magnetic Resonance

P. Aer. Pseudomonas Aeruginosa

PDDF Pair Distance Distribution Function

PEPSI Polynomial Expansions of Protein Structures and Interactions

RTB Rotations and Translations of Blocks

RRT Rapidly exploring Random Trees

SANS Small Angle Neutron Scattering

SAS Small Angle Scattering

SAXS Small Angle X-Ray Scattering

SEC Size Exclusion Chromatography

SLD Scattering Length Density

BSA Bovine Serum Albumin

FFT Fast Fourier Transform

HTH Helix-Turn-Helix

Bibliography

[1] James D Watson and Francis HC Crick. "The structure of DNA". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 18. Cold Spring Harbor Laboratory Press. 1953, pp. 123–131.

[2] Max F Perutz. "X-ray analysis of hemoglobin". In: *Science* 140.3569 (1963), pp. 863–869.

[3] Dorothy Crowfoot et al. "XI. The X-Ray Crystallographic Investigation of the Structure of Penicillin". In: *Chemistry of penicillin*. Princeton University Press, 2015, pp. 310–366.

[4] Otto Kratky, I Pilz, and K Muller. *The world of neglected dimensions, small-angle scattering of X-rays and neutrons of biological macromolecules*. A. Paar KG, 1983.

[5] José Mensa et al. "Antibiotic selection in the treatment of acute invasive infections by Pseudomonas aeruginosa: Guidelines by the Spanish Society of Chemotherapy". In: *Revista Española de Quimioterapia* 31.1 (2018), p. 78.

[6] Xian-Zhi Li, Hiroshi Nikaido, and Keith Poole. "Role of mexA-mexB-oprM in antibiotic efflux in Pseudomonas aeruginosa." In: *Antimicrobial agents and chemotherapy* 39.9 (1995), pp. 1948–1953.

[7] Lea M Sommer, Helle K Johansen, and Søren Molin. "Antibiotic resistance in Pseudomonas aeruginosa and adaptation to complex dynamic environments". In: *Microbial Genomics* (2020), mgen000370.

[8] Jiuxin Qu et al. "Persistent bacterial coinfection of a COVID-19 patient caused by a genetically adapted Pseudomonas aeruginosa chronic colonizer". In: *Frontiers in cellular and infection microbiology* 11 (2021), p. 129.

[9] Kohjiro Saito et al. "Molecular mechanism of MexR-mediated regulation of MexAB–OprM efflux pump expression in Pseudomonas aeruginosa". In: *FEMS microbiology letters* 195.1 (2001), pp. 23–28.

[10] Quach Ngoc Tung et al. "The redox-sensing MarR-type repressor HypS controls hypochlorite and antimicrobial resistance in Mycobacterium smegmatis". In: *Free Radical Biology and Medicine* 147 (2020), pp. 252–261.

[11] Daniel Lim, Keith Poole, and Natalie CJ Strynadka. "Crystal structure of the MexR repressor of themexRAB-oprM multidrug efflux operon of Pseudomonas aeruginosa". In: *Journal of biological chemistry* 277.32 (2002), pp. 29253–29259.

[12] Anne Grove. "MarR family transcription factors". In: *Current biology* 23.4 (2013), R142–R143.

[13] Bin Pan, Indira Unnikrishnan, and David C LaPorte. "The binding site of the IclR repressor protein overlaps the promoter of aceBAK." In: *Journal of bacteriology* 178.13 (1996), pp. 3982–3984.

[14] Lizhen Gui, Alden Sunnarborg, and David C LaPorte. "Regulated expression of a repressor protein: FadR activates iclR." In: *Journal of bacteriology* 178.15 (1996), pp. 4704–4709.

[15] Mark S Wilke et al. "The crystal structure of MexR from Pseudomonas aeruginosa in complex with its antirepressor ArmR". In: *Proceedings of the National Academy of Sciences* 105.39 (2008), pp. 14832–14837.

[16] Hao Chen et al. “The Pseudomonas aeruginosa multidrug efflux regulator MexR uses an oxidation-sensing mechanism”. In: *Proceedings of the National Academy of Sciences* 105.36 (2008), pp. 13586–13591.

[17] Hao Chen et al. “Structural insight into the oxidation-sensing mechanism of the antibiotic resistance of regulator MexR”. In: *EMBO reports* 11.9 (2010), pp. 685–690.

