Structural and Conformational Analysis of B-cell Epitopes – component to guide peptide vaccine design

Saba Ferdous

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Structural and Molecular Biology

December, 2017

Declaration of Authorship

I, Saba Ferdous, declare that this thesis titled, 'Structural and Conformational Analysis of B-cell Epitopes − component to guide peptide vaccine design' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself

Signed: _____

Date: _____

*Dedicated to my beautiful **Family***

Abstract

Peptide vaccines have many potential advantages including low cost, lack of need for cold-chain storage and safety. However, it is well known that approximately 90% of B-cell Epitopes (BCEs) are discontinuous in nature making it difficult to mimic them for creating vaccines. To perform a detailed structural analysis of these epitopes, they needs to be mapped onto antigen structures that are complexed with antibody. In order to obtain a clean dataset of antibody-antigen complex crystal structures, a pipeline was designed to process automatically and clean the antibody related structures from the PDB. To store this processed antibody structural data, a database (AbDb) was built and made available online.

The degree of discontinuity in B-cell epitopes and their conformational nature was studied by mapping epitopes in the antibody-antigen dataset. The discontinuity of B-cell epitopes was analysed by defining extended 'regions' (R, consisting of at least 3 antibody-contacting residues each separated by $\leq$ 3 residues) and small fragments (F, antibody-contacting residues that do not satisfy the requirements for a region). Secondly, an algorithm was developed to classify region shape as linear, curved or folded.

Molecular dynamics simulations were carried out on isolated epitope regions (wild type and mutant peptides). The mutant peptides have been designed by mutating non-contacting and hydrophobic residues in epitopes. Two types of mutations (hydrophobic to alanine and hydrophobic to glutamine) have been studied using molecular dynamics simulations. Furthermore, the effect of end-capping on wild type and mutant epitope regions has been studied. Simulation studies were carried out on 5 linear and 5 folded shape regions. Out of these, 2 epitopes (one linear and one folded), along with their mutants and derivatives, were tested experimentally for conformational stability by CD spectroscopy and NMR. The binding of isolated epitopes with antibody was also validated by ELISA and SPR.

Acknowledgements

Pursuing this PhD has been a wonderful and truly life-changing experience for me and it would not have been possible without the support and guidance that I received from many people.

Firstly, I offer my sincere gratitude to my supervisor, Dr. Andrew Martin for his invaluable guidance, support, patience and encouragement throughout my research. I am truly indebted for his massive help in securing PhD funding and giving me the opportunity to work with him and learn how to do science. His significant patience and help during this thesis writing is worth appreciation. The impact of his hard work on me will last forever and will be helpful throught my life and career.

This thesis was co-funded by UCL, UCB and the Schlumberger Foundation. I would like to thank all of these organisations for their generous support and in particular UCB for providing me with the opportunity of experiencing the industrial environment, therefore enabling me to carry out laboratory experiments and discussing my work with wonderful scientists. I would like to thank Andy Popplewell, Jiye Shi, Terry Baker and Seb Kelm for valuable advice and support on experimental work. During my time at UCB, I received great help from Leo Bowsher, Kieran Dawkins, Oliver Durrant, Christine Prosser, Geofrey Odede and Kerry Tyson, and I am truly grateful for that.

I would like to express my gratitude to members of the Martin group. A special thanks goes to Tom Northey for providing me the grounds to get started with programing and Francesco for help in MD simulation experiments. I would also like to thank everyone up in Room 636, in particular Sayoni, Milli and Su Dat for making me smile most of the time.

A special thanks to Mrs. Handan Taibi for looking after me for three years in London. I will never forget the time I spent with Zeynab and Hajer. A great thanks to Mrs. Rila Choudary for taking care of me in the most depressing phase of my PhD. A bunch of friends who regularly checked on me and made me feel better during the write-up phase deserve my deepest gratitude. A great thanks to my best friends Zainab and Usman for making me feel better in the most stressful times.

Last but not the least, a heartfelt thanks to my wonderful parents for their love, support and trust of my abilities. An unexpressable special thanks to Mr. Anwar Zahid for his continuous support since the very start of my PhD journey. I would have never been here without his favour of convincing everyone to achieve the biggest dream of my life. A great thanks to Mr. Mohammad Imran for his appealing encouragment. A massive thanks to my childhood friend and a very special nephew, Haseeb Arsalan for his frequent visits to see me in London. I will never forget the care I got from my sisters (Tahira, Zohra, Rubina and Khadija) in England. There was no single day when I would miss an evening call from my sister, Tahira, (an acting Mummy) which was such an addiction and driving force for me. I am very much grateful for her support in each decision of my life. The role of my sister, Khadija to start my university education has significant impact on defining who I am today. A great thanks goes to sisters Zohra and Rubina for their kindness and love. I would have never been able to overcome my loneliness without everyone's kind support. At times, cute words from my niece, Rida and nephew, Abeer helped me to feel great.

Contents

List of Figures

List of Tables

Abbreviations

AbDb	Antibody Structure Database
APC	Antigen Presenting Cells
AMBER	Assisted Model Building with Energy Refinement
BSA	Bovine Serum Albumin
PB	Phosphate Buffered
CD	Circular Dichroism
CDR	Complementarity Determining Regions
CSI	Chemical Shift Index
CTL	Cytotoxic T Lymphocytes
CV	Column Volume
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic Acid
DSS	Dimethyl Silapentane Sulfonic acid
EDTA	Ethylenediaminetetraacetic Acid
EDC	1-Ethyl-3-(3-dimethylaminopropyl)-carbodiimide
ELISA	Enzyme Linked Immunosorbent Assay
EM	Energy Minimization
ESI	Electro Spray Ionization
Fab	antigen-binding Fragment
FC	Flow Cell
FID	Free Induction Decay
FW	Frame Work

GRAVY Grand Average of hydropathicity

GuHCl Guanidine Hydrochloride

HCV Hepatitis C Virus

HEPES Hydroxy Ethyl Piperazine Ethane Sulfonic acid

HIV Human Immunodeficiency Virus

hLFA-1 human Leukocyte Function associated Antigen-1

HPV Human Papilloma Virus

HPLC High Performance Liquid Chromatography

HSQC Heteronuclear Single Quantum Coherence

IMGT international ImMunoGeneTics information system

K_D Dissociation Constant

LDS Lithium Dodecyl Sulfate

LINCS Linear Constraint Solver

MD Molecular dynamics

MHC Major Histocompatibility Complex

MPP Minimum Perturbation Protocol

MRE Mean Residue molar Ellipticity

NIH National Institute of Health

NHS N-hydroxysuccinimide

NMR Nuclear Magnetic Resonance

NOESY Nuclear Overhauser Effect Spectroscopy

NPT Number of atoms, Pressure, Temperature

NVT Number of atoms, Volume, Temperature

PBS Phosphate Buffered Saline

PBS/T Phosphate Buffered Saline Tween

PDB Protein Data Bank

PEG4 Poly Ethylene Glycol

PEM Photo Elastic Modulator

PES Potential Energy Surface

PME Particle Mesh Ewald

PMT Photo Multiplier Tube

QTOF Quadrupole Time of Flight

RF Radio Frequency

RI Refractive Index

RU Response Unit

SASA Solvent Accessible Surface Area

SACS Summary of Antibody Crystal Structures

SAbDab Structural Antibody Database

SDS Sodium Dodecyl Sulfate

SPR Surface Plasmon Resonance

TCR T-cell Receptor

TFA Trifluoroacetic acid

TFE Trifluoroethanol

TIR Total Internal Reflection

TOCSY Total Correlation Spectroscopy

TMB Tetra Methyl Benzidine

TMS Tetra Methyl Silane

TSP Trimethyl Silyl Propionate

TY Tryptone Yeast

UV Ultra Violet

VHH Variable domain of Heavy chain antibody

WT Wild Type

WTG Wild Type with Glycine linker

WTX Wild Type with Extended ends

WTSS Wild Type with diSulphide bond

Chapter 1

Introduction

Vaccination has provided highly successful protection against major infectious diseases, resulting in the global eradication of diseases such as small pox and poliomyelitis [1–4]. Despite this significant success, vaccines for many infectious diseases are still elusive and several methods are being considered to design novel immunogenic vaccines in the hope of achieving an efficacious response. Traditionally, vaccine development has involved the delivery of live attenuated or inactivated viruses or bacteria by injection. However, such vaccines (that include the whole organism) may cause a detrimental immune response owing to unnecessary proteins present in the vaccine formulation [5]. In principle, the humoral and cell-mediated immune response is dependent on only a few specific proteins. Therefore, the additional proteins present in a vaccine formulation may result in other unwanted host responses such as allergenic and/or reactogenic immune responses [6]. This has led to a focus on using a single protein (or a few proteins) from a micro-organism to induce the protective immune response [5, 6]. However, even a single protein contains many epitopes (regions of a molecule to which an antibody binds), some of which may be responsible for protective immunity while others may lead to an undesired immune response. This has led to an attractive alternative strategy of de-

veloping 'peptide vaccines' containing short peptide fragments (epitopes), that are capable of inducing specific immune responses, consequently avoiding allergenic and/or reactogenic responses [7]. This reductionist approach guides the precise vaccine formulation by using epitope-based peptides [8,9] which are representative of the minimal immunogenic region of an antigen and have the potential to elicit a precise immune response [10]. For development of such vaccines, understanding antibody/epitope interactions is required to understand the ability of epitopes to be recognised by a specific antibody [11]. Moreover, epitope characterisation and mapping has wide application in the design of immuno-therapeutics and diagnostic kits [12, 13].

So far, there are no human peptide/epitope based vaccines on the market owing to challenges associated with peptide stability, delivery and diversity of human immunogenetics. However, their emergence in the pharmaceutical arena and human therapeutics marketplace is expected in the near future [10]. This is because peptide-based vaccines have moved forward from preclinical to human studies over the past decade and this clinical progression has allowed scientists to develop robust paradigms for the use of peptides as vaccines. However, it is fundamentally important to understand the molcular basis of such vaccines.

1.1 Molecular Basis for Peptide Vaccine Development

In order to understand the molecular basis of peptide based vaccines, it is vital to comprehend the immune system which is a very complex topic. This section explains the basics of the immune system and its components.

1.1.1 Overview of the Immune System

The immune system offers two types of immunity: 1) Innate immunity — a non-specific type of protection where a first line of defence is provided by skin, mucous membranes and stomach acid while a second line of defence is provided by the inflammatory response and phagocytes; 2) Adaptive immunity — a specific response against a pathogen provided by leukocytes (T lymphocytes and B lymphocytes). The adaptive immune system constitutes two parts, one is responsible for the cell-mediated (cytotoxic) immune response and the other for the humoral immune response. In the cell-mediated immune response, cytotoxic T lymphocytes (CTLs; also known as CD8+ T cells/killer cells or T_k cells) remove infected cells through direct cytotoxic action on the targets. In the humoral immune response, B lymphocytes (also know as B-cells) are the main players and generate antibodies in response to pathogen-derived peptides. T-helper cells (T_h cells; also known as CD4+ T cells) play their role in both of these processes.

1.1.2 Major Histocompatibility Complex (MHC) Class I and II

Cytotoxic T leukocytes (T_k cells) recognise infected cells through the identification of major histocompatibility complex class I (MHC class I) molecules, bound to peptides derived from pathogenic proteins expressed within the cell. MHC-I is present on the surface of every nucleated cell whereas MHC-II is present only on the surface of antigen-presenting cells (APCs; such as B-cells, dendritic cells and macrophages) and the peptides bound to it are derived from extracellular proteins [10].

T-helper (T_h) cells identify MHC class II molecules bound to peptides derived from extracellular proteins, however, their activation requires a signal that is provided by the interaction of CD28 (present on the surface of T_h-cells) with CD80

and CD86 (present on the surface of APCs) which are co-stimulatory proteins. The T cell receptor (TCR) on the surface of cytotoxic T cells forms an MHC-I/peptide-epitope complex while T helper cells form the MHC-II/peptide-epitope complex to activate B lymphocytes. These interactions are assisted by CD8 and CD4 co-receptors respectively. In summary, T cells not only assist in the activation of B-cells to secrete antibodies and macrophages to ingest the infected cells, but also help in the activation of cytotoxic T cells to kill the infected cells. This complicated interaction of these epitope (peptide) dependent recognition processes results in the activation of immune responses that control infections and tumorigenesis (Figure 1.1).

1.1.3 Lymphocytes and Epitopes

T-cells and B-cells respond to distinct and different epitopes derived from the same organism. In general, T cells identify much shorter linear epitopes whereas B-cells and antibodies recognize longer conformational epitopes. Almost 90% of B-cell epitopes are conformational in nature with the assembly of amino acid residues brought together by protein folding [14]. In principle, this conformational nature of epitopes requires the full folded polypeptide for their action.

The ultimate goal of vaccination is to induce an immune response by selectively activating antigen specific B-cell and T-cells. A vaccine should have two antigenic epitopes: 1) a T_h-epitope 2) an epitope capable of inducing specific B-cell or CTL responses [10]. In some cases, these two epitopes can overlap in an antigen sequence, while in some instances they might be present in distinct regions of the antigen or present in different antigens from the same pathogen [10].

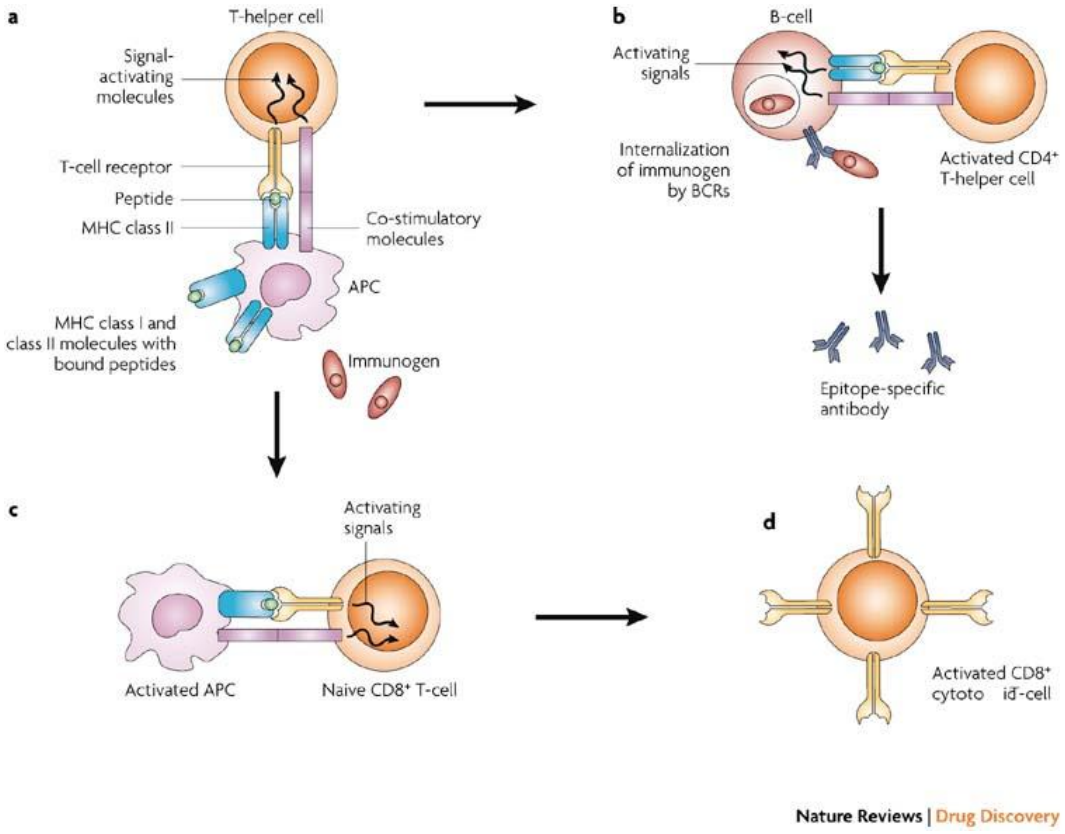


Figure 1.1: The interplay of B-cells, T cells and Antigen-presenting cell in the immune response. The first step in generating an antibody response is uptake of an antigen by the antigen-presenting cell (APC). The antigens then undergo proteolysis to break them into small peptides, some of these peptides then bind to MHC class II molecules followed by their transportation to the surface of the APC. T helper cells have receptors capable of identifying and interacting with the peptide/MHC-II complexes. On this interaction, an APC becomes activated. a) Activation of T helper cells. b) Activation of B-cells by the interaction of T helper cells with peptide/MHC-II complex. The peptide is derived from immunogen by its internalization. This results in the differentiation of B-cells into plasma cells that are capable of secreting antibodies of the same specificity as that of the immunogen's receptor. c) An activated T helper cell can stimulate some APCs to stimulate naive CD8+ T cells. d) This results in activation of CD8+ (cytotoxic killer cells) that are able to identify and kill the infected cells displaying peptide/MHC-I complex (Figure is reproduced from Purcell et al. [10]).

1.1.4 Antibodies

Antibodies (also known as Immunoglobulins) are Y-shaped proteins secreted by B-cells, used by the immune system to bind foreign antigens. Antibodies specifically identify antigens on the pathogens and neutralize them by binding a unique part (epitope) on their surface. At least 10^9 different antibodies can be produced by B-cells to identify a large number of different antigens. This enormous diversity of antibodies is due to the remarkable reordering of immunoglobulin gene segments in B cells that are responsible for the production of antibodies. These antibodies can neutralize the biological activity of their target antigens on binding, or elicit a down stream immune response. The exceptional specificity and affinity of antibodies has provided the basis for a wide range of therapeutic applications, as well research and medical technologies.

1.1.4.1 Structure

Among several distinct classes of Immunoglobins, IgG is the most frequent class in higher mammals. IgG consists of four polypeptide chains; two identical heavy chains (about 440 amino acids in length) and light chains (about 220 amino acids in length) which are held together with disulphide bonds and non-covalent interactions. Each of the heavy chains is comprised of one variable (V_H) and three constant (C_H1, C_H2 and C_H3) domains whereas each light chain is comprised of one variable (V_L) and one constant (C_L) domain (Figure 1.2). These variable and constant domains share sequence and structural similarity where both are comprised of two anti-parallel β-sheets forming a β-sandwich. Each antibody molecule possesses two antigen binding sites which lie in the variable domains of the antibody. They are comprised of complementarity determining regions (CDRs), allowing each immunoglobulin molecule to bind specifically to diverse antigens. The CDRs con-

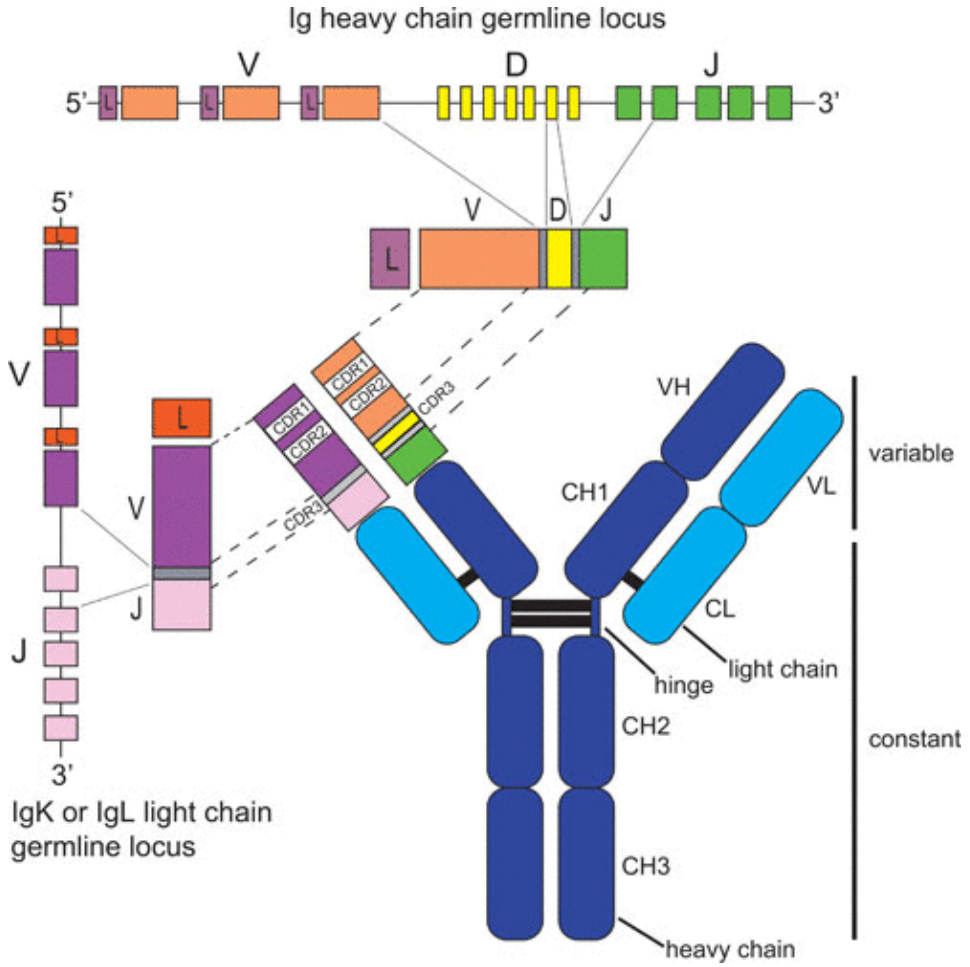


Figure 1.2: Antibody structure and genetic coding. Heavy chain domains are shown in dark blue and light chain domains are shown in light blue. The labelled variable region corresponds to the *Fv* region of the antibody. The CDRs are shown on the light and heavy chains' variable domain. The V(D)J segments of DNA encode the variable domain of the light or heavy chain. The Figure reproduced from Boyd and Joshi [16].

stitute approximately 50-70 amino acid residues from six loops (three in each of the heavy and light chains), also known as hypervariable loops. The CDRs vary greatly in sequence and length among antibodies enabling them to bind specific antigens [15]. The CDRs in the heavy chains are designated as CDR-H1, CDR-H2 and CDR-H3 whilst those in the light chains are named CDR-L1, CDR-L2 and CDR-L3.

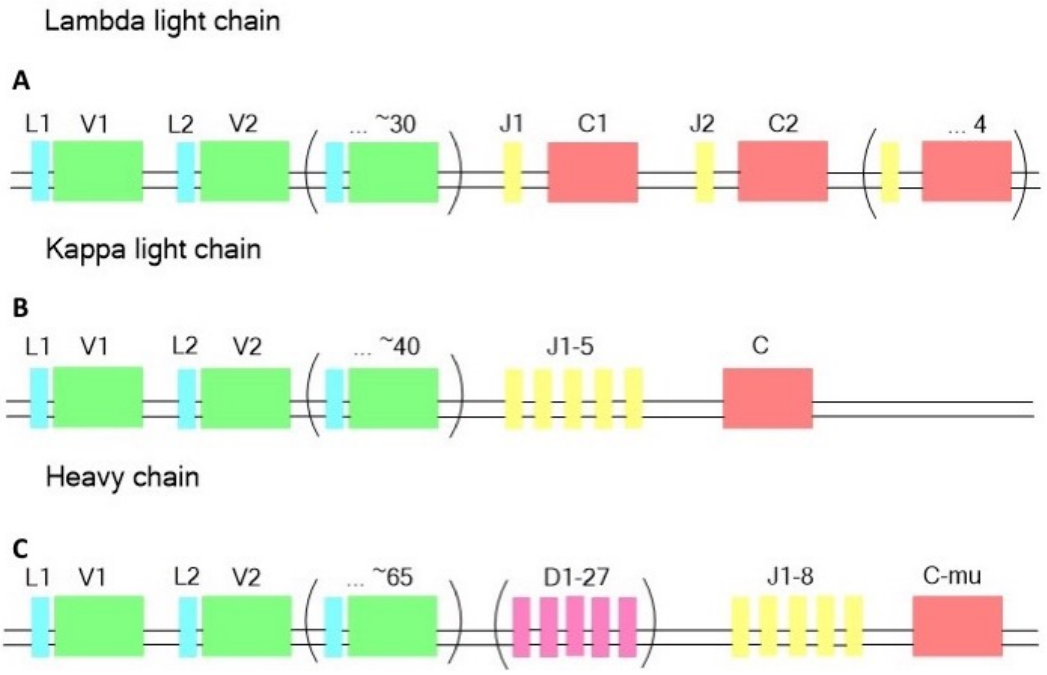


Figure 1.3: Antibody Genetics of the light and heavy chain variable domain. A-B) The two loci for the light chain; Lambda and Kappa, showing the V and J recombinations linked to the C segment. C) The VDJ recombination in the heavy chain.

1.1.4.2 Genetics

The immune system has a robust means of generating diverse and distinctive variable domains, formed from a collection of gene segments (V and J in light chain; V, D and J in heavy chain) which are spliced randomly to produce diversity. In the light chain, there are two loci in the genome, namely kappa and lambda. In both cases, a V (variable) and a J (joining) segment are chosen at random followed by the linking to the constant domain. A similar random selection of V and J segments occurs in the heavy chain with an additional D (diversity) segment, introduced between V and J (Figure 1.2). The addition of a D segment in the heavy chain means that heavy chain sequences are more diverse than their light chain counterparts and the peptide encoded by the D segment falls completely within CDR-H3 making this the most variable of the CDRs. This splicing is often referred as 'V(D)J recombination' and occurs at the DNA level.

Figure 1.3 shows the number and arrangement of V(D)J segments in light and heavy chains. These segments of DNA encode the variable domain of the light or heavy chain protein sequence. The L region adjacent to the V region encodes the leader sequence that is chopped off the final protein. The C segment of the DNA encodes the constant domains (1 in the light chain, 3 or 4 in the heavy chain depending on the antibody class). Based on these DNA segments, the diversity of antibodies can be estimated by calculating the possible combinations of these segments as:

Lambda light chain 30V x 4J = 120

Kappa light chain 40V x 5J = 200

Heavy chain 65V x 27D x 8J = 14,040

The heavy chain shows an additional variabilty because of the imprecise splicing of the D segment which leads to frameshifts within the D segment and the insertion of the additional bases at the junction. This introduces as much as 300x variability and results in 4,212,000 possible combinations in heavy chain. Thus, this provides an estimation of 1,347,840,000 different antibodies formed by the combination of any light and heavy chain pair.

Additional variation in heavy chain 14,040 x 300 = 4,212,000

Antibody diversity 4,212,000 x 120 + 4,212,000 x 200= 1,347,840,000

B-cells produce 5 different types of antibodies, depending on the differences in the constant region of the heavy chain, to perform different roles in the immune system. These antibody classes include IgM (μ), IgD (δ), IgG (γ), IgA (α) and IgE (ε) and contain different gene segment that encode that antibody class. In humans, nine such loci exist which determine the class of antibody. Initially, B-cells transcibe an mRNA having VDJ, μ and δ sequences that can be alternatively spliced to

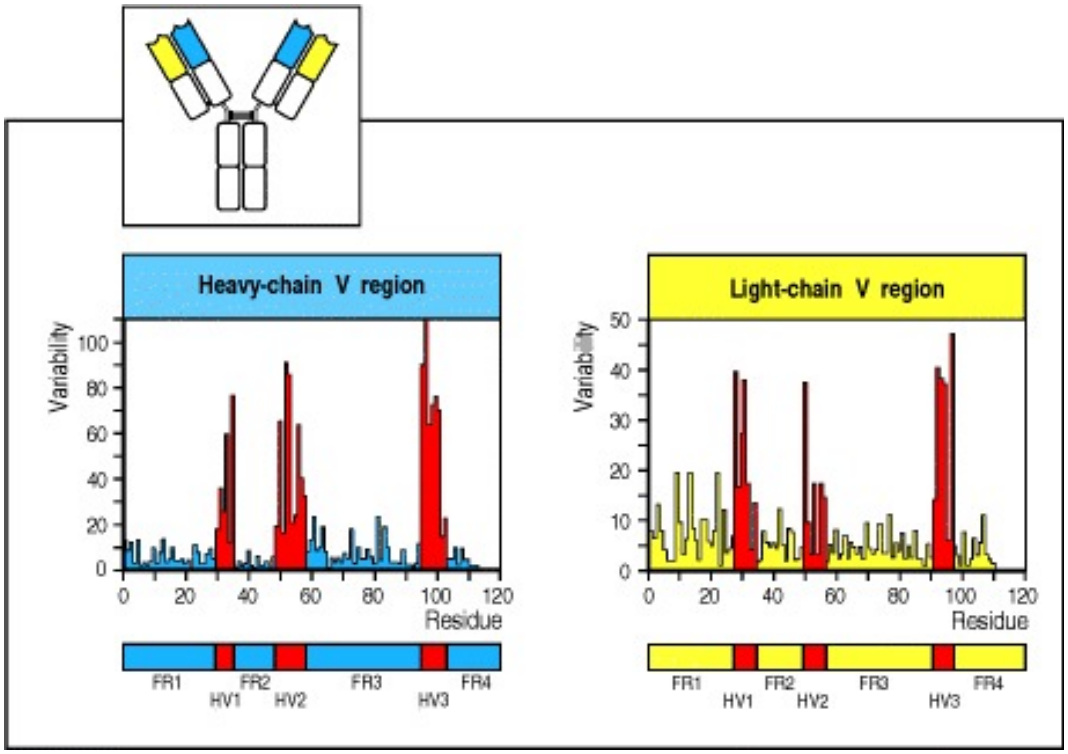


Figure 1.4: The Wu and Kabat variability plot showing the comparison of the amino acid sequences of several light and heavy chain variable domains. The variability of the sequence is plotted against the amino acid residue number. The variability is the ratio of the number of different amino acids observed at a position and the frequency of the most common amino acid. The hypervariable regions (HV1, HV2 and HV3) are shown in red separated by less variable framework regions (FW1, FW2 and FW3). The Figure has been repoduced from Janeway et al. [17].

produce either IgM or IgD antibody. At a later stage during B-cell development, a different heavy chain gene locus can be brought into proximity with the VDJ region resulting in class switching by DNA recombination at switch regions. Like VDJ recombination, recombination events at switch region also occur at the DNA level and result in the permanent loss of intervening constant regions.

1.1.4.3 Variability and Numbering of Variable Domains

At the protein level, the variable domains consist of hypervariable regions (or CDRs) separated by framework regions (FWs) are also present which are relatively conserved across the set of antibodies and provide support to the CDRs. Consequently the variability is not evenly distributed throughout the variable domain, but

it is concentrated in different sections of the variable domain. A variabilty plot was created by Wu and Kabat [15] where the distribution of variable amino acids was plotted by comparing the amino acid sequences of different antibody variable domain (Figure 1.4).

On such a similar set of homologous proteins, a standardized sequence position numbering can be applied in order to simplify the description, analysis and comparison of members of the set. The first standardized numbering scheme (the Kabat scheme) for antibodies was introduced by Kabat *et al.* [18]. This numbering scheme was based only on sequence alignment data because of the unavailability of structural data for antibodies at that time [18]. In 1987, Chothia and Lesk introduced a new numbering scheme (the Chothia scheme) that refined the Kabat scheme using structural data of CDR regions correcting the sites of insertions and deletions (indels) in CDR-L1 and CDR-H1 as compared with the Kabat numbering scheme [19]. In 2008, Abhinandan and Martin introduced a third numbering scheme (the Martin scheme) that not only considered the structural data of CDRs, but also used framework region information. They suggested that the sites of the insertions in some framework regions in the Kabat and Chothia schemes are incorrect [20] and introduced a corrected version of the numbering scheme. In addition to these, two numbering schemes, IMGT [21] and AHo [22], are also available which unify numbering across antibody light and heavy chains, and T-cell receptor α and β chains. The IMGT numbering scheme is not structurally correct while the AHo scheme is. Neither makes use of insertion codes; instead an expected range of position numbers is provided based on the currently available data. Therefore, a lack of insertion codes may cause problems in these schemes on discovery of extreme-length insertions in future.

1.2 Epitopes

Epitopes are regions on an antigen recognised by antibodies and T-cells. On recognition by these cells, an immune response is induced. The immune system has the unique characteristic of recognising non-self proteins (i.e. proteins of foreign organisms) where epitopes are always derived from these non-self proteins except in case of auto-immune disease. One can define B-cell epitopes as antigenic determinants which are recognized by paratopes (the combining site of an antibody). As described previously, epitopes fall into two categories: T-cell epitopes and B-cell epitopes. T-cell epitopes are parts of internalized and processed antigens that are presented to T lymphocytes in association with MHC via the T-cell receptor. In contrast, B-cell epitopes [23] are recognized as three dimensional structures on the surface of native antigens. This chapter will discuss only B-cell epitopes, henceforth generally referred to simply as 'epitopes'.

Types of Epitopes

B-cell epitopes are classified as either continuous, (comprised of a linear segment of an antigen) or discontinuous (also known as conformational), which consist of discontinuous segments of the sequence coming together in the 3D fold of the protein (Figure 1.5). The difference between these two types is not straightforward because discontinuous epitopes often encompass extended regions of sequential residues which can be regarded as continuous epitopes within the larger discontinuous epitope [23].

1.2.1 Continuous Epitopes

A continuous (or sequential) epitope has a single consecutive stretch of amino acids in the protein. In order to evaluate the contribution of each residue to the epitope,

the binding affinity of epitope derivatives with single amino acid substitutions has been measured [24]. It was found that most continuous epitopes contain residues that seem not to be involved in the binding interaction and can be substituted by other residue without showing any effect on the affinity [24]. These continuous epitopes can therefore be regarded as structurally discontinuous, however the substituted residues cannot be removed because of their role as a scaffold. The irreplaceable residues with an effect on binding affinity are categorised as 'functional epitopes' because of the free energy of interaction contributed by each residue [25].

An isolated peptide segment of a protein may not be a true mimic of an antigenic region because it may fail to maintain the same conformation as in the folded protein [23].

1.2.2 Discontinuous Epitopes

A discontinuous (or conformational) epitope is composed of several segments of sequence that are distantly separated in sequence but physically juxtaposed on the protein surface through three-dimensional protein folding. Approximately 90% of B-cell epitopes [14, 26, 27] are conformational and this makes their prediction very challenging. The antigenicity of these epitopes is highly dependent on the native conformation of the protein. Therefore, a discontinuous epitope cannot be isolated as an independent entity from the rest of the molecule in which it is integrated and cannot show an independent binding activity outside of the protein context [23].

This fact has brought serious limitations to the functional characterization of conformational epitopes since their definition depends on structure which is mainly obtained by analysis of antigen-antibody complexes using X-ray crystallography. Using these experimental complex structures, discontinuous epitopes can be obtained by identifying the set of atoms of an antigen making contacts with the atoms

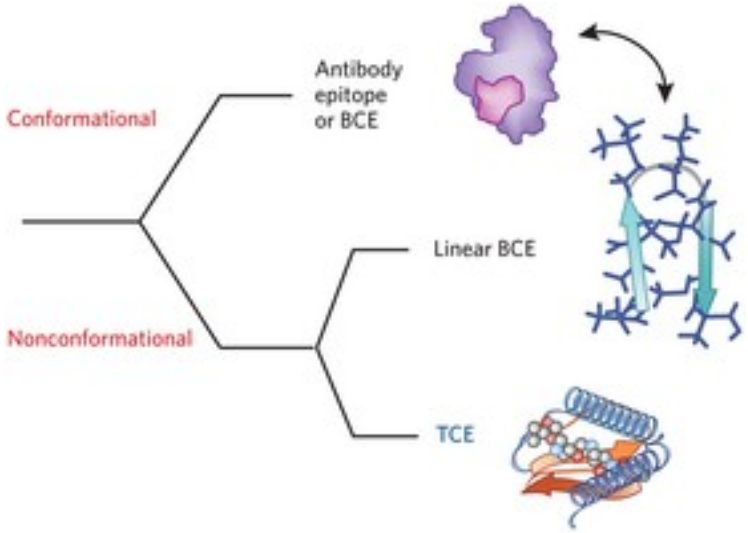


Figure 1.5: Peptide epitopes as conformational and non conformational. B-cell epitopes (BCEs) are either isolated peptides or parts of surface of whole antigens. They are mostly conformational unlike T cell epitopes (TCE) (Figure reproduced from [32].

of an antibody (typically within 4Å). By this definition, a discontinuous epitope consists of 10—22 residues which constitute two to five separate segments.

Considering the above fact, extensive structural studies are needed to evaluate the possibility of mimicking the native conformation of discontinuous epitopes. In principle this could be done by introducing linkers between discontinuous regions or mounting such regions on a scaffold. So far, there has been very little success in reconstituting a discontinuous epitope by synthesis [28–31].

1.3 Synthetic Peptides as B-Cell Vaccines

Over 45 years ago, a German scientist, Anderer, used a small peptide (part of the coat protein of tobacco mosaic virus) to study the immune response [33]. He observed that antibodies raised against a small peptide cross reacted with the viral protein and successfully neutralised the infection. The resulting conjecture that a peptide can mimic the 3D structure of the whole protein could not be verified because of the unavailability of structural data at that time. However, this suggested

the potential of peptide epitopes in vaccine design. Such synthetic peptides and antibodies against them have many potential applications in diagnosis of infectious and autoimmune disorders as well as in development of peptide-based vaccines [33].

The notion of epitope-based vaccines is based on identification and chemical synthesis of B-cell and T-cell epitopes which have the ability to induce a specific immune response. Correct epitope identification and immunization with a precise epitope (potential vaccine), corresponding to the binding site of a potent neutralizing antibody, would induce the generation of similar antibodies against the vaccine. Such a vaccine is referred to as a 'B-cell epitope-based vaccine' [11]. The discontinuous nature of B-cell epitopes has made their identification and prediction challenging compared with linear epitopes and has limited the performance of several B-cell prediction methods [27, 34–40]. This could be attributed to an incomplete understanding of conformational B-cell epitopes and it is therefore important to investigate the 3D structures of B-cell epitopes and gain insights into the segments/regions of which they are comprised.

Several studies have been performed to develop synthetic prophylactic and therapeutic vaccines against various bacterial, viral and parasitic infections. A considerable number of peptide vaccines are under development, such as vaccines for human immunodeficiency virus (HIV) [41], hepatitis C (HCV) [42], malaria [43], foot and mouth disease [44], swine fever [45], influenza [46], anthrax [47], human papilloma virus (HPV) [48], as well as therapeutic anti-cancer vaccines [49–51] for pancreatic cancer, melanoma, non-small cell lung cancer, advanced hepatocellular carcinoma, cutaneous T-cell lymphoma and B-Cell chronic lymphocytic leukemia.

So far, thousands of peptides have been examined pre-clinically and over 300 peptide vaccines have progressed to Phase I clinical trails. The clinical trials

database of the US National Institute of Health (NIH) ClinicalTrials.gov upto the end of August 2017, showed 511 clinical studies of peptide vaccines for preventive and therapeutic purposes had been registered. Currently, 336 candidate vaccines are at Phase I and 246 at Phase II stages of development. Out of a total of 511 studies, only 13 studies have progressed to the Phase III stage. Interestingly, all these 13 studies are based on multiple types of cancers. Recently, in February 2017, one of the peptide vaccine (based on Human Papillomavirus) entered Phase IV clinical trials.

1.3.1 The advantages of Epitope-Based Vaccines

In spite of the fact that, presently, there is only one commercial epitope-based vaccine (for human papillomavirus — HPV) which has recently reached in phase IV (July 2017), it is believed that there are several potential advantages associated with their development. Some of these are discussed here.

1.3.1.1 Specifying The Immune Response

A pathogen contains a plethora of antigens, each composed of multiple potential epitopes. Undoubtedly, protective immunity does not rely on all antigens or epitopes. Moreover, it is also believed that an individual cannot induce a strong immune response to all potential epitopes. In fact, natural repositories of antibodies in an individual's polyclonal serum (in spite of possessing variation complexity), corresponds to a relatively limited collection of epitopes that may not encompass those that are the most useful for protection. Thus, the immune system could be driven towards generating a specific repertoire of antibodies, by creating a cocktail of epitopes with the ability to initiate proven protection i.e. 'hand picked to do the job' [11]. Epitope-based vaccines aim to reduce the number of ineffective antibodies in an infected individual, thus, enhancing the specific activity of neutralizing

antibodies.

1.3.1.2 Exclusion of Undesirable Epitopes

In addition to the enormous range of antibodies that may be generated in response to a pathogen, the presence of deleterious antibodies against the pathogen may bring undesirable effects.

So far, two types of deleterious effects have been investigated. Firstly, the infection of macrophages by Fc-receptor-mediated endocytosis of immunocomplexed pathogens has been attributed to the generation of antibodies against non-effective epitopes [52]. Secondly, an autoimmune response could be initiated owing to the possibilty of epitope mimetics of native host proteins [53, 54]. For instance, a vaccine for Lyme disease was developed against *Borrelia burgdorferi*, its causative agent. It was observed that outer surface protein A (OspA) of the bacterium encompasses a short 9-mer sequence, which is similar to the human leukocyte function associated antigen-1 (hLFA-1). As a consequence, the Lyme disease vaccine seemed to induce an autoimmune response to hLFA-1, leading to arthritic symptoms. Targeted excision of this 9-mer peptide epitope from the OspA has been proposed to resolve this problem [55–57]. Certainly, epitope-based vaccines would enable the rational design of epitope cocktails by removal of epitopes that could be the cause of auto-immune responses.

1.3.1.3 Improving Immunity

Not all the antigens and epitopes of a pathogen have the ability to elicit an immune response by producing antibodies. Very often, epitopes seem to be dominant, but it is not necessarily the case that dominant epitopes correspond to the most effective neutralizing ones [58]. In fact, a pathogen is under natural selection pressure to obscure its ‘weak points’ and distract the host’s immune system by evolving epitopes

with little protective ability. This is done by creating 'baits' that are potentially hydrophilic and surface-accessible but susceptible to constant genetic variations [59]. For instance, the five variable loops of HIV-1 gp120 are immensely immunogenic, but because of their variable nature, HIV escapes immune surveillance by constantly changing the binding capacity of antibodies mounted after first exposure [60].

It is anticipated that epitope-based vaccines would help to overcome this problem. It is expected that selection of particular epitopes would elicit strong neutralization rather than their natural surface accessibility. Some epitopes can evolve through natural selection to be less immunogenic. For example, the epitopes of HIV-1 gp120 that must work in receptor recognition are needed to correspond to the constant structure of the host (such as CD4 and CCR5) [61]. These same epitopes in gp120 are comparatively less immunogenic because they are buried, leading to the production of relatively fewer antibodies against them in the natural development of disease [60]. Ideally, by segregation of these epitopes and presenting them as intact entities, i.e. immunizing with these in the absence of the rest of virus, enhanced immunogenicity could be acquired. Such a reductionist approach to produce epitope-based vaccines would provide more focused, functional and efficient immune responses towards them.

Moreover, immunogenicity of epitope-based vaccines can be enhanced by incorporating lipid, carbohydrate and phosphate groups in a controlled fashion. This may also readily increase stability and solubility [10].

1.3.1.4 Cost Effectiveness

The development of vaccines can be technically difficult and biohazardous when pathogens are required to be cultured in large quantities. In addition, different pathogens require different production protocols, specific conditions and reagents

making the procedure very costly. Manufacture of epitope-based vaccines as synthetic peptides would be less complicated, much safer and cheaper. Moreover, these synthetic peptides can easily be characterized and analysed by using analytical techniques such as chromatography and mass spectrometry.

1.3.1.5 Ease of Storage

A further major advantage of epitope-based vaccines is that they do not require cold-chain facilities for storage, transport and distribution.

1.3.2 Epitope-Based Vaccine Production

There are three steps in the process of epitope-based vaccine production. The first step is epitope mapping (also called epitope discovery), the most important and challenging. This step involves the identification of immunogenic epitopes in the antigen. Once such an epitope has been identified, the second step is reconstruction of the epitope into a functional immunogen. Since the contiguous surface of B-cell epitopes is generally comprised of a few discontinuous segments that are brought together in the protein structure by folding, the reconstruction of an epitope must take this fact into consideration. The contacting residues need to be positioned in a proper spatial orientation. Again, this is not an easy task as the challenge is to retain the 3D conformation. There are two possibilities to engineer an epitope. It could either be done by joining or stapling together the antigenic segments, supported by some sort of scaffolding or, alternatively, the engineered epitope is based on components that are functionally similar, but structurally unrelated mimetics. In the later case, functional moieties of an epitope are effectively placed in space by using alternative residues and chemistries to reconstruct a landscape of the engineered epitope, but with different composition. Eventually, a simpler, but comprehensive approach could be derived to engineer the partial structural segments of an epitope.

For instance, in an epitope comprised of a series of anti-parallel β strands forming a number of juxtaposed β hairpins, the construction of each β hairpin separately could be considered instead of attempting to encompass a whole epitope [11]. Once an epitope is reconstructed, the third step in the production of an epitope-based vaccine is developing effective immuogens that are capable of safe delivery to the host and which produce a specific immune response. For this purpose, particulate carriers are required for delivery and adjuvanting [11].

1.3.3 Challenges in Peptide Mimetics

Of the many challenges in developing peptide-based vaccines, one is to identify the epitope with reasonable accuracy and precision. The most effective way of identifying the epitope is by studying a crystal structure of the antibody-antigen complex. While designing a peptide to mimic a continuous epitope may be relatively straightforward, designing a peptide that mimics the structure of a discontinuous epitope is clearly a much more difficult problem and requires a deeper understanding of discontinuous epitopes at a structural level. While generally thought to be a minor effect in protein antigens, another potential challenge is the possibility of conformational change on formation of the antibody-antigen complex [62, 63]. Such changes are more common when antibodies bind short peptides, but can also occur with protein antigens [64]. Another problem is that a peptide epitope, taken out of the context of the whole protein, will not necessarily adopt the same conformation as the native/whole protein antigen and consequently may fail to induce an immune response which generates antibodies that cross-react with the native protein. If a peptide adopts a conformation more similar to the conformation that it has in the native protein, it is more likely to activate a B-cell response that generates specific antibodies that will being to whole antigen. There are various approaches to induce

peptides to fold correctly [65], however any approach which involves the integration of conformational elements from discontinuous epitopes requires complete knowledge of the 3D structure of the native protein. The conformational dependency of B-cells for their activation has led to elucidating the structural nature and composition of B-cell epitopes and their binding interactions with antibodies. In order to address these conformational challenges, a detailed analysis of 3D structures of discontinuous epitopes is needed to understand both the level of discontinuity and the conformational nature of the continuous stretches of a discontinuous epitope.

1.3.4 Engineering Peptides for Vaccine Design

There are several approaches that have been adopted to enhance the effectiveness of synthetic peptide therapeutics. These strategies include glycosylation, amino acid sequence modification and substitution and cyclization. Out of these, amino acid substitutions within the sequence of a peptide epitope are highly significant in vaccine design owing to the possibility of adopting favourable conformations leading to the generation of a potent immune response. Recently, cyclization of linear peptides has become an attractive tool to provide conformationally more restricted and biostable analogues [66].

1.4 Aims and Objectives

In order to explore the conformational challenges associated with discontinuous B-cell epitopes for the design and use of epitope-based vaccines, there is a definite need for structural analysis of epitopes. Moreover, this analysis requires a dataset comprised of structures of antibody-antigen complexes, however the unavailabilty of a clean dataset provided a need to create such a dataset.

The major aims of this project were:

1. Generation of a dataset containing antibody-antigen structures which could be used to map epitopes (regions making contacts with the antibody) on the surface of the antigen structures.

2. To analyze the level of structural discontinuity of epitopes. While it is well established that the majority of B-cell epitopes are discontinuous, the structural nature of this discontinuity needs to be investigated. A linear epitope, by definition, would consist of a single continuous stretch (or in the terminology used here, a 'region') of peptide, while, the structural composition of a discontinuous epitope can vary much more widely. At one extreme, it could consist of two or more such regions while at the other extreme it could consist of purely scattered residues, not in continuous regions of primary sequences (defined here as 'fragments'). Overall, an epitope could consist of zero or more regions together with zero or more fragments

3. While a region is defined as a continuous stretch of amino acids in the primary sequence, this tells us nothing about the shape of these regions. Thus the third objective is to analyze the shape of the regions to determine whether they are truly linear, curved or form more hairpin-like, or otherwise locally folded structures.

4. To determine the extent to which epitopes can be mimicked using isolated peptides. Clearly this should be straightforward using a true linear epitope and it should, in principle, be possible to link regions into a single peptide. However, within the scope of this thesis, epitopes with only a single region were considered to study the conformational stability of the isolated peptides out of native antigen. Hydrophobic to hydrophilic mutations within these iso-

lated peptides may help to stablise the native conformation as may disulphide stapling and cyclisation of folded peptides. Therefore, the aim is to study five linear and five folded epitope regions and explore stabilising mutations stapling and cyclisation using molecular dynamics simulations.

5. Having found potentially stablising mutations in the selected epitope regions, the next question is whether they will adopt the same conformation *in vitro*. To confirm this, experimental validation needs to be performed.

Chapter 2

Introduction to Methods

Overview

This chapter outlines the general theoretical framework of methods, either computational or experimental, that have been applied in this thesis. The computational methods include molecular dynamics simulations used to explore the stability of isolated peptides and their mutants (*in silico*) while experimental methods include Circular Dichroism (CD) spectroscopy and Nuclear Magnetic Resonance (NMR) used to study the structural stability of selected peptides and Surface Plasmon Resonance (SPR) which was employed to study the binding association of peptides with antibody (*in vitro*).

2.1 Computational Techniques

2.1.1 Molecular Dynamics

Molecular dynamics (MD) has been used widely to study the structural dynamic properties of macromolecules such as proteins, peptides and DNA at the atomic level. The MD simulations in this thesis are described as 'classical MD' which means that quantum effects (such as the motion of electrons) are neglected. This

section briefly summarises the equations, principles and approximations on which MD simulation is based.

In principle, the MD methodology is quite simple. First, all of the atoms in a system are assigned initial velocities, coordinates and (partial) charges. A potential is computed using the positions and charges which is used to calculate the force experienced by each of the atoms from which an acceleration can be calculated. A new set of coordinates (positions) and velocities are generated for each of the atoms after a short time step effectively integrating Newton's laws of motion. The new values are then passed back into the first step of calculation and the process is repeated, producing a trajectory of atomic positions and velocities over the time. Further explanation of these principles and approximations has been broken down into three sections which are described below.

2.1.1.1 Approximation I – Born-Oppenheimer

The dynamics of any system is illustrated by the time-dependent Schrödinger equation (Equation 2.1).

$$\mathrm{E}\Psi = \hat{\mathrm{H}}\Psi \tag{2.1}$$

where $\hat{\mathrm{H}}$ represents the Hamiltonian operator which is equal to the sum of the potential and kinetic energy, Ψ is the wave function which is comprised of all particles (coordinates and momenta of both nuclei and electrons) of the system. The low mass of electrons makes them move with a much higher velocity compared with the nuclei. In the Born-Oppenheimer approximation [67], the assumption is made that the motion of atomic nuclei and electrons can be separated and the total wave function is simply the sum of nuclear and electron wave functions (Equation 2.2).

$$\Psi_{tot} = \Psi_{nuclei} + \Psi_{electrons} \tag{2.2}$$

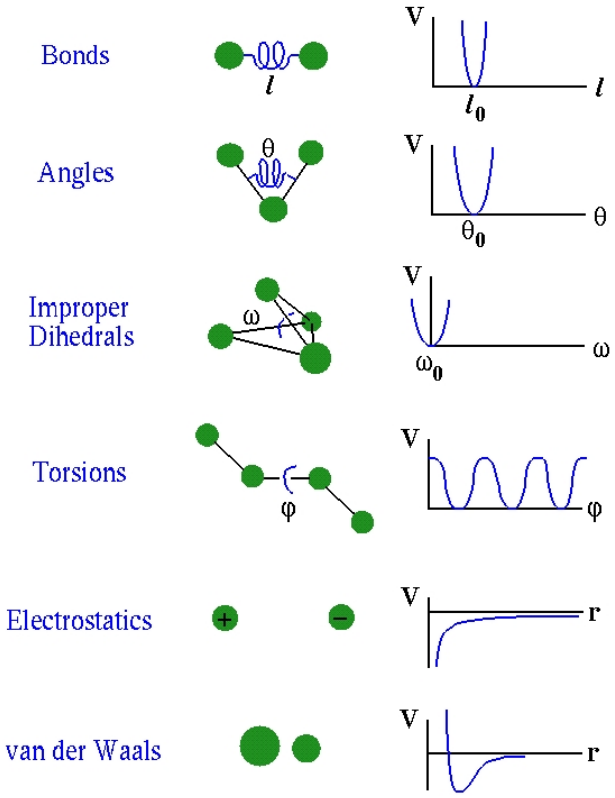


Figure 2.1: Schematic representation of ball-and spring model for potential energy function in MD simulations. The Figure has been reproduced from Steinbach's review on macromolecular simulation [69].

In classical MD, this approximation is taken one step further and it is assumed that the electron motion is dependant only on the nuclear configuration. Consequently, nuclear motion develops gradually on a single potential energy surface (PES) which is coupled with a single electronic quantum state which is acquired by solving the time-independent Schrödinger equation over the range of fixed nuclear geometries. In reality, MD simulations are carried out on a ground state potential energy surface, and the atomic nuclei are treated as classical particles [68]. In other words, nuclei experience electrons as an average field in classical MD.

2.1.1.2 Approximation II – Force Fields

A force field is a collection of mathematical equations that estimates the potential energy of the system and mainly depends on the structural conformation. The physi-

cal system is comprised of collections of atoms that are held together by interatomic forces. Hence, a force field takes account of the forces acting on the system to calculate the potential energy of a molecule. A simple illustration of a force field is a 'ball-and-spring' model where atoms representing balls and connected via bonds (springs) [70] (Figure 2.1).

A typical force field considers bonded interactions, namely energy terms for bonds, angles, dihedrals (impropers and torsions), and non-bonded interactions which include elecrostatics and van der Waals. Thus, a potential energy fuction used in MD simulation can be described by the Equation 2.3:

$$E_{tot} = E_{bonded} + E_{non-bonded} \tag{2.3}$$

where each term is a collection of interactions mentioned above and can be re-written as:

$$E_{bonded} = E_{bond} + E_{angles} + E_{dihedral} \tag{2.4}$$

$$E_{non-bonded} = E_{electrostatic} + E_{van-der-Waals} \tag{2.5}$$

Over the years, a number of force fields have been developed with varying parameterisation strategies, but a typical force field is shown in Equation 2.6 [71–74].

$$\begin{aligned} E_{tot} = \; & \textstyle\sum_{bonds} k_l (l - l_0)^2 + \\ & \textstyle\sum_{angles} k_\theta (\theta - \theta_0)^2 + \\ & \textstyle\sum_{impropers} k_\omega (\omega - \omega_0)^2 + \\ & \textstyle\sum_{torsions} A_n [1 + cos(n\phi - \phi_0)] + \\ & \textstyle\sum_{i<j} \left(\varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_r \varepsilon_0 r_{ij}} \right) \end{aligned} \tag{2.6}$$

where the first 'bonded' sum is over bonds between two atoms; the second sum represents bond angles defined by three atoms; the third and fourth sums are calculated on the basis of four atoms (Figure 2.1). In the 'non-bonded' interactions, atom pairs are parametrised in terms of partial charges, q_i, for Coulombic interactions, and the terms ε_{ij} and r_{ij}^{min} represents the depth and width of the Lennard-Jones potential for atom type i and j, respectively [75]. These force field parameters are derived by fitting to experimental measurements or structural thermodynamics to quantum level calculations [76].

In this thesis, the conformational stability of small peptides and antibody-peptide complexes is studied using the AMBER ff99SB*-ILDN force field [77] which is a modified form of the original AMBER force field [71]. The types of modifications in the applied force field are described in Section 5.2.

2.1.1.3 Approximation III – Classical dynamics and Equations of Motion

In MD simulations the time evolution of interacting individual atoms (resulting MD trajectories) is followed by simultaneous integration of Newton's second law of motion (Equation 2.7 and 2.8).

$$\mathbf{F} = m\mathbf{a} \tag{2.7}$$

$$\mathbf{F}_i = m_i \frac{d^2\mathbf{r}_i(t)}{dt^2} \tag{2.8}$$

where $\mathbf{r}_i$(t) is the position vector of ith atom and is equal to $(x_i(t), y_i(t), z_i(t))$ and $\mathbf{F}_i$ is the force acting upon ith atom at time t and m_i represents the mass of the atom. The force, $\mathbf{F}_i$, on ith atom at position $\mathbf{r}_i$ determines its acceleration a_i which

causes change in the velocity and position of atom *i* over the given time step *dt*. In order to inegrate Newton's equations of motion numerically, the Verlet algorithm is used [78]. The basic aim of this numerical integration is finding an expression that determines positions $\mathbf{r}_i(t + \Delta t)$ at time $(t + \Delta t)$ in terms of known positions at time t. The basic formulation of Verlet algorithm can be derived from the 3^{rd} order Taylor expansion at positions $\mathbf{r}_i(t + \Delta t)$ and $\mathbf{r}_i(t - \Delta t)$ and is shown in Equation 2.9.

$$\mathbf{r}_i(t+\Delta t) \cong 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t^2 \tag{2.9}$$

A relatively smaller integration time step (Δt) must be chosen as compared with the fastest motions of the systems. Because the bond vibrations that involve light atoms such as hydrogen happen within a period of several femtoseconds, this suggests that Δt should be on the subfemtosecond time scale to ensure stability of the integration. However, the fastest and unimportant vibrations can be ignored by applying constraints on the bond lengths to allow larger time steps. The simulations presented in this thesis make use of LINCS algorithm for this purpose and a time step of 2 fs [79].

2.2 Experimental Techniques

2.2.1 Circular Dichroism Spectroscopy (CD)

Circular dichroism spectroscopy is one of the most widely used techniques to characterise the peptide and protein conformation in solution. CD is a measure of the difference between absorption of left-handed and right-handed circularly polarised light which is used to study chiral chromophores. These differences in measurements are highly sensitive to the secondary structure of proteins and peptides. CD allows the estimation of the secondary structure composition of a protein (alpha helices, parallel and anti-parallel beta sheet, turns and disordered regions). In addition, one of the key applications of CD is studying the folding properties and effects of mutations on the conformation and stability of proteins. The presence of an ordered structure in a protein results in a CD spectrum with distinct positive and negative peaks, whereas an unstructured protein shows a relatively flat spectrum. The key peaks are located in the far ultraviolet portion of the spectrum between 190 and 250 nm as shown in Figure 2.2. During the work in this thesis, this technique has been used to study the secondary structure of isolated peptides.

2.2.1.1 Physical Principles

The underlying physical principle is the measurement of CD (as defined above) over a range of wavelengths by the absorption of right and left circularly polarised light by chromophores in chiral molecules. Owing to the chiral nature of biomolecules and in particular, proteins having all chiral amino acids (except glycine), CD is a highly suitable for structural characterisation. The presence of the peptide bond, connecting the residues, provides a ubiquitous chromophore in peptides and proteins. Proteins typically absorb UV-light in the far-UV region (λ = 190-250 nm) resulting in two characteristic absorption bands: a Π-Π^* electronic transition (at

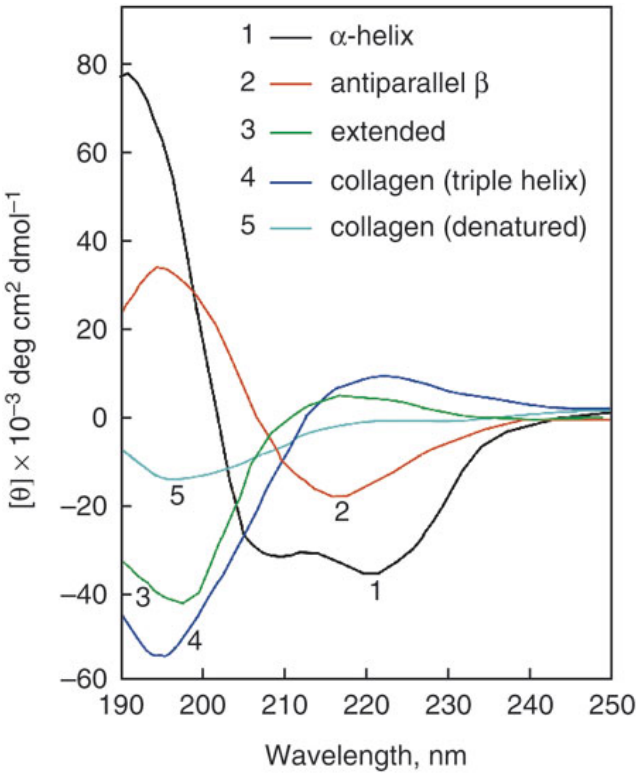


Figure 2.2: CD spectra of poly-L-lysine in the (1) α-helical (black) (2) antiparallel β-sheet (red) conformations, (3) extended/random coil (green) conformation, (4) native triple-helical (blue) and (5) denatured (cyan). The Figure is taken from a review by Greenfield [80].

190 nm) and a n-Π^* electronic transition (at 210 nm) which are associated with aromatic residues and amide bonds, respectively. The sensitivity to amide transitions (n-Π^*), in the range of 190-240 nm, gives rise to the characteristic CD spectra [81, 82]. Figure 2.3 explains the working principle of a CD spectrometer. In order to understand the physics behind CD, circularly polarised components can be described more precisely by the field vectors of right (E_R) and left (E_L) circularly polarised light and their superposition (E_R + E_L), before and after traversing the sample (Figure 2.4).

In the case of linearly polarised light, a combination of right (E_R) and left (E_L) cicularly polarised light with a similar amplitude and wavelength produces plane polarised light. Hence, the superposition of both of these light beams yields a simple

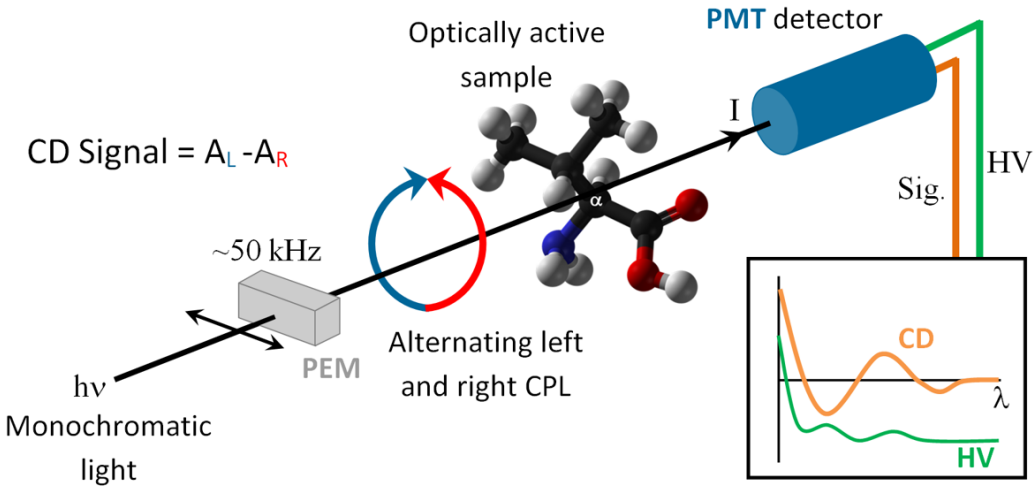


Figure 2.3: The principle behind CD spectroscopy. A monochromatic light beam passes through a Photo Elastic Modulator (PEM) which performs the conversion of a linearly polarised light beam into a circular or alternating left and right handed polarised light beam, traversing the sample and following Lambert-Beer's law, causing the two polarisations to be absorbed differently. Subsequently, the difference in absorption is detected by a Photo Multiplier Tube (PMT), amplified and recorded by a computer. Image obtained from http://www.isa.au.dk/news/rosetta-nov2014.asp

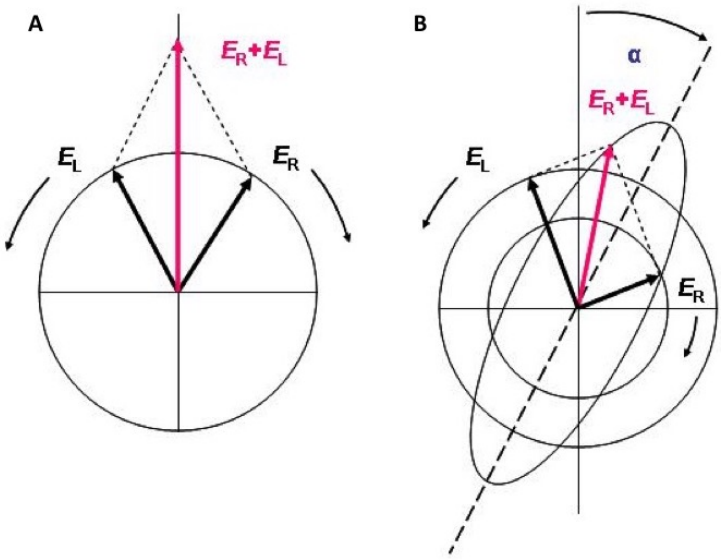


Figure 2.4: A) Linear polarized light, as a superposition of opposite circular polarized light, of equal amplitude and phase, but with opposite handedness. B) Ellipticity (CD) and optical rotation (OR) as a result of varying absorption of the left and right handed polarised components. Image obtained from http://www.ruppweb.org/cd/cdtutorial.htm

line (Figure 2.4A). This happens before the light beam passes through the sample. By contrast, beam passage through an asymmetric sample (optically active) results in absorption to different extents generating an elliptical shape signal instead of a line (Figure 2.4B). The phenomenon of this ellipse is reffered as circular dichroism.

2.2.1.2 Data Analysis

Though modern spectrometers directly measure the difference in absorption ΔA, the widely used and accepted measurement unit for CD is ellipticity, θ. At a particular wavelength, the difference between left (A_L) and right (A_R) polarised light absorption can be described by Equation 2.10

$$\Delta A = A_L - A_R \tag{2.10}$$

The application of Lambert-Beer's law produces:

$$\Delta A = (\varepsilon_L + \varepsilon_R)dc = \Delta\varepsilon dc \tag{2.11}$$

where d is path length in cm, c concentration of the sample in mol/L, ε_L and ε_R represent the molar extinction coefficients of left and right circular polarised light. This results in $\Delta\varepsilon$ circular dichroism which is path length and concentration independent.

In contrast, the ellipticity (θ) is derived from the magnitudes of the electric field vectors, (E_R) and left (E_L) (described above), using the rules of trigonometry. Mathematically, these field vectors are replaced by the square root of the irradiation intensity of the left and right circular polarised light. Here, the application of Lambert-Beer's law provides a complex equation that is solvable by a Taylor series. This produces the following equation:

$$\theta = \Delta A(\frac{ln10}{4})(\frac{180}{\pi}) \tag{2.12}$$

The molar ellipticity $[\theta]$ can be defined to remove the dependence of θ on the path length and concentration:

$$[\theta] = \frac{100\theta}{dc} \tag{2.13}$$

Combining equations 2.11, 2.12 and 2.13, we obtain:

$$[\theta] = 100\Delta\varepsilon(\frac{ln10}{4})(\frac{180}{\pi}) = 3298.2\Delta\varepsilon \tag{2.14}$$

The absolute intensity of a protein is compared with the number of amino acids and commonly given as the mean residue ellipticity and results in Equation 2.15:

$$[\theta]_{mr} = \frac{[\theta]100M}{cdN} \tag{2.15}$$

where M is the molar weight in g/mol, c the concentration in mg/mL, d the path length in cm and N the number of amino acids in the peptide/protein. A slightly different version of this equation has been used in this thesis and is given in Equation 2.16. This is because the peptide concentration was expressed in micromolar.

$$[\theta]_{mr} = \frac{10^6[\theta]}{cdN} \tag{2.16}$$

where c is the peptide concentration in micromolarity, d is the path length in millimetres and N is the number of residues.

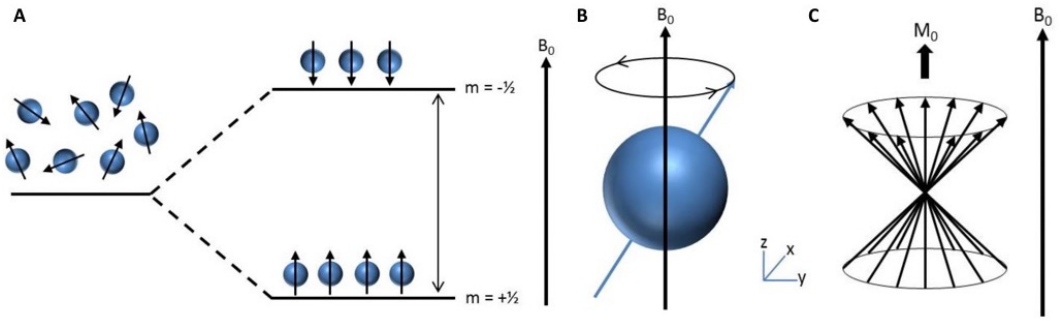


Figure 2.5: A) Zeeman splitting under a magnetic field — nuclei with I = 1/2. B) A single nucleus having angular momentum and the frequency of precession which depends on its gyromagnetic ratio and the strength of the magnetic field. C) The ensemble average magnetisation (M_0) caused by the excess of nuclear spins in the lower energy state as a result of an external magnetic field.

2.2.2 Nuclear Magnetic Resonance (NMR)

Nuclear magnetic resonance spectroscopy is one of the most powerful methods to produce high resolution structural information about biological macromolecules such as proteins and nucleic acids at atomic resolution. In NMR, structure determination is carried out in solution providing a mimic of natural and physiological environment which makes it different from other structure determination techniques such as X-ray crystallography. In this thesis, NMR was carried out on two peptides to study their residue level secondary structure.

2.2.2.1 Spin

The basic principle of NMR rests on the fact that some atomic nuclei have magnetic properties that can be used to produce chemical information. The subatomic particles (electrons, protons and neutrons) possess a spin and NMR exploits this intrinsic atomic property. The spin of an atom is characterised by its 'spin' quantum number (I) which is defined by the nuclear framework of an atom: a nucleus having an equal number of protons and neutrons will have no spin, i.e. I = 0; if the sum of the number of protons and neutrons is odd then it will possess a spin of a half integer (such as 1/2, 3/2, 5/2); if the numbers of protons and neutrons are both odd then this will

result a nuclear spin of an integer (such as 1, 2, 3). A relationship between quantum number (I) and magnetic momentum (μ) of a nucleus is described as a ratio, called the gyromagnetic ratio (γ):.

$$\gamma = \frac{2\pi\mu}{hI} \tag{2.17}$$

where h is Planck's constant. When atoms having a spin are placed in an external magentic field, they are capable of undergoing transitions between nuclear energy spin levels. For protein NMR, the most important nuclei are ^{1}H, ^{13}C and ^{15}N (I = 1/2). A nucleus of spin, I, will have 2I + 1 states which means that the nucleus with I = 1/2 under a magnetic field effect will split into two energy levels (m = 1/2 and m = -1/2). In NMR, this effect is known as the Zeeman effect (Figure 2.5A). The initial population of these energy levels is thermodynamically determined by the Boltzmann distribution which results in a slightly larger nuclei population in the lower energy levels. Electromagnetic radiation can be used to excite the nuclei into higher energy states. The frequency of this radiation needed for transition between two energy levels is proportional to the difference in energy levels. This difference is controlled by the strength of the external magnetic field (shown as B_0 in Figure 2.5) and the intrinsic gyromagnetic ratio of each nucleus. The energy difference can be defined as:

$$\Delta E = \frac{\gamma h B_0}{2\pi} \tag{2.18}$$

The population of each state, N, can be computed by Equation 2.19, where k is the Boltzmann constant, T is the temperature and ΔE is the energy difference. This shows that the ratio between the two states is very small at thermal equilibrium. This small difference in the population results in a net magnetisation (M_0).

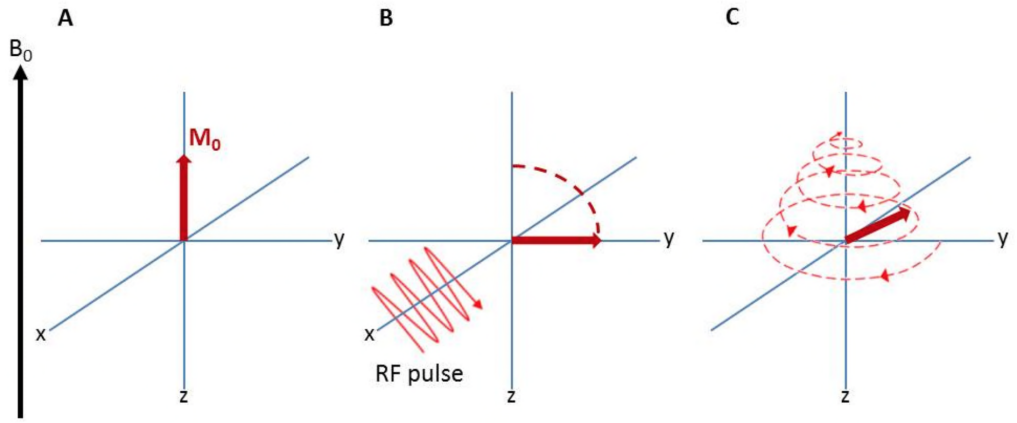


Figure 2.6: A 90° RF pulse. A) The average magnetic momentum (M_0) at equilibrium, B) An RF impulse causing M_0 to flip into the x-y plane, C) M_0 relaxes to the equilibrium position after the RF pulse being precessed around the z-axis.

$$\frac{N_{upper}}{N_{lower}} = e^{-\frac{\Delta E}{kT}} \tag{2.19}$$

2.2.2.2 Net Magnetisation

When a nucleus having intrinsic angular momentum is placed in a magentic field, it will precess around the axis of the magnetic field (Figure 2.5B) with a frequency, ω, known as Larmor frequency:

$$\omega = \gamma B_0 \tag{2.20}$$

All the nuclei of a specific type in an NMR sample will precess at those nuclei's Larmor frequencies producing an average magnetic momentum, M_0 (coming from individual nuclei) with a common angular frequency. Figure 2.5C shows the average magnetisation as M_0 which is aligned with B_0. During an NMR experiment, after placing a sample in the magentic field, M_0 aligns with B_0 until equilibrium is achieved. A radio frequency (RF) electromagnetic pulse is applied to the NMR sample which perturbs the equilibrium. This causes the rotation of M_0 followed by its precession into the x-y plane resulting in the induction of a voltage in the detector

coil that is positioned in this plane. The RF pulse balances the probability of transitions between lower and higher energy levels. Since there is a higher population in the lower energy state, more transitions will occur from lower to higher than from higher to lower leading to an imbalance. After the completion of the pulse, M_0 will restore its previous position aligned with B_0. This longitudinal relaxation is known as T_1 or spin-lattice. The spin-spin relaxation (T_2), which is caused by interactions, redistributes the energy among the components (nuclei) of the spin system, causing the restoration of this relaxation back to the equilibrium, leading to the decay of the NMR signal with time (Figure 2.6). This decaying signal is named FID (free induction decay) which is detected during the NMR experiment. Since all the nuclei influenced by the RF pulse contribute to the FID, the time domain signal needs to be converted to a frequency domain spectrum by Fourier transformation.

