Stability and aggregation-prone conformations of an antibody fragment antigen-binding (Fab)

by

Nuria Codina Castillo

A thesis presented for the degree of

Doctor of Philosophy

in the

Research Department of Biochemical Engineering,

University College London

March 2019

Declaration

I, Nuria Codina Castillo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Antibody-based products have become the main drug class of approved biopharmaceuticals, with over 60 drugs on the market and many more in clinical development. However, many never reach the market because protein aggregates form during manufacturing and storage, which lower the efficacy of the product and may cause immune responses in patients. To date, very little is known about the structural conformers that initiate aggregation. Stability of the humanized fragment antigen-binding (Fab) A33 was first studied using molecular dynamic (MD) simulations under two stresses, low pH and high temperature. Results revealed different unfolding pathways, with C_L domain partially unfolding at low pH, and C_L and V_H at high temperature. These conformational changes exposed different predicted aggregation-prone regions (APR), to suggest different aggregation mechanisms. Further salt bridge analysis provided insights into the ionizable residues likely to get protonated first. Mutational study with FoldX and Rosetta predicted that the constant domain interface can be stabilized further, backed by packing density calculations. To experimentally characterize the aggregation-prone conformers, solution structures of Fab A33 under different conditions of pH and salt concentration, were solved using small angle X-ray scattering (SAXS). SAXS revealed an expanded conformation at pH 5.5 and below, with an R_g increase of 2.2% to 4.1%, that correlated with accelerated aggregation. Scattering data were fitted using 45,000 structures obtained from the atomistic MD simulations under the same conditions, to locate the conformational change at low pH to the C_L domain. The approach was then validated using intra-molecular single-molecule FRET with a dual-labelled Fab as an orthogonal detection method. The conformational changes were found to expose a predicted APR, which forms a mechanistic basis for subsequent aggregation. Overall, these findings provide a means by which aggregation-prone conformers can be determined experimentally, and thus potentially used to guide protein engineering, or ligand binding strategies, with the aim of stabilizing the protein against aggregation.

Impact Statement

The work performed in this PhD thesis falls under the umbrella of stabilization of therapeutic antibody products and characterization of protein aggregation mechanisms. The methodologies and results found here, could be of interest and applicability to both, academic and industrial sectors. To understand the impact of this work, we need to put it into context. Currently, antibody-based products are the most rapidly growing class of pharmaceuticals, because of their high specificity towards their targets (e.g. biomarkers on the surface of cancer cells). Unfortunately, they tend to aggregate during all stages of product development, which leads to decreased efficiency and could elicit an immunological response. Methods for improving the stability of therapeutic antibodies are generally done during the development phase, by trial and error of the composition of the formulated product, which are both costly and time consuming. There is great demand and potential for identifying the drivers of instability across different stress conditions, early in the discovery phase, which will enable the rational engineering of protein scaffolds that are inherently manufacturable. In this context, the first section of this thesis elucidated the stability-limiting regions of the antibody fragment Fab A33 using several computational tools: atomistic molecular dynamics simulations, in-silico mutational analysis by FoldX and Rosetta, predictors of aggregation-prone regions, packing density calculators and analysis of existing Fab sequences. My results identified mutations to those regions that have the potential to stabilize Fab fragments to both thermal and pH-stresses simultaneously. The methodology used here, could greatly improve the developability screening of candidate antibody products for many diseases, such as cancer, chronic inflammatory diseases, infectious diseases, and cardiovascular medicine. These research findings are currently in preparation for submission to the journal PLOS Computational Biology.

The second major research finding in this PhD thesis provided molecular-level insights into the early stages of the aggregation mechanism of Fab A33. Small-angle X-ray scattering revealed that the aggregation of Fab A33 correlated with a slight expansion of native state upon acidification. Little is known about the structures of native conformers that initiate aggregation. Here, I used SAXS analysis at the very limits of its capability, by fitting the data to full molecular dynamics simulations under the same

conditions, to obtain atomistic structural information. This revealed the regions of the Fab undergoing conformational fluctuations. Additionally, I used single-molecule FRET on dual-labelled Fabs as an orthogonal detection method, to confirm the displacement of local regions in Fab under low pH. Finally, the conformational changes were found to expose a predicted aggregation-prone region (APR) as a likely aggregation mechanism. This research highlights the promise of SAXS combined with molecular dynamics simulations to resolve aggregation-prone conformers within native ensembles, particularly for large proteins that are less accessible by NMR. The findings also confirm the importance of combining predictors of aggregating-regions with structural changes in the protein, and adds further evidence to the importance of local unfolded states in the aggregation mechanisms of globular proteins. This work was published in the Journal of Molecular Biology (2019). In a more general picture, protein aggregation also plays a central role in human diseases such as Alzheimer's and Parkinson's diseases. Thus, an understanding of the aggregation-prone conformers and regions involved in protein aggregation is of importance both for stabilizing protein therapeutics and for devising strategies to prevent in vivo aggregation, either via protein engineering or formulation, or for the design of drugs that bind to and stabilize proteins against aggregation.

Publications

Codina, N., Hilton D., Zhang C., Chakroun N., Ahmad S. S., Perkins S. J., and Dalby P. A. (2019) An expanded conformation of an antibody Fab region by X-ray scattering, molecular dynamics, and smFRET identifies an aggregation mechanism. Journal of Molecular Biology, https://doi.org/10.1016/j.

Codina, N., Zhang C., Chakroun N., and Dalby P. A. (2019) Insights into the stability of a therapeutic antibody Fab fragment by molecular dynamics and its stabilization by computational design. [In preparation for submission to PLOS Computational Biology].

Presentations and abstracts

Codina, N., Hilton D., Zhang C., Perkins S. J., and Dalby P. A. Elucidation of an expanded aggregation-prone conformation of Fab using SAXS, MD simulations and smFRET. 32nd Annual Symposium of The Protein Society. 9-12th July 2018, Boston, Massachusetts. (Poster). **Winner of travel award.**

Codina, N., Perkins S. J., and Dalby P. A. Elucidation of an expanded aggregation-prone conformation of Fab using SAXS, MD simulations and smFRET. Royal Society of Chemistry Protein and Peptide Subject Group Early Stage Researcher Meeting. 20th July 2018, Southampton, United Kingdom. (Oral)

Codina, N., Hilton D., Zhang C., Chakroun N., Perkins S. J., and Dalby P. A. Insights into the stability and mechanism of unfolding of a therapeutic fragment antibody by SAXS and MD simulations. PEGS Europe: Protein & Antibody Engineering Summit. 12-16th November 2018, Lisbon, Portugal. (Poster).

Acknowledgements

Firstly, I would like to thank my PhD supervisor, Prof Paul Dalby, for his guidance and support during my studies at UCL. I am very grateful for his mentoring, his approachability, his vast knowledge, and his encouragement and motivation during all the projects undergone in this thesis. I also want to thank our collaborator, Prof Steve Perkins, for his detailed guidance in technical aspects, such as data analysis, and in preparing scientific manuscripts. Looking back, I feel privileged to have had the opportunity to work with both of them, (I will miss them), and I take many great lessons on performing scientific research with me.

My time at UCL has been greatly formative. I felt very welcome at both departments I worked at, the department of Biochemical Engineering and the department of Structural and Molecular Biology, where I met many motivated and inspirational researchers. I would like to thank Dr Cheng Zhang, for his mentoring in protein expression and purification and computational tools, including molecular dynamics. I also want to thank Dr David Hilton and Dr Nesrine Chakroun for laying the basis where my worked followed. I also want to thank Valentina Spiteri for helping me get started in X-ray scattering and many insightful discussions. I extend the gratitude to all my labmates, including Dr Samir Aoudjane, Dr Haoran Yu, Dr Weiluo Lee, Henry Wilkinson, and Gar Kay-Hu. Their camaraderie and friendship made my PhD experience very enjoyable, and I cherish the time spent together with them.

I would like to thank my thesis committee, Dr Robin Curtis and Dr Konstantinos Thalassinos, for their time and dedication in reviewing this thesis.

I am particularly thankful to the Centres for Doctoral Training (CDT) within the Engineering and Physical Sciences Research Council (EPSRC) for the financial support and for giving me the opportunity to come and pursue my studies at UCL.

Last but not least, a heartfelt thanks to my loving family, my supportive friends and Cam.

Contents

List of Figures

List of Tables

List of Abbreviations

3D	Three dimensional
aaRS	Aminoacyl-tRNA synthetase
ADC	Antibody drug conjugate
ADCC	Antibody dependent cellular cytotoxicity
AF	Alexa Fluor
APD	Avalanche photodiode
APR	Aggregation-prone region
BLAST	Basic local alignment search tool
CDC	Complement dependent cytotoxicity
CDR	Complementarity-determining regions
C_H	Constant domain of the heavy chain
C_L	Constant domain of the light chain
CPEC	Circular polymerase extension cloning
cryo-EM	Cryo-electron microscopy
DIBO	Dibenzocyclooctyne
D_{max}	Maximum particle diameter
DNA	deoxyribonucleic acid
DOT	Dissolved oxygen tension
DSSP	dictionary of protein secondary structure
E. coli	Escherichia coli
E_{app}	Apparent FRET transfer efficiency
ELISA	Enzyme-linked immunosorbent assay
ESI mass spec	Electrospray ionization mass spectrometry
ESRF	European Synchrotron Radiation Facility
Fab	Fragment antigen-binding
Fc	Fragment crystallisable
FcRn	Neonatal Fc receptor
FDA	Food and drug administration
FPLC	Fast protein liquid chromatography
FRET	Förster resonance energy transfer
GdmCl	Guanidinium chloride
HAMA	Human anti-mouse antibody

HC	Heavy chain
HPLC	High performance liquid chromatography
Ig	Immunoglobulin
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IS	Ionic strength
LB	Lysogeny broth or Luria broth
LC	Light chain
mAb	Monoclonal antibody
MD	Molecular dynamics
N.A.	Not applicable
NEB	New England Biolabs
NMR	Nuclear magnetic resonance
NSAA	Nonstandard amino acid
ORI	Origin of replication
OS	Occluded surface
OSP	Occluded surface packing value
pAzF	p-azido-l-phenylalanine
PBS	Phosphate buffered saline buffer
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
pI	Isoelectric point
PPG	Polypropylene glycol
R_0	Förster radius
R_g	Radius of gyration
REU	Rosetta Energy Unit
RMSD	Root-mean-square deviation
RMSF	Root mean square fluctuation
SAS	Small-angle scattering
SASA	Solvent accessible surface area
SAXS	Small angle X-ray scattering
scFv	Single-chain variable fragments
SD	Standard deviation
SEC	Size-exclusion chromatography

SEM	Standard error of the mean
smFRET	Single-molecule Förster resonance energy transfer
SS	Secondary structure
TBE	Tris/Borate/EDTA buffer
TCEP	Tris(2-carboxyethyl)phosphine
T_m	Melting temperature
tRNA	Transfer RNA
UV	Ultraviolet
V_H	Variable domain of the heavy chain
Vis	Visible
V_L	Variable domain of the light chain
VMD	Visual molecular dynamics
WT	Wild type

List of Units

°C	degree Celsius
Å	Angstrom
cm	centimetre
fL	femtolitre
fs	femtosecond
g	gram
h	hour
K	Kelvin
kb	kilobase
kcal	kilocalorie
kDa	kilo Dalton
keV	kiloelectronvolt
L	litre
m	meter
µg	microgram
µL	microliter
µm	micrometre
µM	micromolar
µs	microsecond
µW	microwatts
mA	milliampere
mg	milligram
mL	millilitre
mM	millimolar
ms	millisecond
min	minute
M	molar (mol/L)
mol	mole
ng	nanogram
nm	nanometre
ns	nanosecond
pM	picomolar
pmol	picomol
ps	picosecond
rpm	revolutions per minute
s	second
V	volt

Chapter One

Introduction

1.1 Antibody and fragment antigen-binding (Fab)

1.1.1 Brief History of Antibodies

The history of antibodies is linked to that of vaccines. In 1798, Edward Jenner realized that milkmaids that had previously caught cowpox, would later not develop the very similar but more serious disease smallpox. He believed that exposure to cowpox provided protection against smallpox. And thus, in the first vaccination event, Edward Jenner gave fluid from a pustule of a cow infected with cowpox to a young boy, giving him protection to smallpox, and proving that immunity can be gained once a patient has already encountered a pathogen (Riedel 2015). It was not until the 1890 that the mechanism of protection provided by vaccination began to be understood. Emil von Behring and Shibasabura Kitasato demonstrated that serum from infected animals can be used to treat and prevent infection in other animals, in particular they studied diphtheria and tetanus (Kantha 1991). Emil von Behring would later win the Nobel prize in 1901, for the development of serum therapy (Kaufmann 2017). In 1900, Paul Ehrlich proposed the "side-chain theory", in which he hypothesized that cells express a variety of side-chains that can be shed into the blood to bind pathogens. Paul Ehrlich is considered one of the fathers of modern immunology (Winau et al. 2004). He also proposed a model for an antibody molecule in which the antibody was branched and consisted of multiple sites for binding to foreign material, known as antigen, and for the activation of the complement pathway (Davies & Chacko 1993). Paul Ehrlich won the Nobel prize in 1908 in recognition for his work on immunity. This model agreed with the "lock and key" hypothesis for enzymes proposed by Emil Fischer (Lemieux & Spohr 1994). In 1948, Astrid Fagreaus discovered that B cells, in the form of plasma cells, generated antibodies (Silverstein 2004). Further work focused on solving the antibody structure, and Gerald Edelman and Rodney Porter were jointly awarded the Nobel Prize in 1972 for independently discovering the molecular structure of antibodies (Pauling 1940; Porter 1959). The first atomic resolution structure of an antibody fragment was published in 1973 (Poljak et al. 1973). This was quickly followed by the invention of monoclonal antibodies in 1975 by Georges Köhler and César Milstein, who would later win the Nobel prize in 1984 for the discovery of production of monoclonal antibodies (Köhler & Milstein 1975). This marked the start of the modern era of antibody research and discovery.

1.1.2 Antibody Structure

Before introducing the fragment antigen-binding (Fab) of an antibody, it is necessary to introduce full-size antibodies. Antibodies are glycoproteins belonging to the immunoglobulin superfamily that are secreted by B cells to identify and neutralize foreign organisms or antigens. Antibodies circulate in the blood, and when they find an unfamiliar foreign object, such as a virus or bacteria, they bind tightly to its surface. Coating might be enough to prevent infection, if not, antibodies act as markers to alert the immune system and activate the other defensive mechanisms (Davies & Metzger 1983). The basic functional unit of an antibody is an immunoglobulin (Ig) monomer, as is the case for IgG, typically used as prototype to explain the structure of all antibodies (other antibody classes can be multimeric). Antibodies are "Y"-shaped proteins (Figure 1.1), and their structure helps explain their binding specificity and their biological activity (Harris et al. 1992). Antibodies are formed by four chains, two identical long heavy chains (H; 50 kDa each) and two identical short light chains (L; 25 kDa each), linked by disulphide bonds and non-covalent interactions. Each light chain pairs with a heavy chain, and each heavy chain pairs with another heavy chain. The total molecular weight of an IgG is approximately 150 kDa, and its size about 10 nm. The arms of an antibody molecule contain the antigen-binging sites of the antibody, and thus are called fragment antigen-binding (Fab), whereas the stem of the antibody, interacts with effector cells within the immune system to elicit a physiological response, and is called fragment crystallisable (Fc). Thus, an antibody contains two identical Fab fragments (50 kD), with two identical antigen-binding sites at the tips of the Fab arms, and one Fc fragment (50 kD) (Alzari et al. 1988).

By comparing many antibody amino acid sequences, it was found that each light and heavy chain are comprised of a region of high variability at the amino terminal (about 110 first amino acids), called the variable (V) region, and the remaining large region in the carboxyl terminal end was constant in different types of antibodies, called the constant (C) region (Wang et al. 2007). The variable regions contain the antigen-binding site, and varies greatly from one antibody to another. Specifically, variability is concentrated in regions called complementarity-determining regions (CDRs) or hypervariable regions. There are three CDRs in each chain, moving from the amino terminal end they are called CDR1, CDR2, and CDR3, each about 10 amino acids in length. The more conserved amino acids between the CDRs are called framework residues, and compose about 85%

of the variable region. Framework residues define the positioning of the CDRs, and hold them in place. The variable region folds so that the CDRs are exposed on the surface of the chain. When the light and heavy chains are joined, the CDRs of the chains form a cleft that serves as the antigen-binding site. Changes in amino acid residues at the position of this cavity change its shape and thus its specificity, generating millions of antibodies with slightly different antigen-binding sites. This enormous diversity of antibody CDRs on the antigen-binding fragments allows the immune system to recognize an equally wide variety of antigens.

The three-dimensional structure of antibodies is divided in domains, called immunoglobulin domains. Each domain contains about 70-110 amino acids. Each domain has a characteristic tertiary structure consisting of two layers of β-sheets, an inner β-sheet and an outer β-sheet, in a sandwich shape. Each domain contains a disulphide bridge near the center of the domain (Morea et al. 2000). The light chain of IgG has two domains called V_L and C_L. The heavy chain of IgG has four domains, one V_H and three in the C_H region (C_H1, C_H2, and C_H3). Extensive non-covalent interactions occur in the interface between domains V_L and V_H, C_L and C_H1. Where the arms meet, the stem of the Y (between C_H1 and C_H2 domains), is known as the hinge region. The hinge region allows segmental flexibility, which means that the two Fab regions can move relative to one another on antigen binding. There are several disulphide bonds in an antibody molecule, the total number depends on the antibody class. As we have seen, there are disulphide bonds located within chains stabilizing the β-barrel domain fold, and the other disulphide bonds link the different chains, the two heavy chains and the heavy chain to the light chain. Antibodies are glycoproteins, thus they contain carbohydrates, situated within its Fc region, at conserved residues (Maverakis et al. 2015). The attached glycans determine the effector functions (Wright & Morrison 1998). The structure of the glycans determines the affinity of antibodies for Fc receptors, which directs the appropriate immune response for each different type of foreign object they encounter.

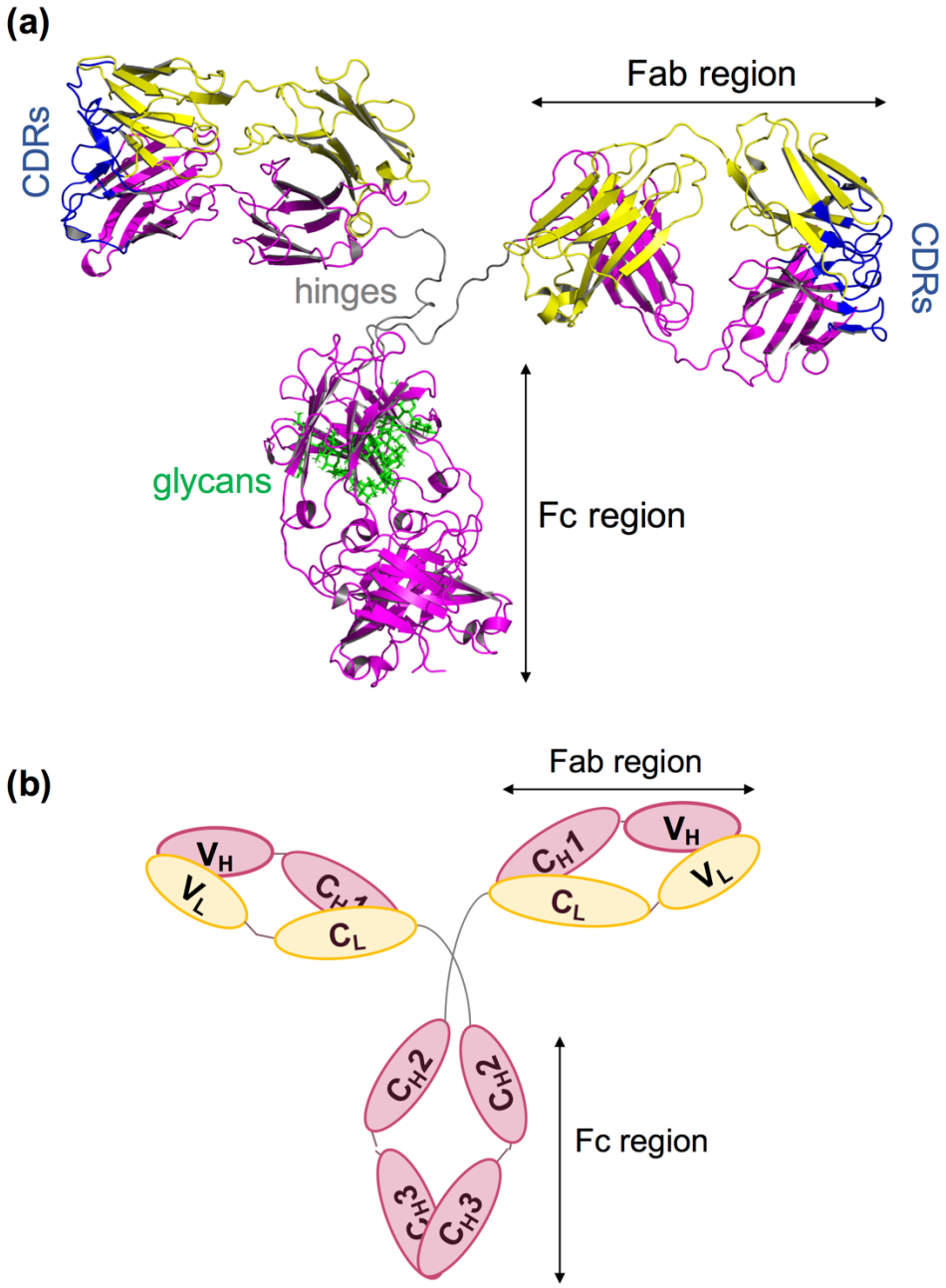

Figure 1.1. Antibody structure. (a) Cartoon representation of a full-size antibody (PDB ID: 1igt). Antibodies have two identical light (L; yellow) and heavy (H; magenta) chains. The arms of an antibody are termed fragment antigen-binding (Fab) and the stem is called fragment crystallizable (Fc). The flexible regions connecting Fabs and Fc are the hinge regions (gray). The antigen-binding region at the complementary determining regions (CDRs; blue), are located at the tips of the Fabs. The glycans are located at the C_H2 domain (green). (b) Schematic representation of a full-size antibody, divided by domains. Each chain is divided into two regions, the variable (V) and constant (C) regions. Light chains have two domains (V_L and C_L) and heavy chains have four domains (V_H, C_H1, C_H2 and C_H3).

1.1.3 Antibody Isotypes and Subtypes

Antibodies themselves can be immunogenic, for example, if humans are immunized with mouse antibodies, our immune system recognizes them as a foreign complex glycoprotein, and we generate antibodies against them. Thus, antibodies contain antigenic determinants that allow us to classify them (Schroeder & Cavacini 2010). The antigenic determinants that divide antibodies into classes are called isotypic determinants. Isotypic determinants are located on the heavy chain, and divide antibodies into classes or isotypes. Humans have five antibody isotypes: IgG, IgA, IgM, IgD, and IgE (Figure 1.2). Specifically, they have different types of crystallizable fragments (Fc) (Woof & Burton 2004). IgG, IgD, and IgE are monomeric (one Ig unit), IgA is either monomeric or dimeric (two Ig units), and IgM is pentameric (five Ig units). Their location in the body, half-life, abundance and function, differ between them. IgG is the most abundant antibody in the blood (70-75% abundance), and has the longest half-life (20-24 days). IgG can enter tissue spaces (for instance the placenta), where it coats antigens, speeding their uptake. IgA (10-15% abundance), concentrates in body fluids to guard the entrances of the body and protect against pathogens. IgA is found in mucous, saliva, tears, milk and intestinal juice. IgM (10% abundance) is the largest antibody, and it tends to remain in the blood, where acts in the early stages of immune response, and it can efficiently kill bacteria (Collins et al. 2002). IgD (1% abundance) remains bound to the membrane of B cells and regulates the activation of cells like basophiles and mast cells. IgE is found in trace amounts in the blood (0.002% abundance) and protects against parasitic worms and allergic reactions.

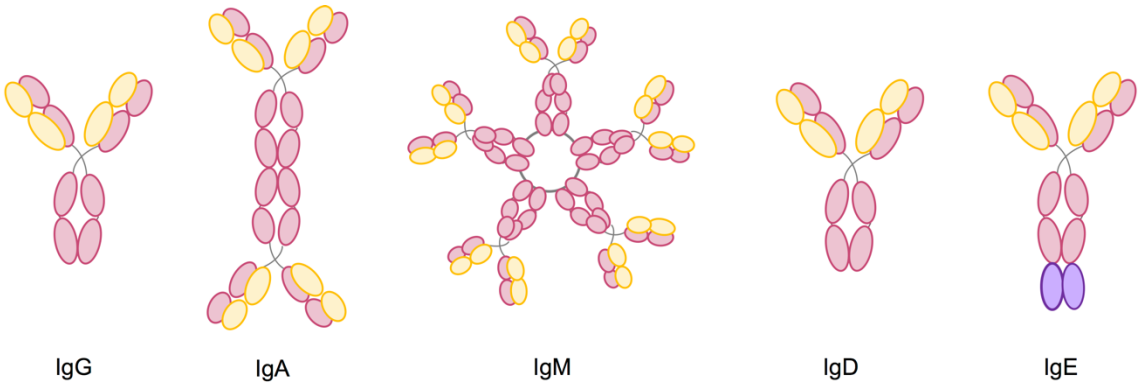

Figure 1.2. Antibody classes or isotypes. Schematic representation of the five immunoglobulin classes or isotypes in humans: IgG, IgA, IgM, IgD, and IgE.

Other isotypic heavy-chain determinants define differences within a class, generating antibody subclasses. IgG has four subclasses, called IgG1, IgG2, IgG3, and IgG4 (Figure 1.3). IgG isotypes differ in the hinge region and the location and number of disulphide bonds. IgG1 and IgG4 contain two inter-chain disulphide bonds in the hinge region, IgG2 has four and IgG3 has eleven. In humans, IgG1, IgG2, IgG3, and IgG4 are found in normal serum in the approximate proportions of 65, 25, 5, and 5%, respectively.

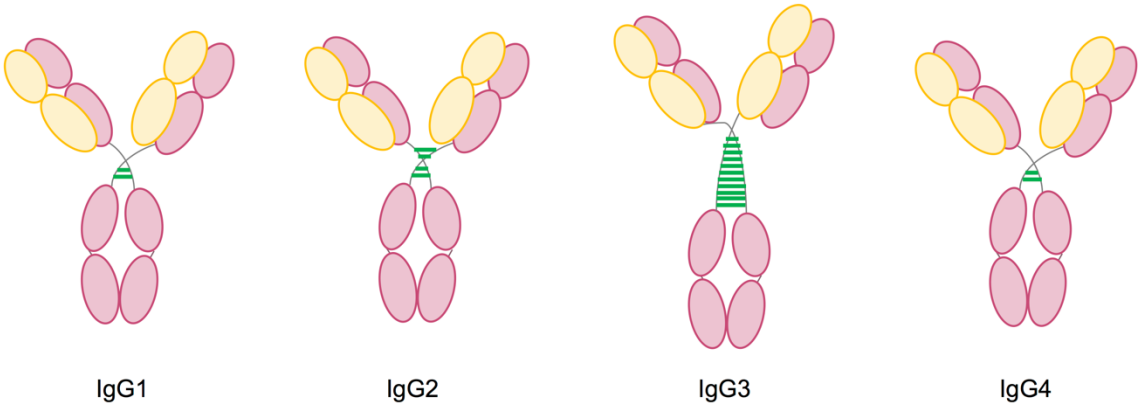

Figure 1.3. IgG antibody subclasses. Schematic representation of the four IgG subclasses: IgG1, IgG2, IgG3, and IgG4. Disulphide bonds are colored green.

Antigenic determinants on light chains classify them as either kappa (κ) or lambda (λ) chains. Light-chain antigenic determinants are not useful in determining antibody class, because κ and λ chains can be associated with any class of heavy chain (all classes and subclasses). A κλ chain combination never occurs on the same antibody molecule.

There are more specific antigenic determinants, called allotypic and idiotypic determinants. Allotypic determinants are carried by only some individuals within a given species and are inherited in a Mendelian fashion (Vidarsson et al. 2014). The genetic variation in allotypic determinants is due to individuals having different alleles. Idiotypic determinants are individual-specific and are located in the antigen-binding site of the antibody, in variable regions of the heavy and light chains.

1.1.4 Antibody Effector Functions

Whereas the Fab arms of the antibody, and specifically the small CDR loops, determine the specificity of the antibody, the Fc region determines the effector functions (Wang et al. 2018). Antibody effector functions can be divided into four major mechanisms: blocking, complement dependent cytotoxicity (CDC), antibody dependent cellular cytotoxicity (ADCC) and phagocytosis (Figure 1.4) (Lu et al. 2018). Blocking refers to binding to block parts of the surface of a bacterial cell or virus to render its attack ineffective. Antibodies alone cannot directly destroy a foreign organism, instead, antibodies mark them for destruction by other defense systems. One option for antibodies to cause cell lysis is by activating the classical complement pathway (CDC). Many antibodies cluster together and interact with C1q of the C1 complex, recruiting the complement membrane attack complex. Antibodies in an antibody-coated target cell also can interact with Fc receptors on effector cells (natural killer cells, macrophages, monocytes and eosinophils) to engage in antibody dependent cellular cytotoxicity. Once Fc receptors in the surface of effector cells bind to the Fc region of an antibody, a signaling pathway is triggered that results in the secretion of cytokines (lytic enzymes, perforin, granzymes and tumour necrosis factor) by the effector cell, which mediate the destruction of the target cell. Lastly, antibodies can stimulate the removal of an antibody-coated target cell by engulfment of a phagocyte (also called opsonization).

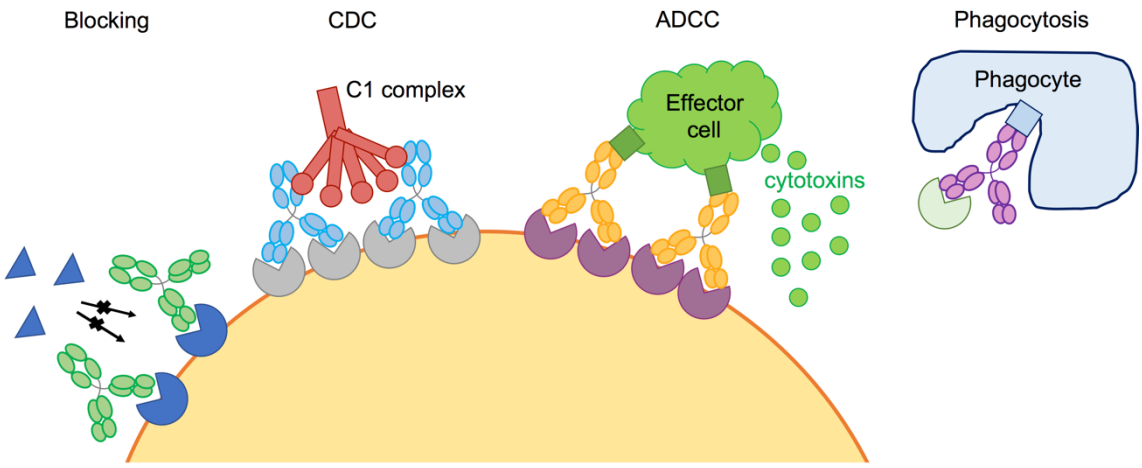

Figure 1.4. Antibody modes of action. Antibody effector functions include: blocking, complement dependent cytotoxicity (CDC), antibody dependent cellular cytotoxicity (ADCC) and phagocytosis.

1.1.5 Humanization

The first FDA-approved antibody to treat a human disease was a mouse antibody, in 1986, to treat kidney transplant rejection. However, it was soon realized that they presented a short serum half-life, they were not able to trigger human effector functions, and more problematic, our immune system recognized the mouse antibody as a foreign protein, and raised a human anti-mouse antibody (HAMA) response (Hwang & Foote 2005). Thus, it was necessary to humanize antibodies by engineering them, to be able to use them as therapeutic agents. The first attempts consisted in human constant domains (C_L and C_H) and mouse variable domains (V_L and V_H), which received the name of chimeric antibodies (about 66% human) (Figure 5.5) (Boulianne et al. 1984; Morrison et al. 1984). Removal of the mouse constant domains removed the most immunogenic part of the antibody, however, chimeric antibodies were still able to generate HAMA responses. In a further attempt to reduce the immunogenicity, only the CDRs of mouse antibodies were grafted onto the human variable region framework, creating humanized antibodies (about 90% human) (Figure 1.5) (Jones et al. 1986; Cheetham et al. 1998). In humanized antibodies, anti-antibody responses have still been noted in patients, however the severity of the response was reduced. Humanized antibodies have been approved and are in the market to treat diseases, such Zenapax (daclizumab), used to prevent organ transplant rejection (Przepiorka et al. 2000). Lastly, fully human antibodies are being engineered using two techniques, phage display and transgenic mice. In phage display, a library of human antibodies is expressed on the surface of phage and subsequently selected and amplified in *E. coli* (McCafferty et al. 1990). In transgenic mice, mice expressing a human antibody repertoire are used. These transgenic mice were generating by replacing the mouse antibody encoding genes with human versions, and the antibodies produced are fully human (Green et al. 1994; Nelson et al. 2010).

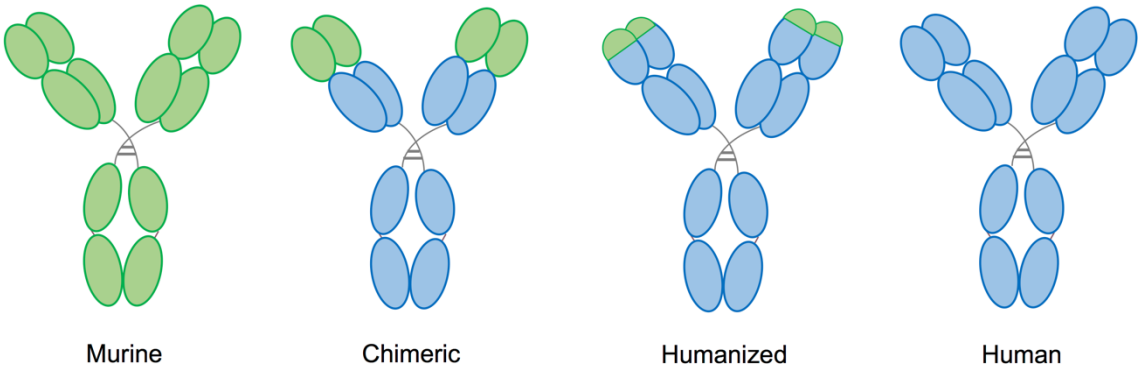

Figure 1.5. Progressive humanization of antibodies. A schematic representation of the advancement from fully mouse antibodies (green), to chimeric antibodies (~66% human), to humanized antibodies (~90% human), and human antibodies (blue).

1.1.6 Antibody Fragments

Full-size antibodies are not always desired for therapeutic applications, and there are advantages in using smaller antibody fragments. One main drawback of full-size antibodies, especially true for anti-cancer antibodies, is that due to their large size, they have difficulties penetrating some tissues, like the physical barriers of solid tumours (Christiansen & Rajasekaran 2004). The use of smaller fragments of these antibodies allows a deeper penetration into these tissues (Nelson 2010). Antibody fragments do not contain the Fc domain, and thus, do not induce Fc-mediated responses. Antibody fragments act by binding and blocking ligands or receptors. In some instances, the Fc mediated effects might not be required or even desired, such as for the treatment of autoimmune disorders, where activation of Fc receptor-expressing cells might be unwanted because of the toxicity associated with cytokine release. In addition, antibody fragments are not glycosylated, which allows their easier and less costly expression in prokaryotic systems (Holliger & Hudson 2005). The Fc domain allows the FcR-mediated recycling, which gives full-size antibodies their long half-life of 7-21 days. In contrast, antibody fragments are rapidly degraded in humans and have shorter elimination half-lives of only hours to days. This shorter circulation half-life can be useful in imaging applications, when the exposure of healthy tissues to radioisotope must be limited, and in cancer therapy, to reduce the prolonged exposure of radiolabelled antibodies (Brekke & Løset 2003). Alternatively, several strategies have been developed to extend the half-life

of antibody fragments, including conjugation to proteins such as albumin and PEGylation (Chapman et al. 1999), which was applied to the FDA approved anti-TNFα Fab, certolizumab pegol.

Manufacture of antibody fragments began in 1980s. Initially fragments were generated by proteolysis of full-size antibodies using enzymes. The enzyme papain cleaves the antibody molecule into two Fab fragments and a Fc fragment. The enzyme pepsin cleaves below the hinge region, generating a $F(ab')_2$ fragment (Figure 1.6), and a pFc' fragment. Later, genetic engineering was used to generate diverse therapeutic antibody fragments (Enever et al. 2009).

Fab fragments were the first antibody fragments to be generated as therapeutics. Fab fragments are monovalent and approximately three times smaller than full antibodies (50 kD). Each Fab is composed of one light (the entire light chain) and one heavy chain (part of the heavy chain), each comprising a variable (V_L and V_H) and a constant (C_L and C_H1) domain (Figure 1.6). The variable domains contain the antigen-binding site at their complementary determining regions (CDRs), formed by three loops in V_L and three loops in C_L. There are five disulphide bonds in Fab, four of them intra-domain and the last one between the light and heavy chains at the hinge region.

The next fragments to be developed were single-chain variable fragments (scFv) (Huston et al. 1988; Monnier et al. 2013). First, only the variable region (Fv) of antibodies was used, consisting of the variable domain of the heavy (V_H) and the variable domain of the light chain (V_L). However, the domains would dissociate, thus introduction of a flexible linker uniting them was necessary, forming a single polypeptide, called single-chain variable fragments (scFv) (Figure 1.6). The peptide linker connects the V_H and V_L domains. scFv are half the size of the Fab fragment (~28 kDa). Due to their small size, scFv have low stability and tend to multimerize (Wilkinson et al. 2009). For this reason, even though scFv retain the antigen binding site of the antibody, its low stability means that Fabs are more often used.

Smaller fragments, such as a single V_H domain, were also tried, however, they rarely retain the affinity of their parent antibody and due to their small size are poorly soluble and prone to aggregation.

Molecular engineers have continued to innovate by creating multimeric fragments with multi-specificity, mainly based on Fab fragments or scFv as building blocks (Cuesta et al. 2010). They have generated diabodies, which can either be bivalent or bispecific (Figure 1.6) (Holliger et al. 1993), and complementary scFvs themselves produced as a single chain (tandem scFvs) among others. Additionally, conjugation of antibodies and fragments to external molecules, such as drugs, forming antibody drug conjugates (ADCs) is also being explored (Dan et al. 2018).

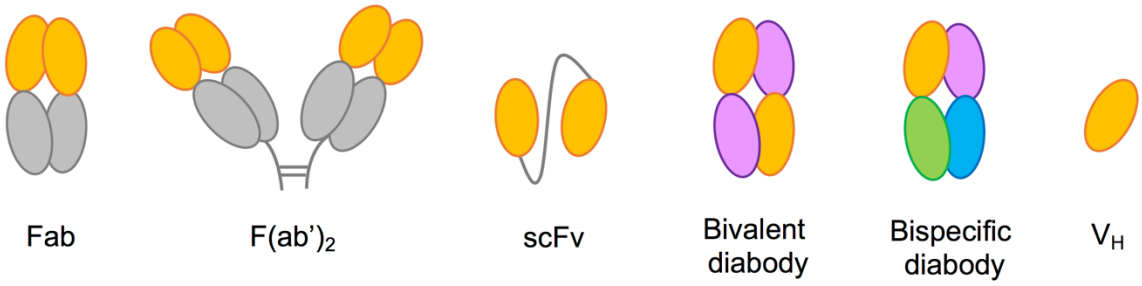

Figure 1.6. Antibody fragments constructs. Schematic representation of some of the most common antibody fragments: Fab, $F(ab')_2$, scFv, bivalent diabody, bispecific diabody and V_H.

1.1.7 Fab A33 sequence and structure

Fab A33 fragment was derived from a murine monoclonal antibody (MAb) A33, by UCB Celltech (Slough, UK). Fab A33 and murine Mab A33 recognize a protein expressed on the surface of colon cancer cells. The antigen is expressed on several human tumour cell lines, including Colo205, ASPC-1 and SW1222 cell lines. Initial studies revealed that murine Mab A33 generated human anti-mouse antibody (HAMA) responses in patients. Thus, murine Mab A33 was humanized and Fab A33 was generated by recombinant DNA technology, cloning into a human kappa light chain and the human heavy chain, IgG1 (King et al. 2001). The original Fab A33 contained a free thiol group in the hinge region, which was mutated to a serine to avoid dimer formation. In this thesis, I refer to Fab A33 C226S, as wild-type Fab A33.

Fab A33 is 442 amino acids in length and has a molecular weight of 47,378 Da. No crystal structure of Fab A33 was available, and a homology model was generated (see Materials and Methods section). Figure 1.7 shows the cartoon representation of the Fab A33 homology model. The sequence of Fab A33 is shown in Figure 1.8. Fab A33 has an estimated pI of 8.76, and an extinction coefficient of 67,435 M^{-1} cm^{-1} at 280 nm, based on calculations from the Expasy ProtParam tool (https://web.expasy.org/protparam/).

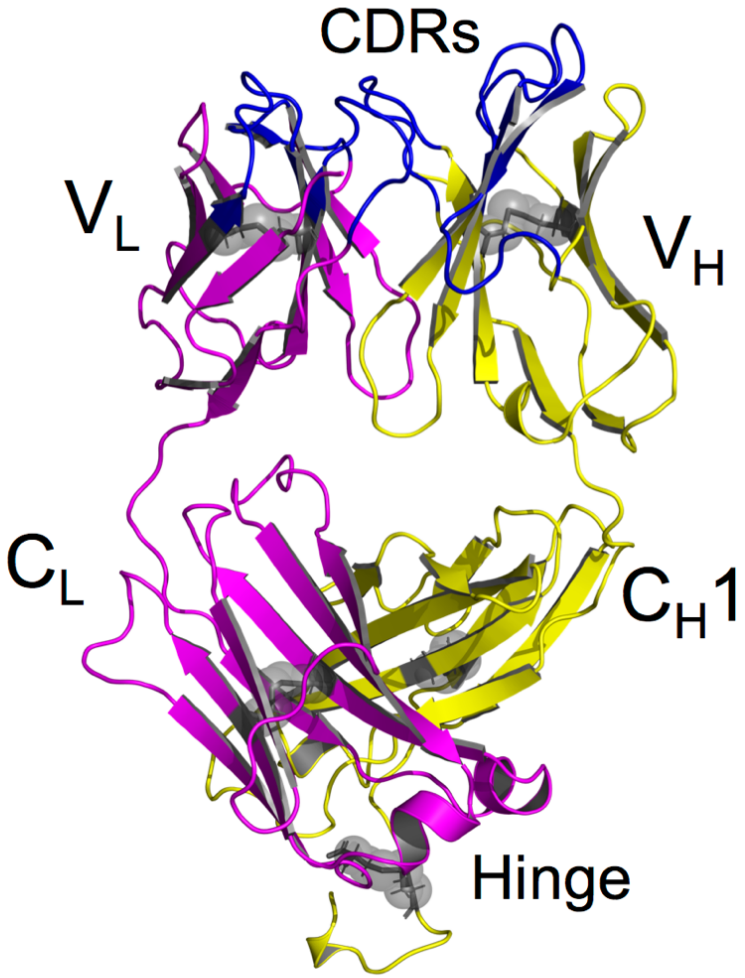

Figure 1.7. Structure of Fab A33. Fab is composed of light (magenta) and heavy (yellow) chains. Each chain contains variable (V_L and V_H) and constant (C_L and C_H1) domains. The antigen-binding region at the complementary determining regions (CDRs; blue), are located in the variable domains. There are five disulphide bonds (gray highlights), four of them being intra-domain in V_L, V_H, C_L and C_H1, and the fifth is at the C-terminus between the light and heavy chains.

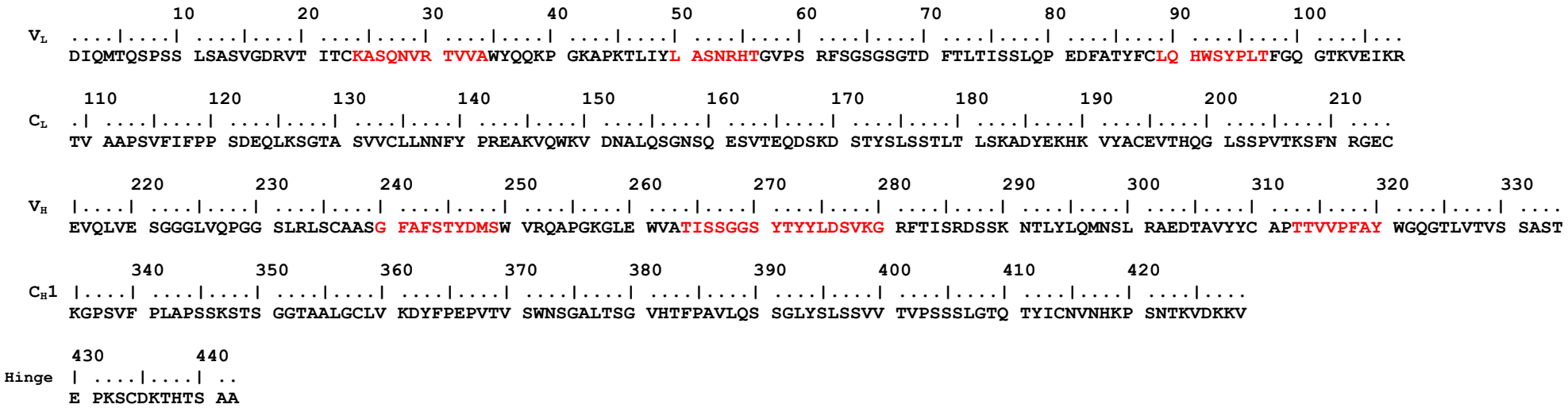

Figure 1.8. Fab A33 sequence. Fab A33 amino acid sequence separated by domains (V_L, C_L, V_H, C_H1 and hinge region). The six CDRs in the V_L and V_H domains are highlighted in red.

1.1.8 Market of antibody-based products

Mainly due to their high specificity and affinity for their targets, full-size antibodies and antibody-based products have been used successfully in the last 30 years to treat many human diseases (Ecker et al. 2015). Antibody products became an attractive choice as therapeutic agents, compared to small molecules, because they are highly specific to their targets and have fewer adverse side-effects (small molecule therapeutics frequently have non-specific interactions) (Smith 2015; Leader et al. 2008). In addition, full-size antibodies have long half-lives in serum due to FcRn recycling (Wang et al. 2008). In terms of modes of action, antibodies bind antigens on the surface or exterior of cells. After binding, the mode of action can include just blocking the function of the antigen by preventing its action. This is the mode of action of antibody fragments, which for example can bind cell surface receptors preventing dimerization and uncontrolled proliferation. Full-size antibodies, upon binding, can recruit the host immune system through the antibody effector functions, to kill the target cells. Alternatively, the antibody-based products can be conjugated with small molecule toxins or radiolabelled isotopes (Chames et al. 2009). This concept was already introduced by Paul Ehrlich, who named them as “magic bullets”, by having antibody-products deliver a cytotoxic payload to the diseased cell specifically on target binding (Strebhardt & Ullrich 2008).

When used for therapeutic purposes, we use monoclonal antibodies (mAbs), which consist of identical antibody molecules (with the same amino acid sequence). The first mAb to be approved by the FDA was in 1986, a murine mAb for the treatment of kidney transplant rejection. Since then, antibody-based products have grown steadily to become the main drug class for new approvals in the pharmaceutical industry. To date, over 60 antibody-based drugs have been approved for therapeutic use (Carter & Lazar 2018). There are over 550 antibodies in clinical development, including more than 50 antibodies in phase III clinical trials. The market for antibody-based products had worldwide revenues of nearly $89 billion in 2016. It is estimated that by 2020, there will be more than 70 antibody products in the market, with world-wide sales of nearly $125 billion (Elvin et al. 2013; Nelson et al. 2010).

Antibody-based products are approved for the treatment of a variety of diseases. At least 30 antibody drugs are indicated for use in oncology, including for the treatment of many prevalent solid and haematological tumours (Weiner et al. 2010; Nelson 2010).

An approximately similar number of antibodies are approved for the treatment of chronic inflammatory or autoimmune diseases (Chan & Carter 2010). A few antibody drugs are being used to treat patients in other areas of medicine including cardiovascular disorders, infectious and ophthalmic diseases, osteoporosis, as well as transplantation. Additionally, antibody-based products are also being used as diagnostic tools, such as for imaging, guide surgeries and detection of cancers (Weiner 2015).

Of the approved antibody-based drugs, IgG is the main molecular format. The remaining approved antibody-based drugs include antibody fragments, bispecific antibodies and antibody conjugates (including those conjugated to cytotoxic drugs, radioisotopes or polyethylene glycol (PEG)). Interestingly, the proportion of antibodies with non-IgG formats is higher for antibodies in early clinical development (Beck et al. 2010). Currently, there are six Fab fragments approved by the FDA for therapeutic applications. ReoPro (abciximab) is a chimeric IgG1 Fab fragment, approved in 1994, which binds to the glycoprotein II3/IIIa receptor on human platelets and inhibits platelet aggregation, for the treatment of blood clot prevention. Lucentis (ranibizumab) is a humanized IgG1 Fab fragment, approved in 2006, which inhibits vascular endothelial growth factor A (VEGF-A), for the treatment of age-related macular degeneration. CroFab (crotalide) is a polyvalent immune Fab (standardized mixture of four different monospecific Fab fragments), approved in 2000 for the treatment of envenomation by four species of North American pit vipers. DigiFab is an anti-digoxin Fab fragment, approved in 2001, for the treatment of digoxin intoxication. Digibind, another digoxin immune Fab, was also approved as digoxin antidote. Lastly, Cimzia (Certolizumab pegol), is a humanized IgG Fab, approved in 2009, which targets TNFα for the treatment of Crohn's disease. There are also examples of Fab fragments being used for diagnostic applications. MyoScint (Imiciromab), is a murine Fab, approved in 1996, which binds human cardiac myosin for imaging myocardial infarction. CEA-scan (Arcitumomab), is a murine Fab, approved in 1996, which binds human CEA for the detection of colorectal cancer.

1.2 Protein Aggregation

1.2.1 Why is it important to study protein aggregation?

Protein therapeutics offer many advantages over small molecule drugs, such as high target specificity and affinity, which has the advantages of high activity at lower concentrations and fewer adverse side effects. There are challenges, however, for proteins to be approved as commercial products due to their low thermodynamic stability (Wang 1999). The thermodynamic stability of the native protein conformation is only about 5-20 kcal/mol in free energy more stable than unfolded, biologically inactive conformations (Chi et al. 2003). Because of this, small changes to the system experienced during the manufacturing process, such as an increase in temperature, a change in pH, a change in salt concentration, shear force through shaking and stirring, or freezing and/or thawing, makes them susceptible to degradation, both chemical and physical (Manning et al. 2010). Chemical degradation involves modifications to covalent bonds, such as deamidation, oxidation, and disulphide bond shuffling (Daugherty & Mrsny 2006). Physical degradation includes protein unfolding, undesirable adsorption to surfaces and aggregation (Wang et al. 2010). The most commonly encountered and troubling manifestation of protein instability is protein aggregation, since it is observed at all stages of product development and aggregates are thought to be the dominant cause of immunogenicity (Wang et al. 2012). In the initial phase of protein production, aggregates are often observed when large amounts of recombinant protein are expressed, and receive the name “inclusion bodies”. These are aggregates of misfolded protein, which represent significant yield losses and the need to be solubilized and refolded to obtain functional, soluble protein. Aggregation is also encountered during the later stages of the manufacturing process, including purification, sterilization, shipping, and storage. The explanation being that relatively small changes of external variables can destabilize the structure of the protein and induce its unfolding, favouring consequent aggregation (Jahn & Radford 2008).

The presence of aggregates causes two major problems for pharmaceuticals, (i) it lowers the potency and efficacy of the therapeutic dose, and (ii) more problematically, trace amounts of aggregates can be hazardous to patients as they may cause severe inflammation or even fatal immune responses (Hermeling et al. 2004). The

immunogenicity of protein aggregates may arise from the formation of new epitopes, either from new quaternary structures in the aggregates, from newly exposed regions due to unfolding (previously buried inside the native protein), or from the formation of repetitive complexes to which the immune system is especially sensitive. Aggregation levels as low as 1% (often not visible to the naked eye) over a 2-year shelf life can render a product clinically unacceptable (Frokjaer & Otzen 2005). For these reasons, protein aggregation represents an unsolved and crucial challenge for the biotechnology industry to address. Currently, our molecular knowledge of the mechanism of protein aggregation is still limited. In order to improve the stability of a therapeutic molecule, screens over many formulation mixtures are performed by varying pH, buffer type, and ionic strength, as well as the addition of excipients, such as sucrose, sorbitol, arginine and mannitol (Parkins & Lashmar 2000). This approach is both expensive and time consuming. A better understanding of the molecular mechanisms of protein aggregation could lead to more efficient and reliable methods to prevent aggregation. For example, proteins could be engineered to be more robust to aggregation or the search for formulation excipients can be performed more rationally. By preventing aggregation, we hope that more therapeutic candidates will be able to reach the market.