[18] Marina Eiting et al. “The mutation G145S in PrfA, a key virulence regulator of Listeria monocytogenes, increases DNA-binding affinity by stabilizing the HTH motif”. In: *Molecular microbiology* 56.2 (2005), pp. 433–446.

[19] Madhanagopal Anandapadamanaban et al. “Mutation-induced population shift in the MexR conformational ensemble disengages DNA binding: a novel mechanism for MarR family derepression”. In: *Structure* 24.8 (2016), pp. 1311–1321.

[20] Minsun Hong et al. “Structure of an OhrR-ohrA operator complex reveals the DNA binding mechanism of the MarR family”. In: *Molecular cell* 20.1 (2005), pp. 131–141.

[21] Kelly Evans, Lateef Adewoye, and Keith Poole. “MexR Repressor of the mexAB-oprMMultidrug Efflux Operon of Pseudomonas aeruginosa: Identification of MexR Binding Sites in the mexA-mexRIntergenic Region”. In: *Journal of bacteriology* 183.3 (2001), pp. 807–812.

[22] Cecilia Andrésen et al. “Critical biophysical properties in the Pseudomonas aeruginosa efflux gene regulator MexR are targeted by mutations conferring multidrug resistance”. In: *Protein science* 19.4 (2010), pp. 680–692.

[23] Inoka C Perera and Anne Grove. “Molecular mechanisms of ligand-mediated attenuation of DNA binding by MarR family transcriptional regulators”. In: *Journal of molecular cell biology* 2.5 (2010), pp. 243–254.

[24] JW Peng, H Yuan, and XS Tan. “Crystal structure of the multiple antibiotic resistance regulator MarR from Clostridium difficile”. In: *Acta Crystallographica Section F: Structural Biology Communications* 73.6 (2017), pp. 363–368.

[25] Thirumananseri Kumarevel. “The MarR family of transcriptional regulators–a structural perspective”. In: *Antibiotic Resistant Bacteria–A Continuous Challenge in the New Millennium* (2012), pp. 403–418.

[26] Zuqin Nie et al. “c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells”. In: *Cell* 151.1 (2012), pp. 68–79.

[27] Valeriya Posternak and Michael D Cole. “Strategically targeting MYC in cancer”. In: *F1000Research* 5 (2016).

[28] Charles Y Lin et al. “Transcriptional amplification in tumor cells with elevated c-Myc”. In: *Cell* 151.1 (2012), pp. 56–67.

[29] Ami Albihn, John Inge Johnsen, and Marie Arsenian Henriksson. “MYC in oncogenesis and as a target for cancer therapies”. In: *Advances in cancer research*. Vol. 107. Elsevier, 2010, pp. 163–224.

[30] Leo Kretzner, Elizabeth M Blackwood, and Robert N Eisenman. “Myc and Max proteins possess distinct transcriptional activities”. In: *Nature* 359.6394 (1992), p. 426.

[31] Satish K Nair and Stephen K Burley. “X-ray structures of Myc-Max and Mad-Max recognizing DNA: molecular bases of regulation by proto-oncogenic transcription factors”. In: *Cell* 112.2 (2003), pp. 193–205.

[32] Adrian R Ferré-D’Amaré et al. “Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain”. In: *Nature* 363.6424 (1993), p. 38.

[33] Dongxue Wang et al. “MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma”. In: *Nucleic acids research* 45.5 (2016), pp. 2396–2407.

[34] Allison Doerr. “Cryo-electron tomography”. In: *Nature Methods* 14.1 (2017), pp. 34–34.

[35] Anthony WP Fitzpatrick et al. “Cryo-EM structures of tau filaments from Alzheimer’s disease”. In: *Nature* 547.7662 (2017), pp. 185–190.

[36] Mark A Herzik, Mengyu Wu, and Gabriel C Lander. “High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM”. In: *Nature Communications* 10.1 (2019), pp. 1–9.

[37] Bruce Alberts et al. “Protein function”. In: *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[38] H Lodish, Zipursky SL BA, et al. “Section 2.2, Noncovalent Bonds”. In: *Molecular Cell Biology* (2000).

[39] Olivier Lichtarge, Henry R Bourne, and Fred E Cohen. “An evolutionary trace method defines binding surfaces common to protein families”. In: *Journal of molecular biology* 257.2 (1996), pp. 342–358.