2.2.2.3 Chemical Shift

Different nuclei possess different gyromagnetic ratios and different Larmor frequencies which leads to unique signals at different frequencies. The Larmor frequency of a specific nucleus is influenced by the presence of electrons around that nucleus. In the NMR experiment, the applied magnetic field induces a local magnetic field around the electronic cloud of the nucleus in the opposite direction to that of the applied field. As a result, the nucleus experiences a small but measurable effect which either can cancel out a small amount of the external magnetic field (which is termed 'shielding'), or strengthen the field (which is called 'deshielding'). These effects cause all the nuclei of similar type, but in different chemical environments, to have slightly different Larmor frequencies. Such differences in the Larmor frequency are known as chemical shifts and are measured in parts per million (ppm) instead of Hz to remove the effect of the static magentic field strength. The following equation

defines chemical shifts:

$$\delta_{ppm} = \frac{\omega_O - \omega_{ref}}{\omega_{ref}} \tag{2.21}$$

where ω_O is the observed frequency and ω_{ref} is the frequency of a reference compound.

2.2.2.4 Two Dimensional (2D) NMR

It is possible to distinguish the differences in chemical shifts of small molecules, however dealing with large macromolecules, such as proteins, provides a high number of chemical shifts that are mostly quite close and overlapping. In order to deal with this problem, two or higher dimensional NMR experiments can be carried out to monitor the chemical shifs across additional frequency axes. These multidimentional experiments can either be homonuclear (same type of nucleus) or heteronuclear (different type of nucleus).

During the work in this thesis, only 2D NMR experiments were carried out and these include TOCSY, NOESY and C-HSQC. These types of experiments have four main phases: **preparation** to initiate and establish the magnetisation; **evolution** where the spins are enabled to precess; **mixing** to allow magnetisation between nuclei; and **detection** of the signal. A 2D experiment is actually a progression of 1D experiments with different magnetisation mechanisms either scalar coupling (via bonds), or dipolar coupling (via space) and different evolution and detection times.

Total Correlation Spectroscopy (TOCSY)

TOCSY is a homonuclear 2D experiment (using ^{1}H nuclear detection) which makes use of isotropic mixing to transfer magnetization between scalar coupled spins that are part of an amino acid residue, giving rise to a characteristic pattern which de-

pends on the side chain leading to the identification of the amino acids. However, there is no scalar coupling across the amide bond i.e. to a different spin system. In other words, each amino acid has a unique spin system that corresponds to the rise of distinct peaks (for all protons) in the ^{1}H-^{1}H TOCSY spectrum. The number of peaks appearing in TOCSY depends on the length of the mixing time. The most certain peaks to arise during a short mixing time are HN-Hα. With an increase in mixing time, connections between of Hβ, Hγ and other side chain protons also appear. The peak patterns allow the type of amino acids present in the protein to be determined. For the identification of spin systems, random coil standard chemical shifts [83] are used as a starting point. The experimental peak pattern in a TOCSY spectrum will deviate if secondary structure is present in the peptide.

Nuclear Overhauser Effect Spectroscopy (NOESY)

NOESY is a homonuclear 2D experiment (using ^{1}H nuclear detection) which depends on the Nuclear Overhauser Effect (NOE) to transfer magnetisation using dipolar interactions through space and scales with the inter-spin distance, *r*. The strength of the NOE is proportional to $1/r^{6}$ and leads to the detection of only protons within a distance of 5 Å. This experiment plays a vital role in protein structure determination and assists in the derivation of the linear sequence of individual spin systems (i.e. amino acids).

Once all the spin systems are assigned to the types of amino acids (using TOCSY), the next stage involves establishing their connection in sequential order in the protein sequence i.e. sequential assignments. The sequential assignments can be difficult in the presence of secondary structure because the helical regions contain sequential amide protons in close spatial proximity. It makes the assignment somewhat harder but also gives information about secondary structure.

Heteronuclear Single Quantum Coherence (HSQC)

HSQC is a heteronuclear 2D experiment which is particularly important for studying protein structures. There are two types of HSQC experiments: 1) ^{1}H-^{13}C HSQC to study the correlation between the aliphatic carbon and its attached protons and unique peaks are obtained for each of the unique carbons (Cα, Cβ, Cδ and so on) and protons in the amino acid. Hence the number of peaks depends on the number of carbons in each spin system. However, there is no magnetization transfer as there is in TOCSY. 2) ^{1}H-^{15}N HSQC is a well-known and frequently performed experiment that provides the correlation between the nitrogen and amide proton. Unlike ^{1}H-^{13}C HSQC, ^{1}H-^{15}N HSQC provides a single peak for each amide group in the protein backbone. In addition to the backbone amide proton peak, the peaks are also produced for sidechains containing nitrogen bound protons such as those in Asn, Arg, Gln, His, Lys and Trp. This spectrum is much simpler than ^{1}H-^{13}C HSQC, TOCSY and NOESY because the latter provide multiple peaks for a single amino acid. This is one of the commonly used experiments because it gives an overview of the entire protein backbone with one experiment by providing a fingerprint that is easy to monitor for changes.

Theoretically, these require isotopic labelling of C and N, however it is possible to collect data with the natural abundance of ^{13}C using a higher sample concentration. This is because ^{13}C is 1.1% naturally abundant as compared with the ^{15}N which has only 0.4% natural abundance. In this thesis, ^{1}H-^{13}C HSQC was carried out and peptides were not isotopically labelled.

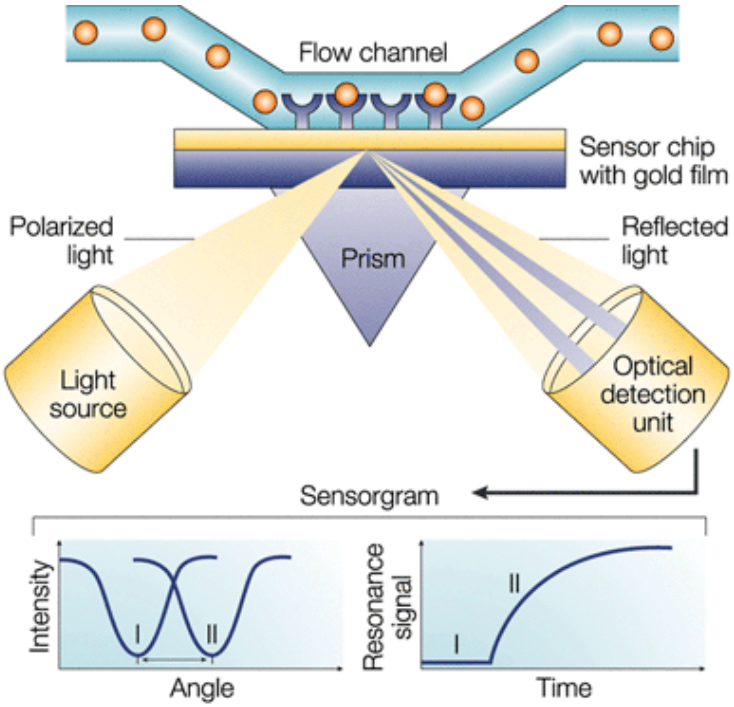


Figure 2.7: Schematic representation of SPR working principle. The change in the refractive index of material near the surface layer of a sensor chip is monitored by the SPR biosensor. An intense beam (shadow) of light is reflected at an angle which relies on the mass of immobilised molecules at the surface. On biomolecular interaction, the mass of the surface layer increases and results in the shifting of SPR angle from I to II. This change in resonant angle (which depends on the mass change) can be detected as a function of time. The Figure has been reproduced from Cooper's review on optical biosensing. [85]

2.2.3 Surface Plasmon Resonance (SPR)

Surface Plasmon Resonance (SPR) in a powerful label-free optical technique that is used for the detection of bi-molecular interactions where one of the molecules is in the mobile phase and the other is immobilised on a thin gold sensor chip [84]. SPR is widely used for studying binding and molecular interactions. In this thesis, SPR has been used to study the binding of peptides with antibody.

2.2.3.1 General Principle of SPR

Surface plasmon resonance is a phenomenon that happens when a photon of incident light strikes a metal surface and reflects at a critical angle of incidence. Within this angle of incidence, the light energy excites the electrons in the metal surface

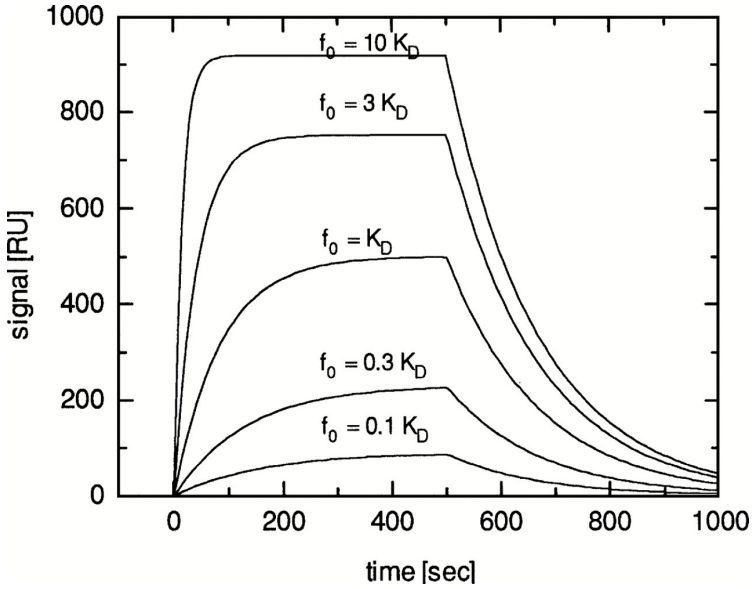


Figure 2.8: Schematic representation of SPR data for a simple 1:1 interaction. The data are fitted to the expected signal for different concentrations of the mobile molecule. The association phase from 0-500 sec, is followed by the dissociation phase from 500-1000 sec. The Figure has been taken from Schuck, 1997 [84].

layer (now referred as plasmons) and consequently they move parallel to the metal surface. The plasmon excitation generates an electric field (with a range of 300 nm in the BiaCore SPR instrument) and electromagnetic waves propagate at the interface of the metal surface and the buffer solution containing the sample (a dielectric medium) [86]. The metal surface is coated with a thin layer of gold which does not let the light refract across the interface and, as a result, total internal reflection (TIR) is observed [87]. Commercial biosensors make use of a high-reflective index glass prism where incident light is used which is reflected out of the prism and the angle where resonance is satisfied is recorded on the detector. The resonance angle depends on the refractive index of the material near the metal surface (Figure 2.7).

A typical biosensor experiment is comprised of multiple steps. Firstly, one of the candidate molecules is attached to the sensor surface. Secondly, the other mobile molecule is introduced at constant concentration into the buffer flow above the sensor chip. As a result of affinity interaction, a probe-target binding happens

which eventually causes an increase in the refractive index at the SPR detector, and hence the association of mobile molecule with the immobile molecule is recorded. The change in signal over time is measured in resonance or response units (RU), where 1 RU is equal to a critical angle shift of 10^{-4} degrees. Thirdly, in the dissociation phase, the time taken to dissociate the mobile molecule is monitored. Finally, chip regeneration is performed to remove any remaining complex from the sensor surface. There are several different strategies that can be used to regenerate a chip. These include usage of low/high pH buffer, high salts, denaturants, etc. However, the final step is not always required as it depends on the type of molecule that fails to dissociate from the chip. In this thesis, binding kinetics were studied by repeating the association, dissociation and regeneration steps with different concentrations of the mobile molecule. The binding kinetics curves are shown in Figure 2.8 and provide insights into the kinetic rate constants and the thermodynamic equilibrium constant of the interaction [84].

Association and dissociation kinetic rate constants are calculated by fitting the data on the sensorgrams of the concentration series with the BIAevaluation software using the 1:1 binding model with RI = 0. The model describes a 1:1 interaction at the surface as shown in the equation 2.22.

$$A + B \leftrightarrow AB \tag{2.22}$$

The affinity between antibody and peptides are calculated with the BIAevaluation software using the steady state 1:1 affinity model. In this model, the equilibrium dissociation constant K_D is computed for a 1:1 interaction using the plot of steady binding levels R_{eq} against analyte concentration (C). The model uses Equation 2.23 as a basis to compute the affinity values using steady state data:

$$R_{eq} = \frac{C \text{ x } R_{max}}{(C + K_D)} + RI \tag{2.23}$$

where R_{max} is the analyte binding capacity of the surface (RU) and RI is the refractive index contribution which serves as an offset on the R_{eq}-axis.

Chapter 3

AbDb: Antibody Structure Database

In order to analyze structures of proteins of a particular class, these need to be extracted from the Protein Databank (PDB). In the case of antibodies, there are a number of special considerations: (i) identifying antibodies in the PDB is not trivial, (ii) they may be crystallized with or without antigen, (iii) for analysis purposes, one is normally only interested in the *Fv* region of the antibody, (iv) structural analysis of epitopes, in particular, requires individual antibody-antigen complexes from a PDB file which may contain multiple copies of the same, or different, antibodies, (v) standard numbering schemes should be applied. Consequently, there is a need for a specialist resource containing pre-numbered non-redundant antibody *Fv* structures with their cognate antigens. This chapter describes the creation of an automatically updated resource, AbDb, which collects the *Fv* regions from antibody structures using information from our SACS database which summarises antibody structures from the PDB. PDB files containing multiple structures are split and numbered and each antibody structure is associated with its antigen where available. Antibody structures with only light or heavy chains have also been processed and sequences of antibodies are compared to identify multiple structures of the same antibody. The data may be queried on the basis of PDB code, and the name or species

of the antibody or antigen, and the complete datasets may be downloaded.

Database URL: `www.bioinf.org.uk/abs/abdb/`

3.1 Introduction

A number of databases contain antibody sequence data [88–94], while others provide summaries of, or access to, structural data [91, 92, 94–96]. Antibody structures currently represent ∼2.2% of the Protein Databank (PDB, May 2017). There are a number of resources that make information about antibody structures available. The first specialised structural resource was **SACS** [95] which simply provides a regularly-updated list of antibody structures from the PDB with some basic annotations. **abYsis** [91] is predominantly an analysis resource; it brings together antibody sequence data from different publicly available data sources including Kabat, EMBL-ENA, the PDB and, optionally, IMGT, providing tools to analyze antibodies within the web-based resource (e.g. numbering [20], canonicals [97, 98], unusual residues, humanness [99], germline source, etc.) **IMGT/3Dstructure-DB** [92] provides an interface for the inspection of antibody structure data including bound protein antigens. **SAbDab** [96] provides a web-based search interface and allows the original antibody PDB files to be downloaded, as well as PDB files with Chothia numbering applied. These numbered files also contain information about the pairing of light and heavy chains and identify associated antigen chains.

In this chapter, the term 'antibody' is used to refer to antibody-derived antigen binding fragments including Bence-Jones light chain dimers and camelid VHHs. Thus the relationship between antibody and antigen chains in a PDB file is complex. Antibodies may consist of a light-chain dimer, a single heavy chain, or a conventional light/heavy complex. Likewise, antigens may consist of small molecules or of one or more protein or non-protein chains and the epitope to which an anti-

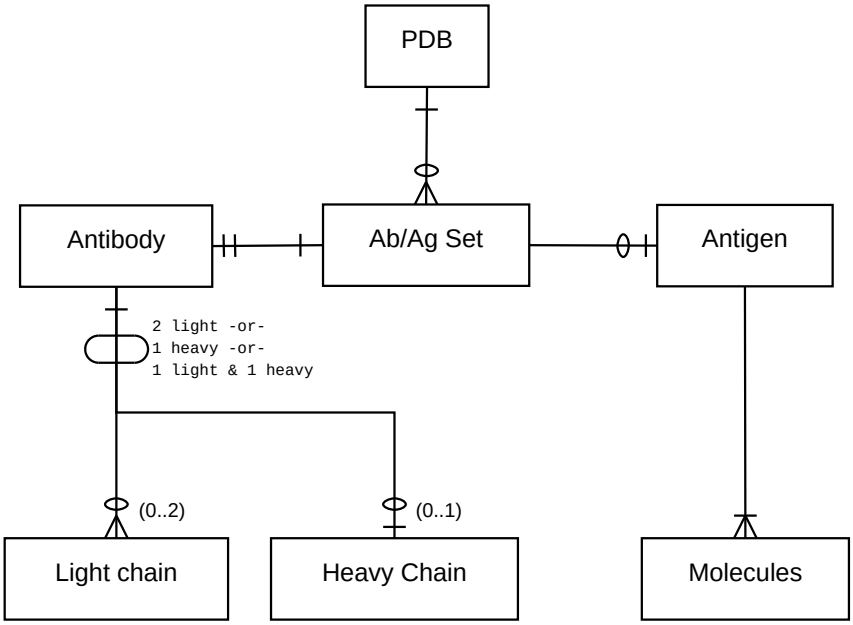


Figure 3.1: A PDB file may have single or multiple antibody-antigen (Ab/Ag) sets. The notation is in Information Engineering Style: a single line indicates one item; two lines indicates exactly one mandatory item; a single line and a circle represents zero or one items; a crowsfoot with a circle indicates zero or more items; a crowsfoot with a line indicates one or more mandatory items. Each Ab/Ag Set contains exactly one antibody. An antibody may be 'complete' (containing one light and one heavy chain), or may be a light-chain dimer or heavy-chain monomer. Each Ab/Ag Set also contains zero or one antigens which contains one or more molecules or chains. Note that the same antigen may take part in more than one Ab/Ag Set.

body binds may span more than one chain. A single PDB file may contain multiple copies of the same antibody or multiple antibodies bound to different parts of the same antigen. Figure 3.1 provides an entity-relationship (ER) diagram describing these scenarios.

Antigens having multiple chains has proved a particular challenge for the analysis of discontinuous epitope that span multiple antigen chains and some studies have simply excluded such examples from the analysis [100]. Another problem with antibody structure related resources is the incorrect identification of 'antibody-binding proteins' as antigens. As a result, epitopes and paratopes, such as those identified by IMGT/3Dstructure-DB, can be incorrect. For example, PDB file 1DEE [101], shown in Figure 3.2, contains a complex of the *Fab* fragment of IgM

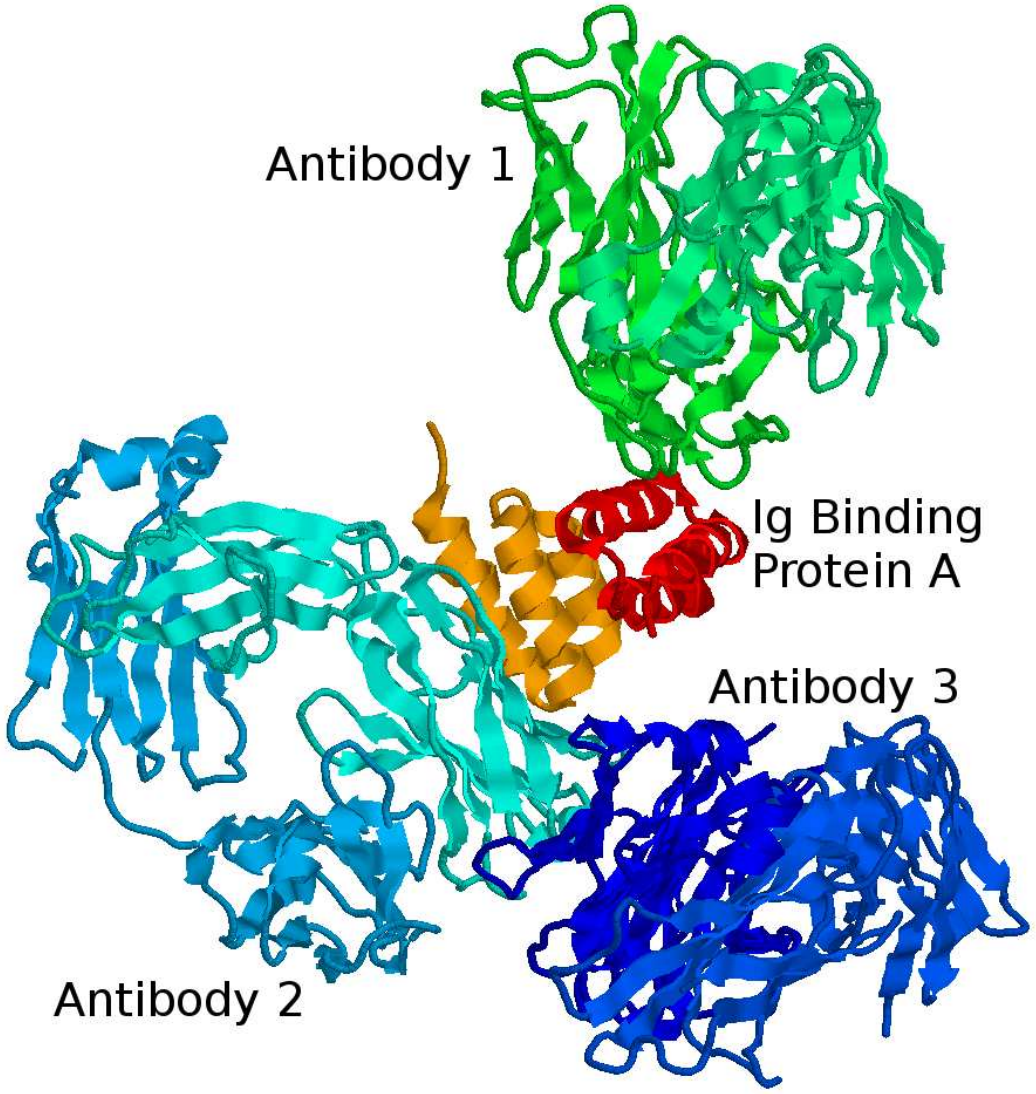


Figure 3.2: The structure of *Staphylococcus aureus* protein A domain D (an immunoglobulin binding protein) complexed with the *Fab* fragment of a human IgM antibody (PDB: 1DEE).

antibody 2A2 and Ig-binding domain D of the *Staphylococcus aureus* virulence factor, protein A (SpA). SpA binds to framework residues of the V_H domain that are highly conserved in human V_H3 derived domains; the CDRs are not involved. Both IMGT/3Dstructure-DB and SAbDab erroneously indicate that this antibody is in a complex with antigen. In addition, antibodies themselves can act as antigens; other antibodies can bind to the framework, or so-called anti-idiotypic antibodies can bind the CDRs. When an antibody binds to the framework there is a clear differentiation

between antibody and antigen. In the case of anti-idiotypic complexes, without additional information, it is impossible to determine which is truly the antibody and which is the antigen.

The automatically maintained resource described in this chapter addresses these and other problems. In particular, the desire was for a resource that would:

- Correctly identify antibody-binding proteins and not treat them as antigens,
- Provide downloadable isolated *Fv* fragments discarding constant domains,
- Provide multiple numbering schemes (Kabat [102], Chothia [103] and Martin [20]),
- Split PDB files containing multiple antibodies into separate files, each containing antigen chains as appropriate. For example, PDB file 4FQR contains 12 copies of the same antibody and these need to be split into separate files; in 3ULU, three different antibodies are bound to the same antigen chain so this must be replicated between the three files,
- Correctly handle cases where the antibody binds multiple chains of an antigen (e.g. 2FEE, 3PJS, 3ZTN),
- Deal with instances of anti-idiotypic antibodies where each antibody is treated both as antibody and as antigen,
- Provide information on redundancy (representative antibodies and lists of non-redundant antibodies and redundant sets),
- Provide information on antibodies that are available both bound and unbound,

- Provide separate datasets for antibody-antigen complexes where the antigen is a protein and where it is a non-protein (e.g. PDB code 4M7J binds carbohydrate, 1A0Q and 1BAF bind hapten, 3VW3 binds DNA, and 2R8S binds RNA),

- Correctly handle light-chain-only antibodies (such as Bence-Jones dimers) and heavy-chain-only (VHH) antibodies,

- Provide a dataset that is as clean and robust as possible at the possible expense of excluding a small percentage of the available structures,

- Provide a web interface that can be used to search by PDB code, antibody or antigen name or species, but most importantly allows the download of complete datasets of free antibodies, antibodies complexed with protein antigens, or antibodies complexed with non-protein antigens, as well as lists of redundant antibodies and complexed/uncomplexed pairs. In addition, it was desired to separate 'complete' antibodies (i.e. paired light and heavy chains), light-chain dimers and heavy-chain only structures.

None of the other available resources deals correctly with all these problems.

3.2 Database Construction and Data Description

3.2.1 Data Processing Pipeline

PDB data are mirrored locally and automatically uncompressed using `ftpmirror` which may be obtained from `github.com/AndrewCRMartin/bioscripts`. Identification of antibody structures in the PDB is not trivial because they may be described as 'antibody' or 'immunoglobulin' and both keywords may be used in other contexts;

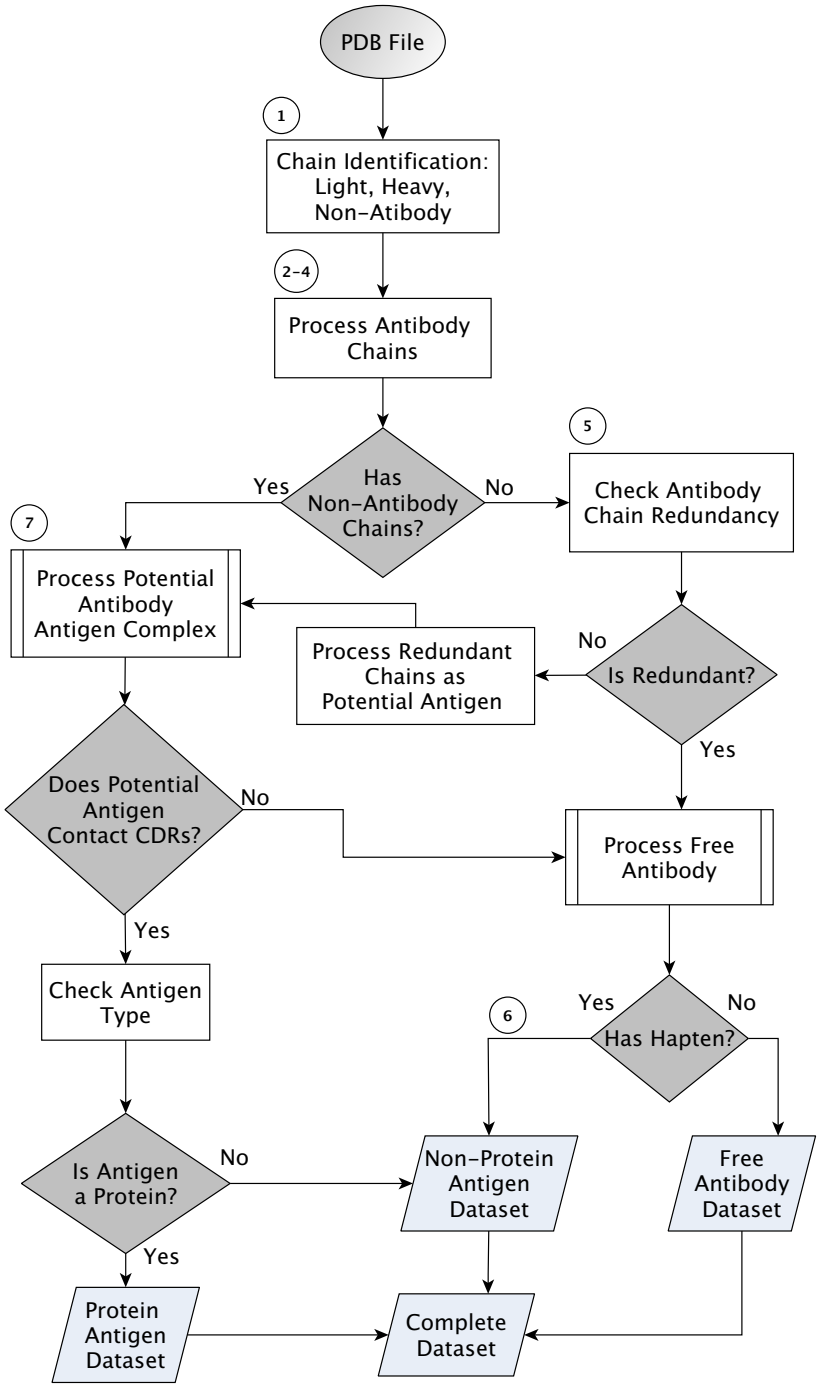


Figure 3.3: Data processing algorithm outline. The circled numbers refer to steps in the text.

without further precautions, simple homology searches will also identify related molecules such as T-cell receptors. Identification of antibody PDB files is handled by an enhanced version of the SACS procedure described previously [95]. SACS data were obtained as an XML file from `www.bioinf.org.uk/abs/sacs/` which is parsed to obtain a list of PDB codes.

Figure 3.3 shows a work flow diagram for which the major processing steps are explained in detail below.

Step 1: Identify Chain Types. Each PDB file is processed to identify the chain types present: light, heavy or antigen. An in-house program, `idabchain` aligns the sequence of each PDB chain with consensus light and heavy chain variable domain sequences. Each chain is provisionally assigned as light or heavy depending which scores higher. The highest score against the light and heavy chain consensus sequences is also recorded. Each chain provisionally identified as light (heavy) is then set to antigen if it does not have the highest score for a match against the light (heavy) chain sequence and its score is less than 80%. If SEQRES records are available, then the sequence is read from both the SEQRES and ATOM records and the higher score is selected — this deals with missing residues in ATOM records and also cases where the SEQRES records contain the leader sequence.

Step 2: Assign Antibody Type. A decision is then made as to whether the antibody is 'complete' (containing both light and heavy chains), light-chain only or heavy-chain only. Complete antibodies and antibodies containing a single chain type are then processed differently to deal with correct chain pairing. In particular, for some cases of light-chain only antibodies, the non-crystallographic symmetry of the two light chains means that only one chain appears in the original PDB file. `pdbsymm` from BiopTools [104] exploits the `REMARK 350 BIOMT` records to

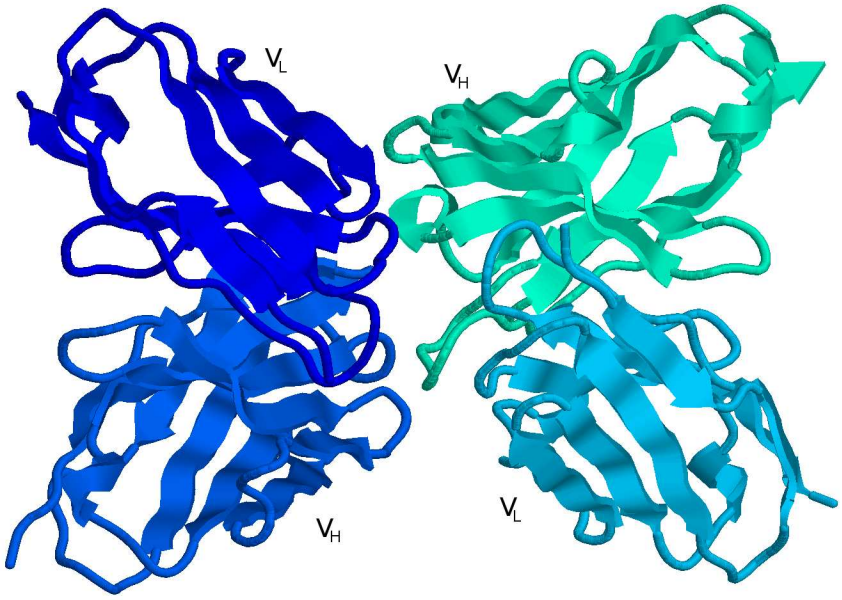


Figure 3.4: An example of an anti-idiotype antibody where one antibody is interacting with the CDRs of another antibody. The processing pipeline generates two structures with each antibody treated as antibody and antigen respectively. (PDB: 1DVF)

construct a new PDB file containing the two copies of the light chain.

Step 3: Splitting and Numbering. In a PDB file, peptide, protein, DNA, and RNA antigens are present as ATOM records with distinct chain labels whereas lipid, carbohydrate and hapten antigens are represented by HETATM records and may have the same chain label as one of the protein chains. The file is then split such that each chain (as defined by the chain label in the PDB file) is placed in its own file and associated HETATMs are removed to be handled later. The files containing the antibody chains are then numbered according to each of the three numbering schemes (Kabat, Chothia and Martin) using `Abnum` [20] and constant domains are discarded.

Step 4: Antibody Reassembly. Most PDB files containing 'complete' antibodies have either duplets of light and heavy chains, or triplets of light, heavy and antigen chains which appear in that order (i.e. light, heavy or light, heavy, antigen). However, this is by no means always the case. Even the presence of L and

H chain labels does not always indicate a pair of light and heavy chains forming an antibody. It is therefore necessary to determine the correct pairs of light and heavy chains from the chain type assignments in the split chains. For heavy-chain-only antibodies, no antibody chain pairing needs to be performed. Light-chain-only antibodies require the two light chains (generated by symmetry if necessary) to be paired and this is done by identifying chains having the Cα of residue L36 in the two chains within 20Å of one another and the Cα of the two L87 residues also within 20Å. For 'complete' antibodies, chains are paired where there is a maximal number of contacts defined as having atom centres within 4Å. For example, PDB file 1DVF (Figure 3.4) is an anti-idiotypic antibody with light chains A,C and heavy chains B,D. Light chain A makes 128 contacts with heavy chain B and 20 contacts with heavy chain D. Light chain C makes 10 contacts with heavy chain B and 117 with heavy chain D. Consequently one antibody is formed from chains A,B while the other is formed from chains C,D.

Step 5: Assigning Antibody/Antibody Complexes. Initially antibodies in PDB files containing no non-antibody chains (i.e. PDB files containing only light and heavy chains) are assigned as free antibodies. However, if the PDB file contains more than one antibody it is possible that one is acting as an antigen for the other. If the sequences of the two antibodies are identical then this will not be the case, but if the sequences are different then they will be forming antibody-antigen pairs (e.g. 1DVF, Figure 3.4). Thus each antibody in turn is reassigned as being a potential 2-chain antigen and processed as such in the next steps. All such antibodies are placed in the protein antigen set.

Step 6: Assigning Free/Complexed Antibodies for HETATM Complexes. If an antibody is currently assigned as uncomplexed, a check is now made for CDR

contacts with HETATM non-protein antigens including haptens, lipids and carbohydrates which were previously removed from the PDB files. These groups are assigned as antigen if there are inter-atom contacts of less than 4Å between any pairs of atom centres, but there are no CONECT records indicating that the HETATM group is covalently bound to the antibody. All such antibody complexes are placed in the non-protein antigen set.

Step 7: Assigning Free/Complexed Antibodies for Protein and Nucleotide Antigens. Antibodies from PDB files containing non-antibody chains (protein, DNA or RNA) are initially assumed to be complexes. However, the correct antigen chain must be identified and, in some cases, while the non-antibody protein is indeed contacting the antibody, it is not interacting with the CDRs of the antibody and should therefore be classified as an antibody-binding protein rather than as an antigen as seen in Figure 3.2. A non-antibody chain is identified as antigen based on the following conditions: (i) more contacts are made with CDRs than with framework; (ii) at least 15 contacts are made with the CDRs. Again, contacts are defined as a distance of ≤ 4Å between any pair of atom centres. Antibodies can bind across antigens which have multiple chains. For example, in 4XI5, the antibody binds to chains L and H from the envelope glycoprotein of *Varicella zoster* virus [105].

Step 8: Annotating the PDB files. The bulk of the standard PDB header is removed and replaced by customised REMARK 950 records (not used by the PDB). During numbering of the antibody chains, the chain labels are replaced by L and H as appropriate. A REMARK 950 record is used to indicate the mapping of the L and H chains to their original chain labels in the source PDB file. Antigen chains will retain their original chain label unless an antigen had the original label L or H, in which case it will be replaced by lowercase l or h respectively. Other key

annotations provided in the REMARK 950 records include: the numbering scheme applied to the antibody structure; the method by which the structure was solved; the resolution, R-factor, and R-Free where appropriate; the antigen and antibody molecule name and species.

Antibodies binding to DNA or RNA antigen chains are added to the non-protein antigen set while others are placed in the protein antigen set. Note that some of the PDB files contain both free and complexed antibodies. The above procedure handles this separating antibody-antigen complexes and free antibodies into their respective datasets. At this stage the pipeline has segregated the data into three antibody types: (i) complete antibodies (heavy and light chains), (ii) light-chain-only antibodies, (iii) heavy-chain-only antibodies. Each of these types is sub-divided into three complex classes: (i) free antibodies, (ii) complexes with protein and (iii) complexes with non-protein (DNA, RNA, hapten, lipid or carbohydrate). Each of these nine datasets is numbered with three numbering schemes making 27 separate datasets. In addition, another 9 datasets are generated for the three antibody types and three numbering schemes, but each set containing all three complex classes.

Thus 36 datasets are made available, but at this stage many of the antibodies may be 'redundant' — i.e. the same antibody appears in several files (originating from either the same PDB file or different files) and may be present both complexed and uncomplexed.

3.2.2 Redundancy Processing

In order to provide information about redundant antibodies, the ATOM records of the numbered antibody structures are used to extract the sequence. Each antibody

pair (both light and heavy chains) is compared on the basis of the residue labels that are present in both sequences. For example, if residue L24 is present in both sequences and the amino acid is different, then the two antibodies would not be regarded as redundant. If residues are missing in one antibody compared with the other these positions are ignored in the comparison. This process is repeated across all pairs of antibodies in order to identify redundant clusters.

In order to select a representative from each cluster, if there were differences in lengths, then the shorter sequences are discarded. From the remaining structures, the highest resolution structure is selected.

The non-redundancy processing is performed across each of the 36 (redundant) datasets described previously such that a non-redundant dataset is produced for each of the redundant sets.

In addition, lists are generated indicating the redundant clusters. 12 lists are provided for each of the three antibody types (complete, light-chain-only and heavy-chain-only) and each of the four complex classes: (free, protein, non-protein, all). In addition, three more lists are provided for antibodies available both free and complexed (for each antibody type: complete, light-chain-only and heavy-chain-only).

3.2.3 Implementation

The system is completely automatic and is implemented in Perl, C and Bash. The main data processing algorithm is implemented in Perl. It extracts PDB codes from the SACS XML file, does all the processing (including calling out to `Abnum` [20] to perform the numbering) and generates separate directories for each dataset of antibody structures. In-house programs to identify antibody chains (`idabchain`) [106] and haptens (`hashapten`) [107] both written by Dr. Andrew Martin and the

a) b)

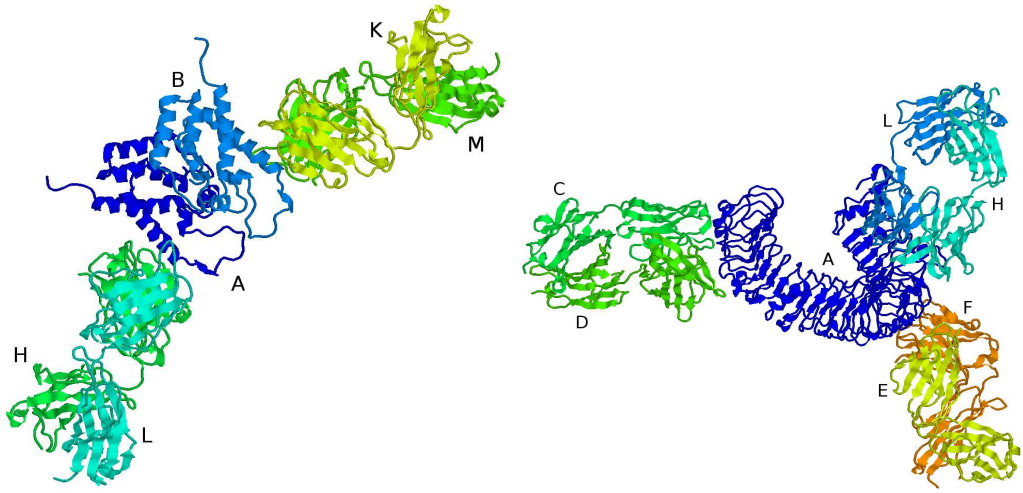


Figure 3.5: a) PDB file 1AFV containing two copies of the same antibody (H/L and K/M) complexed with two copies of the same antigen (A and B), b) PDB file 3ULU containing three different antibodies (C/D, E/F and L/H) interacting with different parts of the same antigen (A).

Table 3.1: Non redundant antibodies in PDB 3ULU along with other redundant antibody strucures.

Query PDB	Query PDB status	Redundant PDB	Redundant PDB status
3ULU_1	Protein complex	3ULV_1	Protein complex
		3NA9_1	Non-protein complex
3ULU_2	Protein complex	3ULV_2	Protein complex
		3ULS_1	Free Antibody
		3ULS_2	Free Antibody
3ULU_3	Protein complex	3ULV_3	Protein complex
		3QPQ_1	Non-protein complex
		3QPQ_2	Non-protein complex
		3QPQ_3	Non-protein complex
		3QPQ_4	Non-protein complex

`Abnum` numbering program are written in C. The web page is generated automatically to provide access to the data. Search options are implemented using Perl CGI scripts.

3.3 Results and Discussion

3.3.1 The Web Interface

The data are available via a web interface at `www.bioinf.org.uk/abs/abdb/`

3.3.1.1 Database Searching

The resource may be searched by PDB code, keyword (antibody or antigen name) and species (of antibody or antigen). The results are presented as a table with all the antibodies that match the search query along with any antibodies that are redundant with these antibodies. The information in the table includes a downloadable antibody structure file for each of the three numbering schemes (Kabat, Chothia and Martin), the antibody status (free or complexed), resolution and R-factor.

PDB files containing antibody-antigen complexes can contain multiple antibody structures (which may or may not be identical sequences) and the main processing pipeline will have split these into separate files. For example, 1AFV contains two copies of the same antibody, FAB25.3, bound to two copies of the human immunodeficiency virus type 1 capsid protein (Figure 3.5a). The main processing pipeline identifies chains L and H as one antibody, chains K and M as a second antibody and identifies chain A as the antigen associated with the first antibody and chain B as associated with the second antibody. The *Fv* region of the first antibody (chains L,H) together with its cognate antigen (chain A) is stored as as 1AFV_1 while the *Fv* of the second antibody and its antigen (chains K,M,B) is stored as 1AFV_2. While 1AFV only contains 'internal redundancy' (i.e. there are two copies of FAB25.3 in the original PDB file, but no other files contain a structure for FAB25.3), some antibodies are present in multiple PDB files. For example, a search for PDB 4KKC will display the two copies (4KKC_1 and 4KKC_2) that are

present in that PDB file together with an additional 70 PDB entries having the same antibody sequence.

On the other hand, 3ULU contains three different antibodies (FAB15, FAB12 and FAB1068) bound to different parts of the same antigen (human toll-like receptor 3 — chain A in the PDB file) as shown in Figure 3.5b. Thus the main processing pipeline generates three files, each containing a copy of the antigen and one of the antibodies: 3ULU_1 (containing chains L, H and A), 3ULU_2 (chains C, D and A) and 3ULU_3 (chains E, F and A). In such cases, redundant antibody information is provided for each of the distinct antibodies (Table 3.1).

In addition to the search by PDB code, the database can be queried by a keyword from the antibody or antigen name. For example, all the structures for the HyHEL antibodies (anti-Hen egg white lysozyme) from the Smith-Gill group [108] can be queried using 'hyhel'. Similarly all the capsid antigen bound antibody structures can be queried by searching for 'capsid'. Data can also be searched by antibody and antigen species and a drop-down menu has been provided with all the species of antibody and antigen observed in the dataset.

3.3.1.2 Data Download

In addition to individual PDB files, a compressed archive of each of the 72 antibody structures datasets (the 36 redundant and 36 non-redundant sets described above) can be downloaded. The 15 lists of redundancy information can also be downloaded.

3.3.2 Database Statistics

A list of 2402 antibody structure PDB codes was obtained from SACS. Of these, 2014 PDB files were successfully processed to generate 3376 PDB files: 2938 complete antibodies, 171 light-chain-only and 267 heavy-chain-only. A detailed break-

Table 3.2: Contents of the AbDb datasets, June 2017.

Datasets	**Complex Type**	**Processed PDB Files**	**Resultant Antibodies**	**Non-Redundant Antibodies**
Complete Antibody	Protein	976	1591	673
	Non-protein	275	374	194
	Free Antibody	580	973	531
	Complete Dataset	1794	2938	1184
Light Chains	Protein	12	17	5
	Non-protein	9	17	5
	Light Only	77	137	48
	Complete Dataset	86	171	52
Heavy Chains	Protein	88	162	74
	Non-protein	5	11	5
	Heavy Only	47	94	51
	Complete Dataset	134	267	121

down of the content of the AbDb database as of June 2017 is shown in Table 3.2.

The SACS database which provides lists of antibodies to be processed by AbDb is a cumulative resource (listing PDB files that have now been obsoleted) that also contains information on structures of antibody *Fc* fragments. Consequently there are a number of files that are not processed by AbDb: 111 *Fc* fragments and 37 obsoleted. In addition, the numbering procedure fails for at least one of the chains in 190 PDB files and 50 files are identified as containing single chain *Fv*s (*scFv*s) with a single chain label. *scFv*s are fused V_H and V_L domains with a peptide linker [109]. Frequently the peptide linker is flexible and is not visible in the PDB file and, in these cases, while part of a single chain, the PDB file often has different chain labels for the V_H and V_L regions. Such examples are processed correctly by AbDb. However, in cases where the two regions have been given the same chain label, the `Abnum` program is not able to pick apart the two regions and therefore does not number the protein correctly. Consequently these files are currently automatically rejected.

3.4 Conclusions

AbDb provides a regularly updated resource with processed antibody structures, not provided by other antibody structure resources. In particular, it provides:

- Processed PDB files containing only the variable domains and split into individual antibodies with cognate antigens (including multi-chain and non-protein antigens),
- PDB files numbered with Kabat, Chothia, and Martin numbering schemes,
- 36 simply classified downloadable datasets: complete antibodies, light-chain-only and heavy-chain-only also split into free antibodies, complexes with protein antigens and complexes with non-protein antigens (all numbered with all three numbering schemes),
- Non-redundant versions of each of the 36 datasets,
- 12 information files describing redundant clusters,
- Information on redundancy when searching by PDB code: a list of all processed PDB files containing redundant antibodies is provided,
- Three information files describing antibodies available both complexed and uncomplexed,
- Non-antibody proteins, not interacting predominantly with the CDRs, are not treated as antigen and the antibody is classified as uncomplexed (e.g. 1DEE),
- Antibodies involved in anti-idiotype interactions are classified both as antibodies and antigens,
- Light-chain dimers regenerated by exploiting non-crystallographic symmetry.

Unfortunately it is almost impossible to predict every variant of antibody structures that may become available in the future. Consequently AbDb is continuously being updated to handle unforeseen circumstances as they are encountered. The desire is to make the dataset clean and robust at the possible expense of missing some of the available structures. Future planned enhancements include improved processing of *scFv*s and single chain *Fab*s (*scFab*s) where a single chain label is used for light and heavy chain segments in the PDB file. Currently, the *idabchain* [106] program which identifies chain types identifies these as hybrid chains, but the AbDb pipeline does not process them further owing to current restrictions in the numbering program *Abnum*. Improvements are needed to *Abnum* to deal correctly with these examples. While there are, as yet, no structures, antibody drugs have been created having *scFv*s fused to the C-terminus of a conventional heavy chain; AbDb would currently ignore the *scFv* fusion and a general method is needed for dealing with fusions. *Abnum* has been designed to fail cleanly with unusual structures rather than provide incorrect numbering, but enhancements to its ability to number these unusual cases are planned. Similarly, the non-crystallographic symmetries have only been exploited only for light-chain-only antibodies but, in future will be expanded to handle all structures.

Chapter 4

Structural Analysis of B-Cell Epitopes

As described in the introduction, peptide vaccines have many potential advantages over conventional ones. However, it is well known that approximately 90% of B-cell Epitopes (BCEs) are discontinuous in nature making it difficult to mimic them for creating vaccines. In this chapter, the degree of discontinuity in B-cell epitopes and their conformational nature is examined. The discontinuity of B-cell epitopes is analyzed by defining extended 'regions' (R, consisting of at least 3 antibody-contacting residues each separated by ≤ 3 residues) and small fragments (F, antibody-contacting residues that do not satisfy the requirements for a region). Secondly, a novel algorithm has been developed that classifies each region's shape as extended, curved or folded. In addition, the conformation of each of the epitopes was studied by investigating the shape and secondary structure composition. These analyses and classifications were designed to give insight into the probability of an isolated peptide being part of an epitope, as well as the conformational stability of isolated peptides and the possibility of peptide-based vaccine design.

As stated in the introduction, unlike T-cell epitopes that are linear continuous

residues, B-cell epitopes tend to be conformational (discontinuous) comprised of multiple sequential segments that are in close spatial proximity in 3D folding of an antigen. This discontinuous nature of B-cell epitopes has made identification and prediction challenging [27, 34–40]. The increase in structural data available for antibody/antigen complexes has provided new opportunities for conformational analysis and characterization of epitopes to understand their properties in detail. Thus far, structural characterization of epitopes has been performed on the basis of solvent accessibility [110, 111], amino acid composition, size [27, 112–115], secondary structure [112, 113, 116], location on the antigen [27, 113, 117] and geometry [113]. A recent study by Kringelum et al. [118] presents a detailed analysis of antigen-antibody interaction surfaces and described the epitope in terms of its size, shape, segmentation, secondary structure, location, orientation relative to the antibody, amino acid composition, amino acid 'co-operativeness' (particular amino acid pairs mediating cooperative antibody-antigen binding) and spatial amino acid composition. This analysis was performed on a relatively large dataset (107 unique antibody-antigen complex structures) compared with previous studies which used smaller datasets (up to 53 unique antibody-antigen complex structures [100, 113]). In terms of shape, Kringelum et al. described B-cell epitopes as flat, oblong or oval based on an analysis of epitope and paratope residues.

A detailed analysis of 3D structures of discontinuous epitope is needed to understand both the level of discontinuity and the conformational nature of the continuous stretches of a discontinuous epitope. This analysis relies on a clean, processed dataset of antibody-antigen complexes as descriped in Chapter 3.

4.1 Dataset Preparation

A non-redundant dataset of 673 unique antibody-antigen structures was obtained from AbDb (www.bioinf.org.uk/abs/abdb) in December 2016 (see Chapter 3). To ensure the presence of protein antigens and not peptide antigens, the dataset was filtered on the basis of antigen length. A threshold of 30 residues was used to define a protein-antigen. 153 peptide-antigen complexes were identified and excluded reducing the dataset to 520 unique antibody-protein antigen complexes. A further 11 antibody-antigen complexes solved using electron diffraction were excluded because of their low resolution, leaving 509 structures. A further 3 complexes were removed owing to incorrect pairing resulting from single chain Fabs in AbDb, missing structural information and incorrect symmetry. Of the remaining 506 complexes, 464 had an antibody bound to a single chain while 42 interfaces spanned multiple antigen chains.

4.2 Defining Epitopes

4.2.1 Epitope Residue Mapping

The structures of antibody-antigen complexes were used to define the epitopes as the set of antigen residues having any atoms in contact with the CDR region of an antibody (Figure 4.1). A contact was defined as a centre-to-centre distance less than 4 Å. Other studies have used similar criteria to map epitopes [27,115,119]. An in-house C program, *chaincontacts* [120] was used to compute inter-chain atomic contacts which provided the list of antigen residues comprising the epitopes and the antibody residues comprising the paratope.

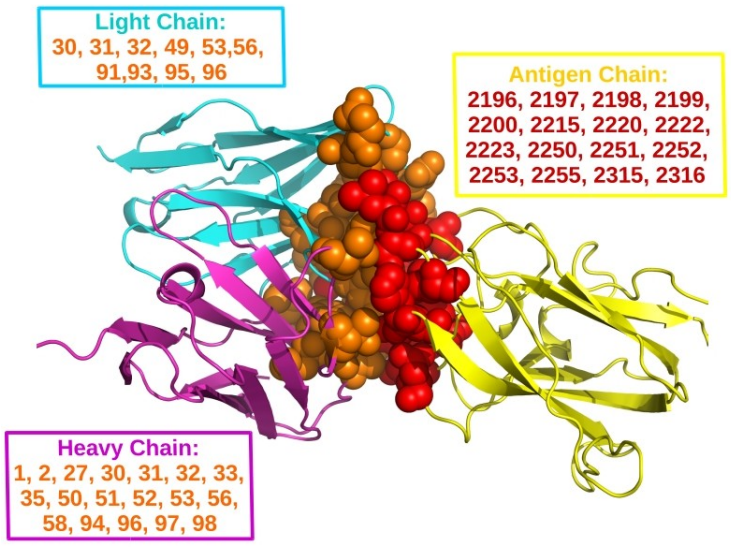


Figure 4.1: Epitope Mapping — Epitope residues (red balls) are shown on the surface of an antigen. The paratope residues (gold balls) on the surface of light and heavy chains are shown providing a binding site for the antigen.

4.2.2 Epitope Structural Discontinuity Determination

The epitopes were separated into two different types of structural elements: regions (R) and fragments (F). Regions were defined as continuous stretches of antigen sequence having at least three residues in contact with antibody with gaps of up to three residues between any pair of contacting residues. Fragments were defined as individual antigen residues that contact the antibody, but do not form part of a region (Figure 4.3). Sivalingam and Shepherd [100] investigated discontinuous epitopes defining regions with no gaps, gaps of three and of five non-contacting residues. Their findings suggest that with the gap of three or five, 85-88% epitopes are comprised of multiple regions.

In this study, a gap of up to 3 non-contacting (non-epitope) residues was chosen on the basis of the structure of alpha helices. The gap of up to three residues allows the inclusion of amino acids which lie on the same face of an α-helix (shown in yellow in Figure 4.2). Code was implemented in Perl to perform this classification. There are several studies that used the idea of including non-epitope residues

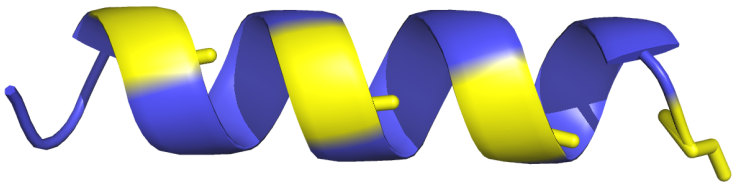


Figure 4.2: A gap of up to three residue in the epitope region. Amino acid residues which lie on the same face of an α-helix are shown in yellow with a spacing of 3 between them.

and obtaining segments of reasonable length to perform analysis [27, 100, 115, 118]. This helps in overcoming the problem of having too many shorter segments and provides a way to characterise the epitope in terms of having multiple small regions/peptides. Moreover, for the purpose of peptide studies, it is not feasible to investigate several very short and non-continuous segments. In short, the allowance of gap/non-epitope residues provides a linear and continuous stretch of residues that can be used effectively for analysis. In the past, different studies have used the terms, 'segment' or 'fragment' for stretches of residues with gaps [100, 113, 118]. In this study, they will be referred to as 'regions' whereas the term 'fragment' will be used for single amino acids that are making contact with antibody, but which do not form part of a region.

4.3 Epitope Analysis - Regions and Fragments

Epitopes were analyzed in terms of the number of regions, number of fragments, region length, longest region length, probability of having other regions if an epitope has a region of a certain length, the relationship of region length with either the number of regions or the number of fragments, epitope size, shape/conformation and secondary structure of all the regions in the epitope dataset. These analyses were performed to assess the overall composition and structural features that characterize an epitope. The size of an epitope was calculated by summing the number

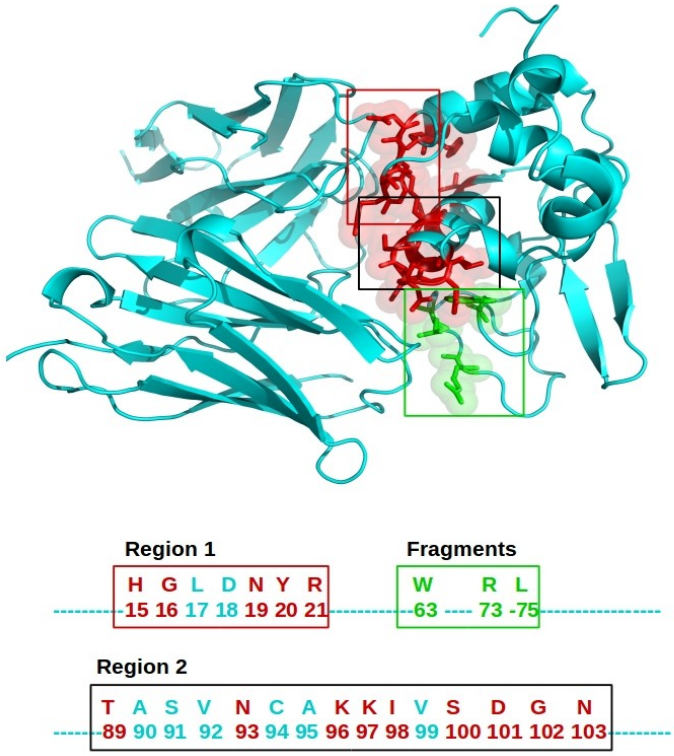


Figure 4.3: Regions and Fragments - The epitope is comprised of two regions and three fragments. Region 1 has a gap of two non-contacting residues (shown in blue) whereas Region 2 has three gaps of up to three non-contacting residues. The contacting residues in the epitope are shown in red.

Table 4.1: Number of regions and fragments in epitopes

	Epitopes	**Regions**	**Fragments**
Single chain	464	1195	919
Multiple chain	42	134	139
Combined	506	1329	1058

of residues that make up the regions and fragments; consequently it may be larger than the number of contacting residues since regions include residues between the contacting residues.

In the dataset of 506 epitopes, a total of 1329 regions and 1058 fragments were observed. Table 4.1 shows the number of regions and fragments in each of the datasets.

Table 4.2: Distribution of regions (R) and fragments (F) in complete epitope dataset.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**Total**
F0	23	45	35	21	1	0	1	0	0	126
F1	13	34	24	19	7	0	0	0	0	97
F2	10	39	41	9	9	1	0	0	0	109
F3	7	25	22	10	3	1	0	0	1	69
F4	9	24	15	6	1	0	0	0	0	55
F5	5	11	4	3	3	1	0	0	0	27
F6	3	4	2	2	0	0	0	0	0	11
F7	0	2	0	1	0	0	0	0	0	3
F8	0	1	0	0	1	0	0	0	0	2
F9	1	2	0	0	0	0	0	0	0	3
F10	0	0	0	1	0	0	0	0	0	1
F11	0	0	0	1	0	0	0	0	0	1
F12	0	0	0	0	0	0	0	0	0	0
F13	0	0	0	0	0	0	0	0	0	0
F14	0	1	0	0	0	0	0	0	0	1
F15	0	0	0	0	0	0	0	0	0	0
F16	0	1	0	0	0	0	0	0	0	1
Total	71	189	143	73	25	3	1	0	1	506

4.3.1 Distribution of Regions and Fragments

The structures of 506 distinct B-cell epitopes were analysed for the distribution of regions and fragments. Among these, most of the epitopes were part of a single chain antigen while some were mapped onto multiple chains antigen. It was observed that the epitopes were composed of one to nine regions and zero to sixteen fragments. In previous studies, without the concept of fragments (single contacting residues) and with a gap of up to 5, a maximum of 11 continuous stretches were reported [100,113]. In our dataset of epitopes, most frequently, epitopes have (in order of frequency) compositions R2F0 $>$ R3F2 $>$ R2F2 $>$ R3F0$\approx$R2F1. In the epitope dataset, one of the epitopes (from PDB, 1TZI) was found with no regions but just fragments (R0F4), and was excluded from the analysis. Overall, 90% of epitopes have up to 5 regions (R1-R5) and 5 fragments (F0-F5) as shown in the Table 4.2. This agrees with previous analyses of smaller datasets of epitopes [100, 113].

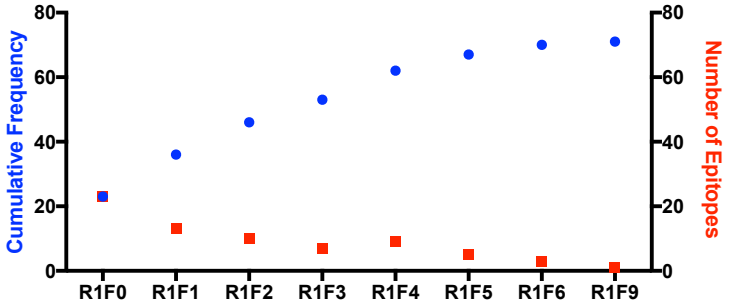


Figure 4.4: Cumulative frequency of epitopes having a single region and different number of fragments.

Using this nomenclature, a linear (non-conformational) B-cell epitope would be R1F0 whereas a conformational B-cell epitope would be composed of more than one region with zero or more fragments. In this study, only 23 epitopes out of 506 were truly linear (R1F0), such that about 95% of epitopes are conformational in this dataset agreeing with several studies that report over 90% of B-cell epitopes are conformational [14, 26, 27]. This suggests that only 5% of epitopes are truly linear so could be mimicked by a simple peptide. However, it is likely that some R1F0-F3 epitopes may also be successfully mimicked.

Of the epitope structures analysed in this study, approximately 14% are comprised of a single region and up to 9 fragments (R1F0-F9). A cumulative plot for all the epitopes with one region is shown in Figure 4.4. Moreover, ~37% of epitopes have 2 regions with up to 16 fragments (R2F0-F14), ~28% of epitopes have 3 regions with up to 6 fragments (R3F0-F6), ~14% are with 4 regions and up to 11 fragments (R4F0-F11) and ~5% of epitopes contain 5 regions with 5 fragments. A very small fraction of the data has over 5 regions and fragments (Table 4.2).

4.3.1.1 Separating Single Chain and Multiple Chain Epitopes

Epitopes can be part of a single chain or multiple chains. The discontinuity in multiple chains may be more diverse and prediction of such epitopes is even more

Table 4.3: Distribution of regions and fragments in single chain dataset

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**Total**
F0	23	44	33	20	1	0	1	0	0	122
F1	13	30	21	18	5	0	0	0	0	87
F2	10	39	40	7	9	1	0	0	0	106
F3	6	23	20	8	2	1	0	0	1	61
F4	9	23	13	5	1	0	0	0	0	51
F5	5	8	2	2	2	1	0	0	0	20
F6	3	4	2	0	0	0	0	0	0	9
F7	0	1	0	1	0	0	0	0	0	2
F8	0	1	0	0	1	0	0	0	0	1
F9	1	2	0	0	0	0	0	0	0	3
F10	0	0	0	0	0	0	0	0	0	0
F11	0	0	0	0	0	0	0	0	0	0
F12	0	0	0	0	0	0	0	0	0	0
F13	0	0	0	0	0	0	0	0	0	0
F14	0	1	0	0	0	0	0	0	0	1
F15	0	0	0	0	0	0	0	0	0	0
F16	0	1	0	0	0	0	0	0	0	1
Total	70	177	131	61	20	3	1	0	1	464

Table 4.4: Distribution of regions and fragments in multiple chain antigen epitope dataset

	R1	**R2**	**R3**	**R4**	**R5**	**Total**
F0	0	1	2	1	0	4
F1	0	4	3	1	2	10
F2	0	0	1	2	0	3
F3	1	2	2	2	1	8
F4	0	1	2	1	0	4
F5	0	3	2	1	1	7
F6	0	0	0	2	0	2
F7	0	1	0	0	0	1
F8	0	0	0	0	1	1
F9	0	0	0	0	0	0
F10	0	0	0	1	0	1
F11	0	0	0	1	0	1
Total	1	12	12	12	5	42

challenging. As described above, In this study, a total of 464 epitopes were mapped to single chain antigens whereas 42 epitopes were part of multiple chains.

In order to check whether the distribution of regions and fragments in both of these datasets is the same, they were investigated separately. Tables 4.3 and Table 4.4 show the data obtained. A conventional chi-squared between single and multiple chain dataset was performed which provided a p-value of 0.0028, suggesting a significant difference in the distribution of the datasets. Furthermore, to know if the single or multiple dataset is more representative of the combined dataset, an unconventional chi-squared test was performed on these data to determine whether the distribution of regions and fragments differs from the combined set. To this end, chi-squared tests were performed on the single and multi chain dataset using expected values calculated from the observed values in the combined dataset. Grouping was performed as necessary to satisfy the requirements of a chi-squared test (no expected < 1 and < 20% less than 5). See Tables A.1, A.2 and A.3 in Appendix A. Equation 4.1 is used to calculate the expecteds for the single (and multiple) chain dataset using observed values from the combined dataset.

$$Exp_{Single} = \frac{Obs_{Com} * Tot_{Single}}{Tot_{Com}} \tag{4.1}$$

A p-value of 0.58 suggests that the single chain dataset is randomly drawn from the combined set while this is not true for multiple chain dataset which has a p-value of 0.0026 suggesting that epitopes in this dataset have a different region/fragment distribution. In particular, single region epitopes are less usual than expected when there are multiple chains and epitopes with two or more regions are more common than expected. Thus each chain tends to contribute at least one region. Both of these chi-squared tests suggest that these 2 datasets should be analysed separately for the

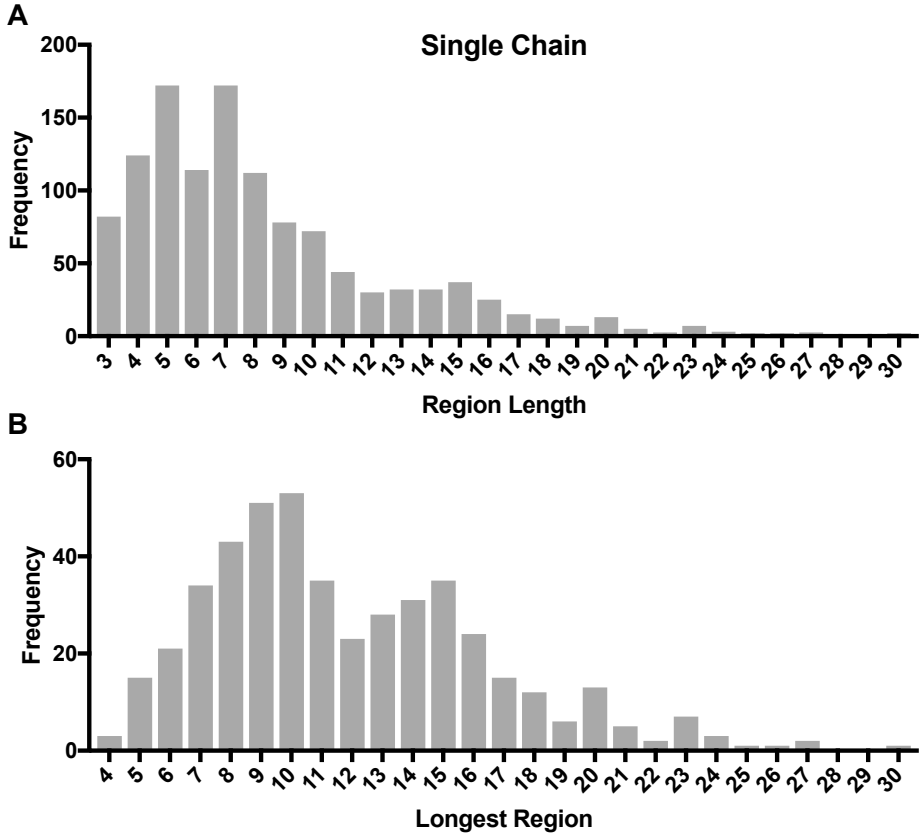


Figure 4.5: A) Distribution of 1195 regions in 464 epitopes (single chain antigen). A region can be as short as 3 residues and as long as 30 residues. B) Distribution of the longest regions in 464 epitopes. The longest region in a given epitope can be in the range of 4 to 30 residues.

rest of the epitope structural analysis.

4.3.2 Length Analysis of Epitopes

4.3.2.1 Length of Regions

Given the limited size of the antigen combining site, it was hypothesized that the length of a region would be inversely correlated with the number of regions, i.e. a longer region would be likely to have fewer other regions. To this end, the length of individual regions was investigated. In the single chain dataset, the region length ranges from three to thirty residues whereas it was found in the range of three to twenty three for the multiple chain dataset. The distribution of these region lengths

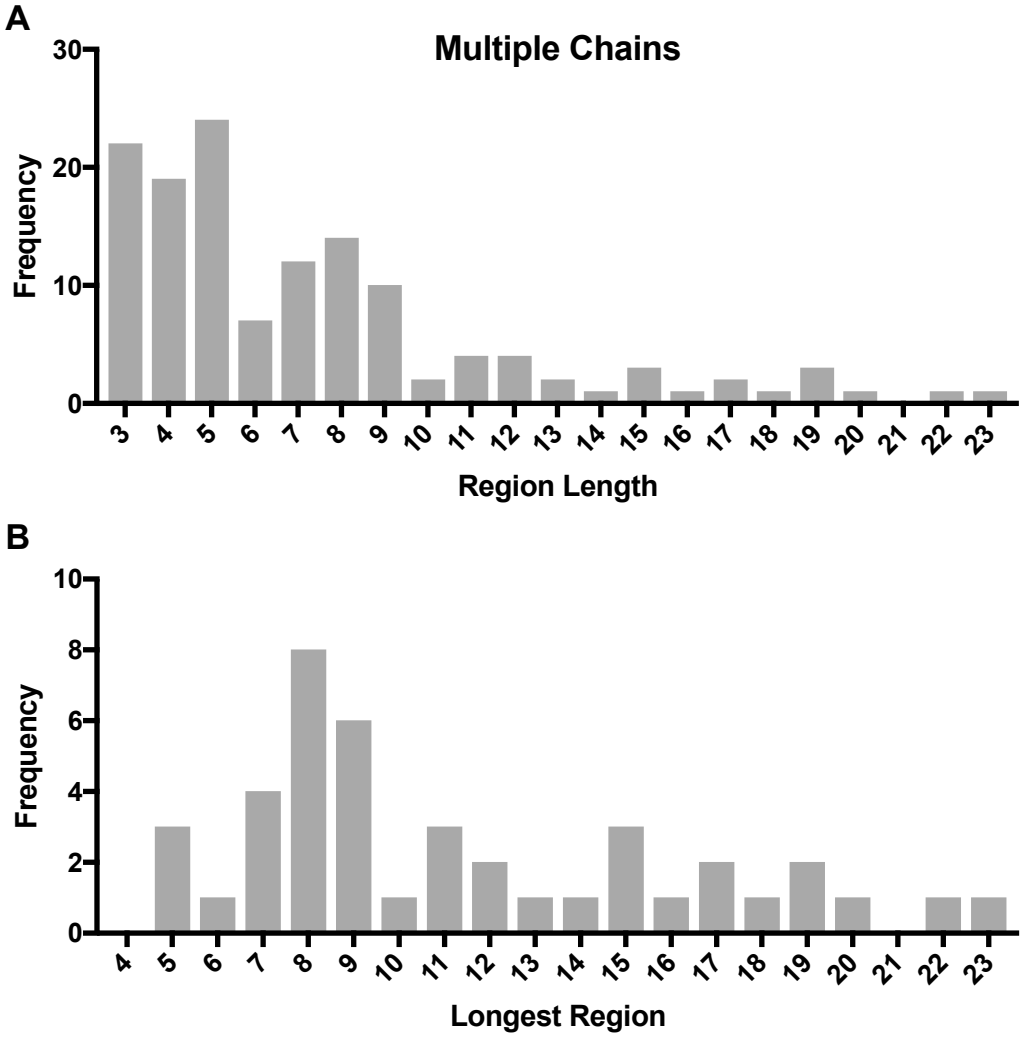


Figure 4.6: A) Distribution of 134 regions in 42 epitopes (multiple chain). A region in multiple chain dataset can be as short as 3 residues and as long as 23 residues. B) Distribution of the longest regions in 42 multi chain epitopes shows that the longest region in this dataset can be in the range of 5 to 23 residues.