Protein misfolding and aggregation also play a central role in many diseases such as Alzheimer's, Parkinson's, Huntington's, amyotrophic lateral sclerosis (ALS) and prion diseases (Chiti & Dobson 2017; Ross & Poirier 2004). In these diseases, a specific peptide or protein that is normally soluble, is deposited as insoluble aggregates (Knowles et al. 2014). For example, in the case of Alzheimer's disease two proteins are involved amyloid-β and tau protein, for Parkinson's is α-synuclein, and for spongiform encephalopathies are prion proteins. The aggregates formed usually consist of the specific misfolded protein in an ordered arrangement of β-sheets, forming fibres. This structure is known as cross-β structure or amyloid (Nelson et al. 2005). Thus, these diseases receive the name amyloidosis, because the common characteristic is the presence of these fibrillary deposits. Even though much has been discovered about these diseases in the past 50 years, the causes (how these protein misfolding diseases initiate) and the mechanisms of how fibrillary aggregates form and how they cause harm in the cell, are still largely unclear. Thus, a better understanding of protein aggregation will help design drug candidates to reverse or inhibit disease.

1.2.2 Mechanisms of protein aggregation

The journal nature describes protein aggregation as "the process by which misfolded proteins adopt a conformation that cause its polymerization into aggregates and organized fibrils". It is becoming clear that aggregation involves at least two steps, conformational changes to the protein native state and assembly of protein molecules into higher order aggregates (Chi et al. 2003; Roberts 2014). In the first step, the native state of the protein undergoes a conformational change to form an aggregation-prone state (Calamai et al. 2005). This intermediate or native-like state is believed to expose aggregation-prone regions, which are normally protected in the native protein, not able to initiate polymerization (Pawar et al. 2005; Khurana et al. 2001). Upon protein unfolding, the aggregation-prone regions increase its solvent exposure, such that in the second and subsequent steps the intermediate is driven by the hydrophobic effect or the propensity of exposed sequences to form cross-β sheets, to associate with other molecules. Different numbers of monomers may associate to form oligomers of various sizes, with the monomer conformations remodelled within oligomers in different ways. Oligomers may in turn grow into larger aggregates, whether amorphous or highly-structured as in the case of amyloid fibrils (Chiti & Dobson 2006; Stefani & Dobson 2003).

The conformational stability of the native protein plays a crucial role in aggregation. The first step, conformational changes to the protein native state, is controlled by the conformational stability of the native protein relative to aggregation-prone states. Energetically, this step is controlled by the free energy of unfolding, ΔG_{unf}, which is the thermodynamic stability of the native protein conformation relative to the thermodynamic stability of the aggregation transition state, though this is often probed indirectly by the free energy of unfolding of the native protein relative to the fully unfolded state. As examples to help illustrate the point, most of the mutations associated with hereditary forms of protein misfolding diseases have been shown to decrease the conformational stability of the globular native fold (ΔG_{unf}) and thus, promote aggregation (Chiti et al. 2003). Addition of chaotropes, such as urea or guanidinium chloride (GdmCl), also destabilize the conformational stability of the native protein, increasing the aggregation rate. On the contrary, by stabilizing the native state, which can be done by protein engineering with stabilizing mutations or the addition of stabilizers such as sucrose to the solution, increases ΔG_{unf} and stabilizes the proteins against aggregation.

The second step, assembly into aggregates, is controlled by the half life or relative population of aggregation-prone states, their ability to form specific intermolecular interactions, and their colloidal stability in terms of intermolecular attractive and repulsive forces, the later reflected in the values of the osmotic second virial coefficient (B_{22}). Assembly processes occur because of attractive intermolecular interactions (Roberts et al. 2014; Arzenšek et al. 2012). The osmotic second virial coefficient (B_{22}) is a thermodynamic solution parameter that directly quantifies overall protein-protein interactions on the molecular level, which include hard-sphere, electrostatic, van der Waals' forces, and all other short-range interactions (Neal et al. 1999). Positive B_{22} values indicate repulsive forces between protein molecules (protein-solvent interactions are favoured over protein-protein interactions). Negative B_{22} values reflect overall attractive forces (protein-protein interactions being favoured over protein-solvent interactions). Thus, another way to prevent aggregation is by using solution conditions that increase B_{22}, such as pH and ionic strength. As an example, when proteins are highly charged, repulsive interactions between proteins stabilize protein solution colloidally, making assembly processes such as aggregation energetically unfavourable (Chakroun et al. 2016). However, the protein might not be active under these conditions.

Two major contributions to interactions between protein molecules in aqueous solutions are Coulombic electrostatic interactions and van der Waals' interactions. The total energy of the interaction is the sum of the repulsive electric double-layer and the attractive van der Waals' forces, modulated by electrolyte concentration (Chi et al. 2003; Roberts et al. 2014). When two protein molecules with the same charge approach each other, there is an energy barrier they need to overcome to come together. At distances shorter than the energy barrier, the molecules experience attractive forces, resulting in their aggregation. If the energy barrier is high, the molecules remain kinetically stable. If the energy barrier is small or negative, particles are colloidally unstable and they come together. This can be achieved by adding high salt concentration, which screens the double-layer repulsion between protein molecules. This can also be achieved when the pH of the solution is near the isoelectric point (pI) of the protein.

Protein aggregation is thus controlled by both, conformational stability and colloidal stability, and depending on the solution conditions either could be rate limiting.

1.2.3 Protein stability

The stability of a protein is defined as the tendency to maintain its native structure (Dobson 2003). For globular proteins in physiological conditions, the thermodynamic stability of the native protein conformation is only about 5-20 kcal/mol in free energy more stable than unfolded, inactive conformations. Thus, although the native structure is energetically favorable, proteins are only marginally stable structures. This small net conformational stability arises from a balance between large stabilizing forces and large destabilizing forces. The driving energy for protein folding is the energetic cost of allowing amino acids with hydrophobic side chains to be exposed to the solvent. As non-polar molecules cannot participate in hydrogen bonding or ionic interactions, water molecules form organized, and thus energetically costly structures around these amino acid side chains to minimize contact. Burying the hydrophobic residues inside the protein core to limit solvent exposure, provides a large gain in configurational entropy of water molecules, which don't have to form these highly organized structures. On average, for protein folding, 85% of non-polar side chains are buried (Lesser & Rose 1990). The other forces contributing to the free energy of folding are hydrophobic interactions, hydrogen bonding, van der Waals' forces, electrostatic forces, disulfide bonds and intrinsic propensities (local peptide interactions). The main force opposing protein folding is the protein's conformational entropy, both local entropy (translational, rotational, and vibrational degrees of freedom on the molecular scale) and non-local entropy (excluded volume and chain configurational freedom), which are reduced in the folded state (Kumar et al. 2011).

The low thermodynamic stability of the native state of proteins can be understood in the context that proteins need to be able to move to fulfil their functions. Protein dynamics range from small atomic fluctuations around an average structure to large-scale reorganizations and conformational changes, which occur on time-scales of femtoseconds to seconds. For instance, proteins such as enzymes need to catalyze reactions, receptors need to bind their ligands and transmit information, or membrane protein channels that only allow certain molecules to go through, cannot be rigid structures, rather they need to be flexible and dynamic entities. The protein native conformation is flexible, and there exists an ensemble of native states.

1.2.4 How solution conditions affect protein stability

Protein aggregation has been found to depend strongly on the protein's solution environment, such as temperature, pH, salt type and concentration, (cosolutes, preservatives, and surfactants), as well as the relative thermodynamic stability of the protein native state.

Solution pH has a strong influence on the aggregation rate. pH affects electrostatic interactions, and the solution pH determines the total charge of the protein. Proteins are often only stable over narrow pH ranges, and outside these ranges aggregation is accelerated. pH affects both, the conformation of the protein and the electrostatic interactions between protein molecules (Chi et al. 2003). First, we consider the effect on the protein conformation. At pH near the isoelectric point (pI) of the protein, the net charge of the protein is almost zero, and proteins possess both positively and negatively charged groups. There are specific interactions, such as salt bridges, that generally stabilize the native state of the protein. At pH far from the pI, the number of charged groups in the protein is increased, resulting in increased charge repulsion that destabilizes the folded conformation, because the charge density on the folded protein is greater than on the unfolded protein (pH-induced unfolding). When the effect of pH on interactions between protein molecules is considered, these effects are reversed. At pH near the pI, anisotropic charge distribution on the protein surface could give rise to dipoles, potentially leading to aggregation. At pH far from the pI, proteins are highly charged, and repulsive interactions between proteins stabilize the protein colloidally, making assembly processes such as aggregation energetically unfavourable (Olsen et al. 2009). However, in this last case, even though protein aggregation is not favoured, the folded native state of the protein was lost. Thus, final formulations of therapeutic proteins are generally at pHs close to the pI of the protein.

The thermodynamic stability of the native protein conformation, characterized by the free energy of unfolding (ΔG_{unf}), typically shows a parabolic profile as a function of temperature. Protein unfolding can occur at both, high and low temperatures (Mahler et al. 2009). The process of cold denaturation is complex and not well understood. High temperatures provide enough energy to the system to promote aggregation. High temperature affects both, the conformation of the protein and the reaction kinetics (Chi et al. 2003). Increasing the temperature increases the thermal kinetic energy of the reactants,

resulting in increased diffusion, increased frequency of collisions, and collisions with enough energy to overcome activation energies. High temperatures also perturb the native protein conformation (Milardi et al. 1994). Initial studies at high temperatures suggested that aggregation takes place from the fully unfolded state of the protein, however, more recent evidence suggest that unfolding can also occur from partially unfolded conformations of the protein. This hypothesis is supported by recent work on Fab A33, where it was found that the melting temperatures (T_m) of the protein under different conditions were only correlated with aggregation kinetics that were determined at temperatures elevated to just below the T_m of the protein, where aggregation from the unfolded state therefore predominated. By contrast, T_m did not correlate with the aggregation kinetics determined at lower storage temperatures, indicating that global unfolding was no longer the cause of aggregation (Chakroun et al. 2016). Additionally, it is usually observed during heating that aggregation starts at temperatures below the equilibrium melting temperature of the protein, where complete unfolding has not yet occurred. For multi-domain proteins, such as antibodies and Fab fragments, thermal denaturation can be domain specific, as different domains have different stabilities (Vermeer & Norde 2000).

Salt type and salt concentration also influence the stability of the protein formulation. Ions modulate the strength of electrostatic interactions between the charged groups. They affect both, conformational stability of the protein by affecting intra-molecular charge-charge interactions, and assembly processes by affecting inter-molecular charge-charge interactions. At low concentrations, the main effect of ions in solution results from charge shielding, which reduces electrostatic interactions, rather than destabilizing the native protein conformation. At high concentrations of certain salts, in addition to the charge-shielding effects, preferential binding of ions can occur, which can destabilize the native state (Arakawa & Timasheff 1982). Salts can be characterized depending on their interactions with water, as kosmotropes (oder-making) and chaotropes (disorder-makers). Kosmotropes tend to be small and have high charge density, which causes water molecules to favourably interact, which also stabilizes proteins. In contrast, chaotropes are larger and poorly hydrated, and have the opposite effect, they disrupt the structure of water, which may cause the denaturation of proteins (Curtis et al. 1998). The effect of different types of salts on the solubility of proteins was studied by Franz Hofmeister, who discovered that certain cations and anions had consistent effects, generating the Hofmeister series (Baldwin 1996; Collins 2004). The series classifies ions

in order of their ability to salt out or salt in proteins. The mechanisms are not entirely clear; however, recent simulation studies seem to indicate that the water molecules directly contacting the proteins may play a crucial role.

Binding ligands, excipients, preservatives or surfactants are also often added to the final formulation of therapeutic proteins, to guarantee its stability. There are certain solutes (sugars, polyols, certain salts such as ammonium sulphate) that stabilize the native protein conformation, whereas there are other solutes (urea and guanidinium chloride (GdmCl)) that favour its unfolding. A proposed explanation for these different effects of solutes on protein conformation, is the differential binding of these solutes towards the folded or unfolded states of the protein. Denaturants such as urea and GdmCl exhibit greater binding to the denatured state of the protein than to the native state, favouring unfolding and aggregation. In contrast, stabilizers such as sucrose and glycerol, are preferentially excluded from the surface of a protein molecule. Preferential exclusion can be interpreted as negative binding. During unfolding, protein surface area increases, leading to more preferential exclusion. The net effect of greater negative binding to the unfolded state is to favour the native state.

1.2.5 Characterization of aggregation-prone conformations

Many proteins aggregate with first-order kinetics, implying a unimolecular rate-limiting step linked to conformational changes or partial unfolding, rather than a rate-limiting bimolecular association of two protein molecules (Chi et al. 2003). It is therefore important to characterize the nature of any conformational changes in the native state that can promote aggregation. These states are believed to expose aggregation-prone regions in the protein, which causes the protein to aggregate. For many years, experiments have tried to characterize the states that precede aggregation. Initial studies suggested that aggregation takes place from the fully-unfolded state of the protein, drawn from early observations of proteins at elevated temperatures. However, increasing evidence suggests that, at temperatures below the melting temperature (T_m) of the protein, aggregation takes place from near-native states, where only partial or transient local-unfolding of the protein occurs (Chakroun et al. 2016).

Many studies have reported the presence of near-native states of proteins that are aggregation-prone (Bemporad & Chiti 2009; Bemporad et al. 2012; Zhuravlev et al. 2014; Uversky et al. 2001; Khurana et al. 2001). Back in 1998, Kendrick et al. studied the aggregation of recombinant human interferon-γ (rhIFN-γ) and elucidated that aggregation proceeds through a transiently expanded conformational species within the native state ensemble (Kendrick et al. 1998). First, they observed experimentally that rhIFN-γ follows a first-order aggregation kinetics and that addition of sucrose stabilizes the protein against aggregation. By combining kinetic analysis with solution thermodynamics, they inferred that only a small (9%) expansion of the native state surface area is needed to form the intermediate state that precedes aggregation. This conformational expansion is only about 30% of that required for the complete unfolding of rhIFN-γ. Additionally, they suggested that sucrose stabilizes rhIFN-γ against aggregation by shifting the equilibrium within the ensemble of rhIFN-γ native conformations to favor the most compact native species over the transiently expanded native species. Similar results were found for human granulocyte colony stimulating factor (rhGCSF), in which the expanded intermediate state preceding aggregation represented only 15% of the change in surface area observed for the completely unfolded conformation (Webb et al. 2001). Furthermore, they did hydrogen-deuterium exchange experiments, which very often show that amide hydrogens buried in the interior of a native protein can exchange with the solvent, result of internal hydrogen bonds breaking and exposing backbone amides to solvent. Interestingly, they observed that the addition of sucrose reduced the rate of H-D exchange for rhGCSF. More recently, only transient local unfolding was found necessary to show faster aggregation for variants of human lysozyme, using hydrogen-deuterium exchange experiments (Canet et al. 2002). Studies on hyperthermophilic acylphosphatase, superoxide dismutase 1, transthyretin, 2-microglobulin and Fyn SH3 also showed that global unfolding was not necessary, and that aggregation could be initiated from locally unfolded states (Chiti & Dobson 2009). NMR was able to resolve a structural folding intermediate of the 6.4 kDa Fyn SH3 domain that was more aggregation-prone than the native state (Neudecker et al. 2012). However, this relied upon mutations that stabilized the folding intermediate, and so the use of NMR to characterize directly pre-aggregational states in unmutated native-ensembles remains very challenging, particularly for larger proteins such as the 48 kDa humanized antibody fragment Fab A33.

1.2.6 Aggregation process

Up until now, we have considered the mechanism and driving forces in the initial stages of aggregation. To form protein aggregates, monomer proteins need to come together and form high molecular weight assemblies. Protein aggregates, either amorphous or fibrillar, are characterized by a high content in β-sheet structures. Aggregates form regardless the secondary structure of the native protein conformation, and the stress experienced (such high temperature, low pH, addition of denaturants, stress to the system) (Hartl & Hayer-Hartl 2009). If aggregation is initiated from near-native aggregation-prone states, the early oligomers formed would retain native-like structure. If aggregation is initiated from fully unfolded or largely unfolded conformations, the initial oligomers formed will be disordered structures. As aggregation proceeds, both types of oligomers experience internal reorganizations to form the β-sheet structure present in the late aggregates (Orte et al. 2008). The stability of these species increases with time, to ultimately form stable fibrils. Amyloid fibrils are thread-like protein aggregates with a core region formed by repetitive arrays of β-sheets oriented perpendicularly to the fibril axis forming a structure known as cross-β (Nelson et al. 2005).

Protein aggregation has been characterised as a nucleation and growth mechanism (Invernizzi et al. 2012; Frieden 2007; Bemporad & Chiti 2012; Tanaka & Komi 2015). Two stages can be differentiated, an initial lag phase, where nucleation takes place, and an exponential phase, where accumulation and fibril growth occurs. During the initial lag phase, the soluble species (usually monomers) associate to form a nuclei, a poorly characterized state which formation influences the overall kinetics of the amyloid reaction. There is an energy barrier for nucleation (thermodynamically disfavoured phase), and once this energy is overcomed, growth of the nucleus occurs. It is known that seeding with pre-formed aggregates, reduces the lag time and promotes aggregation. The oligomers formed at this stage are a highly heterogeneous ensemble of species (Yang et al. 2018). During this phase, the solution remains clear. The lag phase comprises the conversion steps to the first species to display β-sheet conformation. Growth of these oligomers can happen through addition of monomers or polymeric association. The growth phase consists of multiple stages by which soluble species are progressively arranged at the ends of preformed β-sheet rich structures in a thermodynamically favourable process, until eventually the solubility limit is reached, and they precipitate

out of solution. Preformed fibrils can catalyse the formation of new fibrils by fragmentation, branching and/or nucleation on the fibril surface. During the manufacture and storage of therapeutic proteins, soluble aggregates can form. These are not visible to the eye, but if they form a nucleus, it can foster rapid assembly into large aggregates, regarding the product not usable.

Chapter Two

Materials and Methods

2.1 Methods for protein structure determination

Protein structures can be determined using methods including X-ray crystallography, nuclear magnetic resonance (NMR), cryo-electron microscopy (cryo-EM) and small-angle scattering (SAS) techniques of X-rays (SAXS) or neutrons (SANS) combined with atomistic modelling. As of March 2017, the structures deposited in the Protein Data Bank (http://www.rcsb.org) (Westbrook et al. 2003), were determined 89% by crystallography, 9% by NMR, 1% cryo-EM, and <1% by other techniques (Gore et al. 2017). X-ray crystallography, NMR and cryo-EM provide atomic-resolution (0.1 - 0.3 nm) detail of the solved structures. X-ray crystallography provides structures of proteins of all sizes, however, formation of ordered crystals is necessary, to study their diffraction. Certain proteins have proven difficult to crystallize, such as flexible proteins, proteins with large surface carbohydrates and membrane proteins (Nogales & Scheres 2015). NMR allows the study of protein structure in solution, however, NMR encounters problems with macromolecules of high molecular mass (> 30-40 kDa) (Markwick et al. 2008; Frueh et al. 2013; Poppe et al. 2013). Cryo-EM allows the study of proteins in their native state by quickly freezing them hydrated, and allows the study of large, complex and flexible structures (Murata & Wolf 2018; Wang & Wang 2017). SAS techniques determine the solution structures of proteins, from sizes approximately above 15 kDa, but they are low resolution techniques (2 - 4 nm). However, by combining the scattering data with atomistic models, a precision of 0.5 - 1.0 nm can be obtained (Perkins et al. 2008). In this thesis, I used small-angle X-ray scattering (SAXS) combined with atomistic modelling and single-molecule FRET, to study the structure and conformational changes of Fab A33 in different solution conditions, and this will be reviewed in more detail in this chapter.

2.1.1 Small-angle X-ray scattering (SAXS)

SAXS is a diffraction technique that characterizes the structure of proteins in solution. SAXS is normally used when proteins cannot be crystallized or where solution conditions affect the protein structure (Perkins et al. 2008). In this thesis, SAXS allows to study differences in the structure of Fab A33 under different solution conditions, of pH and ionic strength.

2.1.1.1 SAXS experiment - data acquisition

In short, in a SAXS experiment, a focused and monochromatic beam of X-rays (wavelength of 1 Å) traverses a sample of the target protein in solution, and the X-ray photons are scattered by the electrons of the atoms of the protein, which are then detected by a 2D X-ray detector. X-ray photons are scattered by the electrons within the sample. X-ray photons cause electrons to oscillate at the same frequency as the incident wave. These oscillating charges become a dipole, which give rise to a spherical wave being emitted of the same energy and wavelength (elastic scattering). The resulting scattering pattern contains information about the size and shape of the protein (Mertens & Svergun 2010).

In X-ray crystallography, the beam of radiation is diffracted by the electrons in the crystal following Bragg's Law ($\lambda = 2d\sin(\theta)$; where λ is the wavelength, d is the spacing between the planes in the atomic lattice, and 2θ is the angle between the incident ray and the diffraction planes). In SAXS, the protein molecules are randomly orientated in solution. Thus, the diffraction gives a radially-symmetric pattern, and in this case, is described as scattering. Bragg's Law can be adapted to describe scattering, in terms of Q, the scattering vector, which is the difference between the scattered vector (k_s) and the incident vector (k_i), k_s - k_i ($|k_s| = |k_i| = 2\pi/\lambda$). The magnitude of the scattering vector is measured as:

$$Q = 4\pi\sin(\theta)/\lambda \qquad \text{(Eq. 2.1)}$$

where 2θ is the scattering angle and λ is the wavelength. The units of Q are nm^{-1} (inverse of the wavelength) (Perkins et al. 2008). Scattering experiments depend on elastic coherent scattering. Elastic scattering implies that the scattered beam has the same energy as the incident beam, but the direction of propagation has changed. Coherent scattering means that scattered waves interfere to give a single wave in a given direction, for example from two point scatters (Figure 2.1a), which contains information about the structure of the protein. The scattering signal is comprised of the sum of the contributions made by all pairs of scatterers to the scattering pattern. The scattered X-rays are then detected by a 2D flat X-ray detector situated behind the sample.

In a SAXS experiment, the intensities, I(Q), of the scattering curve are measured as a function of the scattering vector Q (Figure 2.1b). The radial average of the scattering pattern about the position of the direct main beam, gives a 1D SAXS profiles via integration, corresponding to the scattering curve I(Q) in reciprocal space. I(Q) can be interpreted as the Fourier transform (from real space to the reciprocal space) of the distance distribution of the point scatterers. The scattering curve of the protein alone is obtained by subtracting the low-scattering curve of the buffer to the high-scattering curve of the protein in the same buffer. If the sample is composed of monodisperse identical molecules, the scattering intensity is the average of a single molecule scattering over all orientations.

In a typical SAXS scattering curve, I(Q) drops off quickly as Q values increase (Svergun & Koch 2002). Smaller scattering angles and smaller Q, correspond to lower resolution, where large distances between scatterers are detected allowing to see the shape of the protein. Higher scattering angles and higher Q, provide higher resolution structural information, corresponding to shorter distances between scatterers. Because X-rays interact with electrons, the scattering intensity is proportional to atomic number. At a scattering angle of zero, the intensity of scattering I(0) is proportional to the molecular mass of the molecule. If the concentration of the protein is high, this leads to interference between X-rays scattered from different molecules, known as inter-particle interference, visible at low angles ($Q < 1$ nm^{-1}).

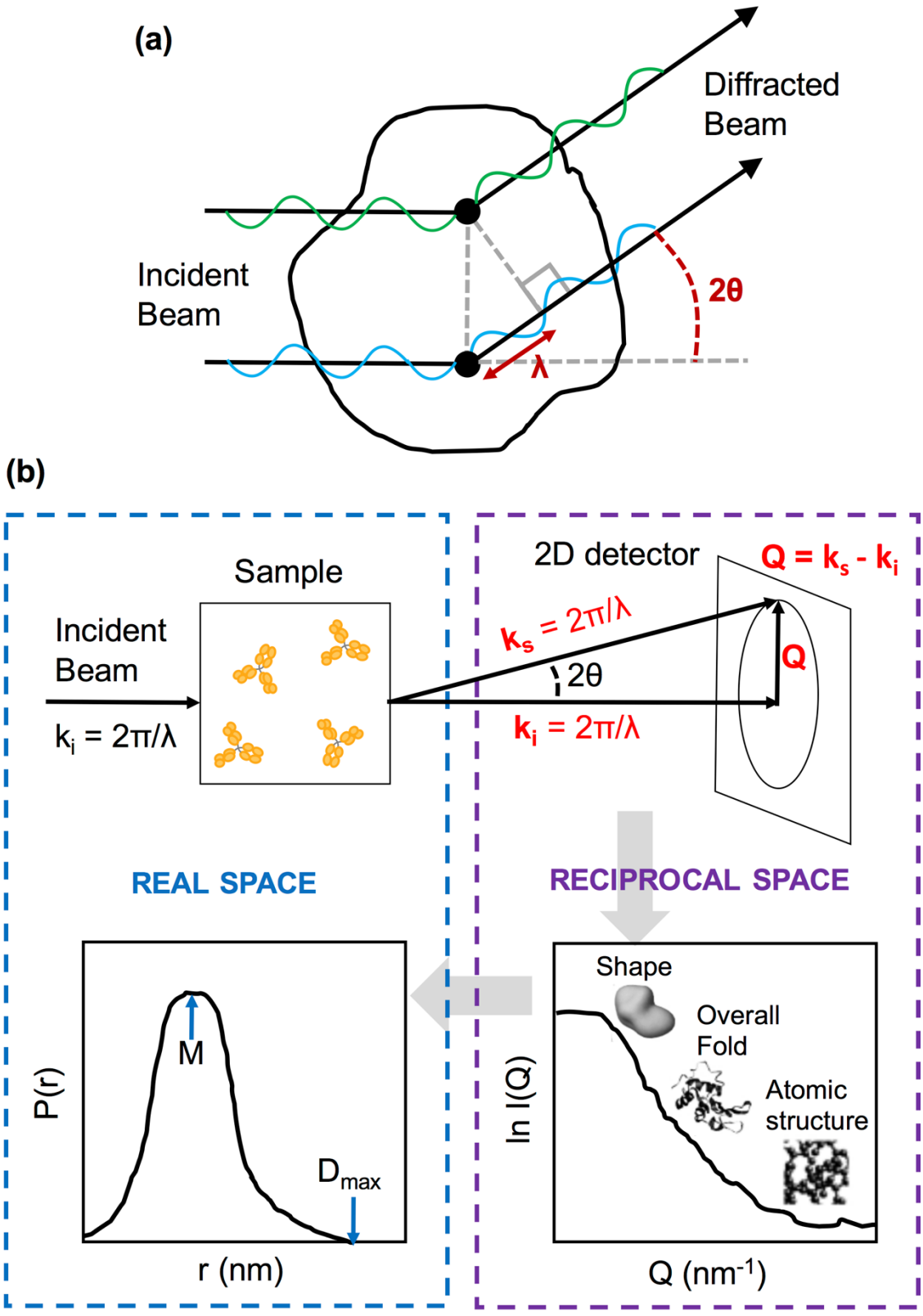
(a)
Diffracted
Beam
Incident
Beam
2θ
λ
(b)
Sample
Incident
Beam
$k_i = 2\pi/\lambda$
2D detector
$Q = k_s - k_i$
$k_s = 2\pi/\lambda$
2θ
Q
$k_i = 2\pi/\lambda$
REAL SPACE
RECIPROCAL SPACE
P(r)
M
D_{max}
r (nm)
ln I(Q)
Shape
Overall
Fold
Atomic
structure
Q (nm^{-1})

Figure 2.1. Schematic representations of a SAXS experiment. (a) An incident beam of monochromatic X-rays is scattered by two point scatterers (•) within a protein. The diffracted beams are in phase with each other but out of step by λ at the scattering angle 2θ, causing constructive interference. The accumulation of these events at low angles values gives rise to the scattering pattern of the molecule. (b) Generation of a characteristic 1D scattering curve, ln I(Q) as a function of Q. The scattering from multiple molecules in random orientations, results in a characteristic intensity distribution in reciprocal space, which for monodisperse solutions of non-interacting molecules is equivalent to a single molecule averaged over all orientations. The scattering curve can be converted back to real space using an inverted Fourier transformation. D_{max} and M correspond to the maximum dimension of the molecule and the most occurring distance between point scatterers.

The intensity at each Q can be calculated by the Debye equation, which describes the geometrical relationship between individual scatterers within the molecules, and takes into account the differential orientations of the particles via rotational averaging in space:

$$I(Q) = \sum_p \sum_q f_p f_q \frac{\sin(r_{pq}Q)}{r_{pq}Q} \quad \text{(Eq. 2.2)}$$

where f_p and f_q are the scattering lengths of the electrons at points p and q within the biomolecule, separated by a distance, r_{pq} (Yang 2014).

SAXS measurements are measured very close to the primary beam, at small angles (typically 0.1 - 10º). This angular range contains information about the shape and size of macromolecules. The scattering of sample and buffer is very similar except for the lowest angles. For scattering experiments, intense beams of X-rays are necessary, because the probability of a diffraction event when an X-ray approaches an electron is very low (10^{-25}). Currently, there are many powerful X-ray synchrotrons, such as ESRF (Grenoble, France) and Diamond (Oxfordshire, UK).

2.1.1.2 SAXS data analysis

Prior to data analysis, the collected scattering curves need to be pre-processed (Boesecke 2007). This includes monitoring for radiation damage and only merging scans that have not been damaged, and buffer subtraction. Two types of data analyses are typically performed first, to obtain the radius of gyration (R_g, being a measure of macromolecular elongation) and the intensity at zero Q (I(0)), this later proportional to the molecular weight (M_w). These analyses are Guinier analysis of low Q values, and distance distribution function analysis P(r) from the Fourier transformation of the full I(Q) scattering curve.

X-ray scattering reveals the hydrated dimensions of the macromolecule. This means that SAXS observes the hydration layer, the layer of water molecules in direct contact with the protein. The electron density of this bound water is higher than that of bulk water, making this detectable by SAXS in the scattering of the protein. Previous work measured a mass ratio of approximately 0.3 gram of bound water per gram of protein (Perkins 1986). Consequently, the molecule appears larger by the presence of this hydration layer.

By taking into account larger Q, more information about the structure of the protein can be extracted. This will be done by combining atomistic models of Fab A33 generated using molecular dynamics simulations under the same experimental conditions, and comparing them to the experimental SAXS data. In this way, the models that best fit the data under each condition will be elucidated.

Guinier Analysis

The low Q region of the scattering curve contains information about the overall dimension of the molecule. Guinier realized that a Maclaurin series expansion could be used to describe the scattering curve. At low Q values, the higher order terms of that expression could be neglected, and the series could be approximated to the first two terms. By rearranging the expression, it takes the form:

$$\ln I(Q) = \ln I(0) - R_g{}^2 Q^2/3 \qquad \text{(Eq. 2.3)}$$

where plotting ln I(Q) as a function of Q^2, gives a straight line, with the slope being proportional to the radius of gyration (R_g) and from the extrapolated intercept at zero Q we can obtain I(0). This received the name Guinier approximation, and for it to hold true, it can only be used at low Q values, between $Q*R_g$ of approximately 0.5 and 1.5 (Perkins et al. 2008).

The linearity of the Guinier approximation informs about the monodispersity of the sample. Intermolecular interactions or aggregation are possible to detect in the lowest Q values, where the curve would curve upwards deviating from the straight line. From the I(0) values, information about relative molecular weights can be obtained. To confirm that no aggregation has taken place, monitoring I(0)/c (c is concentration) as a function of concentration or different solution conditions, should reveal a constant I(0)/c value, and not a dependence with concentration or solution condition (Nan et al. 2013).

Distance distribution function analysis P(r)

Using the full Q range of the scattering curve, an inverse Fourier transformation of the I(Q) scattering curve can be generated, to convert the curve from reciprocal space (nm^{-1}) to real space (nm). The resulting curve receives the name distance distribution function, P(r). P(r) corresponds to the distribution of all the distances r between all the volume elements within the macromolecule, equivalent to a histogram of distances between pairs of points within the molecule. The P(r) curve allows a more direct visualization of the solution structure, which provides information about its maximum dimensions (D_{max}, when P(r) becomes zero at large r) and shape. The maximum value of the curve corresponds to the most common distance occurring within the molecule. If a protein is spherical, the shape of the P(r) curve would be bell-shaped and its maximum would be approximately at $D_{max}/2$, whereas if a protein contains several domains, the P(r) curve would display several shoulders that correspond to intra- and inter- subunit distances (Mertens & Svergun 2017). Additionally, the R_g and I(0) can also be measured from the P(r) curve, providing an alternative confirmation to these values, which should agree with the values calculated with the Guinier analysis.

$$P(r) = \frac{1}{2\pi^2}\int_0^\infty I(Q)Qrsin(Qr)dQ \quad \text{(Eq. 2.4)}$$

The program used to calculate the P(r) function is GNOM software (Franke et al. 2012). P(r) function involves integrals with infinite upper limits. In practice, the experimental scattering curve does not contain infinite values and data points close to the beam stop at low Q are missing. Thus, these integrals are constrained at the molecules maximum diameter, D_{max}.

Atomistic modelling of SAXS data

The atomistic modelling of full SAXS curve to large Q values enables additional information about the solution structure of proteins to be obtained beyond the low resolution R_g and I(0) analyses. One option is *ab initio* modelling, where bead models are created to fit the experimental scattering curve, with resolutions of 2 - 4 nm (Svergun & Koch 2002). The combination of experimental scattering curves with atomistic models, allow resolutions of 0.5 - 1.0 nm to be obtained, and elucidation of the best models that fit the experimental data (Perkins & Bonner 2008; Perkins et al. 2009; Perkins et al. 2016).

In this work, the software SCT was used for the atomistic modelling of the SAXS data (Wright & Perkins 2015). In brief, a large number of atomistic models of the target protein need to be generated first spanning different conformations. These atomistic models are transformed into a lower-resolution model (coarse grained), and a theoretical X-ray scattering curve is calculated for each of the models using the Debye equation adapted to spheres. To compare the theoretical scattering curves to the experimental curves, an R-factor is calculated, where models with the lowest R factors can be identified, representative of the average solution structure (Figure 2.2).

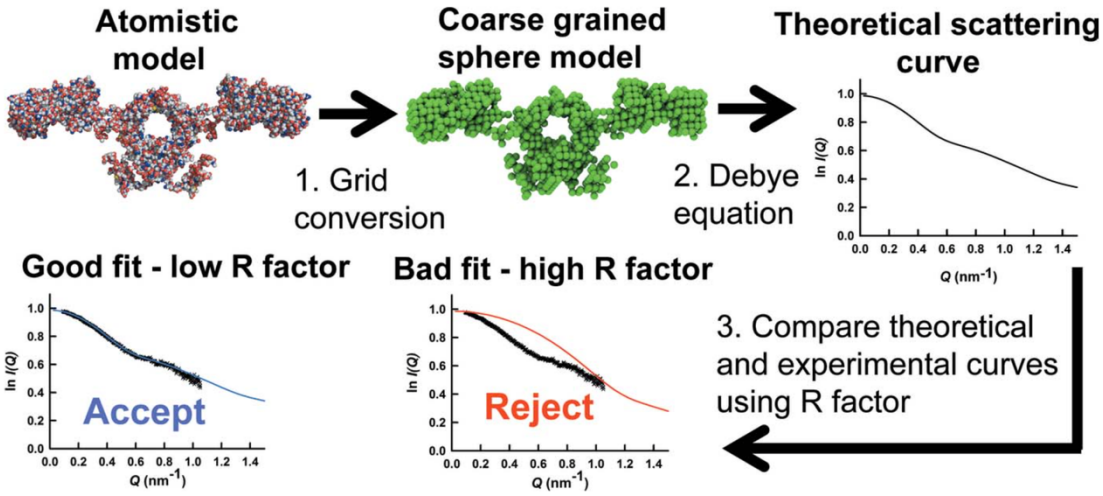

Figure 2.2. Schematic of the steps performed by SCT to identify the best atomistic models that fit the experimental SAXS curves. First, thousands of atomistic structures of the target protein spanning different conformations are generated. A grid transformation is performed on each structure to produce a lower-resolution (coarse-grained) sphere mode with an added hydration layer. A theoretical scattering curve is calculated for each sphere model using the Debye equation adapted to spheres. Theoretical curves are compared to experimental curves using the R factor. Low R factors represent good fits, and are inferred to represent the average solution structure. Figure obtained from (Wright & Perkins 2015).

A large library of atomistic structures in different conformations of the target protein need to be generated first. The starting structure can be either a crystal structure or a structure generated using homology modelling. Libraries of structurally varied models can then be generated using molecular dynamics simulations or Monte Carlo simulations. Whilst a theoretical scattering curve can be calculated from an atomistic structure, this remains a computationally expensive task. SCT reduces the computation demand by converting the initial atomistic model into a coarse-grained structure. This transformation is done by substituting several atoms in the atomistic structure by bigger spheres of diameter less than the resolution of the scattering experiment. A grid of equal divisions is created that contains all atoms within the input structure. If more atoms than a specified cutoff number (usually four) are found inside a grid box, a sphere with a radius of half the grid box, substitutes these atoms in the final sphere model. A layer of water

molecules in the surface of the protein needs to be added in X-ray scattering experiments, because the hydration layer is visible by SAXS. To generate hydrated sphere models, every sphere in the protein surface is surrounded by hydration spheres of the same radius. Excess hydration spheres are then removed to match the hydrated volume of the protein, calculated from its sequence.

Theoretical scattering curves from hydrated sphere models are calculated using the Debye equation adapted to spheres. This equation calculates all the distances from each sphere to the remaining spheres and sums the results. First, a histogram of the distances d between all spheres is constructed. Then, I(Q) curve as a function of Q is obtained from the Debye equation:

$$I_{Theor}(Q) = \frac{I(Q)}{I(0)} = g(Q)\left(n^{-1} + 2n^{-2}\sum_{j=1}^{m} A_j \frac{\sin Qd_j}{Qd_j}\right) \quad \text{(Eq. 2.5)}$$

where d_j is the distance between spheres represented by the jth histogram bin, A_j the number of distances that fall into bin j, m the number of bins in the histogram and n the number of spheres in the model. The squared form factor g(Q) is given by:

$$g(Q) = \frac{[3(\sin Qr - Qr\cos Qr)]^2}{Q^6 r^6} \quad \text{(Eq. 2.6)}$$

where r is the radius of the spheres in the model. If the diameter of the spheres is smaller than the resolution of the experiment, the squared form factor g(Q) becomes almost unchanged in the Q range of interest (Perkins et al. 2008).

Lastly, theoretical scattering curves are compared to the experimental curves in the same Q range, to identify the best fits. To quantify the goodness of the fit, SCT uses the parameter R factor, by analogy with crystallography, using the formula:

$$R = \frac{\Sigma\left\|\|I_{Expt}(Q)\| - \eta\|I_{Theor}(Q)\|\right\|}{\Sigma\|I_{Expt}(Q)\|} x\ 100 \quad \text{(Eq. 2.7)}$$

where η is a scaling factor used to match the theoretical curve to the experimental I(Q). The R factor is expressed as a percentage, with lower values representing better fits. Graphs of the R factor versus R_g values are of great utility in assessing the progression of

a modelling fit analysis. The R_g values of the models can be calculated from Guinier fits of the theoretical scattering curve in the same Q range that used experimentally. The models that better reproduce the experimental data are identified as representative of the average solution structure.

2.1.2 Single-molecule FRET (smFRET)

2.1.2.1 Background to single-molecule Förster Resonance Energy Transfer

In microscopy, the diffraction limit states that it is not possible to focus a laser beam to a spot smaller than about λ/2, which is 200 nm in the far blue (400/2) and 350 nm in the far red (700/2). Thus, in optical microscopy it is not possible to resolve objects that are closer than 200 nm. A typical protein, such as Fab A33 is 10 nm in size, which makes conformational changes not possible to be observed directly. However, they can be detected indirectly using FRET.

In the late 1940s, Theodor Förster proposed the phenomenon now known as Förster resonance energy transfer (FRET). FRET is a mechanism of energy transfer between two light-sensitive molecules (Schuler 2013). It involves a donor and an acceptor molecule. First, the donor fluorophore is excited to an excited electronic state, which can then transfer the energy to a nearby acceptor fluorophore through a non-radiative process (Figure 2.3a). The transfer is based on the concept of treating an excited fluorophore as an oscillating dipole that can undergo an energy exchange with a second dipole having a similar resonance frequency. The efficiency of this energy transfer is inversely proportional to the sixth power of the distance between donor and acceptor (Eq. 2.8), making FRET extremely sensitive to small changes in distance (Deniz et al. 1999). Consequently, FRET measurements can be utilized as an effective "molecular ruler" for determining distances between biomolecules labeled with an appropriate donor and acceptor fluorophore when they are within 2 and 10 nanometers of each other. For FRET to take place, the emission spectra of the donor need to overlap the absorption spectra of the acceptor (Figure 2.3b). The efficiency of the energy transfer is a measure of the fraction of photons absorbed by the donor that are transferred to the acceptor, and has a dependency with the distance donor-acceptor as shown in Eq. 2.8 (Figure 2.4). R_0 is the

Förster radius, the characteristic distance for a given pair of fluorophores that results in a transfer efficiency of 50% (Joo et al. 2008).

$$E(r) = \frac{R_0^6}{R_0^6 + r^6} \qquad \text{(Eq. 2.8)}$$

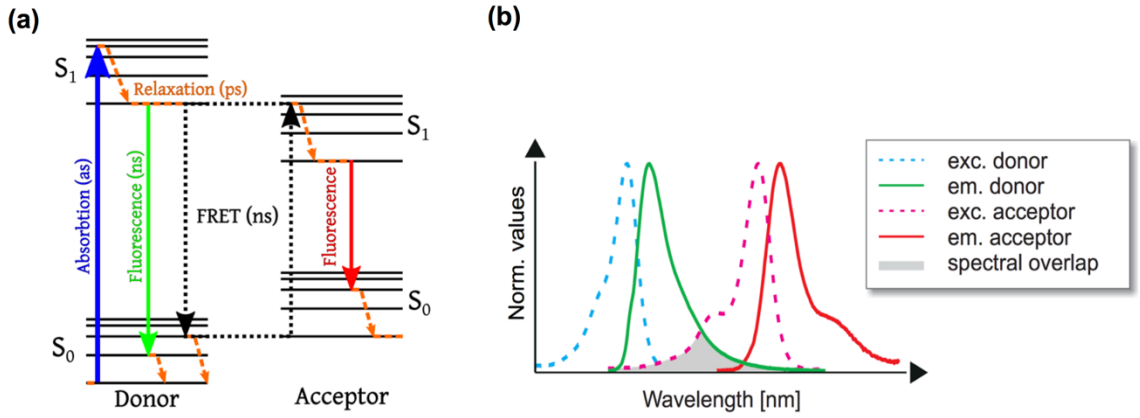

Figure 2.3. Schematic of donor and acceptor spectra's overlap for FRET to occur. (a) FRET visualized through a Jablonski diagram. Donor absorbs a blue photon and gets to the excited state. Donor can loose energy either by emitting a green photon through fluorescence, or by transferring energy to a nearby acceptor through the non-radiative process of FRET. This transfer of energy can only take place if the donor emission overlaps the acceptor excitation. (b) Absorption (dashed lines) and emission (continuous lines) spectres of donor and acceptor fluorophores. The spectral overlap between donor-emission and acceptor-excitation is shown in gray. Images obtained from (Schuler 2013).

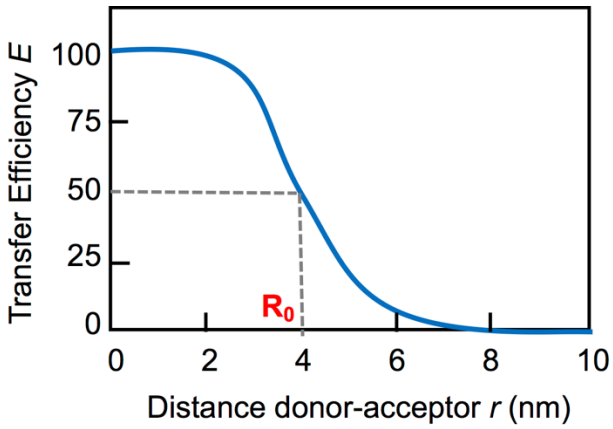

Figure 2.4. Example of the relation between the transfer efficiency and the distance donor-acceptor (*r*) for a given pair of fluorophores.

Experimentally, FRET transfer efficiencies can be determined in different ways (Roy et al. 2008). The more common in smFRET experiments are using fluorescent intensities and fluorescent lifetimes (average time that a fluorophore spends in the excited state before returning to the ground state by emitting a photon). Fluorescent intensities are calculated from the number of photons emitted from the donor and the acceptor fluorophores, n_D and n_A respectively, and the transfer efficiency is calculated as shown in Eq. 2.9. Equivalently, transfer efficiency can be calculated from the fluorescence lifetime of the donor in the presence (τ_{DA}) and absence (τ_D) of the acceptor, as shown in Eq. 2.10.

$$E = \frac{n_A}{n_A + n_D} \quad \text{(Eq. 2.9)}$$

$$E = 1 - \frac{\tau_{DA}}{\tau_D} \quad \text{(Eq. 2.10)}$$

In a typical smFRET experiment a protein is labelled with a donor and an acceptor fluorophore, such that the distance between them is less than 10 nm. In a solution of diffusing single-molecules, if a folded protein resides in the volume illuminated by the focused laser beam, excitation of the donor dye will result in rapid energy transfer to the acceptor dye because the dyes are in close proximity. Consequently, the majority of the fluorescence photons are emitted by the acceptor. Upon unfolding of the protein, the average distance between the donor and acceptor dyes will typically increase. As a result, the energy transfer rate is decreased, and the fraction of photons emitted by the acceptor is lower (Figure 2.5). The changes in fluorescence intensity from donor and acceptor can thus be used to distinguish between different conformational states of a protein (Borgia et al. 2011; Muller-Spath et al. 2010; Hofmann et al. 2010; Merchant et al. 2007).

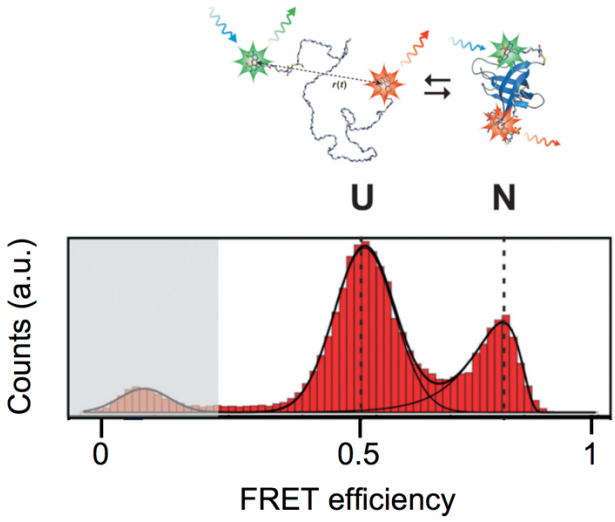

Figure 2.5. FRET efficiency histograms for the identification of different protein conformations. When the protein is folded (N), donor and acceptor fluorophores are near and the FRET energy transfer between them is high. Upon unfolding of the protein (U), the distance between fluorophores increases and less photons from the acceptor are detected, decreasing the FRET efficiency. At a FRET efficiency of zero, the population of molecules without an active acceptor fluorophore is shaded in gray. Image adapted from (Schuler 2013).

FRET combined with single-molecule detection, allows the detection of different conformations of the protein in solution. The challenge to detect a single molecule, is the presence of a huge excess of solvent molecules (~10^{22} water molecules in 1 ml) that contribute to the background, especially by scattering. Detection of a single molecule can be achieved by combining spatial selection and spectral selection. Spatial selection refers to reducing the observation volume as much as possible, as the background is proportional to the number of molecules illuminated. This is achieved by using confocal detection. In a confocal microscope, a pinhole is used to reject out-of-focus light which leads to a detection volume below 1 femtoliter. Spectral selection refers to selecting a detection method with high selectivity for the molecule of interest. Fluorescence allows the selection of a specific molecule by specific absorption and Stokes-shifted emission. For instance, a given fluorophore can be observed by its excitation at a specific wavelength and detection of its emitted light at longer wavelength (Schuler & Hofmann 2013).

2.1.2.2 Confocal single-molecule detection

Given the interest for single-molecule FRET experiments, commercial microscopes have been developed to perform this type of experiments, such as the MicroTime 200 microscope (PicoQuant, Germany). In single-molecule experiments, pulsed picosecond diode lasers are used to excite the fluorophores. The series LDH from PicoQuant emit pulses as short as 50 ps. The lifetime of a fluorophore is on the order of 1-10 ns, so the pulses are much shorter. The laser repetition rate and intensity can be adapted. Typical repetition rates are 40 MHz (25 ns) and 20 MHz (50 ns), this is to guarantee that there is enough time between pulses to collect even the longest times the fluorophore still emits a photon (Wang et al. 2005).

The pulsed laser is directed to the sample for excitation (Figure 2.6a). First, the light goes through the main dichroic that separates excitation light from emission light. Shorter wavelengths will be reflected, whereas longer wavelength traverse through. Light is focused through a high numerical aperture microscope objective to a diffraction limited spot. Emission fluorescent light is then collected through the same objective and goes through a pinhole, to reject out-of-focus light. The combined used of the high numerical aperture objective and the pinhole generate a confocal volume below 1 fL.

Fluorescently labeled proteins, with a FRET donor and acceptor, need to be diffusing in solution at very low concentration, 10-100 pM (10^{-9}-10^{-11} M), to ensure that the probability of two molecules residing in the confocal volume at the same time is negligible (Schuler 2013). When a labeled protein enters the laser beam, the donor dye will get excited and the fluorescence from donor and acceptor will be collected. Fluorescence light from donor and acceptor is separated by wavelength using a dichroic mirror. Finally, fluorescence light is detected using very sensitive detectors capable of counting individual photons, known as avalanche photodiodes (APDs). State-of-the-art counting electronics record the arrival time of every photon with picosecond time resolution (Figure 2.6a).

Each protein is detected as a “burst” of photons, which lasts for about a millisecond, corresponding to the time the protein lasts in traverse the confocal volume (Figure 2.6b). This means that bursts from hundreds to thousands of individual molecules

are typically collected in several minutes to hours. For every detected photon, we know the wavelength and the time of emission relative to the excitation pulse.

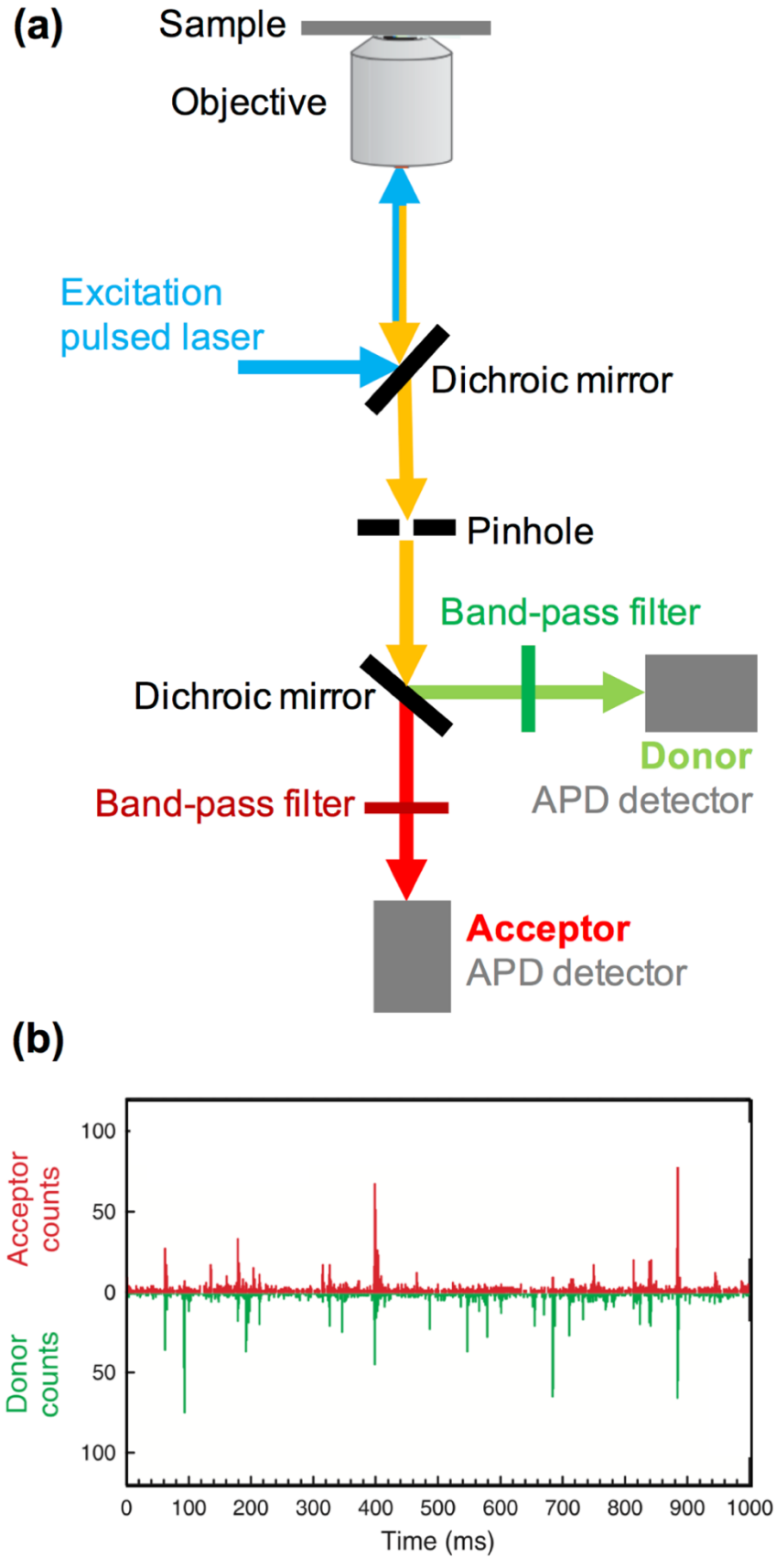

Figure 2.6. Confocal single-molecule set-up to detect FRET of freely diffusing molecules. (a) Components of a smFRET microscope. The excitation pulsed laser is reflected in the main dichroic and goes to the objective where it is focused on the sample. Emitted fluorescent light is collected back through the objective and passes through the main dichroic. Next, it goes through the pinhole and it is separated by wavelength in donor and acceptor emissions with another dichroic mirror. Emission photons are furthered filtered using band-pass filters for detection with avalanche photodiodes detectors. (b) Example of the data record for an smFRET experiment. Each bursts of photons represents a single molecule traversing the confocal volume.