[40] Albert Furrer, Joel F Mesot, and Thierry Strässle. *Neutron scattering in condensed matter physics*. Vol. 4. World Scientific Publishing Company, 2009.

[41] F Sears Varley. “Neutron scattering lengths and cross sectioirn”. In: *Neutron news* 3.3 (1992), pp. 29–37.

[42] G Allen and JS Higgins. “Physico-chemical aspects of neutron studies of molecular motion”. In: *Reports on Progress in Physics* 36.9 (1973), p. 1073.

[43] B Jacrot. “The study of biological structures by neutron scattering from solution”. In: *Reports on progress in physics* 39.10 (1976), p. 911.

[44] HB Stuhrmann. “New method for determination of surface form and internal structure of dissolved globular proteins from snall angle x-ray measurements”. In: *Zeitschrift Fur Physikalische Chemie-Frankfurt* 72.4-6 (1970), pp. 177–+.

[45] Maria Monica Castellanos, Arnold McAuley, and Joseph E Curtis. “Investigating structure and dynamics of proteins in amorphous phases using neutron scattering”. In: *Computational and structural biotechnology journal* 15 (2017), pp. 117–130.

[46] Michel HJ Koch, Patrice Vachette, and Dmitri I Svergun. “Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution”. In: *Quarterly reviews of biophysics* 36.2 (2003), pp. 147–227.

[47] DI Svergun and HB Stuhrmann. “New developments in direct shape determination from small-angle scattering. 1. Theory and model calculations”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 47.6 (1991), pp. 736–744.

[48] DI Svergun et al. “New developments in direct shape determination from small-angle scattering. 2. Uniqueness”. In: *Acta Crystallographica Section A: Foundations of Crystallography* 52.3 (1996), pp. 419–426.

[49] P Chacon et al. “Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm”. In: *Biophysical Journal* 74.6 (1998), pp. 2760–2775.

[50] Dmitri I Svergun and Knud H Nierhaus. “A map of Protein-rRNA distribution in the 70 SEscherichia coli ribosome”. In: *Journal of Biological Chemistry* 275.19 (2000), pp. 14432–14439.

[51] Daniel Franke and Dmitri I Svergun. “DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering”. In: *Journal of applied crystallography* 42.2 (2009), pp. 342–346.

[52] SE Harding and K Jumel. *Current protocols in protein science.* 1998.

[53] AJF Siegert. *On the fluctuations in signals returned by many independently moving scatterers.* Radiation Laboratory, Massachusetts Institute of Technology, 1943.

[54] Ernesto Freire, Obdulio L Mayorga, and Martin Straume. “Isothermal titration calorimetry”. In: *Analytical chemistry* 62.18 (1990), 950A–959A.

[55] Michael M Pierce, CS Raman, and Barry T Nall. “Isothermal titration calorimetry of protein–protein interactions”. In: *Methods* 19.2 (1999), pp. 213–221.

[56] David G Myszka et al. “The ABRF-MIRG’02 study: assembly state, thermodynamic, and kinetic analysis of an enzyme/inhibitor interaction”. In: *Journal of biomolecular techniques: JBT* 14.4 (2003), p. 247.

[57] Jörg Stetefeld, Sean A McKenna, and Trushar R Patel. “Dynamic light scattering: a practical guide and applications in biomedical sciences”. In: *Biophysical reviews* 8.4 (2016), pp. 409–427.

[58] Dennis E Koppel. “Analysis of macromolecular polydispersity in intensity correlation spectroscopy: the method of cumulants”. In: *The Journal of Chemical Physics* 57.11 (1972), pp. 4814–4820.

[59] Szabolcs Fekete et al. “Theory and practice of size exclusion chromatography for the analysis of protein aggregates”. In: *Journal of pharmaceutical and biomedical analysis* 101 (2014), pp. 161–173.

[60] Marion M Bradford. “A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding”. In: *Analytical biochemistry* 72.1-2 (1976), pp. 248–254.

[61] Paul K Glasoe and FA Long. “Use of glass electrodes to measure acidities in deuterium oxide1, 2”. In: *The Journal of Physical Chemistry* 64.1 (1960), pp. 188–190.

[62] Ashley Jordan et al. “SEC-SANS: size exclusion chromatography combined in situ with small-angle neutron scattering”. In: *Journal of applied crystallography* 49.6 (2016), pp. 2015–2020.

[63] Warren M Garrison. “Reaction mechanisms in the radiolysis of peptides, polypeptides, and proteins”. In: *Chemical reviews* 87.2 (1987), pp. 381–398.