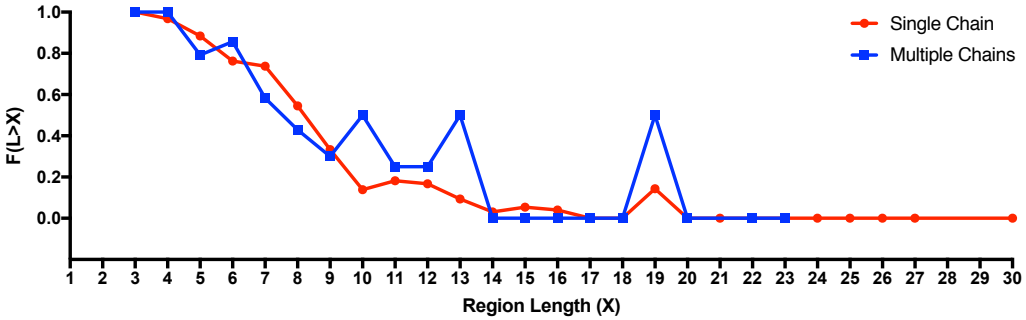


Figure 4.7: Probability of having longer regions than a given region length X.

is shown in Figures 4.5A and 4.6A. 94% of regions are $\leq$ 16 residues long in the single chain dataset. A similar trend was observed by Kringelum et al. where regions of up to 15 residues were seen [118]. In this larger study, while regions of up to 30 amino acid are observed, only about 8% have a length of more than 15 residues. This is likely to be explained by the larger dataset and the fact that a gap of up to 3 non-epitope residues is allowed in our regions compared with only one amino acid in the previous study [118].

The distributions in Figures 4.5 and 4.6 suggest that single chain epitopes tend to have longer regions than multiple chain epitope.

4.3.2.2 Length of the Longest Regions

For epitopes consisting of multiple regions, it is interesting to investigate the length of the longest region in a particular epitope, but previous studies have not done so. The distribution of the longest region was calculated for both datasets. It was found to range from four to thirty residues and from five to twenty three residues in the single and multiple chain datasets respectively (Figures 4.5B and 4.6B). In the single chain dataset, 85% of epitopes were found with the longest region $\leq$ 16 residues whereas the longest regions in the multiple chain dataset appears to be distributed unevenly. However, a chi-squared test on the distribution of both the datasets suggests that the difference is not significant (p-value = 0.48).

4.3.2.3 Probability of a Region Being the Longest

Given the scenario that a region of a certain length is being analysed as a candidate immunogen, it is interesting to know whether there are likely to be other longer regions within the same epitope. In other words, for a given region length, what is the chance of that being the longest region and therefore the major structural component of the epitope? This would allow us to extrapolate the results to epitopes where the antigen structure and only the rough epitope is known (perhaps by scanning mutagenesis). Therefore, to compute the probability, the fraction of epitopes having region length X and also having regions longer than X was calculated as follows and plotted for each possible length of a region in the observed data (Figure 4.7).

$$F(L > X) = \frac{\text{No. of epitopes having } R_X \text{ also having } R_{>X}}{\text{No. of epitopes having } R_X} \tag{4.2}$$

where X is a given region, L is length of region X and R_X represents the region X.

The data show that epitope regions of length 3 or 4 will always be accompained by longer regions. This falls off gradually as region length increases and it becomes statistically unlikely to see longer regions accompanying regions of 13 amino acids or more ($F(L > X)$ falls below 0.05). However there is a peak at length 19 for both single and multiple chains showing that epitopes of this length do tend to be accompained by a longer region. Looking at these examples, it was found that both of the datasets have one such example each (Figure 4.8). In general, however, it can be concluded that if an epitope has a region greater than 13 residues, it is most likely that this is the longest region and are likely to be linear epitopes if only found by scanning mutagenesis.

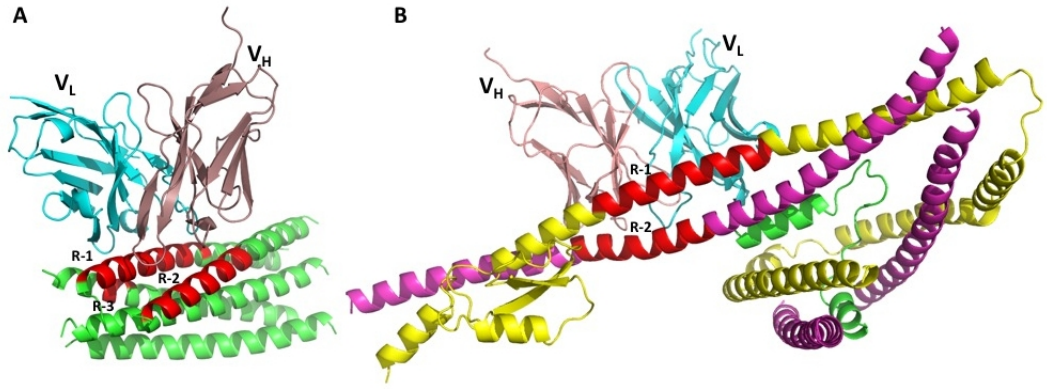


Figure 4.8: Unusal examples in which regions of 19 residues are accompained by longer regions. A) An epitope with three regions; R-1 (19 residues), R-2 (20 residues) and R-3 (4 residues) in single chain dataset from PDB file 3MA9. B) An epitope with two regions; R-1 (23 residues), R-2 (19 residues) from the multiple chain dataset (PDB file 5CWS).

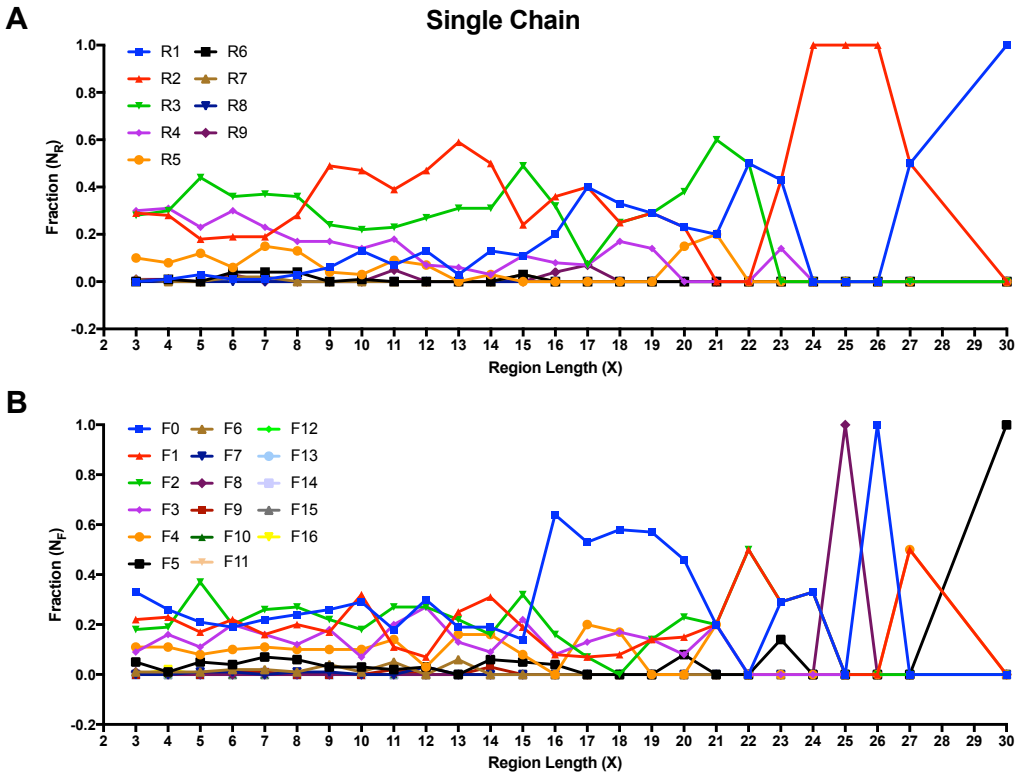


Figure 4.9: A) The fraction of regions ($F(N_R)$), and B) the fraction of fragments ($F(N_F)$), for a given length X, in the single chain dataset. R1-R9 represent the number of regions. F1-F16 represent the number of fragments. The peaks for region length ≥ 24 are an artifact of the very small number of epitopes having such long regions.

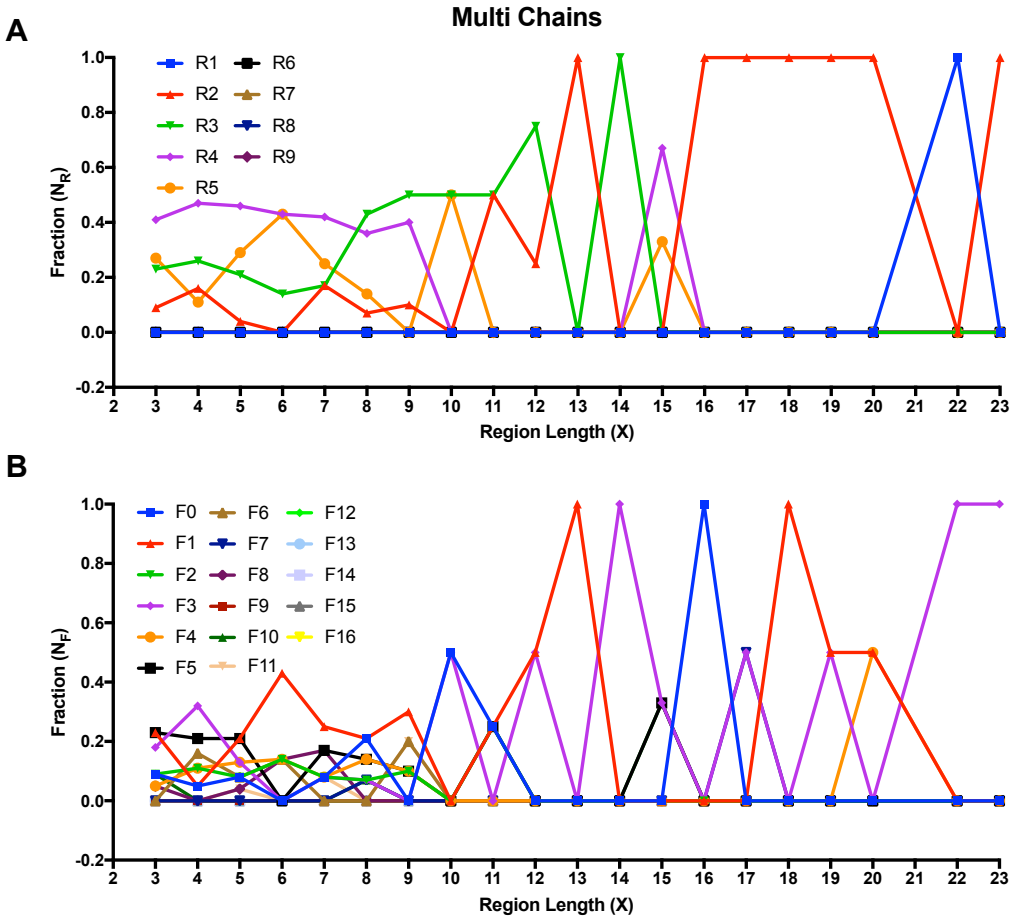


Figure 4.10: A) The fraction of regions (F(N_R)), and B) the fraction of fragments (F(N_F)), for a given length X, in the multiple chain dataset. R1-R9 represent the number of regions, F1-F16 represent the number of fragments. The peaks for region length ≥ 11 are an artifact of the very small number of epitopes having such long regions.

4.3.2.4 Relationship between Region Length and Number of Regions and Fragments

Again, with the aim of identifying regions that are likely to be dominant within epitopes, the correlation between region length and either the number of regions or of fragments was investigated. For any given maximum region length X, the fraction of regions ($F(N_R)$) and fraction of fragments ($F(N_F)$) having a specified length was computed as follows:

$$F(N_R) = \frac{\text{No. of epitopes having } R_X \text{ also having n regions}}{\text{No. of epitopes having } R_X} \tag{4.3}$$

$$F(N_F) = \frac{\text{No. of epitopes having } R_X \text{ also having n fragments}}{\text{No. of epitopes having } R_X} \tag{4.4}$$

where R_X is the given region X, N_R is the number of regions, N_F is the number of fragments and n is 1 to 9 and 0 to 16 for the number of regions and the number of fragments respectively. In the single chain dataset, epitopes having smaller regions tend to have more regions compared with epitopes having longer regions. Epitopes having regions of length between 14 and 23 generally have only one region. Interestingly, for regions of these lengths, a higher fraction of epitopes is observed with no fragments (Figure 4.9). This suggests that linear epitopes (R1F0) mostly lie within this range of region length. Moreover, epitopes having regions with a length of up to 14 residues tend mostly to have 2 or 3 regions and 0 to 4 fragments.

The same analysis for the multiple chain dataset is shown in Figure 4.10. The main difference is having only epitopes with a single region. Most of the epitopes tend to contain 4 regions and up to 5 fragments if they have small regions (3-9 residues long). However, epitopes having regions of length between 8 to 23 tend to have 2 or 3 regions and up to 4 fragments.

4.3.3 Relationship between Regions and Fragments

In general, an epitope with fewer regions may be expected to have more fragments and vice versa. Similarly, the length of regions and number of residues making an epitope might have a relationship with the number of fragments in an epitope. The Pearson correlation coefficient between the number of fragments with either the number of regions in an epitope, the longest region in an epitope, the total

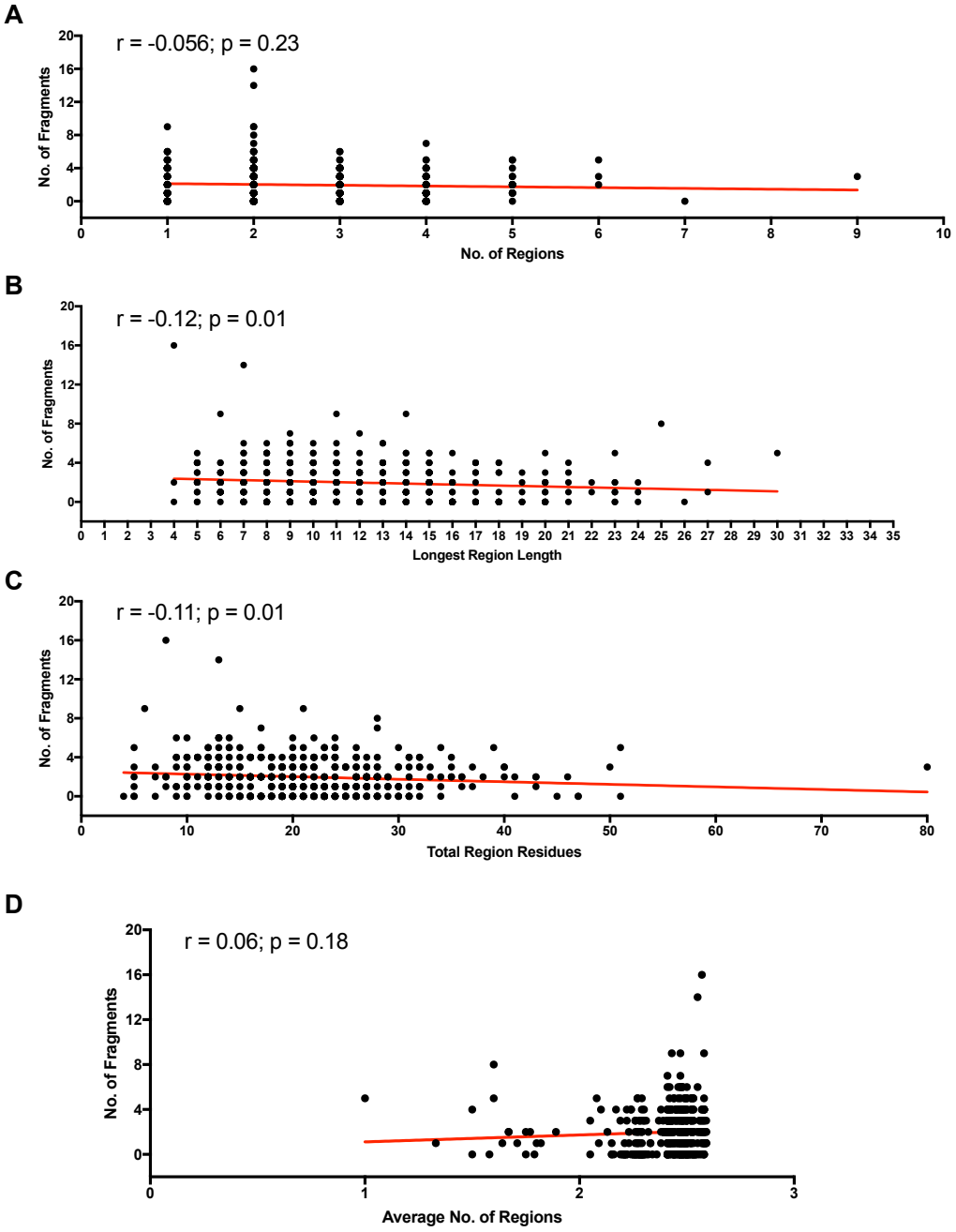


Figure 4.11: Correlation between the number of fragments and A) the number of regions, B) the longest region, C) the number of residues in regions and D) the average number of regions in an epitope in the single chain dataset. No significant correlations were observed.

residues making an epitope and the average number of regions in an epitope was calculated and the data are shown in Figures 4.11 and 4.12 for single and multiple chain datasets respectively. Surprisingly, these data do not provide any evidence for a positive or negative relationship between the number of fragments and the four other variables.

4.4 The Size of Epitopes

The size of an epitope was defined as the total number of residues that constituent regions and fragments. In the single chain dataset, most of the epitopes (77%) are found in the range of 15 to 35 residues. The average size of an epitope in this dataset was found to be $\sim$23 residues. In the multiple chain dataset, most of the epitopes ($\sim$47%) are in the range of 20-30 residues with the smallest epitopes of length 15 and the largest being 45 residues long. The average length of an epitope in this dataset was found to be $\sim$26 residues. Figure 4.13 shows the epitope size distribution. The epitope size in these two datasets was observed to be significantly different (p-value=0.0002, Welch's t-test).

A study conducted by Rubinstein et al. [118] on a dataset of 53 epitopes concluded that 75% of epitopes are in the range of 15-25 residues [121]. Another analysis of 107 epitopes calculated the average size of an epitope as 15 residues. These previous studies do not agree with the findings of the current study, presumably because the epitopes were defined as contacting residues rather than all residues in regions that include non-contacting residues. Moreover, the current study includes 5 times more epitopes.

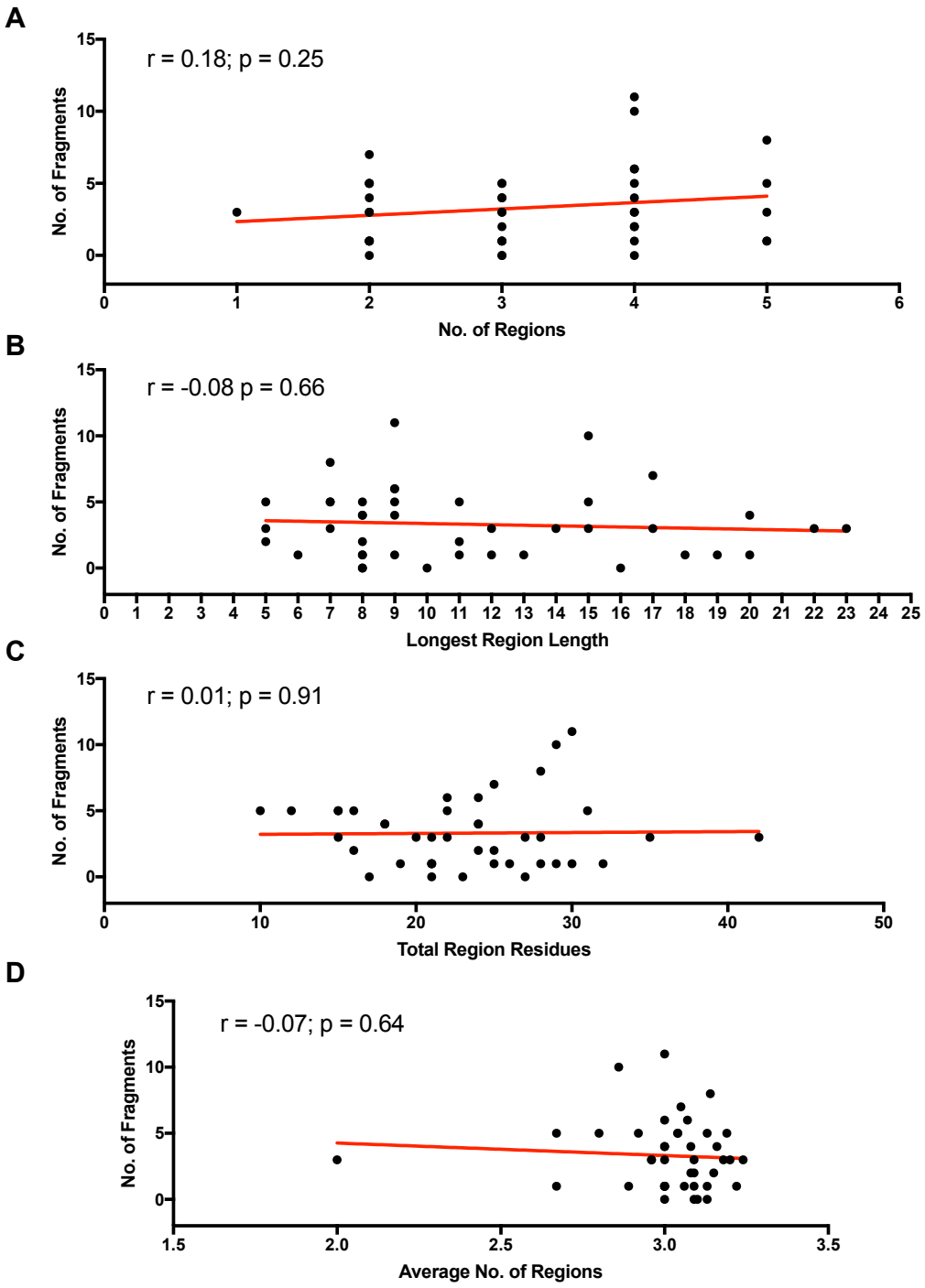


Figure 4.12: Correlation between the number of fragments and A) the number of regions, B) the longest region, C) the number of residues in regions and D) the average number of regions in an epitope in the multiple chain dataset. Only very weak correlations were observed (in B and C).

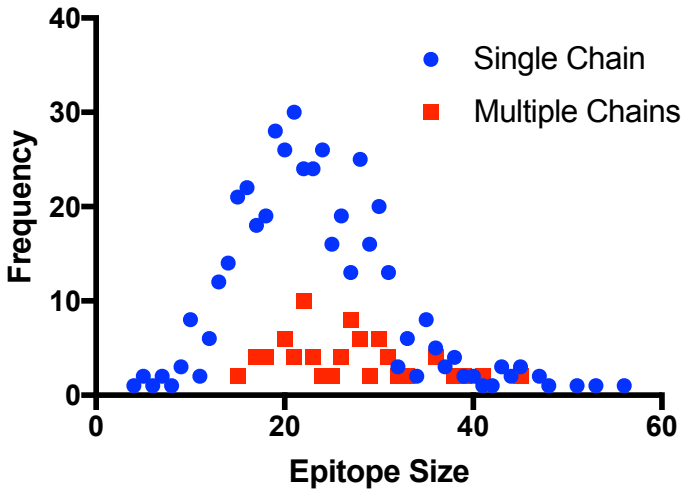


Figure 4.13: Distribution of epitope size in the single and multiple chain datasets. A t-test to compare these distributions shows they are significantly different (p=0.0002).

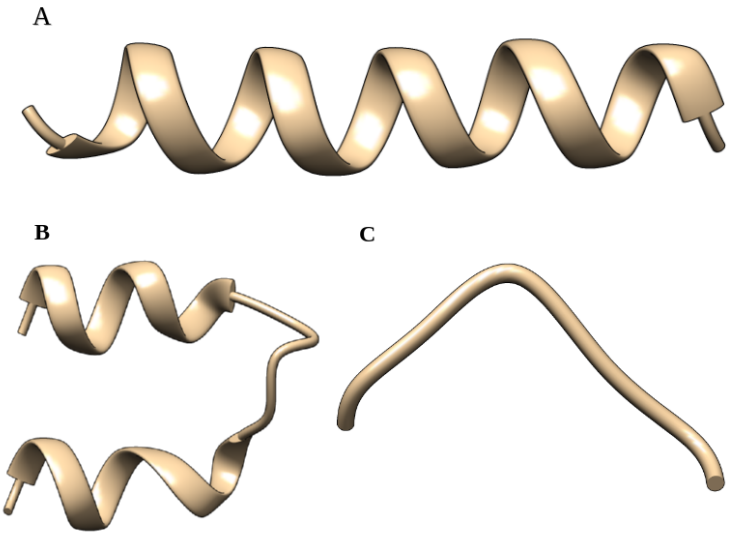


Figure 4.14: Region shapes A) An extended region, B) A folded region, C) A curved region.

4.5 Conformational Analysis of Epitopes – Methods

In order to analyse the conformation (3D fold) and secondary structure of each of the regions, a classification method was developed (in Perl) to classify regions into three different conformations: extended, curved and folded (Figure 4.14). The method also classified each of the regions into three secondary structure classes: helix, strand and coil.

4.5.1 Shape Classification

The peptide shape classification algorithm used a measure of linearity by comparing a given region with an ideal, extended beta strand or alpha helix. Each peptide region was classified as predominantly alpha, beta or coil based on secondary structure assignments performed using an in-house implementation of the Kabsch and Sander [122] method as modified by Smith and Thornton [123]. A threshold of $>$ 60% frequency was used to classify a region into any of the three secondary structure types (helix, strand or coil).

The major steps of the algorithm are shown in Figure 4.15 and described below.

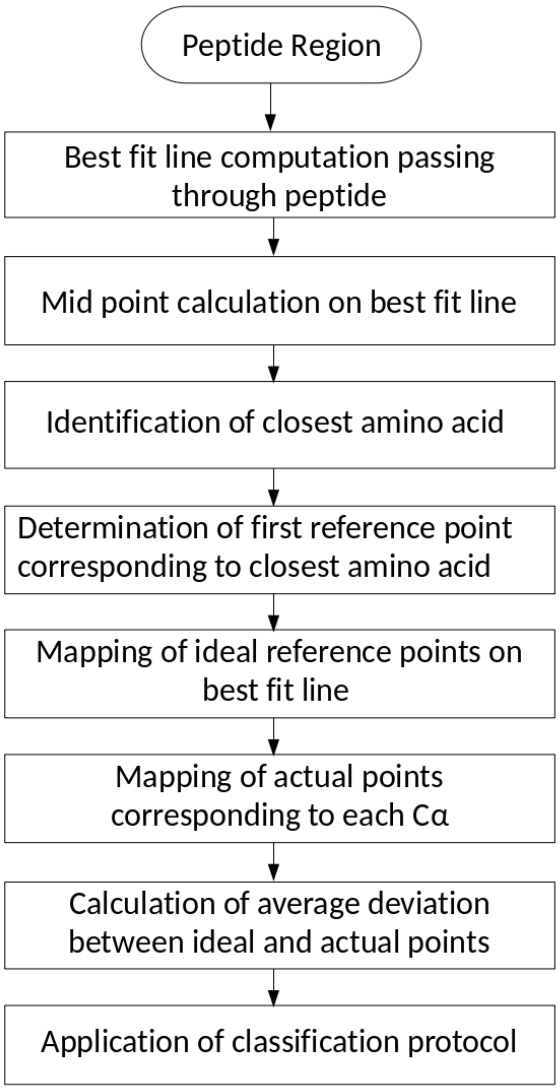


Figure 4.15: Flow chart to describe the steps in shape classification method.

Step 1: A best fit straight line was calculated through the Cα atoms (described below) and converted into a vector (**VL**). A vector for each region (**VR**) was also calculated by using its start and end position (i.e. first and last amino acid of a given region). **VL** was required to be in the same direction as **VR** (Figure 4.16). An angle $< 90°$ between these two vectors confirms that they are in the same direction, if it is not, then **VL** is reversed.

To do this, a C program, *pdbline* [124] was written which draws a best fit line

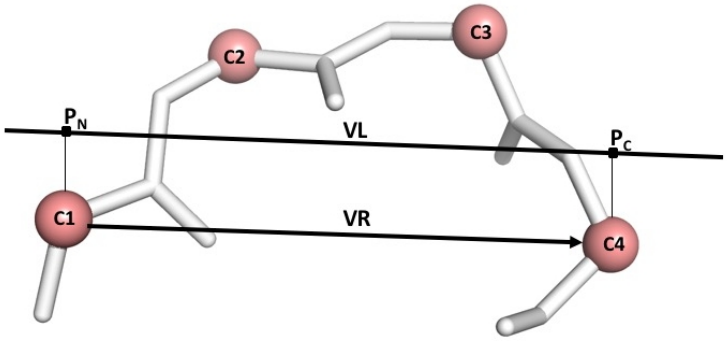


Figure 4.16: The best fit line vector **VL**, in a four residue region, is defined by the termini ($\mathbf{P_N}$ and $\mathbf{P_C}$) of the best fit line which was computed through a set of four residues. The Cα atoms are shown as balls. The region vector **VR** is defined by the first and last residue of the region.

through a specified set of Cα atoms. This program works as follows:

1. Extract the structure (coordinates) of the zone of interest from the PDB file.

2. The centroid of the given set of coordinates is calculated.

3. The covariance matrix and Eigen vectors are computed.

4. Finally, the Eigen vector components of the regression line are used to find other points on the line that pass through the centroid. The Eigen vector with the largest value represents the best fit line passing through the centroid. The first ($\mathbf{P_N}$) and last point ($\mathbf{P_C}$), on the best fit line, are the projections of coordinates of first and last residue onto the Eigen vector. This provides a best fit line for the region as shown in Figure 4.16.

Step 2: The mid point of the best fit line is computed by taking the average of points $\mathbf{P_N}$ and $\mathbf{P_C}$ using Equation 4.5.

$$\mathbf{M} = \frac{\mathbf{P_N} + \mathbf{P_C}}{2} = \left(\frac{x_N + x_C}{2}, \frac{y_N + y_C}{2}, \frac{z_N + z_C}{2} \right) \tag{4.5}$$

where

$\mathbf{P_N}$ = (x_N, y_N, z_N)

$\mathbf{P_C}$ = (x_C, y_C, z_C)

M = (x_M, y_M, z_M)

Step 3: The closest amino acid to the mid point is then identified by calculating the Euclidean distance between the Cα of each amino acid and the mid point using Equation 4.6.

$$d_{min} = \min_{i=1}^{n}(\sqrt{(x_i - x_M)^2 + (y_i - y_M)^2 + (z_i - z_M)^2}) \tag{4.6}$$

Step 4: The closest amino acid (to the mid point) was used as a starting point to map reference points (R_i) onto the best-fit line. Reference points are the positions at which one would expect Cα atoms (Cα, i) to be projected for an ideal extended beta strand (spacing of 3.5 Å) or alpha helix (spacing of 1.5 Å). The number of reference points mapped onto the best fit line is equivalent to the number of residues in the peptide. For example, for a peptide with length *n*, if the starting point for reference point mapping was the i_{th} residue (calculated by finding the closest residue to the mid point), then *i-1* and *n-i* reference points are mapped before and after the start point, respectively. The beta strand spacing of 3.5 Å was used for regions classified as predominantly coil.

The determination of the first starting reference point on the best fit line is shown in Figure 4.17: P_N and P_C are the start and end points of the best fit line; C_i is the closest amino acid to the mid point separated from $\mathbf{P_N P_C}$ by the distance $\boldsymbol{d_c}$. Two vectors **W** (C_i-P_N) and **VL** (P_C-P_N) are computed. The line segment distance ***d*** on the best fit line is computed using Pythagoras's theorem where the magnitude of **W** and $\boldsymbol{d_c}$ are taken as two sides of a right angled triangle. In order to find the point (R_i) along the best fit line at a distance ***d*** from P_N, **VL** was normalised to **U**

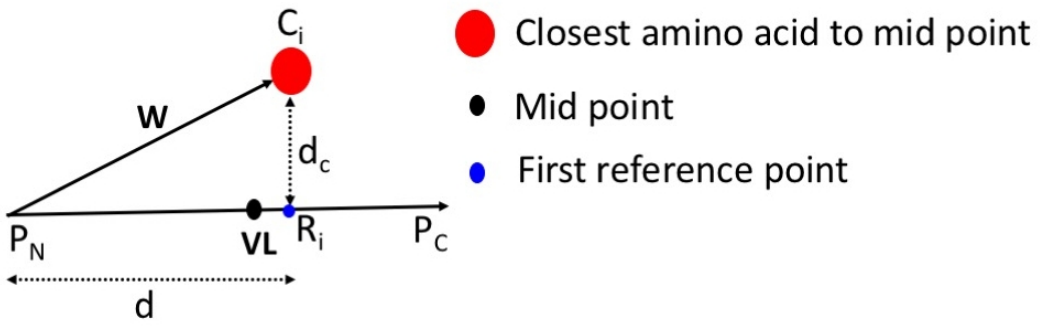


Figure 4.17: Computation of first reference point on the best fit line: P_N and P_C are the start and end points of the best fit line; C_i is the closest amino acid to the mid point seperated by the distance $\boldsymbol{d}_c$. **W** (C_i-P_N) and **VL** (P_C-P_N) are unit vectors. $\boldsymbol{d}$ is a line segment on **VL**.

by using Equation 4.7:

$$\mathbf{U} = \frac{\mathbf{VL}}{\|\mathbf{VL}\|} \tag{4.7}$$

The first reference point R_i is P_N + $\delta\mathbf{U}$ in the direction of P_C, or P_N - $\delta\mathbf{U}$ in the opposite direction. $\delta\mathbf{U}$ is the stepping distance relative to the first reference point on the best fit line. δ represents the $\boldsymbol{d}$ as shown in Figure 4.17. The rest of the reference points are mapped on the basis of the spacing distance as described above.

Step 5: In the next step, the actual points (P_i) corresponding to each Cα in the peptide (C_i) are projected onto the best fit line. To map the actual points onto the best fit line, the above described procedure (computation of the first reference point) has been used where point C_i) in vector **W** corresponds to each Cα projection, in the peptide, on the best fit line (shown in Figure 4.18).

Step 6: The average deviation (**D**) between the reference (R_i) and actual (P_i) points is calculated by finding the difference between them and taking the average. This is used as a measure of the linearity of the peptide (using Equation 4.8 as shown in Figure 4.18).

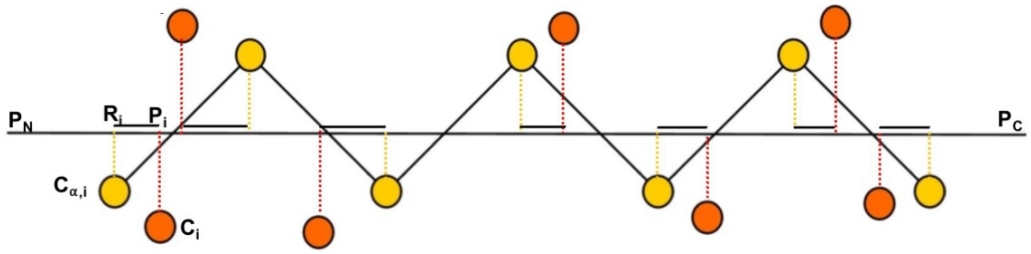


Figure 4.18: Yellow points, $C_{\alpha,i}$ represent the Cα positions of an ideal extended peptide with their projections onto $\mathbf{P_N P_C}$ indicated by R_i with spacing 3.5Å. Red points, C_i represent actual Cα positions with their projections onto $\mathbf{P_N P_C}$ indicated by P_i. The difference between R_i and P_i is computed and used to find the average deviation of actual Cα and ideal Cα positions as a measure of linearity.

$$\mathbf{D} = \frac{\sum_1^n |R_i - P_i|}{n} \tag{4.8}$$

4.5.2 Classification Protocol

Classification cut-offs for the average deviation between an ideal extended conformation and the actual shape were explored using visual analysis. Extended (linear) and non-extended peptides were distinguished on the basis of this deviation cut-off. The non-extended peptides were further classified into curved and folded peptides on the basis of the number of contacts among the residues along the peptide. Peptides were classified as extended if they have length > 4 and a deviation ≤ 1.0 Å or length ≤ 4, and a deviation ≤ 0.5 Å. For peptides with length ≤ 4 and deviation more than 0.5 Å, a 'contact rule' (described in Section 4.5.3) was used. It is most likely that these are small random coils with a curved shape. Longer peptides with an average deviation between 1.0 Å and 2.5 Å are most likely curved or folded, but some were found to be essentially extended peptides with a 'hooked' end (described below). Peptides of length ≥ 6 residues were checked for the presence of a hook.

Some of the extended peptide shapes have a folded hook at one of the end with the rest of the peptide being essentially linear. In such cases, the first residue's de-

viation from ideal was compared with the last residue's deviation. If the deviation of the first residue was more than the last residue then the possible hook was at the start of the peptide otherwise the possible hook was at the end. The average deviation of the whole peptide was recalculated by excluding the first (or last) residue. If the average deviation is still more than 1.0 Å then the process of exclusion of up to three terminal (N-terminus and C-terminus separately) residues and recalculation of average deviation was repeated. If the average deviation is then $<$ 1.0 Å the peptide is classified as extended, otherwise it was checked for curved and folded by the 'contact rule'. Peptides retaining an average deviation $>$ 2.5 Å are tested by the 'contact rule' to choose between curved and folded shapes. The flow chart in the Figure 4.19 outlines the classification protocol.

4.5.3 The Contact Rule

Extended and non-extended peptides were classified on the basis of a deviation cut-off. For further classification of non-extended peptides into curved and folded peptides, a 'contact rule' was devised which works on the basis of the number of contacts among the residues along the peptide. The number of contacts (defined at a distance of $\leq$ 4 Å between any atom centres) was calculated between pairs of residues defined as:

$$n-d : n+i+d$$

where $\boldsymbol{n}$ is the reference position in the peptide, $\boldsymbol{i}$ is the spacing/separation between residues making contact ($\boldsymbol{i} \geq 3$) and $\boldsymbol{d}$ is a step along the residues of the peptide ($\boldsymbol{d} \geq 0$). This equation is iterated over $\boldsymbol{n}$, $\boldsymbol{d}$ and $\boldsymbol{i}$. Figure 4.20 shows pairs of residues making contacts within the set distance criteria.

In order to consider multiple folds in a peptide, a contact threshold (T_C) was used to define the separation between contacting residues to identify local contacts

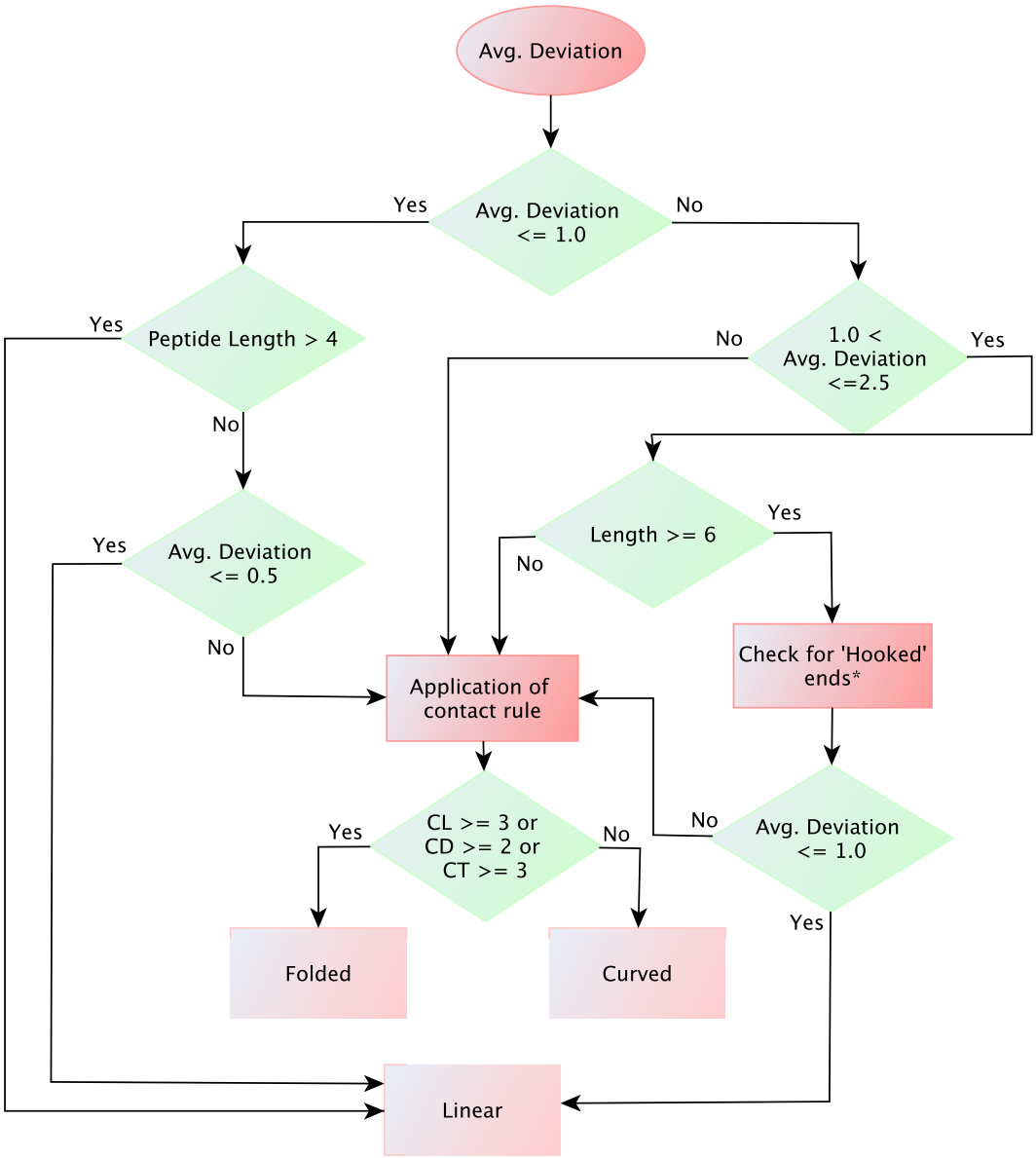


Figure 4.19: Flow chart of the peptide shape classification protocol: C_L, C_D and C_T refer to local, distant and total contacts, respectively, among the residues along the peptide (See Section 4.5.3). * Details of 'Hooked' peptides are described in Section 4.5.2.

($\boldsymbol{i} \leq T_C$) and distant contacts ($\boldsymbol{i} > T_C$). The contact threshold (T_C) is computed by halving the length of the peptide $\boldsymbol{N}$ if it is less than or equal to 12 and assigned as 5 otherwise (Equation 4.9).

$$T_C = \begin{cases} N/2 & if(N \leq 12) \\ 5 & otherwise \end{cases} \tag{4.9}$$

A decision about the shape of a peptide is made on the basis of the number of local (C_L), distant (C_D) or total contacts (C_T). Two or more distant contacts, three or more local contacts, or three or more total (local + distant) contacts leads to the peptide being classified as folded rather than curved (Equation 4.10).

$$Folded = \begin{cases} C_L \geq 3 \\ C_D \geq 2 \\ C_T \geq 3 \end{cases} \tag{4.10}$$

This rule is termed the 'contact rule' and the method is explained in Algorithm 1, below.

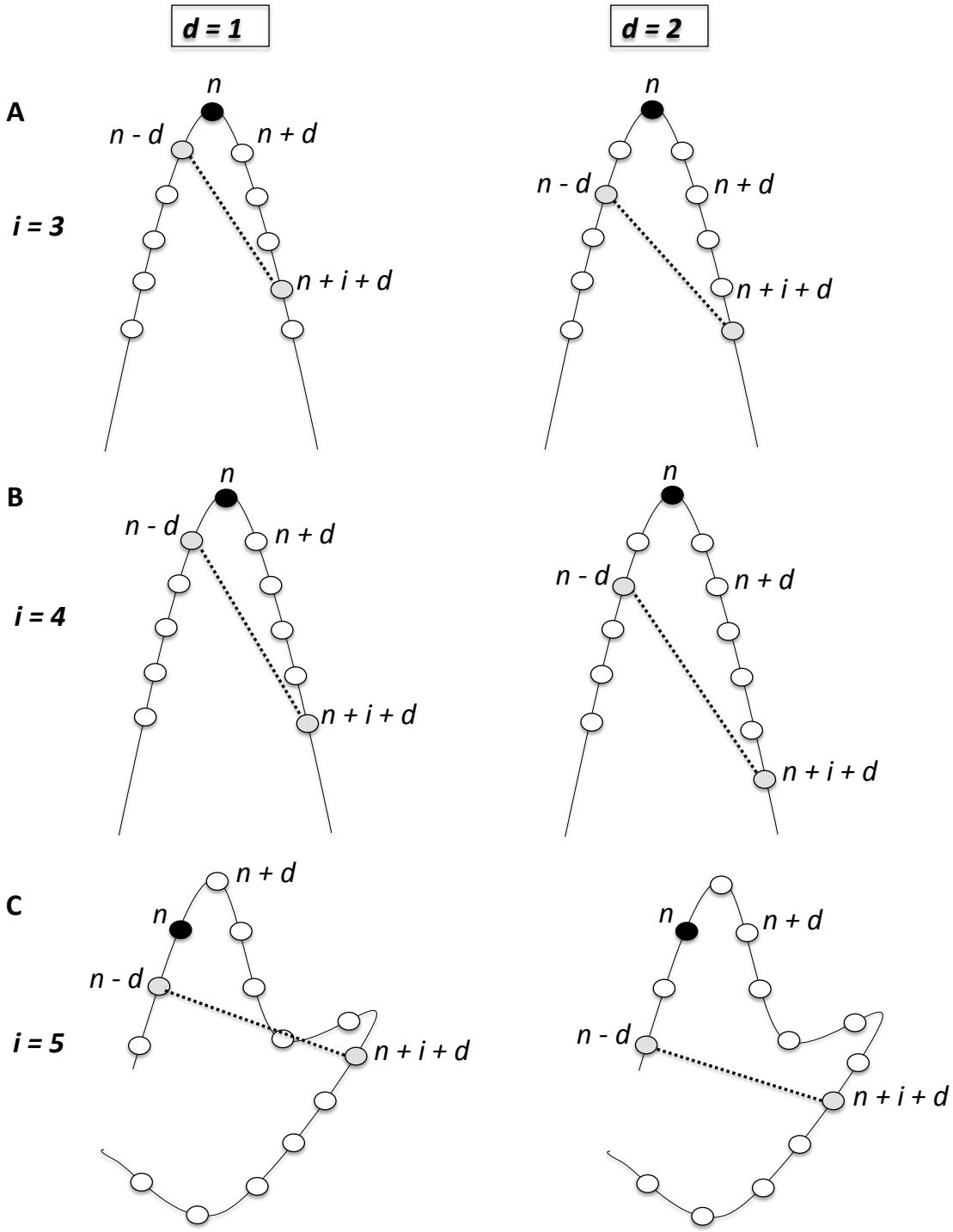


Figure 4.20: Contacts in folded peptides A) A peptide with 11 residues is shown where ***n*** is the reference position to start for checking the pairs of amino acids making contacts. A spacing ***i*** of 3 residues has been shown along the residue that is contacting with a stepping distance ***d*** of one and two respectively. B) The same peptide with a spacing of ***i***, of 4 residues. C) A 13 residue long peptide with multiple folds where the spacing ***i*** depends on the contact threshold (T_C) which is computed on the basis of the length of the peptide (see text).

Algorithm 1 Calculation of contacts for folded and curved shape classification.

```
1:  procedure
        Input
2:      PeptideResidues[]
        Initialization
3:      C_T ← 0
4:      C_L ← 0
5:      C_D ← 0
6:      len ← length of peptide
7:      if len <= 12 then
8:          T_C ← len/2
9:      else
10:         T_C ← 5
11:     end if
12:     for n ← 0, n < len do
13:         for i ← 3, i < len − 3 do
14:             for d ← 0, d < len do
15:                 res1 ← PeptideResidues[n - d]
16:                 res2 ← PeptideResidues[n + i + d]
17:                 if res1 ≥ 0 & res2 < len then
18:                     CalculateAtomicDistance(res1, res2)
19:                     if distance ≤ 4.0 then
20:                         if i ≤ T_C then
21:                             C_L++
22:                         else
23:                             C_D++
24:                         end if
25:                     end if
26:                 end if
27:             end for
28:         end for
29:     end for
30:     C_T ← C_L+C_D
31:     if C_L ≥ 3||C_D ≥ 2||C_T ≥ 3 then
32:         shape ← folded
33:     else
34:         shape ← curved
35:     end if
36: end procedure
```

4.5.4 Statistical Tests

In this study, tests include 2-way chi-squared, 3-way chi-squared, Welch's t-test and Pearson correlation. It is a common practice to use a 2-way chi-squared test to find the significance of dependence between two categorical variables in a single population. In this study, the correlation of three variables needed to be determined. Therefore, a three-way chi-squared was implemented to find the significance of independence among three types of conformational shapes (extended, curved and folded) and secondary structure elements (helix, strand and coil). The method of three-way chi-squared calculation is described below.

4.5.4.1 Calculation of 3D chi-Squared

First, for the null hypothesis, complete independence was assumed between the three variables.

If rows, columns and planes are referred as r, c, p, (with dimensions R, C, P) with each cell containing the observed value o_{rcp} then a total, t, can be defined for a particular row, r, as:

$$t_{r++} = \sum_{c=1}^{C} \sum_{p=1}^{P} o_{rcp}$$

(where a subscript of $+$ indicates summation over the appropriate index)

Similarly for columns and planes:

$$t_{+c+} = \sum_{r=1}^{R} \sum_{p=1}^{P} o_{rcp}$$

$$t_{++p} = \sum_{r=1}^{R} \sum_{c=1}^{C} o_{rcp}$$

The expected value for a given cell, e_{rcp} is then:

$$e_{rcp} = \frac{t_{r++} \times t_{+c+} \times t_{++p}}{N^2}$$

(where N is the total number of observations).

The chi-squared value is then calculated as normal:

$$\chi^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{p=1}^{P} \frac{(o_{rcp} - e_{rcp})^2}{e_{rcp}} \tag{4.11}$$

The number of degrees of freedom, D, is simply:

$$D = (R-1)(C-1)(P-1)$$

The calculation of the expected values is based on information from Lienert and Wolfrum [125], Lin [126] and Li [127].

4.6 Conformational Analysis of Epitope Regions – Results

Several methods [27, 34, 38–40, 121, 128–131] have been developed for the prediction of conformational B-cell epitopes. These methods used 3D structural information of an epitope along with several other features that include amino acid properties, spatial information, surface accessibility and residue clustering. Unfortunately, none of these methods is able to provide good prediction of conformational B-cell epitopes. Therefore, it is important to understand their 3D structural shape. To this end, a detailed analysis was performed of the 3D structure and shape of epitopes. In the dataset of 1329 regions, 1195 regions were extracted from 464 single chain

Table 4.5: Classification of regions in 3 different shapes and sub classification of each of the shape on the basis of secondary structure

	Single Chain Antigens (464)	**Multi chain Antigens (42)**
Extended	**475**	**52**
Helix	152	11
Strand	107	9
Coil	216	32
Curved	**578**	**68**
Helix	28	5
Strand	25	1
Coil	525	62
Folded	**142**	**14**
Helix	22	3
Strand	29	5
Coil	91	6
Total	**1195**	**134**

antigens while 134 regions were identified in 42 antigens comprised of multiple chains. The shape of each of the regions was classified into either extended, curved or folded. Each of the shapes was then further classified by their secondary structure content. Table 4.5 shows the statistics for each of the shape categories with sub-categories of secondary structure forming these regions.

4.6.1 Region Length Analysis in each of the Shapes

The length of each of the regions was investigated in all shape categories as shown in Figure 4.21. Most extended and curved regions are between 3 and 9 residues, whereas folded regions are comprised of 9 to 17 residues. A similar trend can be seen in both datasets of epitopes. t-tests were performed on extended/curved, extended/folded and curved/folded shape pairs which showed that the length distribution of extended/curved regions is not significantly different whereas extended/folded and curved/folded region lengths are significantly different. The smallest and longest region in the extended dataset was 3 and 25 residues

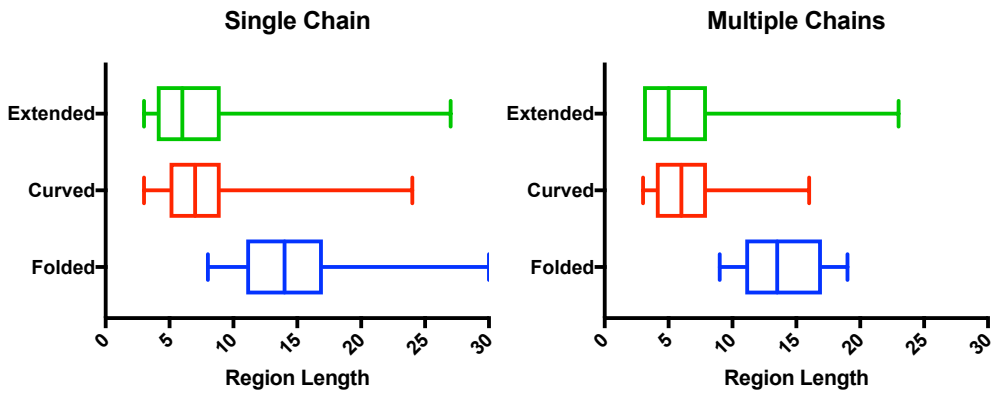


Figure 4.21: The distribution of region lengths in extended, curved and folded shape regions in single and multiple chain datasets. The p-value for extended/curved (0.92), extended/folded (<0.0001) and curved/folded (<0.0001) was computed and was found to be similar in both the single and multiple chain datasets. In the box and whisker diagram, the whiskers represent the minimum to maximum values in the data whereas the box shows the lower quartile, median and upper quartile.

respectively whereas 80% of extended regions are comprised of 3–9 residues. The longest region in the curved dataset was 24 amino acids. Folded regions were 8 to 30 residues long with 73% of regions in the range of 9–17 amino acids.

4.6.2 Distribution of Region Shapes in the Epitope Datasets

Each of the epitopes was investigated for the presence of regions of a particular shape/conformation. The epitopes coming from single and multiple antigen chains were explored separately. The distribution of each of the shapes is shown in Figure 4.22. A chi-squared test was performed on these three shapes of region to determine whether they are randomly distributed in the single chain dataset. A p-value of < 0.0001 confirms that these three region shapes are correlated with one another in the single chain dataset. However, in the multiple chain dataset, a p-value of 0.26 was seen suggesting that region shapes are randomly distributed in this dataset.

Epitopes can either have regions of similar shape (i.e. only extended, curved or folded) or combinations of different shapes (i.e. extended-curved, extended-folded,

curved-folded or extended-curved-folded). In the single chain dataset, 66 epitopes were observed having only extended regions with one instance of 6 extended regions in a single epitope. 86 epitopes were seen with just curved regions (up to 5). There were 36 epitopes with one or two folded regions. However, owing to the presence of more than one type of shape region in an individual epitope, all the possible combinations of shapes were investigated.

4.6.2.1 3-way Comparison of Shapes

In order to investigate every possible combination of region shape (along with region number) in a particular epitope, a 3D contingency table was computed. A total of 126 (3x6x7) combinations were formed due to 0–6, 0–5 and 0–2 number of regions with extended, curved and folded shapes respectively (Table 4.6). The null hypothesis for this 3 way test would be that extended shape regions are independent of curved and folded, and curved are independent of folded, i.e. there is no correlation between any of the shapes. However, for a chi-squared test to be valid, there should be no more than 20% of the expected values below five and no expected values below one [132]. However, the data had lots of very low expected values owing to zero counts (observed values) for several possible combinations of shapes. Consequently, data were grouped as shown in Table A.4 (see Appendix). The 3-way chi-squared test showed a p-value of 0 indicating a strong correlation among the shapes.

The observed and expected values of different combinations showed clear trends. For example, the chance of having E1 (1 extended region) when there are no curved or folded regions is much less likely than expected (p-value = $2.22\text{x}10^{-15}$). However, the chances of having E1 in the presence of C1 or C2 is much more likely than expected (p-value; F0/C1/E1 = $1.92\text{x}10^{-5}$, F0/C2/E1 = $1.72\text{x}10^{-6}$). E2 in the

Table 4.6: Frequency of every possible combination of Folded, curved and extended shape in single chain dataset. F0-F2 shows presence of 0, 1 or 2 folded regions in an epitope. C0–C5 and E0–E6 represents the presence of 0–5 and 0–6 number of regions with curved and extended conformation respectively. Hence, the contingency table (3x6x7) is computed on the basis of maximum possible number of regions of a particular shape making an epitope.

Folded	Curved	Extended	Count	Folded	Curved	Extended	Count	Folded	Curved	Extended	Count	Folded	Curved	Extended	Count
F0	C0	E0	0	F0	C4	E4	0	F1	C3	E1	3	F2	C1	E5	0
F0	C0	E1	20	F0	C4	E5	0	F1	C3	E2	1	F2	C1	E6	0
F0	C0	E2	31	F0	C4	E6	0	F1	C3	E3	0	F2	C2	E0	1
F0	C0	E3	13	F0	C5	E0	4	F1	C3	E4	0	F2	C2	E1	0
F0	C0	E4	0	F0	C5	E1	0	F1	C3	E5	0	F2	C2	E2	0
F0	C0	E5	1	F0	C5	E2	0	F1	C3	E6	0	F2	C2	E3	0
F0	C0	E6	1	F0	C5	E3	0	F1	C4	E0	1	F2	C2	E4	0
F0	C1	E0	18	F0	C5	E4	1	F1	C4	E1	0	F2	C2	E5	0
F0	C1	E1	62	F0	C5	E5	0	F1	C4	E2	1	F2	C2	E6	0
F0	C1	E2	24	F0	C5	E6	0	F1	C4	E3	0	F2	C3	E0	1
F0	C1	E3	15	F1	C0	E0	32	F1	C4	E4	0	F2	C3	E1	0
F0	C1	E4	0	F1	C0	E1	21	F1	C4	E5	0	F2	C3	E2	0
F0	C1	E5	0	F1	C0	E2	11	F1	C4	E6	0	F2	C3	E3	0
F0	C1	E6	0	F1	C0	E3	2	F1	C5	E0	0	F2	C3	E4	0
F0	C2	E0	40	F1	C0	E4	1	F1	C5	E1	0	F2	C3	E5	0
F0	C2	E1	40	F1	C0	E5	0	F1	C5	E2	0	F2	C3	E6	0
F0	C2	E2	13	F1	C0	E6	0	F1	C5	E3	0	F2	C4	E0	0
F0	C2	E3	1	F1	C1	E0	19	F1	C5	E4	0	F2	C4	E1	0
F0	C2	E4	0	F1	C1	E1	13	F1	C5	E5	0	F2	C4	E2	0
F0	C2	E5	0	F1	C1	E2	6	F1	C5	E6	0	F2	C4	E3	0
F0	C2	E6	0	F1	C1	E3	0	F2	C0	E0	4	F2	C4	E4	0
F0	C3	E0	18	F1	C1	E4	0	F2	C0	E1	1	F2	C4	E5	0
F0	C3	E1	13	F1	C1	E5	0	F2	C0	E2	0	F2	C4	E6	0
F0	C3	E2	3	F1	C1	E6	0	F2	C0	E3	0	F2	C5	E0	0
F0	C3	E3	0	F1	C2	E0	11	F2	C0	E4	0	F2	C5	E1	0
F0	C3	E4	0	F1	C2	E1	2	F2	C0	E5	0	F2	C5	E2	0
F0	C3	E5	0	F1	C2	E2	1	F2	C0	E6	0	F2	C5	E3	0
F0	C3	E6	0	F1	C2	E3	0	F2	C1	E0	0	F2	C5	E4	0
F0	C4	E0	6	F1	C2	E4	0	F2	C1	E1	0	F2	C5	E5	0
F0	C4	E1	4	F1	C2	E5	0	F2	C1	E2	0	F2	C5	E6	0
F0	C4	E2	1	F1	C2	E6	0	F2	C1	E3	0				
F0	C4	E3	0	F1	C3	E0	3	F2	C1	E4	0				

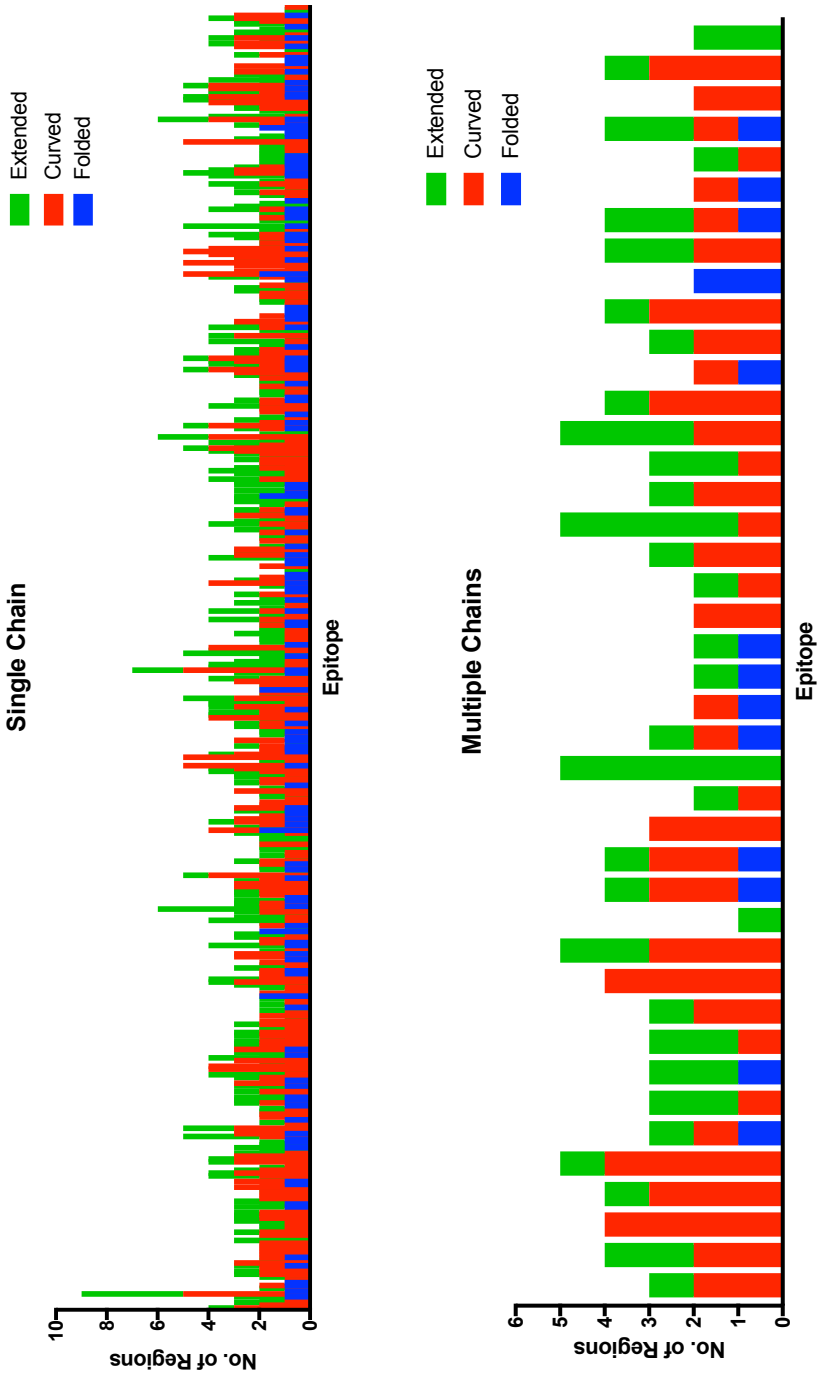


Figure 4.22: The distribution of extended, curved and folded regions in the single (464 unique epitopes) and multiple (42 unique epitopes) chains epitope dataset.

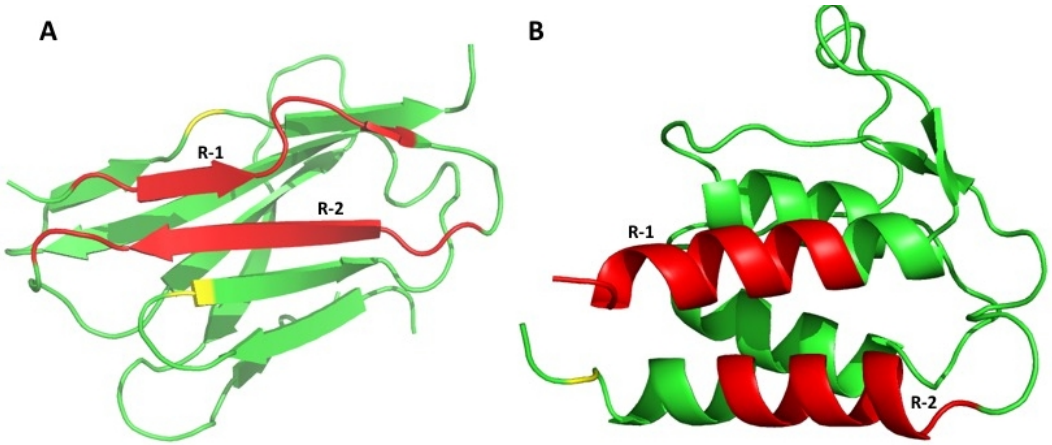


Figure 4.23: Epitopes with 2 extended regions. A) An epitope mapped onto PDB 2ARJ − comprised of 2 extended regions (red) and 2 fragments (yellow). B) An epitope mapped onto PDB 4JLR − comprised of 2 extended regions (red) and 1 fragment (yellow).

absence of any curved and folded regions is much higher than expected whereas E2 in the presence of any curved regions is less likely to occur than expected (p-value = 0). This shows that two extended regions are enough to make an epitope on its own without the contribution of any other shape. Two examples are shown in Figure 4.23. Epitopes, having more than three extended regions, have zero or one curved regions (p-value; F0/C0/E3456 = 0, F0/C0/E3456 = 0.00029).

In the 3D contingency table, it is evident that C1 tends to occur much less frequently than expected by chance when there are no extended or folded regions (p-value = $8.04\text{x}10^{-12}$). However, C1 and E1 occur together much more than expected by chance (p-value = $1.92\text{x}10^{-5}$). For epitopes with C1 but extended regions over 2 (i.e. E23456), there is a negligible difference between observed and expected values (p-value = 0.1). For C2, having E0 and E1 (in the absence of folded) is much more likely to happen as compared to C2 with E23456 regions (p-value; F0/C2/E1 = $1.72\text{x}10^{-6}$, F0/C2/E2 = $6.22\text{x}10^{-7}$, F0/C2/E3456 = $4.65\text{x}10^{-10}$). For epitopes having 3 and more curved regions (i.e. C345) in the absence of E0 and F0 are more likely to occur than expected by chance (p-value = $2.88\text{x}10^{-15}$). How-

ever, it is less likely to have extended regions when there are more than two curved regions (p-value; F0/C345/E2 = $5.33\text{x}10^{-5}$).

F12 (i.e. Folded-1 and Folded-2 together) when there are no curved and extended regions are more likely than expected (p-value = 0). This implies that one or two folded regions are enough to make the structure of an epitope without the contribution of other region shapes. This is because folded regions are normally longer than the other two region shapes (Figure 4.21).

There are 13 epitopes which have all three shapes (E1, C1 and F1) in one epitope but these seem to occur much less than expected by chance (p-value = $1.92\text{x}10^{-5}$). In fact, the probability of any number of curved and extended regions in the presence of folded regions is low.

The significance of each of the above described combinations was calculated using 2x2x2 chi-squared test (see Section 4.5.4.1 for development of this test). These results show that the presence of one shape and the number of regions influences the presence of other shapes.

2-way Comparison of Shapes

The distribution of shapes in the epitope dataset in Figure 4.22 suggests that extended-curved and curved-folded combinations tend to occur more frequently, therefore it was interesting to find out the probability of any two shapes occurring together in one epitope. To this end, a 2-way chi-squared test was performed on extended-curved, extended-folded, and curved-folded shapes.

In the case of extended-curved, epitopes having more than two curved regions (C345) tend to have no extended regions (p-value = $1.60\text{x}10^{-14}$) whereas the chances of C1 and E1 (but not E2) being together are more likely than expected by chance (p-value; C1/E1 = 0.006, C1/E2 = 0.57). However, C345 tends to oc-

cur much less than expected by chance if two extended regions are already there (p-value = $9.58\text{x}10^{-5}$). In the case of E345, there are high chances of having no curved region, i.e. C0 (p-value = $6.70\text{x}10^{-5}$). There are no significant chances of having E345 and C1 or C2345 (p-value = 0.18).

The grouped data of the 2D chi-squared test for extended-folded shows that folded regions (F1 or F2), on their own, are more likely to occur than expected by chance (p-value = $2.35\text{x}10^{-11}$). A similar trend was observed in the case of folded and curved (p-value = $8.49\text{x}10^{-14}$). The significance of the above results was determined by a 2x2 chi-squared test (with Yates correction).

In the multiple chain dataset, a 3-way contingency table (3x5x6) was computed with 90 unique combinations of extended, curved and folded shape regions (Table 4.7). Owing to the very small dataset, comprised of 42 epitopes, a very extensive grouping had to be done. Nevertheless, it did not fulfil the requirements of a chi-squared test. Therefore, it was not possible to find significance. However, by looking at the data, 3, 5 and 1 epitopes were found comprised of solely extended, curved and folded regions respectively. 50% of epitopes were seen with extended (up to 5) and curved (up to 4) together. A very small fraction of folded regions was seen associated with up to 2 extended or curved regions. 6 examples were observed with the presence of all three shapes in an individual epitope (Figure 4.13).

Table 4.7: Frequency of every possible combination of Folded, curved and extended shape in multiple chain dataset. F0-F2 shows presence of 0, 1 or 2 folded regions in an epitope. C0-C4 and E0-E5 represents the presence of 0-4 and 0-5 number of regions with curved and extended conformation respectively. Hence, the contingency table (3x5x6) is computed on the basis of maximum possible number of regions of a particular shape making an epitope.

Folded	Curved	Extended	Count	Folded	Curved	Extended	Count	Folded	Curved	Extended	Count
F0	C0	E0	0	F1	C0	E0	0	F2	C0	E0	1
F0	C0	E1	1	F1	C0	E1	2	F2	C0	E1	0
F0	C0	E2	1	F1	C0	E2	1	F2	C0	E2	0
F0	C0	E3	0	F1	C0	E3	0	F2	C0	E3	0
F0	C0	E4	0	F1	C0	E4	0	F2	C0	E4	0
F0	C0	E5	1	F1	C0	E5	0	F2	C0	E5	0
F0	C1	E0	0	F1	C1	E0	3	F2	C1	E0	0
F0	C1	E1	3	F1	C1	E1	2	F2	C1	E1	0
F0	C1	E2	3	F1	C1	E2	2	F2	C1	E2	0
F0	C1	E3	0	F1	C1	E3	0	F2	C1	E3	0
F0	C1	E4	1	F1	C1	E4	0	F2	C1	E4	0
F0	C1	E5	0	F1	C1	E5	0	F2	C1	E5	0
F0	C2	E0	2	F1	C2	E0	0	F2	C2	E0	0
F0	C2	E1	5	F1	C2	E1	2	F2	C2	E1	0
F0	C2	E2	2	F1	C2	E2	0	F2	C2	E2	0
F0	C2	E3	1	F1	C2	E3	0	F2	C2	E3	0
F0	C2	E4	0	F1	C2	E4	0	F2	C2	E4	0
F0	C2	E5	0	F1	C2	E5	0	F2	C2	E5	0
F0	C3	E0	1	F1	C3	E0	0	F2	C3	E0	0
F0	C3	E1	4	F1	C3	E1	0	F2	C3	E1	0
F0	C3	E2	1	F1	C3	E2	0	F2	C3	E2	0
F0	C3	E3	0	F1	C3	E3	0	F2	C3	E3	0
F0	C3	E4	0	F1	C3	E4	0	F2	C3	E4	0
F0	C3	E5	0	F1	C3	E5	0	F2	C3	E5	0
F0	C4	E0	2	F1	C4	E0	0	F2	C4	E0	0
F0	C4	E1	1	F1	C4	E1	0	F2	C4	E1	0
F0	C4	E2	0	F1	C4	E2	0	F2	C4	E2	0
F0	C4	E3	0	F1	C4	E3	0	F2	C4	E3	0
F0	C4	E4	0	F1	C4	E4	0	F2	C4	E4	0
F0	C4	E5	0	F1	C4	E5	0	F2	C4	E5	0

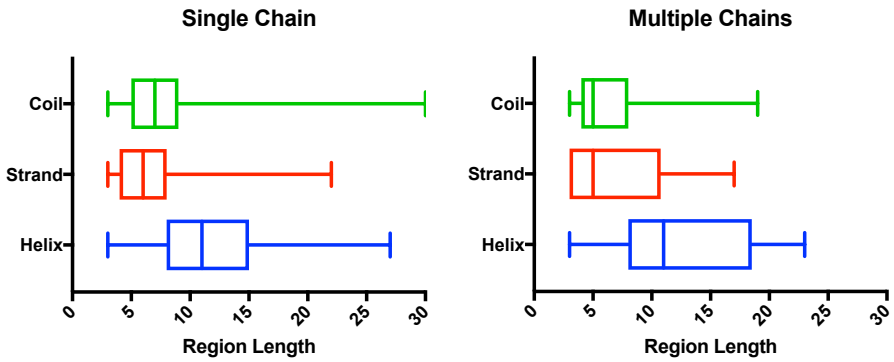


Figure 4.24: The distribution of region lengths in helix, strand and coil structure of regions in the single and multiple chain datasets. The significance of the differences between each pair of secondary structure elements was computed. In the single chain dataset, the length of helix-coil ($p < 0.0001$) helix-strand ($p < 0.0001$), strand-coil ($p < 0.002$) were found to be significantly different. In the multiple chain dataset, the lengths of helix-coil ($p < 0.0008$) and helix-strand ($p < 0.02$) were found to be significantly different, but this was not true for strand-coil ($p = 0.36$). In the box and whisker diagram, the whiskers represent the minimum to maximum values in the data whereas the box shows the lower quartile, median and upper quartile.