To analyze smFRET data, bursts of photons corresponding to a single-molecule traversing the confocal volume need to be identified first. When a single-molecule traverses the detection volume, a large number of photons with short times between them are emitted. In contrast, when no single molecule is present in the detection volume, background is characterized by less photons with longer times between them. This allows the identification and classification of single-molecule events (Ingargiola, Lerner, et al. 2016). For every single-molecule burst identified, first several corrections are applied before calculations of transfer efficiencies. Bursts need to be corrected for background, differences in quantum yields and detection efficiencies of donor and acceptor, cross-talk between the channels, and direct excitation of the acceptor. Lastly, for every single-molecule event, FRET transfer efficiencies are calculated using equation 2.9, where n_D and n_A are the corrected numbers of donor and acceptor photons in the burst, respectively.

2.2 Computational methods for predicting protein stability

In this section, I summarize the main bioinformatics tools used to study the stability of Fab A33. This include molecular dynamic simulations to explore Fab dynamics under different solution conditions, *in silico* mutational studies to calculate free-energies and predict stabilizing mutations, and tools to predict the more aggregation-prone regions in Fab A33.

2.2.1 Homology modelling

When no crystal structure of the target protein is available from experimental techniques (X-ray crystallography or NMR), homology modelling offers a computational solution to generating an atomic-resolution model of the protein. Homology modelling constructs the new structure from the amino acid sequence of the target protein and an available experimental structure of a related homologous protein, called a template. It also receives the name template-based modelling (Ginalski 2006). Homology modelling have been successfully used when sequence identity is greater than 30% (Venselaar et al. 2010). When no amino acid sequence with high similarity is available, it is possible to predict the 3D structure of a protein from scratch, also called *ab initio* (or *de novo*) modelling. It has been shown that protein tertiary structure is better conserved than amino acid sequence. Evolutionarily, protein structure is conserved longer than amino-acid sequence and much longer than the corresponding DNA sequence. This implies that proteins with significant sequence identity, generally show high structural similarity.

Generation of a homology model follows four steps, (i) selection of the template protein structure, (ii) alignment of the amino acid sequences of target and template proteins, (iii) construction of the 3D structure model and (iv) assessment of the quality of the model (Muhammed & Aki-Yalcin 2019). Normally, the first and second steps are done at the same time, by finding an available protein structure with high sequence identity, by using sequence alignment algorithms, such as BLAST, and searching the PDB database. The choice of template structure is important, as the quality of the final

homology model depends on the sequence alignment. The alignment should be inspected manually after to ensure that the most conserved residues are well aligned. Then, homology models can be created using software such as Modeller (Eswar et al. 2006), SWISS-MODEL (Schwede et al. 2003), Phyre2 (Kelley et al. 2015) and Rosetta (Leaver-fay et al. 2011). Generally, sequence identities of 70% and higher, generate RMSDs of ~1-2 Å between the Cα of target and template structures, and this decreases with sequence identity, to only RMSDs of 2-4 Å at 25% sequence identity. Loop regions show particularly high RMSDs, due to their flexibility and the higher sequence variability in these regions. The final homology model can be further studied using molecular dynamics simulations and verified using experimental data.

2.2.2 Molecular dynamic simulations

Molecular dynamics (MD) are computational simulations that allow to study the dynamic evolution of the system, such as a protein in solution, by studying the motion of its atoms. MD simulations provide insights into protein stability, protein conformational changes, protein folding and interactions between molecules (for proteins, DNA, membranes or ligands). MD simulations can also be applied to drug design and structure prediction by simulating folding of the polypeptide chain from random coil (*ab initio*) (Hospital et al. 2015; Toofanny & Daggett 2012).

Usually, for large molecules such as proteins, simulations based on classical Newtonian mechanics are used. Quantum mechanics based modelling is also possible, however, it is normally used for small molecules, as it becomes computationally prohibitively expensive for large molecules. The concept behind MD simulations is that the position and velocity of each atom in the protein can be determined by solving Newton's equations of motion and by using a force field. First, energy functions (force fields) are stablished, which allow us to calculate the force experienced by any atom given the positions of the other atoms, using the equation:

$$F(x) = -\nabla U(x) \qquad \text{(Eq. 2.11)}$$

where x represents coordinates of all atoms, and U is the potential energy function. Once the forces acting on individual atoms are obtained, classical Newton's laws tell us how those forces will affect the motions of the atoms. Using Newton's second law:

$$F = ma \qquad \text{(Eq. 2.12)}$$

where F is force on an atom, m is mass of the atom, and a is the atom's acceleration. From the acceleration it is possible to calculate the velocity of the atom, and with the velocity we can update the atom positions. These equations can only be solved using numerical approaches. A time step shorter than the fastest movements in the protein should be used. Normally, time steps of 1 and 2 femtoseconds (10^{-15} s) for atomistic simulations are used. Thus, the basic algorithm for MD simulations is that at each time step (1 or 2 fs), forces acting on each atom are solved using a force field, and the atoms are moved by updating their velocity and position solving Newton's laws of motion.

A force field (potential function) describes all the interactions between atoms, using mathematical expressions to describe the potential energy (Weiner et al. 1984; Hagler et al. 1974). These include bond lengths (described as springs), bond rotations (described as springs), dihedral angles (described as periodic functions), improper angles (described as springs), van der Waals interactions (described by the Lennard-Jones potentials) and electrostatic interactions (described by Coulomb's law). They consist of both, bonded components, which includes stretching, bending, torsion and improper interactions, and non-bonded components, which includes both electrostatic and Van der Waals. The parameters in these expressions (such as atomic radius, atomic charge, equilibrium bond length, angle and dihedral) are obtained from comparisons of the force field with both, experimental and quantum mechanical data. Many force fields are available to generate MD simulations, and they differ in the way they are parameterized. Three major force fields are used for MD simulations, CHARMM (Brooks et al. 2009), AMBER (Case et al. 2005) and OPLS-AA (Kaminski et al. 2001). Not to confuse with the software packages available to perform MD simulations, which include CHARMM, AMBER, Gromacs (Abraham et al. 2015) and NAMD (Phillips et al. 2005). The dominant package for visualizing the results of the simulations is Visual Molecular Dynamics (VMD) (Humphrey et al. 1996).

There are two types of solvent, explicit and implicit. In an explicit solvent (such as the TIP3P, SPC/E and SPC-f water models), water molecules and ions are included, whereas in an implicit solvent the water effect is modelled (mathematical model to approximate the average effects of the solvent). An explicit solvent is more accurate, but it is more computationally expensive because it adds many more molecules to the simulation. In contrast, an implicit solvent is less accurate but faster. In an implicit solvent, certain phenomena such as conformational changes which reorient water dipoles, bridging water molecules and hydrogen bond fluctuations are neglected (Kleinjung & Fraternali 2014). In an explicit solvent, usually periodic boundary conditions are used, which imply that a water molecule that goes off the left side of the simulation box will come back in the right side.

MD simulations are normally carried out to nanoseconds (10^{-9} s) or microseconds (10^{-6} s). Structural changes in proteins can take place in nanoseconds, microseconds, milliseconds or even longer. However, many time steps need to be calculated for nanoseconds or microseconds trajectories (millions to trillions), which demand a high amount of computational power. Until recently, simulations of 1 microsecond were rare. Advances in computer power have enabled microsecond simulations, but simulation timescales still remain a challenge (Hospital et al. 2015).

Molecular mechanics force fields are inherently approximations, which cause MD simulations to have limitations. For example, hydrogen bonds have a partially quantum mechanical nature, however, they are described as Coulomb interactions of atomic point charges. Every atom is assigned a fixed partial charge at the beginning of the simulation, whereas electron clouds are constantly shifting according to their environments, so that partial charges would be better represented as dynamic. Additionally, covalent bonds cannot break or form during (standard) MD simulations, which implies that the protonation state of acid and basic groups does not change during the simulations.

2.2.3 Computational prediction of protein ΔΔG upon mutation

Computational programs were developed to be able to predict the effect of mutations on the stability of proteins. Specifically, these methods calculate the changes in Gibbs free energy due to a point mutation, ΔΔG, which correspond to the difference between the ΔG of the protein carrying the mutation minus the ΔG of the WT protein. These software allow the prediction of beneficial mutations to enhance protein stability, from point mutations that reduce the free energy of the protein. Among the best known, are FoldX (foldx.crg.es) (Zhang et al. 2012) and Rosetta (www. rosettacommons.org) (Kellogg et al. 2011) software.

Both software require the 3D structure of the protein. Thus, the quality of the protein structure is crucial for accurate calculations. Then, they use energy functions to calculate the Gibbs free energy. FoldX and Rosetta differ in that Rosetta uses a physical energy function to simulate the forces between atoms, whereas FoldX uses statistical potentials and the parameters in the energy calculation were determined in laboratory experiments (from datasets of experimentally characterized protein mutants). Both software are better at predicting stability trends than quantitatively estimating the value of the stabilization; where total energies are not able to predict experimental results. Additionally, the accuracy of these algorithms remains low, that is why they are often combined to find coincident stabilizing mutation predictions (Buß et al. 2018; Wijma et al. 2014).

There are several reasons as to why the accuracy of computational approaches for stability engineering remains low (Magliery 2015). The key forces that underlie protein stability are well understood, such as the burial and tight packing of hydrophobic residues, the ejection of ordered solvent, and the formation of hydrogen bonds and other electrostatic interactions, conformational entropy, and bond strain. But some of them are hard to calculate. For example, the gain in solvent entropy that largely underlies the hydrophobic effect is not explicitly included in these calculations. Further, solvation is hard to consider due to the challenge of polarizability. The free energy of folding, ΔG, between the folded state and the unfolded state, is also challenging to calculate due to the difficulty in modeling the unfolded state. It is also very difficult to model the effects of misfolding, or to account for alternative conformations. Yet, many examples exist where these algorithms have been successfully applied to predict stabilizing mutations that have

been confirmed experimentally. These findings suggest that existing energy functions take into account important stabilizing effects. Core packing is believed to be the dominant factor, which is also correlated to incomputable factors such as solvent entropy.

2.2.4 Predicting aggregation-prone regions

The β-sheet structure typical of aggregates, is composed of hydrophobic amino acids with a high β-sheet propensity and a low net charge. The regions in proteins capable of forming these β-sheet structure are termed aggregation-prone regions (APRs) (Ventura et al. 2004; Pawar et al. 2005). APRs are generally short sequence segments (5-15 amino acids) that display high hydrophobicity, low net charge and a high tendency to form β-structures (De Baets et al. 2014). Also, generally APRs are located in the interior of proteins, protected from the solvent. And if they were to become exposed, they could trigger aggregation by self-associating.

Several methods have been developed to identify APRs in proteins. These methods can be separated in sequence-based and structure-based methods. Sequence-based APR predictors only use the protein sequence as input, equivalent to the fully unfolded state. Predictions are based on either the intrinsic properties of amino acids, or their compatibility with protein structural features in known amyloid fibril structures. Examples include TANGO (Fernandez-Escamilla et al. 2004), AGGRESCAN (Conchillo-Solé et al. 2007), PASTA (Walsh et al. 2014), MetAmyl (Emily et al. 2013), FoldAmyloid (Garbuzynskiy et al. 2010), FishAmyloid (Gasior & Kotulska 2014) and Waltz (Maurer-Stroh et al. 2010). They all consider the hydrophobicity, charge and secondary structure propensity of short sequences of the amino acid sequence. TANGO additionally looks at the propensity to form β-sheet structures instead of other structures such as α-helix, β-turn, β-strand and random coil. AGGRESCAN has an aggregation propensity scale for each of amino acid, calculated from the relative solubility of point mutants of amyloid β-peptides in *E. coli*. FoldAmyloid considers the packing density, where regions with a strong packing density are considered amyloidogenic. PASTA is based on the assumption that β-strands constituting the amyloid fibril have a preference for an in-register parallel or anti-parallel arrangement with minimal energy. Creating a dataset with these strictly defined secondary structures allowed the calculation of a pairing energy for each possible pair of residues, which is then used to score all possible

stretches. Lastly, as these predictions do not always agree, Amylpred2 was built which is a consensus from up to eleven existing algorithms (Tsolis et al. 2013).

Sequence-based APR predictors identify the APR regions in proteins. However, these APRs need to become exposed to the solvent in order to trigger aggregation, either by structural dynamics or partial unfolding. Thus, structure-based methods were developed, which take into account the 3D structure of the protein and identify the APRs likely to become exposed. Examples of structure-based APR predictors include AGGRESCAN 3D (Zambrano et al. 2015), AggScore (Sankar et al. 2018), SAP (Chennamsetty et al. 2009) and Solubis (Van Durme et al. 2016). AGGRESCAN 3D combines the information provided by AGGRESCAN with structural information to identify spatially proximal aggregation-prone regions. Additionally, AGGRESCAN 3D allows fast simulations to explore the flexibility and dynamic regions of the protein. AggScore uses the distribution of hydrophobic and electrostatic patches on the surface of the protein, and additionally taking into account the intensity and relative orientation of these patches to predict the APRs. Lastly, SAP uses atomistic molecular dynamics simulations in combination with a new way to calculate aggregation propensity, which they termed spatial aggregation propensity (SAP). SAP measures the exposed hydrophobicity of certain patches on the protein surface, and this can be monitored during the simulations. SAP identifies the extent of aggregation prone hydrophobic patches exposed on the surface of the protein, which could be natively exposed, exposed due to dynamic fluctuations or due to conformational changes.

2.3 Cloning and protein expression

2.3.1 Cloning and site-directed mutagenesis

2.3.1.1 CPEC as cloning method

In this thesis, I used the method circular polymerase extension cloning (CPEC) to clone Fab A33 into the new expression plasmid pET-29a(+). CPEC is a sequence-independent cloning method, in which overlapping flanking regions are created in the insert(s) and vector using PCR, followed by the CPEC reaction (Quan & Tian 2011). In a typical CPEC reaction, linear double-stranded insert(s) and vector are first heat-denatured, and the resulting single strands are then annealed with their overlapping ends and extended using each other as a template to form double-stranded circular plasmids.

For a successful CPEC, the overlapping regions are designed to have a high and close melting temperatures (within 60-70 °C and ± 2 °C), which eliminates vector reannealing and concatenation of inserts. CPEC primers were designed using the Gibson/Assembly option in SnapGene software (from GSL Biotech; available at snapgene.com), and purchased from Eurofins (Wolverhampton, UK) (Table 2.1). Stock solutions of primers at 100 μM were first prepared, and then aliquots of 10 μM (10x working solution) were prepared and stored at -20 ºC. Primer T_m's were measured using the calculator:
http://www.basic.northwestern.edu/biotools/oligocalc.html

First, insert and vector carrying the overlapping sequences are prepared using PCR (Tables 2.2 and 2.3). The annealing temperature for Q5 polymerase was selected as the T_m of the primer with lowest melting temperature +3 °C. The extension time was measured as 20 s per kilobase of the final product (7.425 kb). The finished PCR product, 5 μL, were run in a 8% agarose gel containing SYBR safe stain (provided as a 10.000x solution), at 120 V for 40 min, in 1X TBE buffer (Sigma Aldrich, USA). I loaded 2 μL of the amplified reaction with the corresponding 6X loading dye (NEB, USA). 1 Kb ladder (NEB, USA) was used as a size reference. Bands were visualized under UV at 302 nm (Alphaimager mini, Protein simple, USA). The molecular weight of the amplified product were confirmed.

Insert DNA was purified, first excised using a scalpel and then digested using QiaQuick gel purification kit (Qiagen, Holland). Vector DNA, was first treated with Dpn I enzyme to digest the original pET-29a(+) empty vector (methylated DNA). 1 µL of DpnI enzyme was added to the PCR amplified vector and incubated at 37 °C for 10 min, and heat inactivated at 80 °C for 20 min. Lastly, purification of the vector was done with E.Z.N.A. Cycle Pure Kit (Omega). DNA concentrations were measured for absorbance at 280 nm on a Nanodrop 2000 (Thermoscientific, USA).

Table 2.1. Primers used to clone Fab A33 into pET-29a(+) using CPEC. Overlap is indicated in blue color.

Primer name	Sequence (5'-3')	*T_m PCR amplification (℃)*	Tm CPEC overlap (°C)
Insert.REV	GGCTTTGTTAGCAGC*GATATGACGACAGGAA*GAGTTTGTAGAAACG	*60.4*	63
Vector.REV	TTCCTGTCGTCATATC*GCTGCTAACAAAGCC*CGAAAGG	*56.7*	
Insert.FOR	TGATGTCGGCGATA*CCATCGGAAGCTG*TGGTATGG	*56.3*	62.8
Vector.FOR	CAGCTTCCGATGG*TATCGCCGACATCA*CCGATGGG	*58.6*	

Table 2.2. PCR setup for amplification of insert and vector containing CPEC overlapping sequences.

	Insert		Vector	
Reagents	Vol (µL)	Final amount	Vol (µL)	Final amount
MilliQ water	32.5		2.5	
5x Q5 Buffer	10	1x	10	1x
DNA (Insert 50 ng/µL; Vector 10 ng/µL)	1	50 ng	1	10 ng
Forward primer (10 µM)	2.5 (Insert.FOR)	0.5 µM	2.5 (Vector.FOR)	0.5 µM
Reverse primer (10 µM)	2.5 (Insert.REV)	0.5 µM	2.5 (Vector.REV)	0.5 µM
dNTPs (10 mM)	1	200 µM	1	200 µM
Q5 DNA Polymerase	0.5	1 unit	0.5	1 unit
Total	50		50	

Table 2.3. PCR conditions for amplification of insert and vector containing CPEC overlapping sequences.

Step	Cycles	Insert	Vector
Initial denature	1	98 °C, 30 s	98 °C, 30 s
1. Denature	2-31	98 °C, 10 s	98 °C, 30 s
2. Anneal		59 °C, 30 s	60 °C, 30 s
3. Extend		72 °C, 60 s	72 °C, 60 s
Final extension	32	72 °C, 5 min	72 °C, 5 min
Hold		4 °C, ∝	4 °C, ∝

Insert and vector were then combined to assemble the final product, where they serve as template of each other. Parameters that can be varied in this reaction are the concentration of vector DNA (typically 5-10 ng/µL), the ratio of insert to vector (typically 1-30:1 insert to vector) and the number of cycles (typically 5-30 cycles) (Speltz & Regan 2013). We used a concentration of 2 ng/µL of vector DNA, and the ratio that best worked for us was 30:1 insert to vector (Table 2.4). We had a vector stock of 100 ng/µL, and we added 1 µL to the reaction to add 100 ng, which correspond to 0.024 pmol of vector DNA. We calculated the µL necessary to add 30 times more mols of insert than vector using Eq. 2.13. Then, we used 10 cycles, through 98 °C denature, 55 °C annealing of the primer homologous sequences and finally 78 °C (Table 2.5).

$$\frac{pmol}{\mu L} = \frac{conc\left(\frac{ng}{\mu L}\right) x\ 1000}{base\ pairs\ x\ 650\ daltons} \qquad \text{(Eq. 2.13)}$$

After the CPEC reaction, the double-stranded circular plasmids with one nick in each strand can be directly transformed into competent host cells, and host DNA repair enzymes seal the nicks in vivo. 5 µL of the CPEC reaction were directly used to transform via heat-shock NEB 10β *E.coli* competent cells (New England Biolabs, Ipswich, US), and plated overnight on 50 µg/mL Kanamycin LB agar plates to form colonies. The following day many colonies grew and some of them were picked to check the CPEC reaction.

The chosen colonies were grown over night in 5 mL of LB media in a 50 mL falcon tube containing 50 μg/mL of kanamycin, incubated at 200 rpm and 37 °C. Part of the overnight culture was used to prepare glycerol stocks and part of the culture to extract the plasmid DNA. To make a glycerol stock, 500 μL of the overnight culture was then mixed with 500 μL of a 50% glycerol solution (v/v) and stored at -80 °C. For plasmid extraction, the overnight culture was first centrifuged at 5000 g for 10 min at 4 °C. The supernatant was discarded, and plasmid DNA was extracted from the cell pellet using QIAprep spin miniprep kits (Qiagen, Holland) and the plasmid was stored at -20 °C. DNA plasmid was also used for sequencing using Source Bioscience (UK) pre-paid voucher Sanger sequencing using a 100 ng/μL and 5μL sample per reaction, using the sequencing primers (Table 2.6).

Table 2.4. CPEC setup.

Reagents (initial concentration)	I:V, 30:1 Vol (μL)	Final amount
MilliQ water	12.5	
5x Q5 Buffer	10	1x
Vector DNA (100 ng/μL)	1	100 ng
Insert DNA (100 ng/μL)	15	
dNTPs (10 mM)	1	200 μM
Q5 DNA Polymerase	0.5	1 unit
Total	50	

Table 2.5 CPEC conditions.

Step	Cycles	Conditions
Initial denature	1	98 °C, 30 s
1. Denature	2-11	98 °C, 10 s
2. Anneal		55 °C, 30 s
3. Extend		72 °C, 2min 29s
Final extension	12	72 °C, 5 min
Hold		4 °C, ∝

Table 2.6. Sequencing primers to confirm the cloning of Fab A33 into pET-29a(+).

Primer name	Sequence (5'-3')	T_m (°C)
pET29Fab_for	AGGAATGGTGCATGCAAGG	51.1
pET29Fab_mid	AGTGGAAGGTGGATAACGC	51.1
T7 term	CTAGTTATTGCTCAGCGG	48

2.3.1.2 Site-directed mutagenesis

Mutations were made using the QuickChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, USA). Mutations were introduced one at a time. Primers were designed using the mutagenesis option in Snapgene, and purchased from Eurofins (Wolverhampton, UK) (Table 2.7). A single point mutation was introduced in a supercoiled double-stranded DNA, using two primers, each complimentary to opposite strands of the vector, where both contain the desired mutation. The primers were first extended in a PCR reaction using PfuUltra HF DNA polymerase (Tables 2.8 and 2.9). The final PCR product was first treated with Dpn I enzyme for 5 min at 37 °C, to digest the original vector that does not contain the mutation and is methylated. The products were run on an agarose gel to check that the PCR worked. Lastly, the final nicked plasmid containing the mutation was transformed into NEB 10β *E.coli* competent cells via heat-shock. The introduction of the mutations was confirmed by sequencing, using the sequencing primers (Table 2.10). To form the double mutants, an additional mutation was incorporated using one of the previous single-mutants and the same PCR conditions as in Tables 2.8 and 2.9. The incorporation of the double mutations was finally confirmed using primers in Table 2.10.

Table 2.7. Primers used to introduce mutations via site-directed mutagenesis. The desired mutation is highlighted in red color.

Mutation		Sequence (5'-3')	T_m (°C)	T_m each side (°C)	
LC-K126pAzF	AAA to TAG (a376t_a377a_a378g)	CCATCTGATGAGCAGTTGTAGTCTGGAACTGCCTCTG CAGAGGCAGTTCCAGACTACAACTGCTCATCAGATGG	67.8	48	46
LC-S156C	TCG to TGC (c467g_g468c)	GGATAACGCCCTCCAATGCGGTAACTCCCAGGAG CTCCTGGGAGTTACCGCATTGGAGGGCGTTATCC	69.2	46	45
HC-S117pAzF	TCT to TAG (c350a_t351g)	CACTGGTGACAGTGTCTTAGGCCTCAACGAAGGGC GCCCTTCGTTGAGGCCTAAGACACTGTCACCAGTG	69.1	47	47

Table 2.8. PCR setup for site-directed mutagenesis reactions.

Reagents	Vol (μL)	Final amount
MilliQ water	40.3	
10x Reaction buffer	5	1x
DNA: pET-29a(+) with Fab (112 ng/μl)	0.5	50 ng
Forward primer (10 μM)	2.5	11 pmol (125 ng)
Reverse primer (10 μM)	2.5	11 pmol (125 ng)
dNTPs (10 mM)	1	200 μM
Quick Change Enzyme	0.5	1 unit
Total	50	

Table 2.9 PCR conditions for site-directed mutagenesis reactions.

Step	Cycles	Conditions
Initial denature	1	95 °C, 2 min
1. Denature	2-19	95 °C, 20 s
2. Anneal		60 °C, 10 s
3. Extend		68 °C, 3 min 43s
Final extension	20	68 °C, 5 min
Hold		4 °C, ∝

Table 2.10. Sequencing primers to confirm the generation of Fab A33 mutants.

Primer name	Sequence (5'-3')	T_m (°C)
Mutations_LC	TCATCTATTTGGCCTCCAAC	49.7
Mutations_HC	TGTGCAGCATCTGGATTC	48

2.3.2 Protein expression and purification

2.3.2.1 Expression of Fab A33 WT and mutants

Plasmid vector pTTOD containing Fab A33 WT (C226S) was transformed via electroporation into W3110 *E. coli* electro-competent cells for protein expression. For Fab A33 mutants to be used in smFRET, plasmid pET-29a(+) (containing Fab gene) and plasmid pEVOL-pAzF (Young et al. 2010), were co-transformed into C321.ΔA.exp ("amberless)" *E. coli* (Lajoie et al. 2013) competent cells via the heat shock method, for expression. Colonies were first grown in agar plates, followed by growth of individual colonies in overnight liquid cultures, and storage as glycerol stocks.

Expression of Fab A33 was done in bioreactors, as they allow for the control of pH, temperature, dissolved oxygen and nutrient concentration, thus reaching much higher cell densities than in shake flasks. First, a pre-culture was grown in 2xPY complex media (16 g/L of phytone, 10 g/L yeast extract and 5 g/L NaCl). In the case of Fab WT, 10 μg/mL of tetracycline were added (to maintain pTTOD plasmid). For Fab A33 mutants, they contained 50 μg/mL kanamycin (to maintain pET-29a plasmid) and 34 μg/mL chloramphenicol (to maintain pEVOL-pAzF plasmid). These pre-cultures were grown in 20 mL in 250 mL shake flasks, inoculated with the respective glycerol stock, and grown at 37 ºC, 250 rpm for approximately 4 h, until the OD_{600} reached 1-2. Then, 2 mL of these cultures were transferred into SM6G defined media (5.2 g/L NaH_2PO_4, 3.3 g/L Na_2HPO_4, 4.4 g/L KCl, 1.04 g/L $MgSO_4$, 4.16 g/L citric acid, 0.25 g/L $CaCl_2$, 112 g/L glycerol, 10 ml/L SM6 elements). As before, the corresponding antibiotics were added. Additionally, 2.5 μg/mL of D-biotin were added to the expression of Fab mutants, because the C321.ΔA.exp strain is auxotrophic for D-biotin, and needs to be supplemented in minimal media. These cultures of 20 mL in 250 mL shake flasks, were incubated at 30 ºC, 200 rpm, for 12-16 h, until an OD_{600} of 4-5 was reached.

Fermentation was done in 250 mL DASbox Mini Bioreactors (Hamburg, Germany). 20 mL of the over-night pre-culture were used to inoculate 150 mL of SM6G media (in a total volume of 170 mL). For Fab A33 WT, no antibiotics were added. For Fab A33 mutants, antibiotics were added (50 μg/mL kanamycin and 34 μg/mL chloramphenicol) to guarantee the maintenance of both plasmids, together with 2.5 μg/mL D-biotin. During the fermentation, a homogeneous internal environment was

maintained at 30 ºC, pH 6.95 and pO_2 above 30%. pH of 6.95 was maintained by the addition of dilute acid and base (10% v/v sulphuric acid or 15% v/v ammonium hydroxide respectively) all coordinated by a Biostat digital control unit. Aeration of the culture medium was sustained by sparging with of sterilized air at 20 L/min ensuring a dissolved oxygen tension (DOT) 30% (assuming the oxygen concentration of the atmosphere is equivalent to 100%). When inadequate aeration levels were reached with a peak impeller rotation rate of 1400 rpm, the feed gas was blended with pure oxygen in a 60:40% v/v ratio respectively. Foaming was mitigated through the addition of polypropylene glycol (PPG) 2000, again controlled using the BIOSTAT digital control unit.

Once an OD_{600} of 40 was reached, a 150 mL magnesium shot (1M Magnesium sulphate heptahydrate solution) was added to the vessel, and the culture temperature reduced to 25 ºC. At this point, for Fab A33 mutants, the nonstandard amino acid pAzF was added to a final concentration of 1 mM (Chem-Impex International, Wood Dale, US). For Fab WT, once the carbon source within the media was finished, indicated by a spike in DOT, fermentation was switch to fed-batch, where 80 %w/w glycerol solution was continually added to the fermenter at 0.7 mL/h. Fab WT expression was induced with addition of 0.2 mM IPTG (final concentration), and the culture was harvested approximately 24 hours post induction. For Fab A33 mutants, no spike in DOT was observed. Thus, once an OD_{600} of 100 was reached, expression of Fab A33 mutants was induced by addition of 0.2 mM IPTG and 0.2% (w/v) and L-(+)-arabinose (inducer for pEVOL-pAzF), (final concentrations). Cells were harvested approximately 24 hours post induction. All cells were pelleted by centrifugation at 10,000 rpm and 4°C for 1h 30 min, before storage of the cell pellet at -20 ºC.

2.3.2.2 Purification of Fab A33 WT and mutants

The recovery of soluble Fab A33 from the *E. coli* periplasm was done via chemical extraction with Tris/EDTA buffer (100 mM Tris pH 7.4, 10 mM EDTA), which destabilize the outer membrane and cell wall while leaving the inner membrane largely intact. Cell pellets were re-suspended in 150 mL of this lysis buffer, and let to react over night at 50 ºC and 100 rpm. A centrifugation step of 1.5 h and 13000 rpm followed, and the Fab A33 rich supernatant was collected for purification.

Fab A33 was purified using the AKTA purifier FPLC system, installed with a XK50 column. The column was packed with Sepharose fast flow protein G resin (GE healthcare), to purify Fab A33 from the other *E. coli* proteins using affinity chromatography. The column was equilibrated with 25 mM sodium phosphate, pH 7.4. The filtered heat lysate was passed through the column, and a washing step was carried out with 3 column volumes of equilibration buffer, followed by two column volumes of equilibration buffer plus 5% v/v isopropanol to remove any hydrophobically bound impurities. Finally, Fab A33 was eluted using 60 mM sodium citrate at pH 3.5, and quickly neutralized by addition of 1 M Tris-HCl, pH 8.5. Fab A33 was buffer-exchanged into the corresponding buffer and concentrated using 10 kDa cut-off centrifugal filters (Merck, Kenilworth, UK). Protein concentration measurements were made using a spectrophotometer, and calculated using the Beer-Lambert law, $c = A/(\varepsilon l)$, where A is the absorbance at 280 nm, l corresponded to the cuvette path length (typically 1 cm), c the unknown protein concentration in mol/L, and ε the molar extinction coefficient of the protein.

Chapter Three

Stability of Fab A33 at low pH and high temperature by molecular dynamics simulations and its stabilization by computational design

3.1 Summary

Protein-based drugs are widely used for the treatment of numerous human diseases. Their development into successful products largely depends on their stability in addition to their specific mode of action. Thus, knowledge about their stability against different stresses, specially early in the development process, is crucial to engineer more stable proteins. In this chapter, I study the structural robustness of Fab A33 using atomistic molecular dynamics (MD) simulations under two stresses, low pH and high temperature. Results revealed that interface contacts between domains were the first to break, prior to domain unfolding. Contacts in the constant interface (C_L-C_H1), were lost quickly during the simulations under both stresses. Notably, FoldX and Rosetta, both agreed that the residues in Fab A33 that can be stabilized the most, were located in this interface. Further support was provided by packing density calculations, which revealed that these residues were under-packed compared to all other inter-domain residues. RMSD calculations and structural alignments showed that at low pH, C_L domain unfolded first, while at high temperature, both C_L and V_H unfolded, revealing different unfolding pathways. These conformational changes exposed different predicted aggregation-prone regions (APR), to suggest different aggregation mechanisms. Salt bridge analysis identified two salt bridges, Glu165-Lys103 and Glu195-Lys149, which possibly drive the conformational change at low pH. They were the most persistent salt bridges at pH 7.0, were not present at pH 3.5, and are both located in the C_L domain. At high temperature, salt bridges broke and reform much quicker and not always with the same partner, contributing to Fab destabilization. Sequence entropy analysis of existing Fab sequences confirmed that there is room for Fab engineering, where certain natural mutations agreed with FoldX and Rosetta predictions. Overall, results from this chapter identified the early stages of unfolding and stability-limiting regions of Fab A33, which can be mutated to engineer more stable Fab fragments.

3.2 Introduction

In the last 30 years, monoclonal antibody products have become the main drug class for new approvals in the pharmaceutical industry (Ecker et al. 2015). To date, over 60 antibody-based drugs are on the market, representing half of the total sales, with over 550 further antibodies in clinical development (Carter & Lazar 2018). They are used as therapeutic drugs to treat human diseases, mainly in oncology, auto-immune diseases and cardiovascular diseases. The use of antibody fragments, such as the antigen-binding antibody fragment (Fab) studied here, brings additional advantages, including deeper tissue penetration due to their smaller size, which has proven beneficial to treat tumours (Nelson 2010). In addition, Fab fragments lack the Fc domain, and thus are not glycosylated which allows simpler and less costly manufacture due to their expression in prokaryotic systems (Enever et al. 2009). However, the lack of the Fc domain leads to their more rapid clearance in humans than for full antibodies.

The stabilization of therapeutic proteins against aggregation remains one of the biggest challenges facing their approval as biopharmaceutical products (Manning et al. 2010; Wang et al. 2010; Wang 2005). Not only their mode of action, but protein stability is a crucial factor to their becoming successful products. Novel antibody products such as Fabs, single-chain variable fragments (scFvs) and bi-specifics are currently being developed and their properties remain largely unknown. Knowledge about the stability of these pharmaceutical products, specially early in the development process, would aid in their engineering and the design of antibody fragments that are aggregation resistant.

Native protein conformations are only marginally stable, and are highly dynamic, hence they are more realistically described as a native ensemble. There is increasing evidence to suggest that under native conditions, aggregation takes place primarily from partially unfolded native-like state (Chiti & Dobson 2009; Neudecker et al. 2012; Canet et al. 2002; Kendrick et al. 1998; Chakroun et al. 2016). Small changes to their environment (e.g., temperature, pH, salt type, salt concentration, cosolutes, preservatives) can destabilize the structure of the protein, and induce unfolding. A loss of the native protein structure, may expose aggregation prone-regions (APR) in the protein, that would normally be shielded in the interior of the protein, and promote aggregation. Different structural regions in the protein may resist differently the destabilization of the external

variable. Thus, determining the conformational changes that a protein experiences under a given set of conditions is important for its stabilization (De Baets et al. 2014; Codina et al. 2019).

Typically, high temperatures perturb the native conformation of the protein to a sufficient degree to promote aggregation. It is observed that aggregation starts at temperatures well below the equilibrium melting temperature (T_m) of the protein, suggesting that partially unfolded conformations are also aggregation-prone (Chi et al. 2003). In addition to their effect on protein conformation, temperature also affects the reaction kinetics, increasing the collision frequency and the number of collisions with enough energy to overcome activation energies, favouring aggregation. Regarding pH, proteins are often stable over narrow ranges of pH, and changes in pH affect the electrostatic interactions. Specific charge interactions, such as salt bridges, that stabilize the native conformation, might be lost, causing a conformation change. In the situation at pH far removed from the isoelectric point (pI) of the protein, an unfolded state of the protein would be favoured where the charge density is lower than in the folded state. pH also has an effect in colloidal stability, where highly charged proteins will repulse each other, not favouring aggregation (Chakroun et al. 2016). In this chapter, I only study the effect of pH and temperature on protein conformation.

Molecular dynamics (MD) simulations have been extensively used to study protein stability (Lindorff-Larsen et al. 2012; Rocco et al. 2008; Salimi et al. 2010; Settanni & Fersht 2008; Collu et al. 2018; Patel & Kuyucak 2017). MD simulations offer atomic resolution to the early conformational events that take place under different conditions. To date, not many studies on antibody fragments have been reported. MD simulations were used to study the potential aggregation liabilities of an antibody Fab fragment, from a human IgG1k antibody, via multiple elevated temperature MD simulations at 300 K, 450 K and 500 K (Buck et al. 2013). Results revealed that domain interfaces deformed prior to the unfolding of individual domains, and from the analysis of the structural changes, they identified two potential aggregation liabilities in the V_H domain of that Fab. Structural deformations in that domain increased the solvent-accessible surface area of the APRs in these regions. The unfolding process of an antibody Fab fragment was also studied using an elastic network model, to reveal that the constant regions are more flexible than the variable regions, and they unfolded earlier (Su et al. 2015). MD simulations at 450 K and 500 K were also used to study the stability-limiting

regions of an antibody single-chain variable fragment (scFv) (Wang & Duan 2011). They found that disruption of the V_L-V_H interface was the first event leading to the unfolding of the native structure of the protein. In contrast to the other works, they found V_H domain to be more thermally resistant than the V_L domain.

Each Fab is composed of one light and one heavy chains (Figure 3.1). Each chain contains a variable (V_L and V_H) and a constant (C_L and C_H1) domains. Each domain has the secondary structure of a β-barrel, also called an immunoglobulin fold, with two layers of β-sheets. Constant domains are formed of seven β-strands and variable domains have two additional β-strands. The variable domains contain the antigen-binding site, also called complementary determining regions (CDRs), formed by three loops in V_L and three lops in C_L. There are five disulphide bonds in Fab, four of them intra-domain and the last one between the light and heavy chains. Individual domains interact with one another, V_L with V_H form the variable region interface (V_L-V_H), and C_L with C_H1 form the constant region interface (C_L-C_H1). Interface contacts are shown in Figure 3.1 and the residues involved in the contacts are listed in Table 3.1. The variable region interface is mainly formed by aromatic side chains that are tightly packed and located at the centre of the interface (six Tyr, two Trp and two Phe), forming hydrophobic interactions. However, less aromatic side chains are involved in the constant region domain interface (four Phe), and no contacts were found between the β-strand 177-180 in C_L domain and the C_H domain at 3.5 Å in our Fab A33 homology model (Figure 3.1).

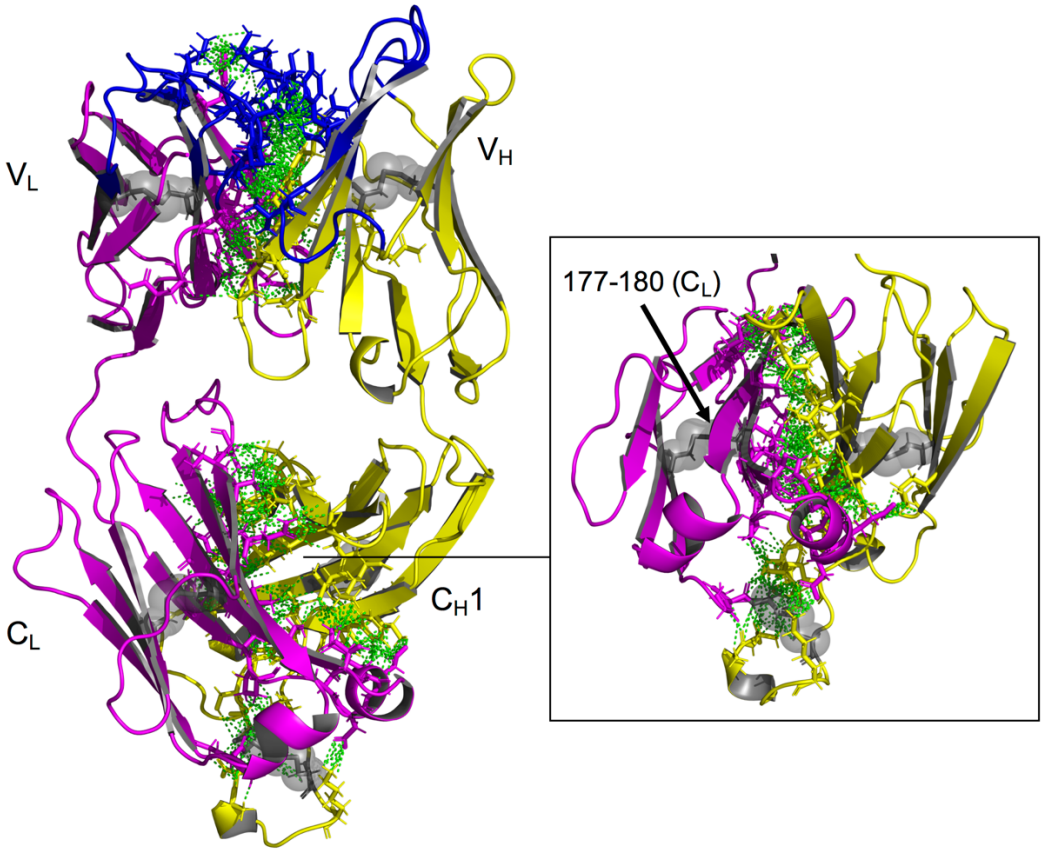

Figure 3.1. Fab A33 structure with interface contacts highlighted. Fab is composed of light (magenta) and heavy (yellow) chains. Each chain contains variable (V_L and V_H) and constant (C_L and C_H1) domains. The antigen-binding region at the complementary determining regions (CDRs; blue), are located in the variable domains. There are five disulfide bonds (gray highlights). Contacts between heavy and light chains within 3.5 Å are indicated with green dashed arrows. β-strand 177-180 in C_L domain does not have contacts with C_H domain, zoom in right-inset.

Table 3.1 Residues located in the interface between light and heavy chains in Fab A33.

V_L	V_H	C_L	C_H	Hinge
V32	E215	F116	F340	K432
Y36	V251	F118	P341	K436
Q38	Q253	P119	L342	A441
A43	G258	S121	A343	A442
P44	L259	D122	P344	
K45	E260	E123	S345	
T46	W261	Q124	S346	
Y49	T264	T129	T353	
L50	Y273	S131	A354	
H55	L275	V133	A355	
T56	Y309	L135	L359	
G57	T313	N137	K361	
F87	T314	Q160	H382	
L89	V315	S162	T383	
H91	P317	V163	F384	
Y94	A319	T164	P385	
P95	Y320	S174	V387	
L96	W321	S176	Q389	
F98	G322	T178	S397	
Q100	Q323	G212	V399	
		E213	T401	
			S404	
			K427	

Here, I report the early unfolding events of Fab A33 at high temperature and low pH, using all-atom MD simulations. A common feature to both stress conditions was that unfolding was initiated by the loss of interfacial contacts between neighboring domains, and that domain unfolding occurred later. However, my results revealed different unfolding pathways for the two stress conditions, leading to partial unfolding of only the C_L domain at low pH, compared to destabilization of both C_L and V_H domains in the high temperature condition. These conformational changes exposed different predicted aggregation-prone regions (APR), which would additionally support divergent aggregation mechanisms. Salt-bridge analysis provided insights into the location of those that were broken most rapidly due to protonation in the low pH simulation, and also showed that high temperature led to an increased fluctuation of salt bridge formation and breaking, more generally throughout the structure. An *in-silico* mutational analysis by both FoldX and Rosetta, predicted that the constant domain interface had the greatest potential for further stabilization, a finding that was also supported by lower packing-density calculations. Taken together, these results determined the stability-limiting regions at low pH and high temperature for Fab A33, and also identified those with the greatest potential for mutations that simultaneously improve stability under both low pH and high temperature conditions.

3.3 Methods

3.3.1 Fab A33 homology model

Fab A33 homology model was built by Dr Cheng Zhang (Zhang et al. 2018), and the method followed is summarized here. The homology model of wild-type Fab A33 was built using Rosetta method "minirosetta", from the crystal structure of human germline antibody 5-51/O12 (PDB ID: 4KMT) and the amino-acid sequence of Fab A33 (Figure 1.7), (Chivian & Baker 2006; Raman et al. 2009). The C226S heavy-chain variant was used to avoid the formation of linked Fab dimers. After residue replacement, 6,811 out of 20,000 structure models retained the five disulphide bonds intact. From these, 1000 structures with the lowest Rosetta Energy Units were selected, and clustered based on their similarities. The largest category in the clustering step contained 573 structures, and the structure with the lowest score in this category was selected as the homology model of Fab A33.

3.3.2 Molecular dynamics simulations

Molecular dynamic (MD) simulations on the Fab A33 homology model were conducted in Gromacs v5.0 (Abraham et al. 2015). MD simulations were carried out at neutral pH and room temperature (pH 7.0 and 300 K) and under two stresses, low pH (pH 3.5 and pH 4.5 at 300 K) and high temperature (pH 7.0 at 340 K and 380 K). Many high temperature simulations are performed at relatively high temperatures (e.g. 500 K), to achieve complete denaturation of the protein, however, in this case, I aimed to partially unfold Fab A33 and detect the regions prone to early unfolding. Simulations were carried out using the OPLS-AA/L all-atom force field (Kaminski et al. 2001; Kortkhonjia et al. 2013; Hu & Jiang 2010; Smith et al. 2015; Yu & Dalby 2018; Yu et al. 2017; Zhang et al. 2018). The Fab PDB file was first converted to a topology file with its five (four intra-chain and one inter-chain) disulphide bonds retained. The protonation state of each residue was entered manually, and these were determined at each pH using the PDB2PQR server, which performed the pKa calculations by PropKa (Li et al. 2005). This gave the following total charges: +9 (pH 7.0), +18 (pH 4.5) and +35 (pH 3.5). The Fab A33 structure was centred in a cubic box with a layer of water up to at least 10.0 Å from the

protein surface. The box was solvated with SPC/E water molecules, Cl^- added to neutralize the net charges, and NaCl added to an ionic strength of 50 mM for all simulations. The system was energy minimized using the steepest descent algorithm (2000 steps) followed by the conjugate gradient method (5000 steps). The solvent and ions around the protein were equilibrated in two phases of position-restricted simulations of the heavy atoms of the protein. First, the desired temperature was reached with 100 ps under NVT ensemble (constant number of particles, volume and temperature) using the velocity rescaling thermostat (based on kinetics energies). In this step, velocity generation took place, using random seeds to generate different initial velocities. Thus, from the same starting structure, different simulations were conducted. Next, the pressure was stabilized to atmospheric pressure with 100 ps under NPT ensemble (constant number of particles, pressure and temperature) using the Parrinello-Rahman barostat. Lastly, MD simulations were carried out for 50 ns in triplicates under the five conditions (pH 7.0 and 300 K; pH 4.5 and 300 K; pH 3.5 and 300 K; pH 7.0 and 340 K; pH 7.0 and 380 K). Jobs were submitted to the UCL Legion High Performance Computing Facility. The time step of the simulations was set to 2 fs and trajectories were saved every 10 ps.

3.3.3 Analysis of MD trajectories

MD trajectories were saved reduced, every 0.2 ns (total of 250 frames). Interface contacts over simulation time were calculated using the native contacts extension of the visual molecular dynamics (VMD) program (Humphrey et al. 1996). A cutoff distance of 4 Å was used in the calculations. Variable domain contacts (V_L-V_H) were calculated between residues 1-108 (V_L) and 215-334 (V_H). Constant domain contacts (C_L-C_H1) were calculated between residues 109-214 (C_L) and 335-442 (C_H1). RMSD of individual domains during the simulations were calculated using the RMSD trajectory tool in VMD. All the structures of the trajectory were first aligned and the RMSD was calculated (no hydrogens included). Domains were V_L (1-108), V_H (215-334), C_L (109 to 214) and C_H1 (335 to 429). Averages and SEM of three independent repeats are shown. Structural alignments of the last 30 ns of the trajectories were also performed using VMD. Secondary structure (SS) assignments of each residue along the trajectory were done using the DSSP module (Touw et al. 2015; Kabsch & Sander 1983). To analyse the loss in β-strand structure, I monitored the percentage of β-sheet SS per residue. These values were summed for each of the 32 β-strands in Fab A33 and differences were calculated

between the unfolding simulations and the reference simulations (pH 7 and 300 K). Lastly, salt bridges were calculated along the trajectories using VMD and a cutoff distance between O and N groups of 3.2 Å. From these, the occurrence (%) of each salt bridge during the simulation was calculated, and averaged for the three independent repeats at each condition.

3.3.4 Aggregation-prone regions (APR) predictions

Aggregation prone regions (APR) of Fab A33 were predicted using PASTA 2.0 (Walsh et al. 2014), TANGO (Fernandez-Escamilla et al. 2004), AGGRESCAN (Conchillo-Solé et al. 2007) and MetAmyl (Emily et al. 2013), using the protein sequence as input. The regions in which three out of the four software identified an APR were selected, resulting in seven APRs (Tsolis et al. 2013). Amylpred2 consensus tool was used to confirm the presence of these APRs (Tsolis et al. 2013). A consensus was created between the four sequence-based software, (Normalized TANGO * 1/4 + Normalized PASTA 2.0 * 1/4 + Normalized AGGRESCAN * 1/4 + Normalized MetAmyl * 1/4), to visualize the aggregation propensity of each residue on Fab A33 structure. To calculate the solvent accessible surface area (SASA) of each APR during the trajectories, the average area per residue over the trajectory was calculated first, using Gromacs analysis tool “sasa”, then summed for each APR.

3.3.5 Mutational study and ΔΔG calculations by FoldX and Rosetta

The effect of mutations on the stability of Fab A33 was studied using FoldX (foldx.crg.es) (Zhang et al. 2012) and the Rosetta method “ddg_monomer” (www. rosettacommons.org) (Kellogg et al. 2011). Both tools predicted the difference in folding free energy, ΔΔG, between the protein carrying a point mutation and the wildtype. Each of the 442 residues in the Fab A33 were mutated to the other 19 possibilities, totalling 8398 single mutants. FoldX was used as a plugin in the graphical interface YASARA (van Durme et al. 2011). The “Repair” command was used first to energy minimize the homology model of Fab A33, by rearranging the amino acid side chains. Next, the “BuildModel” command was used to introduce the point mutations, optimize the structure of the new protein variant, and calculate the stability change upon mutation. Calculations

using the Rosetta "ddg_monomer" method were performed by Dr Cheng Zhang. An example of mutation and option files, listing the parameters of the executable, can be seen in previous work (Zhang et al. 2018). Jobs were submitted to the UCL Legion High-Performance Computing Facility.

3.3.6 Packing Density

Occluded surface (OS) program was used to calculate the atomic packing of Fab A33 (Pattabiraman et al. 1995; Fleming & Richards 2000). The occluded surface packing (OSP) values are useful for identifying regions of loose packing in a protein. OSP value for each residue are calculated from the collection of extended normals (ray-lengths) that extend outward from the molecular surface until they intersect neighbouring van der Waals surface. Analysis of these normals, their respective lengths and the surface area involved in the interaction, defines the packing of each atom in the protein.

3.3.7 Sequence Entropy of Fab sequences

Fab sequences were retrieved from the Protein Data Bank (PDB) (Rose et al. 2013), totalling one hundred light chains and one hundred heavy chains. For light chains, kappa (κ) and lambda (λ) chains were included. For κ light chains, λ light chains and heavy chains, sequences from the species human and mouse were used. Sequence alignment and calculation of the sequence entropy for each residue were calculated using Bioedit (Hall 1999); sequences were aligned using ClustalW within Bioedit. The maximum entropy for 21 possible amino acids (including stop codon) is 3.04 and zero represents a fully conserved residue.

3.4 Results and discussion

3.4.1 Interface contacts, RMSD of individual domains and structural alignments revealed different unfolding pathways at low pH and high temperature

To determine which domains of Fab A33 are more susceptible to unfolding under low pH and high temperature, I first followed the RMSD of each individual domain (V_L, V_H, C_L and C_H1) along the simulations, as changes in RMSD are indicative of a conformational change. Simulations in the unfolding trajectories (pH 3.5 and 4.5 at 300 K, for low pH; pH 7.0 at 340 K and 380 K, for high temperature) are compared to the simulations in the native trajectory (pH 7.0 at 300 K). For every condition of pH and temperature, three independent simulations were performed, and their average RMSD and SEM are shown here. Additionally, structures from the unfolding trajectories (pH 3.5, for low pH; 380 K, for high temperature) are aligned to structures from the native trajectory (pH 7.0 at 300 K), to visualize the structural changes that induvial domains experienced. For each domain alignment, ten structures were taken every 3 ns from each simulation repeat, from the 20-50 ns range at which the RMSD had stabilized. Thus, a total of thirty structures from each stress condition were compared to thirty from the native trajectory. I also monitored the number of interface contacts between domains (V_L-V_H and C_L-C_H1) during the simulations using a cutoff of 4 Å, to understand the temporal relationship between breakage of contacts in each interface, and the unfolding of each domain.