[64] André Guinier, Gérard Fournet, and Kenneth L Yudowitch. “Small-angle scattering of X-rays”. In: (1955).

[65] OTTO Glatter. “The interpretation of real-space information from small-angle scattering experiments”. In: *Journal of Applied Crystallography* 12.2 (1979), pp. 166–175.

[66] Martin A Schroer. “Small angle X-ray scattering studies on proteins under extreme conditions”. PhD thesis. Universitätsbibliothek Technische Universität Dortmund, 2011.

[67] K_ IbeL and HB Stuhrmann. “Comparison of neutron and X-ray scattering of dilute myoglobin solutions”. In: *Journal of molecular biology* 93.2 (1975), pp. 255–265.

[68] Karen Manalastas-Cantos et al. "ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis". In: *Journal of Applied Crystallography* 54.1 (2021).

[69] DI Svergun. "Determination of the regularization parameter in indirect-transform methods using perceptual criteria". In: *Journal of applied crystallography* 25.4 (1992), pp. 495–503.

[70] Andrew E Whitten, Shuzhi Cai, and Jill Trewhella. "MULCh: modules for the analysis of small-angle neutron contrast variation data from biomolecular assemblies". In: *Journal of Applied Crystallography* 41.1 (2008), pp. 222–226.

[71] Sergei Grudinin, Maria Garkavenko, and Andrei Kazennov. "Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles". In: *Acta Crystallographica Section D: Structural Biology* 73.5 (2017), pp. 449–464.

[72] Christian B Anfinsen. "Principles that govern the folding of protein chains". In: *Science* 181.4096 (1973), pp. 223–230.

[73] Szymon Kaczanowski and Piotr Zielenkiewicz. "Why similar protein sequences encode similar three-dimensional structures?" In: *Theoretical Chemistry Accounts* 125.3 (2010), pp. 643–650.

[74] Björn Wallner and Arne Elofsson. "All are not equal: a benchmark of different homology modeling programs". In: *Protein Science* 14.5 (2005), pp. 1315–1327.

[75] William R Pearson. "Using the FASTA program to search protein and DNA sequence databases". In: *Computer Analysis of Sequence Data*. Springer, 1994, pp. 307–331.

[76] Narayanan Eswar et al. "Protein structure modeling with MODELLER". In: *Structural proteomics*. Springer, 2008, pp. 145–159.

[77] Rong Chen, Li Li, and Zhiping Weng. "ZDOCK: an initial-stage protein-docking algorithm". In: *Proteins: Structure, Function, and Bioinformatics* 52.1 (2003), pp. 80–87.

[78] Brian G Pierce, Yuichiro Hourai, and Zhiping Weng. "Accelerating protein docking in ZDOCK using an advanced 3D convolution library". In: *PloS one* 6.9 (2011), e24657.

[79] Howook Hwang et al. "Performance of ZDOCK and ZRANK in CAPRI rounds 13–19". In: *Proteins: Structure, Function, and Bioinformatics* 78.15 (2010), pp. 3104–3110.

[80] Yumeng Yan et al. "HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy". In: *Nucleic acids research* 45.W1 (2017), W365–W373.

[81] Philippe Durand, Georges Trinquier, and Yves-Henri Sanejouand. "A new approach for determining low-frequency normal modes in macromolecules". In: *Biopolymers: Original Research on Biomolecules* 34.6 (1994), pp. 759–771.

[82] Alexandre Hoffmann and Sergei Grudinin. "NOLB: Nonlinear rigid block normal-mode analysis method". In: *Journal of chemical theory and computation* 13.5 (2017), pp. 2123–2134.

[83] Richard Bonneau et al. "Rosetta in CASP4: progress in ab initio protein structure prediction". In: *Proteins: Structure, Function, and Bioinformatics* 45.S5 (2001), pp. 119–126.

[84] SM LaValle and JJ Kuffner. "Proceedings Workshop on the Algorithmic Foundations of Robotics". In: (2001).

[85] Dina Schneidman-Duhovny et al. "FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles". In: *Nucleic acids research* 44.W1 (2016), W424–W429.

[86] Francesco Pesce and Kresten Lindorff-Larsen. "Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data". In: *Biophysical journal* 120.22 (2021), pp. 5124–5135.