4.6.3 Secondary Structure in the Epitope Dataset

The secondary structure of regions forming epitopes, classified into helix, strand and coil was analysed. It is interesting to investigate the length of regions with these three secondary structure elements. It was observed that regions having helix structure tend to be longer than strand and coil regions. This is because of the gaps of 3 amino acids between contacting residues. The length distribution of each of the region structure types is shown in Figure 4.24.

Table 4.5 shows that random coils dominate over helix and strands in regions and that most of these are curved. In the single chain dataset, 70% of regions are coils while 74% of regions are coil in the multiple chain dataset. This agrees with previous studies of epitopes where it was reported that epitopes are enriched by loops and depleted of helices and strands [112, 113]. It is interesting to know how these coil, helix and strand secondary structure elements are distributed in each of the epitopes. The secondary structure class of regions (i.e. the number of re-

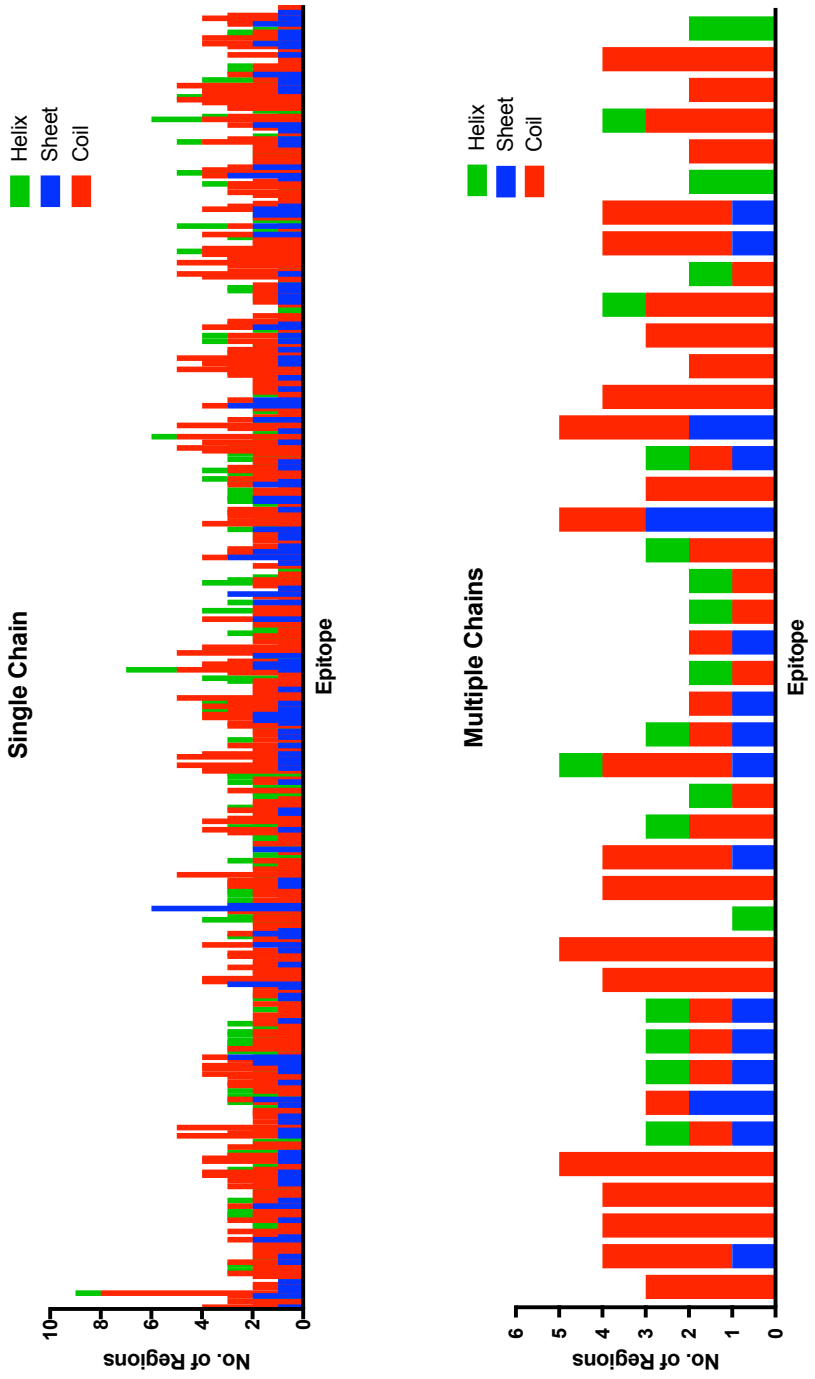


Figure 4.25: The distribution of helix, strand and coil in the regions of each of the epitope in the dataset. The distribution of helix, strand and coil in the regions of the single (464 unique epitopes) and multiple (42 unique epitopes) chains epitope dataset.

gions assigned a given type in an epitope) in epitopes was calculated as shown in Figure 4.25. In the single chain dataset, it was observed that out of 464 epitopes, 184 possessed only random coil regions with a number varying from one to seven. There were 50 epitopes with only helix regions (1-3 regions). However, epitopes consisting only of β-strand regions were less common (19 examples containing up to 6 β-strand regions). More than half (about 54%) of the epitopes were found to be composed of solely helices, strands or coils. The rest of the epitopes had a combination of helix-strand, helix-coil, strand-coil or helix-strand-coil.

The relationship among these three secondary structure elements was also studied by a 3-way chi-squared test and it was found that the presence of one type of secondary structure element influences the presence of the others. A 3-way contingency table (4x7x9) was computed to include every possible combination of helix, strand and coil (Table 4.8 and 4.9). Due to the presence of a lot of zero values for several combinations, grouping was performed that resulted in contingency Table A.5 (shown in Appendix A) with 24 grouped combinations.

The data showed that the chances of having coiled regions of C2–C8 in the absence of any helix (H0) or strand (S0) is much more likely than expected by chance. The relationship of H0, S0 and C2345678 was studied by 2x2x2 chi-squared test (p-value = $1.11\text{x}10^{-16}$). Similarly, the probability of having 1–3 helical regions in the absence of any coiled and strand region is much more likely than expected by chance (p-value = 0). It is very unlikely to have a strand region in an epitope that contains a helical region (p-value = 0). In conclusion, most of the epitopes tend to have regions with the same type of secondary structure element. Different combinations were seen, but more or less frequently than expected by chance.

In the multiple chain dataset, 14 epitopes were observed that formed the whole

Table 4.8: Frequency of every possible combination of helix, strand and curved shape in single chain dataset. H0-H3 shows presence of 0, 1, 2 or 3 helical regions in an epitope. S0-C6 and C0-C8 represents the presence of 0-6 and 0-8 number of regions with strand and coil conformation respectively. Hence, the contingency table (4x7x9) is computed on the basis of maximum possible number of regions of a particular secondary structure making an epitope.

Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count
H0	S0	C0	0	H0	S3	C5	0	H1	S0	C1	35	H1	S3	C6	0
H0	S0	C1	40	H0	S3	C6	0	H1	S0	C2	34	H1	S3	C7	0
H0	S0	C2	78	H0	S3	C7	0	H1	S0	C3	6	H1	S3	C8	0
H0	S0	C3	39	H0	S3	C8	0	H1	S0	C4	3	H1	S4	C0	0
H0	S0	C4	15	H0	S4	C0	0	H1	S0	C5	1	H1	S4	C1	0
H0	S0	C5	12	H0	S4	C1	0	H1	S0	C6	0	H1	S4	C2	0
H0	S0	C6	0	H0	S4	C2	0	H1	S0	C7	0	H1	S4	C3	0
H0	S0	C7	0	H0	S4	C3	0	H1	S0	C8	1	H1	S4	C4	0
H0	S0	C8	0	H0	S4	C4	0	H1	S1	C0	5	H1	S4	C5	0
H0	S1	C0	10	H0	S4	C5	0	H1	S1	C1	7	H1	S4	C6	0
H0	S1	C1	28	H0	S4	C6	0	H1	S1	C2	4	H1	S4	C7	0
H0	S1	C2	23	H0	S4	C7	0	H1	S1	C3	0	H1	S4	C8	0
H0	S1	C3	17	H0	S4	C8	0	H1	S1	C4	0	H1	S5	C0	0
H0	S1	C4	3	H0	S5	C0	0	H1	S1	C5	0	H1	S5	C1	0
H0	S1	C5	0	H0	S5	C1	0	H1	S1	C6	0	H1	S5	C2	0
H0	S1	C6	0	H0	S5	C2	0	H1	S1	C7	0	H1	S5	C3	0
H0	S1	C7	0	H0	S5	C3	0	H1	S1	C8	0	H1	S5	C4	0
H0	S1	C8	0	H0	S5	C4	0	H1	S2	C0	4	H1	S5	C5	0
H0	S2	C0	5	H0	S5	C5	0	H1	S2	C1	0	H1	S5	C6	0
H0	S2	C1	13	H0	S5	C6	0	H1	S2	C2	0	H1	S5	C7	0
H0	S2	C2	10	H0	S5	C7	0	H1	S2	C3	0	H1	S5	C8	0
H0	S2	C3	0	H0	S5	C8	0	H1	S2	C4	0	H1	S6	C0	0
H0	S2	C4	0	H0	S6	C0	1	H1	S2	C5	0	H1	S6	C1	0
H0	S2	C5	0	H0	S6	C1	0	H1	S2	C6	0	H1	S6	C2	0
H0	S2	C6	0	H0	S6	C2	0	H1	S2	C7	0	H1	S6	C3	0
H0	S2	C7	0	H0	S6	C3	0	H1	S2	C8	0	H1	S6	C4	0
H0	S2	C8	0	H0	S6	C4	0	H1	S3	C0	0	H1	S6	C5	0
H0	S3	C0	3	H0	S6	C5	0	H1	S3	C1	0	H1	S6	C6	0
H0	S3	C1	4	H0	S6	C6	0	H1	S3	C2	0	H1	S6	C7	0
H0	S3	C2	0	H0	S6	C7	0	H1	S3	C3	0	H1	S6	C8	0
H0	S3	C3	0	H0	S6	C8	0	H1	S3	C4	0	H2	S0	C0	26
H0	S3	C4	0	H1	S0	C0	20	H1	S3	C5	0	H2	S0	C1	3

Table 4.9: Frequency of every possible combination of helix, strand and curved shape in single chain dataset. H0-H3 shows presence of 0, 1, 2 or 3 helical regions in an epitope. S0-C6 and C0-C8 represents the presence of 0-6 and 0-8 number of regions with strand and coil conformation respectively. Hence, the contingency table (4x7x9) is computed on the basis of maximum possible number of regions of a particular secondary structure making an epitope.

Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count	Helix	Strand	Coil	Count
H2	S0	C2	4	H2	S3	C6	0	H3	S0	C1	1	H3	S3	C5	0
H2	S0	C3	1	H2	S3	C7	0	H3	S0	C2	0	H3	S3	C6	0
H2	S0	C4	1	H2	S3	C8	0	H3	S0	C3	0	H3	S3	C7	0
H2	S0	C5	0	H2	S4	C0	0	H3	S0	C4	0	H3	S3	C8	0
H2	S0	C6	0	H2	S4	C1	0	H3	S0	C5	0	H3	S4	C0	0
H2	S0	C7	0	H2	S4	C2	0	H3	S0	C6	0	H3	S4	C1	0
H2	S0	C8	0	H2	S4	C3	0	H3	S0	C7	0	H3	S4	C2	0
H2	S1	C0	1	H2	S4	C4	0	H3	S0	C8	0	H3	S4	C3	0
H2	S1	C1	0	H2	S4	C5	0	H3	S1	C0	0	H3	S4	C4	0
H2	S1	C2	0	H2	S4	C6	0	H3	S1	C1	0	H3	S4	C5	0
H2	S1	C3	0	H2	S4	C7	0	H3	S1	C2	0	H3	S4	C6	0
H2	S1	C4	1	H2	S4	C8	0	H3	S1	C3	0	H3	S4	C7	0
H2	S1	C5	0	H2	S5	C0	0	H3	S1	C4	0	H3	S4	C8	0
H2	S1	C6	0	H2	S5	C1	0	H3	S1	C5	0	H3	S5	C0	0
H2	S1	C7	0	H2	S5	C2	0	H3	S1	C6	0	H3	S5	C1	0
H2	S1	C8	0	H2	S5	C3	0	H3	S1	C7	0	H3	S5	C2	0
H2	S2	C0	0	H2	S5	C4	0	H3	S1	C8	0	H3	S5	C3	0
H2	S2	C1	1	H2	S5	C5	0	H3	S2	C0	0	H3	S5	C4	0
H2	S2	C2	0	H2	S5	C6	0	H3	S2	C1	0	H3	S5	C5	0
H2	S2	C3	0	H2	S5	C7	0	H3	S2	C2	0	H3	S5	C6	0
H2	S2	C4	0	H2	S5	C8	0	H3	S2	C3	0	H3	S5	C7	0
H2	S2	C5	0	H2	S6	C0	0	H3	S2	C4	0	H3	S5	C8	0
H2	S2	C6	0	H2	S6	C1	0	H3	S2	C5	0	H3	S6	C0	0
H2	S2	C7	0	H2	S6	C2	0	H3	S2	C6	0	H3	S6	C1	0
H2	S2	C8	0	H2	S6	C3	0	H3	S2	C7	0	H3	S6	C2	0
H2	S3	C0	0	H2	S6	C4	0	H3	S2	C8	0	H3	S6	C3	0
H2	S3	C1	0	H2	S6	C5	0	H3	S3	C0	0	H3	S6	C4	0
H2	S3	C2	0	H2	S6	C6	0	H3	S3	C1	0	H3	S6	C5	0
H2	S3	C3	0	H2	S6	C7	0	H3	S3	C2	0	H3	S6	C6	0
H2	S3	C4	0	H2	S6	C8	0	H3	S3	C3	0	H3	S6	C7	0
H2	S3	C5	0	H3	S0	C0	4	H3	S3	C4	0	H3	S6	C8	0

epitope from coil regions (up to 5). There were 3 epitopes comprised of up to 2 helix regions. Since helix regions tend to be long, there tend to be fewer of them (i.e. one or two). An equal number of epitopes was seen with coil-helix and coil-strand combinations. 7 epitopes were seen with all three different types of regions in a single epitope (Figure 4.25). Because of the small dataset, no statistics could be calculated.

4.7 Conclusions

In summary, a detailed structural analysis of B-cell epitopes was performed in terms of epitope structural components, i.e. regions and fragments in 506 unique epitopes. The epitopes from single and multiple chain antigens were examined separately. The results of this study provide sufficient knowledge about the 3D structure of an epitope to guide whether a particular region of an epitope would be able to be used as an immunogen for vaccine design.

According to this analysis, about 95% of B-cell epitopes were found to be conformational. Overall, 90% of epitopes have up to 5 regions (R1–R5) and up to 5 fragments (F0–F5). In terms of length, 94% of regions were seen to be up to 16 residues long, but ranged from 3 to 30 residues. There was no correlation observed between the number of fragments and either the number of regions, the longest region, the total region residues or the average number of regions. If an epitope has a region of up to 14 residues long, there are higher chances of it having other regions that can make part of epitope. Epitopes having regions of between 14 and 23 residues generally have only one region contributing to the epitope structure. Moreover, epitopes having regions with length of up to 14 residues tend mostly to have 2 or 3 regions and 0 to 4 fragments. In terms of epitope size, an average size of 23 and 26 residues was observed in single and multiple chain epitopes respectively.

Conformational analysis of regions forming epitopes informs us that a large proportion of regions are curved or extended. In addition, there is a higher chance than expected of having an epitope contacting one extended and one or two curved regions together (p-value = $1.92\text{x}10^{-5}$). Epitopes with one or two folded regions were more common than expected (p-value = 0). In terms of secondary structure, coiled regions were particularly prevalent in the data. Owing to multiple comparisons tests, the Bonferroni correction was applied to find the significance of independence among the three types of conformational shapes (extended, curved and folded - Table A.4) and secondary structure elements (helix, strand and coil - Table A.5). All the p-values remained significant.

Chapter 5

Molecular Dynamics Simulations of Epitope Regions

Overview

This chapter describes MD simulations of two types of epitope regions: folded (where two α-helices or β-strands are joined by a loop) and extended (linear α-helical structures). In order to explore conformational stability using amino acid substitution mutations in the structure and designing peptide derivatives, the epitope regions have been simulated as both wild type (WT) and mutants. The mutant peptides have been designed either by substitution (the hydrophobic non-contacting residues replaced with alanine and glutamine) or designing peptide derivatives (end-capping, stapling, cyclisation and addition of non-epitope residues at N and C termini). On observing an interesting stabilising effect of a substitution mutation, it has been combined with one or more of the stabilising derivatives. Furthermore, in order to study the association of the stabilising mutant with the antibody, the stabilising mutant and the WT peptides have also been simulated in the presence of antibody.

Introduction

Epitope characterisation into regions and fragments produced a library of peptides (described in Chapter 4). These peptides were classified into three shapes: extended, folded and curved. Extended and folded regions of epitopes are more likely to have proper secondary structure unlike curved regions that are almost exclusively in the forms of loops. Since, the local folding of a polypeptide chain into secondary structure elements (α-helices and β-strands) stabilises the conformation, regions with secondary structure elements have been selected for exploring mutations to stabilise further the native conformation of a region (extracted from the full length antigen protein).

As well as affecting protein folding and stability, mutations are important because they can modify binding/functional sites, stabilising or destabilising interactions [133]. Experimental biologists use site-directed mutagenesis, but this is time consuming, laborious and expensive. The effect of mutations can potentially be predicted using molecular dynamics simulations and MD has been widely used to study protein stability [134]. In order to explore the conformational stability of epitope regions, a wide range of mutations were studied in the selected peptides by using molecular dynamics simulations.

5.1 Simulation Experiments: Mutant Design

For exploring the conformational stability of isolated epitope regions, simulation experiments were planned using five folded and five extended peptides. These peptides were chosen from the epitopes obtained from the shape analysis discussed in Chapter 4. Table 5.1 and 5.2 show information about the selected peptides. These isolated epitope regions contain both antibody contacting and non-contacting

Table 5.1: Folded epitope regions: The peptide sequence is shown with the amino acids contacting antibody coloured in red or orange if hydrophobic. Non-contacting hydrophilic residues are coloured black, while non-contacting hydrophobic residues are shown in blue. These are the targets for making mutations.

PDB	Folded Peptide Sequence	Length	Position	Possible Mutations
4WEB	FKIRMYVGGVEHRLT	15	631-645	3
1ORS	EGHLAGLGLFRLVRLLR	17	107-123	4
4N9G	LSKINDMPITNDQKKLMS	18	68-85	3
4K2U	EKLWEAMLSEHKNNINNCKNI	21	149-169	4
3LHP	NDKAAALGKDKEINWFDISQSLW	23	74-96	3

residues.

Selection of Epitopes Regions for Simulations

The selected folded epitopes are composed of only one region and no fragments where the region is formed of two 2 α-helices or β-strands joined by a loop or turn. Similarly, the extended epitope regions were selected from epitopes with only one extended region and no or few fragments.

Several mutants were designed for each of the selected wild type epitope peptides. An in-house C program (*MutModel*) [135] was used to mutate the side chains of amino acids to be mutated. *MutModel* performs a very simple sidechain replacement using the minimum perturbation protocol (MPP [136]). The sidechain is replaced and then spun around its Chi1 and Chi2 torsion angles to find a position which makes minimal bad contacts.

A number of approaches were followed to explore the design of stable mutants, and are described in detail below.

Table 5.2: Extended epitope regions: The peptide sequence is shown with the amino acids contacting antibody coloured in red or orange if hydrophobic. Non-contacting hydrophilic residues are coloured black, while non-contacting hydrophobic residues are shown in blue. These are the targets for making mutations. The underlined amino acids show non-contacting hydrophilic amino acids that have been mutated as control experiments.

PDB	Linear Peptide Sequence	Length	Position	Possible Mutations
2W9E	GSDYEDRYYRENMHR	15	142-156	2
3EFD	RALHERFDRLERMLDD	16	142-157	6
4M48	YGTNRFSEDIRDMIGFP	17	498-514	3
3P30	HIIYELIEESQKQQEKNEQ	19	640-658	4
1W72	RNMKAHSQTDRANLGTLRGY	20	65-84	4

5.1.1 End Capping

The epitope regions represent an internal section of a protein, and they do not arise from the natural N or C termini of a protein. This means that the peptide ends will have charges which were not there as part of the protein. In such a case, it may be best to block the ends of the extracted peptide, such as by placing an acetyl group at the N-terminus and blocking the C-terminus by an amide group. In addition, the end-capped peptide will be more resistant to breakdown by exopeptidases which degrade the ends.

The selection of peptide end-caps may also depend on the type of study. For example, the ends of α-helices play an important role in helix stabilisation in proteins. Amino acid end capping of short helical peptides in aqueous solution was studied by Forood et al. [137] who observed its role in helix stabilisation. Capping at the N-terminus has a major effect on helix stability whereas capping at the C-terminus has only a minor effect [138]. Several different studies concluded that Asn is the best N-cap, and the stabilizing effect of Asn was found to be equivalent

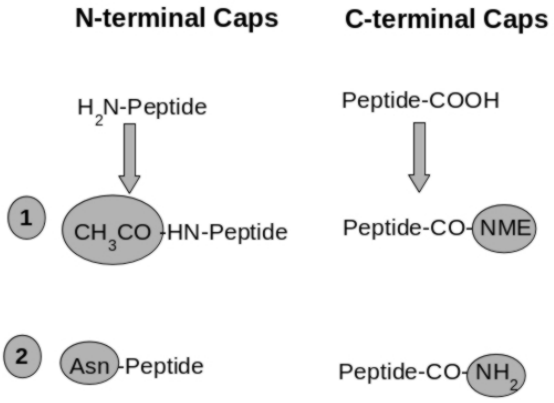


Figure 5.1: Terminal caps used; 1) Acetyl (ACE) group was used as N-cap and methyl amide (NME) was used as C-cap. 2) Asn (N) amino acid was used as N-cap and amide (NH2) was used as C-cap.

to an acetyl group [137–140].

Considering these positive effects of peptide capping, it was decided to use this concept in the computational study of peptides using MD in the hope of stabilising peptides. To this end, two capping schemes were explored; an acetyl (ACE) group as an N-cap and methyl amide (NME) as a C-cap (referred to as 'Cap1') and an Asn as an N-cap and amide group as a C-cap (referred to as 'Cap2').

In order to add Cap1, an extra residue was included at each of the termini and the side chains were mutated to glycine using *MutModel*. The PDB file was then edited manually retaining CA, C and O atom records for the acetyl group at the N-terminus and CA and N for the methyl amide group at the C-terminus. This was followed by the addition of hydrogen atoms (using *Pymol*) giving CH_3CO as the acetyl group and H_3NH as the methyl amide group (Figure 5.1). For Cap2, the extra terminal residues were simply mutated to asparagine and glycine at N and C-terminus respectively. Glycine was manually edited to the amide (NH_2) group as described for Cap1 (Figure 5.1).

5.1.2 Hydrophobic Mutations

In a folded protein, the hydrophobic residues predominantly are found in the core whereas polar residues tend to occur on the surface [141]. An isolated peptide may have hydrophobic amino acids that are buried in the intact protein while exposed on isolation. Consequently, they may result in refolding of the peptides to bring these hydrophobic residues away from solvent.

5.1.2.1 Hydrophobic to Glutamine Mutations

In order to replace hydrophobic residues in isolated peptides with hydrophilic residues, amino acid propensities in alpha helices and beta sheets were considered. Table 5.3 shows the Chou and Fasman amino acid propensity score for alpha helices, beta sheets and turns/coils [142]. It was found that glutamine, being a polar and hydrophilic residue, also has the highest propensity for forming both α-helices and β-sheets. Therefore, all hydrophobic residues not contacting antibody have been mutated to glutamine.

5.1.2.2 Hydrophobic to Alanine Mutations

In small peptides, alanine has a higher tendency to form α-helices [143] and in some studies, alanine was found to have the highest intrinsic preference for the helix interior [144, 145]. Considering these facts, all the non-contacting hydrophobic residues were also mutated to alanine to study their effect on the overall stability of the peptides.

For exploring the stabilising mutations in linear helical regions, these three approaches (end-capping, hydrophobic to alanine and hydrophobic to glutamine) were used. For folded regions, stapling and cyclisation techniques were also investigated.

Table 5.3: Amino acid propensity table. P(a) is propensity of helix forming amino acids, P(b) is propensity of beta sheet forming amino acids and P(turn) is propensity of turn or coil forming amino acids.

P(a)		**P(b)**		**P(turn)**	
Glu	1.51	Val	1.70	Gly	1.56
Met	1.45	Ile	1.60	Asn	1.56
Ala	**1.42**	Tyr	1.47	Pro	1.52
Leu	1.21	Phe	1.38	Asp	1.46
Lys	1.14	Trp	1.37	Ser	1.43
Phe	1.13	Leu	1.30	Cys	1.19
Gln	**1.11**	Thr	1.19	Tyr	1.14
Trp	1.08	Cys	1.19	Lys	1.01
Ile	1.08	**Gln**	**1.10**	Gln	.98
Val	1.06	Met	1.05	Trp	.96
Asp	1.01	Arg	.93	Thr	.96
His	1.00	Asn	.89	His	.95
Arg	.98	His	.87	Arg	.95
Thr	.83	Ala	.83	Glu	.74
Ser	.77	Ser	.75	Ala	.66
Cys	.70	Gly	.75	Phe	.60
Tyr	.69	Lys	.74	Met	.60
Asn	.67	Pro	.55	Leu	.59
Pro	.57	Asp	.54	Val	.50
Gly	.57	Glu	.37	Ile	.47

5.1.3 Stapling/Cyclisation

Stapling is a technique that involves adding covalent bonds near the termini of a peptide when these need to be spatially close to one another. There are several experimental stapling techniques that have been used to enhance the biological performance of peptides [146, 147]. Computationally, we have used the following two techniques to staple the free ends of folded peptides.

5.1.3.1 Disulphide Bond Stapling

The stabilising role of disulphide bonds in protein structures has made it an attractive tool to use for protein engineering (for enhancing stability and activity) [148]. The stabilising ability of disulphide bonds has also been studied in peptides [149]. One of the possible challenges in the process of making disulphide linked peptides

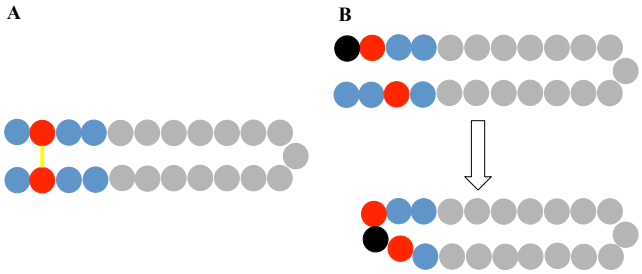


Figure 5.2: Stapling A) disulphide stapling in a folded peptide. The epitope residues are shown grey while non-epitope residues are shown in blue. The cystiene mutations in non-epitope residues are coloured red. The disulphide bond between two cystiene residues is shown in yellow. B) glycine linker in a folded peptide. Non-epitope residues on both termini are shown in blue. The potential positions to add a one residue linker are shown in red while the glycine mutation appears as black. The phi and psi torsion angles of the glycine and other positions outside the epitope are adjusted to create a peptide bond between glycine and one of the potential bonding positions. Extra residues are removed from one of the termini.

is finding sites where cysteine mutations can be introduced where the geometry is optimal for the insertion of a disulphide bridge [150]. An in-house program, *SSSearch* [151], was used that explores potential disulphide bond sites by looking for residues with Cβ distances and Cα distances in specified ranges. The program uses a default distance range of 3.90 - 8.30 and 2.8 - 4.6Å for Cα and Cβ respectively.

SSSearch provides a list of residue pairs with the Cα and Cβ distances between the residues within the required range. Chosen pairs of residues were mutated to cysteines using the in-house program *SSBond* [152] which substitutes cysteines at two sites and performs a conformational search by spinning the Cα-Cβ torsion angles of the two residues in an attempt to optimise the Sγ-Sγ separation (optimum distance 2.03Å).

In the experiments with folded peptides, potential disulphide bond sites were explored outside the original epitope region. For this reason, the peptide was extended by 4-5 residues at the N and C-termini (Figure 5.2A).

5.1.3.2 Cyclisation by Glycine Linker

A folded peptide can be cyclised by the addition of a peptide bond using glycine as a linker between the N and C termini. In the present study, glycine linkers of one or two residues have been used to link the ends of the folded epitope regions covalently. In order to place the glycine linker outside the epitope region, the original peptide was extended by the addition of 3 extra residues at each of the termini. An in-house program, *AddLinker* [153] was used to find positions within a given peptide structure where linkers of 1, 2 or 3 glycines could be placed. This program simply looks for residues with Cα atoms spaced by a distance of 2.7–7.1Å (1 residue linker), 2.6-9.2Å (2 residues linker) and 2.3–9.7Å (3 residues linker). These distance ranges were obtained from a set of high resolution PDB files. A pair of positions outside the epitope that was suitable to place a linker of one (or two) residues was chosen. The adjacent residue distal to the N-terminal identified position was mutated to glycine and other distal residues were removed. Torsion angles were then adjusted to bring the glycine closer to the other residue of the identified pair. It was ensured that the distance between the N atom of the glycine and C atom of C-terminal residue is appropriate to form a peptide bond as a linker (Figure 5.2B).

5.2 Simulation Experiment Methodology: all-atom

For all simulation experiments, the starting structure was either the wild type (or mutant) peptide, a region that forms a part of an epitope (details of epitope regions are described in Chapter 4), or a full length antigen structure from which the selected epitope was extracted, or an antibody-epitope complex. These structures were checked for missing backbone and side chain atoms. In the case of missing atoms, these were added either by *MutModel* [135] or using online tools available at the PDB_HYDRO mutation and solvation server [154] which fixes multiple missing atoms in the given PDB file. The input files were prepared in multiple steps using several GROMACS packages [155–159]. The details of each simulation step and package are described in sections 5.2.1-5.2.4.

There are several force fields available in GROMACS including AMBER, CHARMM, GROMOS and OPLS. Over the course of time, these force fields have been revised by modification of torsion potentials associated with dihedral angles and side chains [160]. For example, the original AMBER force field has gone through several improvements that include modifications in backbone potentials resulting in the Amber ff99SB* force field [161, 162] and modifications in side chain potentials for four types of amino acids resulting in the Amber ff99SB-ILDN force field [163]. Later, the side chain modifications (ILDN) were combined with the backbone potential modifications (ff99SB*) to produce an optimised force field, ff99SB*-ILDN [77]. Lindorff-Larsen and colleagues [160] evaluated eight force fields for the comparison of experimental data and molecular dynamics simulation. They concluded that the peptide folding results of ff99SB*-ILDN and CHARMM22* are in good agreement with the experimental data. For this reason, the Amber ff99SB*-ILDN force field was chosen to perform the molecular dynam-

ics simulations in this study. CHARMM22 was rejected because it generally tends to over stabilise helices [160, 164] and most of the selected peptides in this study were α-helical peptides. The recommended water model for the Amber ff99SB*-ILDN force field is TIP3P [165] .

5.2.1 Peptide Preparation: Topology and Box Creation

A GROMACS tool, *pdb2gmx* was used to generate a topology file for a given structure. The topology file contains molecular information about bonded (bonds, angles, and dihedrals) and non-bonded parameters (atom types and charges) taken from the chosen forcefield. This tool converts a PDB file to a GROMACS .gro file. A position restraint file (.itp) is also generated to hold the heavy atoms in place during equilibration (pressure and temperature coupling). Another tool, *editconf*, was used to create a triclinic box around the peptide. The distance to the edge of box was kept at 1.0 nm to prevent the interaction of protein with its periodic image. The box was filled with the solvent by using the tool *solvate*. A generic equilibrated TIP3P 3-point solvent model (spc216) was used. The peptide in a water box may have a positive or negative charge that needs to be neutralized and counter-ions were added using the tools *grompp* and *genion*.

5.2.2 Energy Minimization

Inappropriate structural geometry and clashes may result in high energies during dynamics. Therefore, any such issues must be resolved by performing energy minimization (EM) of the peptide before starting dynamics. The steepest decent algorithm (with 50000 steps) was used, with a maximum step-size of 0.01 nm and a maximum force of 1000 kJ/mol/nm. For the treatment of long range electrostatic and Van der Waals interactions, particle-mesh Ewald (PME) [166] and twin range

cut-offs were used with a cut-off distance of 1 nm. For neighbour searching, the Verlet cutoff-scheme was selected and xyz periodic boundary conditions were chosen.

5.2.3 Equilibration

In order to place the solvent and ions around the protein uniformly, the system must be equilibrated at a desired temperature and pressure. This was done in two stages, 1) temperature coupling and, 2) pressure coupling. For temperature coupling, an NVT ensemble (with constant number of particles, volume, and temperature) was simulated for 200 ps to obtain the desired temperature of 310 K. In the next stage, pressure coupling was performed for another 200 ps with an NPT ensemble (with a constant number of particles, pressure, and temperature) to reach the desired pressure of 1 bar. The conformation of the peptide was held fixed using explicit position restraints.

Bond constraints were applied using the LINCS algorithm [79] and the Berendsen thermostat [167], which couples the system to an external bath, was used for temperature coupling. For pressure coupling, the Parrinello-Rahman barostat [168], which couples the system to an external bath by letting the volume and shape of the simulation box fluctuate, was used. Initial velocities were generated using a random seed generator. Another option for the bond constraints algorithm is SHAKE [169] which is known for its accuracy, but which is too slow because it resets bonds to defined values one bond at a time. The LINCS algorithm is three to four times faster than SHAKE as it uses Lagrange multipliers for constraint force modelling.

5.2.4 Production MD

Once the system was equilibrated, the production run for MD was executed. Unlike the equilibration phases, the position restraints from the peptide were released to allow free dynamics. The parameters for the production phase were essentially similar to the equilibration phase except for the temperature coupling ensemble where the Berendsen thermostat was replaced by the Nose-Hoover thermostat. This was done to produce a better kinetic ensemble. The Berendsen thermostat is very efficient in relaxing the system to the desired temperature but once the system has reached equilibration then the exploration of the correct canonical ensemble becomes more significant [170, 171]. During the MD production, a time step of 2 fs was used and conformations were recorded every 10 ps.

In order to prepare the input files automatically (Section 5.2.1–5.2.3) for production runs, an in-house script (*doitGROMACS.sh*) [172] was used. All the simulations were run on a node of the GPU based cluster, EMERALD [173], consisting of two 6-core X5650 Intel Xeons and three or eight 512-core M2090 NVIDIA GPUs. On the EMERALD cluster, a typical run time for peptide simulation was 180-200 ns/day whereas it was 45-50 ns/day for peptide-antibody complex simulation.

5.2.5 Analysis

The major trajectory processing takes account of correct periodicity and removal of water. Multiple bash scripts were written to automate the post-processing of a large number of trajectories. In molecular dynamics, it is a common practice to use RMSD as a function of time to measure protein stability, but this measure was found to be misleading for these small peptides owing to flexible ends. Therefore, it was decided to compute the secondary structure assignments for each of the amino

acids during the simulation and calculate the percentage of time each amino acid spent in the initial conformation. This was done as follows:

Step 1: Each frame (conformation recorded every 10 ps) of the trajectory was extracted and passed to the in-house program *pdbsecstr* that uses an implementation of the Kabsch and Sander [122] method as modified by Smith and Thornton (unpublished results) to be less strict at the ends of secondary structure elements. The secondary structure assignments were calculated for each frame of the trajectory and stored in a text file.

Step 2: The secondary structure assignments from the text file were used to compute the percentage of time each of the residues spent in its initial conformation. This was calculated using Equations 5.1 and 5.2:

$$S = \sum_{i}^{N} s_i / N \tag{5.1}$$

$$s_i = \frac{n_{0,i} \times t_s \times 100}{t_T} \tag{5.2}$$

where S is the overall stability; s_i is the stability of residue i, $n_{0,i}$ is the number of times residue i is seen in the original conformation during the simulation with time step t_s (10 ps), t_T is the total time in ps and N is the number of amino acids. Values of s_i for each residue and S for the peptide were saved in a text file for further analysis.

Thus, secondary structure has been used as a measure of structural stability. Most of the available methods [174–176] for secondary structure assignments use the notation of H or h for α-helix, G or g for 3-10 helix, I or i for π-helix, E or e for β-strand, B for isolated β-bridge, T for turn, S for bend and C for coil, the lower

case assignment being the less perfect versions allowed by Smith and Thornton. Because the calculation of secondary structure is very sensitive, some flexibility was allowed. Thus, H, h, g, G, i and I (helix) assignments were considered equivalent, as were E, e, B and b (strand B/b being used to indicate the direction of a bridge) and C, S and T (unstructured). Additional flexibility was allowed at the junctions of ordered secondary structures and unstructured regions. Amino acids with borderline helix (h) or strand (e) initial conformations were grouped with unstructured as well as with helix or strand. For example, a residue initially as 'h' was regarded as maintaining its conformation if it changed to H, G, g, I and i (helical assignments), or to C, S or T (unstructured assignments).

Statistical tests and plots were performed with GraphPad Prism (Version 7.00 for Mac, GraphPad Software, La Jolla California USA, www.graphpad.com).

Antibody-Epitope Complex Simulation

In order to study the behaviour of an epitope region in the presence of an antibody (whether the WT and the mutant maintains the binding association), the wild type and stabilised mutant were simulated in the presence of antibody (light and heavy chain). Some additional processing of this complex was needed before performing molecular dynamics. GROMACS cannot handle the insertions in antibody numbering (e.g. 30A,30B,30C,30D,30E and 30F), so, light and heavy chains were renumbered before creating the topology file. The disulphide bonds between light and heavy chains were made using the GROMACS tool, *pdb2gmx*. The rest of the simulation methodology was the same.

Initially, constraints were applied to the antibody chains to hold the atoms static and only allow the peptide to move during the production MD. It was thought that by doing so, computational time would be saved. However, the LINCS algorithm

cannot be used in combination with the freeze-groups (freezegrps) option. Having no constraints on bond lengths means that the time step for simulation needs to be reduced from 2 fs to 0.5 fs which, consequently, increases the computational time by up to a factor of 5. Therefore, the antibody and epitope complex was simulated without freezing the antibody.

Stapled Peptides Simulation

For disulphide bond stapled peptides, GROMACS needs to be informed explicitly about the location of disulphides. However, GROMACS has the builtin functionality of making disulphide bonds using *pdb2gmx* (with flag -ss and -ter) which requires the two sulphur atoms to be within a certain distance and tolerance of up to 10% in order to be identified as a disulphide. The distance between the sulphur atoms is saved in a file, *specbond.dat* (present in the working directory), and can be edited depending on how far the sulphur atoms are in the structure.

For cyclised peptides (with a glycine linker), the *specbond.dat* file requires the distance between the N atom of the glycine and C atom of the terminal residue to link the residues. At this stage, the toplogy file needs manual editing to remove the two hydrogen atoms from the N terminus leaving NH (previously NH3) and renaming the terminal atoms; namely N3 to N (at the N-terminus) and O2 to O (at the C-terminus). The connections of these removed hydrogen atoms with bonds, pairs, angles and dihedrals also need to be removed from the topology file followed by its renumbering. The renumbering of the topology file was performed using a bash script *renumtop.sh* which was obtained from the GROMACS online archive (`www.gromacs.org/Downloads/User_contributions/Other_software/gmx_top_tools.tgz`). The rest of the simulation methodology was the same.

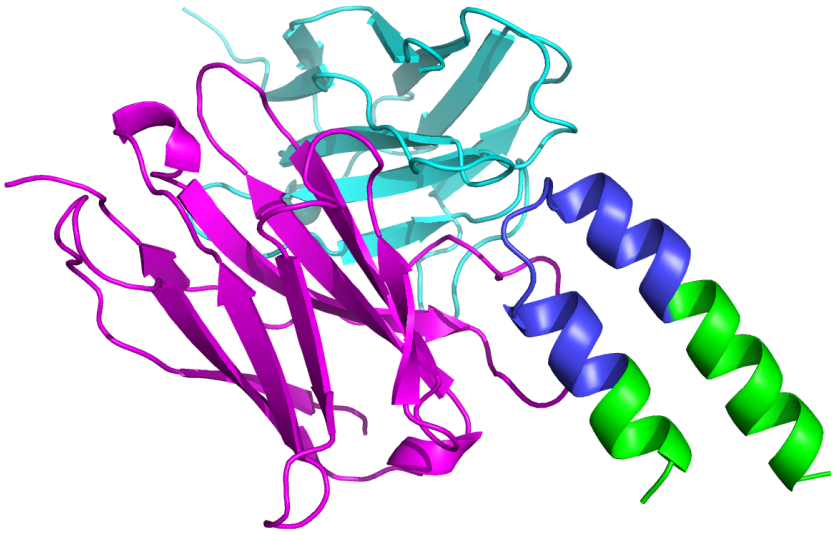


Figure 5.3: 4N9G – Crystal structure of a computationally designed RSV-presenting epitope scaffold and its elicited antibody 17HD9. Light (cyan) and heavy (violet) chains bound with full length epitope scaffold. The epitope (at positions 68-85) mapped on the basis of antibody-antigen contacts is shown in blue.

5.3 Results

5.3.1 Molecular Dynamics Simulation of Folded Regions

5.3.1.1 4N9G – Helix-turn-Helix Epitope

The epitope was identified in the antibody-antigen complex 4N9G [177] and extracted from the antigen (Figure 5.3). The epitope is an α-helix-turn-helix, comprised of 18 residues. Of these, 13 residues make direct contacts with antibody. In the remaining 5 non-contacting amino acids, 2 are hydrophilic while the remaining 3 are hydrophobic. These three non-contacting, hydrophobic amino acids provide sites for alanine and glutamine mutations (Table 5.1). In addition, peptide derivatives have been designed by end-capping, disulphide bond stapling, cyclisation by glycine linker and extension of the WT termini. Along with the WT epitope, a total of 21 peptides were simulated for 500 ns (Table 5.4). The simulations were repeated three times to produce replicates. This resulted in 66 simulation trajectories

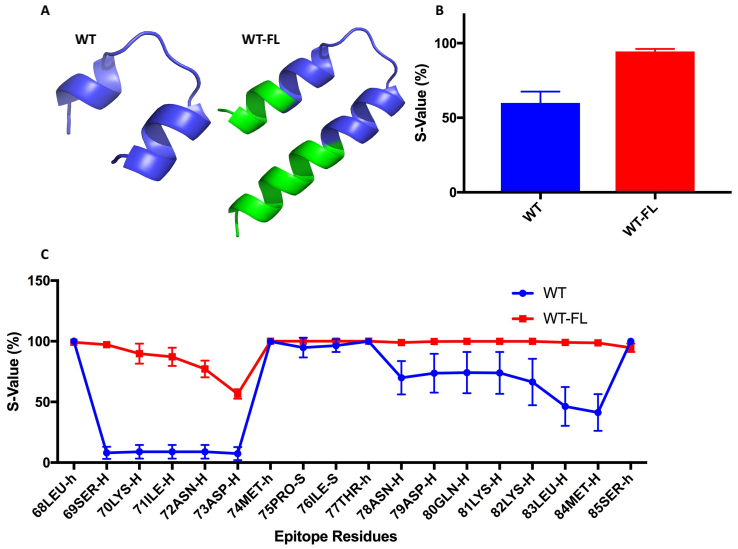


Figure 5.4: (A) 4N9G – The WT epitope (blue) in the full length protein (WT-FL), (B) The S-value of the WT peptide (∼60% and WT-FL (∼94%), (C) The s_i-value of each residue in the epitope during 500 ns simulations (3 replicates). A Welch's t-test provided a p-value = 0.01 for WT and WT-FL which suggests that the epitope is significantly more rigid in the full length antigen protein. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

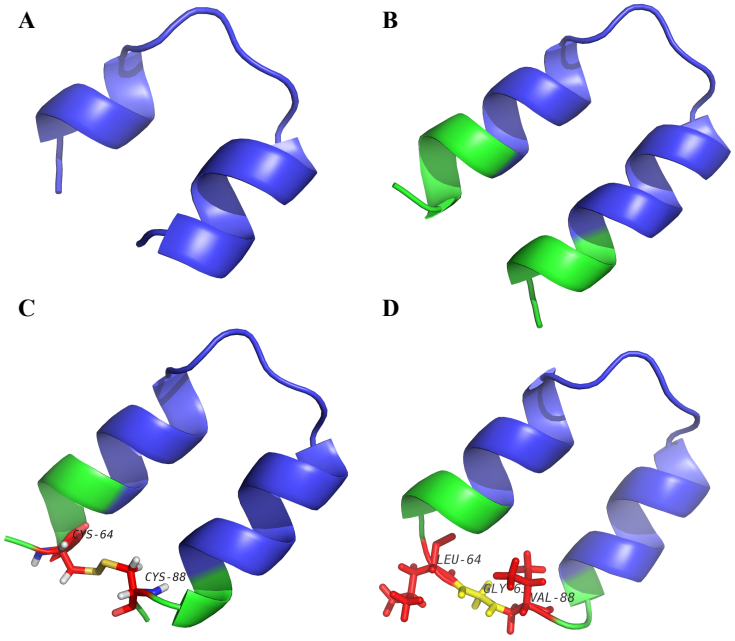


Figure 5.5: 4N9G – helix-turn-helix epitope structure **(A)** The WT peptide at positions 68-85 in the antigen. **(B)** The WTX peptide at positions 63-90 – extended at both ends by 5 residues. The epitope is shown in blue and the extended termini in green. **(C)** The WTSS peptide with a disulphide bond between the 2 cysteines at positions 64 and 88 (sticks); the epitope (blue) with extended ends (green). **(D)** The WTG peptide with glycine linker (yellow sticks) added between amino acids L64 and V88 (red sticks).

which were investigated for secondary structure variation over the period of 500 ns. For all the peptides, the S and s_i stability parameters (Equations 5.1 and 5.2) were calculated and the S-values are shown in Table 5.4.

In order to estimate the level of rigidity of the WT epitope region in the full length antigen chain (RSV-presenting epitope scaffold), the full length antigen was simulated for 500 ns (3 replicates). An S-value of $\sim$94% was found for the epitope (positions 68-85) which provides an insight into the stabilty of epitope when it is part of the intact antigen compared with an S-value of $\sim$60% when the epitope region was taken out of the full length structure (see Figure 5.4). Thus, the aim of stabilising mutations is to try to push the S-value towards 94%.

Table 5.4: 4N9G – Helix-turn-helix epitope (mutant and derivative peptides). Th peptides were simulated for 500 ns (3 replicates).The S-value for each peptide represents the average time it has spent in its initial conformation. The mean of S-value is computed for 3 replicates in each peptide. Bold numbers show the peptides with the interesting results (s_i values for these is shown in Figure 5.6).

Caps		**Alanine Mutations**			**Glutamine Mutations**			**Stapling (WTSS)**		**Linker (WTG)**	**Extended Ends (WTX)**	**S-Value**
Ace/NME	ASN/NH2	I71A	L83A	M84A	I71Q	L83Q	M84Q	L64C:V88C	L64C:L89C	L64:Gly:V88	R63-K90	
-	-	-	-	-	-	-	-	-	-	-	-	**59.95**
✓	-	-	-	-	-	-	-	-	-	-	-	73.42
-	✓	-	-	-	-	-	-	-	-	-	-	60.41
-	-	✓	-	-	-	-	-	-	-	-	-	**71.06**
-	-	-	✓	-	-	-	-	-	-	-	-	63.98
-	-	-	-	✓	-	-	-	-	-	-	-	64.86
-	-	✓	✓	✓	-	-	-	-	-	-	-	49.89
-	-	-	-	-	✓	-	-	-	-	-	-	63.89
-	-	-	-	-	-	✓	-	-	-	-	-	57.94
-	-	-	-	-	-	-	✓	-	-	-	-	54.57
-	-	-	-	-	✓	✓	✓	-	-	-	-	52.37
-	-	-	-	-	-	-	-	✓	-	-	-	81.07
✓	-	-	-	-	-	-	-	✓	-	-	-	82.70
-	✓	-	-	-	-	-	-	✓	-	-	-	81.56
-	-	-	-	-	-	-	-	-	✓	-	-	**87.14**
✓	-	-	-	-	-	-	-	-	✓	-	-	82.77
-	✓	-	-	-	-	-	-	-	✓	-	-	**90.55**
-	-	✓	-	-	-	-	-	-	✓	-	-	**84.73**
✓	-	✓	-	-	-	-	-	-	✓	-	-	84.37
-	✓	✓	-	-	-	-	-	-	✓	-	-	**87.35**
-	-	-	-	-	-	-	-	-	-	✓	-	80.17
-	-	-	-	-	-	-	-	-	-	-	✓	77.26

Table 5.4 shows that Cap1 stablised the WT conformation by about 13% (S $\approx$ 73%) whereas no effect was seen with Cap2 (S $\approx$ 60%). Among all the alanine mutations, I71A stablised the conformation by nearly 10% (S $\approx$ 71%) whereas glutamine mutations only destabilised the structure although the destabilising effect of both the glutamine mutants was not found to be statistically significant (p-value $>$ 0.05).

In order to study the effect of stapling and the addition of a linker, the WT peptide was extended at both termini by inclusion of 5 non-epitope residues. The additional ends of the peptides were searched for potential disulphide bond positions and these positions were mutated to cysteine (as explained in Section 5.1.3.1) to form a disulphide bond. Two possible pairs of residues (L64:V88 and L64:L89) were chosen to introduce a disulphide bond. The stapled peptides were also end-capped to explore any effect of capping on these. Furthermore, the WT peptide was cyclised by the addition of a single residue glycine linker. A pair of residues (L64:V88), in the extended ends, was identified as a potential target for the insertion of a glycine linker (as explained in Section 5.1.3.2). Figure 5.5 shows the WT epitope, WT with extended termini (WTX), WT with disulphide bond (WTSS) and WT with a glycine linker (WTG).

The simulation data show that stapling the ends of peptides with a disulphide bond has stabilised the peptide to a considerable extent. An increase of $\sim$6% was seen when the disulphide stapling was between C64:C89 (WTSS2 S $\approx$ 87%) compared with C64:C88 (WTSS1 S $\approx$ 81%). End-capping on the C64:C88 stapled peptide did not show any remarkable improvement while Cap2 on the C64:C89 stapled peptide increased the S-value by nearly 3% (S $\approx$ 91% approaching the full-length stability, S $\approx$ 94%). WTG also shows an increase in stability (S $\approx$ 80%),

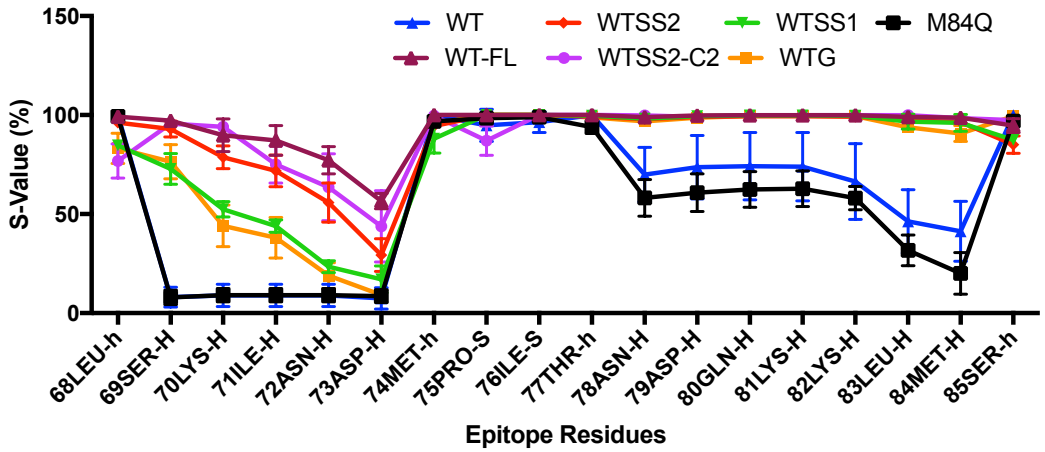


Figure 5.6: 4N9G – Residue level stability during 500 ns simulation (three replicates). WTSS1 refers to a disulphide bond at L64C:V88C and WTSS2 at L64C:L89C positions. C2 refers to the Asn/NH2 cap (cap2). WTG label is used for the WT, cyclised using a glycine linker. WT-FL refers to the epitope region with in the full length antigen protein. In order to use a bigger sample for statisics, the s_i values of 3 replicates were used to compute the p-value. A p-value < 0.0001 was calculated for all the pairs except WT/M84Q. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

but much lower than WTSS2. Surprisingly, WTX was also significantly stabilised (S $\approx$ 77%) compared with WT. A Welch's t-test was performed to test the significance of stability enhancement and a p-value < 0.05 was observed for WT/WTSS1, WT/WTSS2, WT/WTG and WT/WTX pairs which suggests that the stapling, cyclisation and extension of the WT's termini have significantly stabilised the native conformation. Having seen the stablising effect of I71A (S $\approx$ 71%) and WTSS2 (S $\approx$ 87%) separately on the WT epitope (S $\approx$ 60%), a combined mutant of I71A and C64:C89 disulphide bond was produced and end-capped. It was expected that stability might improve by combining two stablising mutants, but its S-value did not exceed than the S-value of WTSS2. An increase of 3% in the S-value was seen with Cap2 which is similar when WT was stapled in the presence of Cap2. The increase in stability of mutant (I71A + SS2) was statistically significant (p-value = 0.02).

In order to examine the behaviour of each of the peptide residues during simulation, the percentage of time for each of the residues that it had spent in the initial

conformation (s_i in Equation 5.2), was computed. Figure 5.6 shows that the first helix in the WT α-helical peptide (residues 69-73) did not maintain its conformation during the simulation whereas the second helix was found to be considerably more stable. It is clear that disulphide bond stapling and cyclisation have stabilised the first helix as well as further stabilised the second helix. The maximum improvement in the stability of the first helix was seen when the peptide was stapled at the L64C:L89C positions along with Cap2 (Asn/NH2 cap). As mentioned earlier, the glutamine mutations destabilised the peptide, and M84Q is shown as an example in Figure 5.6.

Epitope Simulation in the Presence of Antibody

It was expected that the WT peptide would bind with antibody, but a mutation in the structure may prevent this binding. Therefore, the WT (S $\approx$ 60%) and the most stablilised mutant, WTSS2-C2, (S $\approx$ 90%) were simulated in the presence of antibody for 500 ns. The trajectories were subjected to visual analysis and the S-values were computed for both the peptides when they were simulated with the antibody (Figure 5.7). It was observed that the WT and WTSS2-C2 peptides not only maintained the binding with the antibody and did not fall off, but also showed an increase in the S-value compared with the free peptides: WT (S $\approx$ 73% compared with S $\approx$ 60% for free peptide) and WTSS2-C2 (S $\approx$ 93% compared with S $\approx$ 91%)

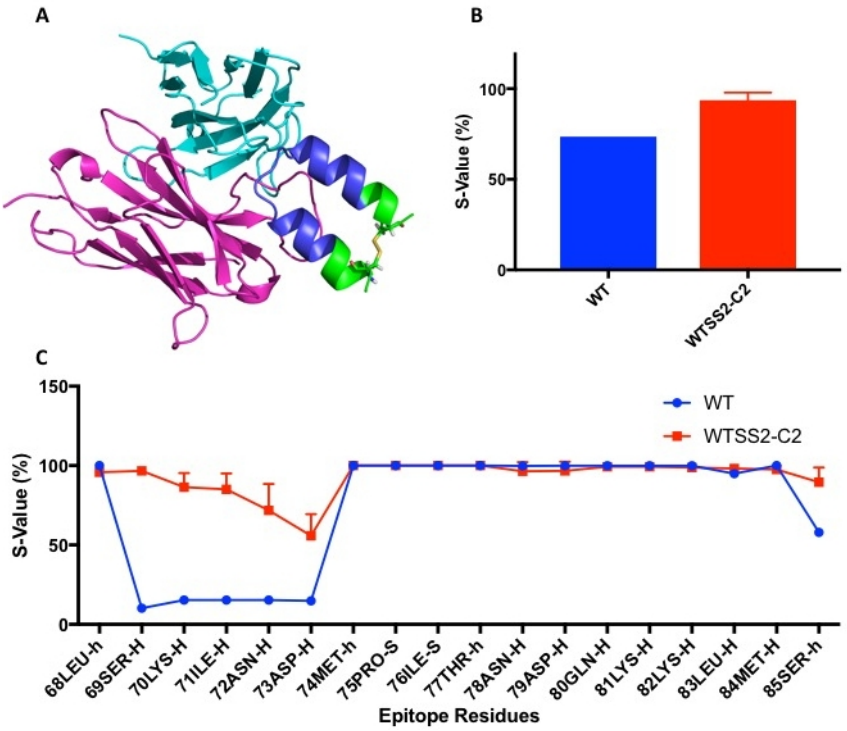


Figure 5.7: (A) 4N9G − The antibody/WTSS2-C2 complex simulated for 500 ns (3 replicates) (B) The S-value of the WT peptide (∼73%) and WTSS2-C2 (∼93%), (C) The s_i-value of each residue in the epitope during 500 ns simulations (WT simulated once and WTSS2-C2 simulated 3 times). The explanation of secondary structure labels on x-axis is given in section 5.2.5.

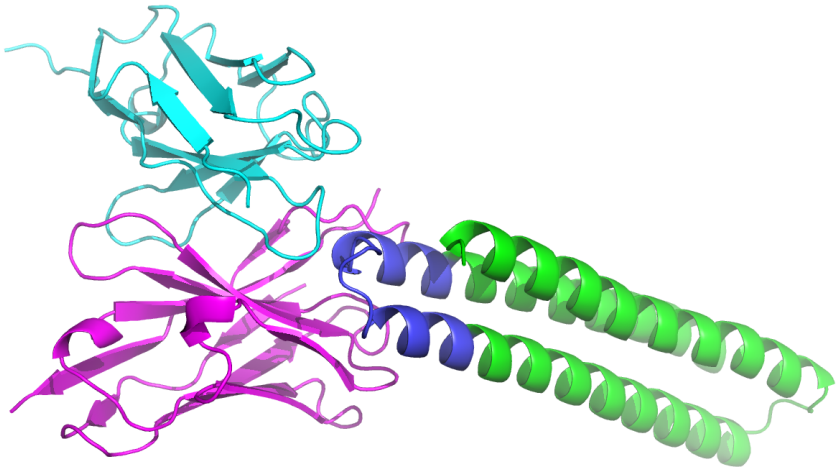


Figure 5.8: 3LHP – Crystal structure of HIV epitope-scaffold 4E10 Fv complex. Light (cyan) and heavy (violet) chains bound with full length epitope scaffold. The epitope (at positions 74-96) mapped on the basis of antibody-antigen contacts is shown in blue.

5.3.1.2 3LHP – Helix-turn-Helix Epitope

In order to study similar types of mutations in another α-helix-turn-helix epitope, the epitope was identified on the antibody-antigen complex structure 3LHP [178] and was found to be comprised of 23 residues (Figure 5.8). Of these, 15 residues were found to be contacting antibody and therefore cannot be mutated. Of the remaining 8 non-contacting residues, three were hydrophobic and candidates for alanine and glutamine substitutions (Table 5.1). Two of these are already alanine so only one of the non-contacting amino acids was mutated to alanine while all three were mutated to glutamine. Peptide derivatives were also designed to study the effect of end-capping, disulphide stapling, cyclisation by glycine linker and extended termini. In order to staple the WT peptide, both of the termini of the WT were extended by 4 residues, and the residues at positions A73 and V98 were found to be suitable sites for the substitution of cysteines. In addition, the WT peptide was cyclised by the addition of a glycine linker between G70 and V98 (Figure 5.9). A total of 14 peptides (WT and 13 mutants and derivatives) were studied by molecular dynamics. Each of the peptides was simulated for 500 ns and three replicates were

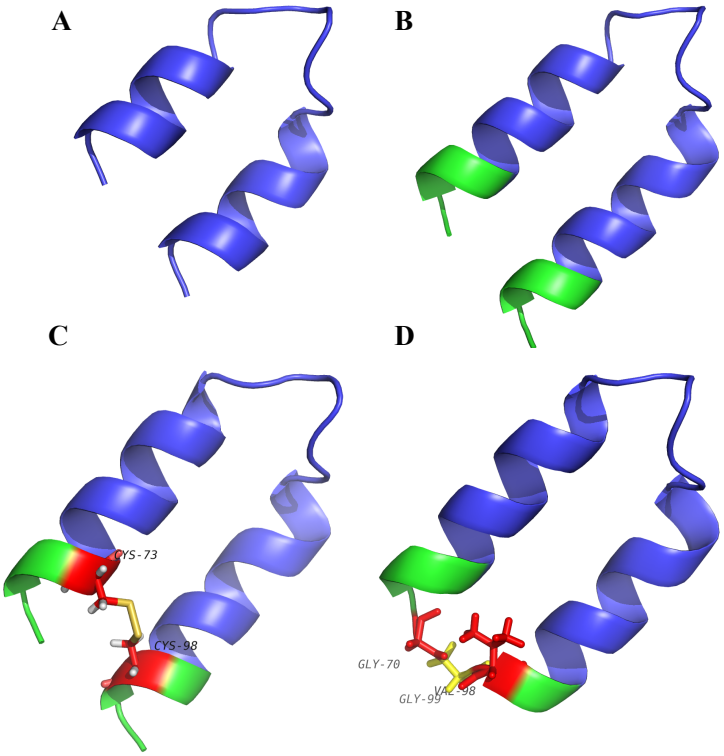


Figure 5.9: 3LHP – helix-turn-helix epitope structure. **(A)** The WT peptide at positions 74-96 in the antigen. **(B)** The WTX peptide at positions 70-100 – extended at both ends by 4 residues. The epitope is shown in blue and the extended termini in green. **(C)** The WTSS peptide with a disulphide bond between the 2 cysteines at positions 73 and 98 (sticks); the epitope (blue) with extended ends (green). **(D)** The WTG peptide with glycine linker (yellow sticks) added between amino acids G70 and V98 (red sticks).

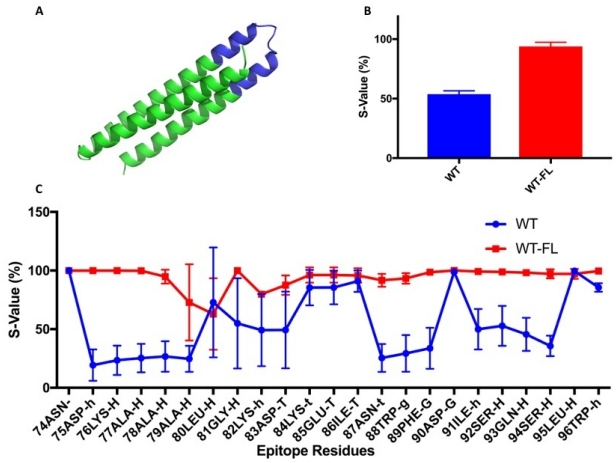


Figure 5.10: (A) 3LHP – The WT epitope (blue) in the full length protein (WT-FL), (B) The S-value of the WT peptide (53%) and WT-FL (93%), (C) The s_i-value of each residue in the epitope during 500 ns simulations (3 replicates). The epitope was found to be significantly more rigid and stable in the full length antigen protein (p-value = 0.0001). The explanation of secondary structure labels on x-axis is given in section 5.2.5.

generated, and the 42 resulting trajectories were analysed for stability.

The full length antigen was simulated for 500 ns (3 replicates) and the S-value was computed for the epitope (at positions 74-96). An average S-value of about 93% was observed which suggests that in full length antigen, the epitope is not very flexible and stays more stable than the epitope when it is taken out of the antigen where it has an S-value of 53% (Figure 5.10).

Table 5.5: 3LHP – Helix-turn-helix epitope (mutant and derivative peptides). Th peptides were simulated for 500 ns (3 replicates).The S-value for each peptide represents the average time it has spent in its initial conformation. The mean of S-value is computed for 3 replicates in each peptide. Bold numbers show the peptides with the interesting results (s_i values for these is shown in Figure 5.11).

Caps		**Alanine Mutations**	**Glutamine Mutations**			**WTSS**	**WTG**	**WTX**	**S-Value**
Ace/NME	ASN/NH2	L80A	A77Q	A79Q	L80Q	A73C:V98C	G71:Gly:V98	G70-K100	
-	-	-	-	-	-	-	-	-	**53.75**
✓	-	-	-	-	-	-	-	-	58.83
-	✓	-	-	-	-	-	-	-	53.80
-	-	✓	-	-	-	-	-	-	53.92
-	-	-	✓	-	-	-	-	-	53.08
-	-	-	-	✓	-	-	-	-	49.60
-	-	-	-	-	✓	-	-	-	56.38
-	-	-	✓	✓	✓	-	-	-	53.23
-	-	-	-	-	-	✓	-	-	**79.14**
✓	-	-				✓	-	-	**79.98**
-	✓	-				✓	-	-	73.14
-	-	-	-	-	-	-	✓	-	**81.51**
-	-	-	-	✓	-	-	✓	-	62.01
-	-	-	-	-	-	-	-	✓	55.07

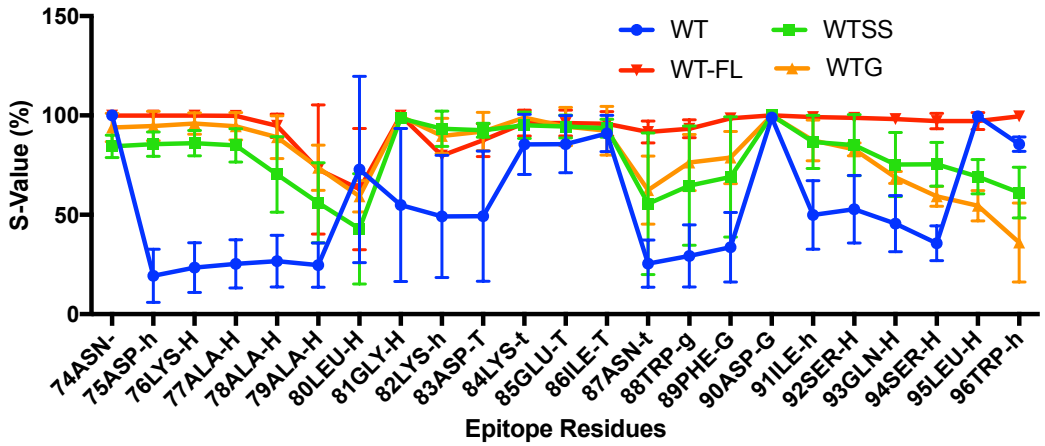


Figure 5.11: 3LHP – Helix-turn-helix mutant peptides simulated for 500 ns simulations (3 replicates). The s_i value represents the time each residue in the peptide has spent in its initial conformation. The s_i values of WT, WTSS, WTG and WT-FL (epitope in full length antigen) are shown. WTSS refers to the peptide stapled with a disulphide bond at positions A73C:V98C. The peptide with a glycine linker is labelled WTG. A welch's t-test resulted in a p-value < 0.0001 when s_i values from 3 replicates were used as a sample for t-test. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

From Table 5.5, the WT peptide spent about 53% of its time in its initial conformation during 500 ns simulations. Capping the WT peptide did not help in maintaining the structure. Alanine and glutamine mutations did not stabilise or destabilise the conformation. The decrease in the S-value of mutant A70Q suggests a destabilising effect of the mutation, but it was not found to be statistically significant. A significant increase in stability was seen in the peptides that were stapled (WTSS) and cyclised (WTG) (Welch's t-test p < 0.0001 Figure 5.11). End-capping of the stapled peptide did not have any significant effect on stability. The WT with extended termini (WTX) had no effect on the conformation (p-value = 0.7). The destabilising mutation A79Q was combined with stabilising mutant WTG and it was found that the cyclisation could not overcome the effect of the mutation (p-value = 0.09).

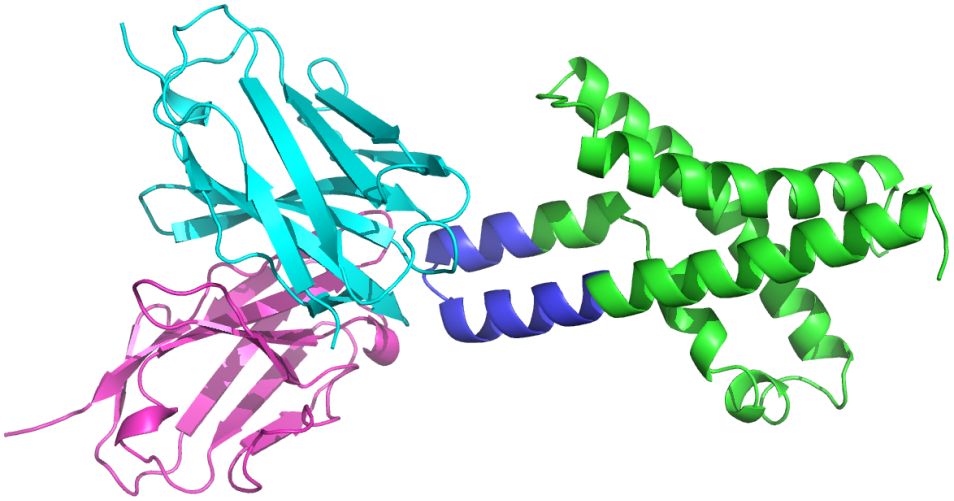


Figure 5.12: 1ORS– X-ray structure of the KvAP potassium channel voltage sensor in complex with an Fab. Light (cyan) and heavy (violet) chains bound with full length antigen potassium channel. The epitope (at positions 107-123) mapped on the basis of antibody-antigen contacts is shown in blue.

5.3.1.3 1ORS – Helix-loop-Helix Epitope

This epitope was mapped on the antigen, KvAP potassium channel voltage sensor (Figure 5.12) [179] and is comprised of 17 residues; 10 contacting and 7 non-contacting residues. Of these non-contacting residues, 4 were hydrophobic and potential sites for mutations (Table 5.1). The 500 ns simulation replicates were only produced for WT, stapled, cyclised and extended termini peptides whereas the alanine and glutamine mutant and the end-capped WT peptides were simulated only once. A total of 18 peptides (the WT and 17 mutants) produced 30 simulation trajectories which were analysed for conformational stability.

The level of epitope rigidity was studied in the full length antigen and it was found that the epitope region retained its initial conformation for about 97% of the simulation time which suggests that the epitope is not flexible in the full length protein. The WT epitope region has an S-value of about 52% and was therefore quite unstable on its own (Figure 5.13).