First, the effect of low pH upon Fab A33 structure was considered (Figure 3.2). Regarding the number of interfacial contacts, almost no change was observed in the variable region (V_L-V_H) between pH 7.0, maintaining 333 ± 24 contacts (discarding the first frame), and pH 3.5, maintaining 309 ± 24 contacts (Figure 3.2a). By contrast, a loss of interfacial contacts in the constant region (C_L-C_H1) was observed between pH 7.0, 335 ± 17 contacts, and pH 3.5, 265 ± 12 contacts (Figure 3.2b). Interestingly, this loss of constant region interfacial contacts at low pH takes place very quickly, with pH 7.0 retaining 384 ± 14 contacts after 5 ns of the simulation, while simulations at pH 3.5 only retained 270 ± 11 contacts. This could be attributed to the lack of a well-defined hydrophobic core in the C_L-C_H1 interface, resulting in numerous early-disrupted contacts. Notably, C_L was the only domain to show a noticeable conformational change at low pH,

revealed as an increase in RMSD from 1.8 ± 0.1 Å at pH 7.0 (calculated between 20-50 ns of the simulation), to 2.2 ± 0.1 Å at pH 3.5 (Figure 3.2e). This domain displacement occurred in the first 20 ns, after many interface contacts had already been lost with respect to pH 7.0, which suggests that destabilization of the C_L-C_H1 interface preceded and potentially accelerated the unfolding of the C_L domain. The other domains (V_L, V_H and C_H1) did not unfold significantly during the low pH simulations (Figure 3.2c,g,i). Structural alignments confirm this result, showing remarkable good alignments between structures of V_L, V_H and C_H1 at pH 7.0 and 3.5 (Figure 3.2d,h,j). Alignments of the CL domain at pH 7.0 and pH 3.5 revealed a slight displacement at low pH, especially visible in the loop regions (Fig 2F). These findings agreed with previous experimental work, which combined SAXS, atomistic modelling and smFRET to reveal the displacement of the C_L domain in Fab A33 at low pH (Codina et al. 2019).

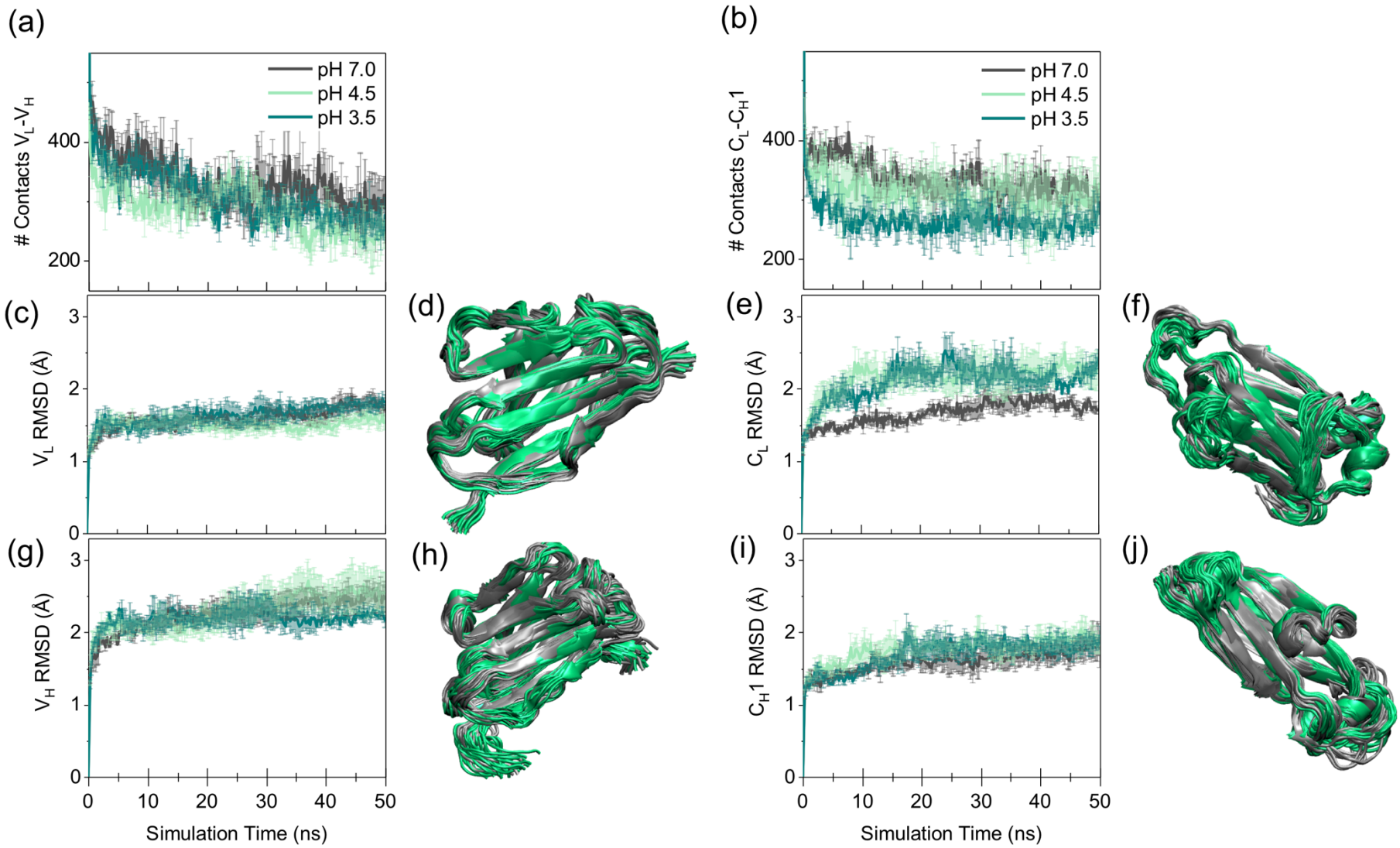

Figure 3.2. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0, 4.5 and 3.5 (all 300 K).

Figure 3.2. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0, 4.5 and 3.5 (all 300 K). (a,b) Contacts between light and heavy chains within 4.0 Å with simulation time, for variable (V_L-V_H) and constant (C_L-C_H) regions, respectively, pH values as labelled. (c, e, g, i) RMSD of individual domains with simulation time for V_L, V_H, C_L and C_H1, respectively, pH values as labelled. In all cases, the average of three independent simulations is shown with the SEM. (d, f, h, j) Alignments of structures from simulations at pH 7.0 and 3.5 for V_L, V_H, C_L and C_H1, respectively. Ten structures from the last 30 ns of each simulation were used, totalling thirty structures from pH 7.0 and thirty from pH 3.5.

Next, the effect of high temperature on Fab A33 was studied (Figure 3.3). MD simulations are commonly run at temperatures as high as 500 K to attempt to fully denature the protein. Here, I aimed to capture the early thermal unfolding events of Fab A33, which involve only partial unfolding of the protein. For this reason, and to reflect experimental conditions more closely, lower temperatures of 340 K and 380 K were used in the simulations. Interfacial contacts were found to break across both the variable and the constant regions, with high temperature (Figure 3.3a,b). At 380 K, contacts in the variable interface only averaged 220 ± 24 contacts and in the constant interface 204 ± 13 contacts. Contacts in the constant domains were found to break earlier than the variable domains, with only 218 ± 14 present after 5 ns of the simulations at 380 K. This is consistent with previous reports, which also found the constant region interface lose a larger fraction of its total interface contacts consistently faster than the variable region interface at high temperature, and also that domain unfolding occurred later than the loss of interfacial contacts. Overall, more contacts were broken at both interfaces with high temperature than with low pH (Buck et al. 2013). While V_L and C_H1 experienced only small domain displacements (Figure 3.3c,i), clear domain unfolding was observed for C_L and V_H (Figure 3.3e,g). At 380 K from 20 to 50 ns, the V_H domain displayed an increase in the RMSD from 2.4 Å at pH 7.0 to 3.2 Å at pH 3.5, and C_L from 1.8 Å at pH 7.0 to 2.4 Å at pH 3.5 (all average ± 0.1 Å). In these cases, many interface contacts were also lost with respect to pH 7.0 and 300 K, before the unfolding of individual structural domains, again consistent with destabilization of the interface contributing to the loss of stability of the individual domains. For both V_L and C_H1 structures from the simulations at 300 K and 380 K aligned well (Figure 3.3d,j), whereas for the V_H and C_L (Figure 3.3h,f) the

domains were structurally perturbed at the higher temperature. The V_H domain experienced a displacement of the loops on the N-terminal region, including the three CDR loops (Figure 3.3h). Differences in the C_L domain at high temperature were found in the loops and within an internal β-strand (Figure 3.3f). This was consistent with previous work which identified instability and structural changes in the V_H domain of another Fab at high temperature (Buck et al. 2013). Taken together, these findings suggest a different unfolding pathway for Fab A33 at low pH and at high temperature.

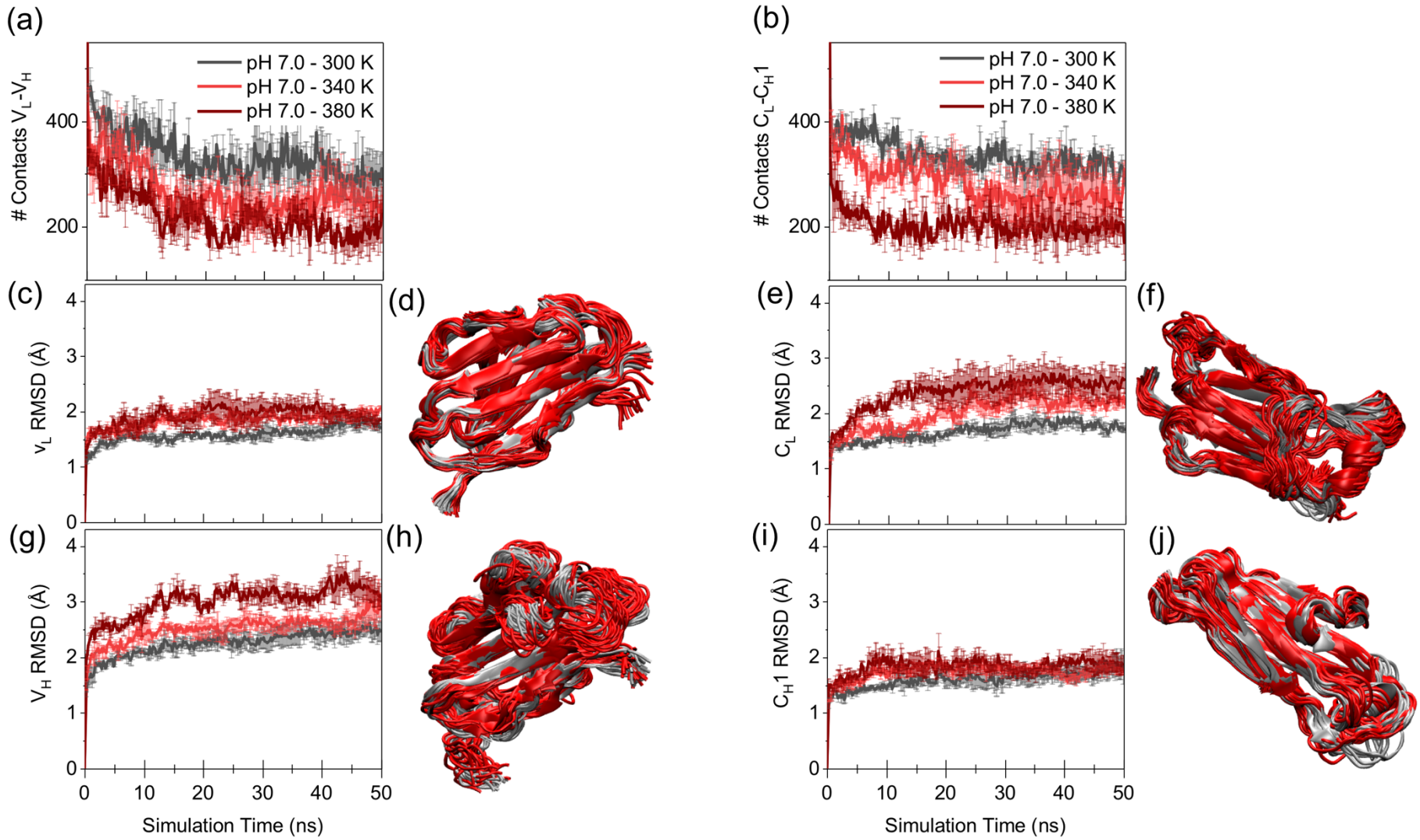

Figure 3.3. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0 and temperatures 300 K, 340 K and 380 K.

Figure 3.3. Interface contacts, RMSD of individual domains and structural alignments for simulations at pH 7.0 and temperatures 300 K, 340 K and 380 K. (a,b) Contacts between light and heavy chains within 4.0 Å with simulation time, for variable (V_L-V_H) and constant (C_L-C_H) regions, respectively, temperature values as labelled. (c, e, g, i) RMSD of individual domains with simulation time for V_L, V_H, C_L and C_H1, respectively, temperature values as labelled. In all cases, the average of three independent simulations is shown with the SEM as error. (d, f, h, j) Alignments of structures from simulations at temperatures of 300 K and 380 K for V_L, V_H, C_L and C_H1, respectively. Ten structures from the last 30 ns of each simulation were used, totalling thirty structures from 300 K and thirty from 380 K.

3.4.2 Loss in β-strand secondary structure confirms regions of unfolding

The unfolding of individual domains was also followed by their loss in secondary structure (SS); specifically, I monitored the change in β-strand structure. Fab domains have a β-barrel structure with two layers of β-sheets, an inner β-sheet and an outer β-sheet, each composed of several β-strands. Constant domains are composed of seven β-strands named A to G, while variable domains contain two more strands, a total of nine, with the two additional strands termed C' and C" (Figure 3.4a). To calculate the loss in β-strand structure for each of the strands, the secondary structure of each residue in Fab A33 was followed first during the simulations (Figure 3.5). From this, the percentage of β-strand per residue was calculated, and summed for each of the 32 β-strands in Fab A33. This value was averaged for each of the three repeats at each condition. Lastly, a percentage change in β-strand SS was calculated, to analyze the loss with regards to the reference simulations (pH 7 and 300 K).

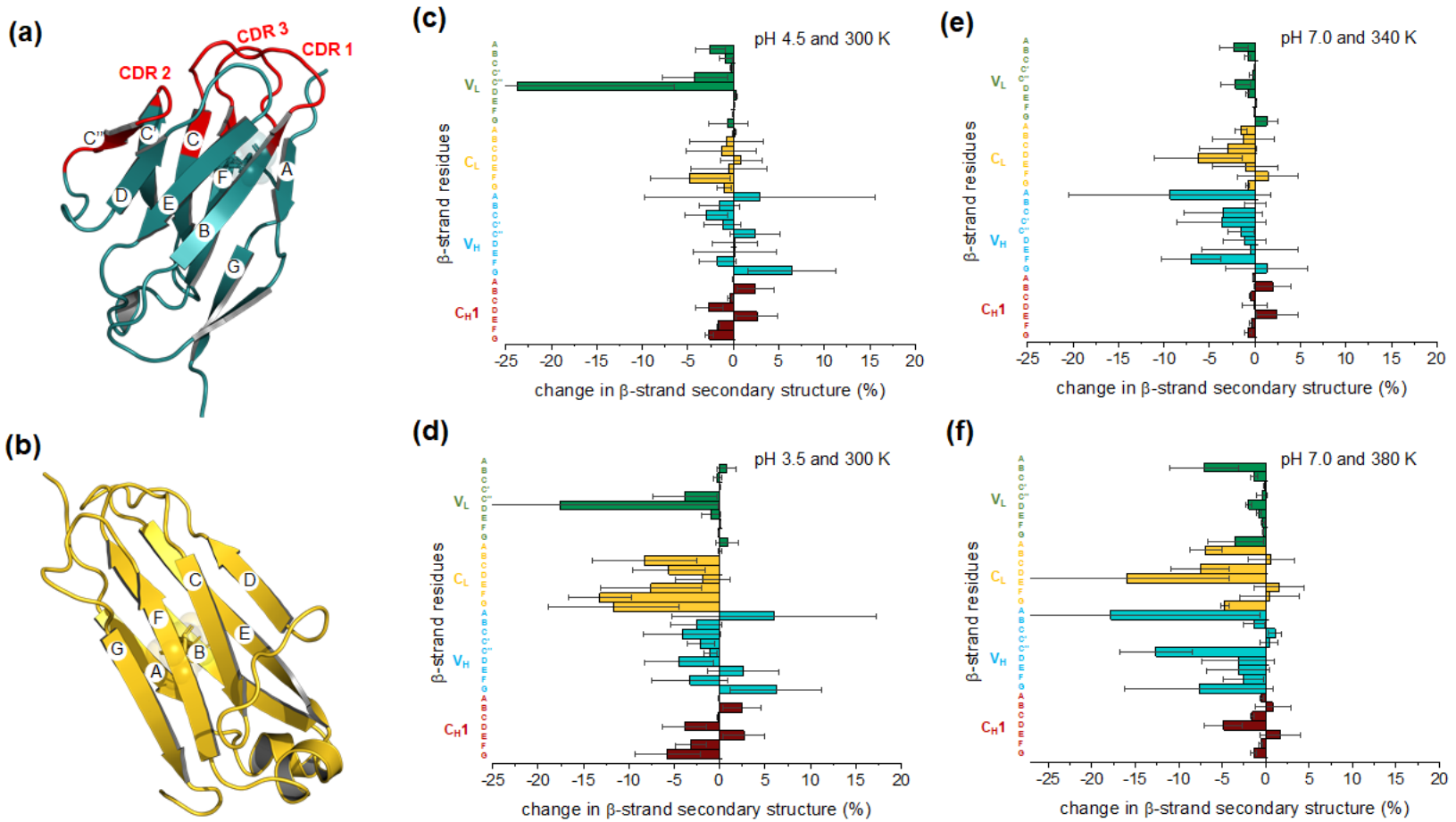

Figure 3.4. Loss of secondary structure for each of the 32 β-strands of Fab A33. (a,b) Strand order shown by lettering (A-G) for variable and constant domains, respectively. (c, d, e, f) Percentage increase/decrease in β-strand secondary structure for each strand in Fab during the simulations, respect to pH 7.0 and 300 K, for (c) pH 4.5 and 300 K, (d) pH 3.5 and 300 K, (e) pH 7.0 and 340 K, and (f) pH 7.0 and 380 K. Error bars are the same and equal for positive and negative values.

At pH 4.5, there are no significant losses in β-strand SS in any of the domains, to the exception of the β-strand C" of the V_L domain (Figure 3.4c). This strand also showed a big error, representative of high variability between repeats. C" is the shortest strand, and is located at the extreme of the outer β-sheet connecting the CDR-2 loop; which suggests this is a flexible region and it might have lost its SS in some simulations. At pH 3.5, the C_L domain had an overall loss in secondary structure content, confirming the results found in the previous section (Figure 3.4d). Strands F (-13 ± 3 %) and G (-12 ± 7 %) of the C_L domain showed the highest β-sheet structure loss, both located in the outer β-sheet. Strands B (-8 ± 6 %), C (-6 ± 4 %) and E (-8 ± 6 %) also experienced significant losses.

At pH 7.0 and 340 K, the losses in β-strand SS are small, however, the regions more likely to destabilize as the temperature increases, can start to be identified. Strand D (-6 ± 3 %) in the C_L domain and F (-7 ± 4 %) in the V_H domain, displayed significant losses. At 380 K, these losses were more noticeable, and located in the C_L and V_H domains, consistent with the unfolding described in the previous section. Many strands in C_L domain show significant losses, A (-7 ± 2 %), C (-8 ± 3 %), D (-16 ± 12 %) and G (-5 ± 1 %), located at the extremes of the inner and outer β-sheets. The V_H domain also showed high losses of β-strand SS. However, of these strands A (-18 ± 17 %) and G (-8 ± 9 %) also showed high variability between repeats. Interestingly, these same two strands in V_H were previously to deform at high temperature in a different Fab (Buck et al. 2013). Strand C" (-13 ± 4 %) of the V_H domain also showed a significant loss of β-strand content.

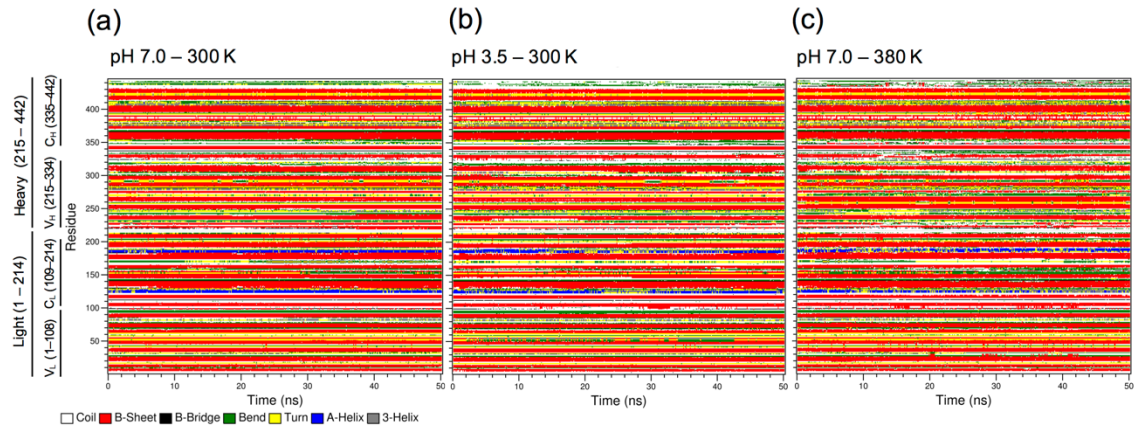

Figure 3.5. Secondary structure (SS) of each residue in Fab A33 with simulation time, calculated using DSSP. Representative SS evolution are shown for (a) pH 7.0 and 300K, (b) pH 3.5 and 300K and (c) pH 7.0 and 380 K, secondary structure type as indicated in the legend.

3.4.3 Salt bridge analysis identifies key stabilizing salt bridges

To identify the ionisable residues that drive the conformational change at low pH, a salt bridge analysis was performed. Salt bridges were identified over the simulation time for all the MD simulations carried out, using an O-N bond distance cutoff of 3.2 Å. From these, the occurrence (%) of each salt bridge during the simulation was calculated, and averaged for the three independent repeats at each condition. Lastly, the most persistent salt bridges at each condition were highlighted in the Fab A33 structure (Figure 3.6).

At pH 7.0 and 300 K, a total of 36 salt bridges were present. Interestingly, many of these salt bridges were flexible and able to form with different partners during a single trajectory, such as Asp122 which partnered with both Lys126 and Lys436, or Asp304 which paired with both Arg252 and Arg281. This is consistent with previous work, which found that salt bridges break and reform, and not always with the same partner (Kortkhonjia et al. 2013). The most persistent (as % of time present) salt bridges at pH 7.0 were Glu165-Lys103 (67 ± 4 %), Glu195-Lys149 (65 ± 4 %), Asp82-Arg61 (61 ± 14 %), Asp151-His189 (58 ± 24 %), and Asp304-Arg252 (55 ± 27 %) (Figure 3.6a,d,g). At low pH, pH 3.5 and 300 K, a total of 27 salt bridges were observed, but most of them were very short lived. The more persistent salt bridges at pH 3.5 were Asp151-His189 (86 ± 6 %), Asp82-Arg61 (78 ± 11 %), Asp122-Lys126 (73 ± 1 %), Asp362-Lys335 (56

± 6 %) (Figure 3.6b,e,h). The protonation state at the end of the pH 3.5 simulations was calculated again using these Fab conformations, which revealed these salt bridges to be still present due to predicted pKa values for these aspartates of below 3.5. Comparison of the salt bridges at pH 7.0 and 3.5, indicated the presence of two salt bridges that potentially trigger the conformational change observed at low pH, and thus the loss of Fab A33 stability. Glu165-Lys103 and Glu195-Lys149, were the most persistent contacts at pH 7.0, but were not present at pH 3.5. Glu165-Lys103 bridges the C_L domain to the V_L domain, and Glu195-Lys149 bridges the outer β-strands C and F of the C_L domain. Loss of these salt bridges at low pH, would therefore destabilize the C_L domain, and promote the observed C_L domain displacement.

At high temperature, pH 7.0 and 380 K, a total of 45 salt bridges were observed. The greater number than at 300 K, reflects an increased conformational flexibility of many salt bridges at higher temperature, in which they often broke, but then reformed with a different partner. Indeed, at the high temperature, salt bridges broke and reformed much faster (Figure 3.6c). At 380 K, the total time present for the most persistent salt bridges observed at 300 K, had decreased to 47 ± 9 % for Glu165–Lys103, 58 ± 3 % for Glu195-Lys149, 21 ± 7 % for Asp151-His189, and 43 ± 22 % for Asp304-Arg252. However, Asp82-Arg61 increased in occurrence to 70 ± 3 % (Figure 3.6c,f,i). These findings indicate that the increased dynamics at high temperature, results in constant rupture and formation of salt bridges, and this transient disruption leaves Fab A33 more susceptible to unfolding.

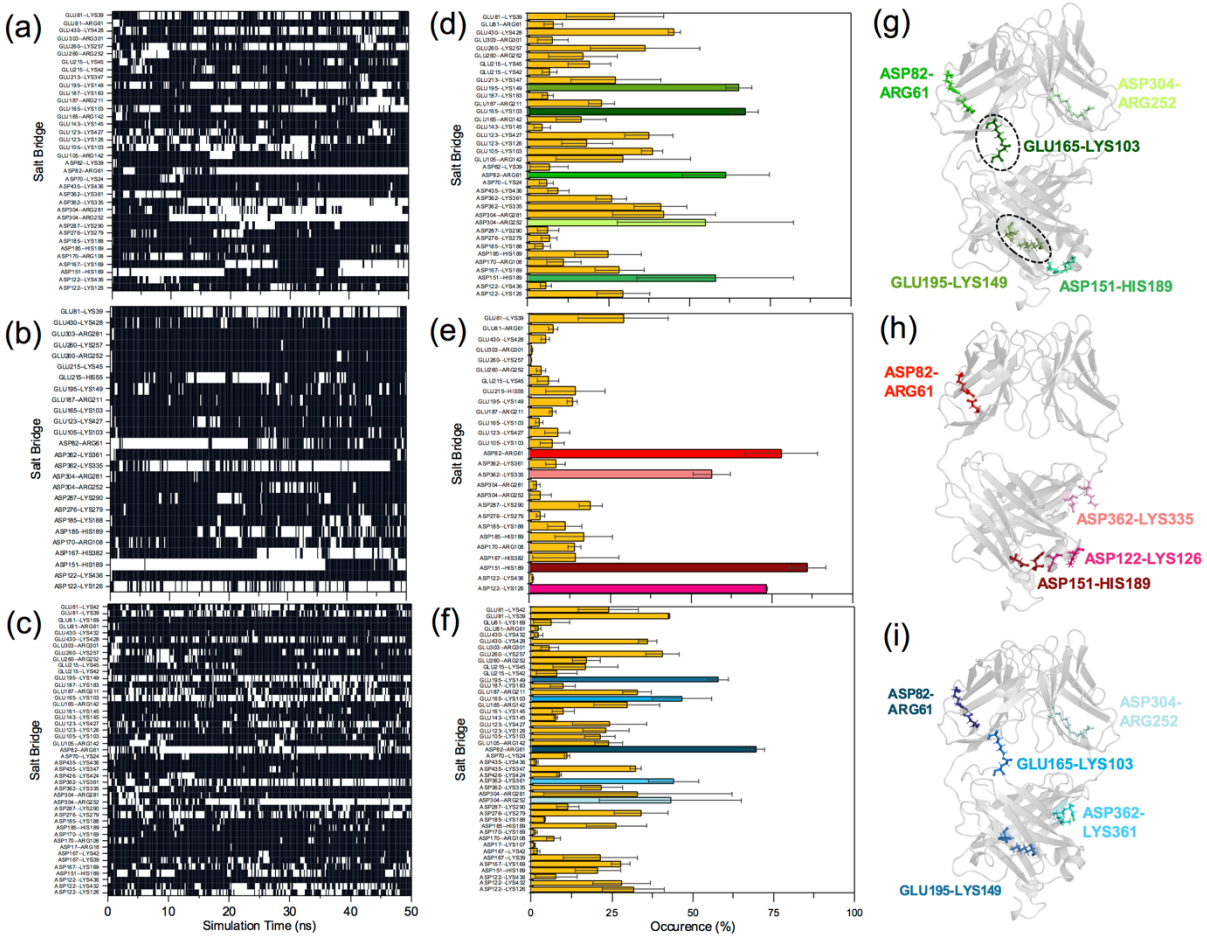

Figure 3.6. Salt bridge analysis. (a, b, c) Salt bridges formed during the simulation time for representative MD simulations at (a) pH 7.0 and 300 K, (b) pH 3.5 and 300 K and (c) pH 7.0 and 380 K. Presence of a salt bridge is indicated in white and absence in black. (d, e, f) List of salt bridges and its occurrence (%) for simulations at (d) pH 7.0 and 300 K, (e) pH 3.5 and 300 K and (f) pH 7.0 and 380 K. Values shown are the average of three independent simulations with their SEM as error. The more persistent salt bridges are highlighted for pH 7.0 (green), pH 3.5 (red) and pH 7.0 and 380 K (blue). (g, h, i) The more persistent salt bridges are mapped into the Fab A33 structure. Two key stabilising salt bridges (Glu165-Lys103 and Glu195-Lys149) are highlighted in a dashed circle.

3.4.4 Solvent exposure of different aggregation-prone regions promotes different aggregation pathways for low pH and high temperature

The aggregation pathways of Fab A33 at low pH and high temperature at pH 7.0, are already known to result in different aggregate morphologies (Chakroun et al. 2016). Here I explored whether the two conditions also exposed different aggregation-prone regions (APRs). Computational biology tools were used to predict the regions in proteins most likely to form and stabilize the cross-β structure characteristic of aggregates. These aggregation-prone regions (APRs) are mostly hydrophobic, possess a low net charge, and have a high propensity to form β-sheets. Several methods have been developed to predict the presence of APRs in a protein. The first methods only used the protein sequence as input, this being equivalent to the fully unfolded state. Predictions were based on either the intrinsic properties of amino acids, or their compatibility with protein structural features in known amyloid fibril structures. Examples include TANGO (Fernandez-Escamilla et al. 2004), AGGRESCAN (Conchillo-Solé et al. 2007), PASTA (Walsh et al. 2014), MetAmyl (Emily et al. 2013), FoldAmyloid (Garbuzynskiy et al. 2010), FishAmyloid (Gasior & Kotulska 2014) and Waltz (Maurer-Stroh et al. 2010). As these predictions do not always agree, Amylpred2 generates a consensus from up to eleven existing algorithms (Tsolis et al. 2013). However, it is known that APRs are frequently buried inside the hydrophobic core of globular proteins, and so their ability to trigger aggregation would depend upon solvent accessibility, i.e. the potential of the APR to become solvent exposed through structural dynamics or partial unfolding. Thus, more recent methods include aspects of the protein structure to predict APRs, including AGGRESCAN 3D (Zambrano et al. 2015), AggScore (Sankar et al. 2018), SAP (Chennamsetty et al. 2009) and Solubis (Van Durme et al. 2016).

Here, I want to compare the solvent accessibility of APRs in Fab A33, between the MD simulations at the unfolding conditions and at the reference trajectory. Thus, I used sequence-based APR predictors to determine the APRs in Fab A33, and determined their solvent accessible surface area (SASA) in the simulations, for relative comparisons. Four sequence-based APR predictors were used in total, PASTA 2.0, TANGO, AGGRESCAN and MetAmyl, to predict the APRs in Fab A33. APRs, where three out of the four predictors identified an aggregation-prone sequence and were selected (Figure 3.7a). Seven segments showed the highest aggregation propensity values, namely residues 31-36, 47-51, 114-118 and 129-139 in the light chain and residues 261-165, 325-

329 and 387-402 in the heavy chain. Additionally, these APRs were confirmed using Amylpred2, which identified the same APRs in addition to others (Figure 3.7a).

To display the aggregation propensity of every residue in Fab A33, a consensus was created between the four sequence-based methods. Each aggregation propensity was normalized between 0 and 1, and weighted equally (Figure 3.7b). The consensus aggregation propensities were mapped into the Fab A33 homology as shown in Figure 3.8. Red represented high aggregation propensities and blue low aggregation propensities. The seven APRs were co-located as three regions of largely buried β-strands within the folded structure, and all were protected from the solvent.

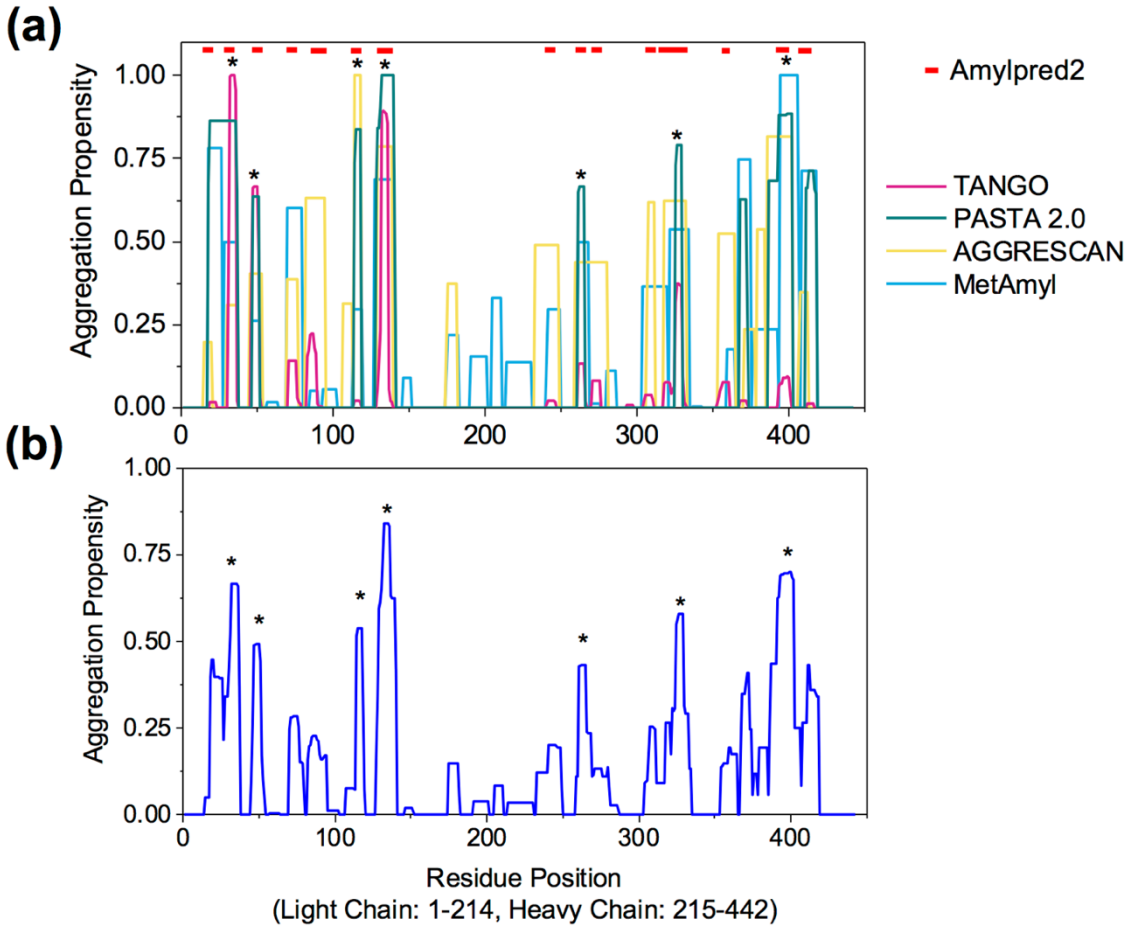

Figure 3.7. Prediction of aggregation-prone regions (APR) in Fab A33 using sequence-based predictors. (a) The aggregation propensity for each residue in Fab A33 was predicted using PASTA 2.0, TANGO, AGGRESCAN and MetAmyl algorithms. These are colour-coded as shown. The sequence regions in which three out of the four predictors agreed, were selected and highlighted with asterisks. Additionally, fifteen

APRs were predicted with the consensus software Amylpred2, shown as red horizontal lines on top of the graph. (b) Consensus score for the four algorithms, where the four scores were each normalised, then summed with equal weights. The asterisks highlight the seven most aggregation-prone regions in Fab A33 in the light chain (residues 31-36, 47-51, 114-118 and 129-139) and in the heavy chain (residues 261-165, 325-329 and 387-402).

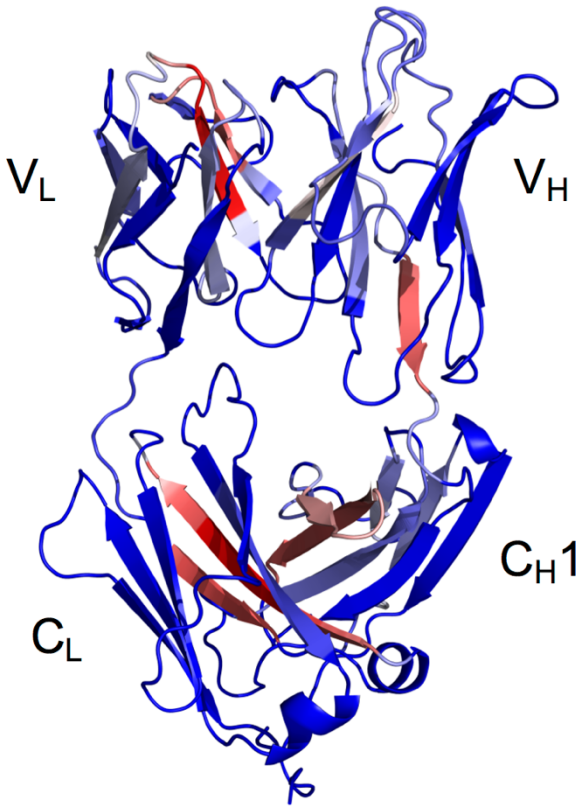

Figure 3.8. Aggregation prone regions in Fab A33. The consensus aggregation propensity values of residues in Fab A33 were added as B-factors to the PDB file for the Fab A33 homology model. Regions with greater aggregation propensities are shown in red and reduced propensities in blue.

Exposure of one of these APRs as a result of a conformational change by an environmental stress, has the potential to trigger aggregation. Thus, the SASA of each APR during the simulations was calculated, as well as the difference in solvent accessibility, ΔSASA, between unfolding conditions and the reference simulation (Table 3.2).

Table 3.2 SASA of the APRs in Fab A33 during simulations and SASA differences between unfolding simulations and the reference simulation. Solvent accessible surface area of the seven aggregation-prone regions in Fab A33 during all simulations, and relative differences (ΔSASA) between the unfolding trajectories (pH 3.5 and 4.5 at 300 K, for low pH; pH 7.0 at 340 K and 380 K, for high temperature) and the reference trajectory (pH 7.0 and 300K).

APR	Fab domain	SASA ($Å^2$) pH 7.0 300K	SASA ($Å^2$) pH 4.5 300K	ΔSASA ($Å^2$) pH(4.5-7.0)	SASA ($Å^2$) pH 3.5 300K	ΔSASA ($Å^2$) pH(3.5-7.0)	SASA ($Å^2$) pH 7.0 340K	ΔSASA ($Å^2$) T(340K-300K)	SASA ($Å^2$) pH 7.0 380K	ΔSASA ($Å^2$) T(380K-300K)
31-36	V_L	118 ± 4	112 ± 5	-5 ± 7	115 ± 16	-3 ± 17	120 ± 8	3 ± 9	106 ± 4	-12 ± 6
47-51	V_L	100 ± 1	104 ± 7	4 ± 13	115 ± 20	15 ± 23	101 ± 6	1 ± 13	96 ± 7	-4 ± 13
114-118	C_L	125 ± 4	121 ± 7	-4 ± 8	110 ± 7	-15 ± 8	112 ± 2	-13 ± 5	120 ± 13	-5 ± 13
129-139	C_L	152 ± 8	147 ± 11	-5 ± 13	151 ± 12	0 ± 14	165 ± 10	13 ± 13	172 ± 14	20 ± 16
261-265	V_H	14 ± 2	13 ± 0	-1 ± 2	13 ± 1	-1 ± 3	19 ± 1	5 ± 3	29 ± 5	15 ± 6
325-329	V_H	122 ± 16	115 ± 14	-7 ± 21	119 ± 7	-3 ± 17	88 ± 10	-34 ± 19	120 ± 22	-1 ± 27
387-402	C_H1	552 ± 21	547 ± 25	-5 ± 33	609 ± 12	57 ± 25	544 ± 7	-9 ± 22	489 ± 48	-15 ± 21

At low pH, only one APR (residues 387-402), was found to increase its solvent accessibility significantly at pH 3.5, with an increase of 57 ± 25 Å^2 (10 % increase), (Table 3.2). This APR is located in the C_H1 domain and its exposure can be explained by the C_L domain displacement observed at low pH (Figure 3.9a). At high temperature, two APRs were found to increase their solvent accessibility, APR 261-265 located in V_H and APR 129-139 located in C_L (Figure 3.9b). APR 261-265 increased its SASA 15 ± 6 Å^2 (107 % increase) and APR 129-139 increased its SASA 20 ± 16 Å^2 (13 % increase). The location of these APRs agrees with the domains found to unfold previously at high temperature. Notably, the APRs exposed at low pH and high temperature are different, suggesting the potential to follow different aggregation mechanisms depending on the stress applied.

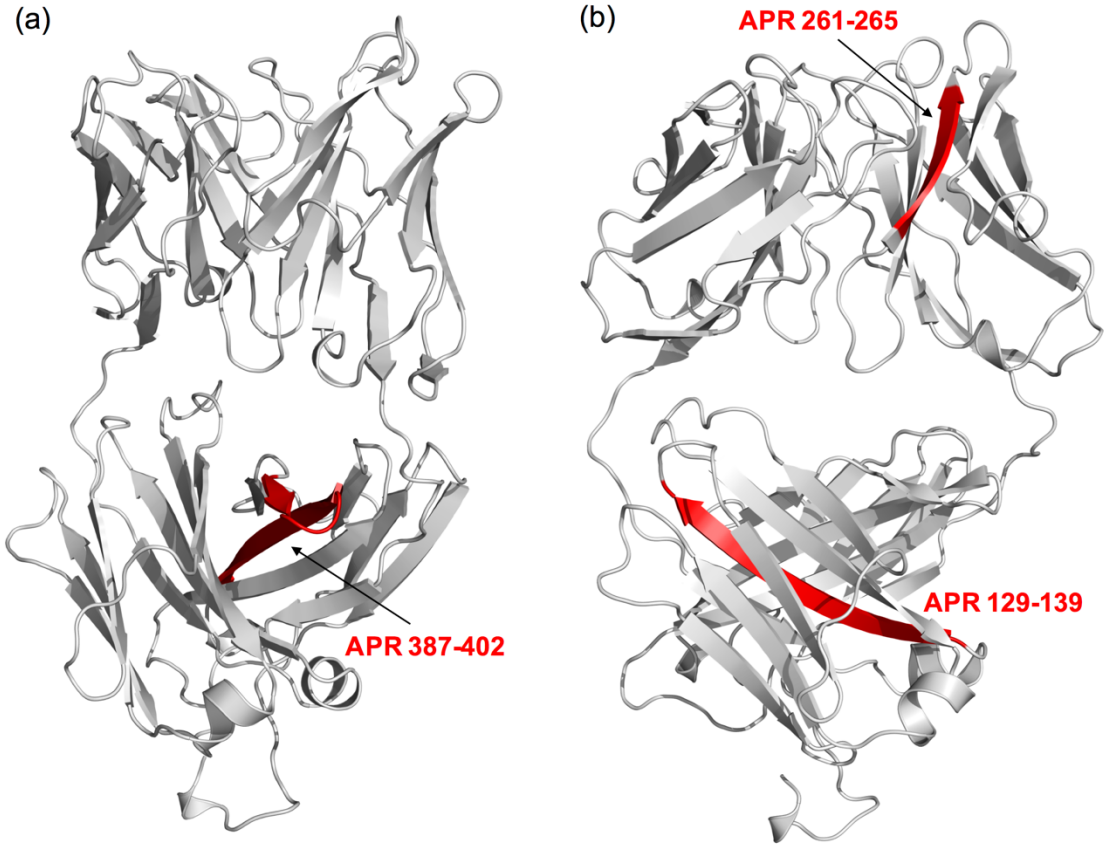

Figure 3.9. Fab A33 predicted APRs that increase its solvent accessibility at low pH and high temperature. (a) APR 387-402 increases its SASA at pH 3.5 and (b) APR 261-265 and APR 129-139 increase its SASA at 380 K (Table 3.2). All mapped in red in Fab A33 homology structure.

3.4.5 FoldX, Rosetta and packing density calculations predict sub-optimal stability of C_L and the C_L-C_H1 interface

Computational tools such as FoldX and Rosetta-ddG (Zhang et al. 2012; Kellogg et al. 2011) predict the relative changes in folding free energy (ΔΔG) between the Gibbs free energies (ΔG) of the wild-type protein and the protein carrying a simulated point mutation, to find those mutations that will most significantly reduce the free energy of the protein. These approaches are often also combined to find consensus predictions (Wijma et al. 2014; Buß et al. 2018). To predict stabilizing mutations in Fab A33, we calculated the ΔΔG with both FoldX and Rosetta-ddG, of all possible single-mutant variants when accessing all 19 other substitutions across the 442 residue positions in Fab A33, totalling 8398 mutations, using the Fab A33 homology model. FoldX identified 1879 of these mutations as stabilizing (22.4 %), while Rosetta-ddG identified 2386 (28.4 %). Stable mutations predicted by both software were 956 (11.4 %), this corresponds to 51% of the mutations identified by FoldX and 40% of the mutations identified by Rosetta. Figure 3.10 shows the correlation between the mutations predicted by FoldX and Rosetta, and Table 3.3 lists the 25 most stabilizing mutations predicted by both algorithms, with their respective ΔΔG values. FoldX reports ΔΔG values in kcal/mol and Rosetta in Rosetta Energy Unit (REU).

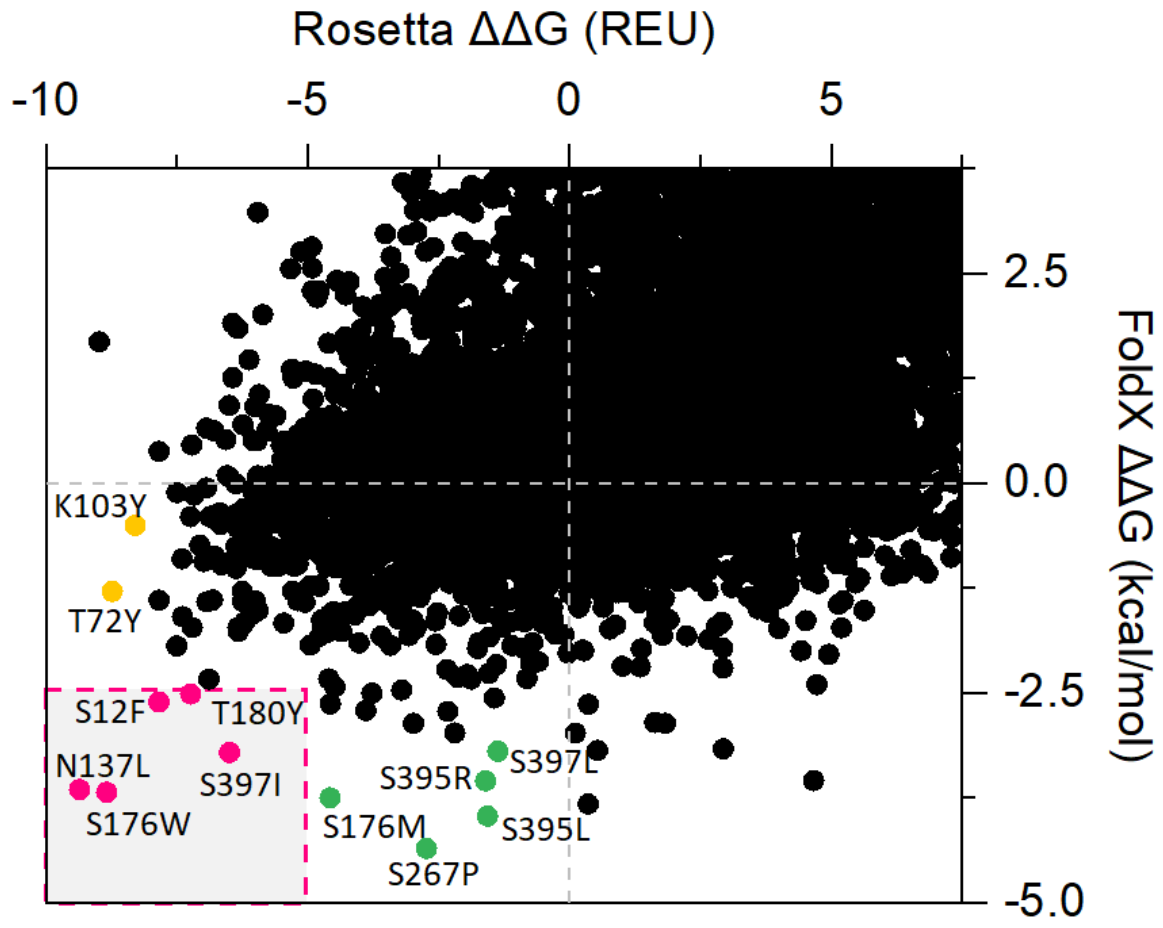

Figure 3.10. Stabilizing mutations predicted by FoldX and Rosetta. Correlation between FoldX and Rosetta predictions. Mutations predicted by both software to be most stabilizing are shown in magenta and highlighted in a gray square on the bottom left. Mutations predicted only by FoldX to be stabilizing are shown in green and mutations predicted only by Rosetta in yellow.

Table 3.3 List of the most stabilizing mutations identified by FoldX and Rosetta-ddG. Mutation and ΔΔG of the 25 most stabilizing mutations predicted by FoldX and Rosetta.

FoldX Mutation	FoldX ΔΔG (kcal/mol)	Rosetta Mutation	Rosetta ΔΔG (REU)
S395L	-4.35	N137L	-9.36
S267P	-3.97	L275H	-8.98
S395M	-3.82	S176W	-8.84
S176M	-3.75	T72Y	-8.73
S176W	-3.68	K103Y	-8.30
N137L	-3.65	T349W	-8.03
S395R	-3.55	S12F	-7.85
N137M	-3.54	S12Q	-7.84
S397I	-3.21	V226Y	-7.84
S397L	-3.19	K103F	-7.50
S176R	-3.18	S203H	-7.50
S395I	-3.16	D426N	-7.40
A254P	-2.99	T180W	-7.40
H382F	-2.98	K103T	-7.24
G336P	-2.86	T180Y	-7.24
S159R	-2.86	S397W	-7.22
S176L	-2.85	S159F	-7.21
S176Y	-2.72	N415Y	-7.17
N137I	-2.71	D70I	-7.06
S397V	-2.63	D70F	-6.98
S12Y	-2.63	V316Y	-6.93
S12F	-2.61	S374H	-6.93
S171G	-2.56	S121P	-6.93
T180Y	-2.51	S345P	-6.89
S395V	-2.50	T283F	-6.87

Figure 3.11a compares the greatest stabilization predicted by FoldX and Rosetta, at each of the 442 residues in Fab A33, regardless of the specific mutation selected by each algorithm. The residues in which both software, FoldX and Rosetta-ddG, agree that could be stabilized the most, are located on the bottom-left of the graph, highlighted in magenta. These residues correspond to S176, N137, S397, S159, S12 and T180. All the predicted mutations are to more hydrophobic amino acids, such as Trp, Leu, Ile, Phe and Tyr (Figure 3.10 and Table 3.3). Four of these six mutations, (S176, N137, S397 and T180), are located in the constant domain interface, between C_L and C_H1 domains (Figure 3.11b). These findings suggest that there is room for further stabilization of the C_L-C_H1 interface. Furthermore, S159 is in the C_L domain, interacting with an outer β-strands, and S12 is in the V_L domain interacting with the C_L domain (Figure 3.11b). Thus overall, the C_L domain has a relatively high potential for stabilization, through repacking of the C_L-C_H1 interface, within the C_L domain itself, and also through improved interaction between C_L and V_L. This is consistent with the MD simulations which found the displacement of C_L away from the interface with C_H, and subsequent unfolding of the C_L domain, to be critical steps in early or partial unfolding.

Lastly, other mutations predicted only by FoldX are highlighted in green, which correspond to S395 and S267. S395 is also located in the C_L-C_H1 interface. S267 is in the V_H domain and belongs to CDR2, and so not a good candidate for general framework stabilization due to its role in antigen binding. Mutations predicted only by Rosetta-ddG are highlighted in yellow, these being K103 and T72. These were both located in the V_L domain, but K103 also interacts with the C_L domain, further suggesting that the interactions within and around the C_L domain are the least optimized for stability.