[87] Gerhard Hummer and Jürgen Köfinger. "Bayesian ensemble refinement by replica simulations and reweighting". In: *The Journal of chemical physics* 143.24 (2015), 12B634_1.

Appendix

Genetic algorithm

In computer science and operations research, a genetic algorithm is a metaheuristic inspired by the process of natural picking that belongs to the larger class of evolutionary algorithms. Genetic algorithms are typically used to generate high-quality solutions to optimisation and search problems by relying on biologically inspired operators such as mutation, crossover and selection. A population of candidate solutions to an optimisation problem is evolved toward better solutions in a genetic algorithm. Each prospect solution has a set of properties that can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

The evolution usually starts from a population of randomly generated individuals and is an iterative process, with the population in each iteration called a generation. In each generation, the fitness of every individual is evaluated. The more fit individuals are stochastically selected from the current population, and each individual's genome is modified (recombined and possibly randomly mutated) to form a new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached.

Operon

Operon is a genetic regulatory system found in bacteria and their viruses in which genes coding for functionally related proteins are clustered along the DNA. This feature allows protein synthesis to be controlled coordinately in response to the needs of the cell. By providing the means to produce proteins only when and where they are required, the operon allows the cell to conserve energy (which is an essential part of an organism's life strategy).

A typical operon consists of a group of structural genes that code for enzymes involved in a metabolic pathway, such as the biosynthesis of an amino acid. These genes are located contiguously on a stretch of DNA and are controlled by one promoter (a short segment of DNA to which the RNA polymerase binds to initiate transcription). A single unit of messenger RNA (mRNA) is transcribed from the operon and is subsequently translated into separate proteins.

The promoter is controlled by different regulatory elements that respond to environmental cues. One standard regulation method is carried out by a regulator protein that binds to the operator region, another short segment of DNA found between the promoter and the structural genes. The regulator protein can either block transcription, referred to as a repressor protein, or an activator protein, which can stimulate transcription. Further regulation occurs in some operons: a molecule called an inducer can bind to the repressor, inactivating it, or a repressor may not be able to bind to the operator unless it is bound to another molecule, the corepressor. Some operons are under attenuator control, in which transcription is initiated but is halted before the mRNA is transcribed. This introductory region of the mRNA is called the leader sequence; it includes the attenuator region, which can fold back on itself, forming a stem-and-loop structure that blocks the RNA polymerase from advancing along the DNA.

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

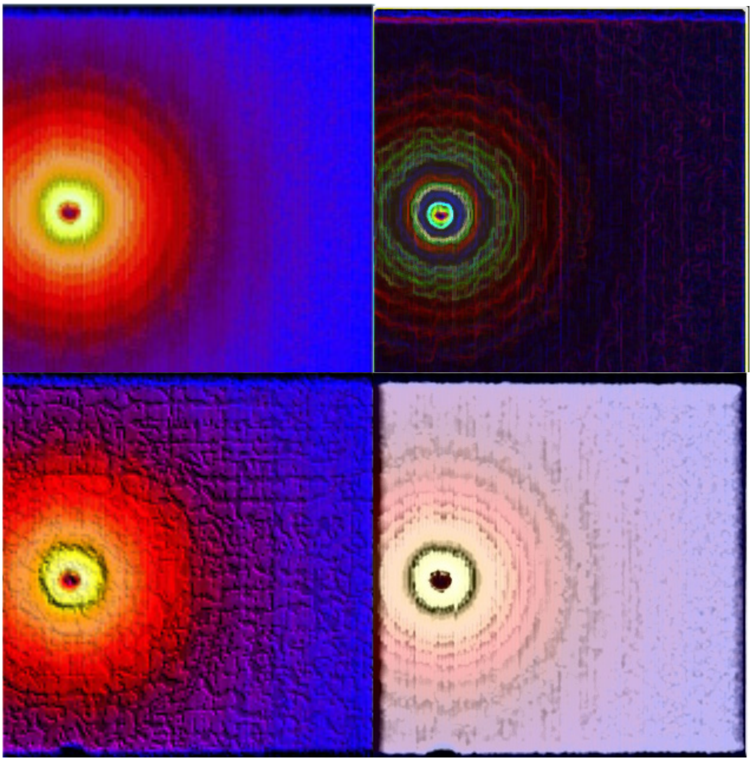
https://doi.org/10.3384/9789179296087

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2155, 2022
Department of Physics, Chemistry and Biology

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se



LINKÖPING
UNIVERSITY