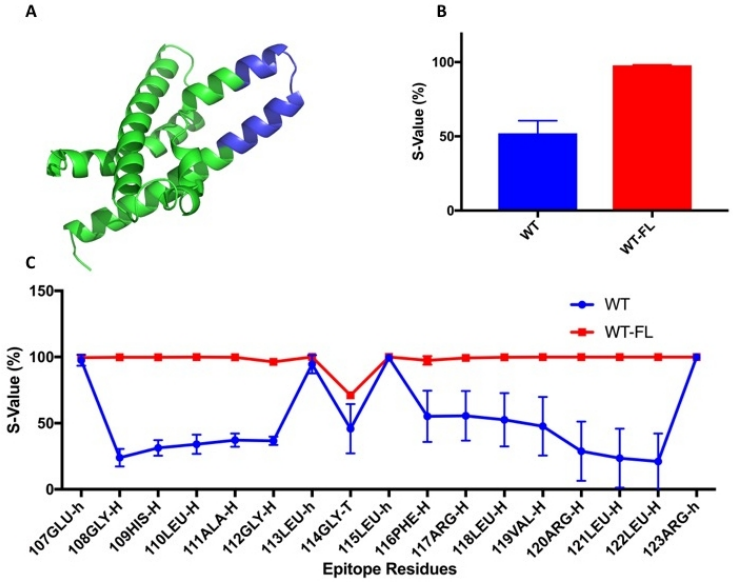


Figure 5.13: (A) 1ORS – The WT epitope (blue) in the full length protein (WT-FL) (B) The S-value of the WT peptide (52%) and WT-FL (97.81%) (C) The s_i-value of each residue in the epitope during 500 ns simulations (3 replicates). A p-value = 0.01 for WT/WT-FL suggests that the epitope in WT-FL was rigid compared with the WT. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

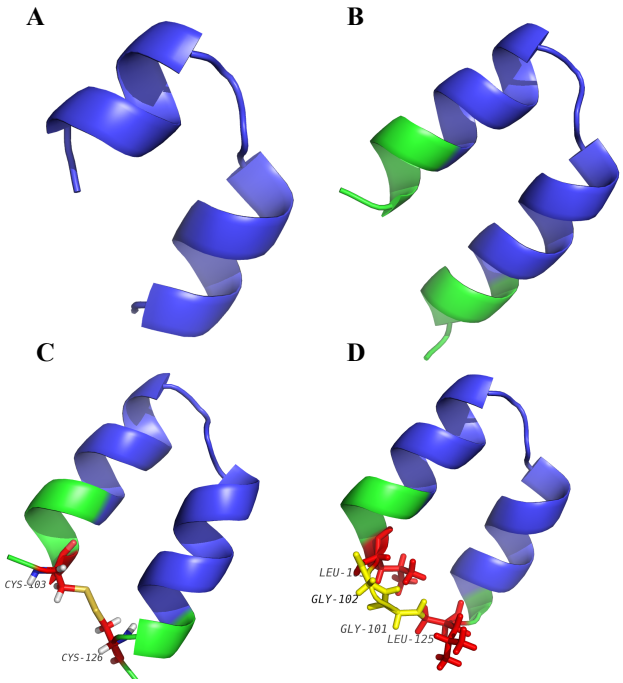


Figure 5.14: 1ORS – Helix-loop-helix structure. **(A)** The WT peptide at positions 107-123 in the antigen. **(B)** The WTX peptide at positions 102-128 – extended at both ends by 5 residues. The epitope is shown in blue and the extended termini in green. **(C)** The WTSS peptide with a disulphide bond between the 2 cysteines at positions 103 and 126 (sticks); the epitope (blue) with extended ends (green). **(D)** The WTG peptide with a glycine linker (yellow sticks) added between amino acids L103 and L125 (red sticks).

Table 5.6: 1ORS − Helix-loop-helix epitope (mutant and derivative peptides). Th peptides were simulated for 1000 ns.The S-value for each peptide represents the average time it has spent in its initial conformation. The mean of S-value is computed for 3 replicates in each peptide. Bold numbers show the peptides with the interesting results (s_i values for these is shown in Figure 5.15.

Caps		**Alanine Mutations**				**Glutamine Mutations**				**WTSS**		**WTG**	**WTX**	**S-Value**
Ace/NME	ASN/NH2	L110A	L118A	L121A	L122A	L110Q	L118Q	L121Q	L122Q	L103C:L125C	L103C:L126C	L103:Gly(2):L125	L102-L128	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	**52.04**
✓	-	-	-	-	-	-	-	-	-	-	-	-	-	61.73
-	✓	-	-	-	-	-	-	-	-	-	-	-	-	68.70
-	-	✓	-	-	-	-	-	-	-	-	-	-	-	49.51
-	-	-	✓	-	-	-	-	-	-	-	-	-	-	46.75
-	-	-	-	✓	-	-	-	-	-	-	-	-	-	51.59
-	-	-	-	-	✓	-	-	-	-	-	-	-	-	53.59
-	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	56.49
-	-	-	-	-	-	✓	-	-	-	-	-	-	-	54.08
-	-	-	-	-	-	-	✓	-	-	-	-	-	-	43.70
-	-	-	-	-	-	-	-	✓	-	-	-	-	-	37.57
-	-	-	-	-	-	-	-	-	✓	-	-	-	-	34.53
-	-	-	-	-	-	✓	✓	✓	✓	-	-	-	-	45.32
-	-	-	-	-	-	-	-	-	-	✓	-	-	-	**85.10**
-	-	-	-	-	-	-	-	-	-	-	✓	-	-	**97.39**
-	-	-	-	-	-	-	-	-	-	-	-	✓	-	**93.0**
-	-	-	-	-	-	-	-	-	-	-	-	-	✓	**97.51**
-	-	-	-	-	-	✓	-	-	-	-	✓	-	-	**76.16**

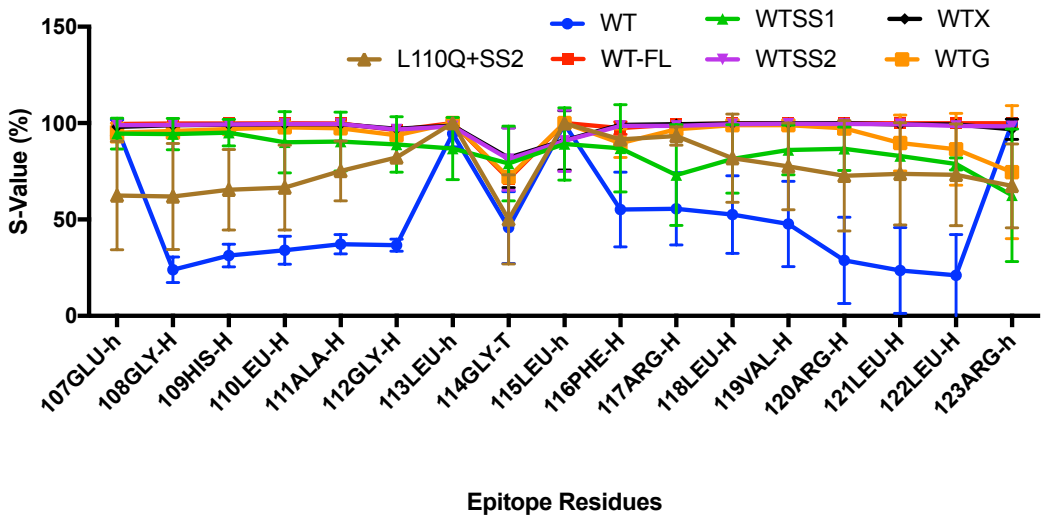


Figure 5.15: 1ORS – Residue level stability during 500 ns simulation. WTSS1 and WTSS2 refer to the WT peptide stapled with a disulphide bond at positions L103C:L125C and L103C:R126C, respectively. The WT peptide with glycine linker is labelled as WTG. The WTX label shows the WT peptide with 5 residues terminal extension. Significance of these results was calculated by using s_i values from 3 replicates and a p-value < 0.0001 was found. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

The data in Table 5.6 show that Cap2 on the WT has stablised the conformation (p-value = 0.02) whereas alanine and glutamine mutations either destabilised or had no effect on stability. The epitope was explored for plausible positions where cysteines could be substituted. Two different pairs of amino acids, L103:L125 and L103:R126, were selected for cysteine muations. It was observed that the disulphide bond between L103C and L126C (WTSS2) stabilised the conformation better than between L103C and L125C (WTSS1), and an increase of 12% in stability was observed with the former which means that the stapling at positions L103 and R126 resulting in an S-value of about 97%. However, the stabilising effect of stapling, with both pairs, was found to be significant (WT/WTSS1 and WT/WTSS2; p-value = 0.01). Cyclisation of this peptide (WTG) required a glycine linker of 2 residues between positions L102 and L128 which resulted in an S-value of 93%. A p-value of 0.004, for WT/WTG, suggests that the cyclisation has stabilised the native conformation to a significant extent. The epitope was extended (WTX) with 5 residues

at both termini to study the effect of additional non-epitope residues on the stability of the epitope. Interestingly, the WT with extended ends significantly retained its native conformation (p-value = 0.01) for over 97% of the time during simulation which is similar to the epitope stability within the context of the full length antigen (Figure 5.15). The WT, stapled, cyclised and extended epitopes are shown in Figure 5.14.

One of the mutants, L110Q, which had an S-value of 54% (slightly more than the WT) was also stapled at positions L103C and R126C, but surprisingly did not show the same level of stability as observed with the WT. This suggests that the stapling works better on the WT sequence as compared to the mutant (Figure 5.15). A Welch's t-test using the S-value of 3 replicates (WT/L110Q + SS2) was found to be insignificant (p-value=0.1). However, the results were found significant when a bigger sample (s_i values from 3 replicates) was used for the t-test (p-value < 0.0001). This suggests that while the overall conformation in not significantly stabilised, different parts of the peptide are stabilised or destabilised.

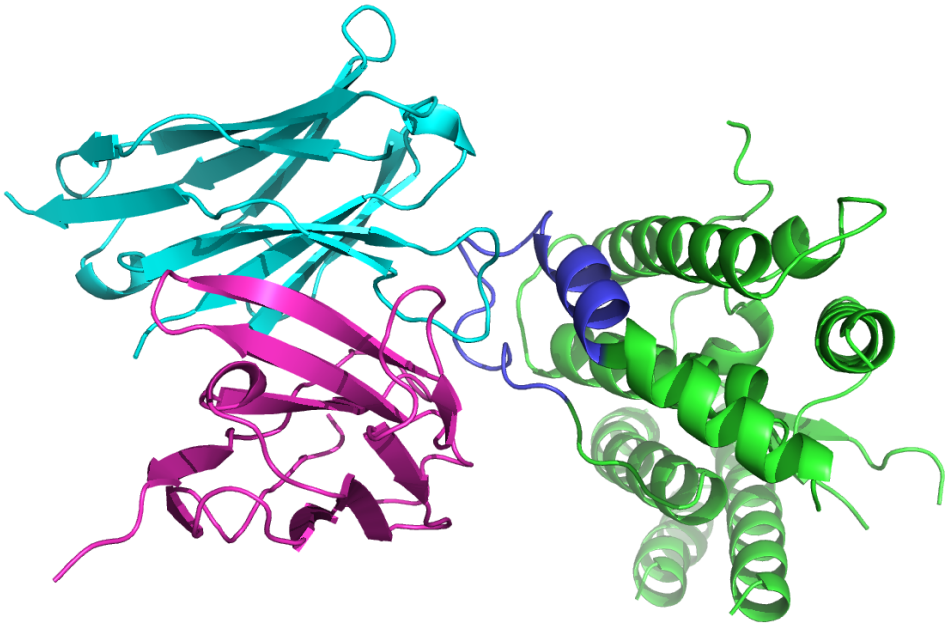


Figure 5.16: 4K2U – Crystal structure of PfEBA-175 F1 in complex with R218 antibody Fab fragment. Light (cyan) and heavy (violet) chains bound with full length antigen potassium channel . The epitope (at positions 149-169) mapped on the basis of antibody-antigen contacts is shown in blue.

5.3.1.4 4K2U – Helix-turn-Loop Epitope

The epitope in 4K2U (Erythrocyte binding antigen complexed with a R218 antibody Fab fragment [180]) forms a helix-turn-loop structure in a folded conformation and consists of 21 residues. The first 9 residues at the N-terminus form the helix (Figure 5.16). There are 12 contacting residues and among the nine non-contacting residues, 4 are hydrophobic (Table 5.1). Like other epitopes described above, epitope mutant and derivatives were studied using simulations. A total of 17 peptides (WT and 16 mutant and derivatives) were simulated for 500 ns with triplicates of each. This generated 51 trajectories which were analysed for conformational stability.

Knowing the fact that the epitope is composed of a helix and a loop, it was expected that the epitope would have a lower stability than when being part of full length antigen. Therefore, to test this, the antigen was simulated for 500 ns (3 repli-

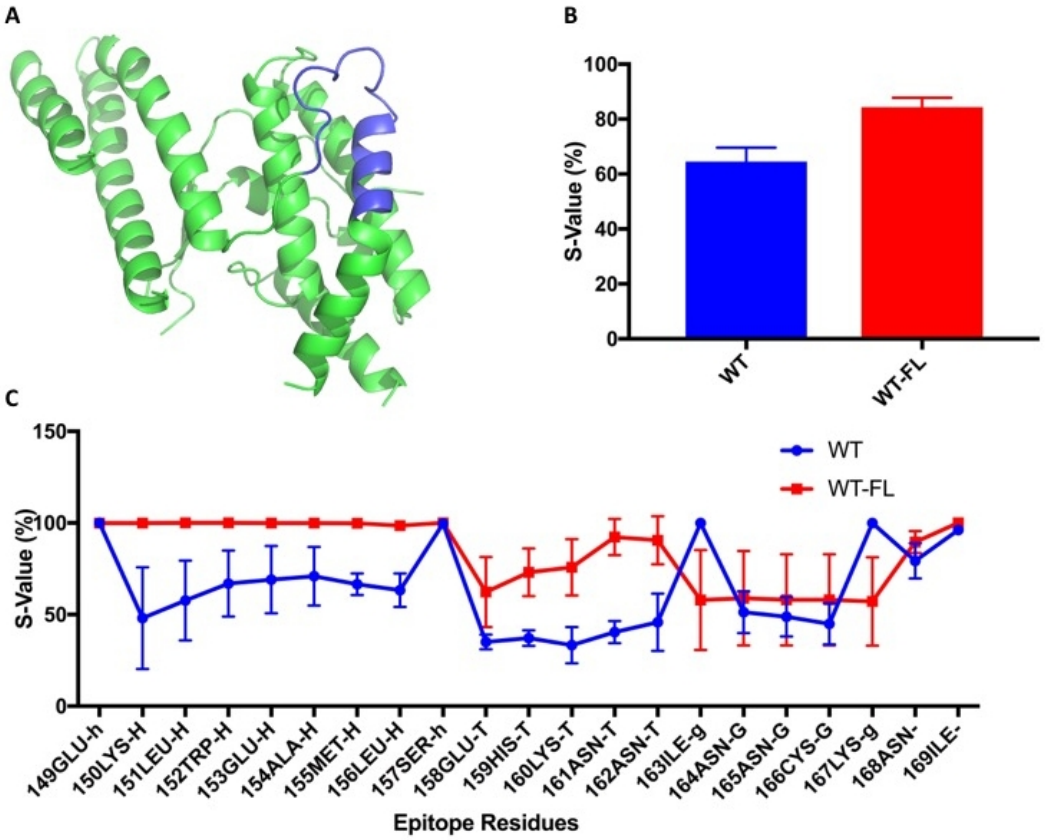


Figure 5.17: (A) 4K2U – The WT epitope (blue) in the full length protein (WT-FL), (B) The S-value of the WT peptide (64.51%) and WT-FL (84.38%), (C) The s_i-value of each residue in the epitope during 500 ns simulations (3 replicates). A p-value < 0.0001 for WT/WT-FL suggests that the epitope is less flexible, being part of the full length antigen, as compared to the the WT epitope when extracted. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

cates) and the S-value of the epitope was investigated. An S-value of approximately 84% was observed when the epitope was part of the native antigen whereas an S-value of about 64% was observed when the epitope region was simulated on its own (Figure 5.17). This shows that the epitope was only partially rigid within the context of full length antigen and the maximum expected conformational stability of any mutant is about 84%.

Table 5.7: 4K2U – Helix-turn-loop epitope (mutant and derivative peptides). Th peptides were simulated for 500 ns (3 replicates).The S-value for each peptide represents the average time it has spent in its initial conformation. The mean of S-value is computed for 3 replicates in each peptide. Bold numbers show the peptides with the interesting results (s_i values for these is shown in Figure 5.19.

Caps		**Alanine Mutations**			**Glutamine Mutations**				**WTSS**	**WTG**	**WTX**	**S-Value**
Ace/NME	ASN/NH2	L151A	M155A	C166A	L151Q	A154Q	M155Q	C166Q	L145C:E172C	N145:Gly:E172	W142-L174	
-	-	-	-	-	-	-	-	-	-	-	-	**64.51**
✓	-	-	-	-	-	-	-	-	-	-	-	59.03
-	✓	-	-	-	-	-	-	-	-	-	-	61.99
-	-	✓	-	-	-	-	-	-	-	-	-	63.49
-	-	-	✓	-	-	-	-	-	-	-	-	61.46
-	-	-	-	✓	-	-	-	-	-	-	-	59.62
-	-	✓	✓	✓	-	-	-	-	-	-	-	68.82
-	-	-	-	-	✓	-	-	-	-	-	-	58.42
-	-	-	-	-	-	✓	-	-	-	-	-	63.15
-	-	-	-	-	-	-	✓	-	-	-	-	58.71
-	-	-	-	-	-	-	-	✓	-	-	-	**71.22**
-	-	-	-	-	✓	✓	✓	✓	-	-	-	61.27
-	-	-	-	-	-	-	-	-	✓	-	-	74.11
-	-	-	-	-	-	-	-	✓	✓	-	-	71.0
-	-	-	-	-	-	-	-	-	-	✓	-	**81.87**
-	-	-	-	-	-	-	-	-	-	-	✓	**80.72**
-	-	-	-	-	-	-	-	✓	-	-	✓	**78.08**

Table 5.7 lists the S-value of all the peptides and shows that end-capping on the WT peptide has a small destabilising effect, but this was not statistically significant. Likewise, alanine and glutamine mutations did not improve the conformation except for the C166Q mutation where an increase of 7% in the S-value was observed. However, this increase in the S-value was not statistically significant.

Two epitope derivatives, WTSS and WTG were designed by the substitution of two cysteines and addition of a glycine linker at positions L145 and E172. The WTX peptide was designed by the addition of 5 residues at each of the termini (Figure 5.18). The highest level of stability ($\sim$81%) was observed when the peptide was cyclised with a glycine linker and the increase in stability was found to be significant (WT/WTG; p-value = 0.004). Since the epitope was stable for 84% of the total simulation time within the full length protein therefore this implies that the peptide remains close to its native conformation when cyclised. S-values of 80% and 74% for WTX and WTSS show the peptides' conformational stability and the s_i-values show that the helix at positions 149 to 157 was retaining its structure as compared to the WT where the N-terminal helix was quite unstable. However, the turn and loop region of the peptide remain flexible for all of these fairly stable peptides (Figure 5.19). A Welch's t-test on WT/WTX was found to be significant (p-value = 0.004) while that was not the case with WT/WTSS (p-value = 0.054). Two additional mutants were designed by combining the mutation C166Q with either disulphide stapling or exended ends. For both of the peptides, the S-value for combined mutations did not exceed the individual mutations which suggests that the stapling and extension of termini has a better stabilising effect on the WT than on the mutant.

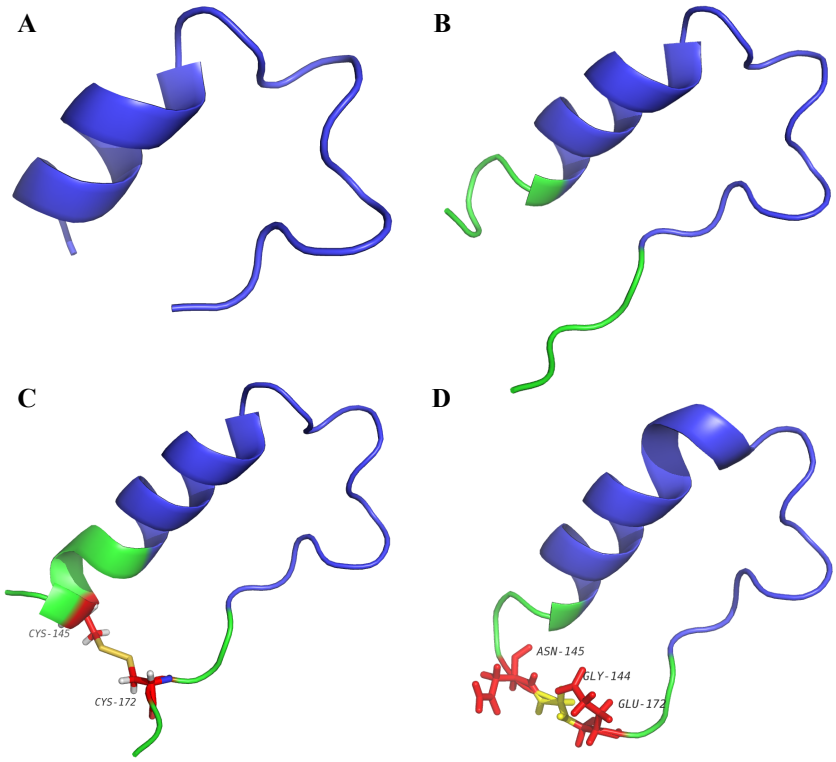


Figure 5.18: 4K2U – Helix-turn-loop epitope structure. **(A)** The WT peptide at positions 149-169 in the antigen. **(B)** The WTX peptide at positions 144-174 – extended at both ends by 5 residues. The epitope is shown in blue and the extended termini in green. **(C)** The WTSS peptide with a disulphide bond between the 2 cysteines at positions 145 and 172 (sticks); the epitope (blue) with extended ends (green). **(D)** The WTG peptide with a glycine linker (yellow sticks) added between amino acids N145 and E172 (red sticks).

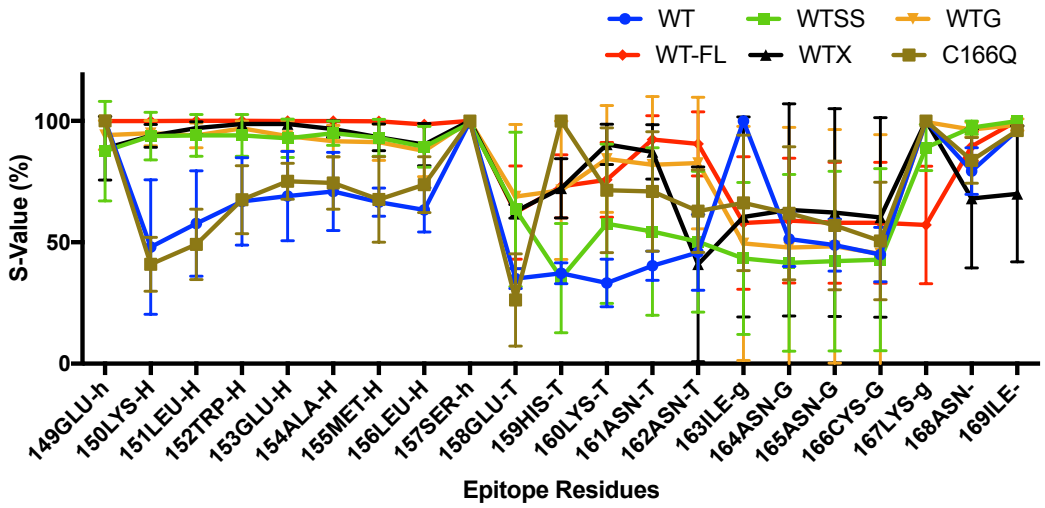


Figure 5.19: 4K2U – Residue level stability during 500 ns simulation (3 replicates). The WTSS refers to the peptide stapled with disulphide bond at positions L145C and E172C. The peptide with glycine linker is labelled as WTG. The WTX label shows WT peptide with extended terminus. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

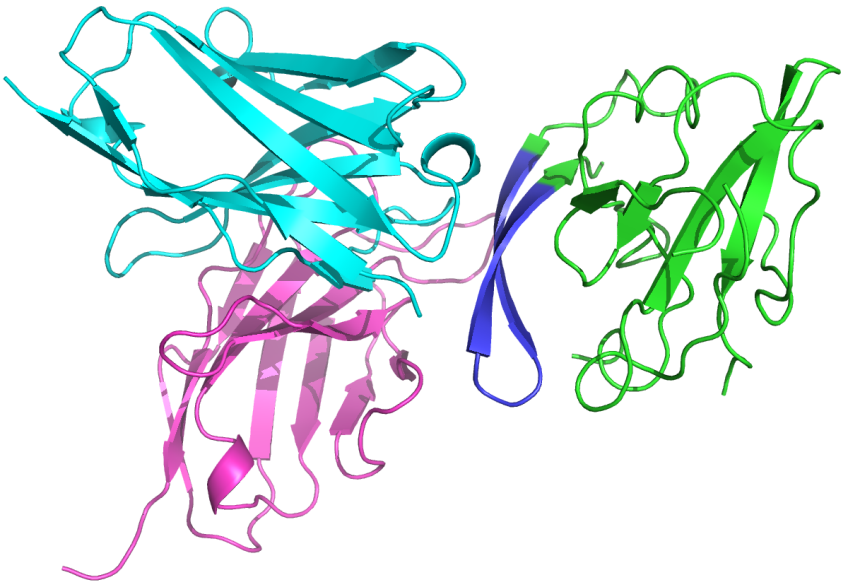


Figure 5.20: 4WEB – Structure of the core ectodomain of the hepatitis C virus envelope glycoprotein 2 bound with a Fab. Light (cyan) and heavy (violet) chains bound with full length hepatitis C virus envelope glycoprotein 2. The epitope (at positions 631-645) mapped on the basis of antibody-antigen contacts is shown in blue.

5.3.1.5 4WEB – β-Strand-loop-β-Strand

A β-strand-loop-β-strand epitope (at positions 631-645) was identified in the E2 core domain of the Hepatitis C virus protein that was bound to a Fab fragment (Figure 5.20) [181]. The epitope has a length of 15 residues, 9 of which contact the antibody. Of the non-contacting residues, 3 are hydrophobic and are potential sites for alanine and glutamine mutations (Table 5.2). These mutations produced 8 mutant peptides, 6 with alanine and glutamine mutations each on an individual site and 2 with three mutations at once. In addition, derivative peptides were produced by addition of end-caps, disulphide bond stapling, addition of a glycine linker and extension of each of the termini in the WT peptide. A total of 29 peptides (WT and 28 mutants and derivatives) were studied by molecular dynamics. The simulations were carried out for 500 ns with 10 replicates of each of the peptides. The resulting 290 trajectories were analysed for structural stability.

Table 5.8: 4WEB $-$ β-Strand-loop-β-strand, folded Epitope - mutant and derivative peptides. The position of mutation and the average time (S-value) a peptide has spent in its initial conformation is shown. S-value represents the mean of 10 replicates of each of the peptide. Bold numbers (S in Equation 5.1) show the peptides with the most interesting results (Residue level stability for these is shown in Figure 5.22)

Caps		**Alanine Mutations**			**Glutamine Mutations**			**WTSS**	**WTG**	**WTX**	**S-Value**
Ace/NME	ASN/NH2	I633A	M635A	V637A	I633Q	M635Q	V637Q	Y628C:A647C	Y628:Gly:N649	N627-N649	
-	-	-	-	-	-	-	-	-	-	-	**95.43**
✓	-	-	-	-	-	-	-	-	-	-	94.87
-	✓	-	-	-	-	-	-	-	-	-	93.65
-	-	-	-	-	-	-	-	✓	-	-	**97.74**
✓	-	-	-	-	-	-	-	✓	-	-	96.45
-	✓	-	-	-	-	-	-	✓	-	-	96.69
-	-	-	-	-	-	-	-	-	✓	-	**97.37**
-	-	-	-	-	-	-	-	-	-	✓	**96.74**
✓	-	-	-	-	-	-	-	-	-	✓	95.83
-	✓	-	-	-	-	-	-	-	-	✓	94.22
-	-	✓	-	-	-	-	-	-	-	-	83.83
-	-	-	✓	-	-	-	-	-	-	-	**81.41**
-	✓	-	✓	-	-	-	-	-	-	-	80.67
-	-	-	✓	-	-	-	-	✓	-	-	**95.54**
-	-	-	✓	-	-	-	-	-	✓	-	**95.97**
-	-	-	-	✓	-	-	-	-	-	-	85.65
-	✓	-	-	✓	-	-	-	-	-	-	79.71
-	-	-	-	✓	-	-	-	✓	-	-	94.20
-	-	-	-	✓	-	-	-	-	-	✓	94.40
-	-	✓	✓	✓	-	-	-	-	-	-	75.62
-	-	-	-	-	✓	-	-	-	-	-	**95.81**
-	✓	-	-	-	✓	-	-	-	-	-	87.50
-	-	-	-	-	✓	-	-	✓	-	-	94.80
-	-	-	-	-	✓	-	-	-	✓	-	94.10
-	-	-	-	-	-	✓	-	-	-	-	85.13
-	-	-	-	-	-	-	✓	-	-	-	**84.57**
-	✓	-	-	-	-	-	✓	-	-	-	89.01
-	-	-	-	-	-	-	✓	✓	-	-	95.81
-	-	-	-	-	✓	✓	✓	-	-	-	82.23

Figure 5.21 shows the structure of the WT peptide, WT with extended N and C termini (WTX) using a 4 residue extension, stapled (WTSS) and cyclised (WTG) by the addition of single glycine linker in the WT epitope.

Table 5.8 shows that the WT epitope spent about 95% of its time in its initial conformation and was fairly stable in solution, being able to keep its β-strand-loop-β-strand conformation as an isolated peptide. It is evident that the presence of three hydrophobic residues in the first β-strand did not cause instability in the epitope structure. End-capping did not further stabilise or destabilise the epitope. Stapling (WTSS) and cyclisation (WTG) of the WT peptide slightly improved the stability at the ends as shown in Figure 5.22 which shows s_i of all the residues in the peptides. A Welch's t-test was applied on WT/WTSS and WT/WTG and a p-value = 0.006 indicating a significant increase in stability for both. The extended termini peptides (WTX) showed a slight improvement in the stability compared with the WT peptide, however this was found to be insignificant (p-value = 0.06). Figure 5.22A also includes the peptide simulation results when it is part of the full length antigen where an S-value of 97% was observered which is only 2% more than the WT peptide when it is taken out of the antigen. This suggests that the epitope region is fairly rigid and stable.

Surprisingly, the hydrophobic to alanine and glutamine mutations significantly destabilised the peptide (p-value < 0.05). I633A, M635A and V637A mutations caused a decrease of 10-15% in the S-value (Table 5.8). Of these mutations, M635A and V637A, were selected and derivative peptides were designed by end-capping, stapling and cyclisation or terminal extension. The end-capping did not help improving the conformation, but the disulphide bond stapling and addition of glycine linker in the destabilised mutant peptide restored the conformational stability to a

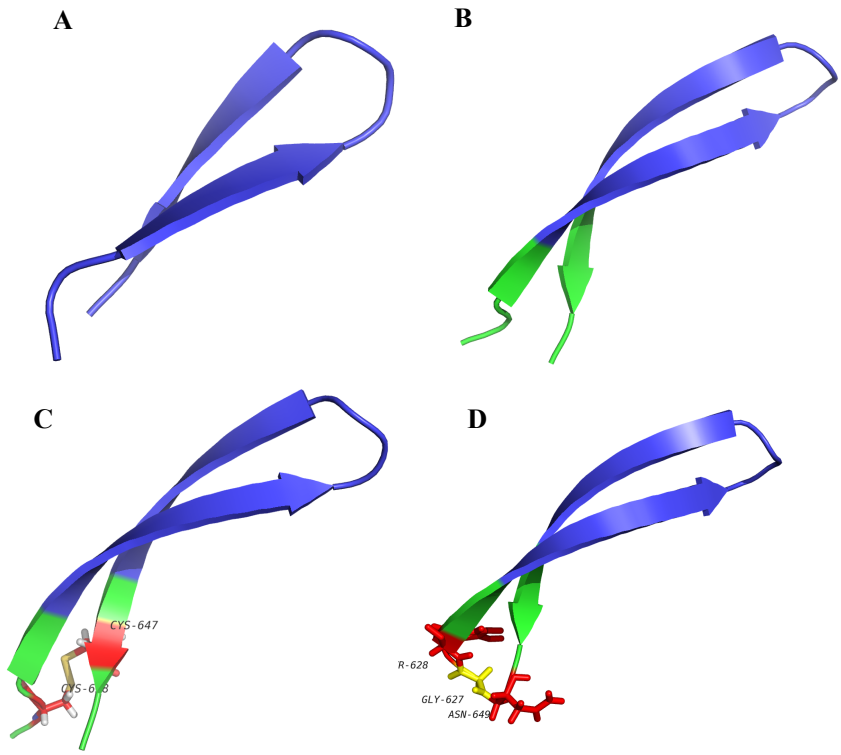


Figure 5.21: 4WEB – β-strand-loop-β-strand structure. A) The WT peptide at position 631-645 in the antigen. B) The WTX peptide at position 627-649 – extended at both ends by 4 residues. The epitope is shown in blue and the extended termini in green. C) The WTSS peptide with a disulphide bond between the 2 cysteines at positions 628 and 647 (sticks); the epitope (blue) with extended ends (green). D) The WTG peptide with glycine linker (yellow sticks) added between amino acids Y628 and N649 (red sticks).

level similar to the WT (Figure 5.22). Of the glutamine mutations, I633Q, M635Q and V637Q, mutation I633Q did not have any destabilising effect with respect to the WT. However, end-capping of I633Q caused a decrease of approximately 8% in S-value, but stapling and cyclisation restored the stability. In the case of M635Q and V637Q, a behaviour (destabilisation) similar to M635A was observed.

Epitope Simulation in the Presence of Antibody

In order to study the behaviour of the epitope and its derivative peptides in the presence of the antibody, WT, WTSS, WTX, and WTG peptides were simulated in complex with Fab. This was to confirm that the stapling and cyslisation do not cause the epitope to dissociate from the antibody. The simulation was carried out

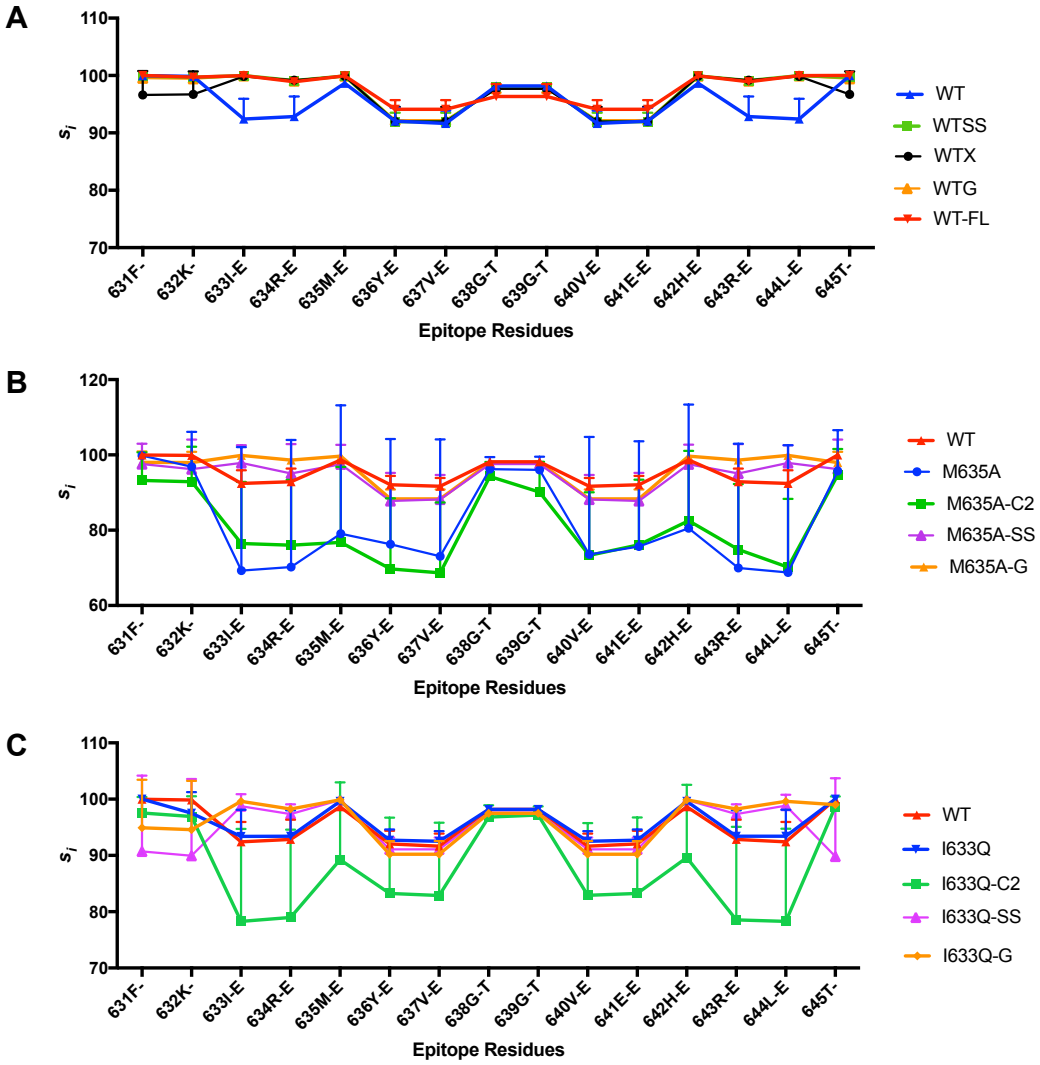


Figure 5.22: 4WEB – Residue level stability, s_i, during 500 ns simulations (10 replicates). **(A)** WT, WTSS (WT stapled at positions Y628C:A647C), WTX (WT extended by 4 residues at each terminus), WTG (cyclised WT with glycine linker between positions Y628 and N649) and WT-FL (epitope in the full length antigen) showing the stability of each residue during the simulations. **(B)** Mutant M635A, along with its derivatives, capped, stapled and cyclised peptides, in comparison with WT. C2 represents the cap Asn/NH2. **(C)** The effect of stapling and cyclisation of the mutant I633Q compared with WT. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

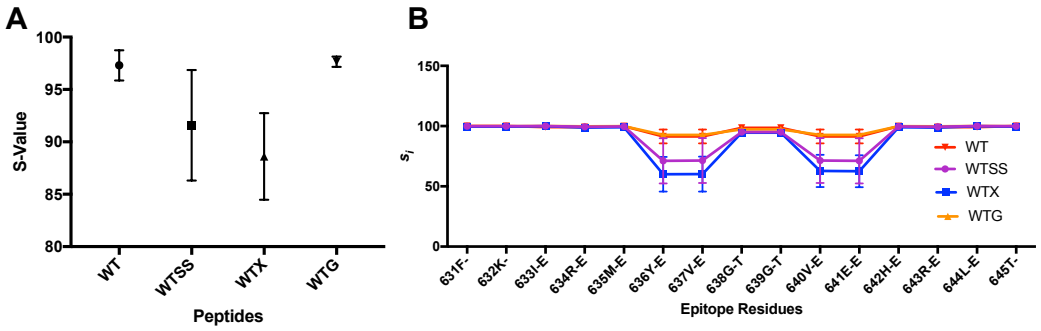


Figure 5.23: Peptide simulation in the presence of the Fab for 500 ns (5 replicates). **(A)** The S-value for each peptide and **(B)** s_i plotted against residue number for WT, WTSS and WTG.

for 500 ns and 5 replicates were produced resulting in 20 trajectories. The trajectories were visually analysed using VMD to ensure that all the peptides were able to maintain their interaction and binding with the Fab. It was observed that all four of these peptides were able to retain their native conformation as well as epitope-antibody interaction over the 500 ns simulations. The peptides from the trajectory were extracted and the S-value was computed as shown in Figure 5.23. Compared with Figure 5.22A, the WT and WTG peptides show increased stability at residues 636, 637, 640 and 641. This is not seen in WTSS and WTX suggesting that these peptides do not interact as effectively.

5.3.2 Summary of Folded Regions Simulations

Table 5.9 summarises the effect of mutations studied using simulations. It is evident that the end-capping, alanine and glutamine mutations have not shown any stabilising effects on any of the epitopes except for 1ORS where end-capping (Cap2) has singnificantly improved the conformation. However, the disulphide bond stapling and cyclisation by glycine linker has worked the best in retaining the native conformation. Stapling did not prove effective in the case of 4K2U, the potential reason being that the flexible C-terminus that was so mobile during simulation that it could not significantly retain the conformation whereas cyclisation was effective on all of

Table 5.9: Summary of the effects of mutations on folded epitope regions. The significantly stabilised mutant types are shown with a tick while others are marked with a cross.

Epitope	Capping	Alanine Mutations	Glutamine Mutations	Stapling	Cyclisation	Extended Ends
4N9G	×	×	×	✓	✓	✓
3LHP	×	×	×	✓	✓	×
1ORS	✓	×	×	✓	✓	✓
4K2U	×	×	×	×	✓	✓
4WEB	×	×	×	✓	✓	✓

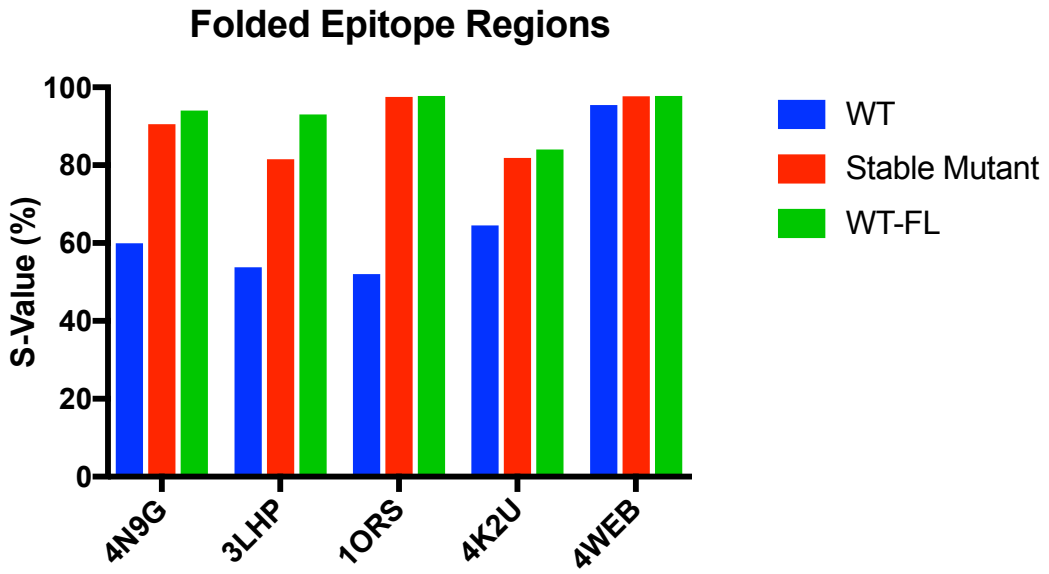


Figure 5.24: Folded epitope regions' simulations for 500 ns (3 replicates). An average S-value for the WT (on its own), stable mutant (the most stable among all the mutant peptides), and WT-FL (the epitope extracted from the simulations of the full length antigen protein.)

the folded epitope regions. Of these five epitopes, four were able to maintain the native structure by extension of 4-5 non-epitope residues at N and C termini. In conclusion, stapling, cyclisation and extension of termini could be used as potential strategies to enhance conformational stability in folded epitopes.

Interestingly, stapling, cyclisation and extension of termini is more effective on the WT peptide compared with alanine or glutamine mutants. In 4N9G, 3LHP, 4K2U and 4WEB, some of the mutant peptides were stapled and cyclised and a decrease in S-value was observed (for example, I71A+stapling in Table 5.4) which suggets that the stapling and cyclisation should be used on the WT peptide.

Figure 5.24 shows that stability has been achieved for all the epitopes and the most stable mutant has reached a similar level of stability to what is seen in the full length WT protein (WT-FL). The most stable mutant peptides from 4N9G and 4WEB were also simulated in the presence of antibody and it was found that the mutations (stapling, cyclisation and extended termini) did not prevent epitope binding with antibody. Indeed they showed an increased stability when simulated in the presence of antibody (Figure 5.7 and 5.23).

10 replicates were obtained for all the mutant peptides of 4WEB which was selected for experimental validation of their conformational stability. In this epitope region, WT, WTG, WTX and V637A were chosen to perform laboratory experiments.

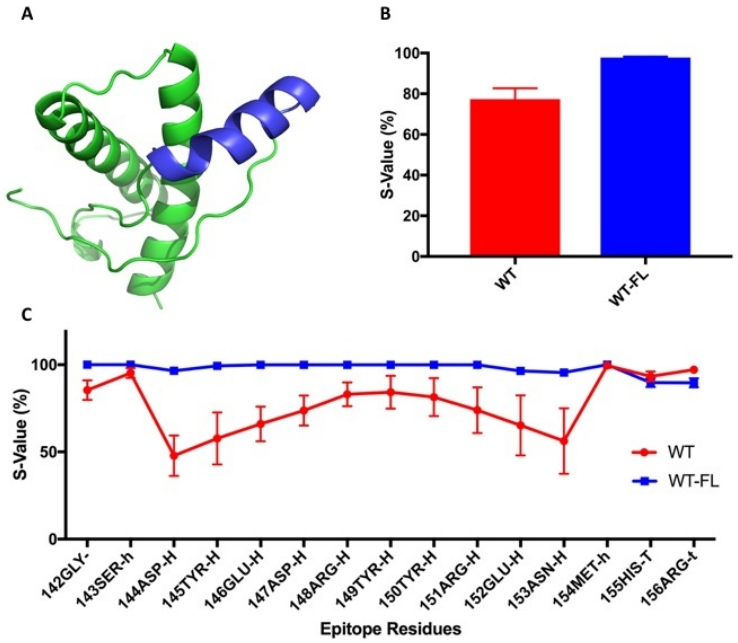


Figure 5.25: (A) The WT epitope (blue) in the full length protein (WT-FL), (B) The S-value of the WT peptide (77%) and WT-FL (97%) (C) s_i-value of each residue in the epitope during 500 ns simulations (10 and 3 replicates for WT and WT-FL respectively). The explanation of secondary structure labels on x-axis is given in section 5.2.5.

5.3.3 Molecular Dynamic Simulation of Extended Regions

As with the folded epitope regions described above, 5 extended helical epitope regions were chosen to study the effect of alanine and glutamine mutations and end-capping on peptide stability. Experiments were performed to explore the stability of isolated extended epitope regions using MD simulations.

5.3.3.1 2W9E – Extended α-Helical Epitope

Human prion protein (PrP), PDB file 2W9E, contains an epitope at positions 142-156, which provides a binding site for the Fab fragment of monoclonal antibody ICSM 18 [182]. In the native protein, this epitope is comprised of an extended α-helix (shown in Figure 6.2) with a length of 15 residues, 12 of which make direct contact with the antibody. Of the non-contacting residues, one is hydrophobic (methionine) and could be mutated in an attempt to increase stability. In addition to this hydrophobic non-contacting methionine, a hydrophilic non-contacting tyrosine was

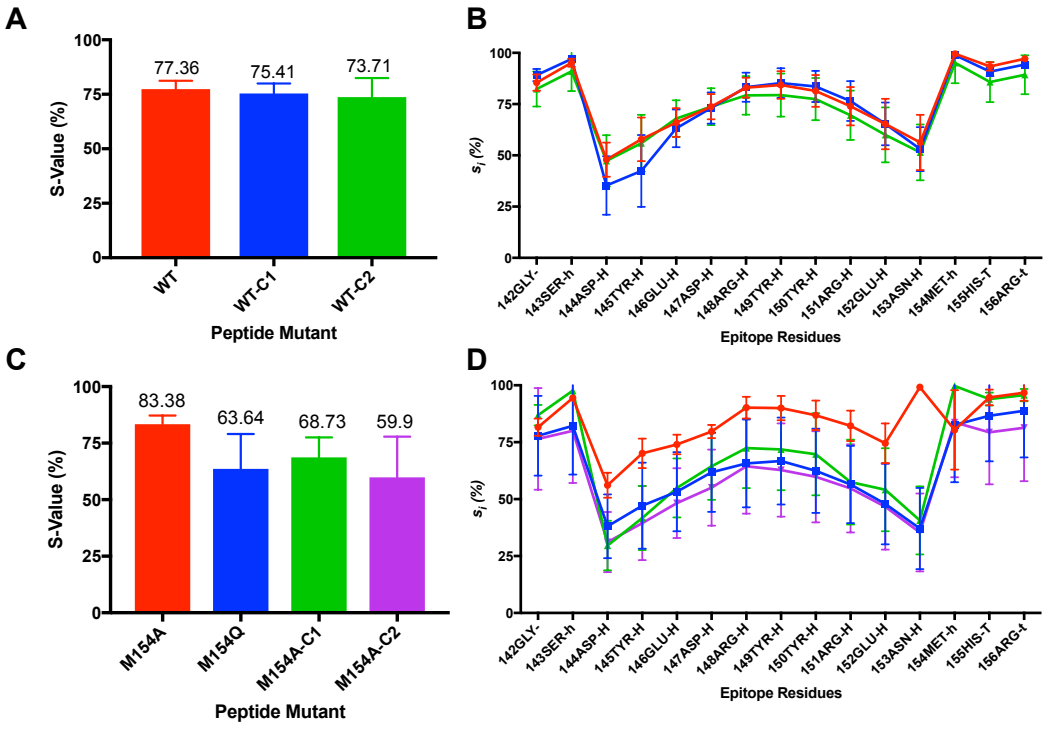


Figure 5.26: 2W9E – extended α-helical mutant peptides simulated for 500 ns (10 replicates). **(A)** An average S-value is shown for WT, WT-C1, WT-C2, **(C)** M154A, M154A-C1, M154A-C2 and M154Q. C1 and C2 represents Ace/NME and Asn/NH2 caps respectively. The error bars show the mean with 95% confidence interval. The p-value for WT/WT-C1 and WT/WT-C2 is 0.5 and 0.40, respectively, - showing no effect of caps on WT. A p-value of 0.02 was observed for WT/M154A - showing the significant increase in the stability of M154A. The mutation M154Q and caps on M154A show significant instability compared with the WT. The p-values for WT/M154Q, WT/M154A-C1 and WT/M154A-C2 was found 0.03, 0.07 and 0.04 respectively. Likewise, M154A/M154-C1 and M154A/M154-C2 have p-value of 0.005 and 0.01 which shows that end-capping has significantly destabilised the stable mutant. **(B)** The s_i values of peptide residues for WT, WT-C1 WT-C2, **(D)** M154A, M154A-C1, M154A-C2 and M154Q. The error bars represent mean and error with 95% confidence interval. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

also mutated as a control experiment.

The level of epitope flexibity within the full length protein antigen was studied by carrying out a 500 ns simulations (3 replicates) and it was observed that the epitope region spent 97% of its time in its native conformation (Figure 5.25). Consequently a stability of over 97% cannot be expected by studying mutations to the extracted peptide.

Table 5.10: 2W9E – Extended α-helical epitope. The peptides were simulated for 500 ns (10 replicates) and 1000 ns (only once). The resulting 110 trajectories (of 10x500 ns) from simulation of 11 peptides were analysed for the S-value. The S-value for each mutant peptide represents the average time it has spent in its initial conformation. The S-value of the WT and the most interesting mutant peptides are shown in bold while it is underlined for the control experiments.

Caps		**Alanine Mutations**		**Glutamine Mutations**		**S-Value (500 ns)***	**S-Value (1000 ns)****
Ace/NME	ASN/NH2	Y150A	M154A	Y150Q	M154Q		
-	-	-	-	-	-	**77.36**	**77.43**
✓	-	-	-	-	-	75.41	77.16
-	✓	-	-	-	-	73.71	79.42
-	-	✓	-	-	-	72.45	76.06
-	-	-	✓	-	-	**83.38**	**88.57**
✓	-	-	✓	-	-	68.73	75.12
-	✓	-	✓	-	-	59.90	47.80
-	-	✓	✓	-	-	**65.08**	**58.92**
-	-	-	-	✓	-	67.38	67.29
-	-	-	-	-	✓	63.64	60.45
-	-	-	-	✓	✓	67.45	67.45

* The S-value is an average of 10 replicates simulated for 500 ns .
** The S-value calculated from the single simulation experiment for 1000 ns.

Table 5.10 shows the average S-value for 11 peptides (WT and 10 mutants) obtained from 10 replicates of 500 ns simulations. On average, the M154A mutant peptide spent approximately 83% of its time in its initial conformation which is 6% more than the time the WT peptide spent in its initial conformation. A Welch's t-test was applied to confirm the significance of this difference, and a p-value of 0.021 was observed showing the increase in the S-value of M154A is significant. The end-capped versions of both the WT and M154A did not perform better than the un-capped. Interestingly, the end-caps on M154A destabilised the peptide by reducing the effect of the mutation, and the reduction in stability was found to be significant (M154A/M154-C1 p-value = 0.005, M154A/M154-C2 p-value = 0.016). The simulation control experiments with mutants Y150A, Y150Q, Y150A+M154A and Y150Q+M154Q were performed and the results were compared with the WT but they were not found to be significantly different from the WT (p-value > 0.05). A Welch's t-test on M154A and Y150A+M154A resulted in a p-value of 0.002 which suggests that the presence of the Y150A mutation has removed the stabilising effect of M154A. Figure 5.26 shows the S-value of the peptides and s_i values of peptide residues.

Peptide Simulation in the Presence of Antibody

It was expected that the WT epitope region would show a stable binding association with the antibody during the simulations while a mutation could prevent the peptide from binding. Therefore, a number of mutant peptides (WT-C2, M154A, M154Q, M154A-C1 and M154A-C2) and the WT peptide were simulated for 500 ns (10 replicates) in the presence of the antibody. Figure 5.27 shows the S-value and s_i values of these 6 peptides in the presence of antibody.

Visual analysis of the trajectories showed that all the peptides stayed in close

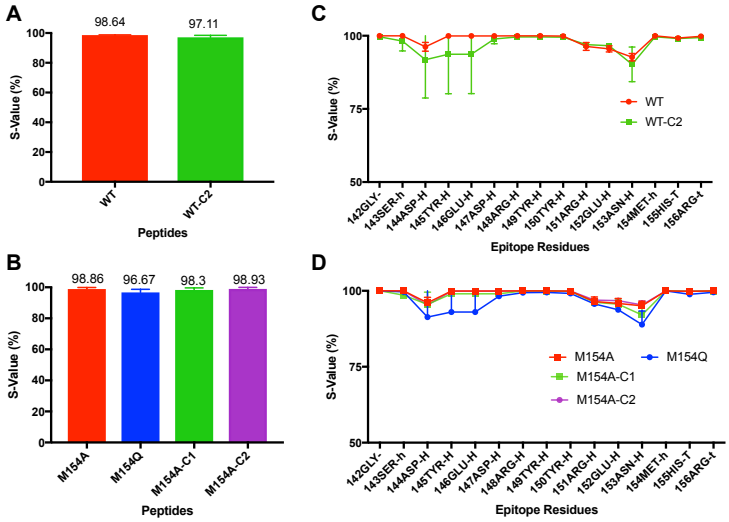


Figure 5.27: 2W9E – extended α-helical mutant peptides simulated in the presence of the antibody for 500 ns (10 replicates). **(A)** The S-value for WT and WT-C2, **(B)** M154A, M154Q, M154A-C1 and M154-C2. C1 and C2 represents Ace/NME and Asn/NH2 caps respectively. The error bars show the mean with 95% confidence interval. P-value of mutant peptides compared with WT was found insignificant - showing no notable difference in the S-value. **(C)** The s_i values of peptide residues for WT and WT-C2, **(D)** M154A, M154Q, M154A-C1 and M154A-C2. The error bars represent mean and error with 95% confidence interval. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

proximity to the antibody binding site and did not fall off. An increase in the S-value from 77% to 98% was observed when the WT peptide was simulated in the presence of antibody and the S-value of all the peptides in the presence of the antibody was also between 96% and 98% suggesting that they are all able to bind successfully, showing increased stability on binding (Figure 5.27). Compared with free peptide (Figure 5.26B), the WT, WT-C2, M154A, M154A/C1 and M154A/C2 peptides show increased stability at residues 144, 145, 152 and 153. Notably the increase in stability of these residues is much less in M154Q (the unstable mutant), and a considerable fluctuation can be seen at these positions (Figure 5.27).

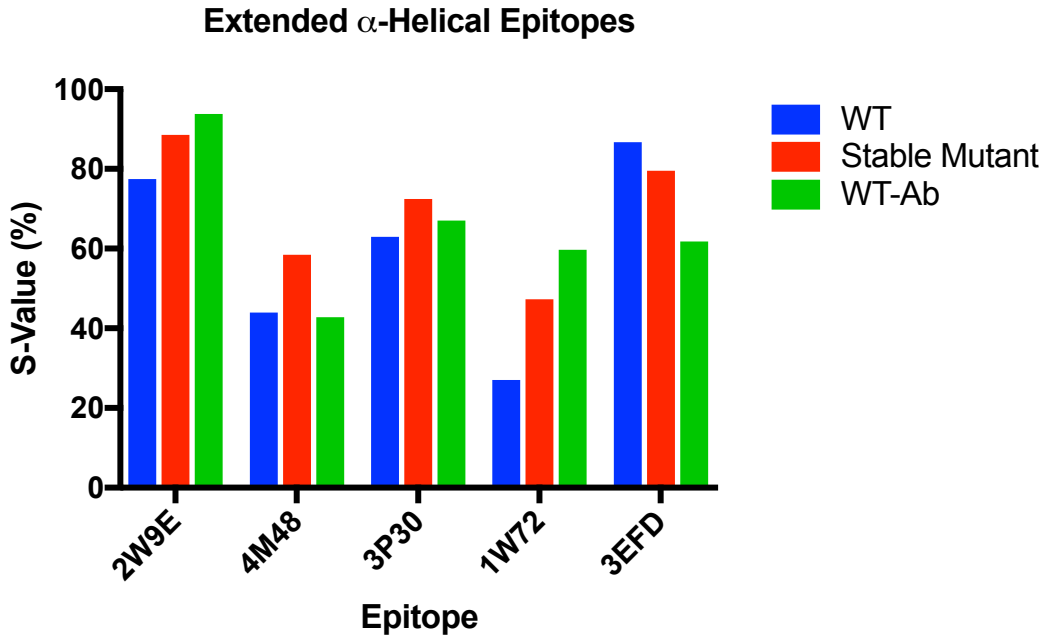


Figure 5.28: Extended α-helical epitope single simulations for 1000 ns. The S-value for WT (on its own), stable mutant (the most stable among all the mutant peptides), and WT-Ab (the WT simulated in the presence of the antibody) is shown. The labels represent the PDB code of the antibody-antigen complex that is used to map and extract the epitope.

5.3.4 Summary of Other Extended α-Helical Epitopes

At the start of this project, 5 extended epitope regions from 2W9E, 4M48, 3P30, 1W72 and 3EFD were selected with the intention of producing replicates of each peptide and its associated mutant peptides. Unfortunately, time limitations meant that this was not possible with the exception of 2W9E (discussed in Section 5.3.3.1). However, results were obtained for single 1000 ns simulations for 2W9E and the four other epitope regions (WT and mutant peptide) and analysed. While no statistics could be performed owing to the lack of replicates, the data for 2W9E suggest that the single 1000 ns simulations are representative of the 500 ns replicates (Table 5.10). Results are shown in Tables 5.11, 5.12, 5.13 and 5.14.

These data show that it was also possible to identify stabilising mutations for 4M48, 3P30 and 1W72 (Figure 5.28). However, 3EFD was an exception where

Table 5.11: 4M48 − extended α-helical mutant peptides. The peptides were simulated for 1000 ns (only once). The WT with end-caps and three hydrophobic residues were explored to study the effects of mutations on conformational stability. The S-value for each mutant peptide represents the average time it has spent in its initial conformation. The WT and stabilised mutant peptides (M510A and M510Q) and their capped versions are shown in bold, and their s_i values are plotted in Figure 5.29. The end-capping on the WT type peptide has increased the stability 9-12%. Two mutations M510A and M510Q showed an increase of 10-15% in the S-value. Therefore, to explore any positive effect of end-capping on these mutant peptides, they were also end-capped and simulated. The end-capping on WT, M510A and M510Q shows that, the cap1 has similar effect on these three peptides which is regardless of mutation.

Caps		**Alanine Mutations**			**Glutamine Mutations**			**S-Value**
Ace/NME	ASN/NH2	F503A	I507A	M510A	F503Q	I507Q	M510Q	
-	-	-	-	-	-	-	-	**43.96**
✓	-	-	-	-	-	-	-	**55.96**
-	✓	-	-	-	-	-	-	52.46
-	-	✓	-	-	-	-	-	42.16
-	-	-	✓	-	-	-	-	46.56
-	-	-	-	✓	-	-	-	**58.46**
✓	-	-	-	✓	-	-	-	**55.39**
-	✓	-	-	✓	-	-	-	41.83
-	-	✓	✓	✓	-	-	-	36.52
-	-	-	-	-	✓	-	-	50.92
-	-	-	-	-	-	✓	-	50.55
-	-	-	-	-	-	-	✓	**53.26**
✓	-	-	-	-	-	-	✓	**52.54**
-	✓	-	-	-	-	-	✓	44.27
-	-	-	-	-	✓	✓	✓	44.34

the WT peptide was the most stable and all other mutations destabilised the native conformation.

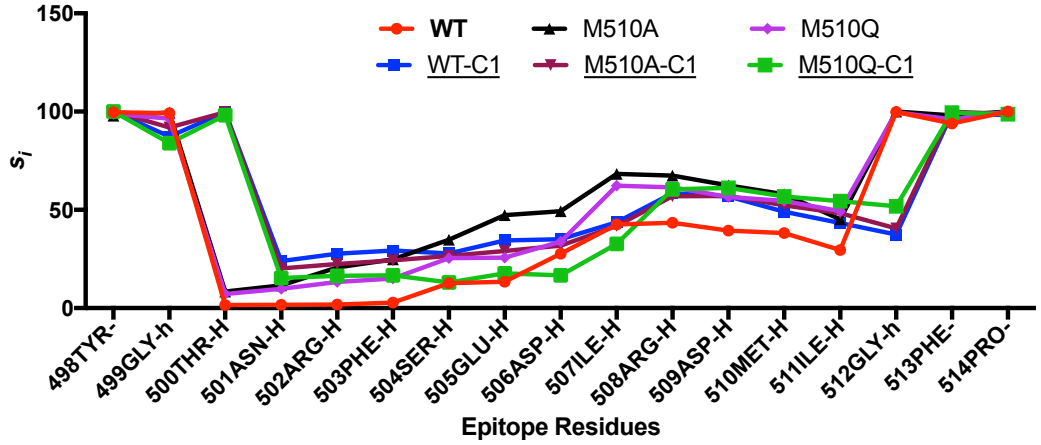


Figure 5.29: 4M48 – extended α-helical peptides. The s_i value represents the time each residue in the peptide has spent in its initial conformation during the simulation. The s_i values of WT, M510A, M510Q and their capped versions are shown. C1 represents Ace/NME cap. The mutation, M510A at the C-terminus, has slightly better retained the overall conformation. Moreover, the S-value of the WT peptide simulation in the presence of the antibody (42.75%) is arbitrarily the same as compared to the value in the absence of antibody (peptide on its own) which is 43.97%. During simulation, the peptide was seen to remain bound to the antibody. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

Table 5.12: 3P30 – extended α-helical mutant peptides. The peptides were simulated for 1000 ns (only once). The WT with end-caps and four hydrophobic residues were studied for the effects of mutations on conformational stability. The S-value for each mutant peptide represents the average time it has spent in its initial conformation. The end-caps on the WT peptide has destabilised the conformation. Among all the mutations, I646Q has increased the S-value to 10% as compared with the S-value of the WT. The end-caps on this stabilising mutant has similar destabilising effect as on the WT. The peptides with an increased S-value compared with the WT and the respective destabilising mutant peptides are shown in bold and their s_i values are plotted in Figure 5.30.

Caps		**Alanine Mutations**				**Glutamine Mutations**				**S-Value**
Ace/NME	ASN/NH2	I641A	I642A	L645A	I646A	I641Q	I642Q	L645Q	I646Q	
-	-	-	-	-	-	-	-	-	-	**62.92**
✓	-	-	-	-	-	-	-	-	-	**58.81**
-	✓	-	-	-	-	-	-	-	-	57.06
-	-	✓	-	-	-	-	-	-	-	49.97
-	-	-	✓	-	-	-	-	-	-	59.77
-	-	-	-	✓	-	-	-	-	-	53.97
-	-	-	-	-	✓	-	-	-	-	**65.20**
-	-	✓	✓	✓	✓	-	-	-	-	28.81
-	-	-	-	-	-	✓	-	-	-	66.07
-	-	-	-	-	-	-	✓			58.19
-	-	-	-	-	-	-	-	✓		65.16
-	-	-	-	-	-	-	-	-	✓	**72.48**
✓	-	-	-	-	-	-	-	-	✓	**54.43**
-	✓	-	-	-	-	-	-	-	✓	31.94
-	-	-	-	-	-	✓	✓	✓	✓	59.64

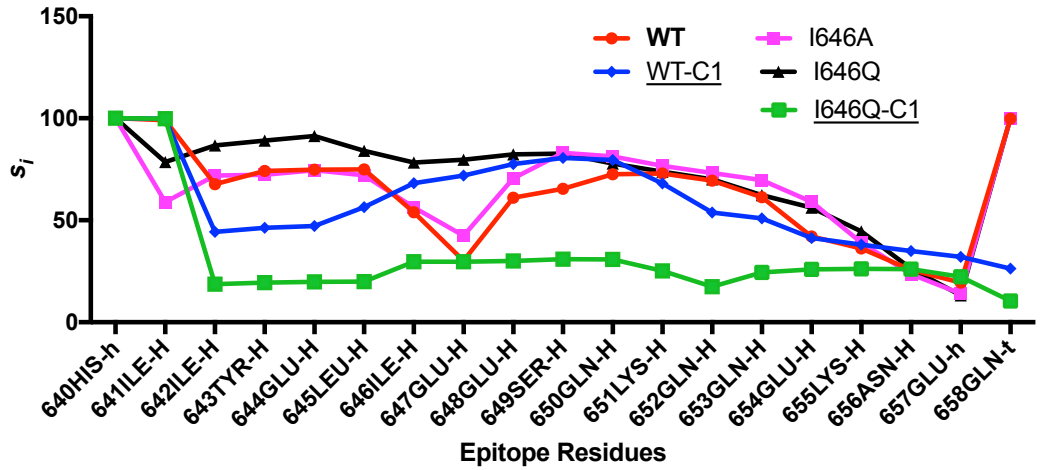


Figure 5.30: 3P30 – extended α-helical peptides. The s_i value represents the time each residue in the peptide has spent in its initial conformation during the simulation. The s_i values of WT, WT-C1, I646A, I646Q and I646Q-C1 are shown. C1 represents Ace/NME cap. The mutant I646Q appears to be the stable one. However, the caps on this stable mutant has destabilised it. Moreover, not a significant difference in S-value was observed when peptide was either simulated on its own or in the presence of the antibody. The visual analysis of antibody-peptide simulation showed that the peptide did not fall-off the antibody during simulation. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

Table 5.13: 1W72 – extended α-helical mutant peptides. The peptides were simulated for 1000 ns (only once). The WT with end-caps and four hydrophobic residues were studied for the effects of mutations on conformational stability. The S-value for each mutant peptide represents the average time it has spent in its initial conformation. The WT appears highly unstable and there is no effects of C1 whereas C2 has destabilising effect. The S-value of M67A, G83A and M67Q shows an increase of 8-20% and shown in bold. Their s_i values are plotted in Figure 5.31. The similar effect of caps was observed as it was seen in the WT. C1 and C2 represent Ace/NME and ASN/NH2 caps respectively.

Caps		**Alanine Mutations**				**Glutamine Mutations**				**S-Value**
Ace/NME	ASN/NH2	M67A	L78A	L81A	G83A	M67Q	L78Q	L81Q	G83Q	
-	-	-	-	-	-	-	-	-	-	**27.04**
✓	-	-	-	-	-	-	-	-	-	27.97
-	✓	-	-	-	-	-	-	-	-	23.82
-	-	✓	-	-	-	-	-	-	-	32.12
✓	-	✓	-	-	-	-	-	-	-	**35.73**
-	✓	✓	-	-	-	-	-	-	-	25.59
-	-	-	✓	-	-	-	-	-	-	27.20
-	-	-	-	✓	-	-	-	-	-	29.44
-	-	-	-	-	✓	-	-	-	-	**43.69**
✓	-	-	-	-	✓	-	-	-	-	44.15
-	✓	-	-	-	✓	-	-	-	-	34.0
-	-	✓	✓	✓	✓	-	-	-	-	28.92
-	-	-	-	-	-	✓	-	-	-	**47.25**
✓	-	-	-	-	-	✓	-	-	-	33.33
-	✓	-	-	-	-	✓	-	-	-	16.38
-	-	-	-	-	-	-	✓	-	-	27.0
-	-	-	-	-	-	-	-	✓	-	42.55
-	-	-	-	-	-	-	-	-	✓	37.79
-	-	-	-	-	-	✓	✓	✓	✓	31.31

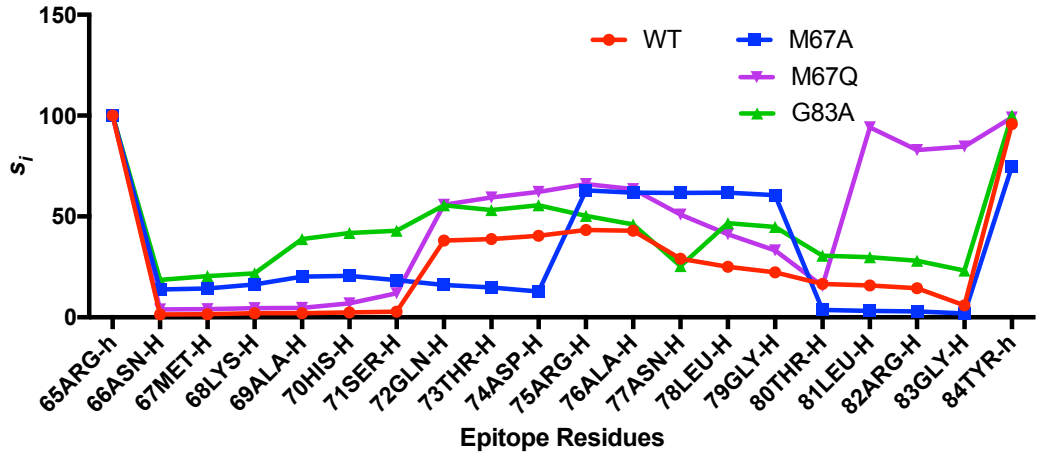


Figure 5.31: 1W72 – extended α-helical peptides. The s_i value represents the time each residue in the peptide has spent in its initial conformation during the simulation. The s_i value of WT, M67A, G83A, and M67Q are shown. The mutations, M67A and M67Q, appear to bring stability from positions 73-79 while G83A stabilises the entire peptide to a considerable extent as compared with the WT. The visual analysis of antibody-peptide simulation showed that the peptide did not fall-off the antibody during simulation. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

Table 5.14: 3EFD – extended α-helical mutant peptides. The peptides were simulated for 1000 ns (only once). The WT with end-caps and four hydrophobic residues were studied for the effects of mutations on conformational stability. The S-value for each mutant peptide represents the average time it has spent in its initial conformation. The S-value of approximately 86% was observed for WT, while none of the mutation in the WT peptide resulted in an increase in the stability, and apparently end-capping further destabilised the WT and mutant close to the WT conformation (F148Q and L155Q). The S-value of these and shown in bold and their s_i values are plotted in Figure 5.32.

Caps		**Alanine Mutations**					**Glutamine Mutations**					**S-Value**
Ace/NME	ASN/NH2	L144A	F148A	L151A	M154A	L155A	L144Q	F148Q	L151Q	M154Q	L155Q	
-	-	-	-	-	-	-	-	-	-	-	-	**86.69**
✓	-	-	-	-	-	-	-	-	-	-		**63.19**
-	✓	-	-	-	-	-		-	-	-	-	51.96
-	-	✓	-	-	-	-	-	-	-	-	-	22.56
-	-	-	✓	-	-	-	-	-	-	-	-	54.42
-	-	-	-	✓	-	-	-	-	-	-	-	43.65
-	-	-	-	-	✓	-	-	-	-	-	-	48.26
-	-	-	-	-		✓	-	-	-	-	-	37.27
-	-	✓	✓	✓	✓	✓	-	-	-	-	-	31.96
-	-	-	-	-	-	-	✓	-	-	-	-	48.92
-	-	-	-	-	-	-	-	✓	-	-	-	**70.95**
✓	-	-	-	-	-	-	-	✓	-	-	-	17.70
-	✓	-	-	-	-	-	-	✓	-	-	-	60.41
-	-	-	-	-	-	-	-	-	✓	-	-	34.07
-	-	-	-	-	-	-	-	-	-	✓	-	18.90
-	-	-	-	-	-	-	-	-	-	-	✓	**79.55**
✓	-	-	-	-	-	-	-	-	-	-	✓	70.02
-	✓	-	-	-	-	-	-	-	-	-	✓	59.64
-	-	-	-	-	-	-	✓	✓	✓	✓	✓	30.11

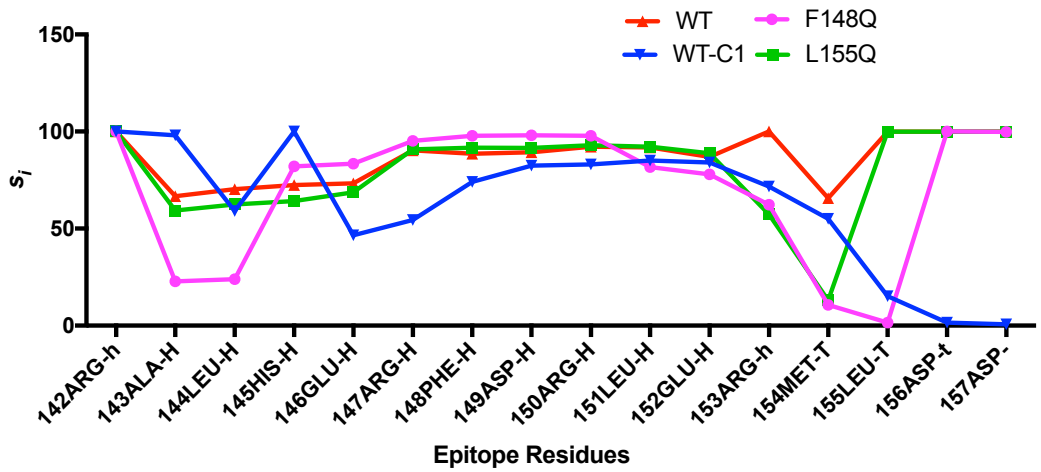


Figure 5.32: 3EFD – extended α-helical peptides. The s_i value represents the time each residue in the peptide has spent in its initial conformation during the simulation. The s_i values of WT, WT-C1 (Ace/NME cap), F148Q and L155Q are shown. The stabilising effect of F148Q is observed at positions 145-150. The WT peptide did not fall off when simulated in the presence of the antibody but was not conformationally as stable as it was on its own. The explanation of secondary structure labels on x-axis is given in section 5.2.5.