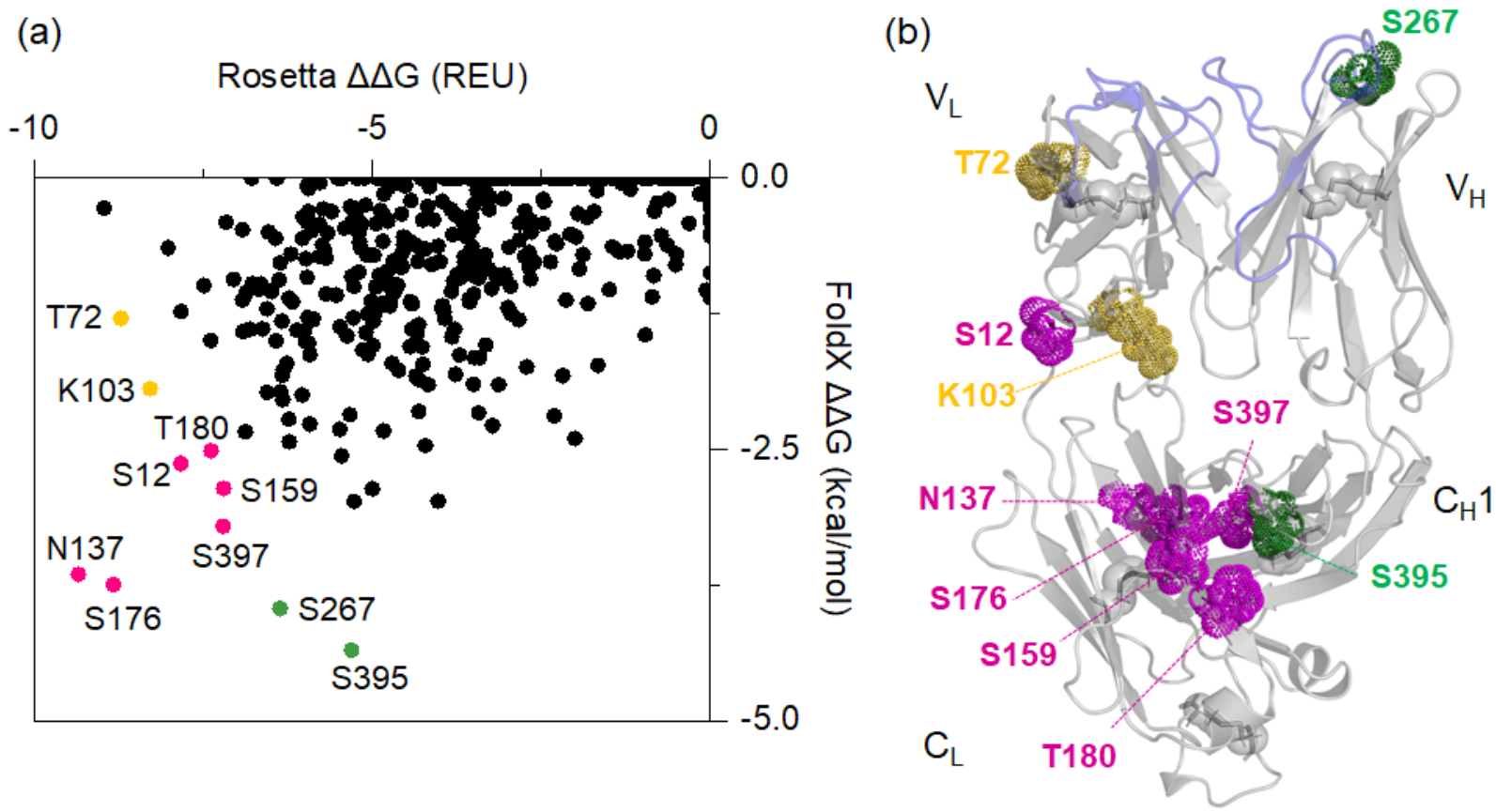

Figure 3.11. Predicted residues that can be stabilized further by FoldX and Rosetta-ddG. (a) Correlation between FoldX and Rosetta predictions. Residues predicted by both software to be most stabilizing are shown in magenta on the bottom left. Residues predicted only by FoldX to be stabilizing are shown in green and residues predicted only by Rosetta in yellow. (b) Residues predicted to be stabilized further the most are mapped in Fab A33 structure, following the same colour scheme as in (a).

The packing density of each residue in Fab A33 was calculated using the package occluded surface (OS) software, which calculates occluded surface and atomic packing (Pattabiraman et al. 1995; Fleming & Richards 2000). The occluded surface packing value of each atom is calculated from normal vectors that extend outward from the atom surface until they intersect a neighbouring van der Waals surface (Figure 3.12). This value is 0.0 for completely exposed residues and 1.0 where 100% of molecular surface is in contact with other van der Waals surface. Thus, the OSP value allows to identify regions of loose packing in the protein. The average OSP for all 28 β-strand residues within domain interfaces (V_L-V_H and C_L-C_H1) was 0.49 ± 0.01 (OSP values shown in Table 3.4). By contrast, the average OSP of the five constant-domain interface residues (S176, N137, S397, T180, and S395), identified by FoldX and Rosetta as having high stabilization potential, was 0.41 ± 0.05 (OSP values shown in Table 3.4). This can be visualized in Figure 3.13, where OSP values were mapped in the structure of Fab A33, with red to indicate high packing density, and blue to indicate low packing density. β-strand residues within domain interfaces were highlighted as sticks (Figure 3.13a). Residues in the constant interface (C_L-C_H1) were lighter colored than residues in the variable interface (V_L-V_H), indicating less tight packing of the constant interface. An insight of the residues identified by FoldX and Rosetta is provided in Figure 3.13b, where a lighter color than the residues in the variable interface was also observed. This result shows that the predicted residues are under-packed, and therefore have the potential to be mutated to pack the C_L-C_H1 interface more tightly.

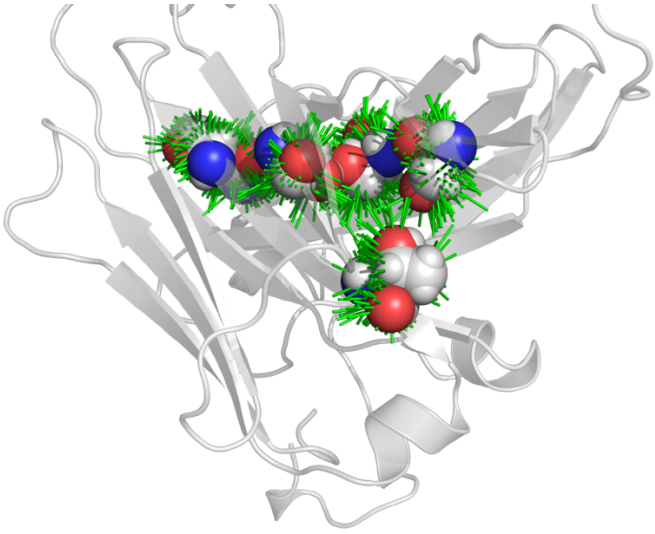

Figure 3.12. Normals used to calculate the packing of each atom in Fab A33 using Occluded Surface software. To calculate the occluded surface packing (OSP) value for each residue, normals that extend from the surface outward until they intersect a neighboring van der Waals surface are used. The normals used to calculate the OSP value of the inter-domain residues identified by FoldX and Rosetta (S176, N137, S397, T180, and S395), are shown.

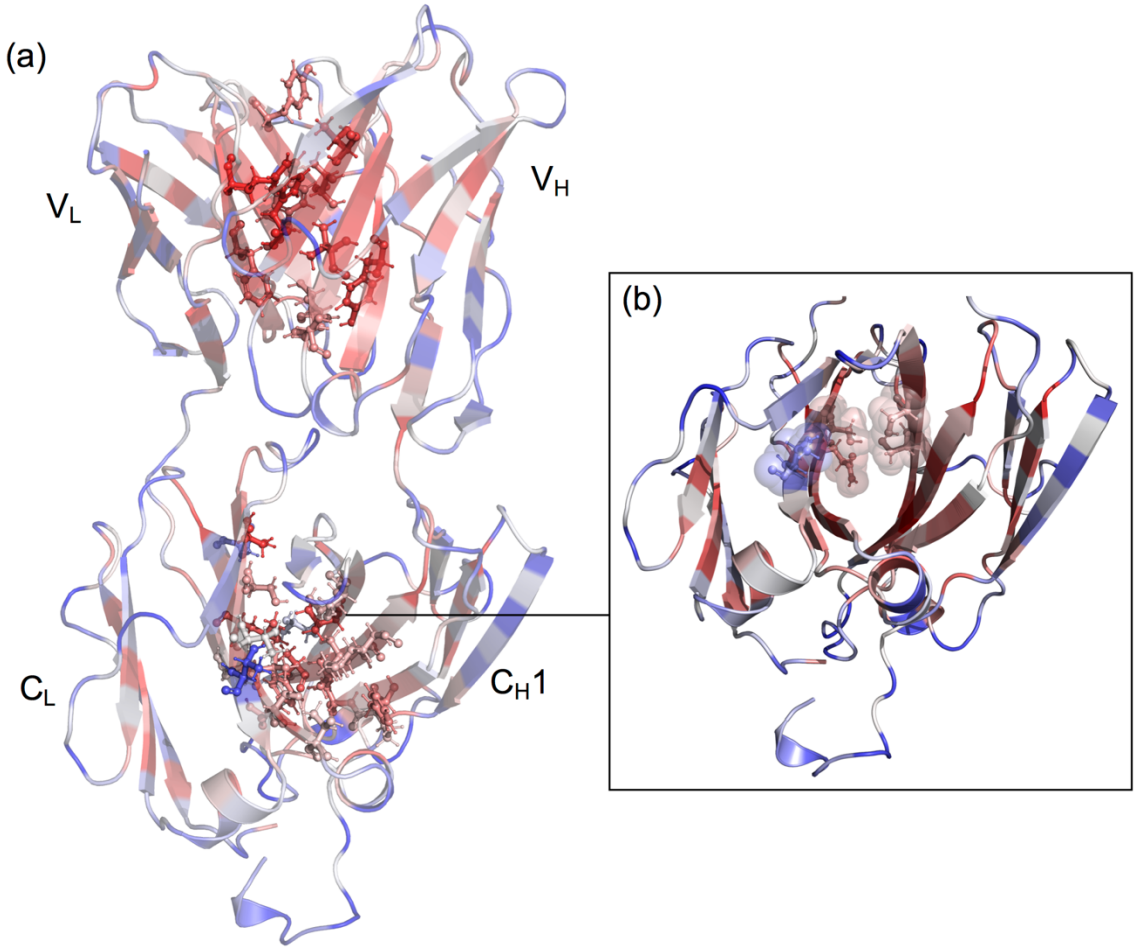

Figure 3.13. Packing density of every residue in Fab A33, computed using Occluded Surface. (a) The occluded surface packing (OSP) values were added as B-factors to the PDB file for the Fab A33 homology model. High packing values are shown in red and low values in blue. Residues in β-strands within domain interfaces (V_L-V_H and C_L-C_H1) are highlighted in sticks and ball representation. (b) Residues identified by FoldX and Rosetta that could be stabilised further (S176, N137, S397, T180, and S395) are highlighted in sticks and ball, and sphere representation.

Table 3.4. Packing indicated by the occluded surface packing (OSP) value of the residues located in β-strands within domain interfaces (V_L-V_H and C_L-C_H1) of Fab A33 homology model. OSP values were calculated using the occluded surface software.

Domain	Residue	OSP value
V_L	Y36	0.524
	Q38	0.468
	T46	0.518
	Y49	0.488
	F87	0.529
	L89	0.596
C_L	F116	0.408
	F118	0.494
	T129	0.425
	S131	0.437
	V133	0.469
	L135	0.535
	N137 *	0.486
	S174	0.576
	S176 *	0.464
	T178	0.384
	T180 *	0.201
V_H	V251	0.558
	Q253	0.477
	W261	0.598
	T264	0.571
	Y309	0.569
C_H1	F340	0.495
	P341	0.420
	L342	0.493
	A354	0.385
	A355	0.540
	L359	0.461
	K361	0.437
	S395 *	0.435
	S397 *	0.440
	V399	0.555
	T401	0.344

*, residues identified by FoldX and Rosetta in the constant domain interface that can be stabilized further.

3.4.6 Comparison to natural sequence variations in Fabs

The natural variability of Fab sequences was identified from within one hundred light chains and one hundred heavy chains curated from those available in the Protein Data Bank (Rose et al. 2013). Sequence alignment and sequence entropy calculations for each residue were obtained using Bioedit (Hall 1999). An entropy of zero indicates a fully conserved residue, whereas 3.04 is the maximum entropy, originating from 21 possibilities (all amino acids plus the stop codon). There is significant positional bias in the sequence variability of Fabs due to the hypervariability of the CDRs, the presence of kappa (κ) and lambda (λ) light chain isotypes, allotypic diversity across individuals, and idiotypic variability within the variable domains of individuals. The sequence entropy analysis (Figure 3.14) clearly shows this, with the highest sequence entropy (>2), for the six CDRs, and a slightly lower variability on average within the C_H domain. Similarly, the higher sequence entropy on average for the C_L domain, compared to the C_H domain results from the grouping of kappa (κ) and lambda (λ) light chain isotypes.

Even though framework residues have more restricted variability, many natural variations are observed that may affect stability. Except for the fully conserved S176, the sites predicted as having the most potential for stabilizing mutations by both FoldX and Rosetta, had natural variations, with sequence entropy values of N137: 0.74, S397: 0.15, S159: 0.96, S12: 0.96, T180: 0.43. This was also true for mutations identified only by FoldX (S267: 1.69 and S395: 0.69), and only by Rosetta-ddG (K103: 0.16, T72: 0.91). Comparisons between the existing mutations and the stabilizing mutations suggested by FoldX and Rosetta are shown in Table 3.5. In general, the mutations found in existing Fabs were conservative changes to residues with properties similar to the original residue. For instance, S397 and S395 are only naturally mutated to Thr, whereas T180 can also be Ser, and K103 can be Arg. By contrast, FoldX and Rosetta predictions were typically from polar to more hydrophobic residues, typically to Ile, Leu, Trp, Val, and Tyr. A few suggested mutations were also found naturally, such as N137I, S12Y, and S267P. S159 also shows the potential to be mutated to more hydrophobic residues, Val and Met. Overall, this analysis shows that despite their low natural variability, many residues in the constant domains had significant scope for stabilization through non-natural mutations.

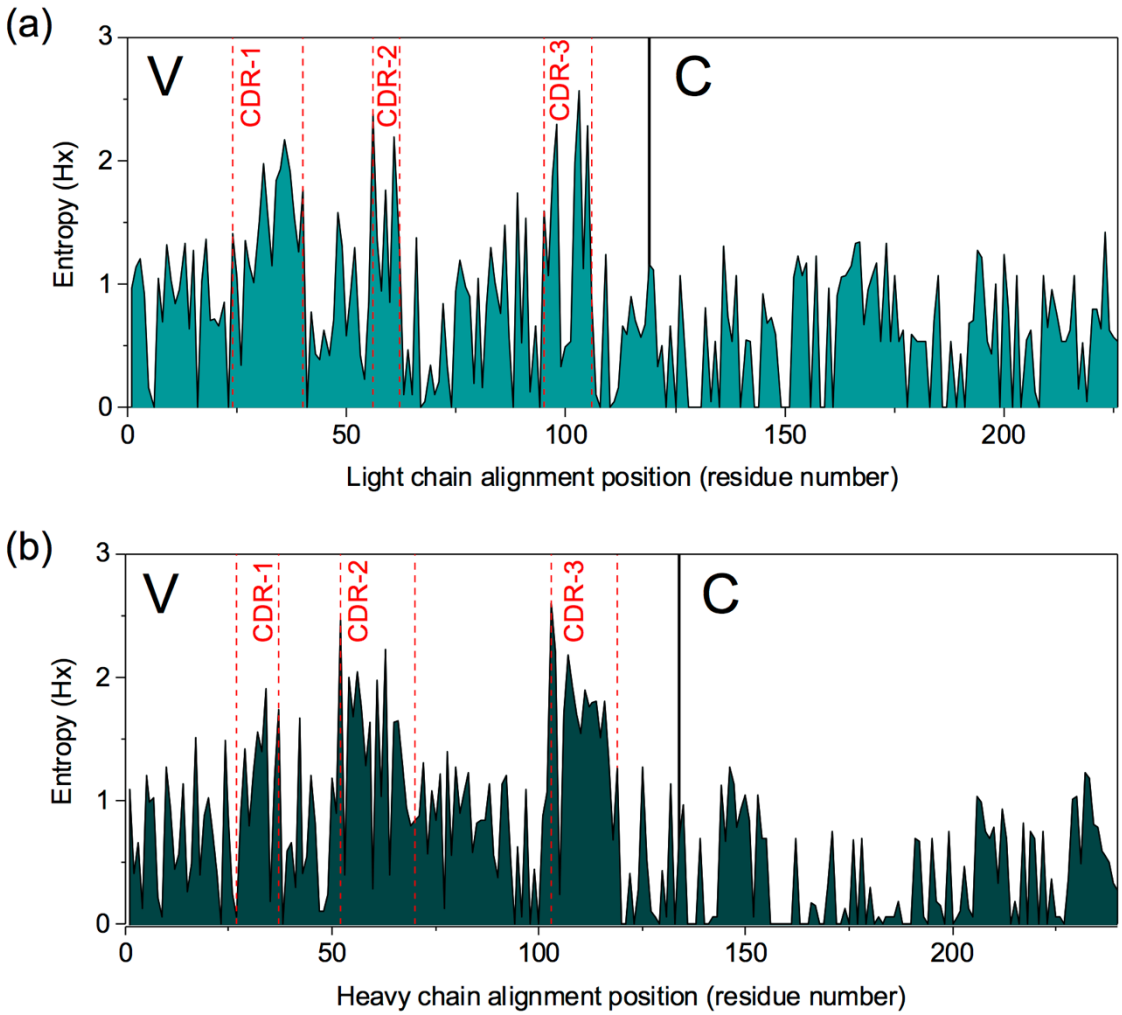

Figure 3.14. Sequence entropy of Fab sequences. (a) Entropy (Hx) of one hundred Fab light chains, including κ and λ light chains. (b) Entropy (Hx) of one hundred Fab heavy chains. Light and heavy chains are from human and mouse species. Alignment and entropy calculation were done with Bioedit. Variable domains are indicated with a (V) and constant domains with a (C), separated by a vertical line. CDRs are indicated in red.

Table 3.5 Comparison between the mutations in existing human and mouse Fabs and the stabilizing mutations suggested by FoldX and Rosetta.

Original Residue	Light chain				Heavy Chain		Suggested mutation by FoldX and Rosetta
	Kappa (κ)		Lambda (λ)		Human	Mouse	
	Human	Mouse	Human	Mouse			
S176	-	-	-	-	N.A.	N.A.	W, M, R, L Y
N137	-	-	-	T, I	N.A.	N.A.	L, M, I
S397	N.A.	N.A.	N.A.	N.A.	-	T	I, L, W, V
S159	-	V	V	M	N.A.	N.A.	R, F
S12	-	Y, P, A, T	-	T	N.A.	N.A.	F, Q, Y
T180	-	-	S	-	N.A.	N.A.	Y, W
S395	N.A.	N.A.	N.A.	N.A.	-	T	L, M, R, I, V
S267	N.A.	N.A.	N.A.	N.A.	P, W, Y, T, G, D, N, A	P, I, G, N, D, Q, L	P
K103	R	-	-	-	N.A.	N.A.	Y, F, T
T72	S	S	S	A	N.A.	N.A.	Y

N.A. (Not Applicable), mutation does not apply to the chain (light or heavy); -, no existing mutations were found on that chain.

3.5 Conclusions

Antibody-based products are the main class of approved biopharmaceuticals, due to their high target specificity. However, there are many barriers to their successful development into therapeutics, with protein aggregation being perhaps the most common and challenging to prevent. There is a need to identify potential instabilities of therapeutic proteins during their early development, particularly against stresses that they will encounter during manufacture, storage and delivery. This would allow their early elimination from further development, or otherwise rational mutagenesis into more stable products. In this context, I have elucidated the first unfolding events that take place on a humanized Fab A33 using atomistic MD simulations, and compared these to predictions of potentially stabilising mutations using computational tools.

Simulations showed that contacts at the interface between domains (V_L-V_H and C_L-C_H1) were lost before individual domains unfolded. Interfacial contacts in the constant domain specifically, were the least stable, which were lost very quickly during the simulations under both stresses, low pH and high temperature. In line with these results, FoldX and Rosetta agreed that the residues that can be stabilized the most, are located in the constant domain interface. Further validation was provided by packing density calculations, which revealed that the residues identified by the stability predictors, were under packed relative to the other residues located in the interface between domains. Based on these findings, I speculate that improvement of Fab A33 stability should start at the constant domain interface. Only one of the top mutations suggested by FoldX and Rosetta, N137I was found to be present in the analysis of natural variation within existing Fab sequences. However, there was significant scope for improvement through mutating the interfacial residues S176, N137, S397, T180, and S395, to the suggested hydrophobic residues.

The further goal could be to improve the stability of the individual domains. The C_L domain was found to unfold at both, low pH and high temperature. Salt bridge analyses identified two key salt bridges that can be at the heart of this domain unfolding at low pH, Glu165-Lys103 and Glu195-Lys149. Glu165-Lys103 bridges the C_L domain to the V_L domain, and Glu195-Lys149 is located in outer β-strands of the C_L domain, bridging the β-strands C and F. FoldX and Rosetta also identified stabilizing mutations in the C_L

domain. To stabilize the interaction between the C_L and V_L domain, S12 and K103 were identified to be mutated to hydrophobic residues. Interestingly, the mutation S12 to Tyr is found naturally. In the C_L domain, S159 was identified, which interacts with an outer β-strand, suggesting this interaction can also be improved. Lastly, the C_H1 domain was also found to unfold at high temperature. The only mutation identified in this domain is S267, to a Pro, which notably is found naturally. Overall, the results found with MD simulations and stabilizing software predictors strongly agree in the domains of Fab A33 that can be stabilized further.

In order to gain insights into the mechanisms by which aggregation might occur, APRs in Fab A33 were identified, and their solvent accessibilities were compared. All APRs in Fab A33 are located in the interior of the protein, however, at low pH and high temperature the SASA of certain APRs increased. Notably, different APRs were exposed under both stresses, suggesting that different aggregation mechanisms occur under each stress. This result stresses the importance of identifying the stability of a protein under the different stresses it might encounter. Taken together, this work provides insights into the stability and robustness of the therapeutically relevant Fab A33, and offers a path to the engineering and design of a more aggregation resistant antibody fragment.

Chapter Four

X-ray scattering and atomistic modelling identify an expanded conformation of Fab A33 at low pH that reveals an aggregation mechanism

4.1 Summary

To prevent aggregation, the mechanism of protein aggregation needs to be understood. This begins with elucidation of the conformational states that lead to aggregation; however, very little is known about the structures of native conformers that initiate aggregation. While several structures of final aggregated states - notably amyloids - are available in the literature, very little is known about the structures of native-like states predicted to mediate the onset of aggregation. Here, I combined small-angle X-ray scattering (SAXS) and atomistic modeling, to characterize an aggregation-prone state of Fab A33 at low pH. SAXS showed that Fab A33 adopted a more expanded conformation at acidic pH (5.5, 4.5 and 3.5) compared to neutral pH (7.0 and 9.0), with radius of gyration increases of between 2.2% and 4.1%. The same conditions lead to accelerated aggregation, indicating that the expanded conformations were more aggregation-prone. To maximize the resolution of SAXS, I took a novel approach that fitted the data to 45,000 structures obtained from fully atomistic molecular dynamics simulations of the entire molecule under the same conditions. This revealed the regions of the Fab undergoing conformational fluctuations, and located the conformational changes in the native state to the constant domain of the light chain (C_L). Lastly, the conformational changes were found to expose a predicted aggregation-prone region (APR) which forms a mechanistic basis for subsequent aggregation. The structural elucidation of aggregation-prone native-ensemble conformers using SAXS atomistic modelling provide a means by which aggregation-prone conformational states can be readily determined experimentally, and used to guide rational approaches to stabilize proteins against aggregation.

4.2 Introduction

Aggregates are the manifestation of the protein's physical instability, and are problematic because they lower the activity of the therapeutic drug and increase its immunogenic potential (Manning et al. 2010; Wang 2005). The most widely accepted aggregation mechanism involves two steps: (i) a conformational change to the protein's native state and (ii) assembly of protein molecules into aggregates (Chi et al. 2003). In the first step, the native state of the protein experiences a conformational change to form an aggregation-competent specie. This intermediate is believed to expose aggregation-prone regions, which are normally shielded from the solvent in the native protein (De Baets et al. 2014). In the second and subsequent steps, the intermediate is driven by the hydrophobic effect or the propensity of exposed sequences to form cross-β sheets to associate with other protein molecules (Roberts 2014). Energetically, the first step is controlled by the conformational stability of the native protein relative to aggregation-prone states, though this is often probed indirectly by the free energy of unfolding ΔG_{unf} of the native protein relative to the fully unfolded state. The second step, assembly into aggregates, is controlled by the persistence time, or relative population of aggregation-prone states, their ability to form specific intermolecular interactions, and their colloidal stability in terms of intermolecular attractive and repulsive forces. Over the years, it has been found that many of the proteins studied in the desired solution conditions follow a first order aggregation kinetics. This implies that the rate-limiting step in these conditions is unimolecular, such as a conformational change, rather than a bimolecular reaction (Chi et al. 2003). Thus, based on these findings, elucidating the conformational states that lead to aggregation is crucial to preventing protein aggregation.

Initial studies on protein aggregation suggested that aggregation takes place from the fully unfolded state, derived from studies at elevated temperatures. Increasing evidence suggests that, at temperatures below the melting temperature (T_m) of the protein, aggregation takes place from near-native states, where only partial unfolding of the protein has taken place (Robinson et al. 2018). This hypothesis is supported by recent work on Fab A33, which found that the T_m of the protein is only a good predictor of aggregation rate at temperatures close to the T_m of the protein, where aggregation from the unfolded state predominates (Chakroun et al. 2016). At temperatures below the T_m of

the protein, T_m was not a good predictor of aggregation rate, since global unfolding is not necessary for aggregation to happen.

Over the years, many studies have reported on the presence of near-native states of proteins that are aggregation-prone. A combined analysis of kinetics and solution thermodynamics of recombinant human interferon-γ, found that only a 9% expansion of the native-state surface area was necessary to form the intermediate state that preceded aggregation (Kendrick et al. 1998). Similar results were found for human granulocyte colony stimulating factor, in which they found that the expanded intermediate state preceding aggregation represented only a 15% of the surface area of the completely unfolded conformation (Krishnan et al. 2002). More recently, only transient local unfolding was found necessary to show faster aggregation for variants of human lysozyme, using hydrogen/deuterium exchange experiments (Canet et al. 2002). Studies on hyperthermophilic acylphosphatase, superoxide dismutase 1, transthyretin, 2-microglobulin and Fyn SH3 also showed that global unfolding was not necessary, and that aggregation could be initiated from locally unfolded states (Chiti & Dobson 2009; Neudecker et al. 2012). NMR was able to resolve a structural folding intermediate of the 6.4kDa Fyn SH3 domain that was more aggregation-prone than the native state (Neudecker et al. 2012). However, this relied upon mutations that stabilized the folding intermediate, and so the use of NMR to characterize directly pre-aggregational states in unmutated native-ensembles remains very challenging, particularly for larger proteins such as the 48 kDa humanized antibody fragment Fab A33.

While hydrogen-deuterium exchange by NMR or mass spectrometry can map changes in native ensemble dynamics, it remains challenging to characterize the structural conformers being sampled, particularly for large proteins. In this chapter, I characterized an aggregation prone species of Fab A33 at low pH, using small angle X-ray scattering (SAXS) in combination with atomistic structures generated using Molecular Dynamic (MD) simulations. SAXS is a diffraction technique used to characterize macromolecules in solution. It is particularly useful to study changes in protein structure due to different solution conditions. Results showed that the room temperature aggregation of Fab A33 occurred through local unfolding into more expanded native-like conformers. The intermediate state preceding aggregation was very similar in structure to the native state but, it was more expanded, it had local regions with increased flexibility and had increased exposure of hydrophobic residues that favor aggregation.

4.3 Methods

4.3.1 Cloning, site-directed mutagenesis, expression and purification of Fab A33

The gene coding for Fab A33 was generously supplied by UCB Celltech (Slough, UK) in the plasmid pTTOD in *E. coli* W3110. The original gene contained an unpaired cysteine at position 226, in the hinge region. To avoid formation of Fab dimers, this cysteine was mutated to a serine, C226S, and I refer to this variant as wild-type Fab A33. WT Fab A33 was expressed and purified as described in previous works (Chakroun et al. 2016; Hilton 2015).

4.3.2 Acquisition of small-angle X-ray scattering data

Data acquisition was performed by Dr David Hilton (Hilton 2015), and it is summarized here. Scattering data was collected for Fab A33 at 1.0 mg/ml and 20 °C, for a range of pH and ionic strengths (IS). Specifically, five pH 3.5, 4.5, 5,5. 7.0 and 9.0, and four IS 20, 50, 150 and 250 mM. Buffers at pH 3.5, 4.5 and 5.5 were 20 mM sodium acetate, at pH 7.0 in 20 mM sodium phosphate, and at pH 9.0 in 20 mM Tris.HCl buffer. The ionic strengths of each buffer were modulated through the addition of sodium chloride (NaCl), to the exception of IS 20 mM were no NaCl was added.

Small-angle X-ray scattering measurements were carried out on beamline BM29 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. The beamline energy ranged from 7 - 15 keV with a storage ring current between 166 mA to 195 mA. Scattering data, I(Q), was collected using a 2D Pilatus detector located 2.867 m from the sample, which combined with a wavelength of 0.09919 nm enabled a Q-range of 0.025 - 5nm^{-1} to be accessed. Samples were stored at 20 °C and loaded in 50 µL portions using a quartz flow-through capillary (diameter 1.833 mm, wall thickness 0.02 mm). Between samples the capillary was automatically washed with ddH_2O, then detergent before flushing with ddH_2O. Data were collected in 10 successive 1 second frames, to minimize the effects of radiation damage, with pre- and post- sample pure buffer measurements for subsequent background subtraction. The 2D data were automatically normalized to an absolute scale, calibrated using the scattering profile of water, and azimuthally averaged

to obtain a 1D intensity profile. Profiles with observable radiation damage were discarded prior to averaging and buffer subtraction. Basic manipulations of the experimental small angle scattering profiles were performed using PRIMUS. This included the subtraction of buffer profiles from those of their corresponding protein samples and averaging of triplicate scattering data.

4.3.3 Analysis of small-angle X-ray scattering data

In a SAXS measurement, intensities I(Q) of the scattering curve are measured as a function of Q (where $Q = 4\pi \sin(\theta)/\lambda$; 2 θ is the scattering angle and λ is the wavelength). Two analyses were performed afterwards. They both found the radius of gyration (R_g) of the protein and the molecular weight from the forward scattered intensity at zero angle I(0). Guinier analyses only use the data at low Q values, up to a Q*Rg of 1.5 for globular proteins (where the approximation is still valid). Guinier analyses consist in plotting ln I(Q) as a function of Q^2, and performing a linear fit.

$$\ln I(Q) = \ln I(0) - {R_g}^2 Q^2/3 \qquad \text{(Eq. 4.1)}$$

The second analysis, uses the full Q range of the scattering data. It consists on the Fourier transformation of the scattering data I(Q) in reciprocal space into real space, to give the distance distribution function P(r). P(r) curve represents all the distances between all the volume elements in the protein. For example, the maximum in the curve corresponds to the most commonly occurring distance, and the maximum r represents the length of the protein. I used the program GNOM for the calculation of P(r) curves and obtain R_g and I(0) (Franke et al. 2012).

4.3.4 MD simulations to generate Fab A33 conformations at different pH

Molecular dynamic (MD) simulations on the Fab A33 homology model were carried out using Gromacs v5.0 (Abraham et al. 2015) and the OPLS-AA/L all-atom force field (Kaminski et al. 2001; Kortkhonjia et al. 2013; Hu & Jiang 2010; Smith et al. 2015; Yu & Dalby 2018; Yu et al. 2017; Zhang et al. 2018), as in the previous chapter. MD

simulations were carried at five pH, 3.5, 4.5, 5.5, 7.0 and 9.0, all at an ionic strength of 50 mM. These simulations were generated at three temperatures, 300 K, 340 K and 380 K, to increase the range of energy to the system and generate more variable structures. In addition, three independent simulations were carried out for all conditions. The protonation state of each residue were determined at each pH using the PDB2PQR server, which performed the pKa calculations by PropKa (Li et al. 2005). Calculated total charge of Fab A33 at each pH were: +35 (pH 3.5), +18 (pH 4.5), +12 (pH 5.5), +9 (pH 7.0) and +5 (pH 9.0). As described in the previous chapter, Fab A33 was placed in a cubic box, solvated with water molecules, and ions were added to neutralize the system and adjust the ionic strength. The system was energy minimized and equilibrated under NVT and NPT ensembles. Lastly, 50 ns MD simulations were performed on the UCL Legion High Performance Computing Facility. The time step of the simulations was set to 2 fs, trajectories were saved every 10 ps, and analyses were performed using standard Gromacs tools. To generate atomistic structures to use in combination with the SAXS data, snapshots were saved for every 50 ns simulation, every 10 ps, totaling 5,000 structures per trajectory. At each pH, every simulation was performed in triplicate and at three different temperatures. Thus, in total, 45,000 models were generated for each pH value to be compared to the SAXS experimental data.

4.3.5 Atomistic modeling of SAXS data using SCT

SCT is an open source software designed for the comparison of experimental X-ray scattering curves to generated plausible structures of the protein in solution (Wright & Perkins 2015). SAXS is a low resolution diffraction technique, where, if no constraints are applied, the structural resolution is 2 - 4 nm. However, when SAXS is combined with atomistic models that provide structural constraints, the resolution is improved to 0.5 - 1.0 nm. The approach consists in computing theoretical SAXS curve for each generated model, and these are compared to the experimental curves. A subset of best-fit models is identified to represent the average solution structure.

I used the structures generated during MD simulations, to compare to experimental SAXS data. A theoretical X-ray scattering curve was calculated for each of the models (45,000 models per pH). First, a coarse-grained model needs to be constructed from the atomistic structures, to ease demand of processing power. This was done by

placing the models in a grid of boxes and replacing it with spheres. I used a standard box side of 0.54 nm and a cutoff of 4 atoms (selected using a structure at the end of a pH 3.5 simulation, one of the most extended models). After, a hydration shell of 0.3 g of water per gram of protein was added to the models because SAXS visualizes the layer of water in contact with the protein. Second, theoretical scattering curve were calculated using the Debye's Law adapted to spheres (computing all the distances r from each sphere to the remaining spheres and summing the results). Third, experimental and theoretical curves were compared using the R factor (by analogy with crystallography, ow R factors represent the better fit structures). Q range: 0.37-1.6 nm^{-1}.

$$R = \frac{\Sigma\left\|\|I_{Expt}(Q)\|-\eta\|I_{Theor}(Q)\|\right\|}{\Sigma\|I_{Expt}(Q)\|} x\ 100 \qquad \text{(Eq. 4.2)}$$

4.3.6 Aggregation-prediction regions software

Aggregation-prone regions (APR) of Fab A33 were predicted using the sequence-based APR predictors PASTA 2.0 (Walsh et al. 2014), TANGO (Fernandez-Escamilla et al. 2004), AGGRESCAN (Conchillo-Solé et al. 2007) and MetAmyl (Emily et al. 2013), and Amylpred2 consensus tool (Tsolis et al. 2013) was used to confirm the presence of these APRs, as in Chapter 3. The difference in solvent accessibility of the APRs in the SAXS best-fit structures at pH 7.0 and pH 3.5 were analysed using Pymol software.

4.4 Results and discussion

The small-angle X-ray scattering data presented in this chapter was collected by Dr David Hilton (Hilton 2015). I have performed the analysis of this data.

4.4.1 SAXS identified an expanded aggregation-prone conformation of Fab A33 at acidic pH

Small-angle X-ray scattering (SAXS) is a diffraction technique that characterizes the structure of a protein in solution (Perkins et al. 2009). By studying the protein in different solution conditions, it is possible to elucidate changes in its conformation due to solution conditions. In this chapter, I investigated the effect of pH and salt concentration upon Fab A33 structure. X-ray scattering curves were acquired for Fab A33 at 1 mg/ml in 20 different solution conditions; five pH: 3.5, 4.5, 5,5. 7.0 and 9.0 and four NaCl concentrations: 0, 50, 150 and 250 mM. Scattering data was first analyzed to obtain the radius of gyration (R_g) and the intensity at zero Q (I(0)), the latter being proportional to the molecular weight. R_g and I(0) can be obtained using two analyses, Guinier analysis of the low Q region and pair density distribution P(r) analysis from the Fourier transform of the full scattering curve. Both analyses found consistent results (Figure 4.1 and 4.2).

Guinier plots from the low Q region of the scattering curve revealed the presence of minor amounts of aggregates, specially for the samples below pH 7.0 (red; Figure 4.1a). The presence of these aggregates can be seen at the lowest Q values, where I(Q) intensities increase and curve upward. Notably, it was also observed that if the data was fitted to larger Q values (green; Figure 4.1a), linear non-aggregated Guinier plots with satisfactory $Q*R_g$ ranges were identified that were distinct from the Guinier fits for aggregated Fab A33. This analysis was confirmed by monitoring I(0)/c of the non-aggregated Guiniers (larger Q region) (Figure 4.1b), (I(0) obtained from extrapolation of the line fit in the Q range of 0.37-0.5 nm^{-1} (green)). The molecular weight of the species present in the sample is proportional to I(0)/c, thus, if no aggregates are present, I(0) should be the same for all the conditions measured. Figure 4.1b showed I(0)/c to be similar at around 40 (39 ± 3) $(mg/ml)^{-1}$ for almost all 20 solution conditions. This outcome indicated that the R_g values of monomeric Fab A33 could be determined independently from its aggregation. In this way, two ranges were fitted: a larger Q range to determine

the radius of gyration of Fab A33 in solution, 0.37-0.5 nm^{-1}, and a shorter Q range to estimate the amount of aggregation present in the samples, 0.14-0.3 nm^{-1}. The parameter used to monitor the amount of aggregate in the sample was $\Delta I(0)/c$ [= $(I(0)_{Q:\ 0.14\text{-}0.3\ nm^{-1}} - I(0)_{Q:\ 0.37\text{-}0.5\ nm^{-1}})/c$], as previously used (Nan et al. 2013), to subtract the non-aggregated Guiniers from the aggregated Guiniers and gain an estimate of the amount of aggregation that had taken place.

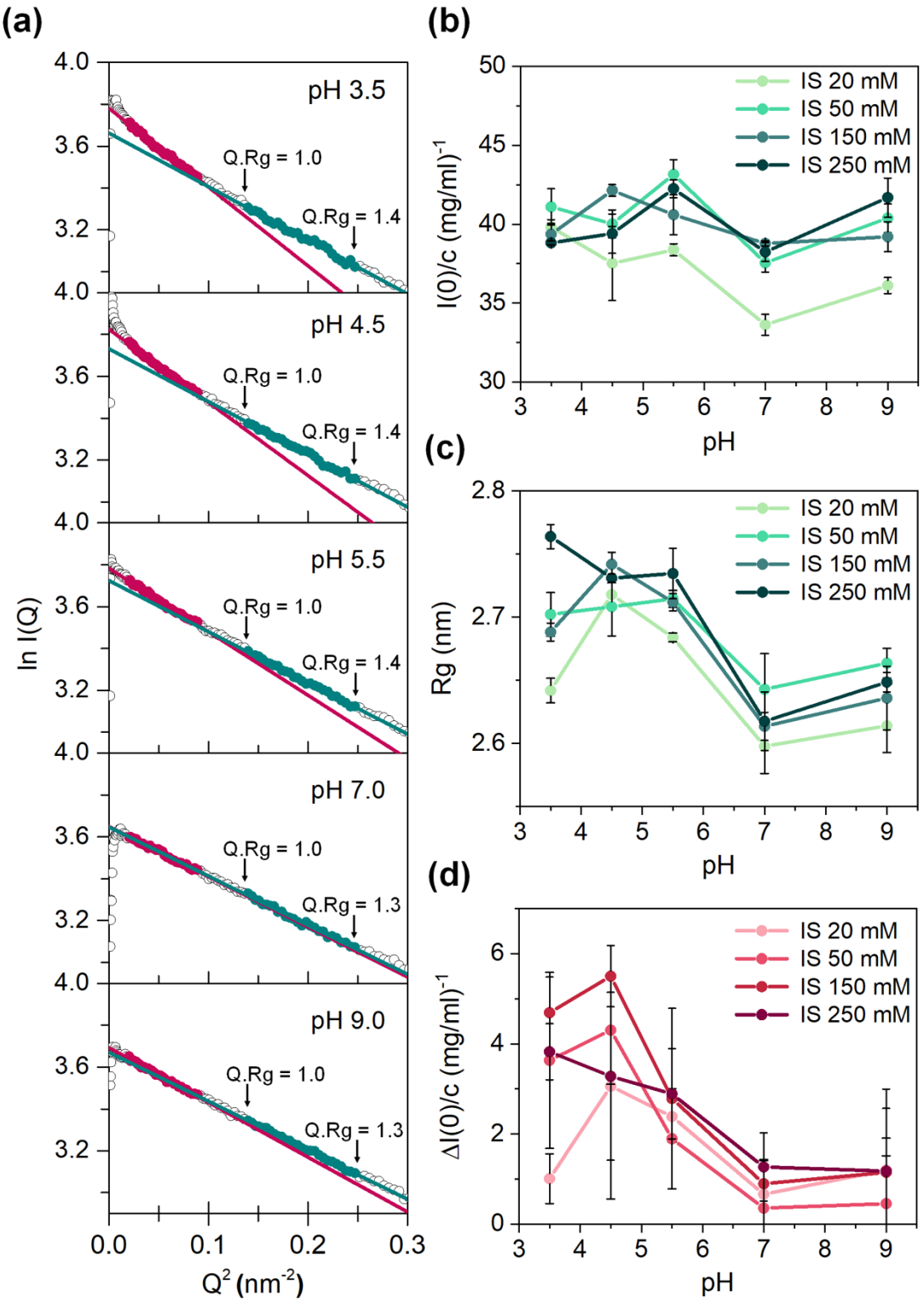

Figure 4.1. SAXS Guinier analyses. Twenty experimental conditions were studied for Fab A33 using five pH (3.5, 4.5, 5.5, 7.0, 9.0) and four ionic strengths (20, 50, 150, 250 mM). (a) Guinier plots of ln I(Q) vs. Q^2 gave the R_g and I(0) values. Five representative fits are shown for each of pH 3.5, 4.5, 5.5, 7.0 and 9.0 in an ionic strength of 50 mM. The fits for native Fab A33 were determined using the Q range of 0.37-0.5 nm^{-1} (green) and those for aggregated Fab A33 was determined from the Q range of 0.14-0.3 nm^{-1} (red). (b) I(0) values for native Fab A33, where I(0)/c is proportional to the molecular weight, and error bars are the SEM of three measurements. (c) R_g values for native Fab A33 for each of the 20 experimental conditions studied, with error bars are the SEM of three measurements. (d) The amount of aggregate present was determined from ΔI(0)/c, defined as $(I(0)_{Q:\ 0.14\text{-}0.3\ nm^{-1}} - I(0)_{Q:\ 0.37\text{-}0.5\ nm^{-1}})/c$.

SAXS showed that Fab A33 adopted a more expanded conformation at acidic pH (5.5, 4.5 and 3.5) compared to neutral pH (7.0 and 9.0) (Figure 4.1c). The radius of gyration increased from 2.62 nm at pH 7.0 and 2.64 nm at pH 9.0, to 2.70 nm at pH 3.5, 2.73 nm at pH 4.5, and 2.71 nm at pH 5.5 (mean SEM of 0.01 nm). These correspond to R_g increases of between 2.2 % and 4.1 % from neutral pH (7.0 and 9.0) to acidic pH (5.5, 4.5 and 3.5). In addition, SAXS data allowed the study of the effect of pH and salt concentration upon protein conformation. Results showed that pH had a bigger effect on the conformation of Fab A33 than salt concentration. Whereas a change from neutral pH to acidic pH, induced unfolding of Fab A33 to a more expanded conformation, an increase in IS from 0 to 250 mM had little effect on R_g (e.g., R_g increased at pH 7.0 from 2.60 nm to 2.64 nm, and at pH 4.5 from 2.72 nm to 2.73 nm (mean SEM of 0.01 nm)). These results are consistent with previous reports, which suggest that pH had a bigger influence on the conformation of the protein (Sahin et al. 2010).

I used ΔI(0)/c as a reporter of protein aggregation. SAXS detected the presence of small amounts of aggregates in the samples at acidic pH (5.5, 4.5 and 3.5), in contrast to the samples at neutral pH (7.0 and 9.0) where no aggregates were found (Figure 4.1d). The same conditions that found an expanded conformation of Fab A33 lead to accelerated aggregation, indicating that the expanded conformation of Fab A33 is aggregation-prone. Interestingly, no aggregates were detected in the pH 3.5 samples at a low IS of 20 mM, unlike for the pH 4.5 and 5.5 samples. This is probably due to the fact that at pH 3.5 Fab

A33 is highly positively protonated, and these charges prevent aggregation of protein molecules, stabilizing Fab A33 colloidally. When salt concentration is increased (Figure 4.1d, pH 3.5 and IS of 50, 150 and 250 mM), the long-range repulsions between charged Fab molecules become shielded, which favors aggregation. These results are consistent with the Fab A33 aggregation kinetics observed previously (Chakroun et al. 2016), for which the pH 3.5 samples at low IS also aggregated much more slowly than at higher pH or IS. These findings also suggest that the addition of salt mainly contributes to charge shielding, and does not destabilize the native conformation sufficiently to induce global unfolding.

The distance distribution functions P(r) also provided an alternative route for calculating the R_g and I(0). Figure 4.2b showed I(0)/c to be similar at around 40 $(mg/ml)^{-1}$, for almost all 20 solution conditions, further indicating that the R_g values of monomeric Fab A33 could be determined independently from its aggregation. P(r) analysis found consistent results to the Guinier analyses, where A33 adopted a more expanded conformation at acidic pH (5.5, 4.5 and 3.5) compared to neutral pH (7.0 and 9.0). The radius of gyration increased from 2.62 nm at pH 7.0 and 2.64 nm at pH 9.0, to 2.71 nm at pH 3.5, 2.70 nm at pH 4.5, and 2.70 nm at pH 5.5 (mean SEM of 0.01 nm), (Figure 4.2c).

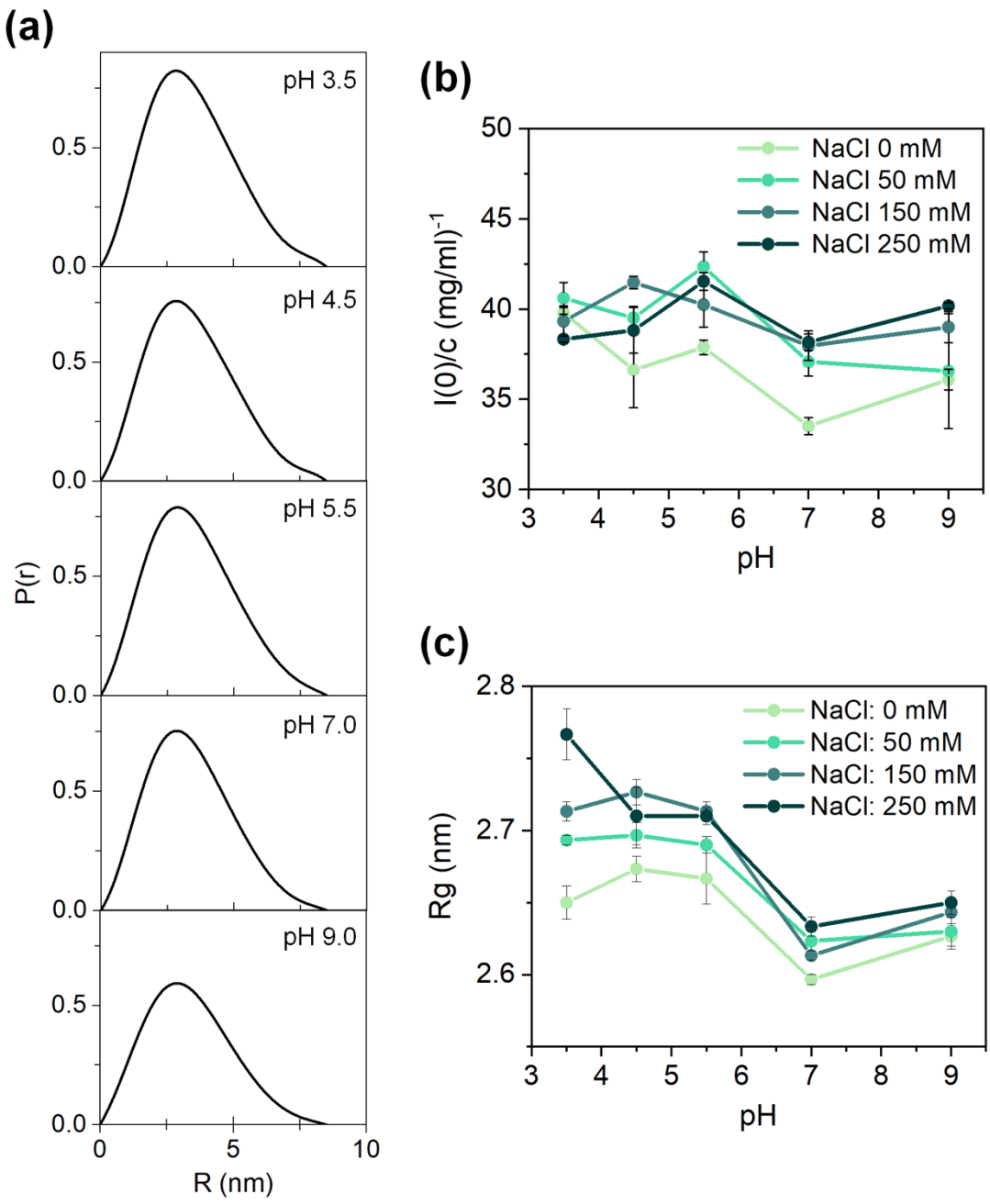

Figure 4.2. pH and ionic strength dependence of the P(r) curves. Twenty experimental conditions were studied using five pH values (3.5, 4.5, 5.5, 7.0, 9.0) and four ionic strengths (20, 50, 150, 250 mM). (a) Representative examples of the P(r) curves at five pH values as labelled; the five curves correspond to an ionic strength of 50 mM. The P(r) curves were calculated using the Q range of 0.3-2 nm^{-1}. (b, c) The I(0) and radius of gyration R_g values for Fab A33 are shown for each of the 20 experimental conditions studied. Errors are the SEM of three repeats.

4.4.2 Correlation of Fab A33 radius of gyration and aggregation rate

SAXS elucidated the presence of a more expanded conformation of Fab A33 at pH below 7.0, that correlated with the presence of small amounts of aggregates also detected by SAXS. To corroborate if the presence of this expanded conformation correlates with aggregation propensity, I combined the results obtained for the radius of gyration of Fab A33 using SAXS with previously reported aggregation kinetics in our lab, for the same experimental conditions (pH 3.5, 4.5, 5.5, 7.0 and 9.0; Ionic Strength: 20, 50, 150 and 250 mM; at 23 ˚C) (Chakroun et al. 2016). Aggregation kinetics were obtained by monitoring monomer loss using SEC-HPLC. Here, we show the initial rates of aggregation (v), measured from the first 20% of Fab A33 monomer loss. The initial rates of monomer loss as a function of pH and ionic strength were increased to 0.027 0.003 % day^{-1} at pH 3.5-5.5, compared to rates of 0.009 0.0018 % day^{-1} at pH 7-9 (Figure 4.3).

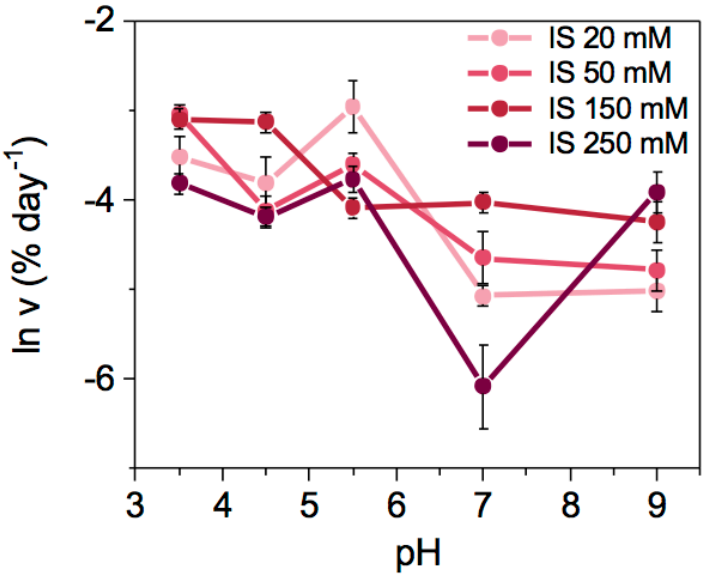

Figure 4.3. Aggregation rates as a function of pH. The initial rates of aggregation v in units of % day^{-1} for Fab A33 at 23 ºC were reported using SEC-HPLC for twenty experimental conditions based on five pH values of 3.5, 4.5, 5.5, 7.0 and 9.0 and four ionic strengths of 20, 50, 150, 250 mM. Errors are the SEM of three experimental repeats.

As found previously with SAXS, I found a correlation between R_g and aggregation rate. The experimental conditions that cause an increase in R_g (pH 3.5, 4.5, and 5.5) also resulted in faster aggregation rates than at neutral pH (7.0 and 9.0) (Figure 4.4). These results confirm that the expanded conformation of Fab is more aggregation-prone.