5.3.5 Summary of Extented Regions Simulation

The data from these 5 extended peptide simulations show that it is possible to explore potentially stabilising mutations by substituting hydrophobic amino acids with hydrophilic. Of these 5 peptides, 3EFD was an exception where none of the mutations helped to retain the native conformation and the WT itself was the most stable. As shown in Figure 5.28, the S-value increase between the WT (blue) and the most stable mutants (red), for the other 4 peptides, is observed to be between 10–20%. The effect of mutation on the stability of an individual amino acid was investigated by taking account of s_i values of individual amino acids in the peptides. Interestingly, among these five sets of epitope regions, three are stabilised by methionine to alanine substitutions at either the N or C-terminus. This suggests that the long hydrophobic side chain of methionine may have been involved in the collapse of the native conformation. This has been further investigated by looking at the relative solvent accessible surface area (SASA) of methionine and alanine during MD simulations of WT and M154A peptides Table 5.15. By looking at the SASA at different times, it is seen that the methionine in the WT peptide appears to be very mobile and is going through phases where it is clearly trying to bury itself and then being exposed, presumably suggesting that the rest of the structure is destablised when the methionine is buried. However, the SASA of alanine does not change very much suggesting a more stable conformation.

The effect of end-caps on N and C-termini was also analysed. End-capping mostly destabilised the WT and mutants in all of these epitope regions with the exception of 4M48 where both the end caps on the WT showed an increase in S-value while this was not seen in the capped mutant peptides. One would expect that if the cap has a stabilising effect on the WT, a similar effect should be seen in the

Table 5.15: The relative solvent accessible surface area of methionine and alanine during MD simulations of WT and M154A peptides (extended helical peptide - 2W9E).

Time (ns)	Methionine	Alanine
0	60.82	53.46
100	40.45	45.49
200	94.79	37.94
300	98.76	38.07
400	113.98	30.21
500	29.51	40.58

capped mutants. However, these observations suggest that end-capping of extended epitope regions does not help in stabilising the native conformation.

10 replicates were obtained for all the mutant peptides of 2W9E which was selected for experimental validation of their conformational stability. In this epitope region, M154A was found to be a stabilising mutation, and consistent results were obtained in replicate experiments suggesting the effect of the mutation was real. A selection of interesting peptides were also simulated in the presence of antibody showing that all mutants appeared able to bind. However, replicates could not be produced for 4M48, 3P30, 1W72 and 3EFD because of time constraints and, therefore, statistical analysis was not possible. While it is hard to conclude the real effect of stabilising mutations from a single experiment, the evidence from 2W9E suggests that the single 1000 ns simulations are reflective of the 10x500 ns replicates suggesting that all but 3EFD could be stabilised. Combining these results with other strategies, such as stapling and terminal extension of epitope regions (by inclusion of non-epitope residues) more experiments can be planned in the future to explore conformational stability.

5.4 Discussion

As described in sections 5.3.2 and 5.3.5, MD simulations have been used as a tool to explore the conformational stability of 5 folded and 5 extended epitope regions. Moreover, the effects of mutations have been studied which provides structural insight for exploring stabilising mutations. The end-capping of both types of conformations did not help in stabilising the peptides in general and the termini in particular. In the case of the folded conformation, although they did not improve it, caps did not significantly destabilise the structure, unlike the extended conformation where the caps have actually disrupted the WT conformation in most of the examples. While the caps do not seem to help in stabilising the folded conformations, their addition at the ends of the isolated peptides may be advantageous since they provide exo-peptidase resistance.

The alanine and glutamine mutations in the folded conformation did not have any significant effect on the stability whereas in the extended conformation both types of mutations have contributed to an increase of 10–20% in the stability (Figure 5.28). In several studies, it was found that alanine has the highest intrinsic preference for the helix interior [144, 145]. This effect was observed in extended peptides where methionine to alanine substitutions showed an increase in the S-value in 3 peptides. Furthermore, alanine's tendency to form α-helices [143] is supported here in isolated peptides. The stabilising effect of methionine to alanine substitutions suggests that the smaller side chain of alanine may have prevented the collapse of the native conformation.

While effective in stabilising folded peptides, the use of disulphide stapling experimentally might bring some problems, i.e. some difficulty in ensuring that peptides are actually stapled rather than forming polymers (and the possibility that

this would occur over time). However the review by Fairlie and Araujo [183] suggests that it has been done successfully. The possibility of cysteine polymerisation can be ruled out by applying other chemical stapling techniques that have been adopted for α-helix stabilisation. These approaches involve side-chain crosslinking via hydrocarbon [184], triazole [185], lactam [186] and azobenzene [187] staples.

In this work, the cyclisation (by disulphide bonds and glycine linkers) and terminal extension of folded peptides have been demonstrated to help retain the native conformation of multiple epitope regions suggesting that any epitope region in a folded conformation can be studied experimentally using feasible stapling approaches without going through extensive molecular dynamics simulations.

Chapter 6

Experimental Studies of Epitope Regions

Overview

This chapter presents the work that was undertaken at UCB Celltech to validate experimentally the conformational stability of 2 epitopes (along with their selective mutant and derivative peptides) and their binding with antibody. Two sets of epitopes were selected for experimental work: 1) a folded β-loop-β (two β-strands joined by a loop) epitope, and 2) an extended α-helical epitope. Epitopes were synthesized as synthetic peptides and their secondary structure stability was studied using circular dichroism (CD) and nuclear magnetic resonance (NMR). Two Fabs were expressed and purified to study antibody specific binding with each epitope. The binding of antibody and epitopes was studied by surface plasmon resonance (SPR) and ELISA assays.

The two selected peptides represent epitopes from the Hepatitis C coat glycoprotein, E2 and the human prion protein. Hepatitis C infection, caused by Hepatitis C virus (HCV), is a global health concern that affects a population of over 160 mil-

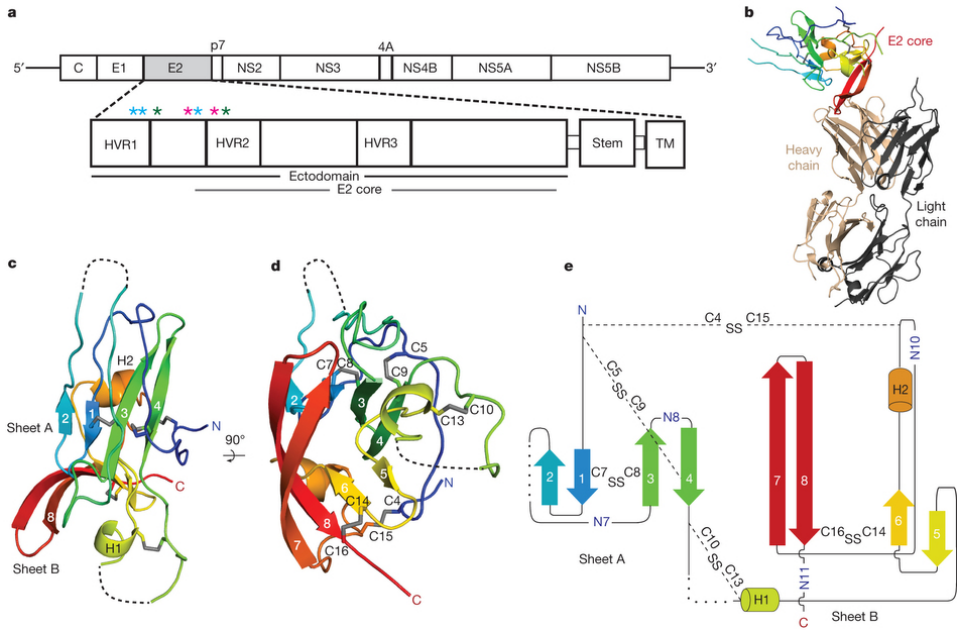


Figure 6.1: a) HCV genome and E2 domain organisation, b-d) Ribbon diagram of the E2 core domain bound to Fab 2A12, e) Topology diagram of E2 core domain, detailing secondary structure elements, disulphide bonds. Figure has been taken from Nature letters [181].

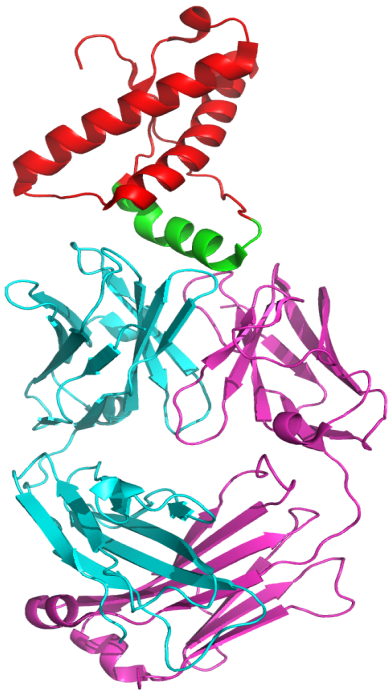
lion [188]. So far, there is no vaccine available to protect against this infection. HCV is an enveloped virus with two surface glycoproteins, E1 and E2. E2 provides the receptor binding site for the host cells and is found to be a target for neutralizing antibodies [189, 190]. The E2 core domain is comprised of β-strands and random coil with two small α-helices [181] and it binds with a Fab fragment via an epitope at positions 631-645 (β-strand 7 and 8 in Figure 6.1). The structure of this core domain could provide insights into HCV entry and will assist in HCV vaccine development (PDB code 4WEB) [181].

A

```
        10         20         30         40         50
MANLGCWMLV LFVATWSDLG LCKKRPKPGG WNTGGSRYPG QGSPGGNRYP
        60         70         80         90        100
PQGGGGWGQP HGGGWGQPHG GGWGQPHGGG WGQPHGGGWG QGGGTHSQWN
       110        120        130        140        150
KPSKPKTNMK HMAGAAAAGA VVGGLGGYML GSAMSRPIIH FGSDYEDRYY
       160        170        180        190        200
RENMHRYPNQ VYYRPMDEYS NQNNFVHDCV NITIKQHTVT TTTKGENFTE
       210        220        230        240        250
TDVKMMERVV EQMCITQYER ESQAYYQRGS SMVLFSSPPV ILLISFLIFL

IVG
```

B



C

Alias	Mutation	Epitope Sequence (142-156)
WT	None	GSDYEDRYYRENMHR
WT-C2	WT-Cap2	N-GSDYEDRYYRENMHR-amide
M154A	M154A	GSDYEDRYYRENAHR
M154A-C1	M154A-Cap1	Ace-GSDYEDRYYRENAHR-NME
M154A-C2	M154A-Cap2	N-GSDYEDRYYRENAHR-amide

Figure 6.2: A) Human prion protein sequence (253 amino acid). The red coloured sequence corresponds to the structure in the PDB (125-223). The epitope sequence (142-156) is shown in green. B) Human prion protein (red) in complex with the Fab fragment of monoclonal antibody ICSM 18 (PDB code: 2W9E). The epitope in the 3D structure is shown in green. C) The WT epitope along with its mutant and derivative peptides. Cap1 refers to acetyl and methyl amide groups at N and C terminus, respectively. Cap2 refers to an additional Asparagine at the N terminus and an amide group at the C terminus.

Human prion protein (PrP) encompasses an epitope at positions 142-156, which provides a binding site for Fab fragment of monoclonal antibody ICSM 18 (PDB code 2W9E) [182]. In the native protein, this epitope is comprised of an α-helix (shown in Figure 6.2).

6.1 Materials and Methods

6.1.1 Reagents and Materials

All reagents and materials used were supplied by Sigma Aldrich (UK) and Fisher Scientific (UK) unless otherwise specified. All the experimental resources (reagents

Table 6.1: Vectors used for DNA transformation and protein expression

Plasmid	Description
ICSM18-pMmFabnh	ICSM 18 anti-PRP therapeutic Fab heavy chain sequence
ICSM18-pMmCk	ICSM 18 anti-PRP therapeutic Fab light chain sequence
4WEB-pMmFabnh	Mouse Fab heavy chain sequence
4WEB-pMmCk	Mouse Fab light chain sequence

and equipment) were used at UCB labs (Slough, UK).

6.1.2 Vectors

Table 6.1 provides a list of plasmids that were cloned for light and heavy chain sequences of 2 different Fabs, synthesised by DNA2.0 Inc. (now called ATUM, USA). The Fasta sequences of the light and heavy chains of the Fabs (ICSM 18 anti-PRP [182] and Mouse Fab [181]) were taken from the PDB.

6.1.3 Transformation

Plasmids, cloned with the light and heavy chain sequences of Fabs, were supplied by DNA2.0 Inc., as lyophilized pellets. The plasmids were re-suspended in 50 µl water. In order to carry out transformation of these plasmids into competent *E.coli* cells, a 10 µl aliquot of sub-cloning grade XL-1 blue *E.coli* cells (Agilent) was gently thawed on ice. 1 µl of re-suspended plasmid DNA was added to the competent cells followed by ice incubation for 30 minutes. This mixture of plasmid DNA and competent cells was heat shocked at 42°C for 45 seconds and returned back to ice for one minute. 125 µl of SOC medium was added to the cells and incubated at 37°C in a shaking incubator for 90 minutes. 30 µl of the bacterial culture was plated out onto LB-agar plates containing 30 µg/ml kanamycin. The plates were incubated overnight at 37°C on a shaking platform.

6.1.4 Giga Prep

Large scale plasmid DNA preparation for protein expression by transient transfection was achieved using a plasmid DNA Giga prep kit (Qiagen). Briefly, 500 ml of media (250 ml of LB and 250 ml of TY broth), containing 30 μg/ml kanamycin was inoculated with a single transformant colony and grown overnight at 37°C.

The bacterial culture was harvested by centrifugation at 3500 rpm for 10 minutes at 4°C. The bacterial cell pellet was re-suspended in 125 ml of Buffer P1 containing RNAse A. The cells were lysed by the addition of 125 ml of Buffer P2, and the sample was gently mixed by inverting the sealed tube 4-6 times, and incubated at room temperature for up to 5 minutes. Buffer P2 contains NaOH and SDS. The sample was then neutralised by the addition of 125 ml of Buffer P3 containing potassium acetate and mixed gently by inverting the sealed tube 4–6 times. The lysate was incubated at room temperature for 10 minutes then filtered using a QIAfilter Cartridge to remove precipitated chromosomal DNA, RNases and proteins. The eluate, containing plasmid DNA was applied to an equilibrated QIAGEN-tip, and allowed to enter the resin by gravity flow. Plasmid DNA, bound to the resin, was washed with 600 ml of QC Buffer, and then eluted with 100 ml QF buffer. The DNA was precipitated by adding 70 ml of isopropanol followed by centrifugation at 4000 rpm for 30 minutes at 4°C. The DNA was washed with 10 ml 70% ethanol and air dried for 10-20 minutes and redissolved in 1 ml distilled water. The concentration of DNA was determined by absorbance at 260 nm (A_{260}) using nanodrop.

6.1.5 DNA Sequencing

In order to carry out a DNA sequencing reaction, a PCR mix of 7.5 μl reaction volume was prepared, comprising of 1 μl DNA (diluted to 0.5 mg/ml), 2 μl sequencing buffer, 1 μl primer (5 μM), 0.3 μl premix (polymerase, dNTPs, ddNTPs - fluores-

cently labelled), and 3.2 µl sterile dH_2O. The PCR mix was prepared separately for forward and reverse primers. These PCR mixtures were amplified using the PCR parameters given in Table 6.2.

Table 6.2: PCR Parameters

Step	**Temperature (°C)**	**Time (s)**	**Cycle**
1	96	30	1
2	96	10	35
3	50	5	35
4	60	4	35

DNA was then precipitated by the addition of 0.75 µl 3M sodium acetate (pH 4.6) and 19.23 µl of 95% ethanol. Samples were centrifuged at 1500 rpm for 45 minutes, following incubation at room temperature for 15 minutes. The pellets were washed with 70% ethanol, air dried and re-suspended in 10 µl formamide. The samples were analysed on a 3100 DNA sequencer (Applied Biosystems).

6.1.5.1 DNA Sequencing Analysis

Sequencing data were analysed using Vector NTI Advance (version 11). Forward and reverse DNA contigs were assembled and processed by ContigExpress (a package in Vector NTI). The DNA sequence was saved and translated to protein sequence by using the Translate tool.

6.1.6 Transfection

6.1.6.1 Cultivating CHOS-XE

The CHOS-XE cell lines, predominantly UCB proprietary, are used for their capacity to produce biological active complex protein such as monoclonal antibodies efficiently to allow performing early screening for large numbers of drug candidates and selection of proteins of interest and to speed up the decision making pro-

cess. The transfection and expression protocol, developed by UCB cell culture and protein expression lab, was followed. A Suspension of CHOS-XE cells was pre-adapted to CDCHO media (Invitrogen) supplemented with 2mM (100X) glutamax. Cells were maintained in logarithmic growth phase agitated at 140 rpm on a shaker incubator (Kuner AG, Birsfelden, Switzerland) and cultured at 37°C supplemented with 8% CO_2.

6.1.6.2 Electroporation Transfection

Prior to transfection, the cell numbers and viability were determined using a CEDEX cell counter (Innovatis AG. Bielefeld, Germany). Viability of 99.8% ($2x10^8$ cells/ml) was measured. The cells were harvested at 1400 rpm for 10 minutes. The pelleted cells were re-suspended in sterile Earls Balanced Salts Solution and spun at 1400 rpm for a further 10 minutes. Supernatant was discarded and pellets were re-suspended to the desired cell density. Vector DNA at a final concentration of 400 μg for ($2x10^8$ cells/ml was mixed and 800 μl pipetted into Biorad cuvettes and electroporated using an in-house electroporation system.

Transfected cells were transferred directly into 3L Erlenmeyer flasks containing PROCHO media enriched with 2 mM glutamax, 0.75 mM sodium butyrate (n-butric acid sodium salt, Sigma B-5887), antibiotic antimitotic (100X) solution (1 in 500), 2.8% Feed A and 0.4% Feed B [PAA]-GE) added at 0 hr. Cells were then cultured in a Kuhner shaker incubator set at 37°C, 8% CO_2 and 140 rpm shaking. The temperature was dropped to 32°C 24 hours post transfection for a further 13 days culture.

6.1.6.3 Harvesting Mammalian Cells

At day 14, cultures were transferred to tubes and supernatant separated from the cells after centrifugation for 30 minutes at 4000 rpm. Retained supernatants were

further filtered through 0.22 μm SARTO BRAN P (Millipore) followed by 0.22 μm Gamma gold filters. Final expression levels were determined by Protein G-HPLC using CDP870hFab 1mg/ml as a standard. The clarified supernatant was stored in the cold room at 4°C ready for purification.

6.1.7 Antibody Purification

6.1.7.1 Protein G Purification of Mammalian Supernatants

The AKTAxpress system (Amersham Biosciences UK Limited) was used for purification of Fabs using a protein G column. The method used for purification of Fabs has been developed by the UCB antibody purification group. As a first step, all the AKTA Basic lines were flushed with 0.1 M NaOH and left for at least 1 hour to remove any endotoxin. The traces of NaOH from lines were removed by flowing with PBS buffer (pH7.4) and 0.1 M glycine-HCl buffer (pH2.7).

To perform affinity chromatography (capture step), a 100 ml protein G (GammaBind Plus Sepharose) column was attached to the AKTAxpress system and cleaned using 2-3 CVs (Column Volumes) of PBS (pH7.4), 6 M GuHCl and 10% methanol at 20 ml/min. The column was pre-equilibrated until stable baselines of UV absorption and conductivity were obtained. The titre supernatant was loaded onto the protein G column at 20 ml/min and fractions collected. The column was washed with 2 to 3 CVs of PBS (pH7.4) at 20 ml/min until the baseline was returned to zero. The elution was then performed using 0.1 M glycine-HCl buffer (pH2.7) at 20 ml/min followed by collection of 10 ml fractions. All the elution peaks were pooled and neutralised by the addition of 1/25 total volume of 2M Tris-HCl pH8.5. These pooled neutralised peaks were analysed on a UV spectrophotometer (Cary 50 UV-Vis, Agilent Technologies) at 280 nm as a protein G pool and on a G3000 HPLC column to determine the monomer/aggregate levels.

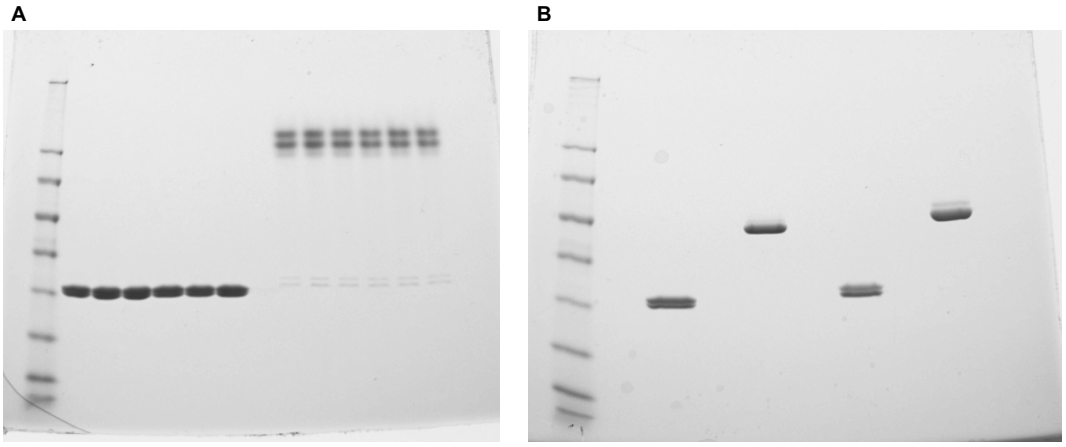


Figure 6.3: SDS Page analysis of ICSM-18 (a therapeutic antibody sequence extracted from PDB 2W9E) and mouse Fab (sequence extracted from PDB 4WEB). A) 6 randomly picked intermediate fractions for each of the Fab B) All the fractions were combined into a final Fab sample. 2 bands (one under reducing and other under non-reducing condition) for each of the Fab showing the successful purification.

6.1.7.2 SEC-HPLC analysis

The analytical size exclusion chromatography (polishing step) was carried out by connecting a 150 ml super loop, containing a protein G pool, to the AKTAxpress system. The fractions collected from the size exclusion run were analysed on G3000 HPLC. Clean fractions were then filtered using a 0.22 μm filter to produce the final sample.

6.1.7.3 SDS-PAGE analysis

The intermediate fractions and final sample (shown in Figure 6.3A and B) were analysed by SDS-PAGE under reducing and non-reducing conditions. A reaction mixture of 200 μl of protein sample, 4 μl of LDS buffer and 2.5 μl of reducing agent was loaded onto the gel. To monitor the protein migration on the gel, a constant voltage of 200 V was applied for 45 minutes. The gel was stained with Coomassie Blue followed by de-staining. The stained gel was analysed by an imaging system (ImageQuant LAS 4000, GE Healthcare Life Sciences).

Table 6.3: Folded Peptide — Epitope from 4WEB (antibody-antigen complex)

Alias	Peptide Mutation	Peptide Sequence (631–645)	Molecular Mass (Av)
WT	None	FKIRMYVGGVEHRLT	1806.16
V637A	V637A	FKIRMYAGGVEHRLT	1778.11
WTX	WT-Extended Ends	NYTIFKIRMYVGGVEHRLTAACN	2657.090
WTG	WT-Cyclised	Gly-YTIFKIRMYVGGVEHRLTAACN	2582.023

Table 6.4: Extended Peptide — Epitope from 2W9E (antibody-antigen complex)

Alias	Peptide Mutation	Peptide Sequence (142–156)	Molecular Mass (Av)
WT	None	GSDYEDRYYRENMHR	1991.064
WT-C2	WT-Cap2	N-GSDYEDRYYRENMHR-amide	2121.210
M154A	M154A	GSDYEDRYYRENAHR	1930.970
M154A-C1	M154A-Cap1	Ace-GSDYEDRYYRENAHR-NME	2003.060
M154A-C2	M154A-Cap2	N-GSDYEDRYYRENAHR-amide	2061.090

6.1.8 Peptide Synthesis

Tables 6.3 and 6.4 show the peptide sequences that were synthesised by PeptideSynthetics (Peptide Protein Research, Ltd. UK). These synthetic peptides were purified in Acetonitrile and water containing 0.1% Trifluoroacetic acid prior to lyophilisation and provided as lyophilised off-white powdered solid. The purity of these compounds was measured by HPLC and mass spectrometry was performed to confirm the molecular masses. The purity of all the synthetic peptides was found to be >98%. The mass spectrometry data agreed with the theoretical molecular masses and is shown in Tables 6.3 and 6.4.

6.1.9 Circular Dichroism (CD) Spectroscopy

In order to measure secondary structure of the peptides, circular dichroism spectra were recorded using a Chirascan CD Spectrometer (Applied Photophysics Ltd., UK). Far UV CD spectra were recorded between 185 and 260 nm using 1 nm bandwidth, 0.2 mm pathlength, 1 second per 0.5 nm sampling time (time per point taken)

and 20°C temperature. Peptide samples (100 μM) were prepared in 20 mM sodium phosphate buffer, pH7, 150 mM NaF. Three separate peptide sample dilutions were prepared to record the spectra. Three scans for each sample dilution were taken. Hence, 9 spectra for each peptide were collected to apply statistics (average and standard error) to the helix fraction. 10 mM peptide stock solutions were made in water and the peptide concentration was measured by UV absorbance using Nanodrop.

Raw CD spectra (in millidegrees/mdeg), after blank subtraction of buffer, were converted to mean residue ellipticity (in degrees squared centimetre per decimole $[\theta]$) by using Equation 6.1 [191].

$$[\theta]_m r = \frac{10^6 * mdeg([\theta])}{cdN} \tag{6.1}$$

where c is the peptide concentration in micromolarity, d is the path length in millimetres and N is number of residues in the peptide. For helical peptides, fraction helix values were obtained, at 222 and 208 nm, by converting mean residue ellipticity ($[\theta]_{obs}$) by using Equation 6.2 [191].

$$F_{helix} = \frac{[\theta]_{obs} - [\theta]_{coil}}{[\theta]_{helix} - [\theta]_{coil}} \tag{6.2}$$

where $[\theta]_{helix}$ shows the mean residue ellipticity of a perfect helix $[-42500\,(1-(3/n))]$ where n is the number of residues in the peptide and $[\theta]_{coil}$ represents a perfect random coil $(+640)$ [191].

6.1.10 NMR

6.1.10.1 NMR Sample Preparation

The helical peptides were dissolved in NMR buffer (20 mM sodium phosphate, pH6.4, 450 µl) and 10% D_2O (50 µl) and mixed by vortexing. The samples were spun down and transferred to 5 mm NMR tubes ready for analysis. NMR samples were stored at 4°C when not in use.

6.1.10.2 NMR Data Acquisition

A Bruker 600 MHz Avance III HD spectrometer with a 5 mm QCI-F cryoprobe was used for all NMR experiments. The data were acquired at 37°C. NMR spectra were obtained from 500 ml samples of 2 mM peptides. 2D spectra were acquired to make sequence-specific ^{15}N, ^{13}C and ^{1}H resonance assignments. 2D spectra for ^{1}H-^{1}H TOCSY [192] with a mixing time of 60 ms and ^{1}H-^{1}H NOESY [193] with an NOE mixing time of 400 ms were recorded. In both of these experiments, the acquisition times for ^{1}H were 85 ms in F_1 and 300 ms in F_2. The data were collected for approximately 9-10 hours (16 scans).

2D spectra for ^{1}H-^{15}N HSQC [194] and ^{1}H-^{13}C HSQC [194] were also recorded in 10% D_2O to identify non-exchanged amide and carbon protons, including hydrogen bonded amide and carbon protons. For the ^{1}H-^{15}N HSQC experiment, acquisition time for ^{15}N was 55 ms in F_1 and 90 ms in F_2 for ^{1}H. The data for this experiment were collected for nearly 11 hours (128 scans). In the ^{1}H-^{13}C HSQC , acquisition time for ^{13}C was 9 ms in F_1 and 80 ms in F_2 for ^{1}H. The experiment was performed for about 3 hours (32 scans). The spectra were processed using the Topspin 3.2 software (Bruker BioSpin).

6.1.10.3 NMR Data Analysis for Sequence-Specific Assignments

Spectra were analysed using the Sparky package (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco).

6.1.10.4 Secondary Structure Prediction from NMR Chemical Shifts

In proteins, a strong correlation has been observed between protein local structure and chemical shifts and hence the chemical shifts can be used to derive the secondary structure. The Chemical Shift Index (CSI) method [195] was used to predict the type of secondary structure of each residue in the peptides. The CSI method predicts secondary structure by comparing ^{13}C (Cα and Cβ) and ^{1}H (Hα and HN) chemical shifts to those of random coils. The difference between observed and experimental chemical shifts was calculated manually using Equation 6.3.

$$\Delta\delta = \delta_{observed} - \delta_{random\ coil} \tag{6.3}$$

The experimental chemical shifts for random coils were taken from the literature [83]. The direction of the chemical shift difference, upfield or downfield of the random coil values, is indicative of atoms being located in areas of α-helix or β-strands. For extended helical peptide, Hα, HN, Cα and Cβ chemical shifts were used to predict secondary structure. A simple algorithm is used for secondary structure analysis. If the observed chemical shift is greater than the experimental chemical shift value of the random coil then a number of '+1' is assigned to that residue; '-1' is assigned if the observed chemical shift is smaller; '0' is assigned if the observed value is found to be within the expected range. An alpha helix is designated when four or more consecutive residues are found with '-1' Hα and/or '+1'

Cα. In contrast, a beta-strand is designated when three or more sequential residues are assigned with '+1' Hα and/or '-1' Cα. Any other regions are defined as coil.

6.1.11 Mass Spectrometry

The molecular masses of synthetic peptides were confirmed by mass spectrometry, was done using a Synapt G2 instrument with electrospray ionisation (ESI) and the QTOF detection method (developed at UCB Celltech labs). The experiment was performed by Andrew Ball at UCB.

6.1.12 ELISA

An ELISA assay was performed to validate the binding of antibody to peptide. ELISA plate wells were coated with 100 µl of 10 µg/ml (5 fold dilution series) peptide in 100 mM sodium carbonate coating buffer and left overnight at 4°C. After incubation, wells were washed three times with 300 µl of PBS/T. After washing, 300 µl of blocking buffer, PBS containing 1% BSA (PB), was added into wells and left for one hour at room temperature. After another wash step, 100 µl of 10 µg/ml antibody was added to the wells for one hour at room temperature. The wash step was repeated and 100 µl of peroxidase-conjugated AffiniPure $F(ab')_2$ fragment goat, anti-mouse IgG $F(ab')_2$ fragment specific conjugate (1:5000 in PBS) was added to the wells and left at room temperature for one hour. The detection antibody, used in this experiment, was supplied by Jackson ImmunoResearch. After a final wash step, 100 µl of TMB substrate was added and colour development was quenched with 50 µl of 2 M H_2SO_4 before absorbance values were read at 450–630 nm. ELISA samples were prepared in duplicates.

6.1.13 Surface Plasmon Resonance (SPR)

Kinetics and affinities of antibody binding to peptides were determined by surface plasmon resonance (SPR) on a Biacore T200 (GE Healthcare, UK) at 20°C with HBS-EP+ buffer 1x (0.01 M HEPES, 0.15 M NaCl, 0.003 M EDTA and 0.05% v/v Surfactant P20) as a running buffer. HBS-EP+ buffer 1x was prepared from HBS-EP+ buffer 10x stock

6.1.13.1 Fab Immobilisation on Sensor Chip

A CM5 sensor chip (GE Healthcare, Uppsala, Sweden), with four flow cells, was used to immobilise Fab (diluted in 10 mM Acetate pH 5.0 buffer) at different ligand densities as described in detail in Section 6.2.3.1. In general, the chip was prepared for Fab immobilisation. Firstly, all the flow cells on the chip were preconditioned by injecting twice 10 µl of 100 mM HCl, 50 mM NaOH, 0.5% SDS and 100 mM H_3PO_4 at a flow rate of 100 µl/min for 10 seconds. Secondly, flow cell 1 (the blank reference) was activated by injecting the mixture of EDC/NHS at a flow rate of 10 µl for 7 minutes and then deactivated with Ethanolamine at a flow rate of 10 µl for 10 minutes to provide a reference surface. Thirdly, the remaining 3 flow cells were activated with EDC/NHS followed by Fab injections. Different flow cells were immobilised with different ligand densities by varying the Fab concentration, injection time or acetate buffer pH. Finally, deactivation of flow cells was performed by Ethanolamine to cap the carboxymethyl ends. In order to study kinetics, peptides were diluted in HBS-EP+ running buffer at concentrations in the range of 3.125 nM to 100 nM (two-fold dilution series) and injected at a flow rate of 30 µl/min for 120 seconds over the immobilized Fab. Blank sensorgrams were also recorded by injection of the same volume of HBS-EP+ buffer over the immobilized Fab. Replicate data were collected at the 50 nM concentration. During method optimization,

it was observed that the peptide dissociated completely in 40 minutes. Therefore, a time of 40 minutes was given between each injection to dissociate the peptide completely from the chip surface and chip regeneration was not needed. Association and dissociation kinetic rate constants were calculated by fitting the data on the sensorgrams of the concentration series with the BIAevaluation software using the 1:1 binding model with RI = 0. This model was chosen because it is known that the interaction is expected to be 1:1 and the peptide only has one epitope for the Fab.

6.1.13.2 Peptide Immobilisation on Sensor Chip

Different biotin-conjugated peptides were immobilised on different flow cells of streptavidin-coated sensor chip SA (GE Healthcare, Uppsala, Sweden). For biotin conjugation, a reaction mixture was prepared by addition of 10 mM Biotin (NHS-PEG4-Biotin in 50 mM sodium phosphate, pH6.5), 10 mM peptide (fresh stock solution was prepared owing to precipitation/insolubility problems) and 50 mM sodium phosphate, pH6.5 buffer. The ratio of biotin to peptide (2:1) and the reaction pH were chosen to favour labelling of only the N-terminal amino group, therefore leaving the remainder of the peptide accessible to bind the Fab. A blank reaction was also prepared in a similar way replacing peptide with phosphate buffer. This reaction mixture was incubated for 24 hours at 4°C. The reaction was quenched by the addition of 0.5 M Tris-HCl, pH7.5 and left for 1 hour at room temperature.

The streptavidin-coated sensor chip was cleaned three times with 1 M NaCl with 50 mM NaOH at a flow rate of 100 µl/min for 60 seconds. The blank reaction mixture was diluted in HBS-EP+ buffer and injected into the reference flow cell while 100 nM biotin-conjugated peptides (diluted in HBS-EP+ buffer) were injected into active flow cells at a flow rate of 5 µl/min. The blank reaction mixture was injected to the same total volume as the peptides because the remaining free biotin

in the peptide reactions which has not become attached to peptide will also bind to the streptavidin on the chip surface. It is therefore important that the same amount of free biotin is attached to the reference flow cell in case this has an impact on the binding of the Fab when it is injected. An immobilisation of 30-75 response units was obtained for different peptides on different flow cells. In order to study affinity of these peptides with antibody, different concentrations of Fab in the range of 800 μM down to 12.5 μM were prepared by serial dilution in HBS-EP+ buffer and injected over the immobilised peptides. Blank sensorgrams were also recorded by injection of the same volume of HBS-EP+ buffer over the immobilized peptides. The affinity between antibody and peptides was calculated with the BIAevaluation software using the steady state 1:1 affinity model because binding was not strong enough to allow kinetic analysis.

6.2 Results - Folded β-Strand Epitope

The epitope has been isolated from the E2 core domain (shown in Figure 6.1, PDB code 4WEB) and synthesised along with its mutant and derivative peptides (details given in sections 5.1 and 6.1.8). In order to study conformational stability of these peptides, CD spectroscopy was performed.

6.2.1 Peptide Solubility Issues

Unfortunately these peptides had severe solubility issues. Table 6.5 shows the sequences of the isolated epitope and its mutant and derivative peptides along with the number of hydrophobic and charged amino acids. The Grand average of hydropathicity (GRAVY) is a measure of a peptide or protein's hydrophobicity or hydrophilicity [196]. These measures are combined in a hydropathy index and reflect the tendency of a protein to be hydrophobic or hydrophilic. A relatively hydropho-

bic peptide or protein will have a positive value of hydropathy.

Although many peptides may easily dissolve in aqueous solutions, a frequently encountered problem is very low solubility or even insolubility, particularly for peptides with a considerable number of hydrophobic amino acids. Hence, it is a challenging task to determine a suitable solvent to dissolve a hydrophobic peptide. For use in a patient, a very limited choice is available.

In this peptide set, initially water was used to make a stock solution of 10 mM. The WT and mutant (V637A) appeared to dissolve in water. There was no initial precipitation, however the WT peptide appeared to precipitate out after several weeks of refrigeration whereas V637A stayed soluble. The aqueous solution of extended (WTX) and cyclised (WTG) peptides formed a highly viscous and jelly-like substance. According to the manufacturer's peptide solubility guidelines [197], this behaviour could be for two reasons. Firstly, peptides containing unpaired cysteines tend to aggregate and both the WTX and WTG have a cysteine at the C terminus. Secondly, peptides with D, E, H, K, N, Q, R, S, T, Y are capable of intermolecular cross linking (i.e. building hydrogen bonds) [197]. Either of these factors together with the hydrophobicity could have contributed to the jelly-like appearance of these peptides in aqueous solution. Clearly there is a fine balance between hydrophobicity and excess hydrophilic residues that can lead to cross-linking. However, a number of different options (depending on the experiments) were considered to solve the solubility issues that included the use of DMSO (Dimethyl sulfoxide), TFA (Trifluoroacetic acid) and TFE (Trifluoroethanol).

6.2.2 CD Spectroscopy of β-loop-β Epitope

Since the epitope in the E2 core domain exhibits an antiparallel β-sheet structure in the full-length protein, a strong β-sheet signal was expected in isolated epitopes

Table 6.5: Epitope sequence (4WEB) and properties — Number of hydrophobic and charged amino acids along with the hydrophobicity score (Gravy) and solubility in water.

Peptide	Sequence	Number of		GRAVY	Solubility in water
		Hydrophobic	Charged		
WT	FKIRMYVGGVEHRLT	6	4	-0.067	Soluble but precipitated out in a few weeks
V637A	FKIRMYAGGVEHRLT	6	4	-0.227	Soluble
WTX	NYTIFKIRMYVGGVEHRLTAACN	9	4	0.026	Jelly-like
WTG	GYTIFKIRMYVGGVEHRLTAACN	9	4	0.161	Jelly-like

if they retain their native conformation when extracted from the full length protein. The characteristic CD spectrum for antiparallel beta-sheet proteins has a strong negative peak at 218 nm and a positive peak at 195 nm [198]. To allow easy comparison between CD spectra, it is a common practice to convert machine units of millidegrees to mean residue molar ellipticity (MRE) which normalizes the signal for the number of residues and the protein concentration. The unit conversion equation is explained in the methods Section 6.1.9.

6.2.2.1 Conditions and Optimisation of β-loop-β Epitope

CD, on such small isolated peptides, was not commonly performed in the lab at UCB, so conditions needed to be optimised. All the initial optimisation was done on a manual Chirascan CD Spectrometer using a cuvette with 1.0 mm pathlength. Initially, WT and V637A peptides were discovered to be soluble in water and an important question to be answered was whether they were forming their native β-sheet structure in buffer and/or water, or whether a solvent was needed to stimulate structure formation. To this end, CD spectra were taken by diluting peptide in either water, phosphate buffer (20 mM sodium phosphate, 150 mM NaF, pH7.0) or TFE. A challenging part of obtaining high quality CD spectra is the optimisation of protein concentration. To explore the correct peptide concentration, a range of concentrations was tested (5 μM, 10 μM, 25 μM, 50 μM and 100 μM peptide in phosphate buffer). At the higher concentrations, no unexpected drops in signal were observed confirming that it is unlikely that the peptide was aggregating in the analysed concentration range. Moreover, a reasonable CD signal was observed at 50 μM concentration (in 20 mM sodium phosphate, 150 mM NaF, pH7.0), therefore all further optimisation was done by keeping this concentration of peptide constant. The effect of temperature on peptide structure was also studied by performing a time course

experiment (temperature ramp) where a CD spectrum was recorded from 4°C to 90°C. It was observed that there was a steady loss of structure at higher temperature, but it was difficult to characterise with CD. Therefore, 20°C (a generally used temperature for peptides) was chosen to perform the rest of the CD experiments.

The theoretical pI of the WT peptide was calculated as 9.99. So a pH screening in the range of 3.18-10.94 (below and above the pI of the peptide) was performed to observe any change in the signal. No considerable difference in the CD signal was observed at different pH (data not shown). Therefore, all the experiments were performed using phosphate buffer (20 mM sodium phosphate, 150 mM NaF, pH7.0). After CD condition optimisation using the manual spectrometer, the peptide samples were re-run using a Chirascan autosampling system. The cuvette in this system had a pathlength of 0.5 mm therefore, 100 μM peptide concentration was used. The increase in concentration on using a smaller pathlength cuvette is in accordance with Lambert-Beer's law (Section 2.2.1.2).

The condition optimisation was done on WT peptide which initially appeared to be easily solubilised in water, and therefore the solubility issues remained undiscovered until the other peptides (WTX and WTG) were received from the supplier. The WTX peptide formed a gel in aqueous solution while the lyophilized WTG peptide was completely insoluble in water. TFA was tested to solubilise it but unsuccessfully; the peptide was lyophilised and the stock solution for WTG was prepared in 10% TFE instead.

6.2.2.2 CD Spectra of β-loop-β Epitope

Considering peptide solubility issues, CD spectra for all four peptides were obtained in buffer as well as with 10% TFE (Figure 6.4). The WT and V637A peptides appeared as random coil in buffer while the structure was altered on addition of 10%

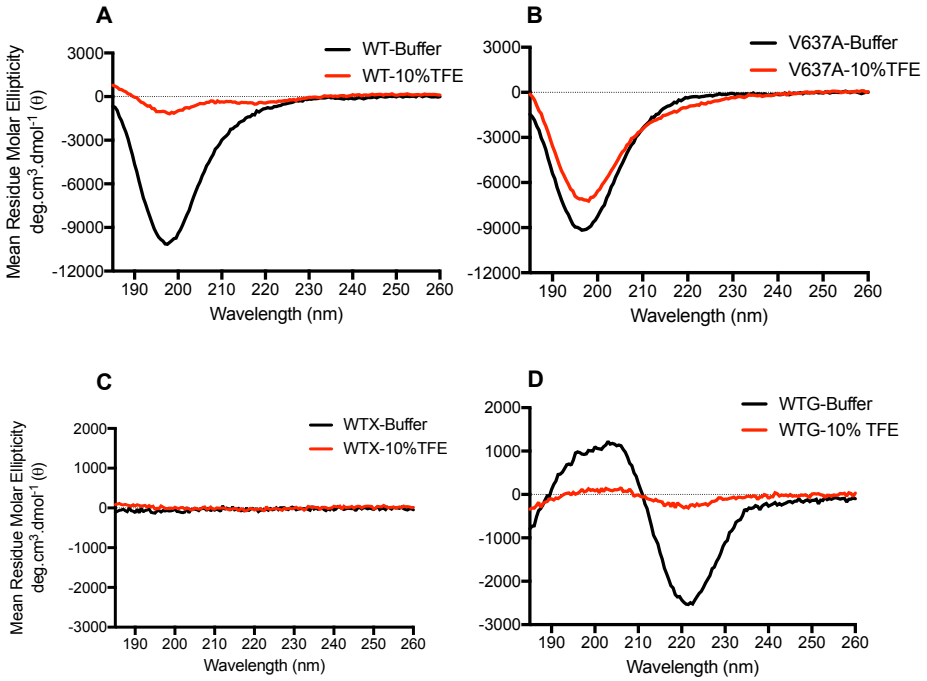


Figure 6.4: Circular dichroism spectra of E2 core domain epitope (631-645) in 20 mM sodium phosphate, 150 mM NaF, pH7.0 buffer and with 10% TFE at 20°C. A) The CD spectra of WT epitope between 185 and 260 nm for a peptide concentration of 100 μM. B-D) The spectra of V637A, WTX (WT with extended ends) and WTG (WT with cyclised ends).

TFE. The WTX peptide showed a spectrum that was equivalent to the blank spectrum. This is because the peptide, that was gel like in stock solution, precipitated out when it was diluted in buffer. Even, the addition of 10% TFE in the buffer could not stop the precipitation of the peptide. The cyclised peptide (WTG) did not precipitate on dilution in buffer and the CD spectrum showed the characteristic spectrum of β-sheet with a positive and negative peak at about 200 and 220 nm respectively (the ideal β-sheet peaks are expected at 195 and 218 nm). The reason for this shifting could be a slight change in conformation due to cyclisation (addition of the glycine linker) of the beta strands. However, the spectrum suggests the presence of beta sheet structure. The structure was affected by addition of 10% TFE. i.e. a considerable decrease in the signal was seen (Figure 6.4).

Having seen some structure in the cyclised peptide (WTG) and no structure in

WT, a TFE titration was performed on these peptides to see the effect of higher concentration of TFE (Figure 6.5). At 50% TFE concentration, the cyclised peptide's peaks were seen to match the perfect β-sheet conformation with characteristic positive and negative peaks at 195 and 218 nm respectively. The same concentration of TFE started to form some sort of structure in the WT peptide. This seems to be evidence of TFE's ability to stabilise the native structure in peptides. However, very high concentrations of TFE can turn a structure into helix and the relevance of using TFE to stabilize peptides for use as immunogens is questionable .

Further CD experiments could have been performed on the cyclised peptide, but owing to time constraints these were stopped. The solubility issue could have been solved by trying different biological solvents, but again the relevance for use as an immunogen is questionable. Instead, NMR provides residue level structural information unlike CD which provides an average estimation of the structural population. The limited time for experimental work was a hurdle to collecting NMR data for this peptide. However, SPR (Surface Plasmon Resonance) experiments were performed to validate the binding of this epitope with Fab fragment 2A12.

6.2.3 Surface Plasmon Resonance (SPR) of Folded β-Strand Epitope and Fab

The full length E2 core domain binds with Fab fragment 2A12 (Figure 6.1). It was expected that SPR would provide binding information of isolated epitopes from the E2 core domain with Fab. Since CD experiments remained inconclusive regarding structural conformation of the isolated epitope, it was hoped that SPR might answer some questions by confirming the interaction. It was expected that if the peptide has adopted its native conformation, it will bind with antibody.

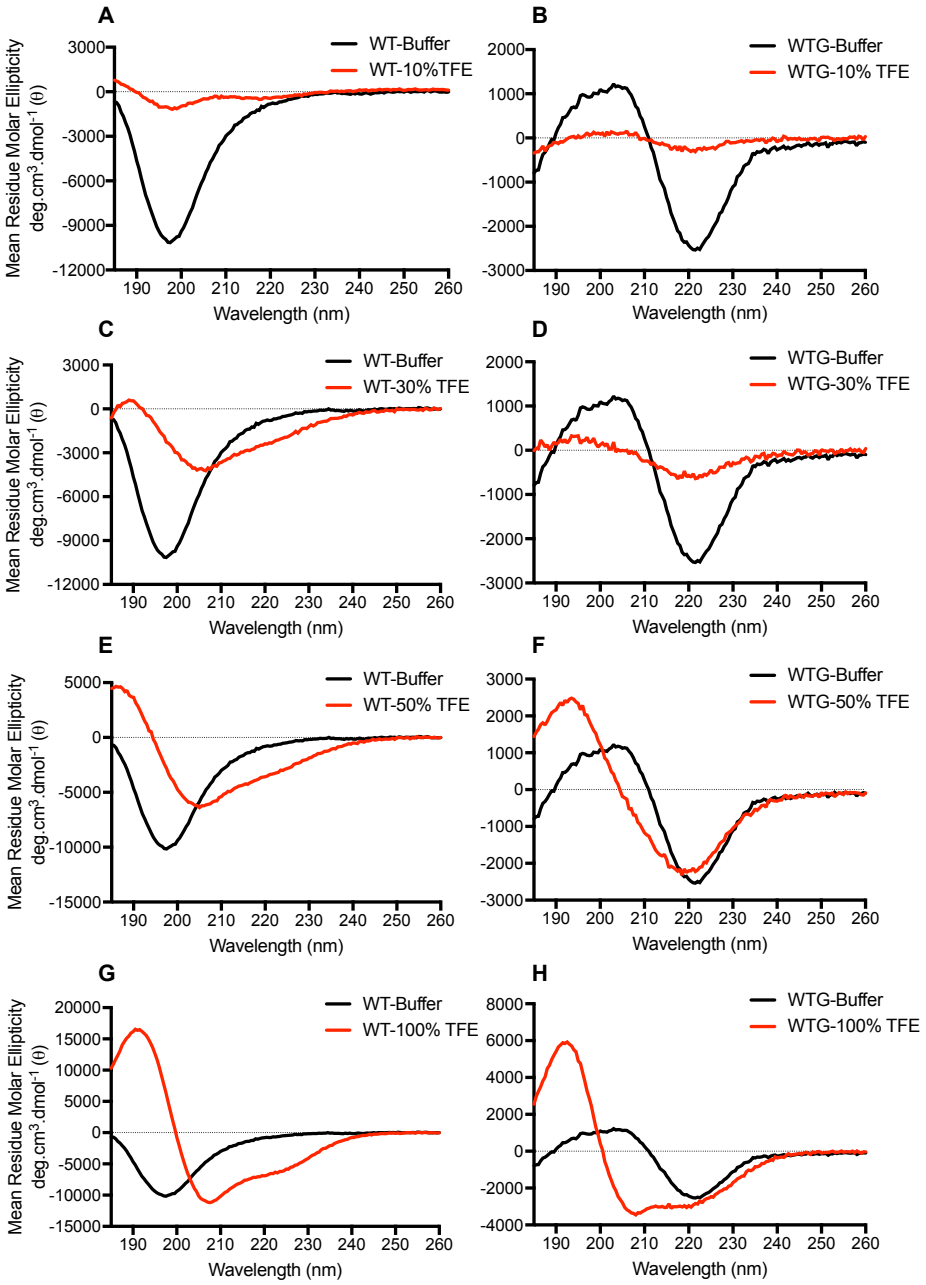


Figure 6.5: Circular dichroism spectra of WT (A, C, E and G) and WTG (B, D, F and H) peptides at different TFE concentrations. For all spectra, 100 μM peptide concentration was used in 20 mM sodium phosphate, 150 mM NaF, pH7.0 buffer at 20°C.

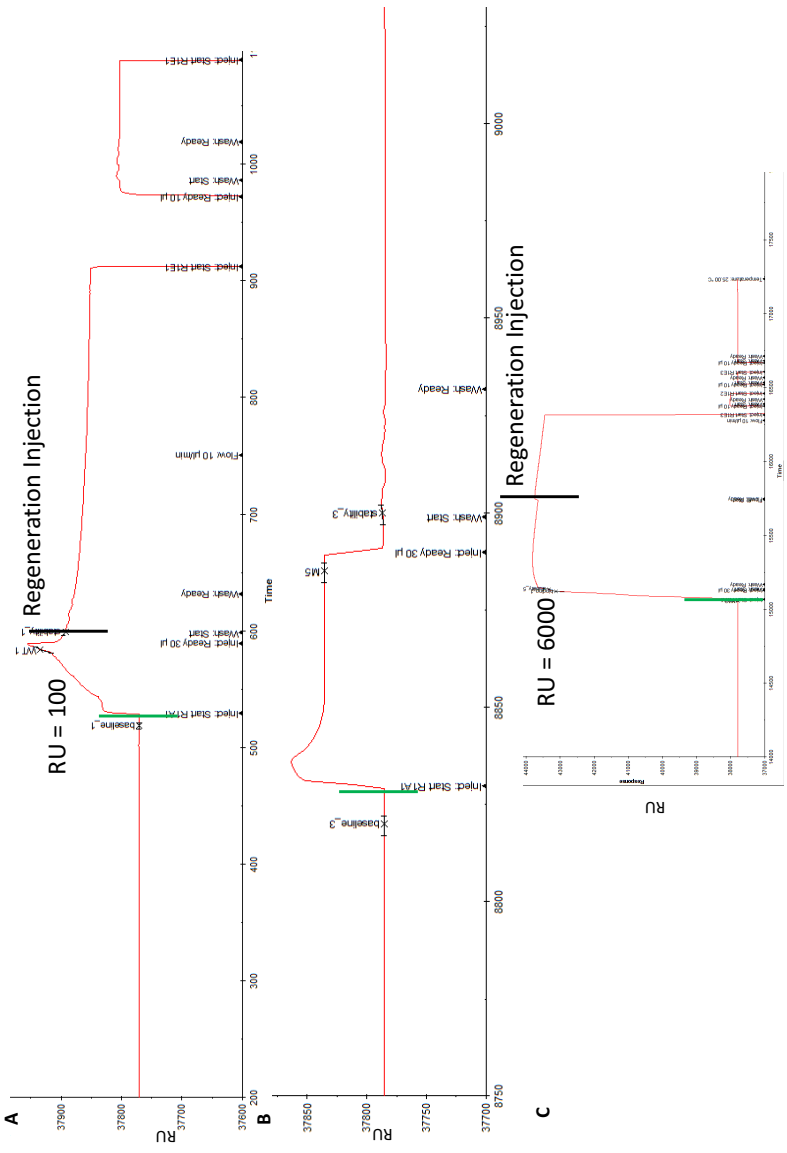


Figure 6.6: Background binding on blank reference flow cell (FC-1). Green bars indicate the injection of peptide while black bars indicate the regeneration injection. A) WT peptide sticking on reference flow cell (100 RU) and not dissociating. Regeneration of the chip was performed to bring the response down to baseline. B) V637A, injected for 60 seconds, dissociates and reaches down to baseline; it does not stick to the chip surface. C) WTX shows reference binding of 6000 RU in 60 seconds. The chip was regenerated to bring the response down to baseline.

6.2.3.1 Fab Immobilisation on the Sensor Chip

In order to study the binding between an antibody and an epitope by SPR, antibody was immobilised on a CM5 sensor chip. The chip was cleaned, activated and capped as explained in the methods. Initially, 100 μg/ml Fab in Acetate buffer (10 mM sodium acetate, pH5.0) was injected into flow cells 2, 3 and 4 of the sensor chip. Flow cell 1 was activated and capped, but Fab was not injected leaving it blank as a reference. It was aimed to immobilise Fab at different target immobilisation levels of 500, 2500 and 5000 response units (RU) for flow cells 2, 3 and 4 respectively. The Fab immobilisation was unsuccessful owing to Fab over-concentration; therefore, to overcome this problem, a 10 fold dilution was performed before trying immobilisation again. By doing so, an immobilisation of about 500 and 1800 response units was obtained on flow cells 2 and 3 respectively. Flow cell 4 was left blank. This amount of Fab, on flow cell 3, was considered enough to check the binding with small peptides. The WT, V637A and WTX peptides were diluted in running buffer (HBS-EP+) at 100 nM concentration. The synthetic WTG peptide was not available at the time of this experiment. To study the association, 100 nM WT peptide was injected into the chip (with immobilised Fab) for 60 sec at a flow rate of 30 μl/min. In an ideal case, there should be no binding on the reference flow cell which confirms that peptide itself is not interacting with the chip's surface. However, in the case of the WT peptide, a background binding of 100 RU was observed on the reference flow cell which suggested that the peptide is making interactions with dextran on the chip surface and is therefore very sticky. Figure 6.6 shows background binding on the reference flow cell for 3 peptides. In spite of observing background binding, the peptides were injected into flow cell 3 (with 1800 RU of Fab) to monitor any possible antibody binding.

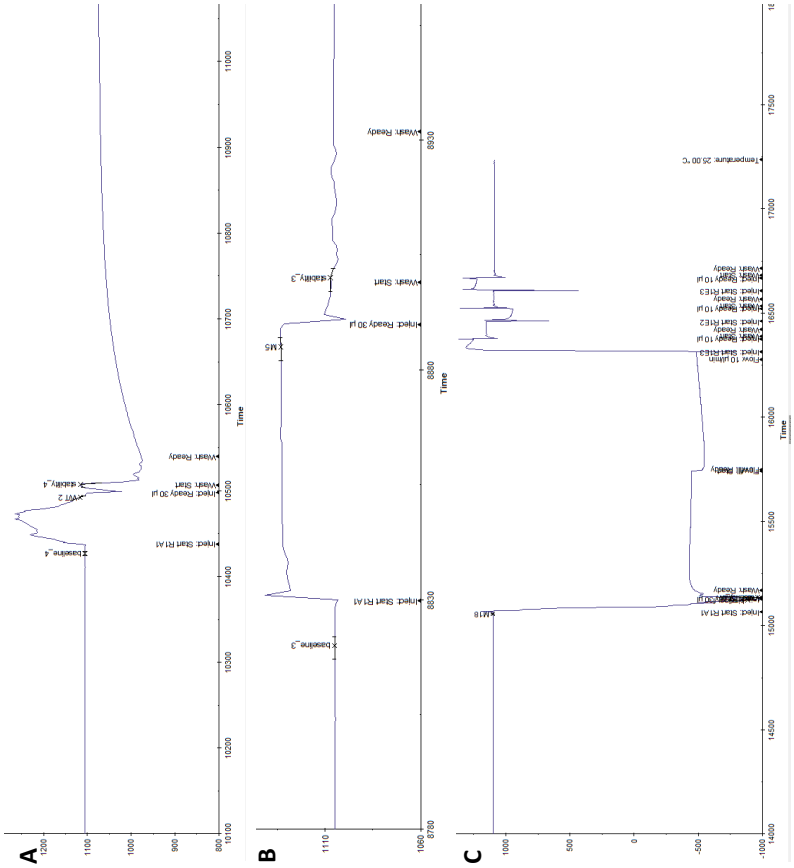


Figure 6.7: Sensorgram to show binding between Fab and epitope on flow cell 3 of a CM5 sensor chip. A) WT peptide injected into the chip for 60 sec at a flow rate of 30 µl/min B) V637A peptide C) WTX peptide.

Figure 6.7 shows the sensorgrams (FC3−FC1, blank subtracted) to monitor the binding on flow cell 3. The sensorgram for WT peptide (Figure 6.7A) shows a signal of weak binding. However, the binding on the blank flow cell suggests that this result might not be real. There is negligible binding seen for V637A. However, WTX has an exceptionally poor sensorgram and this could be due to the peptide's apparent (gel like) insolubility problem. While investigating the background binding on the reference flow cell and the possibility of the peptide reacting with dextran on the CM5 chip, it was decided to use a C1 chip which does not have dextran on the chip surface; the carboxymethyl groups are attached directly to the gold surface. To this end, Fab was immobilised on a C1 chip and peptides were injected into it with the hope of ruling out the possibility of dextran binding with the peptides. Surprisingly, using the dextran-free chip did not stop peptide binding to the reference flow cell.

As another way to investigate dextran interference with the peptide and chip surface, 1 mg/ml and 10 mg/ml of dextran was added to the running buffer (HBS-EP+) and peptide was again injected into the chip. The addition of Dextran in the running buffer should work in almost the exact opposite way to using a C1 chip. The idea is that if the peptide is interacting with dextran on the chip surface, this binding can be prevented by blocking the dextran binding site in the peptide by adding dextran to the running buffer. The dextran should bind to the peptide, thus preventing it from binding to the dextran on the chip surface. However, this approach did not help and background binding was still observed.

Another approach was adopted to overcome the background binding problem by addition of 500 mM NaCl into the running buffer to investigate the peptide behaviour at higher salt concentration. The use of high NaCl concentration is to try to

remove any charge based interaction. Since the chip surface is negatively charged, a non-specific interaction of positively charged analytes with the surface can be obtained. Therefore, the addition of NaCl should shield the charge on the analyte thus reducing this interaction. Unfortunately, this approach also did not stop background binding on the reference flow cell.

At this stage, it was discovered that WT peptide had precipitated out in aqueous stock solution and the WTX peptide's gel-like appearance made us aware of the solubility issues with this set of peptides. Therefore, it was decided to dissolve WT peptide in 100% DMSO and perform SPR using 5% DMSO in the running buffer. For this experiment, a CM5 chip was prepared using the same method (explained in Section 6.1.13.1). 100 μg/ml of Fab was immobilised on flow cell 2 and only WT peptide was diluted to 500 μM in slightly over-concentrated buffer to balance the ratio of DMSO in the running buffer. Therefore, a small amount of over concentrated buffer was prepared separately to dilute the peptide (which was dissolved in 100% DMSO). The peptide was injected into the chip, but similar results were obtained with background binding on the reference flow cell. Other than the solubility issue, high hydrophobicity of the WT peptide could be the cause of the high affinity for the sensor chip, so it was decided to try the opposite approach of immobilising the peptide on the chip surface. To this end, the peptides were biotinylated and immobilised on the chip and the antibody affinity was monitored in the mobile phase. The reason for peptide biotinylation was the short size of peptide. If peptides were directly coupled to the chip surface it is unlikely that they would be able to bind for steric reasons. By adding a biotin tag, the availability of the peptide for binding should be increased as it should be more in solution and more flexible.

6.2.3.2 Peptide Immobilisation on the Sensor Chip

The WT, V637A and WTX peptides were biotinylated and immobilised on streptavidin coated sensor chip SA (as described in methods Section 6.1.13.2). Flow cell 1 was immobilised with blank reaction mixture whereas flow cells 2, 3 and 4 were immobilised with WT, V637A and WTX up to 75, 70 and 30 RUs respectively.

The Fab purified at UCB had a concentration of 33 μM (1.6 mg/ml) but a higher concentration of Fab was needed in the mobile phase of SPR therefore Fab was concentrated and buffer exchange was done to replace the PBS buffer which had been used for Fab purification with 1x HBS-E+ (without Tween P20). Since the running buffer contains 0.05% Tween, the same concentration of Tween was added to the final sample. The UV absorbance of Fab was measured using Nanodrop and concentration was calculated by dividing the absorbance by the extinction coefficient (1.5 for Fab). The concentration of Fab was found to be 38.33 mg/ml. This was converted to Molar units by dividing by the molecular weight of Fab which resulted in 801 μM concentration. Hence, the concentration of Fab was increased from 33 μM to 801 μM.

6.2.3.3 Binding Affinity of Fab and Peptides

The binding affinity of Fab with peptides was assessed by performing a dilution series of Fab in the range of 801 μM down to 12.5 μM. Figure 6.8 shows the binding affinity sensorgram for WT, V637A and WTX. These binding affinity data (i.e. K_D values not being in the instrument's given range) suggest that it might not be possible to monitor the binding kinetics owing to low affinity (Table 6.6).

Steady state analysis as a function of analyte concentrations is the basis for binding affinity determination. The relationship between concentrations of analyte and affinity is described in Section 6.1.13.2. In the Biacore software, affinity

Table 6.6: Dissociation constants of Fab and peptides

Peptide	K_D
WT	$7.99\text{x}10^{-4}$
V637A	$4.20\text{x}10^{-3}$
WTX	$1.07\text{x}10^{-3}$

constants may be calculated using the experimental data from linearised plots (as shown in Figure 6.8) by analysing the data from the plot of the amount of complex against the analyte concentration (as shown in Figure 6.9). The vertical line in these plots represents the value of the calculated equilibrium dissociation constant K_D. A steady plateau curve with a vertical line intersecting the curve within the measured range of analyte concentration is required to conclude that affinity data are valid. Without this, the K_D values calculated by the Biacore software are completely inaccurate because it involves so much extrapolation and therefore could be meaningless. These data show a very weak binding affinity between Fab and WT peptide, and in the case of V637A, there is no binding. An interesting observation was made in the case of WTX where a steady plateau was seen in the range of $12.05-262.8$ μM concentration (Figure 6.9D), but a steep curve for higher concentration (Figure 6.9C). This suggests some sort of binding at lower concentration, but overall affinity data and the K_D value do not prove this. Any conclusion from these data require caution because the experiments were performed only once owing to the solubility issues and limited time at UCB. In order to confirm the reproducible response, more experiments need to be carried out.

6.2.4 Mass Spectrometry

In order to confirm whether the peptides have been successfully tagged with biotin, mass pectrometry was performed on the biotinylated peptides. Figure 6.10 shows that peptides have been biotinylated at up to three positions. In principle, the biotin

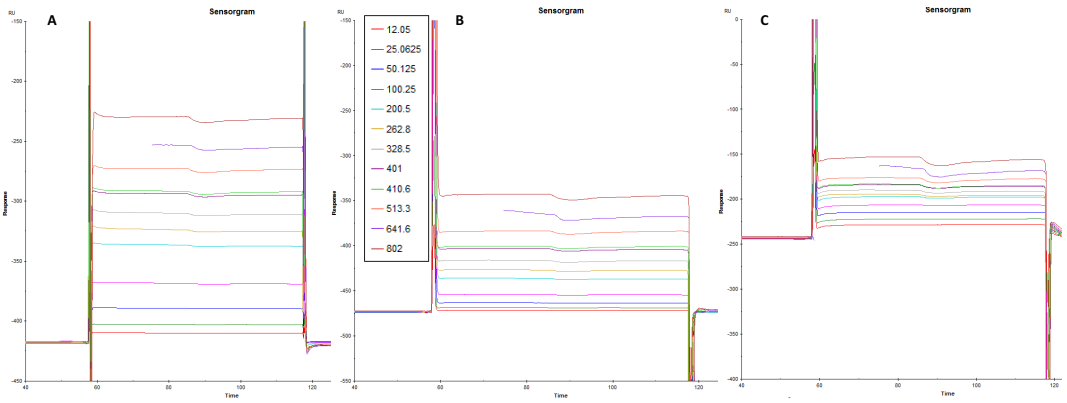


Figure 6.8: Sensorgrams to show binding affinity of Fab with peptides in a concentration range between 801 and 12.05 µM. Some of the data points at 641 µM concentration have been removed because of air in the sample tube. A) WT peptide B) V637A C) WTX

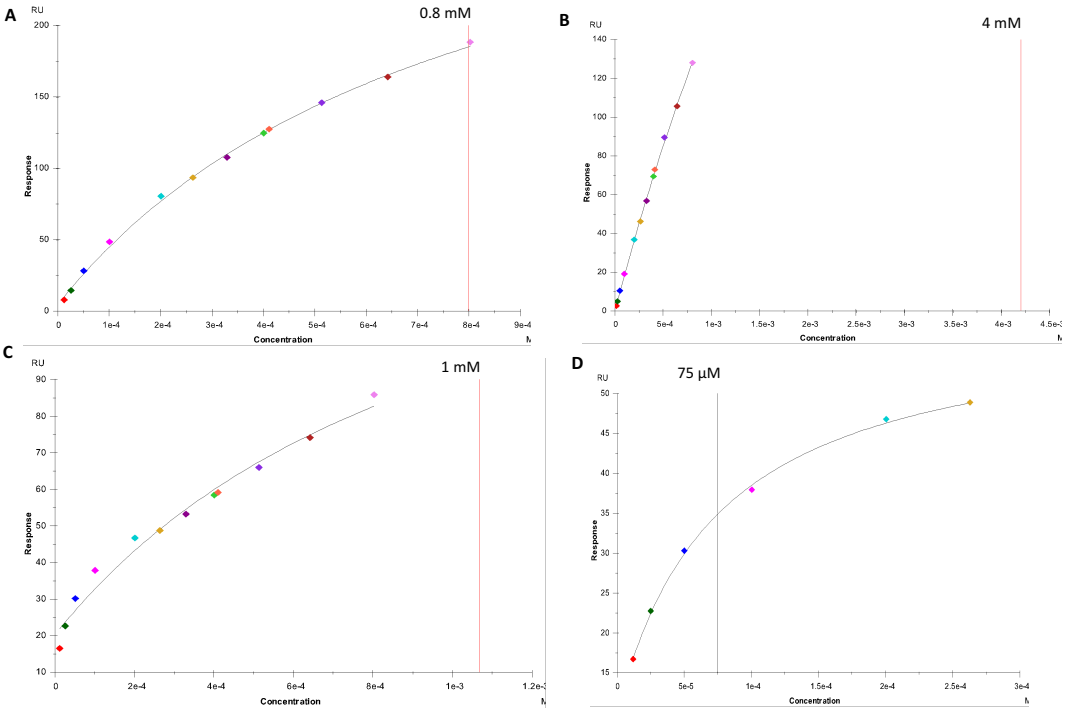


Figure 6.9: Plot of steady state response against analyte concentration using a steady state affinity model for binding affinity determination. The vertical line in these plots represents the value of the calculated equilibrium dissociation constant K_D. A) WT − data points in the range of 802 and 12.05 µM B) V637A − data points in the range of 802 and 12.05 µM C) WTX − data points in the range of 802 and 12.05 µM D) WTX − data points in the range of 262.8 and 12.05 µM.

attaches at amine groups. Considering this and the amino acid sequence of the WT peptide, the probable positions for biotin attachment are the N-terminus, the lysine residue (at the 2nd position) and less readily, at arginine residues (at positions 4 and 13, Table 6.5. One expects the N-terminus and lysine residues to be labelled first followed by labelling of a 3rd biotin at an aginine residue (position 4). However, a very small amount of labelling with a 4^{th} biotin at the second arginine (position 13) was also observed. The peak has not been labelled in the figure because of its very low intensity, however the expected mass for a peptide with a 4th biotin provided evidence for the presence of such peptides.

6.2.5 ELISA

After obtaining inconclusive results from SPR, an ELISA was perfomed to study the potential binding between Fab and peptides (WT and V637A). A half log dilution series, starting from 10 μg/ml (down to 128 pg/ml) concentration of peptide, was performed. 10 μg/ml antibody was used. However, no positive binding data were obtained through ELISA (data not shown).

6.2.6 Summary of Folded Beta Strand Epitope

The WT, V637A, WTX and WTG were studied for conformational stability using CD. In aqueous solution, WT and V637A remained unstructured whereas the WTG retained β-strand conformation. However, because of the peptide solubility issue owing to the presence of 9 hydrophobic amino acids, CD did not work for WTX. The binding of these peptides with the Fab was studied using SPR and ELISA, but the results remained inconclusive. SPR on the WTG peptide could not be performed owing to the limited time at UCB.

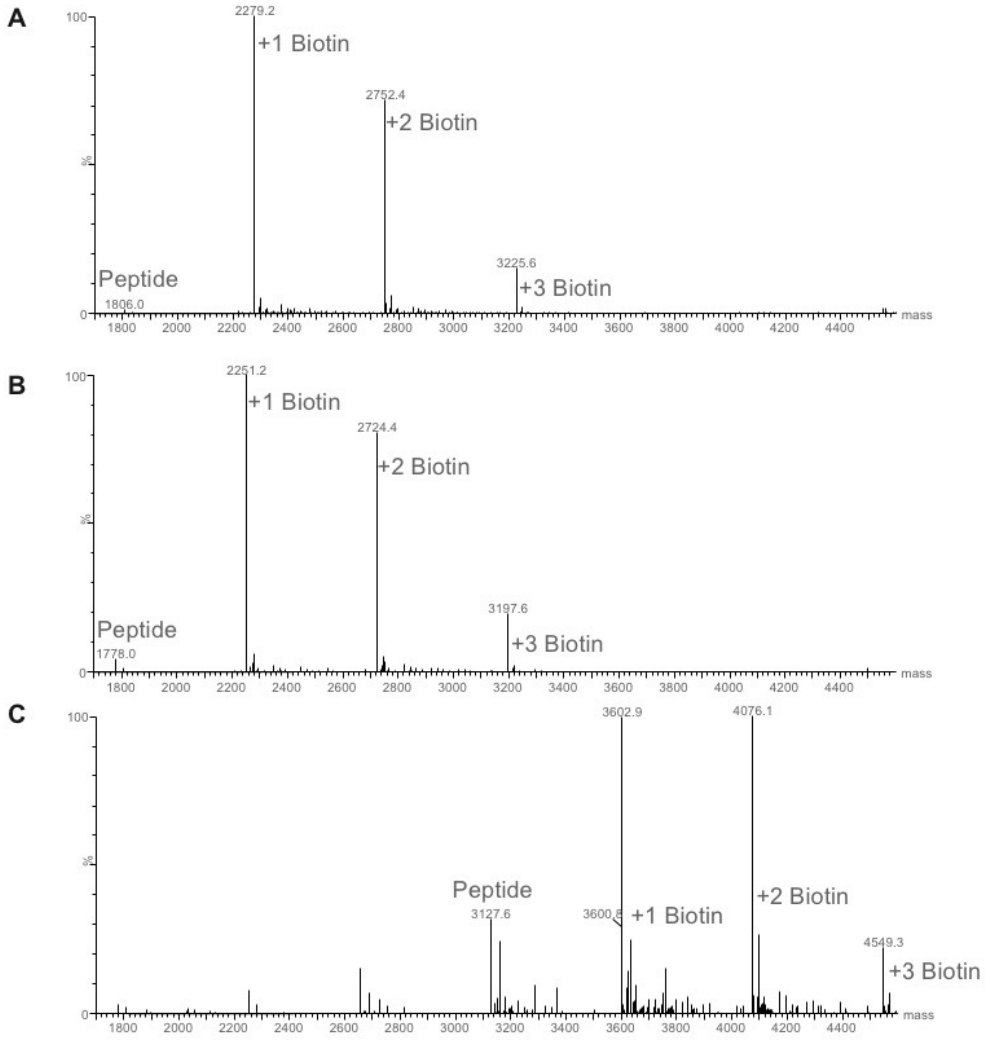


Figure 6.10: Mass spectrometry of biotinlyted peptides. A) WT peptide has been tagged with biotin at three positions, B) V637A, C) WTX

6.3 Results - Extended Helical Peptide

In the native protein, this epitope is comprised of an α-helix (shown in Figure 6.2, PDB code 2W9E). This isolated epitope and its mutant and derivative peptides have been synthetically generated (details given in Section 6.1.8). The secondary structure and conformational stability of these isolated peptides were studied by CD spectroscopy and NMR.

6.3.1 CD Spectroscopy of Extended α-Helical Epitope

In full length human prion protein, the epitope folds into a helical structure (as shown in Figure 6.2). In order to study the conformational stability of this isolated helical epitope, CD was performed and a characteristic helical signal was expected with a strong positive peak at 193 nm and two negative peaks at 208 and 222 nm [199]. As for the beta sheet peptides, all CD data were converted to mean residue molar ellipticity as explained in the methods (See Section 6.1.9).

6.3.1.1 CD Conditions and Optimisation of Extended α-helical Epitope

The CD conditions were optimised in a similar way to that described for the beta sheet peptide. The helical peptide was found to be soluble in water. To confirm the folding and the effect of an aqueous medium on folding, the peptides were dissolved in either water, phosphate buffer (20 mM sodium phosphate, 150 mM NaF, pH7.0) or TFE. The possibility of peptide aggregation was checked by performing a concentration scan (5 μM, 10 μM, 25 μM, 50 μM and 100 μM). The CD signal with different peptide concentrations confirmed that there was no aggregation. At this point, the concentration of 50 μM was selected to perform further optimisation of the experiments. A temperature ramp experiment was performed by recording CD

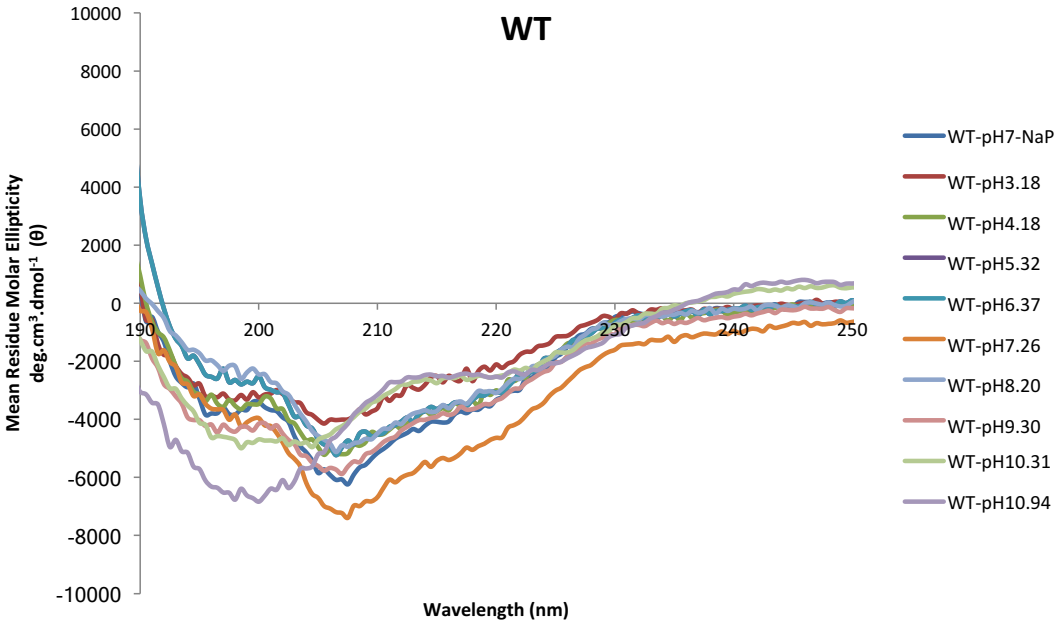


Figure 6.11: pH screening (3.18-10.94) on WT helical peptide in 0.2 M sodium acetate buffer at 20°C.

spectra from 4°C to 90°C. It was observed that the structure was lost at higher temperatures in comparison to the structure that was observed at lower temperatures. Therefore, a temperature of 20°C was chosen during all CD experiments.

The theoretical pI of the WT peptide was calculated as 5.48. So a pH screen in the range of $3.18-10.94$ (above and below the pI of the peptide) was performed to observe any change in the signal. The CD spectra of WT peptide at pH lower than the pI show slightly weaker signals as compared with physiological pH (shown in blue in Figure 6.11). The WT peptide loses any regular structure in a basic environment. On seeing the pH screening results, it was decided to use phosphate buffer (20 mM sodium phosphate, 150 mM NaF, pH7.0) for recording CD spectra.

At physiological pH, a peptide with 50 µM concentration showed a weak negative peak at 222 nm and a strong negative peak at 208 nm but no positive peak (Figure 6.11). Since CD spectroscopy reflects an average of the entire molecular population, it is not possible to determine which specific residues retain the helical

conformation. However, a strong negative peak at 208 nm signals a small fraction of helical population. In summary, this type of CD signal does not characterize a perfect helix or a complete random coil (See Figure 2.2 for the expected signal of a perfect helix or random coil). After CD condition optimisation using a manual spectrometer, the peptide samples were re-run on the Chirascan autosampling system. The cuvette in this system had a pathlength of 0.5 mm, and therefore 100 μM peptide concentration was used. CD spectra of other peptides in the helical epitope set were also studied using the same conditions. The increase in concentration on using the smaller pathlength cuvette is in accordance with the Lambert-Beer's law (Section 2.2.1.2).

6.3.1.2 CD Spectra of the Extended α-Helical Epitope

Figure 6.12A shows the average of 9 spectra (3 scans from 3 different peptide dilutions) for the WT epitope, its mutant and derivative peptides. The WT spectra do not show characteristic random coil because of the complete absence of a strong negative peak at 195 nm. However, the presence of a negative peak at 208 nm (which is one of the peaks indicative of a helix) suggests some sort of helical structure. The WT epitope with terminal caps (WT-C2) shows the strongest CD signal having both negative peaks for a helix. The M154A mutation to the WT shows slightly more structure. However, capping this mutant peptide (M154A-C1 and M154A-C2) did not show a significant rise in signal. The helix fraction at 208 and 222 nm was calculated (Section 6.1.9) for all the peptides (Figure 6.12B).

The alcohol based solvent, TFE (trifluoroethanol) is well known for stabilising the helical content of peptides compared with aqueous solution. The helix stabilisation is the result of a strengthening of the hydrogen bonding interactions in a lower dielectric environment [200]. TFE is also believed to simulate physiological

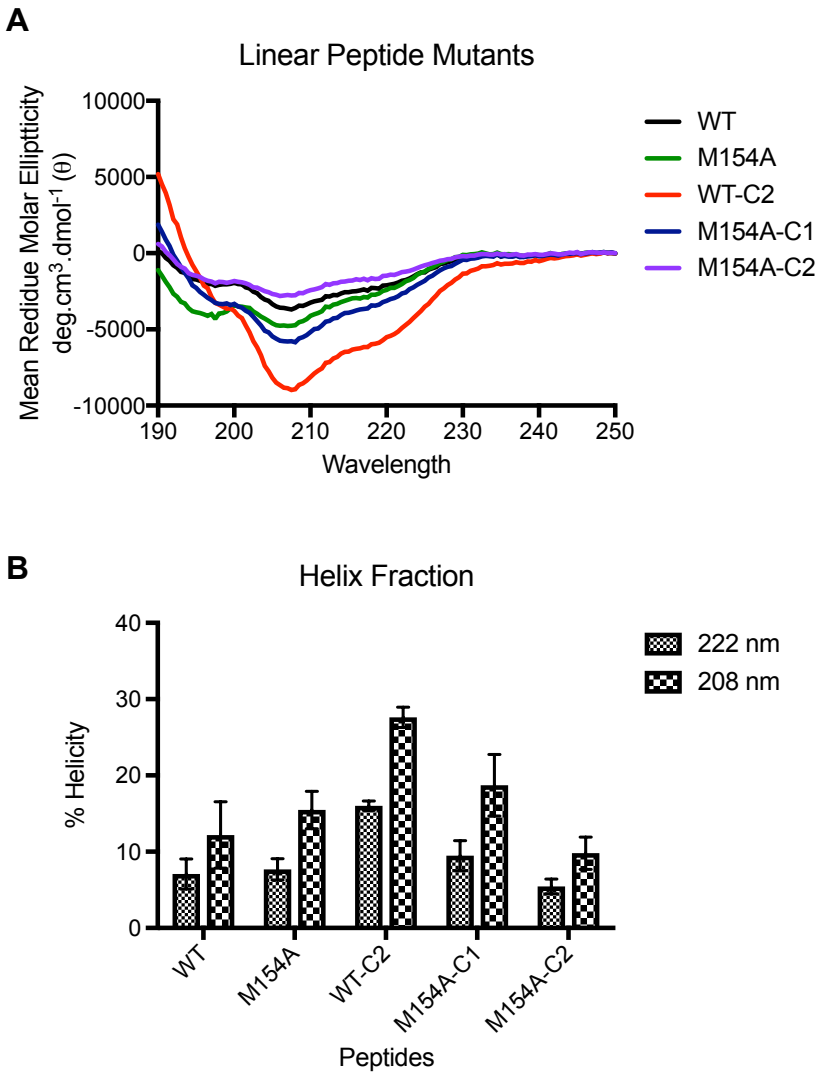


Figure 6.12: A) CD of the extended helical epitope peptide, its mutants and derivative peptides (average of 9 spectra) in 20 mM sodium phosphate, 150 mM NaF, pH7.0 buffer at 20°C. The cap on the WT peptide (WT-C2, coloured red) seems more structured compared with WT epitope (coloured black). The caps of the mutant (M154A, coloured green) did not show such behaviour. The capped mutant peptides (M154A-C1 and M154A-C) are shown in blue and purple. B) Helix Fraction calculated at 208 and 222 nm for all the peptides. Error bars show the 95% confidence interval.

conditions [200, 201] and has been recommended for studying the helicity of small peptides [201, 202].

The helix stabilising/inducing property of TFE is directly proportional to the propensity of amino acids that are capable of making helix therefore it reflects the underlying structural properties of amino acids within a particular protein or peptide [203–205]. Knowing that the WT epitope has helical structure in the full length protein, it was worth experimenting to study the behaviour of peptides in the presence of TFE. TFE titration was performed on the WT and M154A mutant peptide to study the effect of the mutation in the presence of TFE (Figure 6.13 and 6.14). In several previous studies, use of 30% TFE has been reported to stabilise helices [201, 204], and a similar behaviour was observed in this epitope and its mutant peptide. A helical signal appeared at 10% TFE concentration for both WT and M154A peptides which increased with increasing TFE concentration. A similar signal for both the peptides was seen up to 40% TFE, however a considerable difference between WT and M154A peptide signal strength was observed with 50-100% TFE (except with 60% TFE where both gave the same signal). This suggests the mutation might have a higher propensity to form a native-like helical structure at higher concentrations of TFE (Figure 6.15).

It is evident that WT, WT-C2, M154A, M154A-C1 and M154A-C2 in phosphate buffer show a lower population of helix (Figure 6.12). The reason for this weak and intermediate signal between a coil and helix could be attributed to the small length of the peptide. In the case of smaller length peptides, it is most likely that the spectrum will be dominated by end-effects [206]. The length of the WT peptide is 15 residues, if 4 residues on each end are mobile then it is only 6 amino acids that may be involved in forming a helix, but it is difficult to monitor only 2

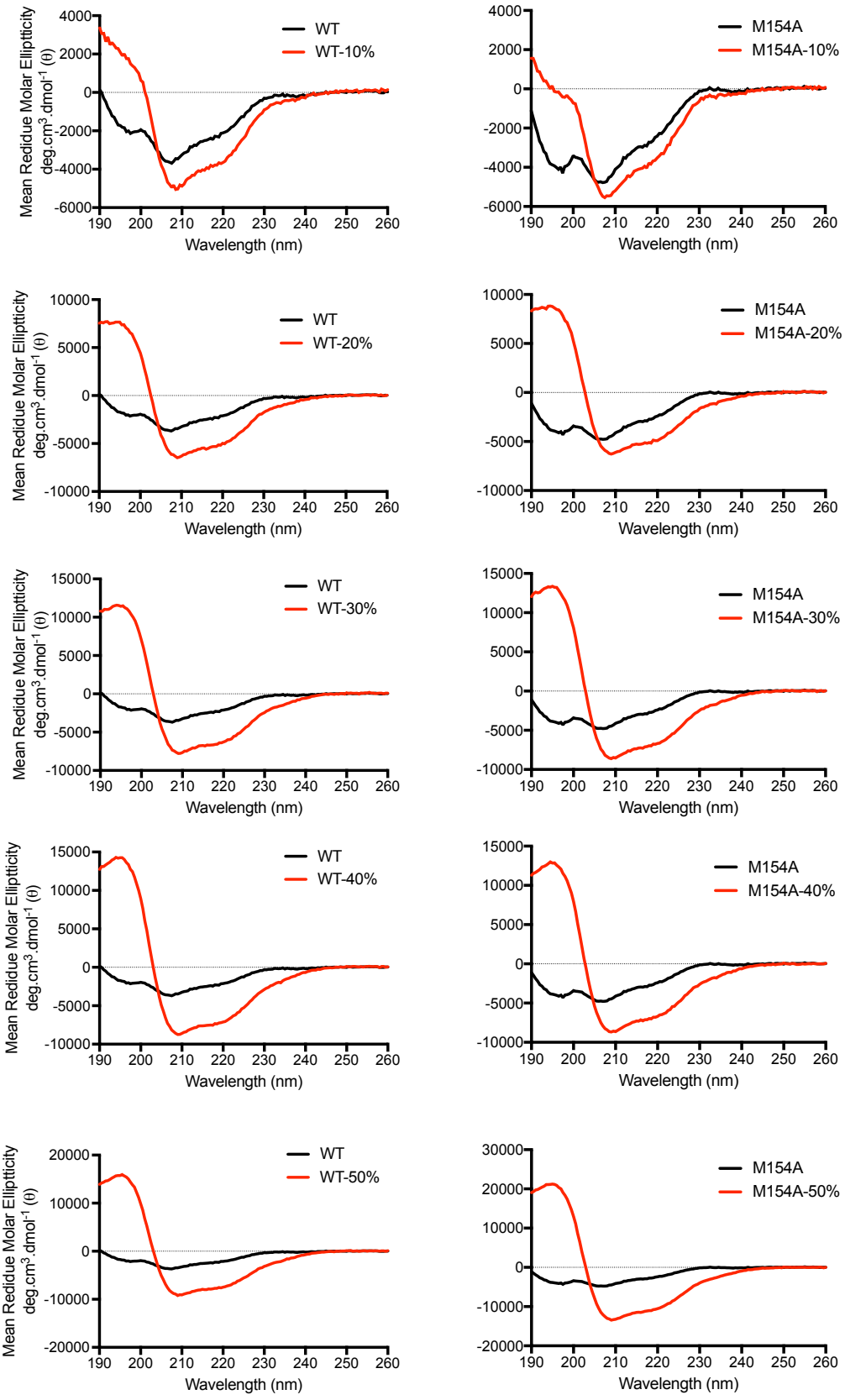


Figure 6.13: TFE titration comparison for WT and M154A at 10-50% TFE concentration.

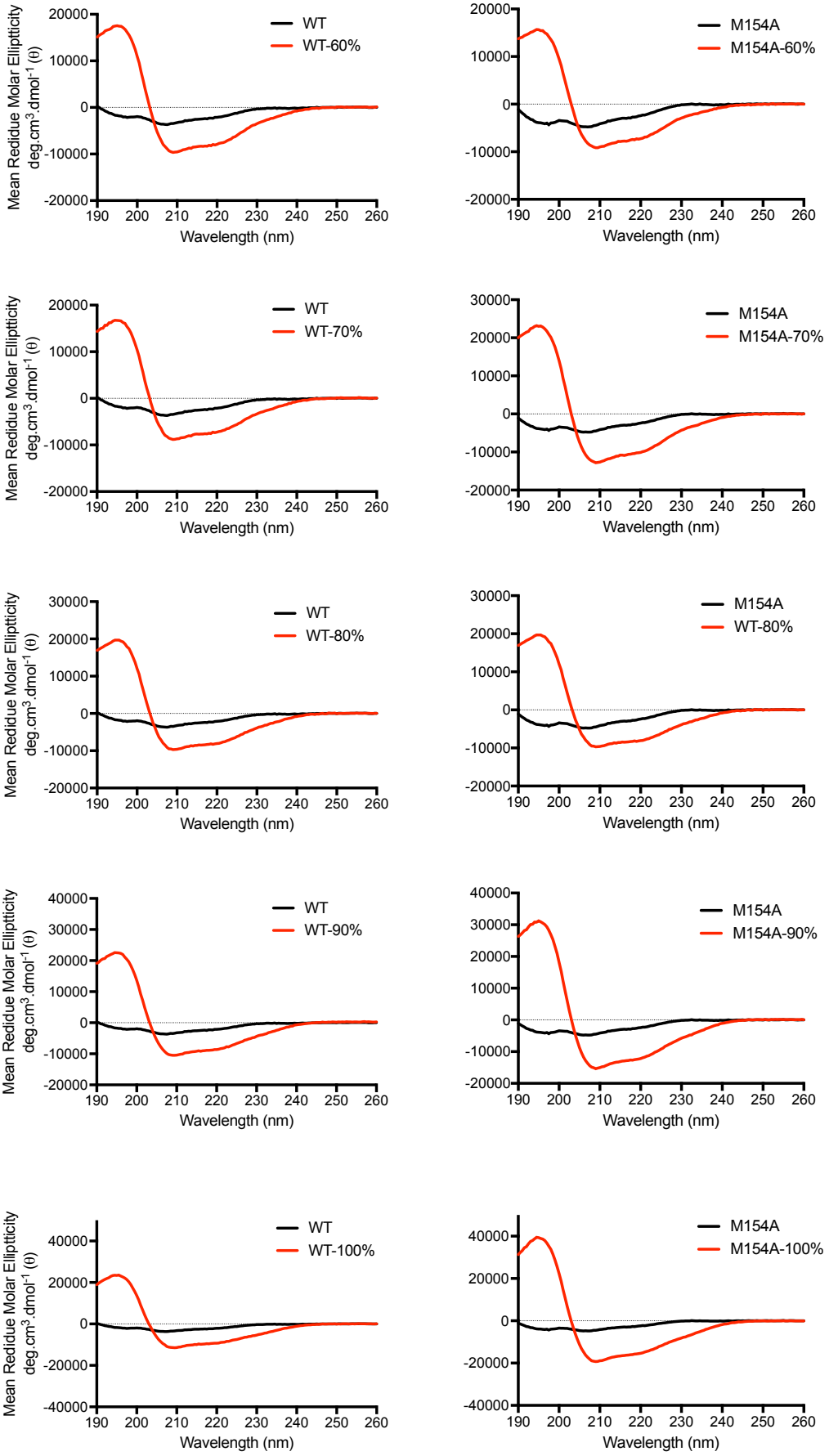


Figure 6.14: TFE titration comparison for WT and M154A at 60-100% TFE concentration.

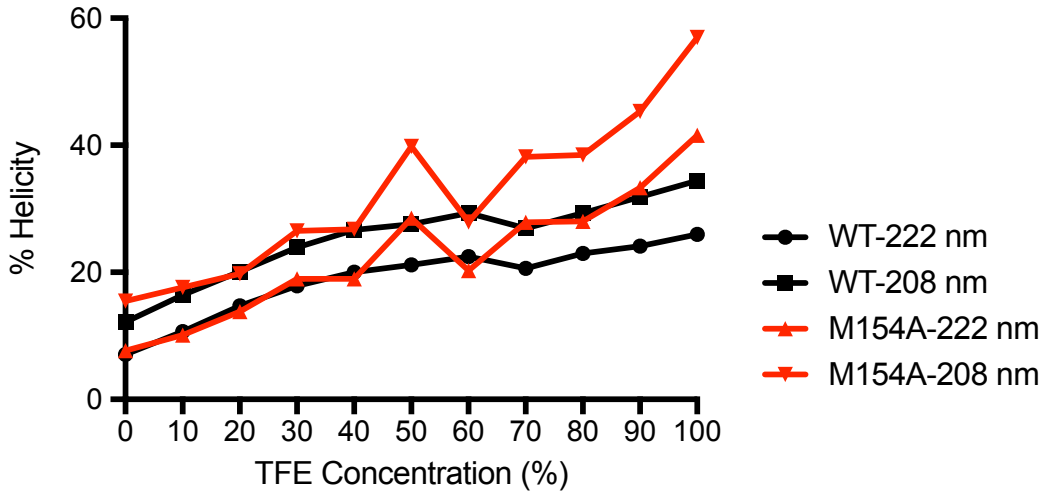


Figure 6.15: Helix Fraction in the WT and M154A at different TFE concentrations. The peaks at 208 and 222 nm are used to estimate the secondary structure of a helix.

turns of a helix using CD. However, the stabilising ability of TFE can provide insight into the folding ability. It has been seen that even 10% TFE was enough to induce an alpha helix, which indicates that the peptide has the potential to retain its native conformation. The effect of mutation could not be seen at lower concentrations of TFE. These results provides motivation to perform NMR studies to collect residue level structural information.

6.3.2 Nuclear Magnetic Resonance (NMR)

This study involved peptide secondary structure determination making use of NMR 2D experiments that included ^{1}H-^{1}H TOCSY, NOESY, ^{1}H-^{13}C HSQC and ^{1}H-^{15}N HSQC. The HSQC experiments were performed at natural abundance because it was not possible to isotopically label the synthetic peptides.

6.3.2.1 NMR Data Acquisition and Processing for Sequence-Specific Assignments

In order to study protein structure by NMR, it is crucial to be able to assign resonances to the protein sequence. For small peptides, this can be achieved using ^{1}H

homonuclear experiments such as total correlation spectroscopy (TOCSY) to assign by residue type and nuclear Overhauser effect spectroscopy (NOESY) in order to assist in sequential assignment. In this study, these were the initial experiments that were performed to gain insight into the peptide's structure. In addition to the ^{1}H homonuclear experiments, ^{1}H-^{13}C HSQC and ^{1}H-^{15}N HSQC (heteronuclear single quantum coherence) spectroscopy experiments were undertaken at natural abundance.

Briefly, TOCSY experiments were used to identify the amino acids and NOESY to assign structure while ^{1}H-^{13}C and ^{1}H-^{15}N HSQC experiments were used for sequence specific backbone assignments and monitoring of the helical conformation.

Table 6.7: Chemical shifts for the WT peptide. NMR experiments and assignments were performed with the enormous help of Christine Prosser and Leo Bowsher at UCB. Assignments 1–15 correspond to positions 142–156 in the protein sequence.

	Cα	**Cβ**	**HN**	**Hα**	**N**
G1	40.7	-	8.602	3.814	-
S2	55.46	61.28	8.604	4.427	115.8
D3	51.65	37.52	8.47	4.496	122.1
Y4	56.24	35.72	8.101	4.315	120.2
E5	54.13	26.61	8.052	4.082	121
D6	51.89	37.62	8.136	4.453	120.4
R7	54.8	27.52	8.005	4.001	120.5
Y8	55.9	35.59	7.964	4.336	119.4
Y9	56.16	35.8	7.857	4.312	120.7
R10	54.13	28	7.922	4.075	121.3
E11	54.66	26.62	8.103	4.088	119.8
N12	50.77	35.99	8.134	4.531	118.2
M13	52.92	29.95	7.964	4.295	119.7
H14	52.59	26.18	8.26	4.58	119.4
R15	54.67	28.52	8.045	4.089	127.6

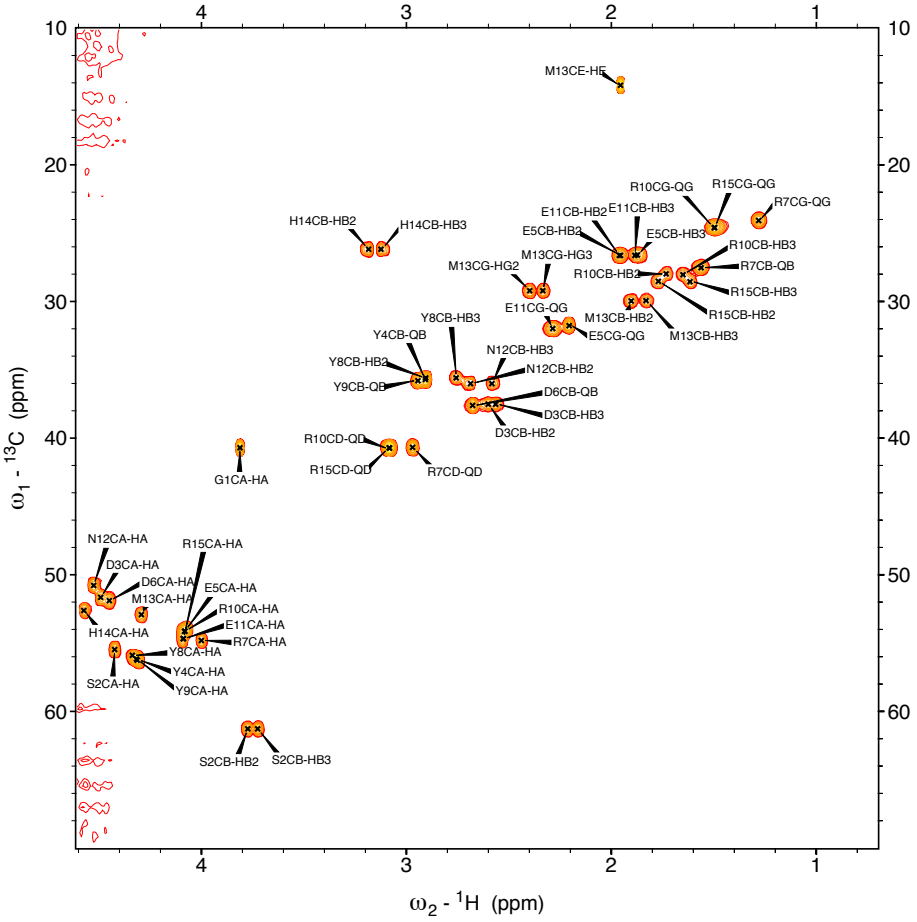


Figure 6.16: ^{1}H-^{13}C HSQC spectrum of the WT helical peptide. The spectrum was recorded at 37°C with a peptide concentration of 2 mM peptides in NMR buffer concentration of 20 mM sodium phosphate, pH 6.4 and 10% D_2O. Assignments 1–15 correspond to positions 142–156 in the protein sequence.

6.3.2.2 Sequence-Specific Assignments for WT and M154A — Helical Peptide

The WT and M154A α-helical peptides gave rise to a well resolved spectrum as illustrated by the ^{1}H-^{13}C HSQC shown in Figures 6.16 and 6.17. Both of the spectra show that the quality of the acquired data is good enough that it allowed for almost complete backbone assignments. Initially, ^{1}H-^{1}H TOCSY (Figure 6.18) was used to perform HN, Hα, Hβ and additional proton assignments. In NMR, it is common to use data from multiple experiments to perform the assignments. This

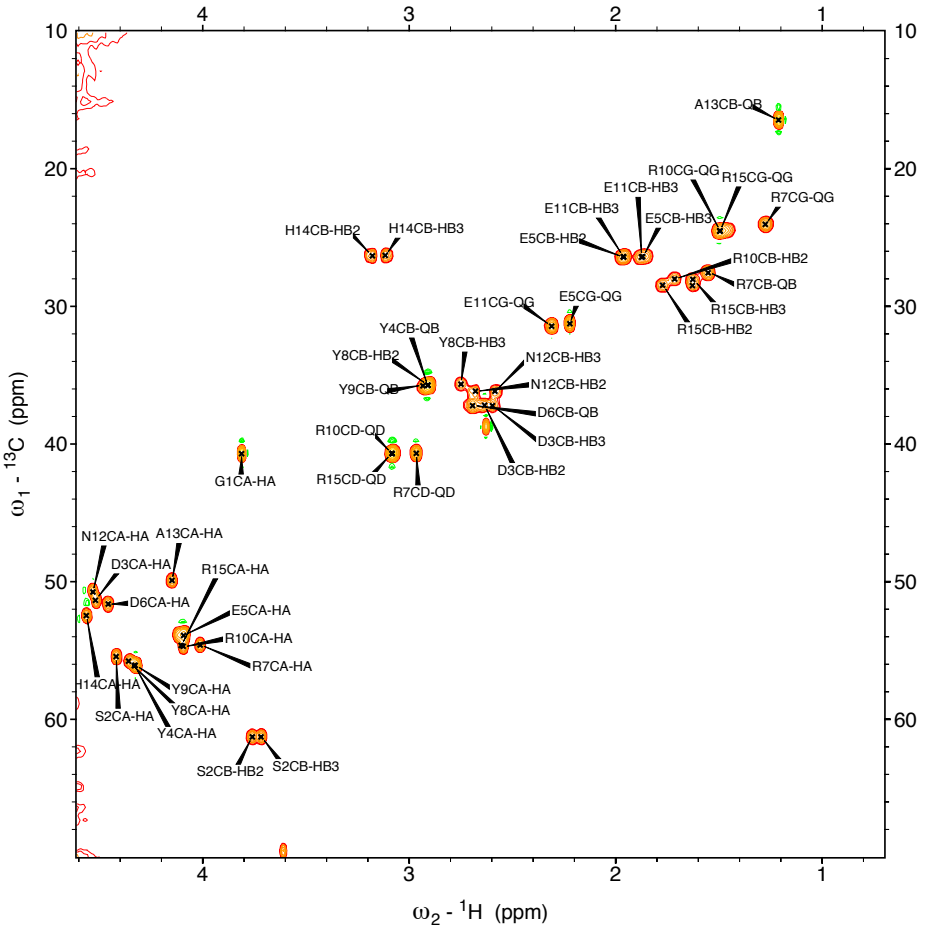


Figure 6.17: ^{1}H-^{13}C HSQC spectrum of the M154A helical peptide. The spectrum was recorded at 37°C with a peptide concentration of 2 mM peptides in NMR buffer concentration of 20 mM sodium phosphate, pH 6.4 and 10% D_2O. Assignments 1–15 correspond to positions 142–156 in the protein sequence.

is because the peaks for a particular amino acid could be missing in one spectrum but present in another. One such example in both of these peptides is glycine. The peak for glycine could not be assigned from the TOCSY, however a combination of C-HSQC and NOESY provided chemical shifts of Cα, Hα and HN. The main priority during assignments was to acquire chemical shifts for backbone atoms (HN, Hα, Cα), but chemical shifts for side chain atoms were also assigned where the peaks were available. All the available assignments for backbone and sidechain atoms are shown in Appendix B, Tables B.1 and B.2. Tables 6.7 and 6.8 show

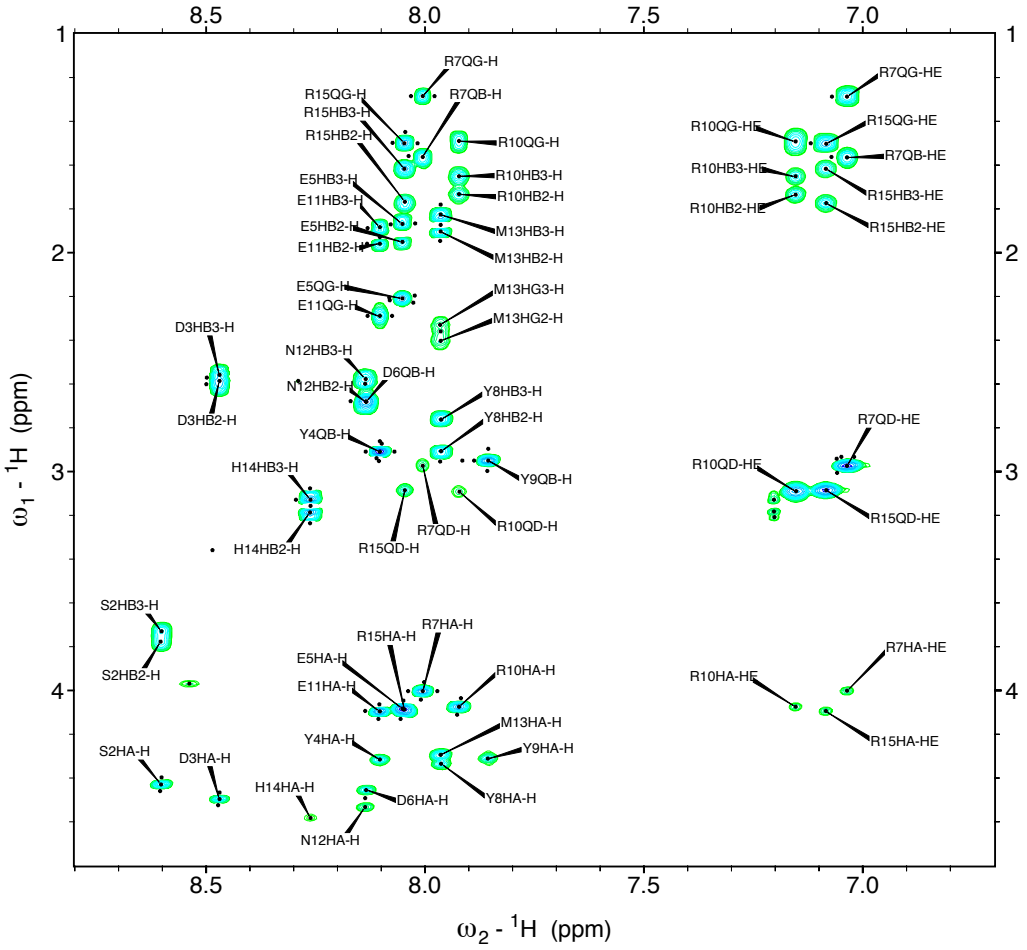


Figure 6.18: ^{1}H-^{1}H TOCSY, spectrum of the WT alpha-helical peptide. The spectrum was recorded at 37°C with a peptide concentration of 2 mM peptides in NMR buffer concentration of 20 mM sodium phosphate, pH 6.4 and 10% D_2O. Assignments 1–15 correspond to positions 142–156 in the protein sequence.

100% of the sequential backbone assignments performed on the WT and M154A peptides. The tables contain chemical shifts of some additional atoms as well as backbone atoms. The chemical shifts for C and N atoms were obtained using ^{1}H-^{13}C and ^{1}H-^{15}N HSQC spectra, respectively, whereas proton shifts were assigned using TOCSY. The linear sequence of these assigned residues was obtained using information from the NOESY spectrum.

For the WT peptide of 15 residues, 15 backbone and 3 side chains (arginine and asparagine) peaks were expected on the ^{1}H-^{15}N HSQC spectrum, but a total of

Table 6.8: Chemical shifts for the M154A peptide. NMR experiments and assignments were performed with the enormous help of Christine Prosser and Leo Bowsher at UCB. Assignments 1–15 correspond to positions 142–156 in the protein sequence.

	Cα	**Cβ**	**HN**	**Hα**	**N**
G1	40.69	-	8.594	3.811	-
S2	55.45	61.26	8.585	4.423	115.7
D3	51.34	37.19	8.469	4.524	121.9
Y4	56.09	35.75	8.112	4.329	120.5
E5	53.89	26.4	8.051	4.095	120.8
D6	51.63	37.22	8.121	4.462	120
R7	54.59	27.56	7.989	4.018	120.5
Y8	55.77	35.67	7.947	4.359	119.4
Y9	56.06	35.76	7.841	4.329	120.9
R10	54.69	28	7.934	4.096	121.9
E11	53.89	26.39	8.114	4.128	120.5
N12	50.74	36.15	8.188	4.537	119
A13	49.91	16.44	7.977	4.154	123.6
H14	52.46	26.31	8.237	4.569	117.5
R15	54.63	28.33	8.056	4.102	127.4

14 backbone and 3 side chains peaks were observed and assigned in correspondence to the HN shifts. The peak for glycine was missing because of being the N-terminal residue. Commonly, the peak for N-terminal residue is not readily seen owing to exchange with the solvent [207].

6.3.2.3 Difference in WT and M154A Peaks

In order to examine the difference in peaks of WT and M154A peptides, the ^{1}H-^{13}C spectrum of the WT was overlaid on the mutant peptide as shown in Figure 6.19A. The Cα – Hα chemical shifts for methionine and alanine are present between 50 and 55 ppm. Likewise, the distinctive peaks of Cβ, Cγ and Cε are also present between 10 and 50 ppm. The rest of the 14 residues have been perfectly overlaid which shows the stability of the chemical shifts for the WT and mutant peptide, M154A. Initially, it was thought that a key peak Cα – Hα of E11 was missing in the mutant peptide while it was present in the WT, but later it was interpreted as

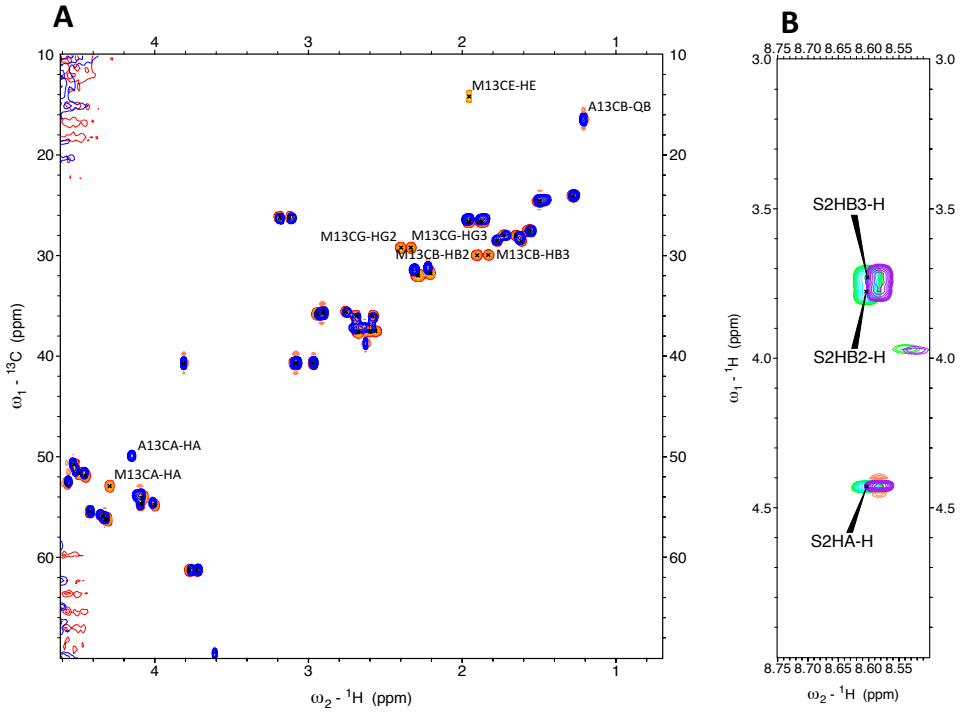


Figure 6.19: A) An ^{1}H-^{13}C HSQC overlay of M154A mutant peptide (blue) on WT peptide (orange). Assignments 1–15 correspond to positions 142–156 in the protein sequence. Thus, position 13 refers to 154 in the protein sequence which is the mutation site. The spectrum was recorded at 37°C with a peptide concentration of 2 mM protein in NMR buffer concentration of 20 mM sodium phosphate, pH 6.4 and 10% D_2O. B) A strip taken from TOCSY overlay of WT (green) and mutant (purple) showing proton shifts of Serine at position 2.

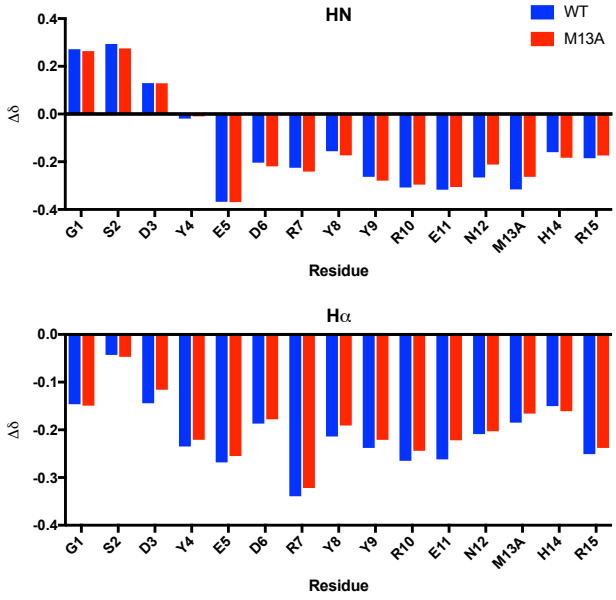


Figure 6.20: Chemical Shift Index using chemical shifts of HN and Hα atoms. Assignments 1–15 correspond to positions 142–156 in the protein sequence. Thus, position 13 refers to 154 in the protein sequence which is the mutation site.

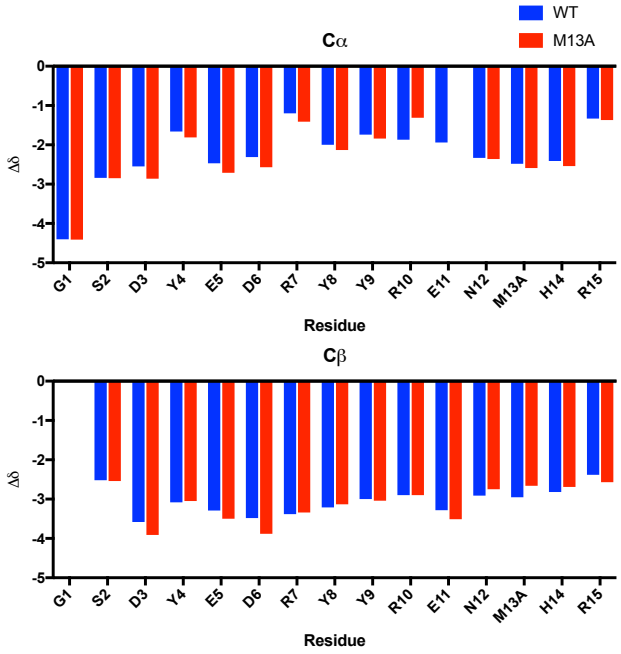


Figure 6.21: Chemical Shift Index using chemical shifts of Cα and Cβ atoms. Assignments 1–15 correspond to positions 142–156 in the protein sequence. Thus, position 13 refers to 154 in the protein sequence which is the mutation site.

an overlap with the peak of E5. A general overlap was observed among Cα – Hα peaks of E5, R10, E11 and R15 (Figures 6.16 and 6.17, and Tables 6.7 and 6.8). The overlay of TOCSY also did not show any inconsistency or considerable deviation of peaks between the spin systems of WT and mutant peptides. An example of the spin system for serine at the second position, from the overlay of WT and M154A spectra, is also shown in Figure 6.19B which shows common peaks from both the spectra.

6.3.2.4 Secondary Structure of WT and M154A Mutant

The Chemical Shift Index (CSI) is a widely accepted procedure to estimate the secondary structure of proteins on the basis of the observed chemical shift differences as compared to the predefined random coil chemical shifts. In order to probe the secondary structure better using the CSI method, the chemical shift differences of several atoms (Hα, Hβ, Cα and Cβ) of all the residues in the peptide were manu-

ally compared with random coil chemical shifts (shown in Figures 6.20 and 6.21). The Hα protons are the most sensitive to conformational changes and shifts are mainly correlated with secondary structure. These have been widely used to study the conformational changes and determination of secondary structure elements in proteins and peptides [195,208,209]. Another method has been developed to predict secondary structure which makes use of Cα and Cβ and ^{13}C chemical shifts [210]. There are also a few studies which have used Cα and Cβ chemical shifts for structural analysis of peptides to probe β-sheets as it is believed that these shifts have a significant correlation with dihedral angles [211]. Neither of these alternative approaches was used here.

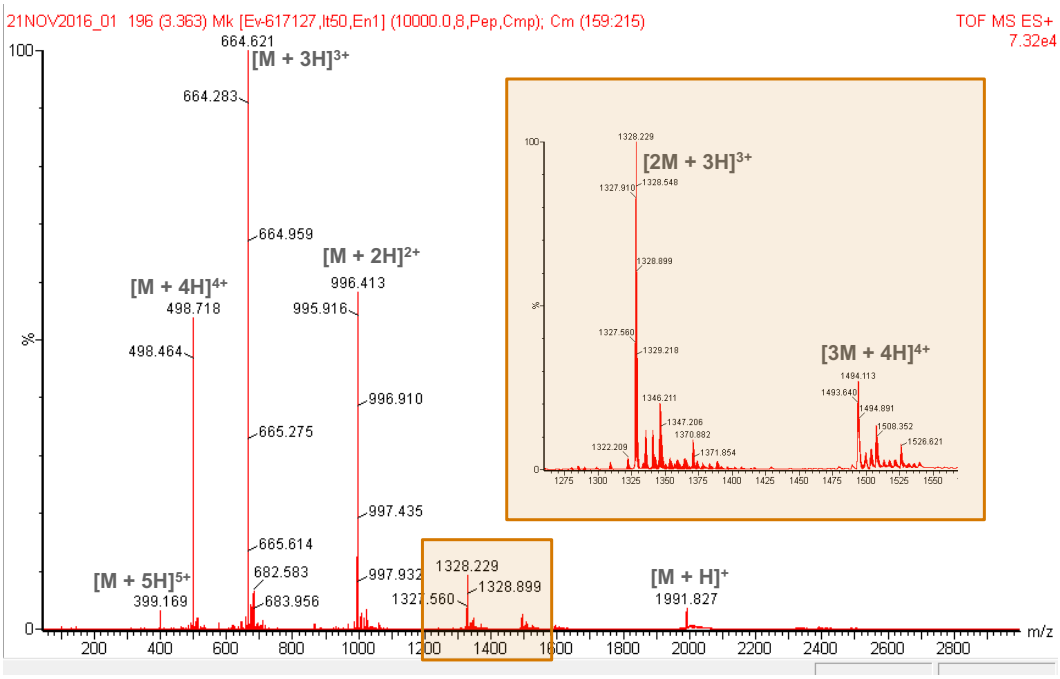


Figure 6.22: Mass spectrometry of WT peptide.

6.3.3 Mass Spectrometry

Before performing NMR on the WT and M154A peptides, the molecular mass of the WT peptide was confirmed by mass spectrometry, and showed agreement with the theoretical mass (shown in Figure 6.22). This was done to make sure that the correct peptide sequence is present.

6.3.4 ELISA

ELISA was carried out on this set of peptides to study their binding with Fab. Figure 6.23 shows the concentration-dependent binding of all peptides with Fab ICSM18. ELISA plates were coated with different amounts of peptide followed by the addition of 10 µg/ml of Fab (details in the methods section). The plate was read at 0, 30 and 60 minute intervals. However, data are only shown for 490 nm after 30 minutes of quenching. The data show that all the peptides bind effectively with antibody. In order to find statistical significance

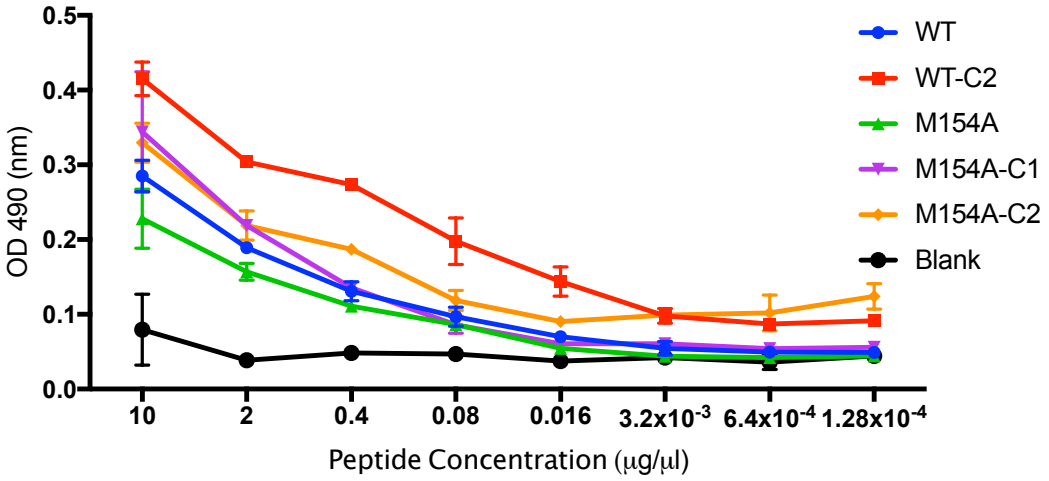


Figure 6.23: Peptide concentration dependent ELISA analysis. All the peptides at 8 different concentrations (5 fold dilution starting from the top concentration of 10 µg/µl) were exposed to 490 nm of UV for 30 minutes. Error bars show standard error of mean. A p-value > 0.05 was calculated for different pairs using Welch's t-test.

among different groups, Welch's t-test was performed on WT/WT-C2, WT/M154A, M154A/M154A-C1 and M154A/M154A-C2 and a p-value > 0.05 was observed suggesting that the difference in binding is not significant. In addition, %coefficient variation (CV) was also computed as a ratio of the standard deviation from the mean, indicating any inconsistencies and inaccuracies in the results. Coefficient variation of less than 20% is considered acceptable. The replicates for these peptides at different concentrations show very little variance, with the CV in the range of 2 to 15% except for a few outliers at higher concentration.

This preliminary binding assay confirmed the binding of Fab with isolated peptides, but further ELISA experiments were not performed at this stage and it was decided to carry out a more sensitive binding assay using surface plasmon resonance (SPR) to provide a means to measure the binding kinetics.

6.3.5 Surface Plasmon Resonance (SPR) of α-Helical Epitope

The binding of the helical peptide with Fab ICSM18, was studied using SPR by immobilising the Fab to the sensor chip surface. The peptides were diluted in the

running buffer and titrated over the ligand surface. The full length prion protein binds with Fab via an epitope at positions 142-156 (Figure 6.2). The isolated epitope is expected to bind to Fab if it adopts its native helical conformation. Similarly, the mutant and other derivative peptides should also bind to Fab if they fold into the correct helix conformation. The interaction of the ligand (immobilised) and analyte (mobile) is monitored by the SPR instrument as a change in the resonance angle over time. From this, association (K_a), dissociation (K_d) and equilibrium rate (K_D) constants can be derived. To this end, SPR was performed on this set of peptides.

6.3.5.1 Optimisation of the SPR Method

A CM5 sensor chip was preconditioned and activated as described in the methods. In order to reach higher ligand densities, it was necessary to increase the ligand concentration and lower the pH of the buffer. To this end, the Fab was immobilised at three different levels (different concentration and pH) on three flow cells of the chip. At first, 25 µg/ml of Fab (diluted in 10 mM acetate buffer, pH 5.0) was injected into flow cell 2 with the aim of obtaining 500 RU. Three injections, each of 120 seconds, provided 570 RU of Fab in flow cell 2. This was followed by injecting ethanolamine to cap the free carboxy methyl ends. Similarly, 100 µg/ml of Fab (diluted in 10 mM acetate buffer, pH 5.0) was immobilised on flow cell 3 with the aim of providing ≈2500 RU. However, only 1890 RUs were obtained for flow cell 3. The same concentration of Fab was placed on flow cell 4, but it was diluted in 10 mM acetate buffer, pH 4.0. An immobilisation of ≈5000 RU was expected but only 2650 RU were obtained on the chip. Flow cell 1 was activated and capped, but left blank to use as a reference.

In order to generate the full binding kinetics for the interaction of Fab with peptides and to acquire the binding constants, the interaction needs to be measured

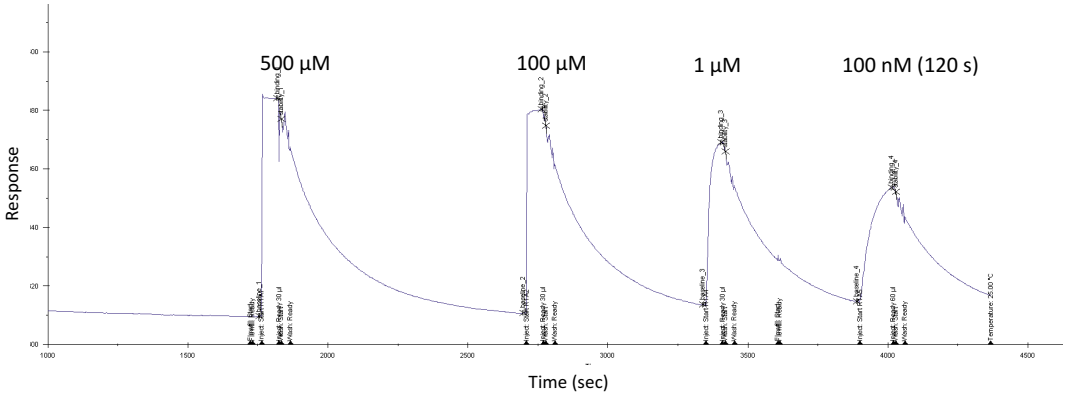


Figure 6.24: SPR profiles for the WT peptide diluted, in HBS-EP+ running buffer, to 500 µM, 100 µM and 1 µM and injected for 60 sec at 30 µl/min. The peptide with 100 nM concentration, injected for 120 sec, shows an acceptable sensorgram showing association and dissociation.

at multiple analyte concentrations. Furthermore, an optimal maximum concentration, from which to perform the dilution series, needs to be determined. Since, in this experiment, the peptide affinity was unknown and the best chance of detecting binding needed to be explored, a very high peptide concentration was chosen to start with and the WT peptide was diluted to 500 µM in HBS-EP+ running buffer. This was injected for 60 seconds at a flow rate of 30 µl/min and binding was monitored. The peptide was further diluted to 100 µM, 1 µM and 100 nM and injected into the chip to detect its binding with Fab. Figure 6.24 shows the sensograms at different peptide concentrations. Having seen the binding data, the 100 nM peptide concentration was selected as the optimal maximum concentration to perform two-fold dilution series because of the high quality shape of the association and dissociation curves (Figure 6.24). It is important for the shape of the curve to have sufficient curvature during the association to allow data fitting. The response must start to plateau in order to give an accurate estimate of the association rate. To ensure that this has been achieved, the injection time was increased. In addition, it is evident from the data that regeneration is not needed as the response returned to the baseline after 40 minutes.

In order to obtain a high quality data set, the kinetic cycle was set up with 6 concentrations (two-fold dilution series starting from 100 nM down to 3.125 nM) and replicates at 50 nM and 0 nM. A time of 40 minutes between each injection was given to dissociate the complex completely. At this stage, only WT peptide was used to study the kinetics.

The sensorgrams at different flow cells and with different peptide concentration, were recorded. The consistent kinetics at different ligand densities provides evidence that there is no mass transport limitation for this peptide (Figure 6.25). After obtaining these data for the WT peptide, flow cell 3 was selected to collect the binding kinetics data for all mutant peptides because a good level of signal for all of the peptides (at different concentrations) was obtained at the lowest immobilisation level. A good response (or signal) at the lowest immobilisation level reduces the chances of mass transport and non-specific binding which can occur at higher concentrations.

6.3.5.2 Binding Kinetics of Peptides

After the optimisation of the Biacore method and obtaining one set of kinetics data for all the peptides, it was decided to repeat the experiment by preparing a new chip and multiple peptide dilutions in order to allow statistical analysis to be applied to the results. 100 µg/ml Fab (diluted in 10 mM acetate buffer, pH 5.0) was immobilised in flow cells 2, 3 and 4. An immobilisation level of 1890 RU was achieved. Two different dilution series for each of the peptides were prepared at 100 nM, 50 nM, 25 nM, 12.5 nM, 6.25 nM and 3.125 nM and injected into 3 different flow cells. This provided 6 replicates for each of the peptides. In addition, one curve was obtained during optimisation as described in the previous section. Hence, there were 7 replicates of each peptide except for WT where there were 9 replicates.

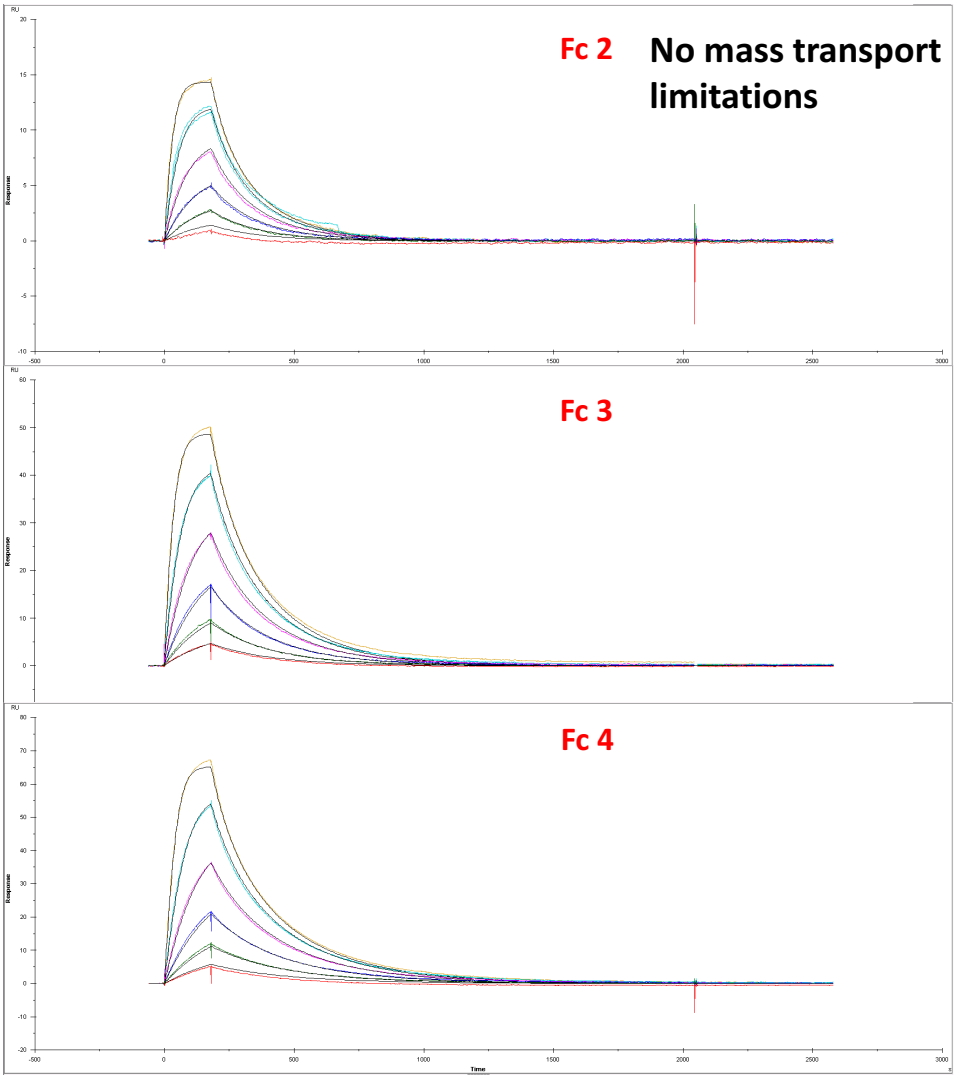


Figure 6.25: WT peptide at 100 nM, 50 nM, 25 nM, 12.5 nM, 6.25 nM and 3.125 nM concentration on different flow cells.

In order to obtain kinetics for the binding between Fab and peptides, a 1:1 binding model was used to fit the experimental data, chosen on the basis of the known one to one interaction between Fab and epitope. During data fitting, 100 nM curves were removed because the fit was considerably improved without them. Figure 6.26 shows the representative sensorgrams and associated fits for each of the peptides. Global fitting of data was performed that included all the concentrations from 50 down to 3.125 nM. The association rate (K_a), dissociation rate (K_d) and equilibrium (K_D) constants, derived from this global fitting are shown in Table 6.9.

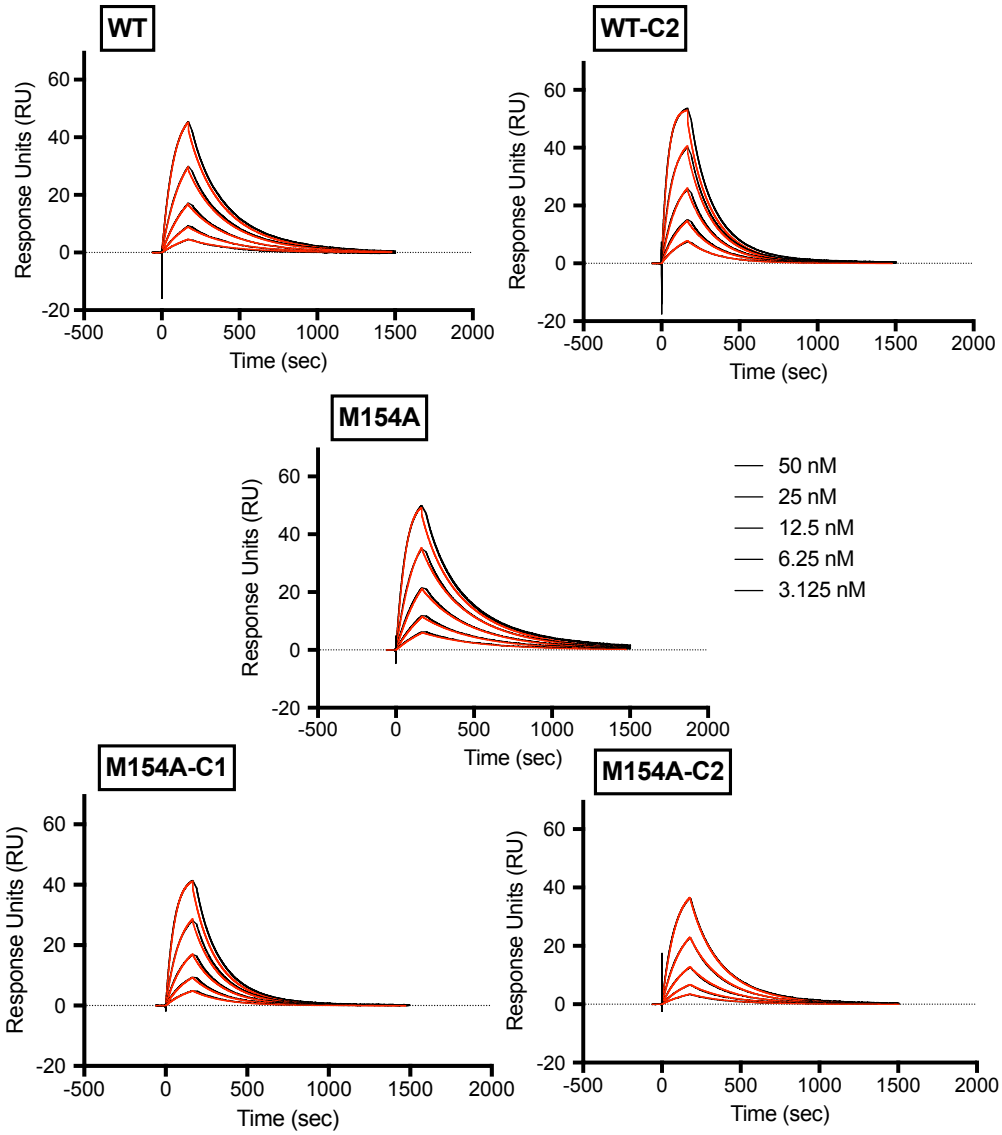


Figure 6.26: SPR curve fitting (global) for different peptides at concentration of 50 nM, 25 nM, 12.5 nM, 6.25 nM and 3.125 nM. Blanks were substracted from each of the concentration curves.

R_{max} is a measure of the activity of surface-attached ligand, and represents the maximum binding capacity of the surface in response units (RU). The theoretical R_{max} can be calculated from the ratio of the molecular weights of analyte and ligand multiplied by the immobilisation level. The Biacore instrument computes the experimental R_{max}, and the ratio of the experimental to theoretical R_{max} can be used to give an idea of the proportion of the protein that is active. The ratio for all of the peptides is approximately the same which tells us that a similar proportion of each of the peptide molecules are capable of binding to the ligand. The M154A-C2 peptide appears to be an exception having a slightly lower ratio; however this decrease was not found to be significant (p-value > 0.05).

Table 6.9: Binding Kinetics: the average of multiple independent replicate experiments. There were 9 replicates for WT whereas these were 7 for the rest of the peptides.

Peptides	K_a (1/Ms)	K_d (1/s)	K_D (M)	$^*\mathbf{R_{max}}$ (RU)	$^{**}\mathbf{R_{max}}$ (RU)	$^*\mathbf{R}_{max}/^{**}\mathbf{R}_{max}$
WT	$4.07\text{x}10^5$	$7.56\text{x}10^{-3}$	$1.87\text{x}10^{-8}$	60.15	69.71	0.86
WT-C2	$5.25\text{x}10^5$	$8.34\text{x}10^{-3}$	$1.61\text{x}10^{-8}$	69.14	75.33	0.92
M154A	$5.41\text{x}10^5$	$6.75\text{x}10^{-3}$	$1.25\text{x}10^{-8}$	58.64	67.57	0.87
M154A-C1	$3.87\text{x}10^5$	$8.70\text{x}10^{-3}$	$2.26\text{x}10^{-8}$	60.07	70.13	0.86
M154A-C2	$2.67\text{x}10^5$	$8.20\text{x}10^{-3}$	$3.09\text{x}10^{-8}$	59.37	72.16	0.82

* Experimental R_{max}
** Theoretical R_{max}

In order to allow comparison of the strength of binding between different peptides and Fab, the association, dissociation and equilibrium constants were plotted and statistical tests were applied to estimate significant differences in binding behaviour. Figure 6.27A shows strong association of M154A followed by WT-C2 and WT whereas M154A-C1 and M154A-C2 have lower rates of association. A Welch's t-test was performed to check the significance between WT/WT-C2 and WT/M154A association. A p-value < 0.05 was found for both of these pairs. This shows that means of WT-C2 and M154A differ from WT significantly and appear to have stronger association compared with the WT suggesting that there is a larger

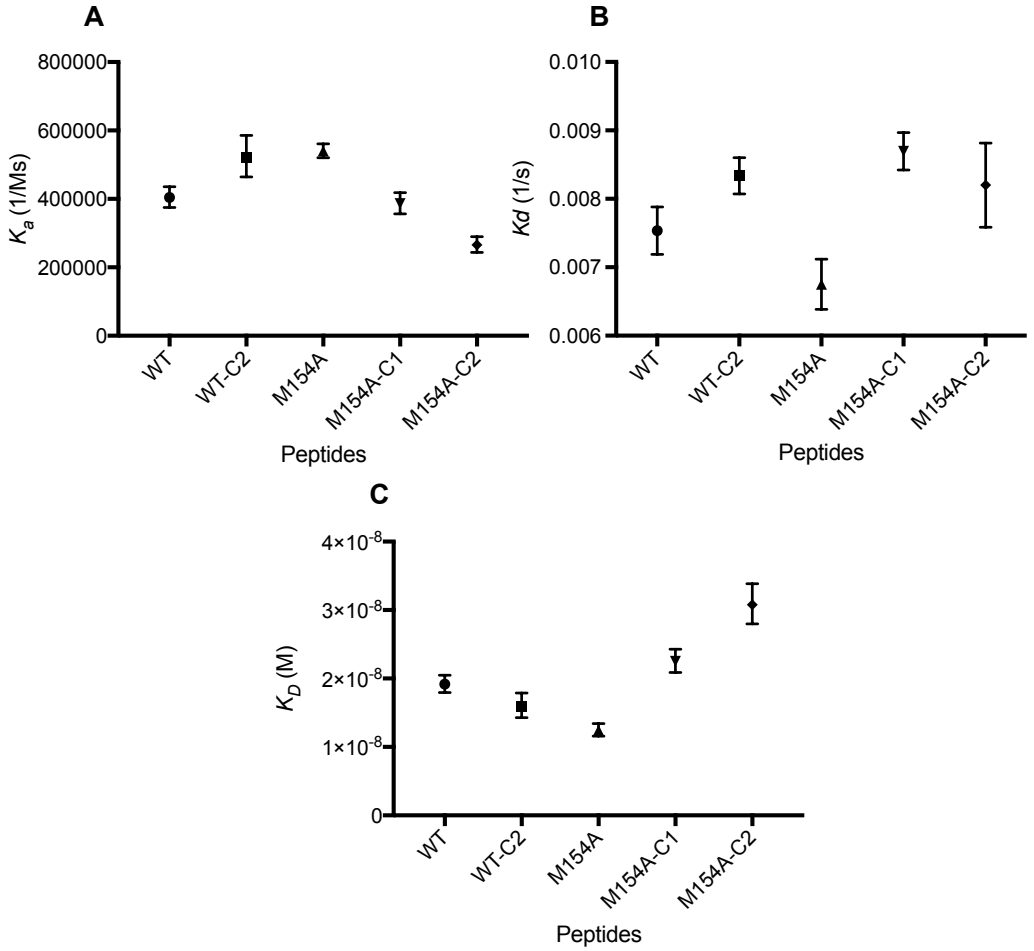


Figure 6.27: Binding kinetics: A) Association rate constants for each peptide. The error bars represent the geometric mean with 95% confidence interval. WT-C2 and M154A binds quickly as compared to WT (p-value < 0.05). B) Dissociation rate constants (mean with 95% CI). M154A dissociates slowly as compared to WT and WT-C2 (p-value < 0.05). C) Equilibrium rate constants (mean with 95% CI). Significant K_D of M154A (p-value = 0.0001) suggests strong binding with Fab.

population of peptide that is correctly folded in the native conformation. In contrast, dissociation rates were plotted and M154A appears to have the lowest off-rate followed by WT while WT-C2 and M154A-C1 and M154A-C2 dissociate at somewhat faster rates (Figure 6.27B). WT-C2 and M154A were found to be significantly different from WT (p-value < 0.05). The equilibrium rate constants were plotted (Figure 6.27C) clearly showing the distinct equilibrium constants for WT-C2 and M154A.

6.3.6 Summary of Extended Helical Peptide

The WT, WT-C2, M154A, M154-C1 and M154-C2 were studied for conformational stability using CD and NMR. In aqueous solution, CD on all of these peptides showed some structured population of helix owing to the presence of a strong negative peak at 208 nm. Cap2 on the WT peptide (WT-C2) was found to be the most structured among all of these peptides (Figure 6.12) and showed a significant difference compared with the WT (p-value = 0.001). The mutation, M154A, did not affect the structure compared with WT (p-value = 0.34), however the Cap1 on M154A significantly improved the conformation compared with either M154A or WT (p-value < 0.05). Interestingly, Cap2 on the M154A significantly destabilised the conformation (p-value = 0.006) which means that the same cap has a different effect on M154A compared with WT. TFE titration was performed on both WT and M154, and it was found that 10% of TFE was enough to stablise the native conformation (Figure 6.13). No change in CD signal of WT and M154 was observed up to 40% TFE concentration whereas a slightly higher helix fraction was observed for M154A at higher TFE concentrations. The presence of helical structure in WT and M154A was confirmed by NMR which showed the presence of α-helix structure in both of these peptides (Figure 6.20). These data suggest that the isolated peptides have the ability to retain their native conformation when taken out of the full length protein.

The binding of these isolated peptides with Fab was studied using ELISA and SPR. Initially binding was confirmed using ELISA, and the SPR data show a strong binding association for all of the peptides. The significantly faster association rate and slower dissociation rate of M154A compared with the WT (p-value < 0.05) suggest that the mutation has the ability to strengthen the interaction of the isolated

peptide (Figure 6.27). The capped versions of M154A also showed binding with the Fab but it was not as strong as WT and M154 (p-value < 0.0001).

6.4 Discussion

6.4.1 Effects of TFE on Peptide Secondary Structure

Isolated peptides (from full length protein) in aqueous medium lack long-distance protein interactions that ensure proper folding and therefore could lose their secondary structure on isolation. This has been observed in both sets of peptide (β-strand and helical) studied experimentally where the WT peptides did not show characteristic structure in aqueous solution. Although there was evidence of some structure in solution, it was not possible to characterise it confidently. The already-known structure of these peptides in the full length protein, the expected behaviour of the isolated peptides in solution and the ability of TFE to stabilise peptide structure suggested the use of TFE to probe the secondary structure of these isolated peptides. There was enough evidence from the literature to use 30% TFE while studying the structure of small peptides [201, 212, 213]. Therefore, in the initial set of experiments, the effect of adding 30% TFE was studied.