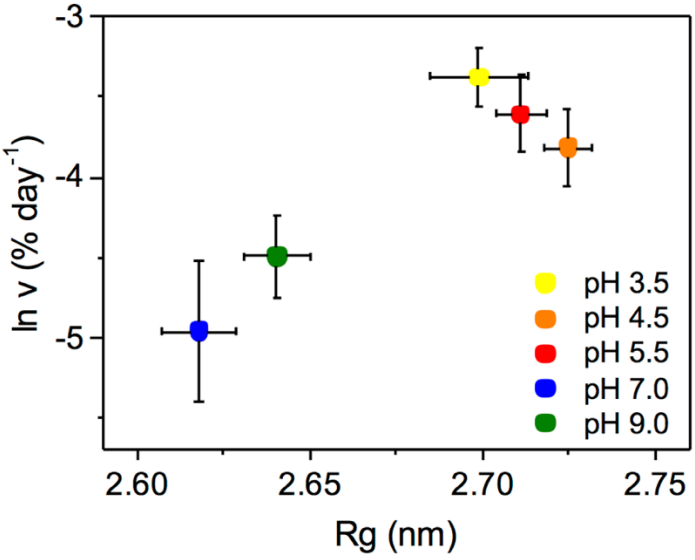

Figure 4.4. Correlation between the R_g values and aggregation rates v. For each pH (see inset), the averaged R_g values for native Fab A33 and the aggregation rates v are shown for the four ionic strengths. Error bars are the SEM.

4.4.3 Molecular Dynamic simulations captured pH-induced unfolding

Fab A33 homology model was used as the starting structure for MD simulations using Gromacs. MD simulations of 50 ns were carried at five pH, 3.5, 4.5, 5.5, 7.0 and 9.0, and one IS, 50 mM. Total charge and protonation state of the ionizable residues at each pH were determined using propKa software (which predicts the pK_a value of the ionizable groups based on the protein structure) (Li et al. 2005). Based on the homology model of Fab A33, propKa determined the following total charges at each pH: +35 (pH 3.5), +18 (pH 4.5), +12 (pH 5.5), +9 (pH 7.0) and +5 (pH 9.0). In addition, simulations were performed at three different temperatures, 300 K, 340 K and 380 K. In this section, I present the results of the simulations at 300 K. For all conditions, three independent simulations were carried out.

R_g of the protein and solvent accessible surface area (SASA) were monitored as a function of simulation time for all the pH (Figure 4.5a,b). For each pH, the three simulation repeats were averaged at every time point of the simulation (10 ps) to show the variability between repeats, and are shown with transparency. These averages were smoothed by window averaging consecutive data, and they are shown in darker color.

Results showed that the simulations were able to capture the increase in the R_g and SASA as the pH decreased, in agreement with SAXS data. The values of R_g derived from simulations were smaller than the R_g values reported previously with SAXS, because no account was taken of the hydration shell visible by SAXS, nonetheless the trends were clear. MD simulations also found an increase in R_g as the pH became more acidic. The MD simulations at pH 7.0 and 9.0 gave an R_g of 2.52 nm and 2.51 nm (mean SEM of 0.01 nm), respectively, whereas at pH 3.5 this increased to 2.58 nm ± 0.02 nm. At pH 4.5 and 5.5, the increase in R_g from simulations was not as noticeable as the increase observed experimentally with SAXS. MD simulations do not update the protonation state of molecules continuously as the protein structure unfolds, and this would limit the rate of structural change during simulation, most critically at pH 4.5-5.5, which overlaps the pK_a range of acidic residues. Results for SASA followed the same trend that for R_g. As the pH of the solution decreased, the solvent accessible surface area of Fab A33 increased. SASA started at a value of 210 nm^2 for pH 7.0 and 9.0, and finished at a value of 220 nm^2 for pH 3.5.

The MD simulations also provided information about protein dynamics potentially down to the level of individual residues. The flexible regions of Fab A33 were assessed using the root mean square fluctuation (RMSF), this being the average distance that a residue moves during the simulation. The RMSF for the last 30 ns of each simulation (20-50 ns) were averaged for each residue, and visualized by color in Figure 4.5c and Figure 4.5d for pH 7.0 and pH 3.5 respectively. Less flexible residues are shown in blue and more flexible residues in red. For pH 7.0, the most flexible regions were found to be the loop regions, followed by the α-helical regions, then the β-strand regions. The highest flexibility was seen in the CDR loops, the C-terminus of the heavy chain, and several loops and α-helices. The β-strands of the C_L and C_H1 domains were more flexible than the V_L and V_H domains. For pH 3.5, Fab A33 was seen to be more flexible than at pH 7.0. In addition to the regions seen to be flexible at pH 7.0, which were also flexible at pH 3.5, both the C_L and C_H1 domains showed increased flexibility at low pH. Taken together, MD simulations found that at low pH, Fab A33 adopted an expanded conformation, with increased solvent surface area and regions with increased flexibility. These are characteristics expected of aggregation-prone conformers.

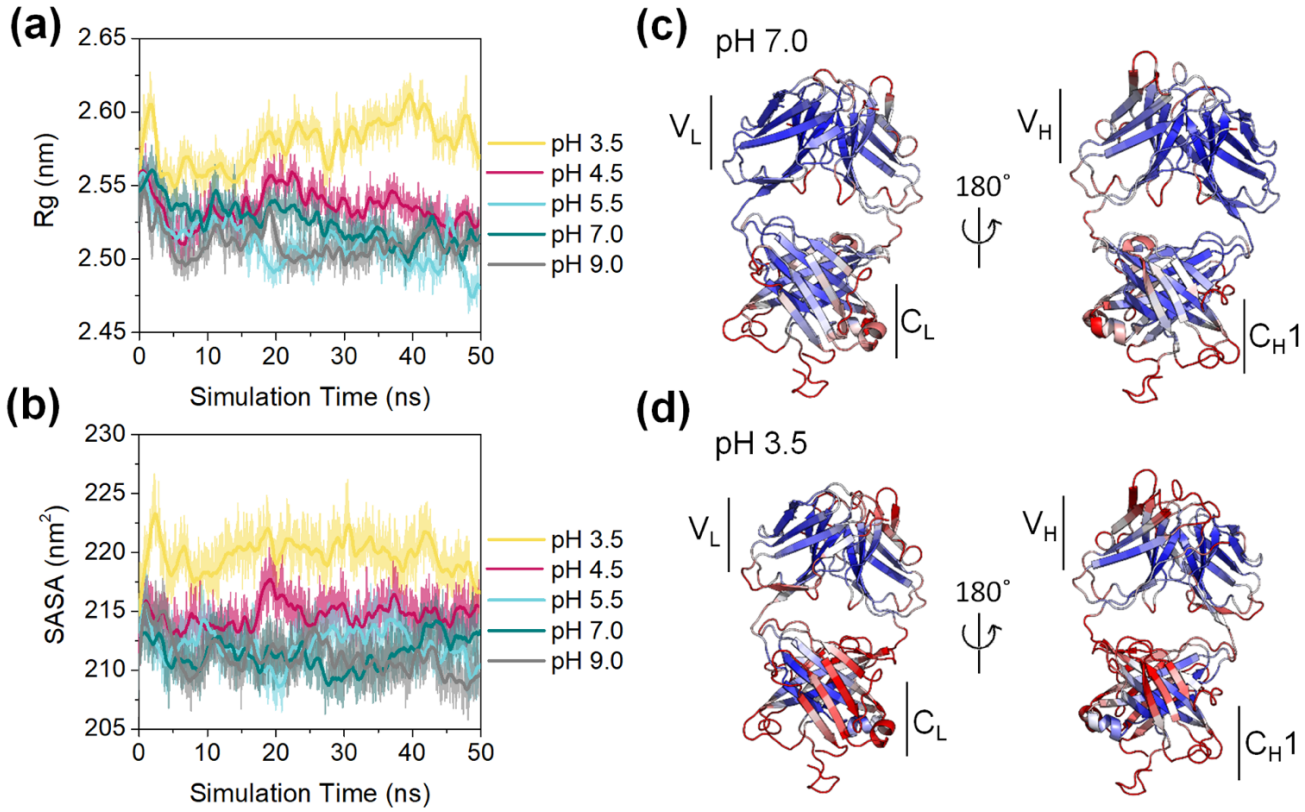

Figure 4.5. MD simulations of native Fab A33 at 300 K. (a, b) The R_g values and solvent accessible surface area (SASA) of Fab A33 are shown as a function of simulation time for five pH values as labelled, using an ionic strength of 50 mM for each. For each pH, three simulation repeats were averaged at every time frame, from which a window average is shown in a darker colour. (c, d) The root mean square fluctuation (RMSF) of the simulations at pH 7.0 and pH 3.5 respectively are shown in blue (low values) and red (high values) to highlight the dynamic regions in the structure. The RMSF values were added as notional B-factors to the PDB file for the Fab A33 homology model.

4.4.4 Atomistic modelling of SAXS data to characterize the expanded conformation

The full X-ray scattering curve contains additional information about the structure of the protein in solution, beyond the radius of gyration. One way to extract this information is to combine SAXS data with atomistic models of the protein in different conformations. A theoretical X-ray scattering curve can be calculated from each of the atomistic models, and these curves can be compared to the experimental X-ray scattering curve. From these, the curves that best fit the experimental SAXS curve are identified. The structures corresponding to these theoretical best curves are accepted as representative of the average solution structure. This has been applied to a range of protein structures (Wright & Perkins 2015; Walker et al. 2017; Rayner et al. 2015).

Each MD simulation above recorded 5,000 structural snapshots of the 50ns simulations at every 10 ps, i.e. 45,000 structural models for each pH value. A theoretical scattering curve was calculated for each of the models from the simulations, and the theoretical curves were compared to an experimental SAXS curve at the same pH and NaCl concentration. To measure how good the fit between theoretical and experimental curves is, the parameter R factor was used, by analogy with crystallography, which monitors the agreement between the curves. The better the fit, the lower the R factor. R factors were calculated by comparing theoretical and experimental curves in the Q range 0.37-1.6 nm^{-1}. Graphs of R factors (goodness of fit) versus R_g values are very helpful to oversee how well the atomistic models are reproducing the experimental data. Thus, theoretical radius of gyrations were also calculated from the theoretical scattering curves, by performing Guinier Analysis on the same Q range that the one used on the experimental curves, Q range 0.37-0.5 nm^{-1}.

Figure 4.6a shows, for each pH, an example of R factor versus R_g graph, which allows us to see how well the 45,000 models compared to that experimental curve. First, all R factor values were below 5%, which indicated very good fits, and thus, models generated during the simulations were close in structure to the Fab A33 solution structure found experimentally. The presence of a minimum at all pH indicated that enough conformations of Fab A33 had been sampled. Interestingly, at pH 5.5, simulations at higher temperature were necessary to characterize the minimum. This was suggested in the previous section, where it was found that simulations at pH 5.5 and 300 K did not

capture the conformational change observed experimentally. The ten best fit models for each pH were highlighted in the graphs and shown in yellow, with R-factors below 2%. R_g found experimentally for each pH are indicated as vertical lines in the graph, with all minima being within one standard deviation of the experimental R_g values, indicating very good fits had been obtained. Notably, minimums of best fit structures were reached at different radius of gyration for each pH. As found previously, Fab A33 had a more compact structure at pH 7.0 and 9.0, and partially unfolded to a more expanded conformation at pH 5.5, 4.5 and 3.5.

To visualize how well the best fits models compare to the experimental data, the X-ray scattering curves of the best models and experimental curves were overlapped (Figure 4.6b). The P(r) curves from the models were also calculated and overlapped to the experimental curves, and are shown on the upper-hand corner. Visual inspection showed very good fits between the X-ray experimental I(Q) (circles) and best-fit modelled I(Q) curves (red lines), and experimental P(r) curves (black lines) and best models (red lines).

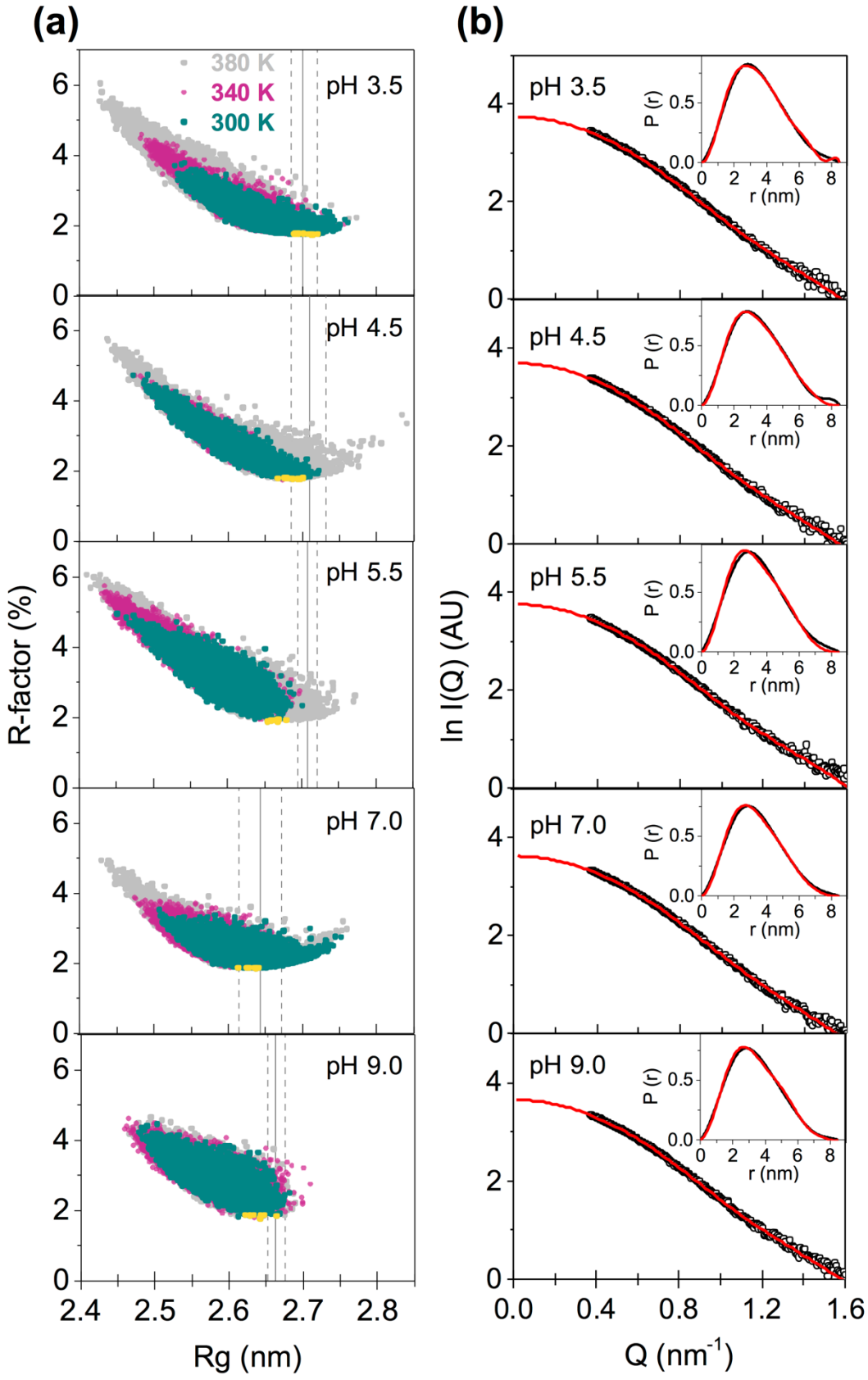

Figure 4.6. Comparison of the SAXS data with the MD simulations.

Figure 4.6. Comparison of the SAXS data with the MD simulations. (a) Comparison of the experimental X-ray scattering curve for native Fab A33 with the structures generated from the MD simulations for the five pH values as shown. MD simulations were carried out using an ionic strength of 50 mM at three temperatures of 300 K (green), 340 K (purple) and 380 K (grey). In total, each experimental SAXS curve was compared against 45,000 simulated structures per pH value. The goodness of fit was monitored using R-factors (Methods). The vertical lines represent the experimental R_g values with their experimental errors (SEM). The R_g value of each model was calculated from the theoretical scattering curve using the same Q-range used experimentally. The 10 best fit models with the lowest R-factors are highlighted in yellow. (b) Comparison of each experimental SAXS scattering curve (black) with its best-fit modelled curve (red). The inset shows the comparison between the experimental and best fit modelled P(r) curves.

To gain insights into the molecular structure of the Fab A33 expanded conformation at low pH, alignments of the sets of ten best-fit structures were performed. First, the ten PDB structures from MD simulations that best fitted the SAXS experimental data for pH 7.0 and IS of 50 mM were considered (Figure 4.7). I observed that the overlap was not perfect, which means that Fab A33 is dynamic in solution and we can gain information about the more flexible regions in its structure. The RMSD of each Fab domain was calculated relative to the structure that best fitted the SAXS experimental data at pH 7.0, as reference. As expected from their antigen-binding role, the CDR loops showed high flexibility, with a median RMSD of 0.18 nm and an interquartile range of 0.13-0.21 nm. Interestingly, the C_H1 domain also showed high flexibility with an RMSD of 0.17 nm and range of 0.10-0.19 nm. In particular, the C_H1 C-terminal -strand connected to the hinge peptide showed wide conformational variability. In the full-length antibody, the hinge is attached to the Fc region, which may provide additional stability. The V_H and C_L domains showed RMSDs and ranges of 0.15 nm (0.14-0.16 nm) and 0.10 nm (0.09-0.12 nm) respectively. The V_L domain showed the least variability with a RMSD and range of 0.08 nm (0.08-0.09 nm) and good alignment of all the β-strands.

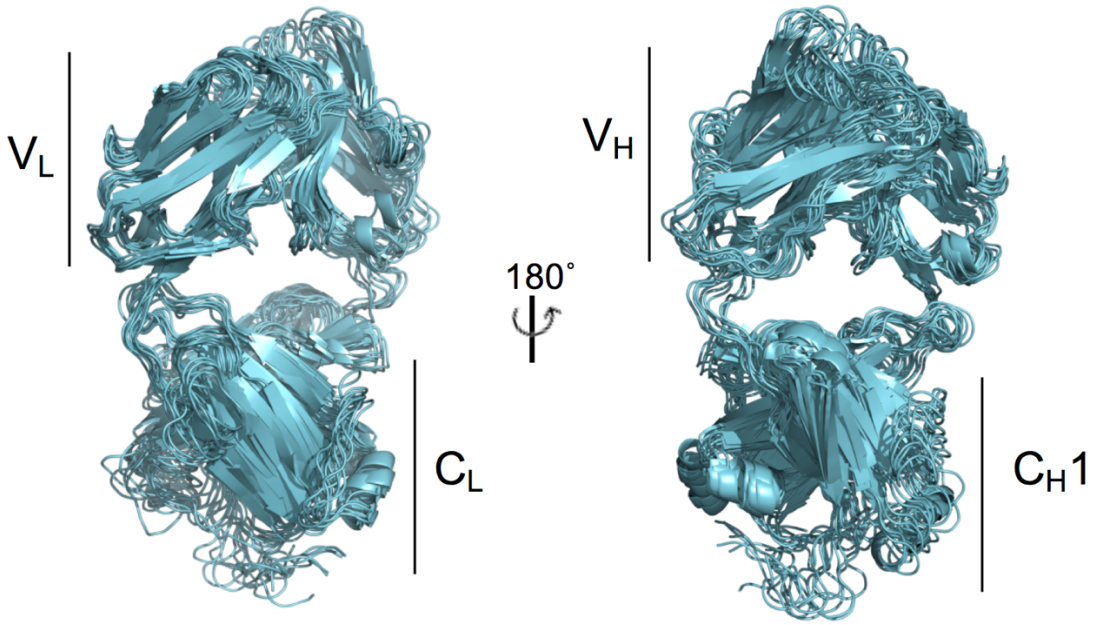

Figure 4.7. Alignment of the best-fit Fab A33 structures at pH 7.0. The ten best-fit simulated structures determined for pH 7.0 and an ionic strength of 50 mM are aligned in a cartoon representation. (a) Alignment of the light chain; (b) Alignment of the heavy chain.

I next aligned the ten best fit structures at pH 7.0 to the ten best fit structures at pH 3.5 (Figure 4.8). These alignments provided structural information about the pH induced conformation change. The RMSDs at pH 3.5 were also calculated relative to the reference best fit structure at pH 7.0. Notably, the C_L domain was the only domain to show a significant increase in RMSD as the pH was decreased. The RMSD and range of the C_L domain increased to 0.16 nm (0.13-0.17 nm) at pH 3.5, compared to 0.10 nm (0.09-0.12 nm) at pH 7.0. The structure alignments revealed a displacement of this domain at low pH (magenta; Figure 4.8a), being more open to solvent at pH 3.5. This pH-dependent domain displacement was clearly visualized in two loops (light chain residues 150-159 and 198-205; arrowed in Figure 4.8a) which connect C_L β-strands. The α-helix at residues 183-188 of the C_L domain was displaced. The α-sheet structure was lost in 8 out of 10 of the C_L best structures at pH 3.5 in residues 144-147, which suggests an increased flexibility in this segment.

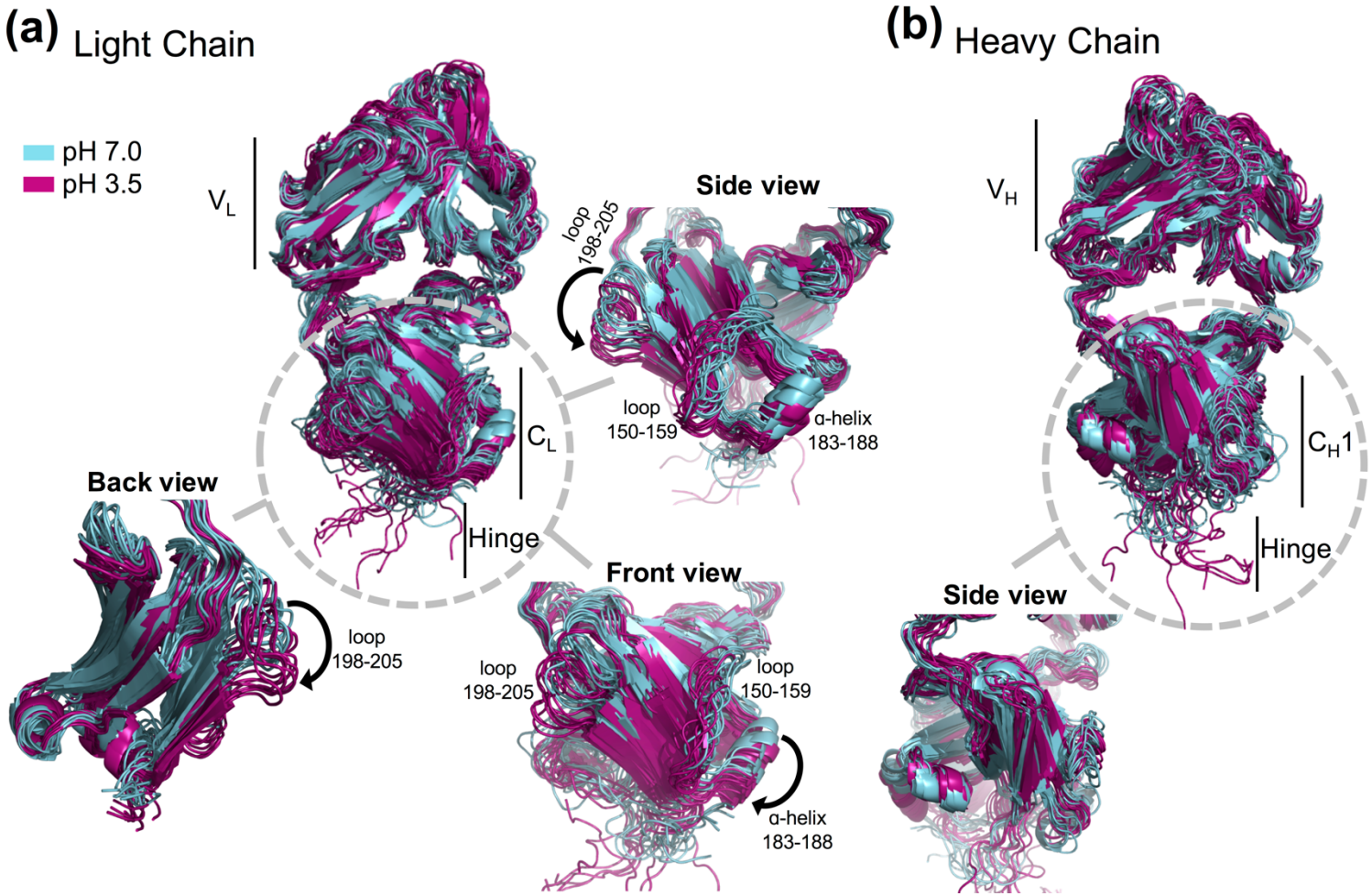

Figure 4.8. Alignment of the best fit Fab A33 structures at pH 7.0 and 3.5. The ten best-fit simulated structures determined for pH 7.0 are shown in cyan and those for pH 3.5 are shown in magenta; both at an ionic strength of 50 mM. (a) Alignment of the light chain in which the C_L domain is highlighted to show its loop and helix displacements at low pH in three views. (b) Alignment of the heavy chain, in which a side view of the C_H1 domain is shown.

No corresponding systematic displacements were found at pH 3.5 for the other three domains, even though these showed comparable RMSDs and ranges of 0.10 nm (0.09-0.10 nm) for the V_L domain, 0.17 nm (0.16-0.17 nm) for the V_H domain and 0.18 nm (0.16-0.20 nm) for the C_H1 domain, each relative to the best fit Fab A33 structure at pH 7.0. As seen at pH 7.0, the C_H1 domain was relatively flexible in sampling a wide range of conformations, particularly in the C-terminal -strand connected to the C-terminal hinge (Figure 4.8b). The hinge itself was highly extended at pH 3.5, and adopted a range of conformations. Additional views of the best-fit structures at pH 7.0, 5.5 and 3.5 in Figure 4.9 provide further visual support for the conformational shift at low pH in the C_L domain.

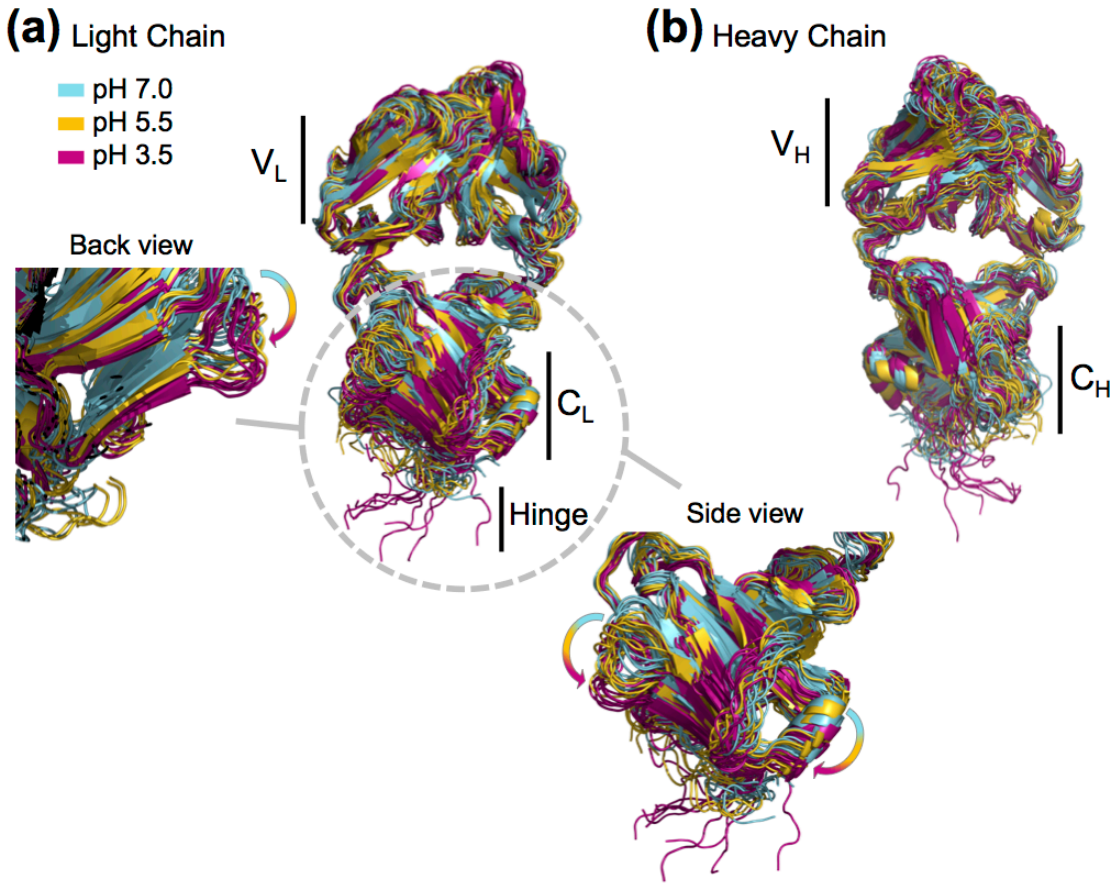

Figure 4.9. Alignment of the SAXS best fit structures at pH 7.0, 5.5 and 3.5. The top ten PDB structures that best fit the SAXS curves at three pH values are superimposed upon each other (cyan, pH 7.0; orange, pH 5.5; and magenta, pH 3.5), all for an ionic strength of 50 mM. (a) Alignment of the 30 light chains only, in which the C_L domains are highlighted with a dashed circle to show the displacement of this domain with pH. Back and side views of this domain are also presented. (b) Alignment of the 30 heavy chains.

In order to obtain a better picture of the displacements experienced by the whole Fab A33 molecule at low pH, distances between the four domains were also measured, using one cysteine in each domain. Six distances were monitored in total (V_L-V_H, C_L-C_H1, V_L- C_L, V_H- C_H1, V_L-C_H1 and V_H- C_L), using the four cysteines located in outer ß-strands (C23, C194, C236, C414), (Figure 4.10). The six distances were calculated for the ten best SAXS fit structures at pH 7.0 and the ten best SAXS fit structures at pH 3.5, and their averages and SEM are reported (Table 4.1). Results confirmed that the displacement at low pH occurred in the C_L domain, given that the only inter-domain distances that increased between pH 7.0 and 3.5, were the distances where the C_L domain was involved. Distances increased between the outer ß-strand cysteines of V_L-C_L (0.29 0.07 nm), V_H-C_L (0.27 0.03 nm) and C_L-C_H1 (0.05 0.02 nm). By contrast, the distances between the other domains did not change significantly.

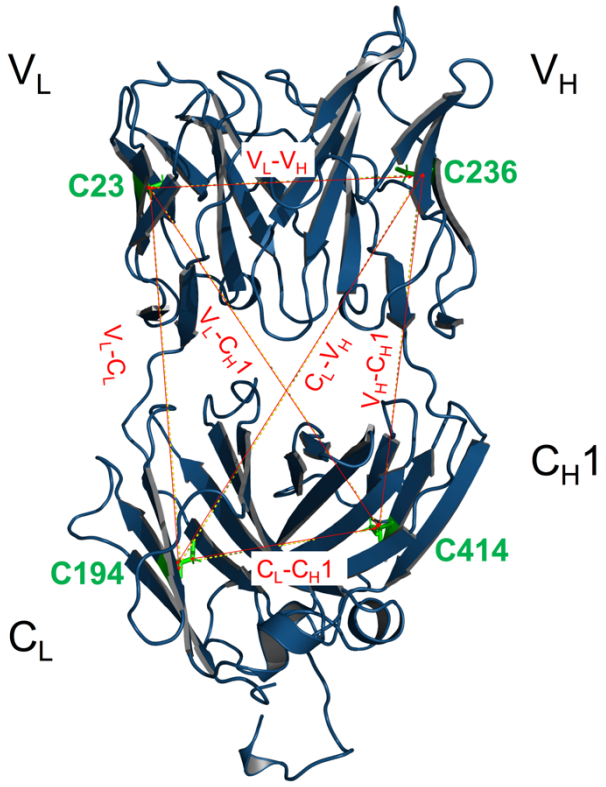

Figure 4.10. Location of the inter-domain distances studied in Table 4.1, in the Fab A33 structure. Six distances were measured between the four Fab domains (V_L-V_H, C_L-C_H1, V_L- C_L, V_H- C_H1, V_L-C_H1 and V_H- C_L) using the four cysteines located in outer ß-strands (C23, C194, C236, C414), for the ten best fist SAXS structures at pH 7.0 and 3.5.

Table 4.1 Inter-domain distance differences between the best SAXS fit structures at pH 7.0 and 3.5, using one cysteine in each domain (V_L, V_H, C_L and C_H1). Six distances were monitored between the four Fab domains (V_L-V_H, C_L-C_H1, V_L- C_L, V_H- C_H1, V_L-C_H1 and V_H- C_L) using the four cysteines (C23, C194, C236, C414) located in the outer ß-strands. These are shown in the Fab A33 structure in Figure 4.10. Averages and SEM shown.

Distance (nm)	pH 7.0	pH 3.5	ΔpH (3.5 - 7.0)
C23 (V_L) - C236 (V_H)	3.00 ± 0.02	2.95 ± 0.01	-0.05 ± 0.03
C194 (C_L) - C414 (C_H1)	2.47 ± 0.02	2.52 ± 0.01	0.05 ± 0.02
C23 (V_L) - C194 (C_L)	4.18 ± 0.02	4.47 ± 0.06	0.29 ± 0.07
C236 (V_H) - C414 (C_H1)	4.11 ± 0.07	4.09 ± 0.06	-0.02 ± 0.09
C23 (V_L) - C414 (C_H1)	4.68 ± 0.06	4.64 ± 0.02	-0.05 ± 0.06
C236 (V_H) - C194 (C_L)	4.99 ± 0.02	5.26 ± 0.02	0.27 ± 0.03

4.4.5 Identification of aggregation-prone regions (APR) suggests an aggregation mechanism

APRs in Fab A33 were determined using sequence-based APR detectors (done in Chapter 3), as I have already identified the experimental solution conformations of Fab A33 in different solution conditions via SAXS atomistic modelling. Thus, the APR predictions were combined with the best experimentally identified structures at pH 7.0 and 3.5 to identify differences in their solvent exposure. Four sequence-based APR predictors were used, PASTA 2.0 (Walsh et al. 2014), TANGO (Fernandez-Escamilla et al. 2004), AGGRESCAN (Conchillo-Solé et al. 2007) and MetAmyl (Emily et al. 2013), and their findings were confirmed with the consensus tool Amylpred2 (Tsolis et al. 2013). Seven APRs were identified in Fab A33 namely residues 31-36, 47-51, 114-118 and 129-139 in the light chain and residues 261-165, 325-329 and 387-402 in the heavy chain.

To display the aggregation-prone regions on the Fab A33 homology model as shown in Figure 4.11a, each aggregation propensity was normalized between 0 and 1, and weighted equally. Red represented high aggregation propensities and blue low aggregation propensities. The seven APRs were co-located as three regions of largely buried β-strands within the folded structure, and all were protected from the solvent. Next, the difference in solvent accessibility of the APRs in the SAXS best-fit structures at pH 7.0 and pH 3.5 were analysed. The solvent accessibility of one APR visibly increased at pH 3.5 due to the displacement of the C_L domain (circled; Figure 4b,c). Quantitatively, the SASA of the seven APRs were calculated for the ten best-fit structures at pH 7.0 and pH 3.5, and summed (Table 4.2). While most APR showed small decreases in solvent accessibility, the APR at residues 387-402 increased by 83 $Å^2$ from 536 ± 43 $Å^2$ at pH 7.0 to 619 ± 39 $Å^2$ at pH 3.5 (3% increase), due to the displacement of the C_L domain at low pH. These data illustrate the potential of combining biophysical methods that determine conformational changes, with sequence-based APR prediction tools, for determining aggregation hotspots. For Fab A33, the aggregation prediction tools suggested a possible molecular explanation for the observed increase in aggregation at low pH as the result of structural instabilities.

Table 4.2 Comparison of the solvent accessible surface area (SASA) for the most aggregation-prone regions in Fab A33 between pH 7.0 and pH 3.5. The SASA of the seven most aggregation-prone regions were computed using PyMol for the top ten SAXS best fit structures at each of pH 7.0 and pH 3.5. The Table reports absolute and relative solvent accessible surface areas, and the differences between the average SASA at the two pH values.

APR	Fab domain	SASA (Å^2) pH 7.0	SASA (Å^2) pH 3.5	ΔSASA (Å^2) pH(3.5-7.0)	SASA (%) pH 7.0	SASA (%) pH 3.5	ΔSASA (%) pH(3.5-7.0)
31-36	V_L	111 ± 20	92 ± 10	-18	12 ± 2	10 ± 1	-2%
47-51	V_L	107 ± 23	90 ± 17	-17	12 ± 3	10 ± 2	-2%
114-118	C_L	115 ± 19	90 ± 12	-25	14 ± 2	12 ± 2	-3%
129-139	C_L	158 ± 12	140 ± 10	-18	9 ± 1	8 ± 1	-1%
261-265	V_H	10 ± 4	14 ± 5	4	1 ± 0	2 ± 1	1%
325-329	V_H	139 ± 18	134 ± 14	-6	23 ± 1	15 ± 2	-8%
387-402	C_H1	536 ± 43	619 ± 39	83	22 ± 2	25 ± 2	3%

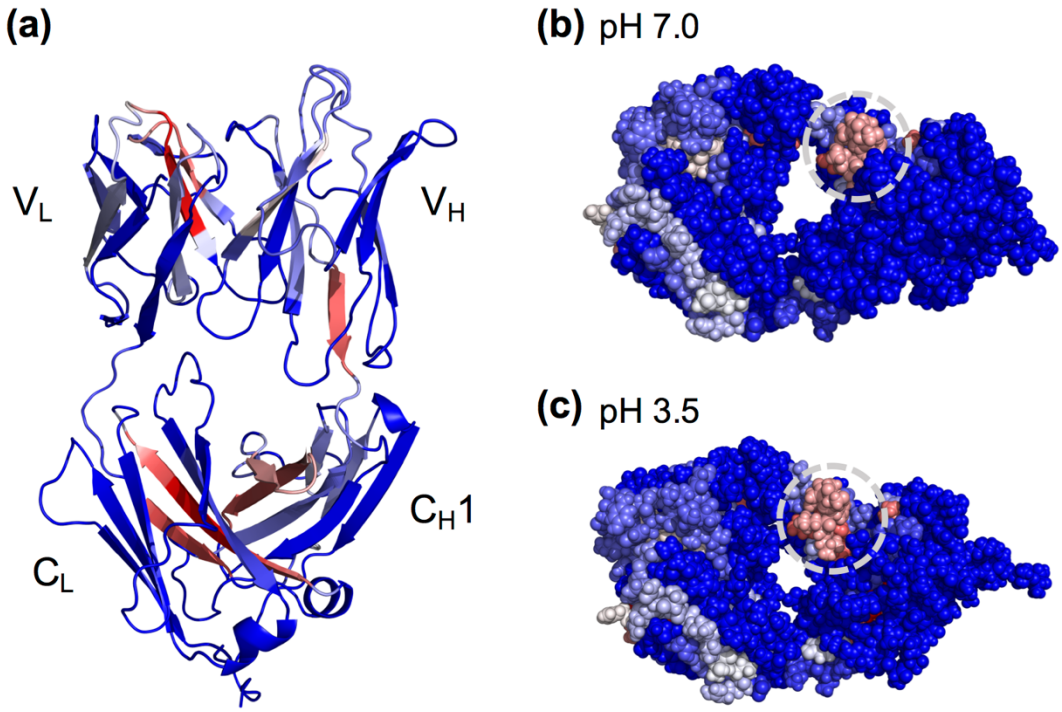

Figure 4.11. Aggregation prone regions in Fab A33. (a) The consensus aggregation propensity of residues in Fab A33 was determined using PASTA 2.0, TANGO, AGGRESCAN and MetAmyl software. Using the native Fab A33 homology model, regions with greater aggregation propensities are shown in red and reduced propensities in blue. (b, c) Aggregation propensities in the SAXS best-fit structure for pH 7.0 and 3.5, respectively, are shown using a CPK spheres representation. The circled residues highlight the increase in SASA of APR 387-402 at pH 3.5 compared to pH 7.0.

4.5 Conclusions

Characterization of the aggregation competent states is necessary to develop a rigorous understanding of the aggregation mechanism and to provide insights into possible approaches to rationally design candidate therapeutics. However, such aggregation-prone conformations for near-native solution conditions have proven most challenging to characterize over the years, and have remained elusive within unmutated native-protein ensembles. In this chapter, I characterized the structural perturbations that take place within the native ensemble of the humanized antibody Fab A33 over a range of different pH and ionic strengths, using SAXS atomistic modelling. Our data inferred the existence of an expanded aggregation-prone conformation of Fab A33, which adopted a more expanded conformation at acidic pH (5.5, 4.5 and 3.5) compared to neutral pH (7.0 and 9.0), with an increase in the R_g of between 2.2% and 4.1%. The presence of this expanded conformation coincided with accelerated aggregation, as small amounts of aggregates were also detected by SAXS and R_g values correlated with previously measured aggregation kinetics, indicating that this expanded species is aggregation prone. To gain insight into the structure of the expanded conformation, SAXS data was combined with atomistic structures generated using MD simulation at the same conditions. Results revealed a displacement of the constant domain of the light chain (C_L) at low pH. This finding adds to the increasing amount of evidence suggesting that aggregation at near native conditions takes place through a state that is only slightly perturbed in structure relative to the native state.

To explain the increased aggregation propensity of the expanded conformations of Fab A33, I used online software to predict the aggregation-prone regions (APR) that are more likely to form cross-β structures found in aggregates. Results showed that all predicted APR were buried in the interior of the protein; however, the SASA of one of them increased with the displacement of C_L at low pH. Based on these findings, I propose an aggregation mechanism for Fab A33. Aggregation takes place through the formation of an aggregation-pone intermediate first, which is characterised by being native-like in structure but expanded relative to the native state (Chiti & Dobson 2009; Bemporad & Chiti 2009). This aggregation-prone intermediate has regions with increased flexibility and increased total SASA. The initial oligomers formed would thus retain high structure similarity to the native state. We hypothesize that in later stages of the aggregation

process, a structural re-arrangement takes place to form the typical cross-β structure of amyloids, as indicated in previous studies (Krishnan et al. 2002; Orte et al. 2008; Iljina et al. 2016). Future work to confirm this proposed aggregation mechanism, could include reducing the aggregation propensity of the exposed APR. Results from this chapter also provided experimental confirmation to the findings from molecular dynamic simulations in Chapter 3. There are several strategies to stabilize the C_L domain, such as mutation of the salt bridges identified by MD simulations (Glu165-Lys103 and Glu195-Lys149), stabilize the constant domain interface as suggested by FoldX and Rosetta to more hydrophobic residues, or lower the aggregation propensity of the predicted exposed APR.

Collectively, this work provides compelling evidence of how local unfolding can lead to transiently-formed structural conformers within the native ensemble that promote aggregation. It also highlights the promise of SAXS combined with molecular dynamics simulations to resolve aggregation-prone conformers within native ensembles, particularly for large proteins that are less accessible by NMR. This also provides a new route to gaining molecular level knowledge of potential target sites for the rational engineering of more stable proteins, either via protein engineering or formulation, or for the design of drugs that bind to and stabilize proteins against aggregation in vivo.

Chapter Five

Characterization of the aggregation-prone conformation of Fab A33 at low pH using single-molecule FRET as an orthogonal technique

5.1 Summary

In this chapter, the C_L domain unfolding observed at low pH in Fab A33 by MD simulations and atomistic modelling of SAXS data, was confirmed using single-molecule Forster Resonance Energy Transfer (smFRET), as an orthogonal method. The non-radiative energy transfer between donor and acceptor fluorophores in FRET is very sensitive to distance changes, specially in the 2 to 10 nm range, which makes smFRET suited to study conformational changes in proteins. Two dual-labelled Fab A33 constructs were generated to probe an intra-C_L separation and a separation between the C_L domain and the heavy-chain linker. Specifically, these were (Dist 1) LC-K126pAzF + LC-S156C, and (Dist 2) HC-S117pAzF + LC-S156C. Each construct contained one nonstandard amino acid (p-azido-l-phenylalanine) and one solvent-exposed cysteine, to attach the fluorophores Alexa Fluor 488 DIBO Alkyne (donor) and Alexa Fluor 594 Maleimide (acceptor), respectively. The confocal detection of freely diffusing molecules was used to obtain the apparent transfer efficiency histograms (E_{app}) of Fab A33 at pH 7.0 and 3.5. smFRET revealed that Dist 1 was unchanged between pH 7.0 and 3.5, while the distance between C_L and the heavy chain linker (Dist 2) increased at pH 3.5, with a decrease in FRET efficiency from E_{app} = 0.87 at pH 7.0 to E_{app} = 0.78 at pH 3.5, confirming the partial unfolding of Fab A33 at low pH. Additionally, the values obtained for apparent transfer efficiencies were highly correlated with distances measured from the best models derived from SAXS and MD simulations. Both methods found that Dist 1 did not change with pH, being 2.5 nm at pH 7.0 and pH 3.5, while Dist 2 increased from 2.8 nm at pH 7.0 to 3.5 nm at pH 3.5. Taken together, the displacement at low pH of the C_L domain was validated by three independent detection methods.

5.2 Introduction

As mentioned in the previous chapters, characterizing the conformational changes in the native state that give rise to aggregation-prone species, is crucial to understanding the mechanisms of protein aggregation. This will allow the search for protein aggregation inhibitors to be done in a rational way, and to ultimately design proteins more robust to aggregation. SAXS revealed that at ambient temperature Fab becomes conformationally expanded in acidic solutions, showing a 2.2% to 4.1% increase in the R_g of the species at pH 5.5 and below; with the presence of these species coinciding with accelerated aggregation. Atomistic modeling of SAXS data and MD simulations revealed that aggregation proceeded through a transient partial unfolding of the native state located on the light chain constant domain (C_L) of Fab A33. To validate these findings, in this chapter I used confocal single molecule detection of FRET-labelled Fab A33 to characterize the partially unfolded aggregation-prone conformer at low pH, as the combined used of single molecule detection and FRET, enables conformational changes in single molecules to be elucidated.

FRET, which stands for Förster resonance energy transfer (or fluorescence resonance energy transfer), involves the energy transfer from an excited donor fluorophore to an acceptor fluorophore through a non-radiative dipole-dipole coupling (Schuler 2013). The efficiency of this energy transfer is inversely proportional to the sixth power of the donor-acceptor distance, thus making FRET extremely sensitive to small changes in the 2 to 10 nm distance range. FRET often receives the name “molecular ruler”, and it is useful to study conformational changes in proteins. The use of single-molecule detection offers the possibility to observe molecules one at a time, and thus, capture molecules that only represent a small fraction of the total number of molecules present, or species that might only be populated briefly (Lerner et al. 2018).

A typical smFRET experiment entails labelling the protein with a donor and acceptor fluorophores at specific locations in the protein, so the distance between them can be monitored (Roy et al. 2008). A very small detection volume (<1 fL) is generated by combining a high numerical aperture objective that focuses the excitation laser beam to a diffraction-limited focal spot, with confocal detection. Very dilute concentrations of the protein are used (10-100 pM), to guarantee that statistically no more than one protein

molecule is present in the detection volume at the same time. During an experiment, protein molecules diffuse freely in solution. When a protein molecule traverses the detection volume, the donor fluorophore is excited, which can transfer energy to the acceptor dye, and both emit fluorescence photons. Donor and acceptor photons are separated by wavelength and detected using single-photon avalanche diode (SPAD) detectors. From the resulting photon record, donor and acceptor fluorescence intensities for each transit event can be extracted, and energy transfer efficiencies calculated. Energy transfer efficiencies can then be related to the distance between the fluorophores, and to different conformational states of the protein (Schuler & Eaton 2008; Hillger et al. 2007).

Not many studies exist on the application of smFRET to study aggregation. Most of these studies have focused on the early sensitive detection of disease related aggregates, and the characterization of the aggregation process by looking at the time-dependent distributions of oligomers, and how this was altered by additional factors. However, few studies exist on the identification and characterization of misfolded proteins as precursors of aggregation. The existing works have focused on disease-related proteins, such as α-synuclein, an intrinsically disordered protein linked with Parkinson's disease. One study on the characterization aggregation-prone states of α-synuclein found that different aggregation conditions affected α-synuclein differently. Low pH promoted the collapse of the C-terminus as indicated by high efficiency between a FRET pair in this region, while positively charged molecules (also shown to promote aggregation) showed only minor effects, suggesting an influence later in the aggregation process (Trexler & Rhoades 2010). Another study found that the binding of α-synuclein to membranes modulated conformational transitions between a natively unfolded state and multiple α-helical structures, which implied that two folded structures are pre-encoded by the α-synuclein amino acid sequence (Ferreon et al. 2009). Lastly, a study comparing wild-type α-synuclein to a disease-associated point mutant of α-synuclein, found that they populated different ensembles, with the wild-type adopting an elongated structure at mM SDS concentrations and the mutant was more flexible and less structures, which may have implications to why this mutant leads to disease (Ferreon et al. 2010).

FRET analysis of protein conformational changes requires a pair of donor and acceptor fluorophores to be attached to the protein at specific locations. Proteins are made of 20 amino acids, and only two of them present sufficient reactivity, the sulfhydryl group of cysteine and the amino group of lysine and the N-terminal amino acid. Most proteins

contain many lysine residues, which makes lysine not suitable for smFRET experiments as proteins would contain multiple labels on them. Cysteine is the most common group used for site-specific fluorophore attachment. Many proteins do not contain any cysteine, what implies that upon introduction of one, a single fluorophore can be attached at that position. If the protein already contains cysteine groups, there are two scenarios. One scenario is that the cysteine groups are not essential for the structure of the protein, and can be mutated to serine groups. Then, a new engineered cysteine can be introduced for labeling. In the second scenario, and the case for Fab A33, many cysteine groups are present in the protein and are important to the structure. In this case nonetheless, advantage can be taken of the different reactivities that cysteines in different locations on the protein possess. Fab A33 contains five disulfide bonds. Four of them are intra-domain (stabilizing the fold of V_L, V_H, C_L and C_H1), and thus are located in the interior of Fab and are not solvent exposed. The fifth disulfide bond bridges the light and heavy chains, and is located at the C-terminal of the constant domains, before the hinge region. In this work, I incorporated an additional solvent exposed cysteine in Fab A33 to attach the acceptor fluorophore. Site-specific labelling was achieved by using mild denaturant conditions, which reduce the solvent-exposed cysteine and potentially the inter-chain disulfide bond, however, by providing time to the protein to re-form the inter-chain disulfide bond after removal of the denaturant and addition of the fluorophore, the correct labelling of Fab A33 was achieved.

To site-specifically attach the donor fluorophore, a reactive nonstandard amino acid (NSAA) was incorporated into Fab A33, p-azido-l-phenylalanine (pAzF) (Figure 5.1). To incorporate a 21st amino acid into a protein, several microbiology components need to be re-engineered (Liu & Schultz 2010). In the cell, translation of mRNA into a protein is done through adapter molecules called transfer RNA (tRNA), which in one end recognize the DNA triplet codon and on the other have attached the corresponding amino acid. Attachment of the amino acid to the tRNA molecule is facilitated by a protein called aminoacyl-tRNA synthetase (aaRS), which recognizes the specific tRNA to its specific amino acid and binds them together. Thus, to incorporate a 21st amino acid, an engineered aaRS / tRNA pair that incorporates the new NSAA in response to a reassigned codon, are needed. To reassign a codon for the new NSAA, we used the amber stop codon (UAG). Amber stop codon is the least used stop codon in *E. coli* (approx. 7% of the time), and it has been shown that cells still grow well after reassignment of this codon to encode a new NSAA. However, in this work I used the engineered *E. coli* C321.ΔA.exp (ID: 49018),

where all 321 UAG terminations have been removed, and replaced by UAA (Lajoie et al. 2013). An orthogonal aaRS / tRNA pair needs to be engineered, where the tRNA is not aminoacylated by any endogenous aaRS and the new aaRS does not aminoacylate any of the endogenous tRNAs. The most successful approach has been to import the pair from a different domain of life, since they have different identity elements. The most commonly used pair is aaRS / tRNATyr from *Methanococcus jannaschii*. First, the tRNA anticodon loop is mutated to CUA. Next, tRNA is engineered to not cross react with any endogenous aaRS. This is achieved by creating a library of mutant tRNA (aa that do not directly interact with aaRS) and submitting them to rounds of positive and negative selection to identify orthogonal tRNA. Finally, aaRS is engineered to only recognize the UAA. This is achieved as well by creating a library of mutant aaRS (randomize residues in the aa binding site) and submitting them to rounds of positive and negative selection.

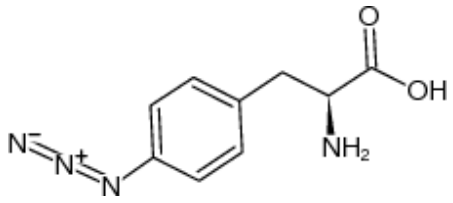

Figure 5.1. Structure on the nonstandard amino acid p-azido-l-phenylalanine (pAzF). pAzF was used for the site-specific labelling of Fab A33 via click chemistry with a fluorophore containing an alkyne moiety.