Although TFE is well known as a helix stabiliser, some studies suggest that it has the ability to stabilise β-sheets or β-hairpins as well [202, 214, 215]. The effect of 30% TFE on stabilisation of α-helices and β-structures was investigated by Roccatano et al. [216] who suggested that the stabilising effect of TFE is due to the accumulation of TFE molecules around the peptide excluding water molecules which in turn removes alternative hydrogen bonding partners and provides a low dielectric environment that promotes the formation of intrapeptide hydrogen bonds. TFE does not disrupt the hydrophobic interactions, but it forms weak interactions

with non-polar residues. Consequently, it contributes towards the stability of secondary structure of peptides rather than denaturing them.

The literature reports a wide range of studies where TFE has been used as a cosolvent in NMR and CD studies [217–219]. Studies have reported the similarity of structure in 30% TFE to those predicted from secondary structure software [220, 221] and to X-ray crystallographic and NMR structures solved without TFE [221–223]. Moreover, the helix stabilising/inducing property of TFE is directly proportional to the propensity of amino acids to form a helix, and therefore it reflects the underlying structural properties of a particular protein or peptide [203–205]. Nevertheless, any interpretation of findings involving TFE requires caution. A moderate level of TFE (up to 30-40%) favours structural stability, but higher levels can have an opposite effect where it may change the structure of a β-sheet to an α-helix by disrupting long range hydrophobic interactions [224–226].

The data in Figure 6.4 show that the isolated β-sheet peptide (WT and V637A) could not maintain its native conformation in aqueous solution, whereas the cyclised peptide (WTG) retained the native conformation in buffer although the peaks were slightly shifted. The addition of 10% TFE did not stabilise or induce any sort of structure in any of these peptides. However, it seems that 40-50% TFE stabilised the structure in the cyclised peptide by shifting the CD peaks to the characteristic maximum and minimum. This was due to the ability of this peptide to sample near-native conformation in buffer; the TFE only stabilised this existing structure. This behaviour was not seen in the WT peptide which was unstructured in buffer. At the same concentration of TFE (30-40%), it started to shift from unstructured to helix-like. A very high concentration of TFE (100%) affected cyclised and WT peptides equally. This behaviour of TFE is not surprising and agrees with the literature

[224–226].

In the case of the helical peptide set, the CD spectrum for WT showed a strong negative peak at 208 nm, a very weak negative peak at 222 nm and a complete absence of a positive peak at 193 nm. The addition of 10% TFE was found to be enough to stabilise the peptide to the extent that the spectrum showed the missing characteristic peaks indicating the presence of α-helix (Figure 6.13). The TFE titration on WT and M154A was performed to study the effect of mutation on peptide folding. For both of these peptides, a similar signal was observed up to 40% TFE concentration. A considerable difference in the signal strength, between WT and M154A from 40-100% TFE concentration was seen (except at 60%). At 100% TFE, for the M154A peptide, twice the signal was observed as compared with the signal for WT. However, this experiment was not enough to characterise the effect of the mutation, so NMR was planned for these two peptides.

In conclusion, these experiments suggest that the peptides show some native-like structure which is stabilised by TFE. In particular, the α-helical peptide, in the absence of TFE, had only a weak signal as flexibility at the ends of the helix would mean only 2 turns of helix were present. In the presence of TFE, the signal is significantly enhanced.

6.4.2 Secondary Structure Characterisation of WT and M154A α-Helical Peptides

Again, the key question to be answered by the NMR experiments was whether the isolated peptides are able to retain their native conformation as seen in the full length protein. The Hα chemical shift difference for the WT and M154A α-helical peptides shows that these isolated peptides are capable of retaining their native α-helical conformation (Figure 6.20). The upfield (negative) direction of Hα and Cβ

shifts strongly supports the presence of helical peptide conformation. However, ideally, a downfield shift (positive) for Cα is expected for helical structure and this was not the case with the WT and M154A peptide (Figure 6.21). It is believed that the reason for the upfield shift of Cα is conduction of the NMR experiments in water instead of using other chemical compounds (such as DSS, TMS, TSP or Dioxane) as a calibration standard.

Another objective of the NMR experiment was to characterise the effect of the mutation M154A in the WT peptide. Considering the data in Figure 6.20, no effect of the mutation can be seen, i.e. it did not appear to stabilise or destabilise the WT conformation.

6.4.3 Binding Analysis of Peptides with Antibody

The binding of the folded β-strand epitope peptides with Fab could not be proved using ELISA or SPR. In SPR, two different approaches were used to study the interaction between the peptides and Fab: (i) immobilisation of Fab on the sensor chip and allowing peptides in the mobile phase and (ii) immobilisation of biotinylated peptides on the sensor chip and Fab being in the mobile phase. The first approach did not work because the peptide was found to stick to the surface of the sensor chip. A number of possible solutions (described in Section 6.2.3.1) were applied, but none of them could stop the background binding of the peptides to the sensor chip. In addition to the solubility issues with this set of peptides, it was assumed that the high frequency of hydrophobic amino acids would have caused the high affinity of peptides for the sensor chip. Consequently, the alternative approach (i.e. immobilisation of peptides on the sensor chip) was applied to monitor the binding affinity of Fab and peptides. However, the results from this approach did not confirm binding and remain inconclusive. The low K_D values could be a result of insolubility

issues and biotinylation on multiple amino acids. However, the mass spectrometry data confirm the presence of peptide with only 1 biotin molecule. Although the results are not quantitative, they clearly suggest that the singly biotinylated peptide is the major species. The biotin attaches at amine groups so it can attach at the N-terminus, at lysine residues and less readily at arginine residues. It is assumed that in this peptide the N-terminus and lysine residue are labeled first and that the 3rd biotin is at one of the arginine residues. Biotinylation at multiple positions was confirmed by mass spectrometry (shown in Figure 6.10). However, it is not possible to know the position at which the single biotin was attached. It is likely to be a mixture of the three different sites which were observed. However this means that there will definitely have been a proportion of peptide which was singly biotinylated at the N-terminus, and it is expected that this population of peptide would still have been able to bind Fab.

The graphs in Figure 6.9 show an increase in the binding of Fab to the peptide on increasing the Fab concentration indicating specific binding. However, the increase in the binding could be due to the fact that the Fab concentration is so high that it is causing non-specific binding of the protein to the peptide or the chip surface rather than measuring a real interaction. In theory, almost any protein would start to show some binding if the concentration is increased sufficiently. Therefore, repeating the experiment with any other Fab could have confirmed the presence or absence of the non-specific binding problem. Interesting data could have been obtained for the WTG peptide (considering the CD data which confirmed the β-sheet structure). Limited time at UCB restricted the ability to obtain data for WTG.

Binding of extended α-helical peptides with the Fab was confirmed by ELISA which was used as a preliminary experiment to decide whether SPR should be car-

ried out. SPR provided positive binding data for all the peptides to compute the binding kinetics. The SPR data in Table 6.9 and Figures 6.26 and 6.27 show that there is a difference in the rate of association and dissociation among the different peptides. A significant difference in the kinetics of the WT and M154A suggests that the differences in orientation of M154A exposed residues in this peptide may have affected the binding with the antibody. This is in contrast to the NMR results which revealed no structural differences between WT and M154A. Cap2 on WT (WT-C2) showed improved binding compared with the WT whereas the caps on the mutant, M154A, showed a significant decrease in the values of binding kinetics. These results agree with the CD data which implies that more structure in WT-C2 may have contributed in an increase in its availibility to the antibody and the destabilised M154A-C2 conformational population may have resulted in decrease in its accessibilty to the antibody. In spite of the kinetic differences among these peptides, all were able to bind strongly with the antibody which proves the presence of native conformation in the isolated peptides.

6.5 Conclusions

This experimental work, on two different sets of epitopes, has given many structural insights and revealed the challenges of working with isolated peptides which require careful consideration for vaccine design. In the case of the β-strand peptide, its hydrophobic nature caused solubility issues which is a common problem when working with peptides in general. However, in future, this issue could be considered as an initial screening step particularly for studying epitopes as isolated peptides. In the case of the linear helical peptide which was quite hydrophilic, experiments (CD, SPR and NMR) worked well and generated useful results.

Chapter 7

Discussion and Conclusions

The main aim of the research in this thesis was to study the structures of B-cell epitopes and explore the associated conformational challenges in using them as immunogens for the design of epitope-based vaccines. In order to achieve this, a dataset of non-redundant epitope structures was needed, however, the extraction of epitopes from antigen surfaces required the availability of experimentally solved antibody-antigen complexes.

Chapter 3 describes the creation of a database (**AbDb**) which contains non-redundant sets of antibody-antigen structures. The antibody structural data available in **SAbDab** [96] could not be used because of the redundancy and, at that time, the lack of standard antibody numbering. A huge effort was put in to achieve a cleaner, non-redundant dataset. The database was extended to incorporate different sets of antibody structures depending on either the presence of types of antibody chains or types of bound antigens. To our knowledge, there is no other resource that provides this type of pre-numbered, non-redundant and well-classified data.

Chapter 4 describes the second stage of the project, a detailed structural analysis of epitopes which resulted in the charaterisation of epitope structural components and their shapes. The shape analysis of epitope regions provided information

about the overall structural composition, shapes and distribution of lengths of regions and number of fragments of epitope regions. This also provided examples of linear (or near-linear) epitopes that could then be analysed and used as potential immunogens. In addition, this analysis provided information about the probability of finding such linear (or near-linear) epitopes and provides information on non-linear epitopes containing two regions which could, in future, be stapled to provide an immunogen for vaccine design.

According to the analysis of 506 unique epitopes, about 5% of B-cell epitopes were found to be continuous. In Chapter 5, the conformational stability of continuous epitopes as isolated peptides was studied to determine the extent to which epitopes can be mimicked using isolated peptides. 10 epitope regions (five extended and five folded) were studied using molecular dynamics (MD) simulations. In order to explore or enhance the stability of isolated epitope regions, the effect of end-capping and hydrophobic to either glutamine or alanine mutations was examined in both the peptides. The ends of folded conformations were also stapled by either disulphide (cysteine) stapling or by the addition of a glycine linker (cyclisation) to help the epitopes retain their native conformation.

The MD analysis of extended epitope regions suggested that it is possible to stabilise the native conformation of isolated peptides using hydrophobic to glutamine or alanine mutations. The end-capping of the extended regions only showed destabilising effects in most of the examples. Interestingly, methionine to alanine mutations in three out of five examples showed an improvement in the stability during the simulation compared with the wild type sequence suggesting that the long hydrophobic side chain of methionine may have been involved in the collapse of the native conformation.

The analysis of folded epitope regions suggested that stapling and cyclisation can significantly enhance the stabilisation compared with the wild type. However, end-capping and hydrophobic to glutamine or alanine mutations did not show any significant improvement in conformational stability. Disulphide stapling involves side-chain cross linking of cysteine residues introduced at selected appropriate positions and the ends of the folded epitope regions were extended to introduce the cysteines. The use of this technique experimentally might introduce some problems as the peptides may polymerize rather than being stapled (or may do so over time). However, a review by Fairlie and Araujo [183] indicates that it has been done successfully. The possibilty of cysteine polymerisation can also be ruled out by applying cyclisation or other chemical stapling techniques that have been adopted for α-helix stabilisation. These approaches involve side-chain crosslinking via hydrocarbons [184], triazole [185], lactam [186] and azobenzene [187] staples.

Cyclisation of folded peptides was studied *in silico* using both disulphide bonds and glycine linkers demonstrating improved stability in all examples suggesting that most folded epitopes can be stapled successfully experimentally without first going through extensive molecular dynamics simulations.

Interestingly, terminal extension of the folded epitope regions without stapling also showed significant stability enhancement in the epitope region suggesting an alternative strategy for stabilising folded epitopes for elliciting an immune response.

Experimental studies were performed on sets of derivative peptides based on one extended α-helical and one folded β-sheet epitope. The experiments on the extended α-helical peptides revealed that they had the ability to retain near-native conformation as isolated peptides.

Comparison of computational and experimental results for the folded β-strand epitope

The WT folded epitope region (derived from 4WEB) contains 6 hydrophobic amino acids. A wide range of mutant peptides were simulated and results have been discussed in Section 5.3.1.5. Among all the studied mutants, WT, V637A, WTX and WTG were selected for experimental studies. The simulation results showed that the WT peptide spent about 95% of simulation time in the initial conformation and the presence of 6 hydrophobic amino acids did not cause the unfolding of the β-strand. The V637A mutant showed a 10% decrease in S-value which was unexpected since alanine is much less hydrophobic than valine alhough it does have a much lower β-strand propensity (Table 5.3). Owing to the extended ends, WTX and WTG contained 9 hydrophobic residues, and simulation showed a stable conformation for these as well.

During experimental studies on this set of peptides, a solubility issue in aqueous solution was discovered. The presence of hydrophobic amino acids was considered as the likely reason for this problem. However, the solubilty issue was adjustable (discussed in Section 6.2.1) and it was possible to collect CD spectra for all of the peptides. Among these, only WTG retained the native β-strand conformation. The binding of WT, V637 and WTX peptides with Fab remained inconclusive, and WTG could not be studied owing to time constrainsts. However, it was expected that this might show positive binding owing to its ability to retain native conformation as an isolated, but cyclised, peptide. Further experiments on this cyclised peptide are needed in future including mutating some of the hydrophobic residues in the extensions to the epitope region.

Comparison of computational and experimental results for the extended α-helical epitope

MD simulation of the extended epitope region (derived from 2W9E) and its mutant peptides showed that the M154A mutation significantly stabilised the peptide compared with the WT. The end-capped versions of both the WT and M154A did not perform better than the un-capped peptide (Section 5.3.3.1). Considering these results from MD simulations, WT, WT-C2, M154A, M154A-C1 and M154A-C2 were chosen to validate the effect of caps and mutation experimentally.

CD experiments showed that all of these extended helical peptides were able to retain their native helix conformation in the presence of 10% TFE. The initial experiments without using TFE showed that the WT-C2 had significantly higher helical structure compared with the WT which does not agree with the MD results where Cap2 did not have any significant effect on the WT conformation. The idea of end-capping was inspired from experimental studies of peptides and was applied computationally but the MD in the case of Cap2 on WT did not appear to predict the effects seen experimentally. Experimentally Cap2 on M154A (M154A-C2) showed a significant decrease in the helix fraction compared with M154A which agrees with the MD data. NMR on WT and M154A proved the retention of helix conformation of these isolated peptides (as predicted by the MD) but no structural differences resulting from the mutation were observed.

SPR data validate the binding affinity of these isolated peptides with Fab and mostly agree with the simulation results. The binding association kinetics of M154A were found to be significantly higher than the WT which agrees with the simulation data where M154A was found to be stable compared with the WT although this was not seen in the NMR. End-capping of the M154A mutant peptide,

did not prevent binding of the peptides to antibody either in the MD simulations or in the SPR experiments. However, during MD simulations of the isolated peptides, end-capping of M154A was observed to have a destabilizing effect. The SPR kinetics support this finding since, when compared with M154A, both M154A-C1 and M154A-C2 showed reduced on-rates (K_a) and increased off-rates (K_d) as would be expected for a peptide that spends less of its time in the required conformation for binding (Figure 6.27).

Compared with the binding kinetics of full length antigen protein described by Antonyuk et al. [182], the isolated peptides had a much stronger binding affinity. Again, these results (potentially the on-rates) suggests that they retain the native conformation in solution. Moreover, the end-capping and mutations did not disrupt the peptide-antibody interaction.

Overall, the experimental studies on the folded epitope regions were not quite as successful as on the extended regions, but the experiments have provided information about the potential problems that may occur during the study of peptides and revealed factors that need to be considered in selecting peptides. In particular, the hydrophobicity of the folded peptide was not considered when selecting it for laboratory experiments. A solubility issue arose owing to the presence of a high number of hydrophobic amino acids and the peptide-antibody binding experiments failed. However, the secondary structure characterisation experiment (CD) worked on three (out of four) peptides, and revealed that none of the peptides was able to maintain its native conformation except the glycine cyclised peptide which showed β-strand structure. Binding could not be validated on this owing to shortage of time at UCB. This suggests that even quite extensive MD simulation is unable to replicate the effects of a large number of hydrophobic amino acids on peptide structure.

The Grand average of hydropathicity (GRAVY) [196] is a measure of total hydrophobicity or hydrophilicity with positive values indicating hydrophobicity. The GRAVY score of -2.50 for the extended peptide shows that it is quite hydrophilic compared with the folded peptide which has a GRAVY score of 0.02. This suggests that GRAVY scores well below zero are important for isolated peptides to be useful and for MD simulation to be predictive of peptide behaviour.

Future work

As described in this chapter, there is a wide range of chemical stapling techniques, designed to stabilise α-helices in extended conformations. These techniques could be explored for helical extended peptides both by MD and experimentally.

Given that the NMR data for the helical extended epitope suggests that it retains its native conformation as an isolated peptide, it would be interesting to take this peptide forward to the immunisation stage in an animal model and see whether it successfully generates antibodies that cross react with and neutralise the native protein. Experimental studies on additional peptides, already studied by MD, should also be carried out and additional peptides should be selected for study.

As well as peptides of two β-strands, it would be interesting to carry out computational studies on folded epitopes that are comprised of two α-helices. Again, it would be interesting to explore cyclization and disulphide stapling in the simulations and take these forward into experimental testing.

Another interesting avenue of research would be to extend the analysis to discontinuous epitopes by joining separate regions (by linker peptides and/or stapling), simulating them computationally to study their conformational stability and take these forward to experiemental work.

Summary

This project has shown that the conformational stability of linear peptides can be predicted with some success using molecular dynamics simulations. This has been demonstrated by studying hydrophobic to alanine and glutamine mutations, disulphide stapling, glycine cyclisation and terminal extension. In particular, methionine to alanine mutations showed enhanced stabilty in extended peptides while stapling, cyclisation and terminal extension showed significant increase in stability of folded peptides.

Experimental work on the folded peptides had only very limited success owing to the solubility issues. These indicated that MD was not a good predictor of peptide behaviour when the hydrophobicity of the peptide was high. However, the success of the experimental work on the linear peptide, which was more hydrophilic, suggests that stapling of folded peptides that are more hydrophilic would be a successful strategy for maintaining native conformation for creating immunogens.

Appendix A

Structural Analysis of B-Cell Epitopes

Table A.1: Grouped data of the complete epitope (combined) dataset. Each cell shows the observed and expected values.

	R0R1	**R2**	**R3-R9**	**Total**
F0F1	36/31.73	79/83.74	108/107.53	223
F2-F16	36/40.27	111/106.26	136/136.47	283
Total	72	190	244	**506**

Table A.2: Grouped data of the single chain epitope dataset. Each cell shows the observed and expected values. The expected values have been calculated using observed from the combined data.

	R0R1	**R2**	**R3-R9**	**Total**
F0F1	36/33.01	74/72.44	99/99.04	209
F2-F16	35/33.01	104/101.79	166/124.71	255
Total	71	178	215	**464**

Table A.3: Grouped data of the multiple chain epitope dataset. Each cell shows the observed and expected values. The expected values have been calculated using observed from the combined data.

	R0R1	**R2**	**R3-R9**	**Total**
F0F1	0/2.99	5/6.56	9/8.96	14
F2-F16	1/2.99	7/9.21	20/11.29	28
Total	1	12	29	**42**

Table A.4: 3D contingency table showing the occurrence of 0–2, 0–5 and 0–6 number of folded, curved and extended regions. The abbreviations of E, C and F are used for extended, curved and folded and the associated number shows the number of respective shape regions. The grouping was performed on the shapes where observed counts were low. This resulted in the grouping of E3456, C345 and F12. The significance of individual under- or over-represented combinations of Folded, Curved and Extended indicated in bold was calculated using a 2x2x2 chi-squared test. A very low p-value was observed for each of these combination.

Folded	Curved	Extended	Observed	Expected	p-value
F0	**C0**	**E1**	**20**	**37.7**	$2.22\text{x}10^{-15}$
F0	**C0**	**E2**	**31**	**19.4**	0
F0	C0	E3456	15	7.4	-
F0	**C1**	**E0**	**18**	**37.9**	$8.04\text{x}10^{-12}$
F0	**C1**	**E1**	**62**	**42.9**	$1.92\text{x}10^{-5}$
F0	C1	E2	24	22.1	-
F0	C1	E3456	15	8.4	-
F0	**C2**	**E0**	**40**	**26.3**	$2.22\text{x}10^{-15}$
F0	**C2**	**E1**	**40**	**29.8**	$1.72\text{x}10^{-6}$
F0	C2	E2	13	15.3	-
F0	C2	E3456	1	5.8	-
F0	**C345**	**E0**	**28**	**14.5**	$2.88\text{x}10^{-15}$
F0	C345	E1	17	16.4	-
F0	C345	E2	4	8.4	-
F0	C345	E3456	1	3.2	-
F12	**C0**	**E0**	**36**	**13.7**	0
F12	C0	E1	22	15.5	-
F12	C0	E2	11	8	-
F12	C0	E3456	3	3	-
F12	C1	E0	19	15.6	-
F12	**C1**	**E1**	**13**	**17.6**	$1.92\text{x}10^{-5}$
F12	C1	E2	6	9.1	-
F12	C1	E3456	0	3.4	-
F12	C2	E0	12	10.8	-
F12	C2	E1	2	12.2	-
F12	C2	E2	1	6.3	-
F12	C2	E3456	0	2.4	-
F12	C345	E0	5	5.9	-
F12	C345	E1	3	6.7	-
F12	C345	E2	2	3.5	-
F12	C345	E3456	0	1.3	-

Table A.5: 3D contingency table showing the occurrence of 0–3, 0–6 and 0–8 number of helical, sheet and coiled regions. The abbreviations of H, S and C are used for helix, sheet and coil and the associated number shows the number of respective secondary structure regions. The grouping was performed on the regions where observed counts were low. This resulted in the grouping of H123, S23456 and C345678. The significance of individual under- or over-represented combinations of Helix, Strand and Coil indicated in bold was calculated using a 2x2x2 chi-squared test. A very low p-value was observed for each of these combination.

Helix	**Strand**	**Coil**	**Observed**	**Expected**	**p-value**
H0	S0	C1	40	59.8	-
H0	S0	C2	78	69.3	-
H0	**S0**	**C345678**	**66**	**45.3**	$1.11\text{x}10^{-16}$
H0	S1	C0	10	10.9	-
H0	**S1**	**C1**	**28**	**18.3**	$1.34\text{x}10^{-5}$
H0	S1	C2	23	21.2	-
H0	**S1**	**C345678**	**20**	**13.8**	$9.80\text{x}10^{-11}$
H0	S23456	C0	9	4.5	-
H0	S23456	C1	17	7.6	-
H0	S23456	C2	10	8.8	-
H0	S23456	C345678	0	5.7	-
H123	**S0**	**C0**	**50**	**19.4**	0
H123	S0	C1	39	32.4	-
H123	S0	C2	38	37.5	-
H123	**S0**	**C345678**	**13**	**24.5**	0
H123	S1	C0	6	5.9	-
H123	S1	C1	7	9.9	-
H123	S1	C2	4	11.5	-
H123	S1	C345678	1	7.5	-
H123	S23456	C0	4	2.5	-
H123	S23456	C1	1	4.1	-
H123	S23456	C2	0	4.7	-
H123	S23456	C345678	0	3.1	-

Appendix B

Experimental Studies of Epitope Regions

Table B.1: Chemical shifts for WT peptide. NMR experiments and assignments were performed with the enormous help of Christine Prosser and Leo Bowsher at UCB. The assignments 1–15 correspond to positions 142–156 in the peptide sequence.

Sequence	**Cα**	**Cβ**	**Cδ**	**Cε**	**Cγ**	**HN**	**Hα**	**Hβ2**	**Hβ3**	**Hε**	**Hγ2**	**Hγ3**	**N**	**NH2**	**Qβ**	**Qδ**	**Qε**	**Qγ**
G1	40.7	-	-	-	-	8.602	3.814	-	-	-	-	-	-	-	-	-	-	-
S2	55.46	61.28	-	-	-	8.604	4.427	3.775	3.728	-	-	-	115.8	-	-	-	-	-
D3	51.65	37.52	-	-	-	8.47	4.496	2.596	2.563	-	-	-	122.1	-	-	-	-	-
Y4	56.24	35.72	-	-	-	8.101	4.315	-	-	-	-	-	120.2	-	2.908	6.984	6.721	-
E5	54.13	26.61	-	-	31.75	8.052	4.082	1.952	1.868	-	-	-	121	-	-	-	-	2.207
D6	51.89	37.62	-	-	-	8.136	4.453	-	-	-	-	-	120.4	-	2.679	-	-	-
R7	54.8	27.52	40.67	-	24.07	8.005	4.001	-	-	7.036	-	-	120.5	-	1.564	2.947	-	1.286
Y8	55.9	35.59	-	-	-	7.964	4.336	2.906	2.759	-	-	-	119.4	-	-	6.903	6.687	-
Y9	56.16	35.8	-	-	-	7.857	4.312	-	-	-	-	-	120.7	-	2.947	7.022	6.731	-
R10	54.13	28	40.74	-	24.62	7.922	4.075	1.736	1.65	7.154	-	-	121.3	-	-	3.091	-	1.493
E11	54.66	26.62	-	-	31.99	8.103	4.088	1.96	1.885	-	-	-	119.8	-	-	-	-	2.288
N12	50.77	35.99	-	-	-	8.134	4.531	2.683	2.579	-	-	-	118.2	7.083	2.585	-	-	-
M13	52.92	29.95	14.17	-	29.22	7.964	4.295	1.902	1.829	1.954	2.399	2.334	119.7	-	-	-	-	-
H14	52.59	26.18	-	-	-	8.26	4.58	3.185	3.126	-	-	-	119.4	-	-	7.399	6.771	-
R15	54.67	28.52	40.71	-	24.57	8.045	4.089	1.767	1.617	7.084	-	-	127.6	-	-	3.088	-	1.5

Table B.2: Chemical shifts for M154A peptide. NMR experiments and assignments were performed with the enormous help of Christine Prosser and Leo Bowsher at UCB. The assignments 1–15 correspond to positions 142–156 in the peptide sequence.

Sequence	**Cα**	**Cβ**	**Cδ**	**Cγ**	**HN**	**Hα**	**Hβ2**	**Hβ3**	**Hε**	**N**	**Qβ**	**Qδ**	**Qγ**
G1	40.69	-	-	-	8.594	3.811	-	-	-	-	-	-	-
S2	55.45	61.26	-	-	8.585	4.423	3.766	3.723	-	115.7	-	-	-
D3	51.34	37.19	-	-	8.469	4.524	2.632	2.596	-	121.9	-	-	-
Y4	56.09	35.75	-	-	8.112	4.329	-	-	-	120.5	2.912	-	-
E5	53.89	26.4	-	31.25	8.051	4.095	1.958	1.868	-	120.8	-	-	2.224
D6	51.63	37.22	-	-	8.121	4.462	-	-	-	120	2.702	-	-
R7	54.59	27.56	40.66	24.05	7.989	4.018	-	-	7.004	120.5	1.555	2.972	1.28
Y8	55.77	35.67	-	-	7.947	4.359	2.905	2.753	-	119.4	-	-	-
Y9	56.06	35.76	-	-	7.841	4.329	-	-	-	120.9	2.932	-	-
R10	54.69	28	40.69	24.53	7.934	4.096	1.718	1.634	7.115	121.9	-	3.085	1.482
E11	-	26.39	-	31.42	8.114	4.128	1.972	1.909	-	120.5	-	-	2.314
N12	50.74	36.15	-	-	8.188	4.537	2.682	2.585	-	119	-	-	-
A13	49.91	16.44	-	-	7.977	4.154	-	-	-	123.6	1.214	-	-
H14	52.46	26.31	-	-	8.237	4.569	3.189	3.116	-	117.5	-	-	-
R15	54.63	28.33	40.66	24.52	8.056	4.102	1.775	1.628	7.076	127.4	-	3.089	1.504

Bibliography

[1] Henderson, D. A., 'Eradication: lessons from the past'. *Bulletin of the World Health Organization*, 76:17, 1998.

[2] Hovi, T., 'Inactivated poliovirus vaccine and the final stages of poliovirus eradication'. *Vaccine*, 19:2268–2272, 2001.

[3] John, T. J., 'The final stages of the global eradication of polio'. *The New England Journal of Medicine*, 343:806, 2000.

[4] Breman, J. G. and Arita, I., 'The confirmation and maintenance of smallpox eradication'. *The New England Journal of Medicine*, 303:1263–1273, 1980.

[5] Thompson, A. L. and Staats, H. F., 'Cytokines: the future of intranasal vaccine adjuvants'. *Journal of Immunology Research*, 2011, 2011.

[6] Petrovsky, N. and Aguilar, J. C., 'Vaccine adjuvants: current state and future trends'. *Immunology and Cell Biology*, 82:488–496, 2004.

[7] Nemchinov, L., Liang, T., Rifaat, M., Mazyad, H., Hadidi, A., and Keith, J., 'Development of a plant-derived subunit vaccine candidate against hepatitis C virus'. *Archives of Virology*, 145:2557–2573, 2000.

[8] Arthur, L. O., Pyle, S. W., Nara, P. L., Bess, J. W., Gonda, M. A., Kelliher, J. C., Gilden, R. V., Robey, W. G., Bolognesi, D. P., and Gallo, R. C., 'Sero-

logical responses in chimpanzees inoculated with human immunodeficiency virus glycoprotein (gp120) subunit vaccine'. *Proceedings of the National Academy of Sciences, USA*, 84:8583–8587, 1987.

[9] Sun, D., Seyer, J., Kovari, I., Sumrada, R., and Taylor, R., 'Localization of protective epitopes within the pilin subunit of the *Vibrio cholerae* toxin-coregulated pilus'. *Infection and Immunity*, 59:114–118, 1991.

[10] Purcell, A. W., McCluskey, J., and Rossjohn, J., 'More than one reason to rethink the use of peptides in vaccine design'. *Nature Reviews Drug Discovery*, 6:404–414, 2007.

[11] Gershoni, J. M., Roitburd-Berman, A., Siman-Tov, D. D., Freund, N. T., and Weiss, Y., 'Epitope mapping'. *BioDrugs*, 21:145–156, 2007.

[12] Irving, M. B., Pan, O., and Scott, J. K., 'Random-peptide libraries and antigen-fragment libraries for epitope mapping and the development of vaccines and diagnostics'. *Current Opinion in Chemical Biology*, 5:314–324, 2001.

[13] Westwood, O. M. and Hay, F. C. *Epitope mapping: a practical approach*. Oxford University Press, 2001.

[14] Theisen, D., Bouche, F., El Kasmi, K., von der Ahe, I., Ammerlaan, W., Demotz, S., and Muller, C., 'Differential antigenicity of recombinant polyepitope-antigens based on loop-and helix-forming B and T cell epitopes'. *Journal of Immunological Methods*, 242:145–157, 2000.

[15] Wu, T. T. and Kabat, E. A., 'An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implica-

tions for antibody complementarity'. *The Journal of Experimental Medicine*, 132:211–250, 1970.

[16] Boyd, S. D. and Joshi, S. A. 'High-throughput DNA sequencing analysis of antibody repertoires'. In *Antibodies for Infectious Diseases*, pages 345–362. American Society of Microbiology, 2015.

[17] Janeway, C. A., Travers, P., Walport, M., and Shlomchik, M. *Immunobiology: the immune system in health and disease. 2005*. Garland Science, 2017.

[18] Kabat, E. A., Te Wu, T., Foeller, C., Perry, H. M., and Gottesman, K. S. *Sequences of proteins of immunological interest*. DIANE publishing, 5^{th} edition, 1992.

[19] Chothia, C. and Lesk, A. M., 'Canonical structures for the hypervariable regions of immunoglobulins'. *Journal of Molecular Biology*, 196:901–917, 1987.

[20] Abhinandan, K. R. and Martin, A. C. R., 'Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains'. *Molecular Immunology*, 45:3832–3839, 2008.

[21] Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G., 'IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains'. *Developmental & Comparative Immunology*, 27:55–77, 2003.

[22] Honegger, A. and Pluckthun, A., 'Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool'. *Journal of Molecular Biology*, 309:657–670, 2001.

[23] Van Regenmortel, M., 'Synthetic peptide vaccines and the search for neutralization B-cell epitopes'. *Open Vaccine Journal*, 2:33–44, 2009.

[24] Getzoff, E. D., Tainer, J. A., Lerner, R. A., and Geysen, H. M., 'The chemistry and mechanism of antibody binding to protein antigens'. *Advances in Immunology*, 43:1–98, 1988.

[25] Cunningham, B. C. and Wells, J. A., 'Comparison of a structural and a functional epitope'. *Journal of Molecular Biology*, 234:554–563, 1993.

[26] Van Regenmortel, M., 'Antigenicity and immunogenicity of synthetic peptides'. *Biologicals*, 29:209–213, 2001.

[27] Haste Andersen, P., Nielsen, M., and Lund, O., 'Prediction of residues in discontinuous B-cell epitopes using protein 3D structures'. *Protein Science*, 15:2558–2567, 2006.

[28] Enshell-Seijffers, D., Denisov, D., Groisman, B., Smelyanski, L., Meyuhas, R., Gross, G., Denisova, G., and Gershoni, J. M., 'The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1'. *Journal of Molecular Biology*, 334:87–101, 2003.

[29] Villén, J., Oliveira, E. D., Núñez, J. I., Molina, N., Sobrino, F., and Andreu, D., 'Towards a multi-site synthetic vaccine to foot-and-mouth disease: addition of discontinuous site peptide mimic increases the neutralization response in immunized animals'. *Vaccine*, 22:3523–3529, 2004.

[30] Dakappagari, N. K., Lute, K. D., Rawale, S., Steele, J. T., Allen, S. D., Phillips, G., Reilly, R. T., and Kaumaya, P. T., 'Conformational HER-2/NEU B-cell epitope peptide vaccine designed to incorporate two native disulfide

bonds enhances tumor cell binding and antitumor activities'. *Journal of Biological Chemistry*, 280:54–63, 2005.

[31] Timmerman, P., Beld, J., Puijk, W. C., and Meloen, R. H., 'Rapid and quantitative cyclization of multiple peptide loops onto synthetic scaffolds for structural mimicry of protein surfaces'. *Chembiochem*, 6:821–824, 2005.

[32] Flower, D. R., 'Designing immunogenic peptides'. *Nature Chemical Biology*, 9:749–753, 2013.

[33] Ponomarenko, J. V. and Van Regenmortel, M. H., 'B-cell epitope prediction'. *Structural Bioinformatics*, pages 849–879, 2009.

[34] Kulkarni-Kale, U., Bhosle, S., and Kolaskar, A. S., 'CEP: a conformational epitope prediction server'. *Nucleic Acids Research*, 33:W168–W171, 2005.

[35] Lo, Y.-T., Pai, T.-W., Wu, W.-K., and Chang, H.-T., 'Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics'. *BMC Bioinformatics*, 14:S3, 2013.

[36] Moreau, V., Fleury, C., Piquer, D., Nguyen, C., Novali, N., Villard, S., Laune, D., Granier, C., and Molina, F., 'PEPOP: computational design of immunogenic peptides'. *BMC Bioinformatics*, 9:71, 2008.

[37] Zhao, L., Wong, L., Lu, L., Hoi, S. C., and Li, J., 'B-cell epitope prediction through a graph model'. *BMC Bioinformatics*, 13:S20, 2012.

[38] Ponomarenko, J., Bui, H.-H., Li, W., Fusseder, N., Bourne, P. E., Sette, A., and Peters, B., 'ElliPro: a new structure-based tool for the prediction of antibody epitopes'. *BMC Bioinformatics*, 9:514, 2008.

[39] Sun, J., Wu, D., Xu, T., Wang, X., Xu, X., Tao, L., Li, Y., and Cao, Z.-W., 'SEPPA: a computational server for spatial epitope prediction of protein antigens'. *Nucleic Acids Research*, 37:W612–W616, 2009.

[40] Sweredoski, M. J. and Baldi, P., 'PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure'. *Bioinformatics*, 24:1459–1460, 2008.

[41] Liu, Y., McNevin, J., Zhao, H., Tebit, D. M., Troyer, R. M., McSweyn, M., Ghosh, A. K., Shriner, D., Arts, E. J., McElrath, M. J., and Mullins, J. I., 'Evolution of human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitopes: fitness-balanced escape'. *Journal of Virology*, 81:12179–12188, 2007.

[42] Kolesanova, E. F., Sanzhakov, M. A., and Kharybin, O. N., 'Development of the schedule for multiple parallel difficult peptide synthesis on pins'. *International journal of peptides*, 2013, 2013.

[43] Epstein, J. E., Giersing, B., Mullen, G., Moorthy, V., and Richie, T. L., 'Malaria vaccines: are we getting closer?'. *Current Opinion in Molecular Therapeutics*, 9:12–24, 2007.

[44] Volpina, O. M., Gelfanov, V. M., Yarov, A., Surovoy, A., Chepurkin, A., and Ivanov, V., 'New virus-specific T-helper epitopes of foot-and-mouth disease viral VP1 protein'. *FEBS letters*, 337:315–315, 1994.

[45] Tarradas, J., Monsó, M., Munoz, M., Rosell, R., Fraile, L., Frías, M. T., Domingo, M., Andreu, D., Sobrino, F., and Ganges, L., 'Partial protection

against classical swine fever virus elicited by dendrimeric vaccine-candidate peptides in domestic pigs'. *Vaccine*, 29:4422–4429, 2011.

[46] Stanekova, Z., Kiraly, J., Stropkovska, A., Mikušková, T., Mucha, V., Kostolanskỳ, F., and Varečková, E., 'Heterosubtypic protective immunity against influenza A virus induced by fusion peptide of the hemagglutinin in comparison to ectodomain of M2 protein'. *Acta Virologica*, 55:61–67, 2010.

[47] Oscherwitz, J., Yu, F., and Cease, K. B., 'A synthetic peptide vaccine directed against the $2\beta2$–$2\beta3$ loop of domain 2 of protective antigen protects rabbits from inhalation anthrax'. *The Journal of Immunology*, 185:3661–3668, 2010.

[48] Solares, A. M., Baladron, I., Ramos, T., Valenzuela, C., Borbon, Z., Fanjull, S., Gonzalez, L., Castillo, D., Esmir, J., Granadillo, M., Batte, A., Cintado, A., Ale, M., Fernandez, D. C. M. E., Ferrer, A., Torrens, I., and Lopez-Saura, P., 'Safety and immunogenicity of a human papillomavirus peptide vaccine (CIGB-228) in women with high-grade cervical intraepithelial neoplasia: first-in-human, proof-of-concept trial'. *ISRN Obstetrics and Gynecology*, 2011, 2011.

[49] Bernhardt, S., Gjertsen, M., Trachsel, S., Møller, M., Eriksen, J., Meo, M., Buanes, T., and Gaudernack, G., 'Telomerase peptide vaccination of patients with non-resectable pancreatic cancer: a dose escalating phase I/II study'. *British Journal of Cancer*, 95:1474–1482, 2006.

[50] Brunsvig, P. F., Aamdal, S., Gjertsen, M. K., Kvalheim, G., Markowski-Grimsrud, C. J., Sve, I., Dyrhaug, M., Trachsel, S., Møller, M., Eriksen, J. A., and Gaudernack, G., 'Telomerase peptide vaccination: a phase I/II study in

patients with non-small cell lung cancer'. *Cancer Immunology, Immunotherapy*, 55:1553–1564, 2006.

[51] Kyte, J. A., Gaudernack, G., Dueland, S., Trachsel, S., Julsrud, L., and Aamdal, S., 'Telomerase peptide vaccination combined with temozolomide: a clinical trial in stage IV melanoma patients'. *Clinical Cancer Research*, 17:4568–4580, 2011.

[52] Füst, G., 'Enhancing antibodies in HIV infection'. *Parasitology*, 115:127–140, 1997.

[53] Bogdanos, D. P., Choudhuri, K., and Vergani, D., 'Molecular mimicry and autoimmune liver disease: virtuous intentions, malign consequences'. *Liver*, 21:225–232, 2001.

[54] Olson, J. K., Croxford, J. L., Calenoff, M. A., Dal Canto, M. C., and Miller, S. D., 'A virus-induced molecular mimicry model of multiple sclerosis'. *Journal of Clinical Investigation*, 108:311, 2001.

[55] Steere, A. C., Gross, D., Meyer, A. L., and Huber, B. T., 'Autoimmune mechanisms in antibiotic treatment-resistant Lyme arthritis'. *Journal of Autoimmunity*, 16:263–268, 2001.

[56] Trollmo, C., Meyer, A. L., Steere, A. C., Hafler, D. A., and Huber, B. T., 'Molecular mimicry in Lyme arthritis demonstrated at the single cell level: LFA-1αL is a partial agonist for outer surface protein A-reactive T-cells'. *The Journal of Immunology*, 166:5286–5291, 2001.

[57] Willett, T. A., Meyer, A. L., Brown, E. L., and Huber, B. T., 'An effective second-generation outer surface protein A-derived Lyme vaccine that elimi-

nates a potentially autoreactive T-cell epitope'. *Proceedings of the National Academy of Sciences, USA*, 101:1303–1308, 2004.

[58] Parren, P. W., Moore, J. P., Burton, D. R., and Sattentau, Q. J., 'The neutralizing antibody response to HIV-1: viral evasion and escape from humoral immunity'. *Aids*, 13:S137–S162, 1999.

[59] Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A., 'Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody'. *Nature*, 393:648–659, 1998.

[60] Wyatt, R., Kwong, P. D., Desjardins, E., Sweet, R. W., Robinson, J., Hendrickson, W. A., and Sodroski, J. G., 'The antigenic structure of the HIV gp120 envelope glycoprotein'. *Nature*, 393:705–711, 1998.

[61] Berkower, I., Smith, G. E., Giri, C., and Murphy, D., 'Human immunodeficiency virus 1. Predominance of a group-specific neutralizing epitope that persists despite genetic variation.'. *The Journal of Experimental Medicine*, 170:1681–1695, 1989.

[62] Berger, C., Weber-Bornhauser, S., Eggenberger, J., Hanes, J., Plückthun, A., and Bosshard, H. R., 'Antigen recognition by conformational selection'. *FEBS Letters*, 450:149–153, 1999.

[63] Bosshard, H. R., 'Molecular recognition by induced fit: how fit is the concept?'. *Physiology*, 16:171–173, 2001.

[64] RiNi, J. M., Schulze-Gahmen, U., and Wilson, I. A., 'Structural evidence for induced fit as a mechanism for antibody-antigen recognition'. *Science*, 255:959–965, 1992.

[65] Purcell, A. W., Zeng, W., Mifsud, N. A., Ely, L. K., Macdonald, W. A., and Jackson, D. C., 'Dissecting the role of peptides in the immune response: theory, practice and the application to vaccine design'. *Journal of Peptide Science*, 9:255–281, 2003.

[66] Davies, J. S., 'The cyclization of peptides and depsipeptides'. *Journal of Peptide Science*, 9:471–501, 2003.

[67] Atkins, P. W. and Friedman, R. S. *Molecular quantum mechanics*. Oxford University Press, 2011.

[68] Nikos, L. D., 'Molecular dynamics beyond the born-oppenheimer approximation: Mixed quantum — classical approaches'. *Computational Nanoscience*, pages 389–409, 2006.

[69] Steinbach, P. J. *Introduction to macromolecular simulation*. National Institutes of Health, 1999.

[70] Dinur, U. and Hagler, A. T., 'New approaches to empirical force fields'. *Reviews in Computational Chemistry*, 2:99–164, 1991.

[71] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A., 'A second generation force field for the simulation of proteins, nucleic acids, and organic molecules'. *Journal of the American Chemical Society*, 117:5179–5197, 1995.

[72] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. A., and Karplus, M., 'CHARMM: a program for macromolecular energy, minimization, and dynamics calculations'. *Journal of Computational Chemistry*, 4:187–217, 1983.

[73] Van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R., and Tironi, I. G., 'Biomolecular simulation: the GROMOS96 manual and user guide'. *Library Manual*, 1996.

[74] Jorgensen, W. L. and Tirado-Rives, J., 'The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin'. *Journal of the American Chemical Society*, 110:1657–1666, 1988.

[75] Lennard-Jones, J., 'On the forces between atoms and ions'. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 109:584–597, 1925.

[76] Jorgensen, W. L. and Tirado-Rives, J., 'Potential energy functions for atomic-level simulations of water and organic and biomolecular systems'. *Proceedings of the National Academy of Sciences, USA*, 102:6665–6670, 2005.

[77] Piana, S., Lindorff-Larsen, K., and Shaw, D. E., 'How robust are protein folding simulations with respect to force field parameterization?'. *Biophysical Journal*, 100:L47–L49, 2011.

[78] Verlet, L., 'Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules'. *Physical Review*, 159:98, 1967.

[79] Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M., 'LINCS: a linear constraint solver for molecular simulations'. *Journal of Computational Chemistry*, 18:1463–1472, 1997.

[80] Greenfield, N. J., 'Using circular dichroism spectra to estimate protein secondary structure'. *Nature Protocols*, 1:2876–2890, 2006.

[81] Kelly, S. M., Jess, T. J., and Price, N. C., 'How to study proteins by circular dichroism'. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1751:119–139, 2005.

[82] Sreerama, N. and Woody, R. W., 'Computation and analysis of protein circular dichroism spectra'. *Methods in Enzymology*, 383:318–351, 2004.

[83] Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D., '1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects'. *Journal of Biomolecular NMR*, 5:67–81, 1995.

[84] Schuck, P., 'Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules'. *Annual Review of Biophysics and Biomolecular Structure*, 26:541–566, 1997.

[85] Cooper, M. A., 'Optical biosensors in drug discovery'. *Nature Reviews. Drug discovery*, 1:515, 2002.

[86] Homola, J., Yee, S. S., and Gauglitz, G., 'Surface plasmon resonance sensors'. *Sensors and Actuators B: Chemical*, 54:3–15, 1999.

[87] Harrick, N. and Beckmann, K. 'Internal reflection spectroscopy'. In *Characterization of Solid Surfaces*, pages 215–245. Springer, 1974.

[88] Johnson, G. and Wu, T. T., 'Kabat database and its applications: future directions'. *Nucleic Acids Research*, 29:205–206, 2001.

[89] Retter, I., Althaus, H. H., Münch, R., and Müller, W., 'VBASE2, an integrative V gene database'. *Nucleic Acids Research*, 33:D671–D674, 2005.

[90] Martin, A. C. R. and Allen, J. 'Bioinformatics tools for antibody engineering'. In Duebel, S., editor, *Handbook of Therapeutic Antibodies Vol 1 (Technologies)*. Wiley-Blackwell, Weinheim, 1st edition, 2007.

[91] Swindells, M. B., Porter, C. T., Couch, M., Hurst, J., Abhinandan, K. R., Nielsen, J. H., Macindoe, G., Hetherington, J., and Martin, A. C. R., 'abYsis: Integrated antibody sequence and structure — management, analysis and prediction'. *Journal of Molecular Biology*, 429:356–364, 2017.

[92] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Ehrenmann, F., Lefranc, G., and Duroux, P., 'IMGT®, the international immunogenetics information system'. *Nucleic Acids Research*, 37:D1006–D1012, 2009.

[93] Chailyan, A., Tramontano, A., and Marcatili, P., 'A database of immunoglobulins with integrated tools: DIGIT'. *Nucleic Acids Research*, 40:D1230–D1234, 2011.

[94] Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., and Dunbrack, R. L., 'PyIgClassify: a database of antibody CDR structural classifications'. *Nucleic Acids Research*, 43:D432–D438, 2015.

[95] Allcorn, L. C. and Martin, A. C. R., 'SACS: Self-maintaining database of antibody crystal structure information'. *Bioinformatics*, 18:175–181, 2002.

[96] Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M., 'SAbDab: the structural antibody database'. *Nucleic Acids Research*, 42:D1140–D1146, 2014.

[97] Al-Lazikani, B., Lesk, A. M., and Chothia, C., 'Standard conformations for the canonical structures of immunoglobulins'. *Journal of Molecular Biology*, 273:927–948, 1997.

[98] Martin, A. C. R. and Thornton, J. M., 'Structural families of loops in homologous proteins: Automatic classification, modelling and application to antibodies'. *Journal of Molecular Biology*, 263:800–815, 1996.

[99] Abhinandan, K. R. and Martin, A. C. R., 'Analyzing the "degree of humanness" of antibody sequences'. *Journal of Molecular Biology*, 369:852–862, 2007.

[100] Sivalingam, G. N. and Shepherd, A. J., 'An analysis of B-cell epitope discontinuity'. *Molecular Immunology*, 51:304–309, 2012.

[101] Graille, M., Stura, E. A., Corper, A. L., Sutton, B. J., Taussig, M. J., Charbonnier, J.-B., and Silverman, G. J., 'Crystal structure of a *Staphylococcus aureus* protein A domain complexed with the Fab fragment of a human IgM antibody: structural basis for recognition of B-cell receptors and superantigen activity'. *Proceedings of the National Academy of Sciences, USA*, 97:5399–5404, 2000.

[102] Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S., and Foeller, C. *Sequences of Proteins of Immunological Interest*. U.S. Department of Health

and Human Services, National Institutes for Health, Bethesda, MD, Fifth edition, 1991.

[103] Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., and Poljak, R. J., 'Conformations of immunoglobulin hypervariable regions'. *Nature*, 342:877–883, 1989.

[104] Porter, C. T. and Martin, A. C. R., 'BiopLib and BiopTools — a C programming library and toolset for manipulating protein structure'. *Bioinformatics*, 31:4017–4019, 2015.

[105] Xing, Y., Oliver, S. L., Nguyen, T., Ciferri, C., Nandi, A., Hickman, J., Giovani, C., Yang, E., Palladino, G., Grose, C., Uematsu, y., Lilja, A. E., Arvin, A. M., and Carfi, A., 'A site of *Varicella zoster* virus vulnerability identified by structural studies of neutralizing antibodies bound to the glycoprotein complex gHgL'. *Proceedings of the National Academy of Sciences, USA*, 112:6056–6061, 2015.

[106] Martin, A. C. R., 'idabchain v2.5'. *UCL*, 2001-2016.

[107] Martin, A. C. R., 'hashapten v1.2'. *UCL*, 2015.

[108] Silverton, E. W., Padlan, E. A., Davies, D. R., Smith-Gill, S., and Potter, M., 'Crystalline monoclonal antibody Fabs complexed to hen egg white lysozyme'. *Journal of Molecular Biology*, 180:761–765, 1984.

[109] Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M.-S., Novotnỳ, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E., and Crea, R., '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, USA*, 85:5879–5883, 1988.

[110] Novotnỳ, J., Handschumacher, M., Haber, E., Bruccoleri, R. E., Carlson, W. B., Fanning, D. W., Smith, J. A., and Rose, G. D., 'Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains)'. *Proceedings of the National Academy of Sciences, USA*, 83:226–230, 1986.

[111] Lollier, V., Denery-Papini, S., Larré, C., and Tessier, D., 'A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures'. *Molecular immunology*, 48:577–585, 2011.

[112] Ofran, Y., Schlessinger, A., and Rost, B., 'Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B-cell epitopes'. *The Journal of Immunology*, 181:6230–6235, 2008.

[113] Rubinstein, N. D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J. M., and Pupko, T., 'Computational characterization of B-cell epitopes'. *Molecular Immunology*, 45:3477–3489, 2008.

[114] Zhao, L. and Li, J., 'Mining for the antibody-antigen interacting associations that predict the B-cell epitopes'. *BMC Structural Biology*, 10:S6, 2010.

[115] Sun, J., Xu, T., Wang, S., Li, G., Wu, D., and Cao, Z., 'Does difference exist between epitope and non-epitope residues?'. *Immunome Research*, 201:1–11, 2011.

[116] Liang, S., Zheng, D., Standley, D. M., Yao, B., Zacharias, M., and Zhang, C., 'EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results'. *BMC Bioinformatics*, 11:381, 2010.

[117] Thornton, J., Edwards, M., Taylor, W., and Barlow, D., 'Location of 'continuous' antigenic determinants in the protruding regions of proteins.'. *The EMBO Journal*, 5:409, 1986.

[118] Kringelum, J. V., Nielsen, M., Padkjær, S. B., and Lund, O., 'Structural analysis of B-cell epitopes in antibody: protein complexes'. *Molecular Immunology*, 53:24–34, 2013.

[119] Ponomarenko, J. V. and Bourne, P. E., 'Antibody-protein interactions: benchmark datasets and prediction tools evaluation'. *BMC Structural Biology*, 7:64, 2007.

[120] Martin, A. C. R., 'chaincontacts v1.3'. *UCL*, 1995-2015.

[121] Rubinstein, N. D., Mayrose, I., Martz, E., and Pupko, T., 'Epitopia: a web-server for predicting B-cell epitopes'. *BMC Bioinformatics*, 10:287, 2009.

[122] Kabsch, W. and Sander, C., 'How good are predictions of protein secondary structure?'. *FEBS letters*, 155:179–182, 1983.

[123] Martin, A. C. R., 'pdbsecstr v1.0'. *UCL*, 2016.

[124] Martin, A. C. R., Raghavan, A., and Ferdous, S., 'pdbline v1.2'. *UCL*, 2014.

[125] Lienert, G. and Wolfrum, C., 'Simplified formulas for three-way chi-square testing'. *Biometrical Journal*, 22:159–167, 1980.

[126] Lin, C. J. 'Analysis of three-way contingency table'. Technical report, 2006.

[127] Li, Q. *STAT 504 - Analysis of Discrete Data*. Penn State Science, 2012.

[128] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J., 'PatchDock and SymmDock: servers for rigid and symmetric docking'. *Nucleic Acids Research*, 33:W363–W367, 2005.

[129] Comeau, S. R., Gatchell, D. W., Vajda, S., and Camacho, C. J., 'ClusPro: an automated docking and discrimination method for the prediction of protein complexes'. *Bioinformatics*, 20:45–50, 2004.

[130] Rubinstein, N. D., Mayrose, I., and Pupko, T., 'A machine-learning approach for predicting B-cell epitopes'. *Molecular Immunology*, 46:840–847, 2009.

[131] Liang, S., Zheng, D., Zhang, C., and Zacharias, M., 'Prediction of antigenic epitopes on protein surfaces by consensus scoring'. *BMC Bioinformatics*, 10:302, 2009.

[132] Dytham, C. *Choosing and using statistics: a biologist's guide*. John Wiley & Sons, 2011.

[133] Studer, R. A., Dessailly, B. H., and Orengo, C. A., 'Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes'. *Biochemical Journal*, 449:581–594, 2013.

[134] Pikkemaat, M. G., Linssen, A. B., Berendsen, H. J., and Janssen, D. B., 'Molecular dynamics simulations as a tool for improving protein stability'. *Protein Engineering*, 15:185–192, 2002.

[135] Martin, A. C. R., 'mutmodel v1.17'. *UCL*, 1996-2011.

[136] Shih, H., Brady, J., and Karplus, M., 'Structure of proteins with single-site mutations: a minimum perturbation approach'. *Proceedings of the National Academy of Sciences, USA*, 82:1697–1700, 1985.

[137] Forood, B., Feliciano, E. J., and Nambiar, K. P., 'Stabilization of α-helical structures in short peptides via end capping.'. *Proceedings of the National Academy of Sciences, USA*, 90:838–842, 1993.

[138] Chakrabartty, A., Doig, A. J., and Baldwin, R. L., 'Helix capping propensities in peptides parallel those in proteins'. *Proceedings of the National Academy of Sciences, USA*, 90:11332–11336, 1993.

[139] Doig, A. J. and Baldwin, R. L., 'N-and C-capping preferences for all 20 amino acids in α-helical peptides'. *Protein Science*, 4:1325–1336, 1995.

[140] Doig, A. J., Stapley, B. J., Macarthur, M. W., and Thornton, J. M., 'Structures of N-termini of helices in proteins'. *Protein Science*, 6:147–155, 1997.

[141] Moelbert, S., Emberly, E., and Tang, C., 'Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins'. *Protein Science*, 13:752–762, 2004.

[142] Fasman, G. D. *Prediction of protein structure and the principles of protein conformation*. Springer Science & Business Media, 2012.

[143] Marqusee, S., Robbins, V. H., and Baldwin, R. L., 'Unusually stable helix formation in short alanine-based peptides'. *Proceedings of the National Academy of Sciences, USA*, 86:5286–5290, 1989.

[144] Marqusee, S. and Baldwin, R. L., 'Helix stabilization by Glu-... Lys+ salt bridges in short peptides of *de novo* design'. *Proceedings of the National Academy of Sciences, USA*, 84:8898–8902, 1987.

[145] Park, S. H., Shalongo, W., and Stellwagen, E., 'Residue helix parameters obtained from dichroic analysis of peptides of defined sequence'. *Biochemistry*, 32:7048–7053, 1993.

[146] Lau, Y. H., De Andrade, P., Wu, Y., and Spring, D. R., 'Peptide stapling techniques based on different macrocyclisation chemistries'. *Chemical Society Reviews*, 44:91–102, 2015.

[147] Tian, Y., Li, J., Zhao, H., Zeng, X., Wang, D., Liu, Q., Niu, X., Huang, X., Xu, N., and Li, Z., 'Stapling of unprotected helical peptides via photo-induced intramolecular thiol–yne hydrothiolation'. *Chemical Science*, 7:3325–3330, 2016.

[148] Zhou, N. E., Kay, C. M., and Hodges, R. S., 'Disulfide bond contribution to protein stability: Positional effects of substitution in the hydrophobic core of the two-stranded α-helical coiled-coil'. *Biochemistry*, 32:3178–3187, 1993.

[149] He, H. T., Gürsoy, R. N., Kupczyk-Subotkowska, L., Tian, J., Williams, T., and Siahaan, T. J., 'Synthesis and chemical stability of a disulfide bond in a model cyclic pentapeptide: Cyclo (1, 4)-cys-gly-phe-cys-gly-OH'. *Journal of Pharmaceutical Sciences*, 95:2222–2234, 2006.

[150] Hazes, B. and Dijkstra, B. W., 'Model building of disulfide bonds in proteins with known three-dimensional structure'. *Protein Engineering, Design and Selection*, 2:119–125, 1988.

[151] Martin, A. C. R., 'sssearch v1.2'. *SciTech Software, DKfz*, 1993-1996.

[152] Martin, A. C. R., 'ssbond v2.2'. *LMB Oxford, DKfz*, 1989-2015.

[153] Martin, A. C. R., 'addlinker v1.0'. *UCL*, 2016.

[154] Azuara, C., Lindahl, E., Koehl, P., Orland, H., and Delarue, M., 'PDB_Hydro: incorporating dipolar solvents with variable density in the Poisson–Boltzmann treatment of macromolecule electrostatics'. *Nucleic Acids Research*, 34:W38–W42, 2006.

[155] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E., 'GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers'. *SoftwareX*, 1:19–25, 2015.

[156] Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B., and Lindahl, E., 'GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit'. *Bioinformatics*, pages 854–854, 2013.

[157] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J., 'Gromacs: fast, flexible, and free'. *Journal of Computational Chemistry*, 26:1701–1718, 2005.

[158] Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E., 'Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation'. *Journal of Chemical Theory and Computation*, 4:435–447, 2008.

[159] Berendsen, H. J., van der Spoel, D., and van Drunen, R., 'Gromacs: a message-passing parallel molecular dynamics implementation'. *Computer Physics Communications*, 91:43–56, 1995.

[160] Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E., 'Systematic validation of protein force fields against experimental data'. *PloS One*, 7:e32131, 2012.

[161] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C., 'Comparison of multiple Amber force fields and development of improved protein backbone parameters'. *Proteins: Structure, Function, and Bioinformatics*, 65:712–725, 2006.

[162] Best, R. B. and Hummer, G., 'Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides'. *The Journal of Physical Chemistry B*, 113:9004–9015, 2009.

[163] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J., Dror, R., and Shaw, D., 'Improved side-chain torsion potentials for the Amber ff99sb protein force field'. *Proteins*, 78:1950–8, 2012.

[164] Rauscher, S., Gapsys, V., Gajda, M. J., Zweckstetter, M., de Groot, B. L., and Grubmuller, H., 'Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment'. *Journal of Chemical Theory and Computation*, 11:5513–5524, 2015.

[165] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L., 'Comparison of simple potential functions for simulating liquid water'. *The Journal of Chemical Physics*, 79:926–935, 1983.

[166] Darden, T., York, D., and Pedersen, L., 'Particle Mesh Ewald: An n log (N) method for Ewald sums in large systems'. *The Journal of Chemical Physics*, 98:10089–10092, 1993.

[167] Berendsen, H. J., Postma, J. v., Van Gunsteren, W. F., DiNola, A., and Haak, J., 'Molecular dynamics with coupling to an external bath'. *The Journal of Chemical Physics*, 81:3684–3690, 1984.

[168] Parrinello, M. and Rahman, A., 'Polymorphic transitions in single crystals: A new molecular dynamics method'. *Journal of Applied Physics*, 52:7182–7190, 1981.

[169] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J., 'Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes'. *Journal of Computational Physics*, 23:327–341, 1977.

[170] Hoover, W. G., 'Canonical dynamics: equilibrium phase-space distributions'. *Physical Review A*, 31:1695, 1985.

[171] Nosé, S., 'A unified formulation of the constant temperature molecular dynamics methods'. *The Journal of Chemical Physics*, 81:511–519, 1984.

[172] Carbone, F. *doitGROMACS V1.6: Script to execute a bunch of stuff with gromacs*, 2013-2016.

[173] Langdon, W. B., 'Initial experiences of the Emerald: e-infrastructure south GPU supercomputer'. *Research Note*, 12:08, 2012.

[174] Kabsch, W. and Sander, C., 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features'. *Biopolymers*, 22:2577–2637, 1983.

[175] Frishman, D. and Argos, P., 'Knowledge-based protein secondary structure assignment'. *Proteins: Structure, Function, and Bioinformatics*, 23:566–579, 1995.

[176] Fodje, M. and Al-Karadaghi, S., 'Occurrence, conformational features and amino acid propensities for the π-helix'. *Protein Engineering*, 15:353–358, 2002.

[177] Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., Macpherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., Klevit, R. E., Graham, B. S., Wyatt, R. T., Baker, D., Strong, R. K., Crowe, J. E., Johnson, P. R., and Schief, W. R., 'Proof of principle for epitope-focused vaccine design'. *Nature*, 507:201, 2014.

[178] Correia, B. E., Ban, Y.-E. A., Holmes, M. A., Xu, H., Ellingson, K., Kraft, Z., Carrico, C., Boni, E., Sather, D. N., Zenobia, C., Burke, K. Y., Bradley-Hewitt, T., Bruhn-Johannsen, J. F., Kalyuzhniy, O., Baker, D., Strong, R. K., Stamatatos, L., and Schief, W. R., 'Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope'. *Structure*, 18:1116–1126, 2010.

[179] Jiang, Y., Lee, A., Chen, J., Ruta, V., Martine, C., Brian, T. C., and Roderick, M., 'X-ray structure of a voltage-dependent K+ channel'. *Nature*, 423:33, 2003.

[180] Chen, E., Paing, M. M., Salinas, N., Sim, B. K. L., and Tolia, N. H., 'Structural and functional basis for inhibition of erythrocyte invasion by antibodies that target *Plasmodium falciparum* EBA-175'. *PLoS Pathogens*, 9:e1003390, 2013.

[181] Khan, A. G., Whidby, J., Miller, M. T., Scarborough, H., Zatorski, A. V., Cygan, A., Price, A. A., Yost, S. A., Bohannon, C. D., Jacob, J., A, G., and J, M., 'Structure of the core ectodomain of the hepatitis C virus envelope glycoprotein 2'. *Nature*, 509:381–384, 2014.

[182] Antonyuk, S., Trevitt, C., Strange, R., Jackson, G., Sangar, D., Batchelor, M., Cooper, S., Fraser, C., Jones, S., Georgiou, T., Khalili-Shirazi, A., Clarke, A. R., Hasnain, S. S., and Collinge, J., 'Crystal structure of human prion protein bound to a therapeutic antibody'. *Proceedings of the National Academy of Sciences, USA*, 106:2554–2558, 2009.

[183] Fairlie, D. P. and Dantas de Araujo, A., 'Stapling peptides using cysteine crosslinking'. *Peptide Science*, 106:843–852, 2016.

[184] Schafmeister, C. E., Po, J., and Verdine, G. L., 'An all-hydrocarbon crosslinking system for enhancing the helicity and metabolic stability of peptides'. *Journal of the American Chemical Society*, 122:5891–5892, 2000.

[185] Tan, Y. S., Lane, D. P., and Verma, C. S., 'Stapled peptide design: principles and roles of computation'. *Drug Discovery Today*, 21:1642–1653, 2016.

[186] Taylor, J. W., 'The synthesis and study of side-chain lactam-bridged peptides'. *Peptide Science*, 66:49–75, 2002.

[187] Kumita, J. R., Smart, O. S., and Woolley, G. A., 'Photo-control of helix content in a short peptide'. *Proceedings of the National Academy of Sciences, USA*, 97:3803–3808, 2000.

[188] Lavanchy, D., 'Evolving epidemiology of hepatitis C virus'. *Clinical Microbiology and Infection*, 17:107–115, 2011.

[189] Scarselli, E., Ansuini, H., Cerino, R., Roccasecca, R. M., Acali, S., Filocamo, G., Traboni, C., Nicosia, A., Cortese, R., and Vitelli, A., 'The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus'. *The EMBO Journal*, 21:5017–5025, 2002.

[190] Sautto, G., Tarr, A. W., Mancini, N., and Clementi, M., 'Structural and antigenic definition of hepatitis C virus E2 glycoprotein epitopes targeted by monoclonal antibodies'. *Clinical and Developmental Immunology*, 2013, 2013.

[191] Myers, J. K., Pace, C. N., and Scholtz, J. M., 'Helix propensities are identical in proteins and peptides'. *Biochemistry*, 36:10923–10929, 1997.

[192] Braunschweiler, L. and Ernst, R., 'Coherence transfer by isotropic mixing: application to proton correlation spectroscopy'. *Journal of Magnetic Resonance (1969)*, 53:521–528, 1983.

[193] Macura, S. and Ernst, R., 'Elucidation of cross relaxation in liquids by two-dimensional NMR spectroscopy'. *Molecular Physics*, 41:95–117, 1980.

[194] Bodenhausen, G. and Ruben, D. J., 'Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy'. *Chemical Physics Letters*, 69:185–189, 1980.

[195] Wishart, D., Sykes, B., and Richards, F., 'The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy'. *Biochemistry*, 31:1647–1651, 1992.

[196] Kyte, J. and Doolittle, R. F., 'A simple method for displaying the hydropathic character of a protein'. *Journal of Molecular Biology*, 157:105–132, 1982.

[197] AnaSpec, I. 'Peptide solubility guidelines'. `https://www.anaspec.com/content/pdfs/PeptidesolubilityguidelinesFinal.pdf`.

[198] Greenfield, N. J. and Fasman, G. D., 'Computed circular dichroism spectra for the evaluation of protein conformation'. *Biochemistry*, 8:4108–4116, 1969.

[199] Holzwarth, G. and Doty, P., 'The ultraviolet circular dichroism of polypeptides'. *Journal of the American Chemical Society*, 87:218–228, 1965.

[200] Tanizaki, S., Clifford, J., Connelly, B. D., and Feig, M., 'Conformational sampling of peptides in cellular environments'. *Biophysical Journal*, 94:747–759, 2008.

[201] Fort, A. G. and Spray, D. C., 'Trifluoroethanol reveals helical propensity at analogous positions in cytoplasmic domains of three connexins'. *Peptide Science*, 92:173–182, 2009.

[202] Hanazono, Y., Takeda, K., and Miki, K., 'Structural studies of the N-terminal fragments of the WW domain: Insights into co-translational folding of a β-sheet protein'. *Scientific Reports*, 6, 2016.

[203] Kemmink, J. and Creighton, T. E., 'Effects of trifluoroethanol on the conformations of peptides representing the entire sequence of bovine pancreatic trypsin inhibitor'. *Biochemistry*, 34:12630–12635, 1995.

[204] Yang, J. J., Buck, M., Pitkeathly, M., Kotik, M., Haynie, D. T., Dobson, C. M., and Radford, S. E., 'Conformational properties of four peptides spanning the sequence of hen lysozyme'. *Journal of Molecular Biology*, 252:483–491, 1995.

[205] Schönbrunner, N., Wey, J., Engels, J., Georg, H., and Kiefhaber, T., 'Native-like β-structure in a trifluoroethanol-induced partially folded state of the all-β-sheet protein tendamistat'. *Journal of Molecular Biology*, 260:432–445, 1996.

[206] Nordén, B., Rodger, A., and Dafforn, T. *Linear dichroism and circular dichroism: a textbook on polarized-light spectroscopy*. Royal Society of Chemistry, Cambridge, UK, 2010.

[207] Bukovsky, E. V. *Fluorinated materials synthesis and characterization for energy storage and energy conversion applications*. PhD thesis, Colorado State University, 2015.

[208] Case, D. A., Dyson, H. J., and Wright, P. E., 'Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies on peptides and proteins'. *Methods in enzymology*, 239:392–416, 1994.

[209] Sharman, G. J., Griffiths-Jones, S. R., Jourdan, M., and Searle, M. S., 'Effects of amino acid ϕ, ψ propensities and secondary structure interactions in

modulating Hα chemical shifts in peptide and protein β-sheet'. *Journal of the American Chemical Society*, 123:12318–12324, 2001.

[210] Wishart, D. S. and Sykes, B. D., 'The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data'. *Journal of Biomolecular NMR*, 4:171–180, 1994.

[211] Santiveri, C. M., Rico, M., and Jiménez, M. A., '13Cα and 13Cβ chemical shifts as a tool to delineate β-hairpin structures in peptides'. *Journal of Biomolecular NMR*, 19:331–345, 2001.

[212] Ramirez-Alvarado, M., Blanco, F. J., Niemann, H., and Serrano, L., 'Role of β-turn residues in β-hairpin formation and stability in designed peptides'. *Journal of Molecular Biology*, 273:898–912, 1997.

[213] Reiersen, H. and Rees, A. R., 'Trifluoroethanol may form a solvent matrix for assisted hydrophobic interactions between peptide side chains'. *Protein Engineering*, 13:739–743, 2000.

[214] Shibata, A., Yamamoto, M., Yamashita, T., Chiou, J. S., Kamaya, H., and Ueda, I., 'Biphasic effects of alcohols on the phase transition of poly (L-lysine) between α-helix and β-sheet conformations'. *Biochemistry*, 31:5728–5733, 1992.

[215] Searle, M. S., Zerella, R., Williams, D. H., and Packman, L. C., 'Native-like β-hairpin structure in an isolated fragment from ferredoxin: NMR and CD studies of solvent effects on the N-terminal 20 residues'. *Protein Engineering*, 9:559–565, 1996.

[216] Roccatano, D., Colombo, G., Fioroni, M., and Mark, A. E., 'Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: a molecular dynamics study'. *Proceedings of the National Academy of Sciences, USA*, 99:12179–12184, 2002.

[217] Alexandrescu, A. T., Ng, Y.-L., and Dobson, C. M., 'Characterization of a trifluoroethanol-induced partially folded state of α-lactalbumin'. *Journal of Molecular Biology*, 235:587–599, 1994.

[218] Goodman, M. and Listowsky, I., 'Conformational aspects of synthetic polypeptides. VI. hypochromic spectral studies of oligo-γ-methyl-l-glutamate peptides'. *Journal of the American Chemical Society*, 84:3770–3771, 1962.

[219] Olofsson, S. and Baltzer, L., 'Structure and dynamics of a designed helix-loop-helix dimer in dilute aqueous trifluoroethanol solution. A strategy for NMR spectroscopic structure determination of molten globules in the rational design of native-like proteins'. *Folding and Design*, 1:347–356, 1996.

[220] Lehrman, S. R., Tuls, J. L., and Lund, M., 'Peptide α-helicity in aqueous trifluoroethanol: correlations with predicted α-helicity and the secondary structure of the corresponding regions of bovine growth hormone'. *Biochemistry*, 29:5590–5596, 1990.

[221] Luidens, M. K., Figge, J., Breese, K., and Vajda, S., 'Predicted and trifluoroethanol-induced α-helicity of polypeptides'. *Biopolymers*, 39:367–376, 1996.

[222] Sonnichsen, F., Van Eyk, J., Hodges, R., and Sykes, B., 'Effect of trifluoroethanol on protein secondary structure: an NMR and CD study using a synthetic actin peptide'. *Biochemistry*, 31:8790–8798, 1992.

[223] Howard, M. J. and Smales, C. M., 'NMR analysis of synthetic human serum albumin α-helix 28 identifies structural distortion upon amadori modification'. *Journal of Biological Chemistry*, 280:22582–22589, 2005.

[224] Fregeau Gallagher, N. L., Sailer, M., Niemczura, W. P., Nakashima, T. T., Stiles, M. E., and Vederas, J. C., 'Three-dimensional structure of leucocin A in trifluoroethanol and dodecylphosphocholine micelles: Spatial location of residues critical for biological activity in type IIa bacteriocins from lactic acid bacteria'. *Biochemistry*, 36:15062–15072, 1997.

[225] Najbar, L. V., Craik, D. J., Wade, J. D., Salvatore, D., and McLeish, M. J., 'Conformational analysis of LYS (11–36), a peptide derived from the β-sheet region of T4 lysozyme, in TFE and SDS'. *Biochemistry*, 36:11525–11533, 1997.

[226] Kumar, S., Modig, K., and Halle, B., 'Trifluoroethanol-induced β–α transition in β-lactoglobulin: hydration and cosolvent binding studied by 2H, 17o, and 19F magnetic relaxation dispersion'. *Biochemistry*, 42:13708–13716, 2003.