In this chapter, an aggregation-prone conformer of Fab A33 at low pH was characterized using smFRET. Two dual-labelled Fab A33 constructs were generated, by recombinant incorporation of an additional solvent exposed cysteine and the NSAA pAzF, in each construct. The correct incorporation of one donor and one acceptor fluorophore at each construct were confirmed with ESI mass spectrometry and UV-vis absorption. Unfolding of Fab A33 was first studied using the denaturant guanidinium chloride (GdmCl), as a control for smFRET studies. Then, the two constructs were studied at pH 7.0 and 3.5, to confirm the displacement of C_L at low pH. Notably, the apparent transfer efficiencies (E_{app}) obtained here correlated highly with the distances obtained from SAXS and MD simulations.

5.3 Methods

5.3.1 Cloning to generate Fab A33 mutants for smFRET

Fab A33 mutants with one nonstandard amino acid (NSAAs), p-azido-l-phenylalanine (pAzF) (Figure 5.1), and one engineered solvent-exposed cysteine were generated to allow attachment of donor and acceptor fluorophores. Two different constructs were generated: (i) LC-K126pAzF + LC-S156C and (ii) HC-S117pAzF + LC-S156C. To incorporate pAzF, two plasmids need to be co-transformed into *E. coli*, the plasmid pTTOD encoding for Fab A33 and the plasmid encoding for the machinery necessary to incorporate pAzF (aaRS / tRNA pair). I used the plasmid pEVOL-pAzF (Plasmid ID: 31186) (Addgene, Cambridge, USA), which encodes an engineered tyrosyl-tRNA synthetase (aaRS) and an amber suppressor tRNA ($tRNA_{CUA}$), derived from *Methanococcus jannaschii*, to incorporate pAzF in response of the amber stop codon (Figure 5.2) (Young et al. 2010; Lim et al. 2015). In order for both plasmids to remain in the dividing *E. coli* cells, the antibiotic resistance and origin of replication (ORI) of both plasmids need to be different. As pTTOD and pEVOL-pAzF had the same origin of replication, p15A, the tac promoter and Fab A33 gene from pTTOD were sub-cloned into pET-29a(+), which has a ColE1 origin of replication, using circular polymerase extension cloning (CPEC) (Figure 5.3) (Quan & Tian 2011). The gene for Fab A33 encodes light and heavy chains separately, and each chain contains an ompA signal sequence in the N-terminal, to allow translocation of the protein to the cellular periplasm (once there, the signal sequence is cleaved by peptidases). CPEC primers were designed using the Gibson/Assembly option in SnapGene, and were (Eurofins, Wolverhampton, UK):
(Insert.REV)
GGCTTTGTTAGCAGCGATATGACGACAGGAAGAGTTTGTAGAAACG
(Vector.REV) TTCCTGTCGTCATATCGCTGCTAACAAAGCCCGAAAGG
(Insert.FOR) TGATGTCGGCGATACCATCGGAAGCTGTGGTATGG
(Vector.FOR) CAGCTTCCGATGGTATCGCCGACATCACCGATGGG.

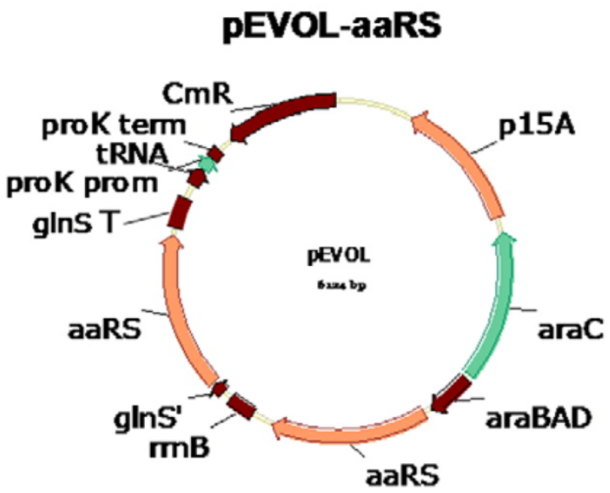

Figure 5.2. Map of the plasmid pEVOL-pAzF (Plasmid ID: 31186 from Addgene). It encodes two copies of an engineered tyrosyl-tRNA synthetase (aaRS) one under control of an arabinose inducible promoter and the other constitutively expressed, and a single copy of the amber suppressor tRNA ($tRNA_{CUA}$) a p15A origin of replication, the chloramphenicol acetyltransferase marker (CmR), and the araC repressor gene (araC), to incorporate pAzF in response of the newly introduced amber stop codon. Image obtained from (Young et al. 2010).

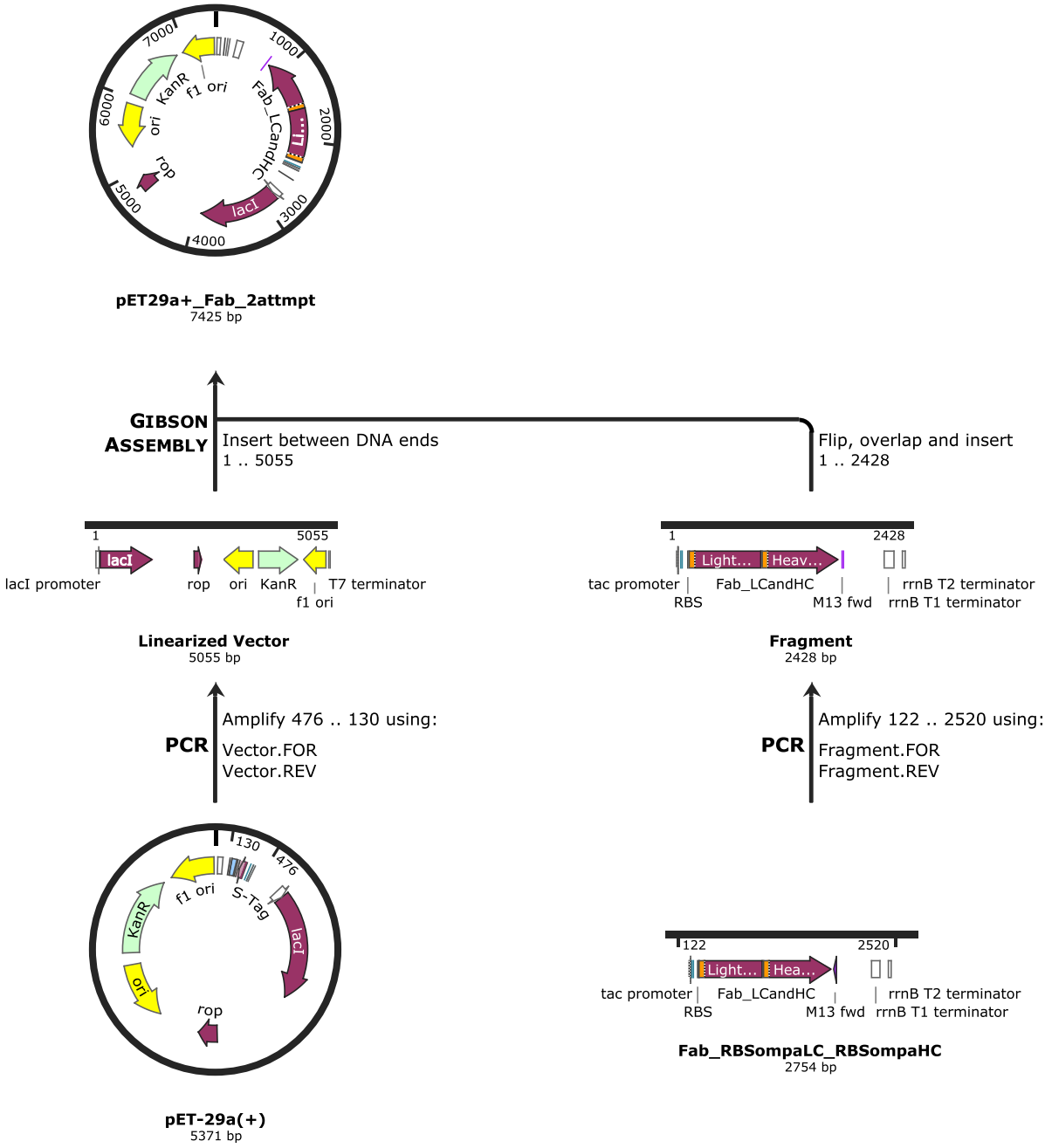

Figure 5.3. Schematic of the cloning steps followed to clone Fab A33 gene into pET-29a(+). The insert is 2428 bp and contains the tac promoter and Fab A33 gene (RBS, ompA, LC and RBS, ompA, HC). The vector is 5055 bp and contains the kanamycin resistance gene and ColE1 ORI. Image generated with SnapGene software (from GSL Biotech).

The insert and vector were amplified by PCR using the CPEC primers and purified using QiaQuick gel purification kit (Qiagen, Hilden, Germany). Final assembly was achieved mixing insert and vector with overlapping fragments in a 30:1 (insert:vector) ratio, with 100 ng of vector and 10 cycles. The assembled product was directly transformed into NEB 10β competent cells (New England Biolabs, Ipswich, US). Final assembly was confirmed by sequencing using Source Bioscience (UK), and the following primers (Eurofins, Wolverhampton, UK):
(pET29Fab_for) AGGAATGGTGCATGCAAGG
(pET29Fab_mid) AGTGGAAGGTGGATAACGC
(T7 term) CTAGTTATTGCTCAGCGG

Following cloning, site-specific mutations were introduced using QuickChange Lightning Site-Directed Mutagenesis (Agilent Technologies, Santa Clara, USA) to form the double mutants: (i) LC-K126pAzF + LC-S156C and (ii) HC-S117pAzF + LC-S156C. In order to incorporate pAzF, I mutated the native codon to the amber stop codon (TAG). Primers for site-directed mutagenesis were designed using the mutagenesis option in SnapGene, and were (Eurofins, Wolverhampton, UK):
(LC-K126pAzF (AAA to TAG).FOR)
CCATCTGATGAGCAGTTGTAGTCTGGAACTGCCTCTG
(LC-K126pAzF (AAA to TAG).REV)
CAGAGGCAGTTCCAGACTACAACTGCTCATCAGATGG
(LC-S156C (TCG to TGC). FOR)
GGATAACGCCCTCCAATGCGGTAACTCCCAGGAG
(LC-S156C (TCG to TGC). REV)
CTCCTGGGAGTTACCGCATTGGAGGGCGTTATCC
(HC-S117pAzF (TCT to TAG).FOR)
CACTGGTGACAGTGTCTTAGGCCTCAACGAAGGGC
(HC-S117pAzF (TCT to TAG).REV)
GCCCTTCGTTGAGGCCTAAGACACTGTCACCAGTG

Lastly, introduction of the mutations was confirmed by sequencing using Source Bioscience (UK), and the following primers (Eurofins, Wolverhampton, UK):
(Mutations_LC) TCATCTATTTGGCCTCCAAC,
(Mutations_HC) TGTGCAGCATCTGGATTC

5.3.2 Expression and purification of Fab A33 mutants

As an expression host, I used the genomically engineered "amberless" *E. coli* (C321.ΔA.exp) (ID: 49018), from Addgene (Cambridge, MA). C321. ΔA.exp was created to improve the incorporation of nonstandard amino acids, by recoding E. *coli* MG1655 strain to have all 321 UAG terminations removed (replaced by UAA) and the release factor 1 (RF1) gene deleted. (Lajoie et al. 2013) Thus, NSAA incorporation does not compete with the termination of translation anymore and it doesn't interfere with cellular process. I co-transformed C321.ΔA.exp with pEVOL-pAzF and pET-29a containing mutant Fab A33. Then, Fab A33 WT and mutants were expressed and purified as described in Chapter 2 (Materials and Methods) section 2.3.2 (Protein expression and purification).

5.3.3 Site-specific labelling of Fab A33

Fab A33 was buffer-exchanged into PBS using 10 kDa cut-off centrifugal filters (Merck, Kenilworth, UK) and adjusted to 0.5 mg/mL. The donor fluorophore dibenzocyclooctyne Alexa Fluor 488 (Thermo Fisher Scientific, Waltham, USA), (Figure 5.4a), was reacted using click chemistry at a 5:1 molar ratio (fluorophore:protein) for 24 h at room temperature with gentle shaking in the dark (Figure 5.5a). To attach the acceptor to the Fab solvent-exposed cysteine using maleimide-thiol chemistry, TCEP was added to 0.5 mM (50-fold molar excess of TCEP to Fab) and incubated for 1.5 h at room temperature (Jevševar et al. 2012). This step regenerates the free cysteine. TCEP is then removed by buffer exchange into PBS and incubation for 24 h to allow reconstitution of the correct disulfide-bridges. Maleimide-activated Alexa Fluor 594 (Figure 5.4b) was added in a 5:1 molar ratio of fluorophore:protein, and incubated for 16-18 h at room temperature (Figure 5.5b). 10 kDa centrifugal filters were used to remove the unreacted dye. The correct labelling of constructs i and ii was confirmed using ESI mass spectrometry and UV-vis absorption.

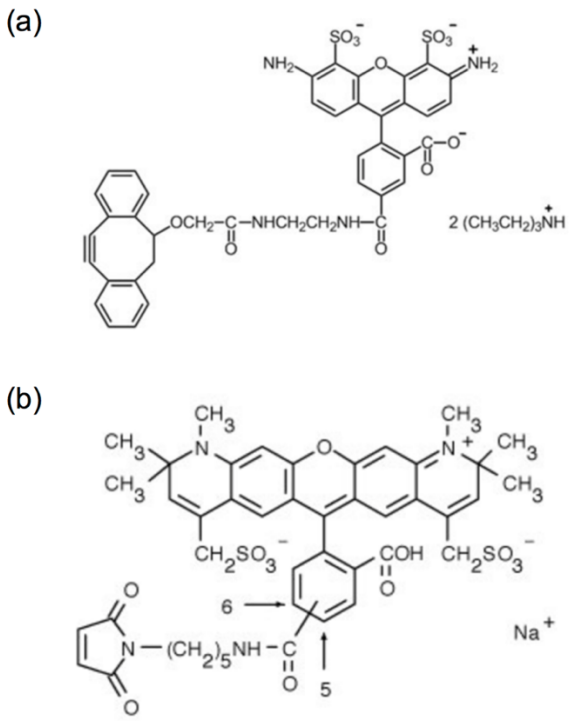

Figure 5.4. Fluorophore structures. (a) Alexa Fluor 488 DIBO alkyne (donor); (b) Alexa Fluor 594 C5 Maleimide (acceptor).

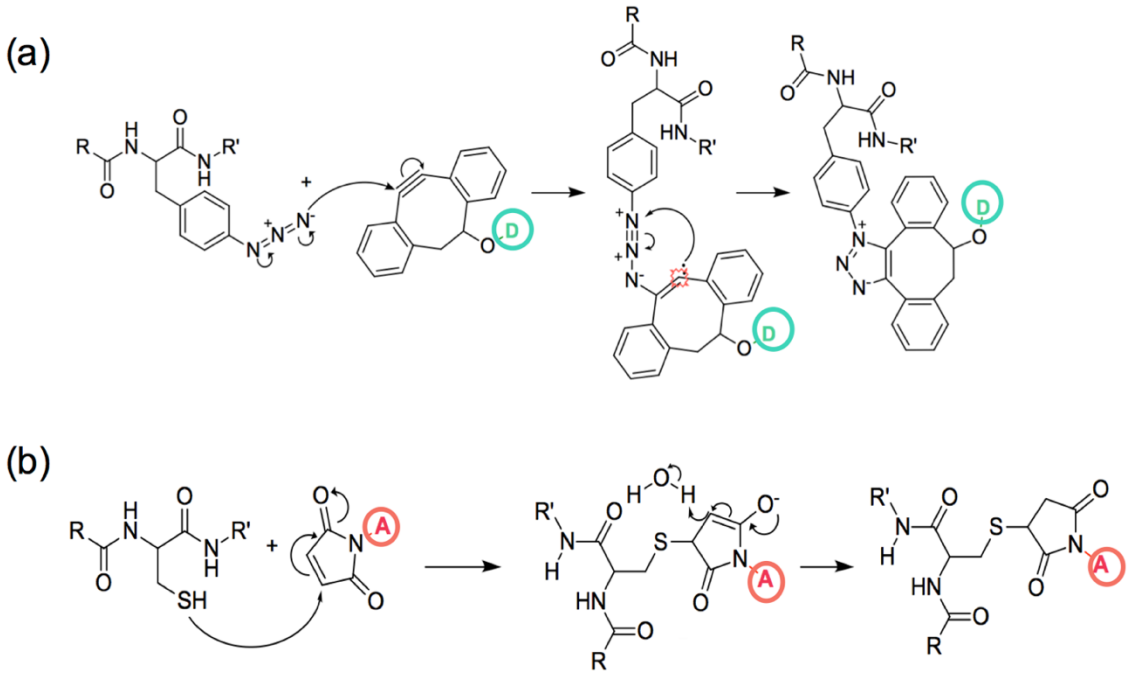

Figure 5.5. Reactions for the site-specific attachments of fluorophores to the protein. (a) click azide/DIBO reaction, where D symbolizes the donor fluorophore; (b) maleimide/thiol reaction, via Michael-type conjugate addition, to form a thioether. A symbolizes the acceptor fluorophore. Reactions drawn with ChemDraw software.

5.3.4 Acquisition of smFRET data using confocal fluorescence spectroscopy

Single-molecule fluorescence measurements were carried out on a MicroTime 200 confocal microscope (PicoQuant, Germany). For excitation, a diode laser at the donor excitation wavelength was used (LDH-D-C-485, PicoQuant, Germany), at 20 MHz (laser pulse every 50 ns) and a laser power of 100 μW at the back aperture of the objective. The laser was focused into the sample solution with an UPlanApo 60x/1.20W objective (Olympus). Measurements were performed by placing the confocal volume 50 μm into the solution relative to the cover slide surface. The fluorescence signal was collected by the same objective and filtered with a 485/595 dual-band dichroic mirror (Chroma Technology). Afterwards, the photons passed through a 100 μm pinhole. Donor and acceptor photons were separated by a second dichroic mirror, 585 DCXR, and further filtered by band-pass filters, ET525/50M for donor, and ET645/75M for acceptor (all Chroma Technology). Finally, photons were detected using two single-photon avalanche photodiodes (SPAD) (PicoQuant). The arrival time of every detected photon was recorded with a HydraHarp 400 counting module (PicoQuant).

Single-molecule measurements were acquired at a protein concentration of <100 pM. The measurements were performed in 20 mM sodium phosphate buffer pH 7.0 and 20 mM sodium citrate buffer pH 3.5, both 50 mM final ionic strength adjusted with NaCl. Despite the low pH, the fluorescence quantum yields of the dyes remained the same (Hofmann et al. 2013). Each sample was measured for 30 min at room temperature.

5.3.5 Analysis of smFRET data

First, the raw data was converted to the Photon-HDF5 file format (.h5) using the open source software Photon-HDF5 (Ingargiola, Laurence, et al. 2016). Next, single-molecule FRET data was analyzed using the open source software FRETBursts (Ingargiola, Lerner, et al. 2016). I followed the steps for background estimation, burst search, burst selection and computation of FRET efficiency histograms. Background rates were calculated first, by plotting a histogram of inter-photon delay times in windows of 30 s. Signal from single-molecules can be differentiated from background because single molecules show short delay times whereas background signal follows a Poisson process that is exponentially distributed. By fitting the long delay times to an exponential, the

background rates were calculated. After background calculation, bursts corresponding to single-molecules traversing the excitation volume were identified. A burst was identified if the rate of photons was 6 times faster than the local background rate, and I used 10 consecutive photons to compute the local count rate. For this calculation, all photons were taken into account (donor and acceptor). After burst identification, corrections were applied. Bursts were corrected for background and donor leakage into the acceptor channel, the later was calculated to be 8%. No acceptor direct excitation and γ-factor correction were applied, thus the conversions of FRET efficiencies to distances was not possible. In this study, I refer to the calculated FRET efficiencies as apparent FRET efficiencies (E_{app}) (Roy et al. 2008; Majumdar et al. 2007; Roy et al. 2009), which allowed the relative comparison between Fab A33 constructs and solution conditions. A size filter was applied to the previous bursts found, where only bursts with more than 30 photons were kept. Lastly, apparent transfer efficiency histograms were calculated for each burst using the expression $E_{app} = n_A/(n_A + n_D)$; where n_D and n_A are the corrected numbers of donor and acceptor photons in the burst, respectively, and apparent FRET efficiencies were fitted to Gaussian functions.

5.4 Results and discussion

5.4.1 Characterization of Fab A33 mutants using Mass Spec and UV-Vis absorption

I generated two constructs to probe the distance intra-C_L domain and between C_L and the heavy chain. Specifically, the constructs were (Dist 1) LC-K126pAzF + LC-S156C, with both fluorophores attached to the C_L domain, and (Dist 2) HC-S117pAzF + LC-S156C, with one fluorophore in the C_L domain and one in the heavy chain linker between variable and constant domains (Figure 5.6). Each construct contained one nonstandard amino acid, pAzF, and one solvent-exposed cysteine, to attach the fluorophores Alexa Fluor 488 (donor) and Alexa Fluor 594 (acceptor), respectively. Certain requirements were taken into consideration when selecting the labelling positions. Fluorophores should initially be positioned at a shorter distance than the Forster radius (R_0, distance at which the energy transfer efficiency is 50%) for that pair of fluorophores, so that an increase in distance due to unfolding can still be captured. The Forster radius for the pair AF-488/AF-594 is 6 nm, and the separation between dyes (Dist 1 and Dist 2) was chosen to be between 2-3 nm. Additional considerations were that the labelling positions were solvent exposed, to increase the labelling efficiency and minimize the effect to the native structure of the protein, and lastly, the selected residues to be mutated did not participate in stabilizing interactions in Fab A33.

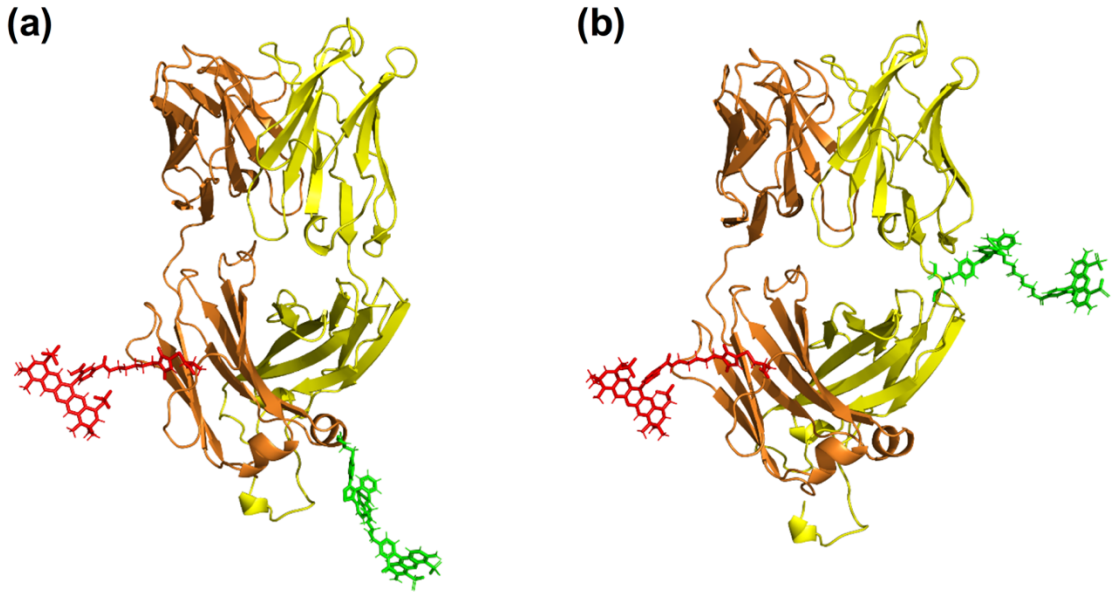

Figure 5.6. Cartoon representation of dual-labelled Fab A33 constructs. (a) (Dist 1) LC-K126pAzF + LC-S156C, to probe an intra-C_L separation. Fab A33 labelled with the donor fluorophore AF-488 DIBO (green) at LC-K126pAzF and the fluorophore acceptor AF-594 maleimide (red) at LC-S156C. The fluorophores are shown in an arbitrary orientation. (b) (Dist 2) HC-S117pAzF + LC-S156C, to probe the distance between C_L and the heavy-chain linker region. Fab A33 labelled with the fluorophore donor AF-488 DIBO (green) at HC-K117pAzF and the acceptor fluorophore AF-594 maleimide (red) at position LC-S156C. The fluorophores are shown in an arbitrary orientation.

First, the expression of wild-type C226S Fab A33 was confirmed using ESI mass spectrometry, at the expected mass of 47,385 g/mol (Figure 5.7a). Incorporation of p-azidophenylalanine was confirmed in the mutant LC-K126pAzF Fab A33 (Figure 5.7b), where the mass increase to 47,445 g/mol corresponded to the mutation of a lysine to pAzF. To confirm the labelling of double-labelled constructs, two experiments were set up. In both, I used the construct termed Dist 1, where Fab A33 was labelled with the fluorophore AF-488 DIBO (donor) at position LC-K126pAzF and the fluorophore AF-594 maleimide (acceptor) at position LC-S156C. Most probably, the additional solvent-exposed cysteine added to Fab A33 was cysteinylated during cell disruption, and thus reduction with a reducing agent is necessary to reconstitute the free thiol group for labelling. The risk of mild reduction is that the disulfide bond bridging light and heavy chains, before the hinge region, might get reduced too, due to its solvent accessibility. This was confirmed in the first scenario, where fluorophore AF-594 maleimide was added at the same time to Fab A33 as the reducing agent TCEP (Figure 5.7c). AF-594 maleimide reacted with the cysteines of the inter-chain disulfide bond, thus separating the light and heavy chains, and resulting in peaks of mass around 25 kDa (Figure 5.7c). To avoid this, in the second scenario, a certain time (24 h) was given after the removal of the reducing agent and prior to the addition of AF-594 maleimide, for the inter-chain disulfide bond to re-form, which should happen quickly due to its spatial proximity (Pepinsky et al. 2011). Chain separation was now not observed, hence leading to peaks around 50 kDa (Figure 5.7d). This method is used to produce, among other products, mono-PEGylated Fabs, using a single hinge cysteine, located in the hinge region, after inter-chain disulfide bridge.

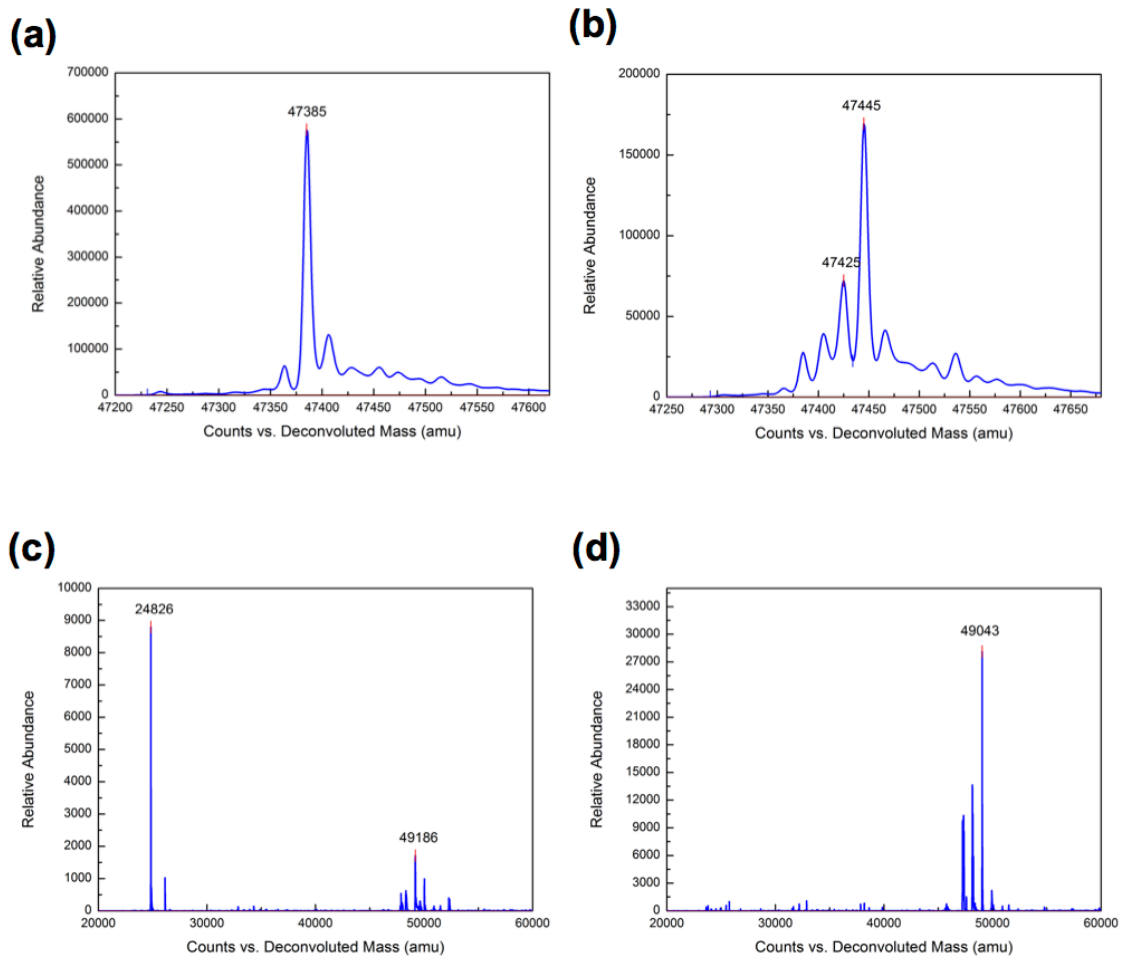

Figure 5.7. ESI mass spectrometry to confirm the labelling steps. (a) Wild-type C226S Fab A33 spectrum; (b) The spectrum of LC-K126pAzF Fab A33 to confirm the incorporation of the nonstandard amino acid p-azidophenylalanine; (c, d) Fab A33 was labelled with the fluorophore AF-488 DIBO (donor) at position LC-K126pAzF and the fluorophore AF-594 maleimide (acceptor) at position LC-S156C. In (c) AF-594 maleimide was added at the same time as TCEP to Fab A33, resulting in the labelling of the cysteines that formed the inter-chain disulfide bond, and the separation of light and heavy chains. In (d) AF-594 maleimide was added after TCEP removal and time was provided for the re-formation of the inter-chain disulfide bond, resulting in labelling of only the solvent-exposed cysteine.

Labelling of Fab A33 with only one donor and one acceptor fluorophores was first confirmed in the mass spectrum of Dist 1 and 2, at the expected mass of 49 g/mol (Figure 5.8a,c). The three peaks observed in the spectrums corresponded from left to right to Fab A33 with no label (~47 g/mol), Fab labelled with one fluorophore (donor or acceptor) (~48 g/mol), and Fab labelled with both fluorophores (~49 g/mol). Further confirmation was provided by the UV-vis absorption spectra (Figure 5.8b,d), where absorption of Fab A33 was seen around 280 nm, absorption of AF-488 at around 488 nm and absorption of AF-594 at around 594 nm, confirming the attachment of both fluorophores to Fab A33.

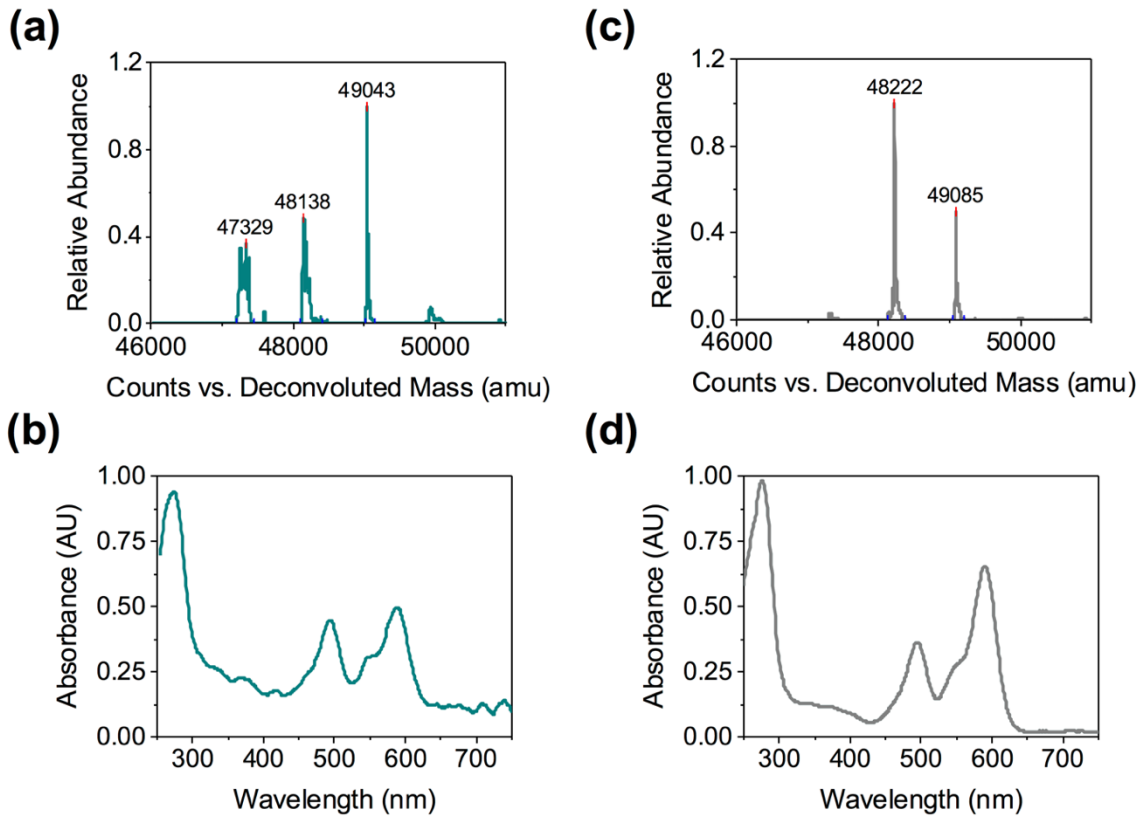

Figure 5.8. ESI mass spectrometry and UV-Vis absorption spectrum of the two double-labelled Fab A33 constructs for smFRET. (a, b) (Dist 1) Fab A33 labelled with the fluorophore AF-488 DIBO (donor) at LC-K126pAzF and the fluorophore AF-594 maleimide (acceptor) at LC-S156C. (a) ESI mass spectrometry (same graph as Figure 5.7d, zoomed), and (b) UV-Vis absorption spectrum. (c, d) (Dist 2) Fab A33 labelled with the fluorophore AF-488 DIBO (donor) at HC-K117pAzF and the fluorophore AF-594 maleimide (acceptor) at position LC-S156C. (c) ESI mass spectrometry and (d) UV-Vis absorption spectrum.

5.4.2 smFRET controls by unfolding Fab A33 using GdmCl as denaturant

FRET is the radiationless transfer of energy from a donor to an acceptor fluorophore, with the efficiency of this energy transfer being very sensitive to changes in distance in the 2-10 nm range. Every time a single molecule of Fab A33 labeled with donor and acceptor fluorophores, diffused through the detection volume of the confocal microscope, a burst of fluorescent photons was emitted. These photons were first separated into donor and acceptor emissions, and later detected with time-correlated single-photon counting electronics that recorded the arrival time of each photon. For every burst of photons, an apparent FRET efficiency (E_{app}) was calculated, which measures the fraction of photons absorbed by the donor that have been transferred to the acceptor. Lastly, apparent FRET transfer efficiency histograms were built from recording many individual events, which display maxima that correspond to subpopulations present in the sample. The peaks, which correspond to the E_{app} of that population, were measured from fitting the histograms to Gaussian functions. Apparent FRET efficiencies can then be related to the separation between the two fluorophores.

The unfolding of Fab A33 was first followed using the denaturant guanidinium chloride (GdmCl), as a control for smFRET experiments. The two dual-labelled Fab A33 constructs were diluted to 100 pM in 20 mM phosphate buffer pH 7.0 containing six different concentrations of GdmCl: 0, 1, 2, 4, 6 M and 6 M with 50 µM TCEP. Raw data with no corrections, is first shown in the form of distributions of inter-photon delays (time between two consecutive photons) (Figure 5.9). Two processes can be identified from these graphs, labelled Fab A33 traversing the detection volume generate the high count signal at short inter-photon delays, while lower count signal at long inter-photon delays originates from the background (detector dark counts, afterpulsing, out-of-focus molecules, sample scattering and impurities). The unfolding of Fab A33 with increasing concentration of GdmCl was observed through the decrease in the acceptor signal at low inter-photon delay times. Decrease in the signal detected from the acceptor fluorophore and an increase in the signal detected by the donor, related to an increase in the distance between fluorophores, result of the unfolding of Fab A33.

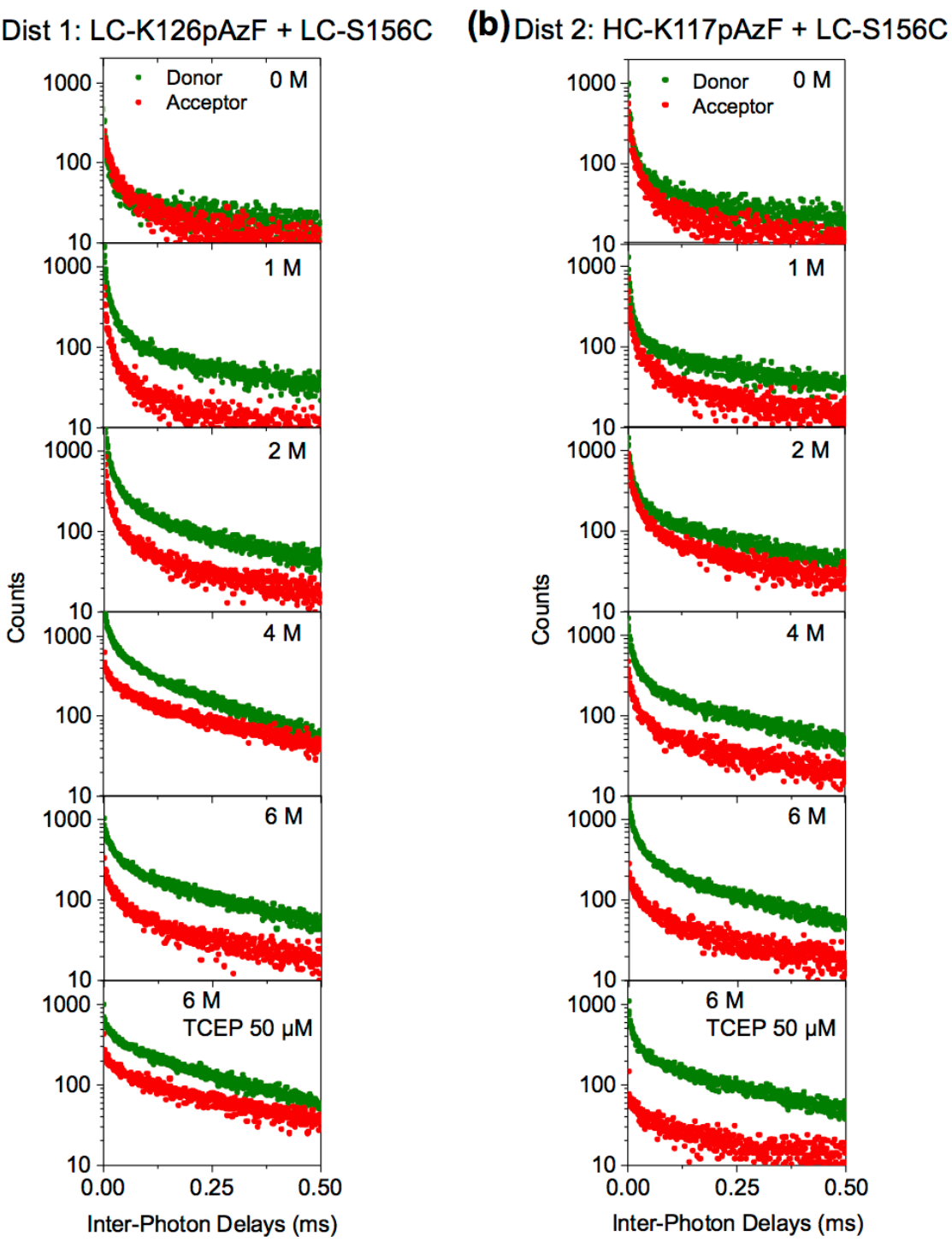

Figure 5.9. Inter-photon delays by smFRET to follow the unfolding of Fab A33 with increasing guanidium chloride concentration. The raw data with no corrections applied are shown in the form of inter-photon delay (time between two consecutive photons) distributions for (a) Dist 1 (Fab A33 labelled at LC-K126 + LC-S156); (b) Dist 2 (Fab A33 labelled at HC-S117 + LC-S156), for six concentrations of GdmCl, top to bottom: 0, 1, 2, 4, 6 M and 6 M with 50 µM TCEP. Photons from the donor are colored green and photons from the acceptor are in red.

Histograms of apparent FRET efficiencies of dual-labelled Fab A33 with increasing GdmCl concentration are shown in Figure 5.10. The peaks at a FRET efficiency of zero correspond to molecules with no active acceptor fluorophore, and these peak backgrounds are in gray. The peaks at high FRET efficiencies correspond to Fab A33 molecules with one donor and one acceptor. To determine their mean transfer efficiencies, these were fitted to Gaussian peak functions (black lines). In the graph with no added GdmCl, the peaks at high efficiency (E_{app} = 0.975 for construct Dist 1 and E_{app} = 0.87 for construct Dist 2) correspond to folded Fab molecules (Figure 5.10a,b top). For both constructs, as the denaturant concentration increased, new peaks with lower apparent transfer efficiencies appeared, corresponding to unfolded states. In Fab A33 labelled at Dist 1, folded and unfolded conformations were detected at GdmCl concentrations of 1 and 2 M. At GdmCl 4 M, only the unfolded state was present, and this peak shifted to lower transfer efficiencies at GdmCl 6 M, indicating an expansion of the unfolded state. When TCEP was added, no peaks were observed, suggesting complete unfolding. In Fab A33 labelled at Dist 2, a slow unfolding of the folded state up to GdmCl 4 M was observed, with a decrease in the transfer efficiency of unfolded state with increased GdmCl concentration. At GdmCl 6 M, complete unfolding of Fab A33 was observed.

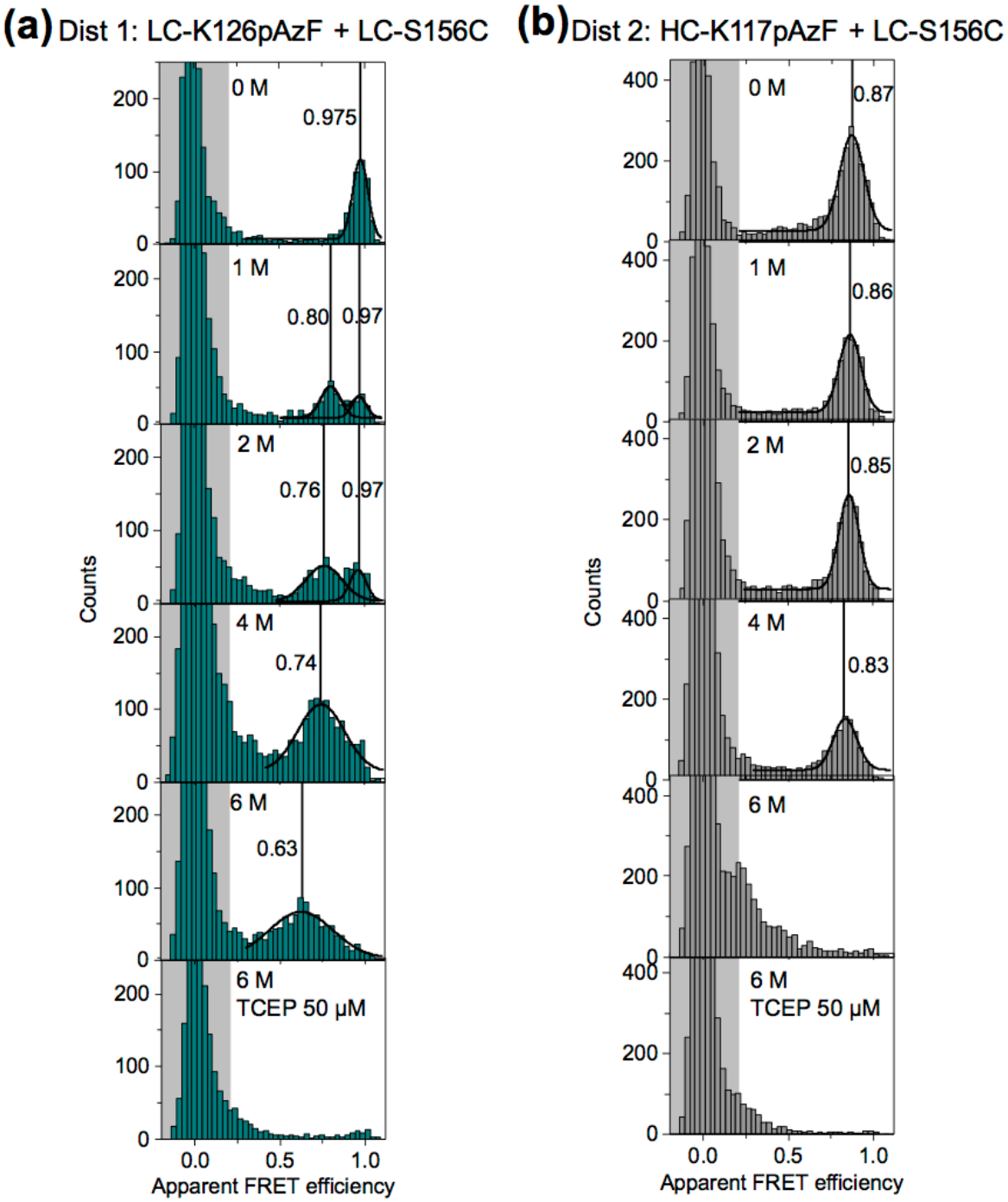

Figure 5.10. FRET efficiency histograms to follow the unfolding of Fab A33 by GdmCl. Apparent FRET efficiency (E_{app}) histograms of (a) Fab labelled at Dist 1 (LC-K126 + LC-S156) and (b) Fab labelled at Dist 2 (HC-S117 + LC-S156) for six concentrations of GdmCl, top to bottom: 0, 1, 2, 4, 6 M and 6 M with 50 µM TCEP. At a FRET efficiency of 0.0, a population of donor-only protein (no acceptor dye) is present, and has been shaded. At higher FRET efficiencies, a population is present that corresponds to Fab A33 with both fluorophores. This population was fitted with a Gaussian function, and the peak is shown by a vertical line.

5.4.3 smFRET to confirm the CL domain displacement at low pH

Confocal detection of freely diffusing molecules was also used to obtain the apparent transfer efficiency histograms of Fab A33 at pH 7.0 and 3.5 in an ionic strength of 50 mM each. As before, raw data are first shown in the form of distributions of inter-photon delays (Figure 5.11). The signature of two processes was observed. At low inter-photon delay times, the presence of labelled Fab A33 is detected, whereas the tail of the distribution corresponds to the background (Ingargiola, Lerner, et al. 2016).

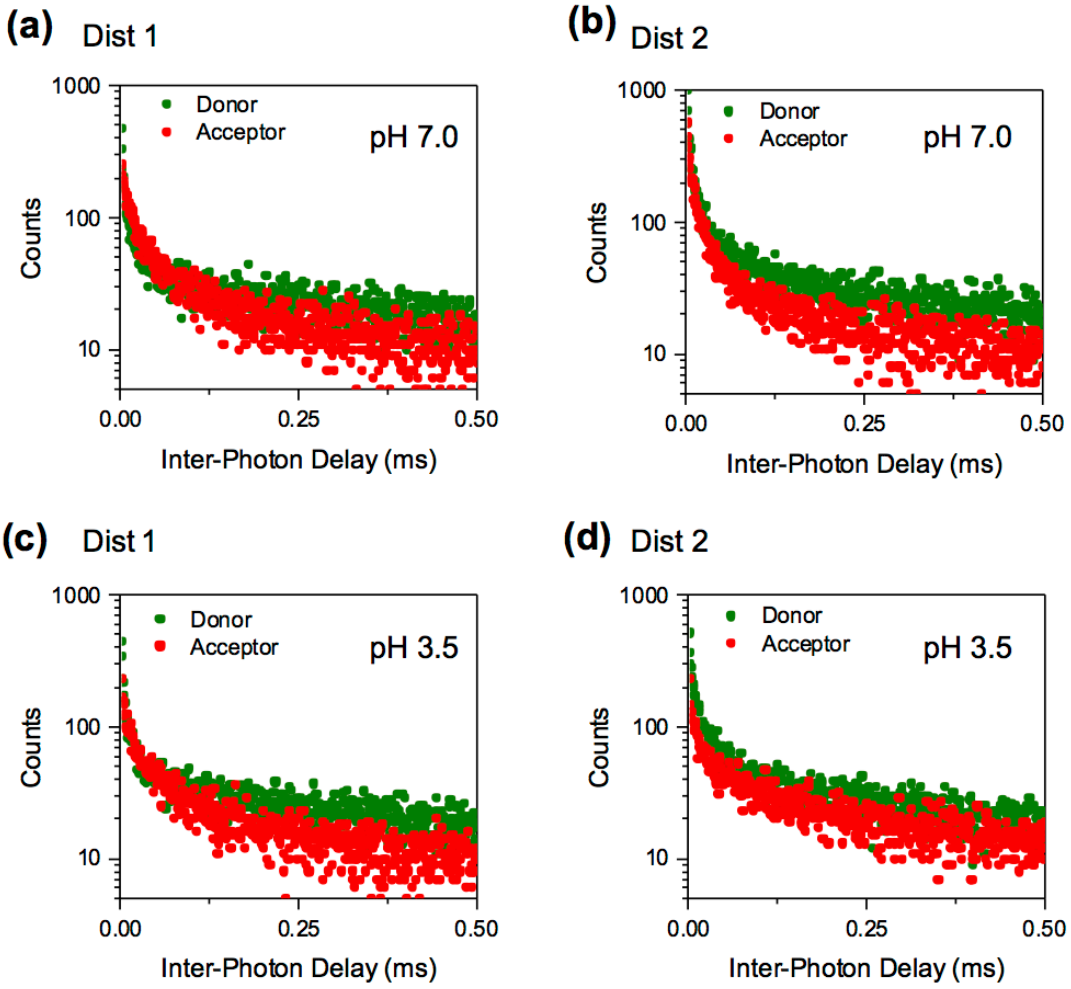

Figure 5.11 Inter-photon delay times by smFRET for double-labelled Fab A33. The raw data with no corrections applied are shown as inter-photon delay (time between two consecutive photons) distributions for (a, b) Dist 1 (Fab A33 labelled at LC-K126 + LC-S156) and (c, d) Dist 2 (Fab A33 labelled at HC-S117 + LC-S156), for each of (a, c) pH 7.0 and (b, d) 3.5 as labelled. Photons from the donor are colored green and photons from the acceptor are in red.

smFRET showed that the intra-C_L distance (Dist 1) did not change with pH, as the same FRET efficiency value (E_{app} = 0.97) was found at pH 7.0 and pH 3.5 (Figure 5.12). In contrast, the distance between C_L and the heavy chain linker (Dist 2) increased at pH 3.5, with a decrease in FRET efficiency from E_{app} = 0.87 at pH 7.0 to E_{app} = 0.78 at pH 3.5 (Figure 5.12). These results confirmed the displacement of the C_L domain and the partial unfolding of Fab A33 at low pH.

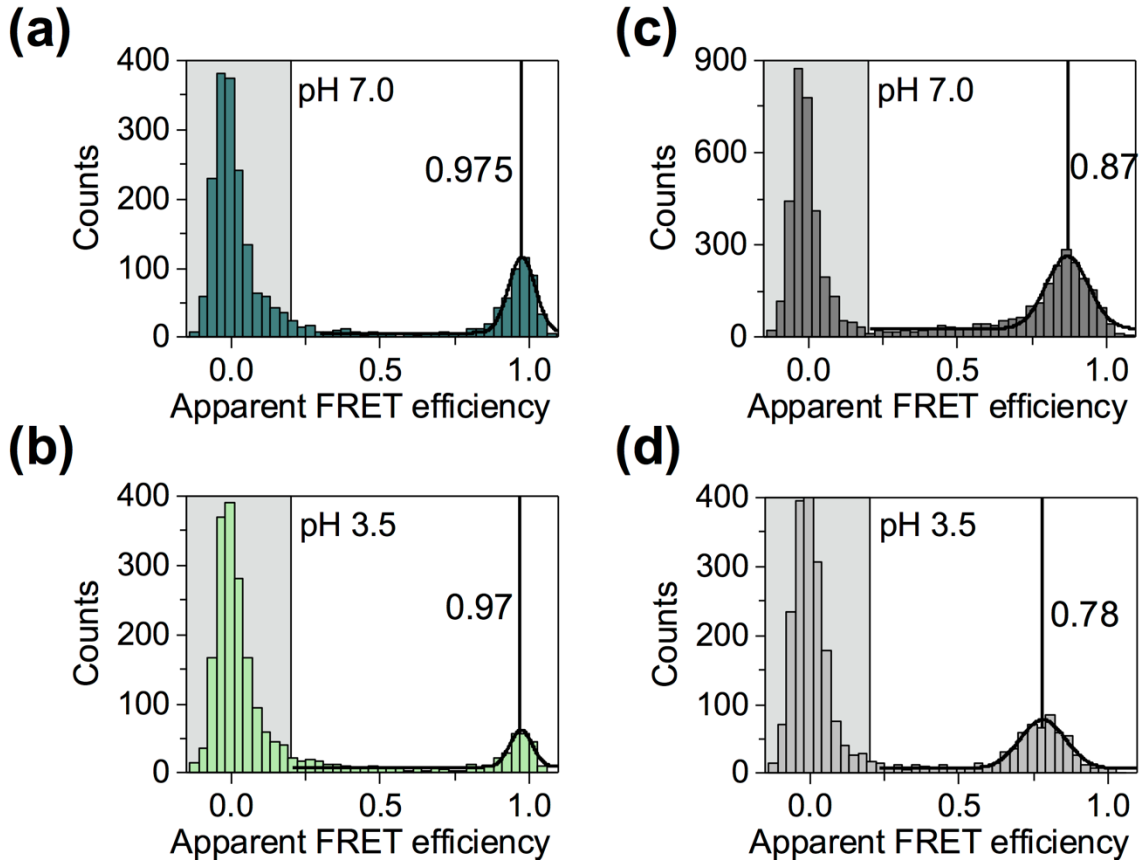

Figure 5.12. FRET efficiency histograms of the two dual-labelled Fab A33 at pH 7.0 and 3.5. Apparent FRET efficiency (E_{app}) histograms of (a, b) Fab labelled at Dist 1 (LC-K126 + LC-S156) (green), and (c, d) Dist 2 (residues HC-S117 and LC-S156) (gray), at (a, c) pH 7.0 (dark color) and (b, d) pH 3.5 (light color). At a FRET efficiency of 0.0, the population of molecules without an active acceptor fluorophore is shaded in gray. At higher FRET efficiencies, there is a population that corresponds to Fab A33 containing both fluorophores. This population was fitted with a Gaussian function and the peak is shown with a vertical line.

5.4.4 Compare FRET efficiencies to distances obtained using SAXS and MD simulations

FRET efficiencies were compared to the distances found using atomistic modelling of the SAXS data and MD simulations. For SAXS, the average and SD of Dist 1 and Dist 2 were measured for the ten best fit structures at pH 7.0 and the ten best fit structures at pH 3.5, using pymol (Figure 5.13a). Dist 1 was unchanged with pH, being 2.5 ± 0.1 nm at pH 7.0 and pH 3.5. However, Dist 2 increased from 2.8 ± 0.4 nm at pH 7.0 to 3.5 ± 0.1 nm at pH 3.5, corresponding to an increased separation between C_L and the heavy chain linker.

Dist 1 and Dist 2 were also monitored during the MD simulations at pH 7.0 and 3.5 (Figure 5.13b). Dist 1 was unchanged during the simulation, while Dist 2 increased from 2.9 ± 0.3 nm at pH 7.0 to 3.3 ± 0.2 nm at pH3.5. Both the atomistic SAXS modelling and the MD simulations confirmed the experimentally observed displacement at low pH of the C_L domain by smFRET.

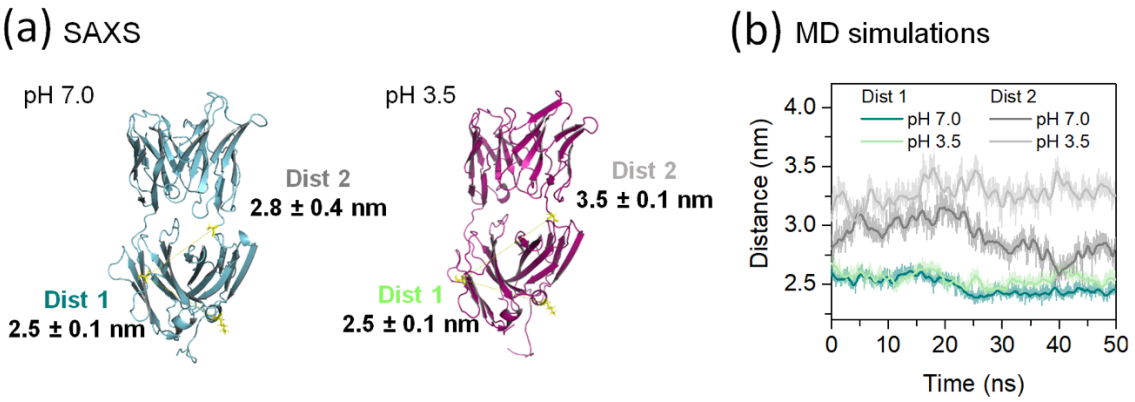

Figure 5.13. Measured distances for the two dual-labelled Fab at pH 7.0 and 3.5, using SAXS atomistic modeling and MD simulations. (a) The averaged Dist 1 and 2 separations and their SD were measured from the ten best-fit SAXS structures at each of pH 7.0 (cyan) and pH 3.5 (magenta). (b) The Dist 1 (green) and Dist 2 (gray) separations as a function of simulation time for pH 7.0 (dark color) and 3.5 (light color) are shown from the MD simulations. Three simulation repeats were averaged at every time frame, from which a window average is shown in a darker color.

5.5 Conclusions

A better mechanistic understanding, in particular the elucidation of pre-aggregational conformational states, is needed to improve protein engineering and formulation strategies for minimizing aggregation. However, very little is known about the structures of native-like states predicted to mediate the onset of aggregation. In chapter 4, we elucidated using SAXS that at ambient temperature Fab A33 becomes conformationally expanded in acidic solutions, and that an increase in the population of these species coincided with a rise in the systems aggregation kinetics suggesting that the expanded conformation is aggregation prone. To elucidate the structural origin of the conformational expansion, I combined SAXS data with atomistic structures of Fab A33 generated using MD simulations, to reveal that the conformational change at low pH takes place in the constant domain of the light chain (C_L). To validate this approach, in this chapter, I used confocal single-molecule FRET to study the protein conformational changes through the position-specific incorporation of fluorescent dyes in Fab A33. FRET is the radiationless transfer of energy from a donor fluorophore to an acceptor in a range of 2-10 nm distances. This transfer of energy is highly sensitive to changes in distance, allowing the study of the separation between the donor and acceptor. By looking at each protein individually, smFRET has the ability to provide insights into the early stages of protein aggregation, such as the nature of the aggregation-prone conformer.

Two dual-labelled constructs of Fab A33 were successfully generated, to monitor the separations termed Dist 1 (residues LC-K126 and LC-S156) and Dist 2 (residues HC-S117 and LC-S156). Dist 1 monitored an intra-C_L separation and Dist 2 a separation between the C_L domain and the heavy-chain linker. Dist 1 and Dist 2 constructs, each contained one nonstandard amino acid (p-azido-l-phenylalanine) and one solvent-exposed cysteine, to attach the donor (Alexa Fluor 488) and acceptor (Alexa Fluor 594) fluorophores. I reported the apparent FRET transfer efficiency (E_{app}), which measures the fraction of photons absorbed by the donor that have been transferred to the acceptor, and was used to report the separation between the fluorophores. smFRET results confirmed the unfolding of C_L at low pH. Dist 1 did not change with pH, with an $E_{app} = 0.97$ at both pH 7.0 and 3.5, however, Dist 2 decreased from $E_{app} = 0.87$ at pH 7.0 to $E_{app} = 0.78$ at pH 3.5, result of the partial unfolding of Fab A33 in this domain. Notably, the values obtained for apparent transfer efficiencies agreed with the with distances measured from the best

models derived from SAXS and MD simulations. SAXS best fits found a distance of 2.5 ± 0.1 nm for Dist 1 at both pH, while Dist 2 increased from 2.8 ± 0.4 nm at pH 7.0 to 3.5 ± 0.1 nm at pH 3.5. Similarly, MD simulations found Dist 1 to be unchanged during the simulation, while Dist 2 increased from 2.9 ± 0.3 nm at pH 7.0 to 3.3 ± 0.2 nm at pH 3.5.

These results highlight the power of single molecule measurements in elucidating the structural changes that take place in the native state that precede aggregation. At low pH, Fab A33 experiences a conformational change to form a partially unfolded intermediate, native-like in character, which is the first step to aggregation, from this aggregation-competent conformation. Thus, results of this work provide a better mechanistic understanding of how aggregation is initiated and propagated, and ultimately provide the tools to protein engineers to design proteins more robust to aggregation and allow the search for protein aggregation inhibitors to be done in a rational way.

Chapter Six

Summary and Future Work

6.1 Summary

In this thesis I have studied the stability and aggregation-prone conformations of a humanized antibody Fab fragment A33, under multiple stresses of pH, ionic strength and temperature, using a combination of computational tools (atomistic molecular dynamics simulations, in-silico mutational analysis by FoldX and Rosetta, predictors of aggregation-prone regions, etc.) and experimental methods (X-Ray Scattering and single-molecule FRET).

The stability and dynamics of Fab A33 was first studied using computational tools, with the aim to elucidate the early unfolding events and stability-limiting regions of this antibody fragment, under stresses it might encounter during its development, such as low pH and high temperature. Results found with MD simulations and stabilizing software predictors strongly agreed in the regions of Fab A33 that can potentially be stabilized further. MD simulations revealed that many contacts were lost in the interface between constant domains (C_L-C_H1) very early in the simulations, under both stresses of low pH and high temperature. Supporting this, calculations by FoldX and Rosetta also both agreed that mutations at this interface had the greatest potential for increasing the stability of Fab A33. Further validation was provided by packing density calculations, which revealed that the residues identified by the stability predictors, were under-packed relative to the other residues located in the interface between domains. At low pH, the C_L domain was found to partially unfold during the simulations, while at high temperature, C_L and V_H were found to unfold, revealing different unfolding pathways depending on the stress experienced. Salt bridge analysis identified the presence of two salt-bridges located in the C_L domain, which probably contribute to the unfolding observed of this domain at low pH, upon protonation. At high temperature, salt bridges broke and reformed very quickly and did not always reform with the same partner, indicative of a different mechanism for Fab A33 destabilization. Overall, my analysis revealed some regions that were common to both thermal and low-pH unfolding, and provided targets for mutation using FoldX and Rosetta that agreed with certain mutations found in existing Fabs.

To experimentally characterize the aggregation-prone conformations, solution structures of Fab A33 under different conditions of pH and salt concentration, were solved

using small angle X-ray scattering (SAXS). SAXS revealed a slight expansion of the native state upon acidification, with an R_g increase of 2.2% to 4.1%. Consistent with previous reports, I found pH to have a bigger effect on the conformation of Fab A33 than salt concentration, which instead, seemed to mainly contribute to charge shielding. Interestingly, the presence of the expanded conformation of Fab A33 coincided with accelerated aggregation, indicating that this conformation was more aggregation-prone. Scattering data were fitted using 45,000 structures obtained from the atomistic MD simulations under the same conditions, and located the conformational change at low pH to the C_L domain. The results were then verified using a complementary method, single-molecule FRET (smFRET) with two dual-labelled Fabs. smFRET confirmed the increase in distance between the C_L domain and the heavy chain linker at pH 3.5 respect to pH 7.0. Lastly, in order to gain insights into the mechanisms by which aggregation might occur, I used online tools to predict the aggregation-prone regions (APR) that are more likely to form the cross-β structures found in aggregates. All APRs in Fab A33 are located in the interior of the protein. However, the displacement of the C_L domain at low pH exposed a predicted APR, which forms a mechanistic basis for subsequent aggregation. Overall, these findings provide a means by which aggregation-prone conformers can be determined experimentally and add further evidence to the importance of partially unfolded states to the aggregation mechanisms of globular proteins.

6.2 Future Work

Based on the findings within this thesis, there are several areas worth exploring in future work; among others, the ideas highlighted below.

Rational mutagenesis of Fab A33 to improve its stability (protein engineering)

Based on the findings presented here, I speculate that stabilization of Fab A33 should start at the constant domain interface (C_L-C_H1). The most stabilizing mutations predicted by FoldX and Rosetta were located in this interface. Only one of the top suggested mutations, N137I, was found to be present in my analysis of natural variation within existing Fab sequences. However, there was significant scope for improvement through mutating the interfacial residues S176, N137, S397, T180, and S395, to the suggested hydrophobic residues (Table 3.5). Next, the C_L domain was found to unfold at both low pH and high temperature. Notably, the remaining top stabilizing mutations found by FoldX and Rosetta were located in this domain. Two mutations were suggested to improve the interaction between C_L and V_L domain, S12 and K103, with S12Y mutation found naturally. In the C_L domain, S159 was identified, which interacts with an outer β-strand, suggesting this interaction can also be improved (Table 3.5). Lastly, the C_H1 domain was found to unfold at high temperature. The only mutation identified in C_H1 domain was S267, identified by FoldX, to S267P, which notably is also found naturally. Interestingly, the mutations suggested here have the potential to stabilize Fab A33 to both, pH and thermal stresses. Not only the effect of single mutations, but the additive effect of combining double and triple mutations will also be interesting to explore.

Experimental proof of the proposed Fab A33 aggregation mechanism

To explain the increased aggregation propensity of the expanded conformations of Fab A33, I used online tools to predict its aggregation-prone regions (APR). APRs are hydrophobic sequences with low net charge and a strong β-sheet propensity, which have the potential to trigger aggregation. At low pH, the unfolding of the C_L domain was found to increase the solvent accessibility of a predicted APR in this region, likely triggering aggregation. To confirm this proposed aggregation mechanism, future work could include the mutation of residues involved in the exposed APR (residues 387-402), to reduce its aggregation propensity, and/or mutagenesis of the ionizable residues that drive the pH-induced change. Here, two salt bridges were identified to to be at the heart of this domain

unfolding at low pH, Glu165-Lys103 and Glu195-Lys149. Glu165-Lys103 bridges the C_L domain to the V_L domain, and Glu195-Lys149 is located in outer β-strands of the C_L domain, bridging β-strands C and F. Site directed mutagenesis of the mentioned residues would ultimately provide insights into their function and role in Fab A33 aggregation.

Characterization of aggregation-prone states of other Fabs and antibody structures

The generality of the results found in this thesis could be studied by elucidating the local changes in the native conformation of other Fabs, that promote protein aggregation. The results would reveal whether the findings found in this thesis are specific to Fab A33, or they have the potential to stabilize a wider range of therapeutic Fab fragments. Additionally, it would be interesting to investigate other antibody-based products, such as $F(ab')_2$ fragments, single-chain variable fragments (scFvs), single domain antibodies (sdAb), bi-specifics and full antibodies. The results would provide insights into the stabilities and aggregation mechanisms of antibody products due to different molecular weights.

Design of excipients and/or ligands that bind and stabilize Fab A33 against aggregation (formulation)

This work has provided the structures of aggregation-prone conformations of Fab A33, which potentially allows the design of ligands that will bind and stabilize Fab A33 against aggregation. Rational drug design is increasingly being done using computer-aided drug design, which requires the accurate structure of the target protein. From there, several computational approaches exist to discover ligand candidates, such as virtual screening (structure- or ligand-based design), de novo design, molecular docking, molecular dynamics simulations, etc. Large number of candidates are then tested in screening libraries for binding. Alternatively, the effect of molecular additives (e.g. formulation excipients) on Fab A33 structure under different solution conditions can be studied using computational tools, such as molecular dynamic simulations or docking software, and experimental techniques such as NMR and hydrogen-deuterium exchange mass spectrometry; as the molecular mechanisms through which excipients stabilize proteins against aggregation remain unclear.

smFRET to follow the aggregation process

smFRET can also be used to study the protein aggregation process, for instance by characterizing the species formed in the oligomerization process. It was seen in this

work that aggregation of Fab A33 at near native conditions proceeds through a partially unfolded expanded conformation that is native-like in structure. Thus, the initial oligomers formed probably retain high structure similarity to the native state. As found in previous studies, in later stages of the aggregation process, a structural re-arrangement might take place to form the typical cross-β structure of amyloids. Oligomers could be distinguished from single-molecules because they have higher fluorescence intensities. Based on the average intensity from a monomer, the approximate number of monomers per oligomer could be extracted. Kinetics of oligomer formation could then be followed, as well as the size of the oligomeric species. Additionally, the internal reorganization experimented by oligomers, from a more native conformation to forming the stable cross-β structure, could be studied in a change in its FRET signature.

Check that antigen-binding activity is retained while engineering Fab A33

The successful development of therapeutic proteins depends critically on achieving stability under a range of conditions, while retaining their specific mode of action. The information available about the antigen of Fab A33 is limited. Fab A33 recognizes a protein expressed on the surface of colon cancer cells. The antigen is expressed on a number of human tumor cell lines, including Colo205, ASPC-1 and SW1222 cell lines. Antigen binding assays could be performed using cells from these human colorectal tumor cell lines. Cells could be incubated in the presence of engineered Fab A33 fragments, and detected by further incubation with a FITC-conjugated antibody that recognizes Fab, and by detection in the FACScan analyser (Becton Dickinson). Alternatively, ELISAs could be developed to test that the antigen-binding activity of Fab A33 was retained.

References

Abraham, M.J. et al., 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2, pp.19–25.

Alzari, P.M., Lascombe, M.B. & Poljak, R.J., 1988. Three-dimensional structure of antibodies. *Annu Rev Immunol*, 6, pp.555–580.

Arakawa, T. & Timasheff, S.N., 1982. Preferential Interactions of Proteins with Salts in Concentrated Solutions. *Biochemistry*, 21, pp.6545–6552.

Arzenšek, D., Kuzman, D. & Podgornik, R., 2012. Colloidal interactions between monoclonal antibodies in aqueous solutions. *Journal of Colloid and Interface Science*, 384(1), pp.207–216.

De Baets, G., Schymkowitz, J. & Rousseau, F., 2014. Predicting aggregation-prone sequences in proteins. *Essays In Biochemistry*, 56, pp.41–52.

Baldwin, R.L., 1996. How Hofmeister ion interactions affect protein stability. *Biophysical Journal*, 71(4), pp.2056–2063.

Beck, A. et al., 2010. Strategies and challenges for the next generation of therapeutic antibodies. *Nature Reviews Immunology*, 10(5), pp.345–352.

Bemporad, F. et al., 2012. Characterizing intermolecular interactions that initiate native-like protein aggregation. *Biophysical Journal*, 102(11), pp.2595–2604.

Bemporad, F. & Chiti, F., 2009. "Native-like aggregation" of the acylphosphatase from Sulfolobus solfataricus and its biological implications. *FEBS Letters*, 583(16), pp.2630–2638.

Bemporad, F. & Chiti, F., 2012. Protein misfolded oligomers: experimental approaches, mechanism of formation, and structure-toxicity relationships. *Chemistry and Biology*, 19(3), pp.315–327.

Boesecke, P., 2007. Reduction of two-dimensional small- and wide-angle X-ray scattering data. *Journal of Applied Crystallography*, 40, pp.S423–S427.

Borgia, M.B. et al., 2011. Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature*, 474(7353), pp.662–665.

Boulianne, G.L., Hozumi, N. & Shulman, M.J., 1984. Production of functional chimaeric mouse/human antibody. *Nature*, 312, pp.643–646.

Brekke, O.H. & Løset, G.Å., 2003. New technologies in the therapeutic antibody development. *Current Opinion in Pharmacology*, 3(5), pp.544–550.

Brooks, B.R. et al., 2009. CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry*, 30(10), pp.1545–1614.

Buck, P.M., Kumar, S. & Singh, S.K., 2013. Insights into the potential aggregation liabilities of the b12 Fab fragment via elevated temperature molecular dynamics. *Protein Engineering, Design and Selection*, 26(3), pp.195–206.

Buß, O., Rudat, J. & Ochsenreither, K., 2018. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Computational and Structural Biotechnology Journal*, 16, pp.25–33.

Calamai, M., Chiti, F. & Dobson, C.M., 2005. Amyloid fibril formation can proceed from different conformations of a partially unfolded protein. *Biophysical Journal*, 89(6), pp.4201–4210.

Canet, D. et al., 2002. Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nature Structural Biology*, 9, pp.308–315.

Carter, P.J. & Lazar, G.A., 2018. Next generation antibody drugs: Pursuit of the "high-hanging fruit." *Nature Reviews Drug Discovery*, 17, pp.197–223.

Case, D.A. et al., 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16), pp.1668–1688.

Chakroun, N. et al., 2016. Mapping the Aggregation Kinetics of a Therapeutic Antibody Fragment. *Molecular Pharmaceutics*, 13(2), pp.307–319.

Chames, P. et al., 2009. Therapeutic antibodies: Successes, limitations and hopes for the future. *British Journal of Pharmacology*, 157(2), pp.220–233.

Chan, A.C. & Carter, P.J., 2010. Therapeutic antibodies for autoimmunity and inflammation. *Nature Reviews Immunology*, 10(5), pp.301–316.

Chapman, A.P. et al., 1999. Therapeutic antibody fragments with prolonged in vivo half-lives. *Nature Biotechnology*, 17(8), pp.783–3.

Cheetham, G.M.. et al., 1998. Crystal Structures of a Rat Anti-CD52 (CAMPATH-1) Therapeutic Antibody Fab Fragment and its Humanized Counterpart. *Journal of Molecular Biology*, 284(1), pp.85–99.

Chennamsetty, N. et al., 2009. Design of therapeutic proteins with enhanced stability. *Proceedings of the National Academy of Sciences*, 106(29), pp.11937–11942.

Chi, E.Y. et al., 2003. Physical stability of proteins in aqueous solution: Mechanism and driving forces in nonnative protein aggregation. *Pharmaceutical Research*, 20(9), pp.1325–1336.

Chiti, F. et al., 2003. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950), pp.805–808.

Chiti, F. & Dobson, C.M., 2009. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology*, 5, pp.15–22.

Chiti, F. & Dobson, C.M., 2017. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annual Review of Biochemistry*, 86(1), pp.27–68.

Chiti, F. & Dobson, C.M., 2006. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75(1), pp.333–366.

Chivian, D. & Baker, D., 2006. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Research*, 34(17), pp.1–18.

Christiansen, J. & Rajasekaran, A.K., 2004. Biological impediments to monoclonal antibody-based cancer immunotherapy. *Molecular cancer therapeutics*, 3(11), pp.1493–1501.

Codina, N. et al., 2019. An Expanded Conformation of an Antibody Fab Region by X-Ray Scattering, Molecular Dynamics and smFRET Identifies an Aggregation Mechanism. *Journal of Molecular Biology*, 431(7), pp.1409–1425.

Collins, C., Tsui, F.W.L. & Shulman, M.J., 2002. Differential activation of human and guinea pig complement by pentameric and hexameric IgM. *European Journal of Immunology*, 32(6), pp.1802–1810.

Collins, K.D., 2004. Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process. *Methods*, 34(3), pp.300–311.

Collu, F. et al., 2018. Probing the early stages of prion protein (PrP) aggregation with atomistic molecular dynamics simulations. *Chemical Communications*, 54(57), pp.8007–8010.

Conchillo-Solé, O. et al., 2007. AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics*, 8(1), p.65.

Cuesta, Á.M. et al., 2010. Multivalent antibodies: when design surpasses evolution. *Trends in Biotechnology*, 28(7), pp.355–362.

Curtis, R.A., Prausnitz, J.M. & Blanch, H.W., 1998. Protein-protein and protein-salt interactions in aqueous protein solutions containing concentrated electrolytes. *Biotechnology and Bioengineering*, 58(4), pp.451–451.

Dan, N. et al., 2018. Antibody-drug conjugates for cancer therapy: Chemistry to clinical implications. *Pharmaceuticals*, 11(2).

Daugherty, A.L. & Mrsny, R.J., 2006. Formulation and delivery issues for monoclonal

antibody therapeutics. *Advanced Drug Delivery Reviews*, 58(5–6), pp.686–706.

Davies, D.R. & Chacko, S., 1993. Antibody structure. *Accounts of Chemical Research*, 26, pp.421–427.

Davies, D.R. & Metzger, H., 1983. Structural Basis of Antibody Function. *Ann. Rev. ImmunoL*, 1, pp.87–117.

Deniz, A.A. et al., 1999. Single-pair fluorescence resonance energy transfer on freely diffusing molecules: observation of Forster distance dependence and subpopulations. *Proceedings of the National Academy of Sciences*, 96(7), pp.3670–3675.

Dobson, M.C., 2003. Protein folding and misfolding. *Nature*, 426(6968), pp.884–890.

van Durme, J. et al., 2011. A graphical interface for the FoldX forcefield. *Bioinformatics*, 27(12), pp.1711–1712.

Van Durme, J. et al., 2016. Solubis: A webserver to reduce protein aggregation through mutation. *Protein Engineering, Design and Selection*, 29(8), pp.285–289.

Ecker, D.M., Jones, S.D. & Levine, H.L., 2015. The therapeutic monoclonal antibody market. *mAbs*, 7(1), pp.9–14.

Elvin, J.G., Couston, R.G. & Van Der Walle, C.F., 2013. Therapeutic antibodies: market considerations, disease targets and bioprocessing. *International Journal of Pharmaceutics*, 440(1), pp.83–98.

Emily, M., Talvas, A. & Delamarche, C., 2013. MetAmyl: A METa-predictor for AMYLoid proteins. *PLoS ONE*, 8(11).

Enever, C. et al., 2009. Next generation immunotherapeutics-honing the magic bullet. *Current Opinion in Biotechnology*, 20(4), pp.405–411.

Eswar, N. et al., 2006. Comparative protein structure modeling using Modeller. In *Curr Protoc Bioinformatics*. pp. 1–47.

Fernandez-Escamilla, A.-M. et al., 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10), pp.1302–1306.

Ferreon, A.C.M. et al., 2010. Alteration of the α-synuclein folding landscape by a mutation related to Parkinson's disease. *Angewandte Chemie - International Edition*, 49(20), pp.3469–3472.

Ferreon, A.C.M. et al., 2009. Interplay of α-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proceedings of the National Academy of Sciences*, 106(14), pp.5645–5650.

Fleming, P.J. & Richards, F.M., 2000. Protein packing: dependence on protein size,

secondary structure and amino acid composition. *Journal of Molecular Biology*, 299(2), pp.487–498.

Franke, D. et al., 2012. New developments in the ATSAS program package for small-angle scattering data analysis. *Journal of Applied Crystallography*, 45(2), pp.342–350.

Frieden, C., 2007. Protein aggregation processes: In search of the mechanism. *Protein Science*, 16(11), pp.2334–2344.

Frokjaer, S. & Otzen, D.E., 2005. Protein drug stability: a formulation challenge. *Nature Reviews Drug Discovery*, 4(4), pp.298–306.

Frueh, D.P. et al., 2013. NMR methods for structural studies of large monomeric and multimeric proteins. *Current Opinion in Structural Biology*, 23(5), pp.734–739.

Garbuzynskiy, S.O., Lobanov, M.Y. & Galzitskaya, O. V., 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3), pp.326–332.

Gasior, P. & Kotulska, M., 2014. FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC Bioinformatics*, 15(1), pp.1–8.

Ginalski, K., 2006. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16(2), pp.172–177.

Gore, S. et al., 2017. Validation of Structures in the Protein Data Bank. *Structure*, 25(12), pp.1916–1927.

Green, L.L. et al., 1994. Antigen–specific human monoclonal antibodies from mice engineered with human Ig heavy and light chain YACs. *Nature Genetics*, 7(1), pp.13–21.

Hagler, A.T., Huler, E. & Lifson, S., 1974. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society*, 96(17), pp.5319–5327.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, pp.95–98.

Harris, L.J. et al., 1992. The three-dimensional structure of an intact monoclonal antibody for canine lymphoma. *Nature*, 360, pp.369–372.

Hartl, F.U. & Hayer-Hartl, M., 2009. Converging concepts of protein folding in vitro and in vivo. *Nature Structural and Molecular Biology*, 16(6), pp.574–581.

Hermeling, S. et al., 2004. Structure-immunogenicity relationships of therapeutic

proteins. *Pharmaceutical Research*, 21(6), pp.897–903.

Hillger, F. et al., 2007. Detection and analysis of protein aggregation with confocal single molecule fluorescence spectroscopy. *Journal of Fluorescence*, 17(6), pp.759–765.

Hilton, D., 2015. *Elucidating the aggregation mechanisms of antibody fragments through biophysical analysis*. University College London.

Hofmann, H. et al., 2010. Single-molecule spectroscopy of protein folding in a chaperonin cage. *Proceedings of the National Academy of Sciences*, 107(26), pp.11793–11798.

Hofmann, H., Nettels, D. & Schuler, B., 2013. Single-molecule spectroscopy of the unexpected collapse of an unfolded protein at low pH. *Journal of Chemical Physics*, 139(12), p.121930.

Holliger, P. & Hudson, P.J., 2005. Engineered antibody fragments and the rise of single domains. *Nature Biotechnology*, 23(9), pp.1126–1136.

Holliger, P., Prospero, T. & Winter, G., 1993. "Diabodies": small bivalent and bispecific antibody fragments. *Proceedings of the National Academy of Sciences*, 90(14), pp.6444–6448.

Hospital, A. et al., 2015. Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem*, 8, pp.37–47.

Hu, Z. & Jiang, J., 2010. Assessment of Biomolecular Force Fields for Molecular Dynamics Simulations in a Protein Crystal. *Journal of computational chemistry*, 31(2), pp.371–380.

Humphrey, W., Dalke, A. & Klaus, S., 1996. VMD: Visual Molecular Dynamics. *Journal of molecular graphics*, 14(1), pp.33–38.

Huston, J.S. et al., 1988. Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in Escherichia coli. *Proceedings of the National Academy of Sciences*, 85(16), pp.5879–5883.

Hwang, W.Y.K. & Foote, J., 2005. Immunogenicity of engineered antibodies. *Methods*, 36(1), pp.3–10.

Iljina, M. et al., 2016. Kinetic model of the aggregation of alpha-synuclein provides insights into prion-like spreading. *Proceedings of the National Academy of Sciences*, 113(9), pp.E1206–E1215.

Ingargiola, A., Lerner, E., et al., 2016. FRETBursts: An open source toolkit for analysis of freely-diffusing single-molecule FRET. *PLoS ONE*, 11(8), pp.1–27.

Ingargiola, A., Laurence, T., et al., 2016. Photon-HDF5: An Open File Format for Timestamp-Based Single-Molecule Fluorescence Experiments. *Biophysical Journal*, 110(1), pp.26–33.

Invernizzi, G. et al., 2012. Protein aggregation: mechanisms and functional consequences. *International Journal of Biochemistry and Cell Biology*, 44(9), pp.1541–1554.

Jahn, T.R. & Radford, S.E., 2008. Folding versus aggregation: polypeptide conformations on competing pathways. *Archives of Biochemistry and Biophysics*, 469(1), pp.100–117.

Jevševar, S., Kusterle, M. & Kenig, M., 2012. *PEGylation of Antibody Fragments for Half-Life Extension* Antibody M. G. Proetzel & H. Ebersbach, eds., Totowa, NJ: Humana Press.

Jones, P.T. et al., 1986. Replacing the complementarity- determining regions in a human antibody with those from a mouse. *Nature*, 321(6069), pp.522–525.

Joo, C. et al., 2008. Advances in Single-Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry*, 77, pp.51–76.

Kabsch, W. & Sander, C., 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), pp.2577–2637.

Kaminski, G.A. et al., 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B*, 105(28), pp.6474–6487.

Kantha, S.S., 1991. A Centennial Review; the 1890 Tetanus Antitoxin Paper of von Behring and Kitasato and the Related Developments. *Keio journal of medicine*, 1, pp.35–39.

Kaufmann, S.H.E., 2017. Remembering Emil von Behring: from Tetanus Treatment to Antibody Cooperation with Phagocytes. *mBio*, 8(1), pp.1–6.

Kelley, L.A. et al., 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10, pp.845–858.

Kellogg, E.H., Leaver-Fay, A. & Baker, D., 2011. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, 79(3), pp.830–838.

Kendrick, B.S. et al., 1998. A transient expansion of the native state precedes aggregation of recombinant human interferon-γ. *Proceedings of the National Academy of Sciences*, 95(24), pp.14142–14146.

Khurana, R. et al., 2001. Partially folded intermediates as critical precursors of light chain amyloid fibrils and amorphous aggregates. *Biochemistry*, 40(12), pp.3525–3535.

King, D.J., Adair, J.R. & Owens, R.J., 2001. Humanized antibodies directed against A33 antigen - Patent No.: US 6,307,026. , p.Celltech R.

Kleinjung, J. & Fraternali, F., 2014. Design and application of implicit solvent models in biomolecular simulations. *Current Opinion in Structural Biology*, 25, pp.126–134.

Knowles, T.P.J., Vendruscolo, M. & Dobson, C.M., 2014. The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell Biology*, 15(6), pp.384–396.

Köhler, G. & Milstein, C., 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature*, 256(5517), pp.495–497.

Kortkhonjia, E. et al., 2013. Probing antibody internal dynamics with fluorescence anisotropy and molecular dynamics simulations. *mAbs*, 5(2), pp.306–322.

Krishnan, S. et al., 2002. Aggregation of Granulocyte Colony Stimulating Factor under Physiological Conditions: Characterization and Thermodynamic Inhibition. *Biochemistry*, 41(20), pp.6422–6431.

Kumar, V. et al., 2011. Impact of short range hydrophobic interactions and long range electrostatic forces on the aggregation kinetics of a monoclonal antibody and a dual-variable domain immunoglobulin at low and high concentrations. *International Journal of Pharmaceutics*, 421(1), pp.82–93.

Lajoie, M.J. et al., 2013. Genomically Recoded Organisms Expand Biological Functions. *Science*, 342(6156), pp.357–360.

Leader, B., Baca, Q.J. & Golan, D.E., 2008. Protein therapeutics: a summary and pharmacological classification. *Nature Reviews Drug Discovery*, 7(1), pp.21–39.

Leaver-fay, A. et al., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymology*, 487, pp.545–574.

Lemieux, R.U. & Spohr, U., 1994. How Emil Fischer Was Led To The Lock and Key Concept for Enzyme Specificity. *Advances in Carbohydrate Chemistry and Biochemistry*, 50, pp.1–20.

Lerner, E. et al., 2018. Toward dynamic structural biology: Two decades of single-molecule förster resonance energy transfer. *Science*, 359(6373), p.eaan1133.

Lesser, G.J. & Rose, G.D., 1990. Hydrophobicity of amino acid subgroups in proteins. *Proteins*, 8(1), pp.6–13.

Li, H., Robertson, A.D. & Jensen, J.H., 2005. Very fast empirical prediction and rationalization of protein pK a values. *Proteins: Structure, Function and Genetics*, 61(4), pp.704–721.

Lim, S.I., Hahn, Y.S. & Kwon, I., 2015. Site-specific albumination of a therapeutic protein with multi-subunit to prolong activity in vivo. *Journal of Controlled Release*, 207, pp.93–100.

Lindorff-Larsen, K. et al., 2012. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *Journal of the American Chemical Society*, 134(8), pp.3787–3791.

Liu, C.C. & Schultz, P.G., 2010. Adding New Chemistries to the Genetic Code. *Annual Review of Biochemistry*, 79(1), pp.413–444.

Lu, L.L. et al., 2018. Beyond binding: Antibody effector functions in infectious diseases. *Nature Reviews Immunology*, 18(1), pp.46–61.

Magliery, T.J., 2015. Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol.*, 33, pp.161–168.

Mahler, H.C. et al., 2009. Protein Aggregation: Pathways, Induction Factors and Analysis. *Journal of Pharmaceutical Sciences*, 98(9), pp.2145–2157.

Majumdar, D.S. et al., 2007. Single-molecule FRET reveals sugar-induced conformational dynamics in LacY. *Proceedings of the National Academy of Sciences*, 104(31), pp.12640–12645.

Manning, M.C. et al., 2010. Stability of protein pharmaceuticals: An update. *Pharmaceutical Research*, 27(4), pp.544–575.

Markwick, P.R.L., Malliavin, T. & Nilges, M., 2008. Structural biology by NMR: structure, dynamics, and interactions. *PLoS Computational Biology*, 4(9), p.e1000168.

Maurer-Stroh, S. et al., 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3), pp.237–242.

Maverakis, E. et al., 2015. Glycans In The Immune System and The Altered Glycan Theory of Autoimmunity: A Critical Review. *Journal of Autoimmunity*, 57(6), pp.1–13.

McCafferty, J. et al., 1990. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, 348(6301), pp.552–554.

Merchant, K.A. et al., 2007. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proceedings of the National Academy of Sciences*, 104(5), pp.1528–1533.

Mertens, H.D.T. & Svergun, D.I., 2017. Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Archives of Biochemistry and Biophysics*, 628, pp.33–41.

Mertens, H.D.T. & Svergun, D.I., 2010. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *Journal of Structural Biology*, 172(1), pp.128–141.

Milardi, D., La Rosa, C. & Grasso, D., 1994. Extended theoretical analysis of irreversible protein thermal unfolding. *Biophysical Chemistry*, 52(3), pp.183–189.

Monnier, P., Vigouroux, R. & Tassew, N., 2013. In Vivo Applications of Single Chain Fv (Variable Domain) (scFv) Fragments. *Antibodies*, 2(2), pp.193–208.

Morea, V., Lesk, A.M. & Tramontano, A., 2000. Antibody modeling: Implications for engineering and design. *Methods*, 20, pp.267–279.

Morrison, S.L. et al., 1984. Chimeric human antibody molecules: Mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, 81(21), pp.6851–6855.

Muhammed, M.T. & Aki-Yalcin, E., 2019. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology and Drug Design*, 93(1), pp.12–20.

Muller-Spath, S. et al., 2010. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 107(33), pp.14609–14614.

Murata, K. & Wolf, M., 2018. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta - General Subjects*, 1862(2), pp.324–334.

Nan, R. et al., 2013. Zinc-induced self-association of complement C3b and Factor H: implications for inflammation and age-related macular degeneration. *Journal of Biological Chemistry*, 288(26), pp.19197–19210.

Neal, B.L. et al., 1999. Why is the osmotic second virial coefficient related to protein crystallization? *Journal of Crystal Growth*, 196(2–4), pp.377–387.

Nelson, A.L., 2010. Antibody fragments: hope and hype. *mAbs*, 2(1), pp.77–83.

Nelson, A.L., Dhimolea, E. & Reichert, J.M., 2010. Development trends for human monoclonal antibody therapeutics. *Nature Reviews Drug Discovery*, 9(10), pp.767–774.

Nelson, R. et al., 2005. Structure of the cross-β spine of amyloid-like fibrils. *Nature*, 435(7043), pp.773–778.

Neudecker, P. et al., 2012. Structure of an intermediate state in protein folding and aggregation. *Science*, 336(6079), pp.362–6.

Nogales, E. & Scheres, S.H., 2015. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol Cell.*, 58(4), pp.677–689.

Olsen, S.N. et al., 2009. Role of electrostatic repulsion on colloidal stability of Bacillus halmapalus alpha-amylase. *Biochimica et Biophysica Acta*, 1794(7), pp.1058–1065.

Orte, A. et al., 2008. Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proceedings of the National Academy of Sciences*, 105(38), pp.14424–14429.

Parkins, D.A. & Lashmar, U.T., 2000. The formulation of biopharmaceutical products. *Pharmaceutical Science and Technology Today*, 3(4), pp.129–137.

Patel, D. & Kuyucak, S., 2017. Computational study of aggregation mechanism in human lysozyme[D67H]. *PLoS ONE*, 12(5), pp.1–17.

Pattabiraman, N., Ward, K.B. & Fleming, P.J., 1995. Occluded molecular surface: analysis of protein packing. *Journal of Molecular Recognition*, 8(6), pp.334–344.

Pauling, L., 1940. A Theory of the Structure and Process of Formation of Antibodies. *Journal of the American Chemical Society*, 62, pp.2643–2657.

Pawar, A.P. et al., 2005. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *Journal of Molecular Biology*, 350(2), pp.379–392.

Pepinsky, R.B. et al., 2011. Production of a PEGylated Fab′ of the anti-LINGO-1 Li33 antibody and assessment of its biochemical and functional properties in vitro and in a rat model of remyelination. *Bioconjugate Chemistry*, 22(2), pp.200–210.

Perkins, S., 1986. Protein volumes and hydration effects. *European Journal of Biochemistry*, 157, pp.169–180.

Perkins, S.J. et al., 2016. Atomistic modelling of scattering data in the Collaborative Computational Project for Small Angle Scattering (CCP-SAS). *Journal of Applied Crystallography*, 49(6), pp.1861–1875.

Perkins, S.J. et al., 2009. Constrained solution scattering modelling of human antibodies and complement proteins reveals novel biological insights. *Journal of The Royal Society Interface*, 6(Suppl 5), pp.S679–S696.

Perkins, S.J. et al., 2008. X-ray and neutron scattering data and their constrained molecular modeling. *Methods in Cell Biology*, 84, pp.375–423.

Perkins, S.J. & Bonner, A., 2008. Structure determinations of human and chimaeric

antibodies by solution scattering and constrained molecular modelling. *Biochemical Society Transactions*, 36(1), pp.37–42.

Phillips, J.C. et al., 2005. Scalable Molecular Dynamics with NAMD. *J Comput Chem*, 26(16), pp.1781–1802.

Poljak, R.J. et al., 1973. Three-Dimensional Structure of the Fab' Fragment of a Human Immunoglobulin at 2,8-A resolution. *Proceedings of the National Academy of Sciences*, 70(12), pp.3305–3310.

Poppe, L. et al., 2013. Profiling formulated monoclonal antibodies by 1H NMR spectroscopy. *Analytical Chemistry*, 85(20), pp.9623–9629.

Porter, R.R., 1959. The hydrolysis of rabbit y-globulin and antibodies with crystalline papain. *Biochem J*, 73, pp.119–126.

Przepiorka, D. et al., 2000. Daclizumab, a humanized anti-interleukin-2 receptor alpha chain antibody, for treatment of acute graft-versus-host disease. *Blood*, 95(1), pp.83–89.

Quan, J. & Tian, J., 2011. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nature Protocols*, 6(2), pp.242–251.

Raman, S. et al., 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, 77(Suppl 9), pp.89–99.

Rayner, L.E. et al., 2015. The solution structures of two human IgG1 antibodies show conformational stability and accommodate their C1q and FcγR ligands. *Journal of Biological Chemistry*, 290(13), pp.8420–8438.

Riedel, S., 2015. Edward Jenner and the History of Smallpox and Vaccination. *Baylor University Medical Center Proceedings*, 18(1), pp.21–25.

Roberts, C.J., 2014. Therapeutic protein aggregation: mechanisms, design, and control. *Trends in Biotechnology*, 32(7), pp.372–380.

Roberts, D. et al., 2014. The Role of Electrostatics in Protein–Protein Interactions of a Monoclonal Antibody. *Molecular Pharmaceutics*, 11(7), pp.2475–2489.

Robinson, M.J. et al., 2018. Tm-Values and Unfolded Fraction Can Predict Aggregation Rates for Granulocyte Colony Stimulating Factor Variant Formulations but Not under Predominantly Native Conditions. *Molecular Pharmaceutics*, 15(1), pp.256–267.

Rocco, A.G. et al., 2008. Characterization of the protein unfolding processes induced by urea and temperature. *Biophysical Journal*, 94(6), pp.2241–2251.

Rose, P.W. et al., 2013. The RCSB Protein Data Bank: new resources for research and

education. *Nucleic Acids Research*, 41(D1), pp.D475–D482.

Ross, C.A. & Poirier, M.A., 2004. Protein aggregation and neurodegenerative disease. *Nature Medicine*, 10, pp.S10-7.

Roy, R. et al., 2009. SSB protein diffusion on single-stranded DNA stimulates RecA filament formation. *Nature*, 461(7267), pp.1092–1097.

Roy, R., Hohng, S. & Ha, T., 2008. A Practical Guide To single-molecule FRET. *Nature Methods*, 5(6), pp.507–516.

Sahin, E. et al., 2010. Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. *Journal of Pharmaceutical Sciences*, 99(12), pp.4830–4848.

Salimi, N.L., Ho, B. & Agard, D.A., 2010. Unfolding simulations reveal the mechanism of extreme unfolding cooperativity in the kinetically stable α-lytic protease. *PLoS Computational Biology*, 6(2), p.e1000689.

Sankar, K. et al., 2018. AggScore: Prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins*, 86(11), pp.1147–1156.

Schroeder, H.W. & Cavacini, L., 2010. Structure and Functions of Immunoglobulins. *J Allergy Clin Immunol.*, 125, p.(2 Suppl 2):S41-S52.

Schuler, B., 2013. Single-molecule FRET of protein structure and dynamics - a primer. *Journal of nanobiotechnology*, 11(Suppl 1), p.S2.

Schuler, B. & Eaton, W.A., 2008. Protein folding studied by single-molecule FRET. *Current Opinion in Structural Biology*, 18(1), pp.16–26.

Schuler, B. & Hofmann, H., 2013. Single-molecule spectroscopy of protein folding dynamics-expanding scope and timescales. *Current Opinion in Structural Biology*, 23(1), pp.36–47.

Schwede, T. et al., 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, 31(13), pp.3381–3385.

Settanni, G. & Fersht, A.R., 2008. High temperature unfolding simulations of the TRPZ1 peptide. *Biophysical Journal*, 94(11), pp.4444–4453.

Silverstein, A.M., 2004. Labeled antigens and antibodies: The evolution of magic markers and magic bullets. *Nature Immunology*, 5(12), pp.1211–1217.

Smith, A.J., 2015. New Horizons in Therapeutic Antibody Discovery: opportunities and challenges versus small-molecule therapeutics. *Journal of Biomolecular Screening*, 20(4), pp.437–453.

Smith, M.D. et al., 2015. Force-Field Induced Bias in the Structure of Aβ21-30: A Comparison of OPLS, AMBER, CHARMM, and GROMOS Force Fields. *Journal*

of Chemical Information and Modeling, 55, pp.2587–2595.

Speltz, E.B. & Regan, L., 2013. White and green screening with circular polymerase extension cloning for easy and reliable cloning. *Protein Science*, 22(6), pp.859–864.

Stefani, M. & Dobson, C.M., 2003. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *Journal of Molecular Medicine*, 81(11), pp.678–699.

Strebhardt, K. & Ullrich, A., 2008. Paul Ehrlich's magic bullet concept: 100 years of progress. *Nature Reviews Cancer*, 8(6), pp.473–480.

Su, J.G. et al., 2015. The intrinsic dynamics and unfolding process of an antibody fab fragment revealed by elastic network model. *International Journal of Molecular Sciences*, 16(12), pp.29720–29731.

Svergun, D.I. & Koch, M.H.J., 2002. Advances in structure analysis using small-angle scattering in solution. *Current Opinion in Structural Biology*, 12(5), pp.654–660.

Tanaka, M. & Komi, Y., 2015. Layers of structure and function in protein aggregation. *Nature Chemical Biology*, 11(6), pp.373–377.

Toofanny, R.D. & Daggett, V., 2012. Understanding protein unfolding from molecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(3), pp.405–423.

Touw, W.G. et al., 2015. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1), pp.D364–D368.

Trexler, A.J. & Rhoades, E., 2010. Single molecule characterization of α-synuclein in aggregation-prone states. *Biophysical Journal*, 99(9), pp.3048–3055.

Tsolis, A.C. et al., 2013. A Consensus Method for the Prediction of "Aggregation-Prone" Peptides in Globular Proteins. *PLoS ONE*, 8(1), pp.1–6.

Uversky, V.N., Li, J. & Fink, A.L., 2001. Evidence for a partially folded intermediate in alpha-synuclein fibril formation. *Journal of Biological Chemistry*, 276(14), pp.10737–10744.

Venselaar, H. et al., 2010. Homology modelling and spectroscopy, a never-ending love story. *European Biophysics Journal*, 39(4), pp.551–563.

Ventura, S. et al., 2004. Short amino acid stretches can mediate amyloid formation in globular proteins: The Src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences*, 101(19), pp.7258–7263.

Vermeer, A.W.P. & Norde, W., 2000. The thermal stability of immunoglobulin: Unfolding and aggregation of a multi-domain protein. *Biophysical Journal*, 78(1),

pp.394–404.

Vidarsson, G., Dekkers, G. & Rispens, T., 2014. IgG subclasses and allotypes: from structure to effector functions. *Frontiers in Immunology*, 5(520), pp.1–17.

Walker, K.T. et al., 2017. Non-linearity of the collagen triple helix in solution and implications for collagen function. *Biochemical Journal*, 474(13), pp.2203–2217.

Walsh, I. et al., 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Research*, 42(W1), pp.301–307.

Wang, H.W. & Wang, J.W., 2017. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science*, 26(1), pp.32–39.

Wang, T. & Duan, Y., 2011. Probing the stability-limiting regions of an antibody single-chain variable fragment: a molecular dynamics simulation study. *Protein Engineering, Design and Selection*, 24(9), pp.649–657.

Wang, W. et al., 2007. Antibody Structure, Instability, and Formulation. *Journal of Pharmaceutical Sciences*, 96(1), pp.1–26.

Wang, W. et al., 2012. Immunogenicity of protein aggregates - concerns and realities. *International Journal of Pharmaceutics*, 431(1–2), pp.1–11.

Wang, W., 1999. Instability, stabilization, and formulation of liquid protein pharmaceuticals. *International Journal of Pharmaceutics*, 185(2), pp.129–88.

Wang, W., 2005. Protein aggregation and its inhibition in biopharmaceutics. *International Journal of Pharmaceutics*, 289(1–2), pp.1–30.

Wang, W., Nema, S. & Teagarden, D., 2010. Protein aggregation-pathways and influencing factors. *International Journal of Pharmaceutics*, 390(2), pp.89–99.

Wang, W., Wang, E.Q. & Balthasar, J.P., 2008. Monoclonal Antibody Pharmacokinetics and Pharmacodynamics. *Clin Pharmacol Ther*, 84(5), pp.548–558.

Wang, X., Mathieu, M. & Brezski, R.J., 2018. IgG Fc engineering to modulate antibody effector functions. *Protein and Cell*, 9(1), pp.63–73.

Wang, Y. et al., 2005. Accurate FRET Measurements within Single Diffusing Biomolecules Using Alternating-Laser Excitation. *Biophysical Journal*, 88(4), pp.2939–2953.

Webb, J.N. et al., 2001. Partial molar volume, surface area, and hydration changes for equilibrium unfolding and formation of aggregation transition state: high-pressure and cosolute studies on recombinant human IFN-gamma. *Proceedings of the National Academy of Sciences*, 98(13), pp.7259–7264.

Weiner, G.J., 2015. Building better monoclonal antibody-based therapeutics. *Nature*

Reviews Cancer, 15(6), pp.361–370.

Weiner, L.M., Surana, R. & Wang, S., 2010. Monoclonal antibodies: Versatile platforms for cancer immunotherapy. *Nature Reviews Immunology*, 10(5), pp.317–327.

Weiner, S.J. et al., 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3), pp.765–784.

Westbrook, J. et al., 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Research*, 31(1), pp.489–491.

Wijma, H.J. et al., 2014. Computationally designed libraries for rapid enzyme stabilization. *Protein Engineering, Design and Selection*, 27(2), pp.49–58.

Wilkinson, I.C. et al., 2009. High Resolution NMR-based Model for the Structure of a scFv-IL-1β Complex: potential for NMR as a key tool in therapeutic antibody design and development. *Journal of Biological Chemistry*, 284(46), pp.31928–31935.

Winau, F., Westphal, O. & Winau, R., 2004. Paul Ehrlich - In search of the magic bullet. *Microbes and Infection*, 6(8), pp.786–789.

Woof, J.M. & Burton, D.R., 2004. Human antibody-Fc receptor interactions illuminated by crystal structures. *Nature Reviews Immunology*, 4(2), pp.89–99.

Wright, A. & Morrison, S.L., 1998. Effect of C2-associated carbohydrate structure on Ig effector function: studies with chimeric mouse-human IgG1 antibodies in glycosylation mutants of Chinese hamster ovary cells. *Journal of immunology*, 160, pp.3393–3402.

Wright, D.W. & Perkins, S.J., 2015. SCT: a suite of programs for comparing atomistic models with small-angle scattering data. *Journal of Applied Crystallography*, 48(3), pp.953–961.

Yang, J. et al., 2018. Direct Observation of Oligomerization by Single Molecule Fluorescence Reveals a Multistep Aggregation Mechanism for the Yeast Prion Protein Ure2. *Journal of the American Chemical Society*, 140(7), pp.2493–2503.

Yang, S., 2014. Methods for SAXS-based Topological Structure Determination of Biomolecular Complexes. *Advanced Matter*, 26(46), pp.7002–7910.

Young, T.S. et al., 2010. An enhanced system for unnatural amino acid mutagenesis in E. coli. *Journal of Molecular Biology*, 395(2), pp.361–374.

Yu, H. et al., 2017. Two strategies to engineer flexible loops for improved enzyme thermostability. *Scientific Reports*, 7(41212), pp.1–15.

Yu, H. & Dalby, P.A., 2018. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proceedings of the National Academy of Sciences*, 115(47), pp.E11043–E11052.

Zambrano, R. et al., 2015. AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*, 43(W1), pp.W306–W313.

Zhang, C. et al., 2018. Computational Design to Reduce Conformational Flexibility and Aggregation Rates of an Antibody Fab Fragment. *Molecular Pharmaceutics*, 15(8), pp.3079–3092.

Zhang, Z. et al., 2012. Predicting folding free energy changes upon single point mutations. *Bioinformatics*, 28(5), pp.664–671.

Zhuravlev, P.I. et al., 2014. Propensity to form amyloid fibrils is encoded as excitations in the free energy landscape of monomeric proteins. *Journal of Molecular Biology*, 426(14), pp.2653–2666.